

UNITED STATES PATENT AND TRADEMARK OFFICE

BEFORE THE PATENT TRIAL AND APPEAL BOARD

MICROSOFT CORPORATION

Petitioner

v.

BRADIUM TECHNOLOGIES LLC

Patent Owner

CASE: To Be Assigned

Patent No. 8,924,506 B2

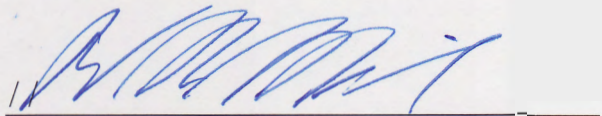
**DECLARATION OF PROF. WILLIAM R. MICHALSON
IN SUPPORT OF PETITION FOR *INTER PARTES* REVIEW
OF U.S. PATENT NO. 8,924,506 B2**

DECLARATION OF PROF. WILLIAM R. MICHALSON
IN SUPPORT OF PETITION FOR INTER PARTES REVIEW
OF U.S. PATENT NO. 8,924,506 B2

I hereby declare that all the statements made in this Declaration are of my own knowledge and true; that all statements made on information and belief are believed to be true; and further that these statements were made with the knowledge that willful false statements and the like so made are punishable by fine or imprisonment, or both, under 18 U.S.C. 1001 and that such willful false statements may jeopardize the validity of the application or any patent issued thereon.

I declare under the penalty of perjury that all statements made in this Declaration are true and correct.

Executed 1/11/ 2016 in DOUGLAS, MA.



William R. Michalson

TABLE OF CONTENTS

	Page
LIST OF APPENDICES	iii
I. INTRODUCTION	1
II. SUMMARY OF OPINIONS	3
III. QUALIFICATIONS AND EXPERIENCE.....	5
A. Education and Work Experience.....	5
B. Compensation.....	8
C. Documents and Other Materials Relied Upon	9
IV. STATEMENT OF LEGAL PRINCIPLES.....	9
A. Claim Construction.....	9
B. Anticipation	10
C. Obviousness.....	10
V. LEVEL OF ORDINARY SKILL IN THE ART	11
VI. TECHNOLOGY BACKGROUND OF THE 506 PATENT	13
A. Data Communications Over the Internet	15
B. Data Communications in Wireless Mobile Systems.....	18
C. Image Tiles and Image Pyramids.....	19
D. Compression of Image Tiles	24
E. Progressive Image Resolution Enhancement.....	25
F. Three-Dimensional Graphics	27
1. Overview of 3D Computer Graphics principles	27
2. Texture	33
3. Virtual Reality Modeling Language (VRML).....	37
G. Mip-Maps	38
H. Storage of image data.....	44
VII. OVERVIEW OF THE 506 PATENT.....	46
VIII. IDENTIFICATION OF THE PRIOR ART AND SUMMARY OF OPINIONS.....	51
A. Reddy.....	52

TABLE OF CONTENTS
(continued)

	Page
B. Hornbacker	53
IX. CLAIM CONSTRUCTION	53
A. “Data Parcel”	54
B. “A mesh” in claim 13	54
C. Other claim terms.....	55
X. UNPATENTABILITY OF THE 506 PATENT CLAIMS.....	55
A. GROUND 1: CLAIMS 1-21 ARE UNPATENTABLE UNDER 35 U.S.C. § 103(a) AS BEING OBVIOUS OVER REDDY IN VIEW OF HORNBACKER.....	55
1. Overview of Reddy and Hornbacker	57
2. Overview of Hornbacker.....	63
3. Motivations to Combine Reddy and Hornbacker.	64
4. Claim 1 is Rendered Obvious by Reddy in view of Hornbacker.....	70
5. Claims 2-7 are Rendered Obvious by Reddy in view of Hornbacker.....	90
6. Claim 8 is Rendered Obvious by Reddy in view of Hornbacker.....	103
7. Claims 9-14 are Rendered Obvious by Reddy in view of Hornbacker.....	115
8. Claim 15 is Rendered Obvious Over Reddy in view of Hornbacker.....	117
9. Claims 16-21 are rendered obvious over Reddy in view of Hornbacker	122

LIST OF APPENDICES

- Appendix A Curriculum Vitae of William R. Michalson
- Appendix B Excerpt of Hanan Samet, *The Design and Analysis of Spatial Data Structures*, University of Maryland (1989)
- Appendix C U.S. Patent No. 5,263,136 (DeAguiar et al)
- Appendix D U.S. Patent 4,972,319 (Delorme)
- Appendix E B. Fuller and I. Richer, *The MAGIC Project: From Vision to Reality*, IEEE Network May/June 1996, pp. 15-25
- Appendix F International Telegraph and Telephone Consultative Committee (“CCITT”) Recommendation T.81, September 1992
- Appendix G Ken Cabeen & Peter Gent, *Image Compression and the Discrete Cosine Transform*
- Appendix H M. Antonini, *Image Coding Using Wavelet Transform*, IEEE Transactions on Image Processing, Vol. 1, No. 2, April 1992.
- Appendix I U.S. Patent No. 5,321,520 (Inga et al)
- Appendix J U.S. Patent No. 6,182,114 (Yap et al.)
- Appendix K U.S. Patent No. 5,179,638 (Dawson et al)
- Appendix L Lance Williams, *Pyramidal Parametrics*, Computer Graphics, vol. 17, no. 3, July 1983
- Appendix M OpenGL Standard Version 1.1, March 1997, available:
<https://www.opengl.org/documentation/specs/version1.1/glspec1.1/node84.html#SECTION00681100000000000000>
- Appendix N H. Hoppe, *Progressive Meshes*, SIGGRAPH '96: Proceedings of the 23rd annual conference on computer graphics and interactive techniques, pp. 99-108
- Appendix O U.S. Patent 5,798,770 (Baldwin)
- Appendix P U.S. Patent No. 5,987,256 (Wu et al)

- Appendix R Boris Rabinovich & Craig Gotsman, *Visualization of Large Terrains in Resource-Limited Computing Environments* (1997)
- Appendix S *User Datagram Protocol (UDP)* (Windows CE 5.0, Microsoft, Available: <https://msdn.microsoft.com/en-us/library/ms885773.aspx> [Accessed April 28, 2015])
- Appendix T OpenGL Standard Version 1.2.1, April 1999, available: <https://www.opengl.org/documentation/specs/version1.2/opengl1.2.1.pdf>
- Appendix X George H. Forman and John Zahorjan, "The challenges of mobile computing," *Computer* vol. 27, no. 4, pp. 38, 47 (April 1994)
- Appendix Y K. Brown and S. Singh, *A Network Architecture for Mobile Computing*, INFOCOM '96, Fifteenth Annual Joint Conference of the IEEE Computer Societies, Networking the Next Generation, Proceedings IEEE vol. 3, pp. 1388-139
- Appendix Z Kreller, B. et al "UMTS: a middleware architecture and mobile API approach," *Personal Communications, IEEE*, vol. 5, no. 2, pp. 32-38 (April 1998)
- Appendix AA Hansen, J. et al, "Real-time synthetic vision cockpit display for general aviation," *AeroSense '99*, International Society for Optics and Photonics, 1999
- Appendix BB U.S. Patent No. 5,760,783 to Migdal et al ("Migdal")
- Appendix EE Theresa-Marie Rhyne, *A Commentary on GeoVRML: A Tool for 3D Representation of GeoReferenced Data on the Web*, *International Journal of Geographic Information Sciences*, issue 4 of volume 13, 1999
- Appendix GG GeoTIFF Format Specification Revision 1.0
- Appendix HH TIFF Revision 6.0, dated June 3, 1992.
- Appendix II FlashPix Format Specification v1.0, dated September 11, 1996
- Appendix KK The Virtual Reality Modeling Language ISO/IEC 14772-1:1997

I. INTRODUCTION

1. My name is William R. Michalson. I am a faculty member at Worcester Polytechnic Institute. I have been engaged by Microsoft Corporation (“Microsoft”) to investigate and opine on certain issues relating to U.S. Patent No. 8,924,506 B2 (the “506 Patent”) entitled “System and methods for network image delivery with dynamic viewing frustum optimized for limited bandwidth communication channels” in Microsoft’s Petition for *Inter Partes* Review of the 506 Patent (“Microsoft IPR Petition”) which requests the Patent Trial and Appeal Board (“PTAB”) to review and cancel all claims of the 506 Patent—claims 1-21 (“Challenged Claims”). I have also been engaged by Microsoft to investigate and opine on certain issues relating to two other patents that are related to the 506 Patent—U.S. Patent Nos. 7,908,343 B2 and 7,139,794 B2—in additional petitions for *inter partes* review by Microsoft. I understand that Bradium Technologies LLC (“Bradium”) is asserting all three patents against Microsoft in an on-going patent infringement lawsuit, No. 1:15-cv-00031-RGA, filed in the U.S. District Court for the District of Delaware on January 9, 2015.

2. I understand that the 506 Patent was assigned from the inventors Isaac Levanon and Yoni Lavi to Inovo Limited on April 3, 2011, and assigned from Inovo Limited to Bradium on June 17, 2013. Bradium is therefore referred to as the “Patent Owner” in this declaration.

3. In this declaration, I will first discuss the technology background related to the 506 Patent and then provide my analyses and opinions on claims 1-21 of the 506 Patent. The discussion of the technology background includes an overview of that technology as it was known before October 1999, which I understand as the earliest invention date of the 506 Patent claimed by the inventors in their inventor declarations submitted to the USPTO during the original prosecution of the 506 Patent's grand-parent patent, U.S. Patent No. 7,644,131. This overview provides some of the bases for my opinions with respect to the 506 Patent.

4. This declaration is based on the information currently available to me. To the extent that additional information becomes available, I reserve the right to continue my investigation and study, which may include a review of documents and information that may be produced, as well as testimony from depositions that may not yet be taken.

5. In forming my opinions, I have relied on information and evidence identified in this declaration, including the 506 Patent, the prosecution history of the 506 Patent, and prior art references listed as Exhibits to the Microsoft IPR Petition and listed as appendices of this declaration. The Appendices to this declaration include a number of references known to those in the art to describe technical concepts relevant to the subject matter of this declaration, and include

(for example) patents, technical publications, and industry standards. In my opinion, an expert or a person of ordinary skill in the art in the subject matter relevant to this declaration would consider each of the Appendices to this declaration relevant to the subject matter of this declaration and would reasonably rely on such materials to form an opinion as to the state of the art prior to December 27, 2000, the interpretation of the prior art references relied upon in Microsoft's petition, and the obviousness of the claims challenged in the petition. I have also relied on my own personal experience in the field of computer graphics, which includes the design and development of computer graphic hardware, software, and display systems.

II. SUMMARY OF OPINIONS

6. Claims 1-21 of the 506 Patent relate to system and method for dynamic visualization of image data transferred through a communications channel. For the reasons explained below, none of the features described in Claims 1-21 of the 506 Patent were novel as of October 1999, nor does the 506 Patent teach a novel and non-obvious way of combining these known features.

7. Claims 1-21 of the 506 Patent relate to well-known technologies in the computer industry such as multi-resolution hierarchical maps, image compression, packetized data transmission, and three-dimensional (3D) graphics rendering. No element of Claims 1-21 is novel, and Claims 1-21 do not bring these

elements together in a way that brings any benefit beyond what a person of ordinary skill in art would expect from the known functions of the individual components. Claims 1-21 describe techniques that were well-known in the field, and combine them in ways that would have been readily apparent to a person of ordinary skill in the art with predictable results.

8. It is my opinion that each of Claims 1-21 is invalid under the patentability standard of 35 U.S.C. § 103 as I understand it and as explained to me by Microsoft's counsel. Within this declaration I discuss specific grounds of invalidity of Claims 1-21; however, my opinion that Claims 1-21 are invalid under 35 U.S.C. § 103 is not limited to these specific grounds, and indeed, it is my opinion that Claims 1-21 would have been invalid in light of the general knowledge of a person of ordinary skill in the art at the time of the alleged invention.

9. For purposes of my analyses in this declaration only, I provide my proposed construction of certain terms in Claims 1-21 in detail in a later part of this declaration.

10. The subsequent sections of this declaration will first provide my qualifications and experience and then describe details of my analyses and observations.

III. QUALIFICATIONS AND EXPERIENCE

A. Education and Work Experience

11. I received a Ph.D. degree in Electrical Engineering in 1989 and a Master of Science degree in Electrical Engineering in 1985 from the Worcester Polytechnic Institute. I received a Bachelor of Science degree in Electrical Engineering from Syracuse University in 1981.

12. I have more than twenty years of experience in the fields of electrical engineering, computer systems, navigation systems, and communications systems. My experience includes the design, implementation and use of geographic information systems (“GIS”), as well as the design, implementation and use of navigation systems relying on GPS and other positioning system technologies. I also have extensive experience in computer communication and data processing systems as well as systems for the efficient transmission of digital images and other data. Additionally, I have experience in the design and implementation of hardware and software systems used to render image data for display.

13. I have published 16 papers in technical journals and 97 papers in technical conferences. I hold eight U.S. patents in the fields of handheld GPS (Global Positioning System), portable geolocation devices, and communication networks. I have also authored one book chapter relating to optical interconnect

networks for massively parallel computers. I became a Senior Member of the Institute of Electrical and Electronics Engineers (IEEE) in 2003.

14. My experience spans from product designs and R&D in industry, teaching, research and development in an educational and research institution to technology consulting to industry. I was an engineer at Raytheon Company for ten years from 1981 to 1991. During this period, I worked on projects related to computer display hardware for various applications, including air traffic control applications.

15. After leaving Raytheon Company, I joined the Worcester Polytechnic Institute and became a full-time faculty member there in 1991. My research at WPI focuses on navigation systems and related technologies. I am the director of WPI's Robot Navigation and Control Laboratory.

16. My research projects at WPI cover various technologies and include (1) a system using tracking and communications technologies to track shipping containers, (2) an automotive based system that combined GPS and map data in an automotive environment, (3) a remote hazard detection system using GPS and radio communications, and (4) a differential GPS system that combined GPS and radio technologies to determine the precise path of vehicles operating off-road during forest operations.

17. I have worked as a consultant in the navigation and communication systems fields, e.g., in the context of space shuttle docking operations, transfer of traffic information to GPS devices, combinations of GPS and cellular communications for tracking purposes, and map-based handheld tracking devices.

18. I am familiar with numerous GIS and mapping products that existed in the market since the late 1980s, including systems and software developed by Etak, Microsoft, DeLorme, and others. In the conduct of my research and other work, I have routinely used commercially available GIS and mapping products and have developed mapping and visualization software for specialized applications. Additionally, I have used and incorporated database systems such as Microsoft Access, Borland Paradox, Oracle, SQL and others in my research and have incorporated database systems into other hardware and software systems for use in storing and retrieving GIS-related data.

19. I have done extensive research work in communications and networking system design, and have worked with all of the digital, analog and software components needed to build communications and navigation systems. My work with communications and networking protocols began in the mid-1980s with TCP/IP over packet radio. I have used these and other communications and networking protocols extensively in conducting my research. In addition, my work on GPS and navigation systems involved implementing low-latency

communications to support differential techniques that allow a GPS receiver to provide more accurate positioning information.

20. I have extensive experience with the development and maintenance of server computers, including the installation and maintenance of web servers and file servers, as well as the design, development, test, and maintenance of web based applications. These applications typically employ C/C++, Java, JavaScript, PHP, HTML, MySQL, and etc. I am also experienced with server-client systems where the client computer exchanges navigation and/or geographical information with server computer through a wired and/or wireless network.

21. My curriculum vitae, which provides a detailed summary of my education, work experience, publication, teaching history, and etc. is attached to this declaration as Appendix A.

B. Compensation

22. I am being compensated for the services I am providing in this and other Microsoft IPR petitions. The compensation is not contingent upon my performance, the outcome of this *inter partes* review or any other proceedings, or any issues involved in or related to this *inter partes* review or any other proceedings.

C. Documents and Other Materials Relied Upon

23. The documents on which I rely for the opinions expressed in this declaration are documents and materials identified in this declaration, including the 506 Patent, patents related to the 506 Patent, the prosecution history for the 506 Patent and other patents related to the 506 Patent, the prior art references and information discussed in this declaration, including the references attached as exhibits of the IPR Petition for the 506 Patent: WO 99/41675 to Cecil V. Hornbacker, III (“Hornbacker”) (Ex. 1003), *TerraVision II: Visualizing Massive Terrain Databases in VRML* by M. Reddy *et al.*, IEEE Computer Graphics and Applications, March/April 1999 (Ex. 1004), and any other references specifically identified in this declaration, in their entirety, even if only portions of these documents are discussed here in an exemplary fashion.

IV. STATEMENT OF LEGAL PRINCIPLES

A. Claim Construction

24. Microsoft’s counsel has advised that, when construing claim terms of an unexpired patent, a claim subject to *inter partes* review receives the “broadest reasonable interpretation in light of the specification of the patent in which it appears.”

B. Anticipation

25. Microsoft's counsel has advised that in order for a patent claim to be valid, the claimed invention must be novel. Microsoft's counsel has further advised that if each and every element of a claim is disclosed in a single prior art reference, then the claimed invention is anticipated, and the invention is not patentable according to pre-AIA 35 U.S.C. § 102 effective before March 16, 2013. In order for an invention in a claim to be anticipated, all of the elements and limitations of the claim must be shown in a single prior reference, arranged as in the claim. A claim is anticipated only if each and every element as set forth in the claim is found, either expressly or inherently described, in a single prior art reference. In order for a reference to inherently disclose a claim limitation, that claim limitation must necessarily be present in the reference.

C. Obviousness

26. Microsoft's counsel has also advised me that obviousness under pre-AIA 35 U.S.C. § 103 effective before March 16, 2013 is a basis for invalidity. I understand that where a prior art reference does not disclose all of the limitations of a given patent claim, that patent claim is invalid if the differences between the claimed subject matter and the prior art reference are such that the claimed subject matter as a whole would have been obvious at the time the invention was made to a person having ordinary skill in the relevant art. Obviousness can be based on a

single prior art reference or a combination of references that either expressly or inherently disclose all limitations of the claimed invention. In an obviousness analysis, it is not necessary to find precise teachings in the prior art directed to the specific subject matter claimed because inferences and creative steps that a person of ordinary skill in the art would employ can be taken into account.

V. LEVEL OF ORDINARY SKILL IN THE ART

27. I understand from Microsoft's counsel that the claims and specification of a patent must be read and construed through the eyes of a person of ordinary skill in the art at the time of the priority date of the claims. I have also been advised that to determine the appropriate level of a person having ordinary skill in the art, the following factors may be considered: (a) the types of problems encountered by those working in the field and prior art solutions thereto; (b) the sophistication of the technology in question, and the rapidity with which innovations occur in the field; (c) the educational level of active workers in the field; and (d) the educational level of the inventor.

28. The "Background" section of the 506 Patent describes a "well recognized problem" of how to reduce the latency for transmitting full resolution images over the Internet on an "as needed" basis, particularly for "complex images" such as "geographic, topographic, and other highly detailed maps." Ex. 1001 at 1:29-46.

29. To solve this problem and to address some perceived issues in the existing art, the 506 Patent discloses a system capable of “optimally presenting image data on client systems with potentially limited processing performance, resources, and communications bandwidth.” *Id.* at 3:46-49. The 506 Patent states that the disclosed technology can achieve faster image transfer by (1) dividing the source image into parcels/tiles, (2) processing the parcels/tiles into a series of progressively lower resolution parcels/tiles, and (3) requesting and transmitting the parcels/tiles needed for a particular viewpoint in a priority order, generally lower-resolution tiles first.

30. In light of the disclosed technology of the 506 Patent, a person of ordinary skill in the art for the 506 Patent would need education or work experience in computer network communications. Because a “common application” of the 506 Patent is to transmit “geographic, topographic, and other highly detailed maps,” (*id.* at 1:41-43), a person of ordinary skill in the art would require some knowledge and experience with geographic information systems (“GIS”).

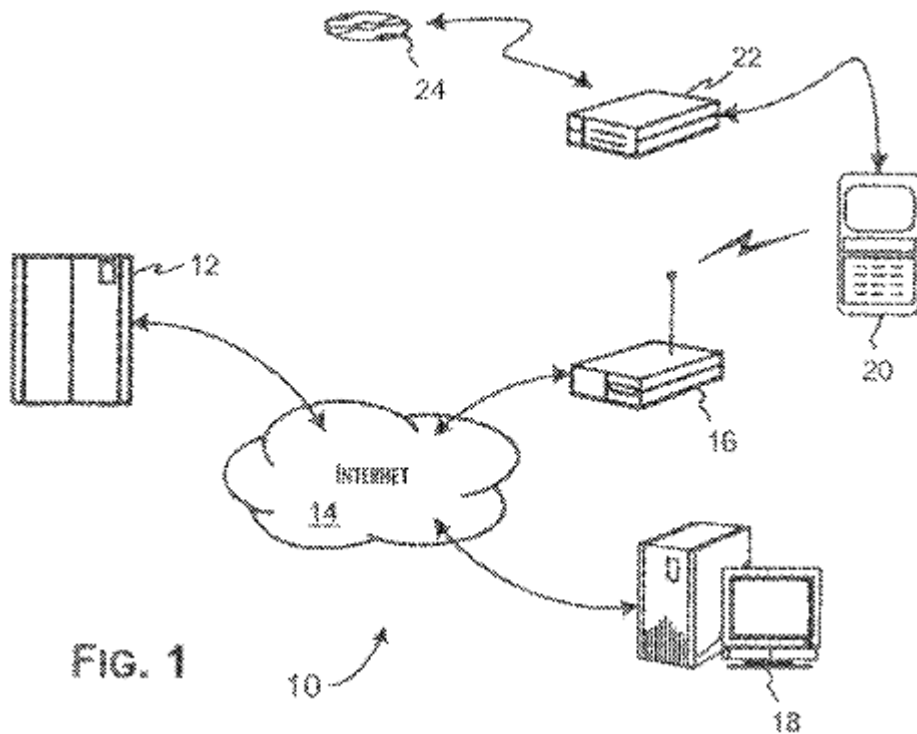
31. Based on the above considerations and factors, it is my opinion that a person having ordinary skill in the art should have a Master of Science or equivalent degree in electrical engineering or computer science, or alternatively a Bachelor of Science or equivalent degree in electrical engineering or computer

science, with at least 5 years of experience in a technical field related to geographic information system (“GIS”) or the transmission of image data over a computer network. This description is approximate and additional educational experience could make up for less work experience and vice versa.

VI. TECHNOLOGY BACKGROUND OF THE 506 PATENT

32. It is my opinion that the 506 Patent recites an obvious and predictable combination of elements that were well-known in the art at the time the 506 Patent was filed and at the time of alleged invention. In this section of my declaration, I provide an overview of some general principles that were understood in the art at the time of filing of the 506 Patent, and therefore would be within the knowledge of a person of ordinary skill in the art. I use certain references (including both patents and non-patent literature) to illustrate the background knowledge of a person of ordinary skill in the art, but the knowledge of a person of ordinary skill in the art at the time regarding the claimed features would not have been limited to these specific references.

32.1 The 506 Patent recites that the “preferred operational environment of the present invention is generally shown in Fig. 1” and links a network server with a client system “through a communications network, such as the Internet 14 generally and various tiers of Internet service providers (ISPs) including a wireless connectivity provider.” Ex. 1002, 5:30-42; Fig. 1:



32.2 Based on my review of the entire specification of the 506 Patent, it appears to me that the inventors describe a system that relied on conventional network connections, including conventional wireless networking methods, and that the underlying means of transmitting data over the Internet or over a wireless network are not emphasized as a point of novelty. In other words, in order to implement the alleged invention in the 506 Patent, a person of ordinary skill in the art would have to rely on existing methods already known in the art of connecting to the Internet and sending data over a wireless connection, since the 506 Patent does not provide any novel teachings about this aspect of the alleged system. This fact is particularly relevant to certain claim limitations which relate to, e.g., the

bandwidth of the communications channel, whether the communications channel is wireless, and the type of client device which operates the 506 Patent's user software, because the ability to connect to the Internet, or connect to the Internet via a wireless channel or on a "small" client device such as a PDA, is something that the 506 Patent assumes that a person of ordinary skill in the art would already know how to do. I considered this relevant to my analysis later in this declaration that these claim limitations are obvious over the references discussed and that a person of ordinary skill in the art would have a reasonable expectation of success implementing the system described by Reddy (which itself describes a laptop computer) on, for example, a Personal Digital Assistant (PDA) or via a wireless connection.

32.3 I provide below a general description of the underlying technology of transmitting data over the Internet and via wireless connections as it existed in 2000 and in before.

A. Data Communications Over the Internet

33. The predominant computer networking technology and set of communications protocols used for most online communications today and prior to the filing of the application for the 506 Patent is known as the Internet Protocol (IP) suite including TCP/IP, named after its two main component protocols: the Transmission Control Protocol and the Internet Protocol. While other protocols,

such as the User Datagram Protocol, or UDP, are also part of the IP suite of protocols, the 506 Patent teaches at 8:12-32 that its preferred embodiment uses TCP/IP to send data packets. In this declaration I do not provide a detailed description of all characteristics of the very well-known TCP/IP protocols, but focus on a few specific aspects of TCP/IP that are pertinent to the claims at issue in the 506 Patent. TCP/IP transmits data between computers in a network using data packets, which are formatted units of data carried by the network as suitably sized blocks. Packets are composed of a header and a payload. The “payload” is the information which the packet is actually intended to convey. The header refers to supplemental data placed at the beginning of a block, and contains information in a standard format such as the sender’s and recipient’s Internet Protocol addresses, a sequence number indicating where in a sequence of packets being transmitted the packet falls, an offset (how far into the data the payload begins) and the protocol governing the format of the payload. The addresses are used to route the packet to its destination, although unlike a circuit-switched connection, different packets going between the same sender and recipient at the same time may take different routes over the network. A rough analogy for data packets is that the header is the “envelope” which contains the address used to deliver the packet, while the payload is the contents of the envelope. The destination computer uses information in the header to place the data packet in its proper place in order, from which the

original data contained in multiple packets can be reassembled. When data segments arrive in the wrong order TCP/IP buffers the out-of-order data until all data can be properly re-ordered and delivered to the application.

34. Before data is transmitted using TCP/IP, the sender and the destination exchange a short series of messages confirming a connection. The connection in this case simply means that the sender and the destination exchange messages to confirm that they are able to exchange messages via the network. When the destination computer receives a packet, it sends a short confirmation message to the sender that the packet has been received. If the confirmation is not received within a certain time period, the sender re-sends the packet. This method avoids losing data in transmission if the transmission of a single packet fails.

35. A common consideration in building online systems is how large to make the data packets. Among other trade-offs, smaller packets may be more likely to reach their destination without loss or error; however, because the header size is similar for a large packet and a small packet, the amount of bandwidth taken up by header overhead increases with the use of smaller packets. In general, there are a variety of packet sizes that may be used in the transfer of a TCP/IP packet from source to destination since lower-level network protocols may have limitations on packet length based on the physical layer or other requirements.

Typically, the network protocols used at any given layer in a protocol stack set the minimum and maximum packet lengths that may be transferred.

B. Data Communications in Wireless Mobile Systems

36. By the late 1990s, it was well-known in the art that digital data could be transmitted by TCP/IP over wireless networks. For example, Appendix X, “The challenges of mobile computing,” *Computer* vol. 27, no. 4, pp. 38, 47 (April 1994) provides an overview of methods for implementing internet connections on mobile devices as of 1994, noting that while wireless networks typically deliver lower bandwidth than wired networks, cellular telephone products of the time could already achieve transmission rates of 9-10 kilobits per second.

37. In another example in 1996, K. Brown and S. Singh, *A Network Architecture for Mobile Computing*, INFOCOM '96, Fifteenth Annual Joint Conference of the IEEE Computer Societies, Networking the Next Generation, Proceedings IEEE vol. 3, pp. 1388-1397 (Appendix Y) describes technologies that integrate wireless or mobile networks with existing fixed data networks such as the Internet. This integration was described by using mobile data protocols to interact and be compatible with the TCP/IP structure of the Internet. This paper described how the Universal Mobile Telecommunications System (UMTS) under development in Europe was expected to offer average mobile data rates of between 1-2 megabytes per second (Mbps) per mobile user. Among other features, “Mobile

users will be able to access their data and other services such as... map services.”

App. Y at 2.

38. Appendix Z, Kreller, B. et al “UMTS: a middleware architecture and mobile API approach,” *Personal Communications, IEEE*, vol. 5, no. 2, pp. 32-38 (April 1998) describes the development of third-generation (3G) mobile networks offering “high-bit-rate data services, guaranteed on-demand bandwidth, and low delays.” *Id.* at 32. To illustrate the development of frameworks to connect mobile telephone networks with existing fixed networks, the authors use the example of a mapping service called “City Guide,” which allows a mobile device to request and download map imagery and other data from a server via hypertext transfer protocol (HTTP) to “provide access to maps describing the current surroundings.” *Id.* at 33. The CityGuide could use JPEG compression and decompression, and could achieve bandwidth of up to 9.6 kbps using the then-existing Global System for Mobile Communications (GSM) cellular data standard. *Id.* at 36, 37. Further, the CityGuide system is an early example of the use of “[w]eb pages to allow instant access to further information about a particular location” by using a browser on a mobile device. *Id.* At 33.

C. Image Tiles and Image Pyramids

39. The 506 Patent describes sub-dividing a high resolution source image into a regular array of image parcels (a.k.a. image tiles), and pre-processing the

image into a series of derivative images of progressively lower resolutions. Ex. 1001 at 6:7-22; Fig. 2. Preferably, the resolution decreases by a factor of four for each derivative image in the series. *Id.* at 6:17-20. Fig. 2 of the six provisional applications to which the 506 Patent claims priority (which is identical in all six provisional applications) best illustrates this image tiling and image pyramid scheme. Ex. 1010 at Fig. 2.

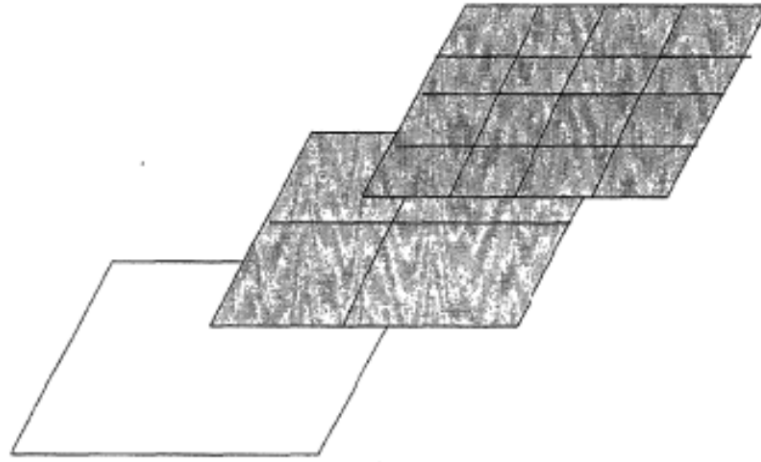


FIG.2

40. This image processing scheme, however, had been developed and widely used long before the 506 Patent's priority date. For example, Hanan Samet's book *The Design and Analysis of Spatial Data Structures* discloses generating an image "pyramid" from a $2^n \times 2^n$ image array, where the pyramid is "a sequence of arrays $\{A(i)\}$ such that $A(i-1)$ is a version of $A(i)$ at half the scale of $A(i)$." App. B, Hanan Samet, *The Design and Analysis of Spatial Data Structures*

at 12 (1989, Reprinted with corrections in January, 1994). Fig. 1.7 in the Samet book is virtually the same as Fig. 2 of the 506 Patent's provisional application.

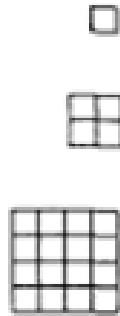
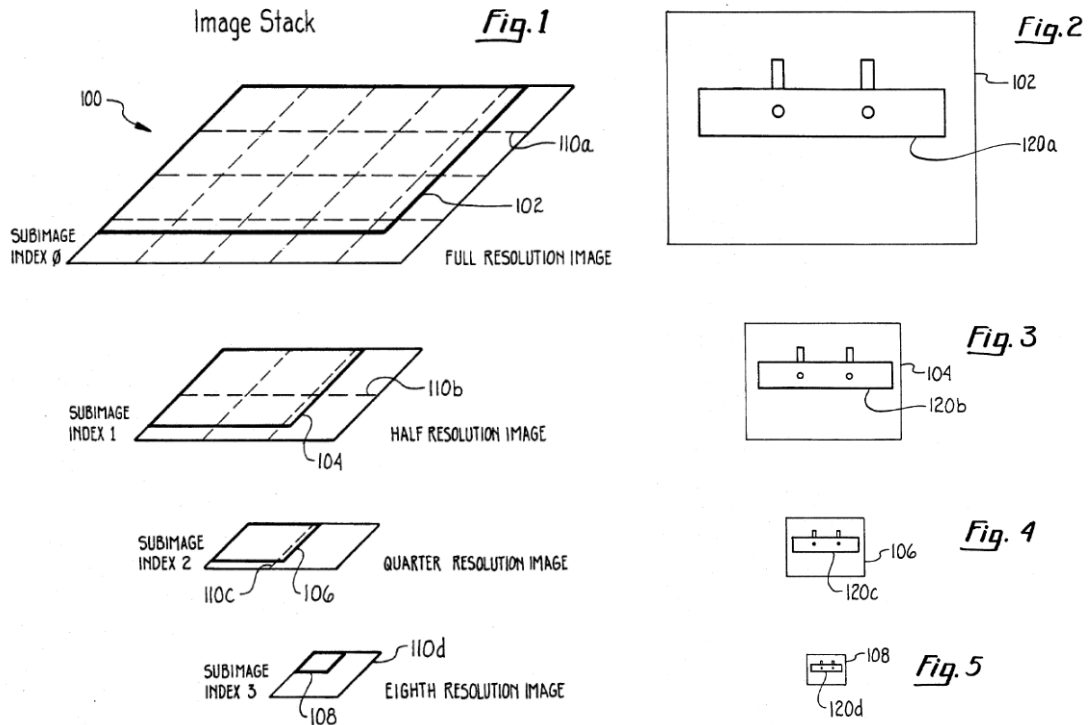
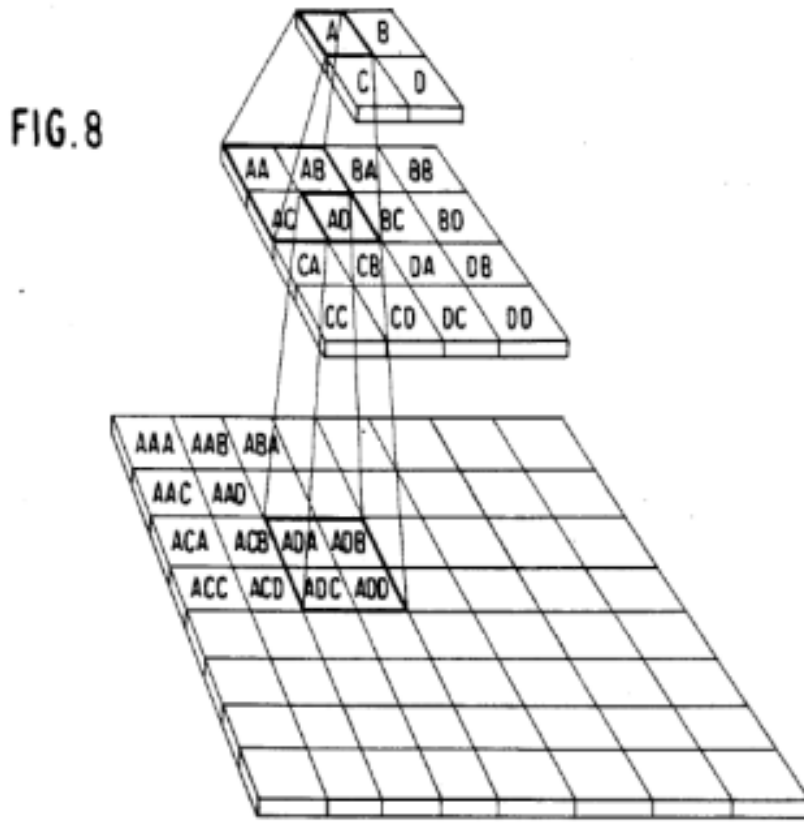


Figure 1.7 Structure of a pyramid having three levels

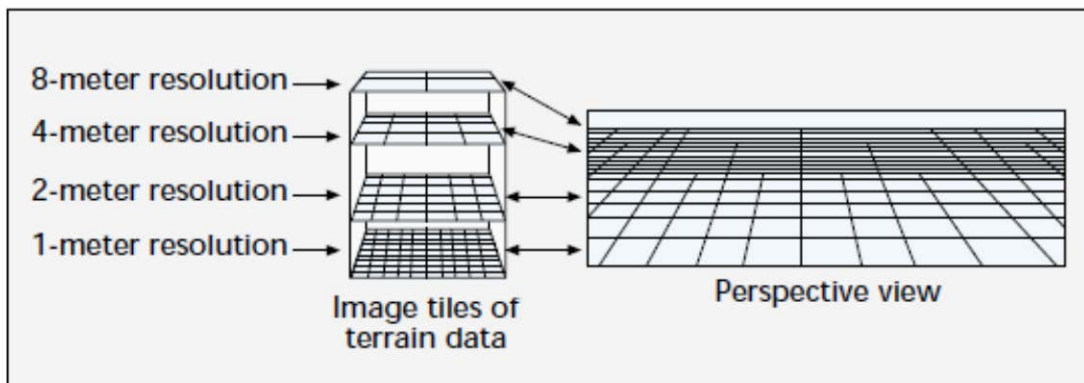
41. In another example, U.S. Patent No. 5,263,136 (DeAguiar et al) filed on April 30, 1991 and issued on November 16, 1993, entitled *System for Managing Tile Images using Multiple Resolutions*, discloses an “image memory management system for tiled images,” where “each source image is stored as a full resolution image and a set of lower-resolution subimages.” App. C at Abstract; Figs. 1 and 2. Suitable applications of the DeAguiar patent’s image tiling and image pyramid scheme include “electrical schematics, topographical maps, satellite images, heating/ventilating/air conditioning (HVAC) drawings, and the like.” *Id.* at col. 6:65-7:2.



42. U.S. Patent 4,972,319 to Delorme, filed on Sept. 25, 1987 and issued on Nov. 20, 1990, also showed that image tiling and image pyramid can be used in mapping applications. Specifically, the Delorme patent discloses a “global mapping system which organizes mapping data into a hierarchy of successive magnitudes or levels for presentation of the mapping data with variable resolution, starting from a first or highest magnitude with lowest resolution and progressing to a last or lowest magnitude with highest resolution.” App. D at Abstract. A pyramid of successively lower resolution image tiles is shown in Fig. 8 of the Delorme 319 patent. *Id.* at Fig. 8.



43. In yet another example, a 1996 article entitled “The MAGIC Project: From Vision to Reality” by Barbara Fuller and Ira Richer (“Fuller”) also shows the image tiling and image pyramid scheme for mapping applications. App. E at Fig. 3.



■ *Figure 3. Relationship between tile resolutions and perspective view.*
 (Source: SRI International)

D. Compression of Image Tiles

44. The 506 Patent discusses that the image tiles can preferably be compressed, e.g., for a fixed compression ratio of 4:1. Ex. 1001 at 6:23-28.

Numerous methods existed, however, long before the 506 Patent's priority date, to compress images for either a variable or fixed ratio.

45. One widely used method of digital image compression is JPEG compression which is based on the Discrete Cosine Transform ("DCT") and is described in the International Telegraph and Telephone Consultative Committee ("CCITT") Recommendation T.81 published in September 1992 (App. F). JPEG compression includes the following main steps: "1. The image is broken into 8x8 blocks of pixels. 2. Working from left to right, top to bottom, the DCT is applied to each block. 3. Each block is compressed through quantization. 4. The array of compressed blocks that constitute the image is stored in a drastically reduced amount of space. 5. When desired, the image is reconstructed through decompression, a process that uses the Inverse Discrete Cosine Transform (IDCT)." App. G at 1, Ken Cabeen & Peter Gent, *Image Compression and the Discrete Cosine Transform*.

46. Another widely used method of digital image compression is based on the wavelet transform. For example, Marc Antonini et al.'s 1992 paper *Image Coding Using Wavelet Transform* discloses a scheme for image compression using

the wavelet transform. *See generally* App. H. In addition, the Antonini paper shows that image compression using wavelet transform not only achieves a good image quality (*id.* at 217-18), but is also suitable for a progressive transmission scheme to “allow the receiver to recognize a picture as quickly as possible at minimum cost.” *Id.* at 218-19.

47. The JPEG 2000 image compression standard, which was designed as the next version of the JPEG Standard to address its identified problems, uses discrete wavelet transform.

E. Progressive Image Resolution Enhancement

48. The progressive image resolution enhancement technique described and claimed in the 506 Patent is one of the “conventional” solutions that have been used to reduce the latency of transmitting complex images over a communications network, as admitted in the “Background of the Invention” section of the 506 Patent. Ex. 1001 at 1:55-65 (“Different conventional systems have been proposed to reduce the latency affect by transmitting the image in highly compressed formats that support **progressive resolution build-up of the image within the current client field of view. . . . Progressive image resolution transmission, typically using a differential resolution method, permits an approximate image to be quickly presented with image details being continuously added over time.**”) (emphasis added).

49. For example, U.S. Patent No. 5,321,520 to Inga et al., filed July 20, 1992 and issued June 14, 1994, discloses a “Progressive Image Enhancement” (“PIE”) method, where a “‘crude’ image is presented to the subscriber” first and then the method “progressively enhance[s] the quality of the presented image” over time. App. I at col. 12:65-13:1. “The longer the user observes a selected image, the ‘better’ the image becomes in the sense of pixel resolution and quantity of gray levels.” *Id.* at col. 13:1-3.

50. The U.S. Patent No. 6,182,114 to Yap et al. was filed January 9, 1998 and issued January 30, 2001. The “Background of the Invention” section of the 506 Patent mentions Yap. The Yap Patent recognizes that “progressive transmission” is an existing approach to solve the problem of “realtime visualization of large scale images over a ‘thinwire’ model of computation,” i.e., over a “low bandwidth line.” App. J at col. 1:47-65. In addition to the traditional progressive transmission method, where the higher resolution of the entire image will be eventually transmitted, the Yap Patent discloses an improved version of progressive transmission, where “resolution is also varied over the physical extent of the image.” *Id.* at col. 2:4-17. Specifically, the Yap Patent discloses that “high resolution data is transmitted at the user’s gaze point but with lower resolution as one moves away from that point.” *Id.* at 2:18-20. The same scheme is used in the 506 Patent.

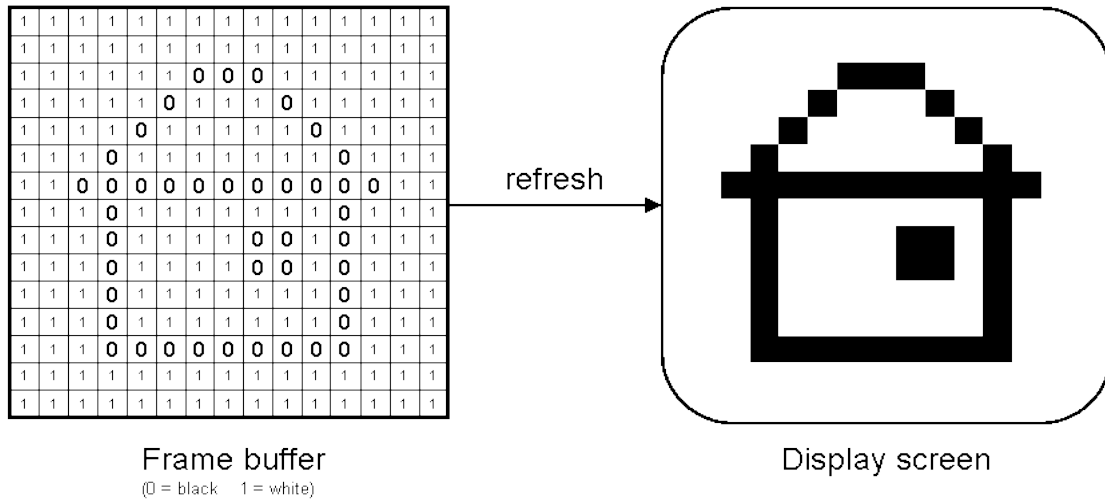
F. Three-Dimensional Graphics

1. Overview of 3D Computer Graphics principles

51. The “field” section of the 506 Patent notes that the claimed invention is designed “to support presentation of high-resolution images subject to dynamic viewing frustums.” Ex. 1002, 1:29-34. The term “frustum” is used in computer graphics to refer to the field of view of a three-dimensional image, and is analogous to the view through a viewfinder of a real-world camera, except that the “camera” is notional and is simply the basis for the calculations that the computer does to render the image. In order to understand this aspect of the 506 Patent and how the prior art relates to it, I discuss below certain concepts in the field of computer graphics. Computer graphics is the art and science of drawing pictures on a display screen using a computer. A picture generated using computer graphics is created from numerical data describing the objects to be drawn. Computer graphics is generally divided into 2D (“two-dimensional”) graphics that only depict images in two dimensions and 3D (“three-dimensional”) graphics that depict images in three dimensions, although by way of representing them on a 2D screen.

52. An image that shows up on a computer display typically corresponds to a large, rectangular, two-dimensional array of values in a computer memory called a frame buffer. An individual location in the frame buffer can hold a color value corresponding to one “dot” or picture element, or pixel for short, on the

display screen. In the simple example shown below, the values in the frame buffer at each pixel are either 0 or 1, which get displayed as black and white, respectively, on the screen.

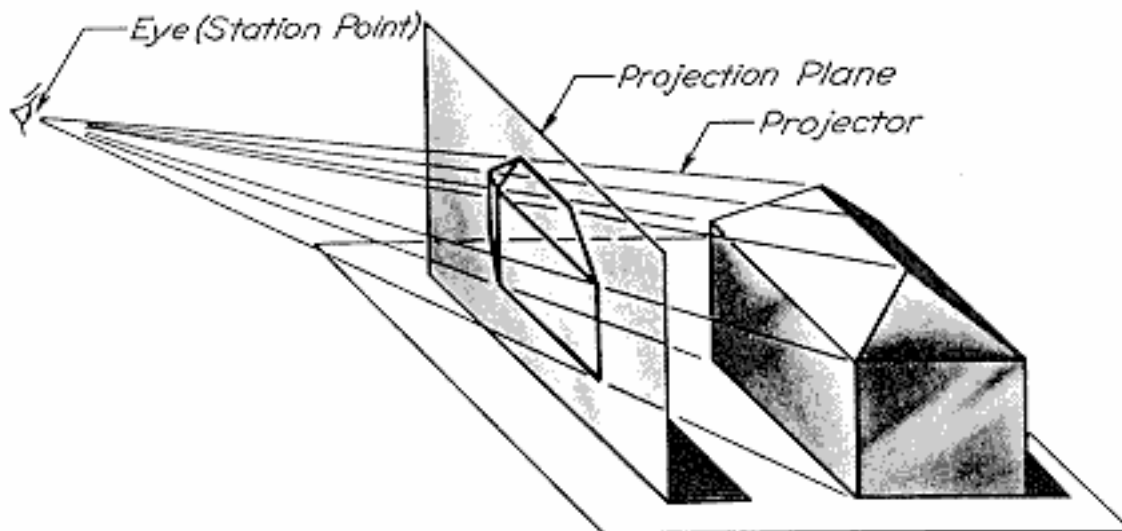


53. In display systems, color values at each pixel are usually either represented by a single number representing shades of gray, or by 3 numbers, R, G, and B, corresponding to red, green, and blue intensity values, for each location on the screen. The computer display is generated by repeatedly "scanning out" the array of numerical pixel values from the frame buffer memory in successive rows at a rate like 60 frames/second, which are used to produce the actual colors seen at each location on the screen.

54. When the computer changes an image displayed on the screen, it updates the corresponding values in the frame buffer. Simply writing a new

number into the frame buffer at a given location results in a new color appearing at that position on the screen starting with the next refresh cycle.

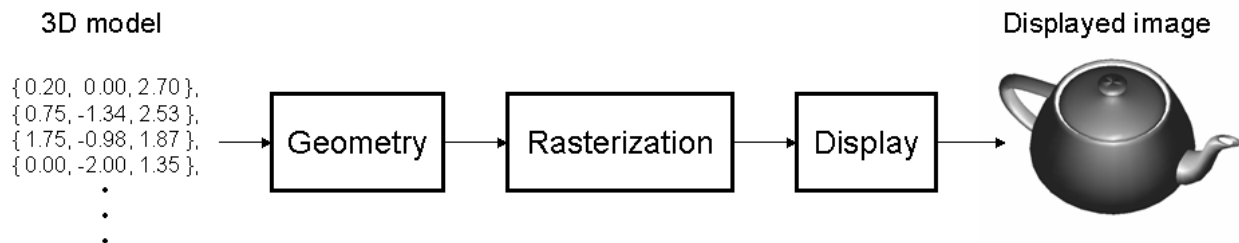
55. Creating an image of a 3D scene involves taking a mathematical description of the objects in a scene, looking at it from a given point of view, and figuring out what colors to draw at all the pixel locations in the frame buffer to create the corresponding image on the screen, as shown in the following example of a 3D house being drawn on a 2D display. All 3D points in the scene are mapped to the corresponding pixels on the screen by projecting along lines of sight, as though the scene was being photographed by a camera onto film. The “projection plane” below is another term for the “viewing frustum” as the term is used in the 506 Patent.



56. The mathematical description of the scene is known as a 3D model. Each 3D object in the 3D model is typically represented using a collection of

geometric “primitives” such as points, lines, and polygons (usually triangles) that make up the object surfaces. In the simple example above, the house might be modeled using 4 polygons for the walls and 4 more polygons for the roof. Each polygon is defined by its vertices or corners. Typically, each vertex is specified using 3D numerical coordinates, X, Y, and Z, for its location in 3D space and R, G, and B values for its color. A mathematical process called rendering is used to model a virtual "camera" looking at the 3D scene from a particular point of view, mathematically project all the 3D polygons into the corresponding 2D pixels in the display, and assign the appropriate colors to them in the frame buffer.

57. Various 3D computer graphics systems are built around the concept of a graphics pipeline. Acting like an assembly line, the graphics pipeline takes in the "raw materials" consisting of the data for the underlying 3D model and processes these through a series of computational steps to produce the image displayed on the 2D screen. In its simplest form, a graphics pipeline is described as having a series of three general phases, geometry, rasterization, and display, as shown in the diagram below:



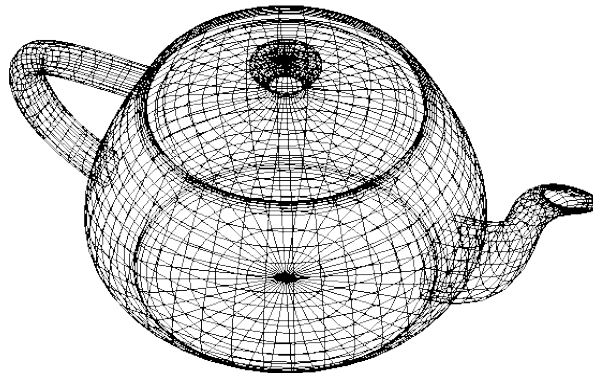
58. In the above diagram, the 3D model represents the X, Y, Z coordinates and R, G, B color values for all the 3D polygons that make up the objects in the desired scene, along with certain other information such as the location of the camera, light sources, display boundaries, etc. Geometry refers to the calculations performed to mathematically transform the 3D coordinates of all the polygons in the 3D model into corresponding screen coordinates given the location and orientation of the imaginary camera viewing the scene. Rasterization refers to the computations that determine all of the 2D pixel locations that will be visible within each 3D polygon and the colors those pixels should have. Display is then the process of writing the 2D pixel color values into the frame buffer and thereby causing the corresponding image to be displayed on the screen.

59. Rasterizing a polygon generally involves three main tasks: determining which pixels fall within the polygon (scan conversion), determining which of these pixels are visible on the screen (visible-surface determination), and determining what color to assign each visible pixel (shading).

60. The process of determining which pixels within the scan converted polygons will actually be visible on the screen is known as visible-surface determination. Depending on the direction from which the scene is viewed by the virtual camera, certain polygons (or portions thereof) may be occluded by other polygons and not visible on the screen, such as the back wall of the house in the

example shown above. One common way to solve the visible surface problem is to write the RGB value for a pixel into the frame buffer only if its 3D position is closer to the camera than what may have been previously written into that same location by another polygon, much as an oil painting is painted in layers from back to front. This involves what is known in the art as depth buffering or z-buffering, that is, keeping track of the depth or "Z" value (distance from the viewer's eye) currently residing at each pixel.

61. Once the scan conversion and rasterization processes are complete, the graphics program must assign colors to each visible pixel, a process that has evolved substantially over the history of computer graphics and depends on the level of realism desired in the resulting image. In the simple example below, only the scan converted pixels that make up the edges of each polygon are drawn with black pixels.



Example of a Wireframe Image

62. The resulting collection of polygons approximating the three-dimensional object is sometimes referred to as a “mesh” or “polygon mesh.” The use of meshes, including foundational work performed by Microsoft to represent three-dimensional objects at varying levels of detail, is described in detail in Hughes Hoppe’s 1996 paper “Progressive Meshes,” (App. N), which was published in the SIGGRAPH ’96: Proceedings of the 23rd annual conference on computer graphics and interactive techniques, pp. 99-108, and is also available online at <http://research.microsoft.com/en-us/um/people/hoppe/pm.pdf>.

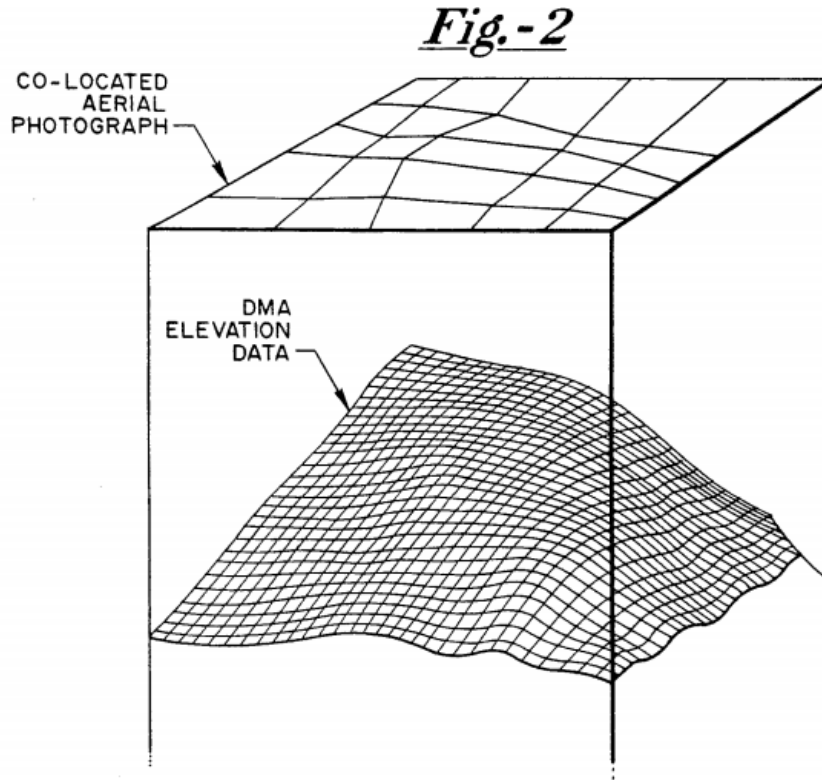
2. Texture

63. In 1974, “texture mapping” was developed as a further improvement in adding detail to objects or images. Texture mapping involves applying a 2D image or function approximating some real-world material like wood, bricks, fabric, marble, or a checkerboard, to the surface of polygons (*i.e.*, the mesh) as in the image shown below. The “pixels” of a texture map are often referred to as texture elements or texels to distinguish them from the pixels of the resulting image. Texture mapping is like applying wallpaper or a decal to a surface. It is possible to construct a brick wall by carefully drawing many 3D bricks, which takes a lot of work, or one can simply paste a photograph of a brick wall onto an otherwise flat wall, which is easier and looks like a brick wall if you don’t look too close. Texture mapping has become standard in 3D graphics systems to use texture

mapping to quickly fill in realistic detail for many of the objects in a 3D scene, especially floors, walls, sky, and other background areas. Textures can be either color images, or can be monochrome images used to modulate the untextured color of the polygon.

64. Some textures may be generic- for example, a 3D graphics rendering program might re-use a “wood” texture for all objects represented as wood. In this scenario, the texture is essentially a “wallpaper” with a repeating pattern applied to certain objects or surfaces within the field of view.

65. Textures may also be unique, or specific to a particular surface or object. This is often the case when photographs are used as textures. For example, when satellite or aerial photographs are used in a 3D rendering of a landscape, the specific portions of the imagery that correspond to a particular location are mapped onto the terrain at that location. For example, U.S. Patent No. 5,179,638 to Dawson et al., assigned to Honeywell, Inc., (App. K) illustrates how an aerial photograph can be used as “texture data” and mapped onto co-located digital elevation data, as shown in Fig. 2:



66. This technique of using aerial imagery as a texture applied to a three-dimensional model of terrain is also known as a “synthetic view.” Synthetic view technology can be used in aviation to provide a pilot operating at night or bad weather with a synthesized view of the terrain around them based on actual position (e.g. derived from GPS). Appendix AA, Hansen, J. et al, “Real-time synthetic vision cockpit display for general aviation,” AeroSense ’99, International Society for Optics and Photonics, 1999, describes such a system. In the figure below, the bottom portion of the figure shows a wire-frame diagram illustrating the three-dimensional model of terrain, while the top image shows the synthetic view created by rendering satellite imagery on the terrain model:

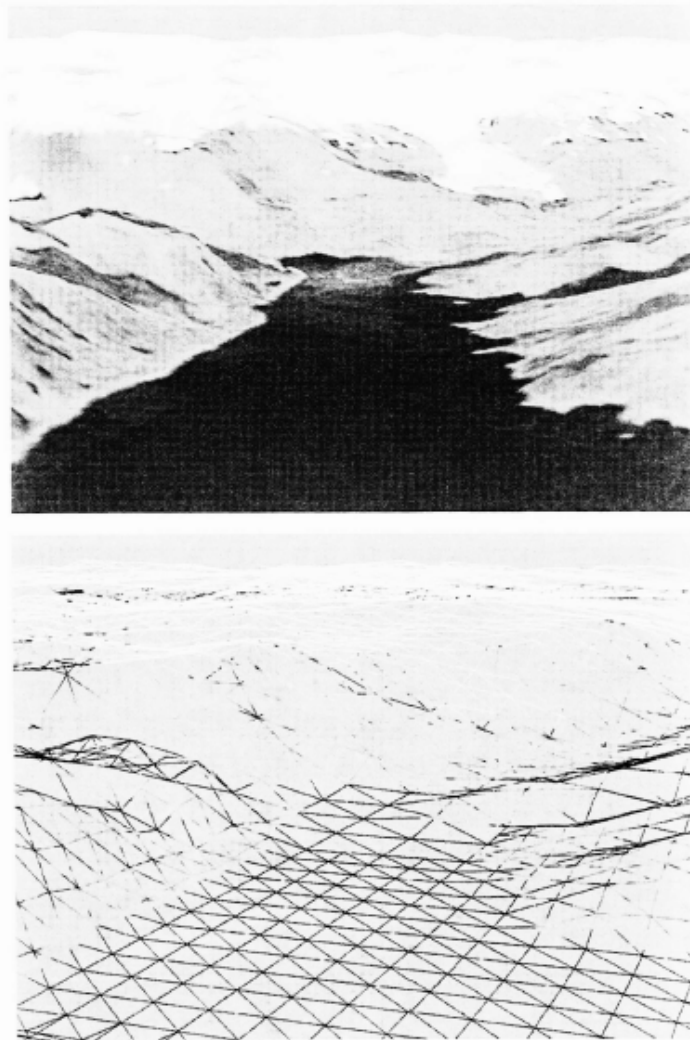
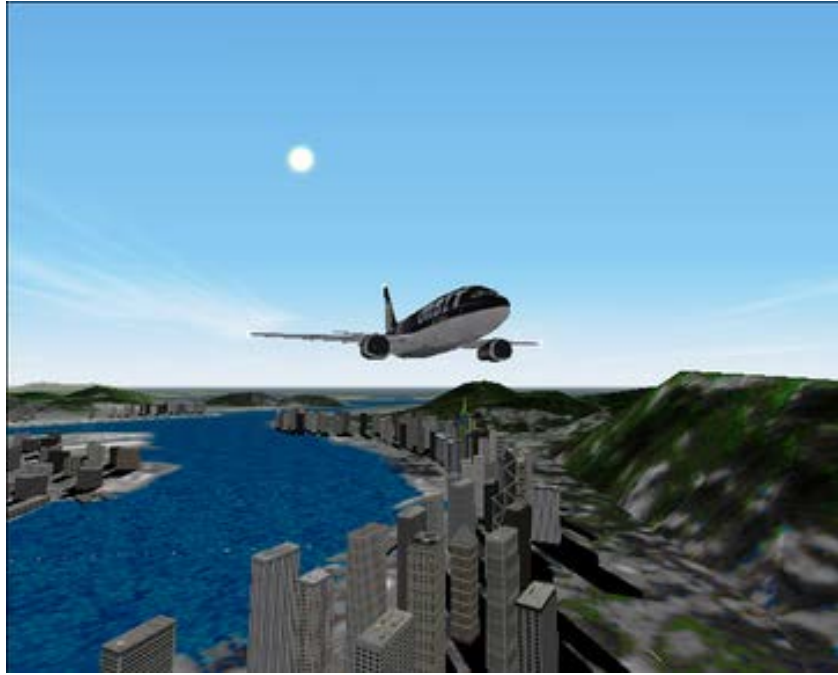


Figure 5. Terrain with extreme amounts of structure can be accommodated with high fidelity. The bottom graphic is a wire frame image of the Alaskan coastline. On the top is the fully rendered scene with imagery.

67. Microsoft used a similar technique with its popular Flight Simulator series of computer games, starting with Flight Simulator 1995. Flight Simulator utilized a real-time 3D rendering of terrain features with textures generated from a variety of sources, including satellite imagery. The figure below illustrates a 3D perspective view generated in Flight Simulator 2000, which was actually released in late 1999:



68.

3. Virtual Reality Modeling Language (VRML)

69. VRML was a well-known industry standard that defined a file format for describing interactive 3D objects and worlds (App. KK). The 1997 release version of the VRML standard, commonly known as VRML97, specified a file format by which visual information, e.g., the texture and 3D modeling discussed herein, can be stored on a computer and viewed.

70. A user wishing to view visual information stored in the VRML format could typically use either a VRML viewer program, a VRML browser or a plug-in that worked with an off-the-shelf web browser. Because VRML was designed to operate in a client-server system, e.g., conveying visual information via web pages, a typical network-based VRML application would rely on an underlying Internet mechanism of hypertext data requests to fetch VRML data from a server to a user

computer. One advantage of VRML file format was that it allowed for hyperlinking to other media such as text and images, and thus provided a mechanism by which a user could navigate through a visual scene being viewed, while in the background, the VRML application would fetch and/or render additional data.

G. Mip-Maps

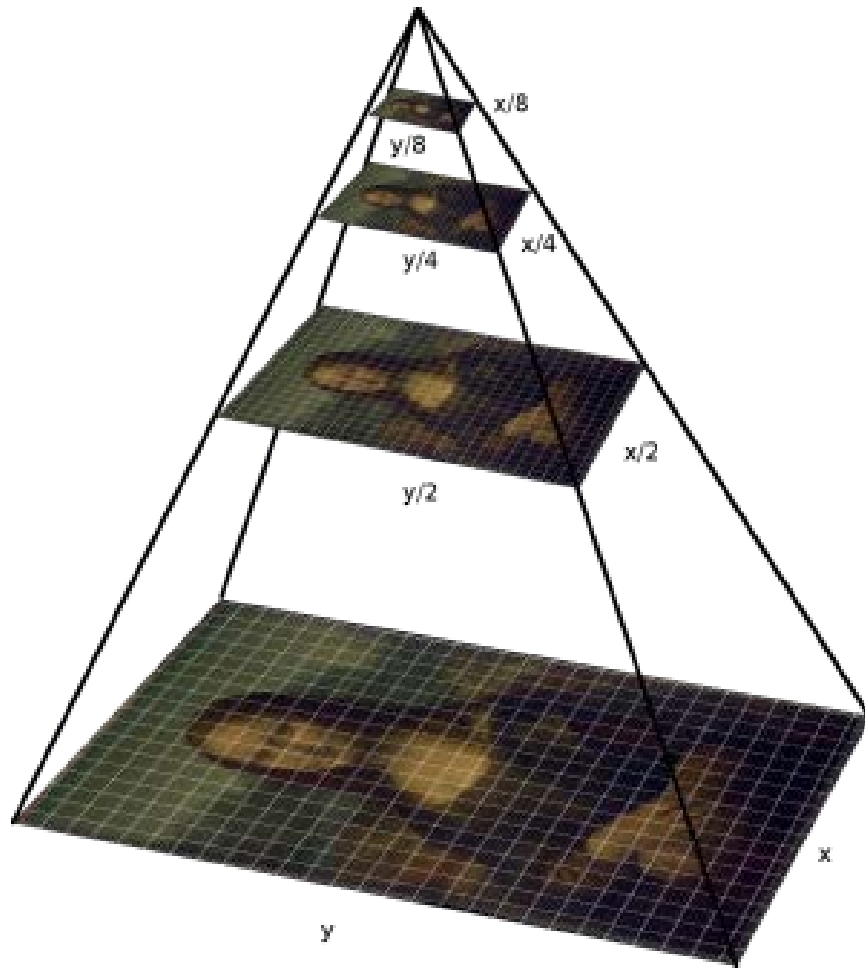
71. The provisional applications that the 506 Patent claims priority to described the use of “mip-maps” as “surface textures when rendering a two-dimensional representation of a three-dimensional scene.” *See, e.g.*, Ex. 1010 at 7-9. This mip-mapping technology, however, has been used for rendering surface textures since 1979, more than two decades before the filing date of the provisional applications to which the 506 Patent claims priority. App. L at 2, Lance Williams, *Pyramidal Parametrics*, Computer Graphics, vol. 17, no. 3, July 1983.

72. The term “mip” derives from the Latin phrase “multum in parvo” meaning “many things in a small place.” The term was adopted by Lance Williams in his 1983 paper, which indicated that “the mip-mapping technology has been used successfully to bandlimit texture mapping . . . since 1979.” *Id.* Mip-mapping has been adopted in several versions of the OpenGL Standard prior to the filing date of the 506 Patent including OpenGL 1.1 released in March 1997 and OpenGL 1.2.1 released in April 1999. App. M, “Mipmapping” section of the

OpenGL 1.1 Standard;¹ App. T at 129-131. OpenGL (GL stands for “graphics library”) is a 3D graphics standard originally designed and released by Silicon Graphics, Inc. of Mountain View, CA. As defined in OpenGL 1.1, a mipmap is “an ordered set of arrays representing the same image; each array has a resolution lower than the previous one. If the texture has dimensions $2^n \times 2^m$, then there are $\max\{n, m\} + 1$ mipmap arrays. The first array is the original texture with dimensions $2^n \times 2^m$. Each subsequent array has dimensions $2^{(k-1)} \times 2^{(l-1)}$ where $2^k \times 2^l$ are the dimensions of the previous array.” *Id.*

73. An illustration of the mipmap pyramid is shown below. *See* Photospector Blog, <http://photospector.com/gigapixels/>.

¹ Available online at <https://www.opengl.org/documentation/specs/version1.1/glspec1.1/node84.html#SECTION00681100000000000000>



74. By the late 1990s, mipmaps were commonly used in 3D graphics applications, among other purposes, to render object textures at varying levels of detail based on the proximity of the object to the simulated viewpoint. For example, it would ordinarily be preferable to display an object in close proximity to the viewpoint at a high level of detail where the display is capable of showing a high level of texture detail, whereas for more distant objects a lower level of resolution is preferable because the display screen is unlikely to be capable of displaying the highest-resolution texture at a great distance, and because lower-resolution textures

require far less system resources and bandwidth to retrieve, load and render. U.S. Patent No. 5,760,783 to Migdal et al ("Migdal," App. BB) is a patent from Silicon Graphics which describes how mipmaps may be used to render textures- including satellite or aerial photographs used as terrain textures for large maps, such as a flight simulator application. App. BB, 9:5-17, 10:14-19. Migdal illustrates how mipmaps at higher levels of detail may be used for points closer to the viewpoint and lower levels of detail for more distant objects. For example, Fig. 4C of Migdal illustrates a perspective view of regions of three different level of detail maps aligned with a center line from an eyesight location:

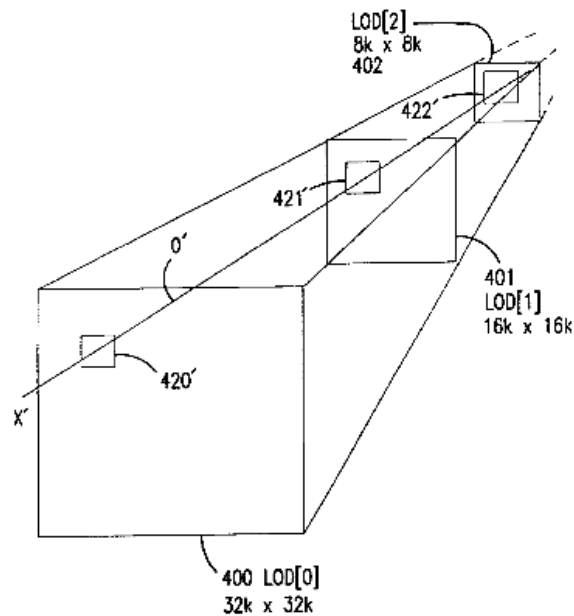
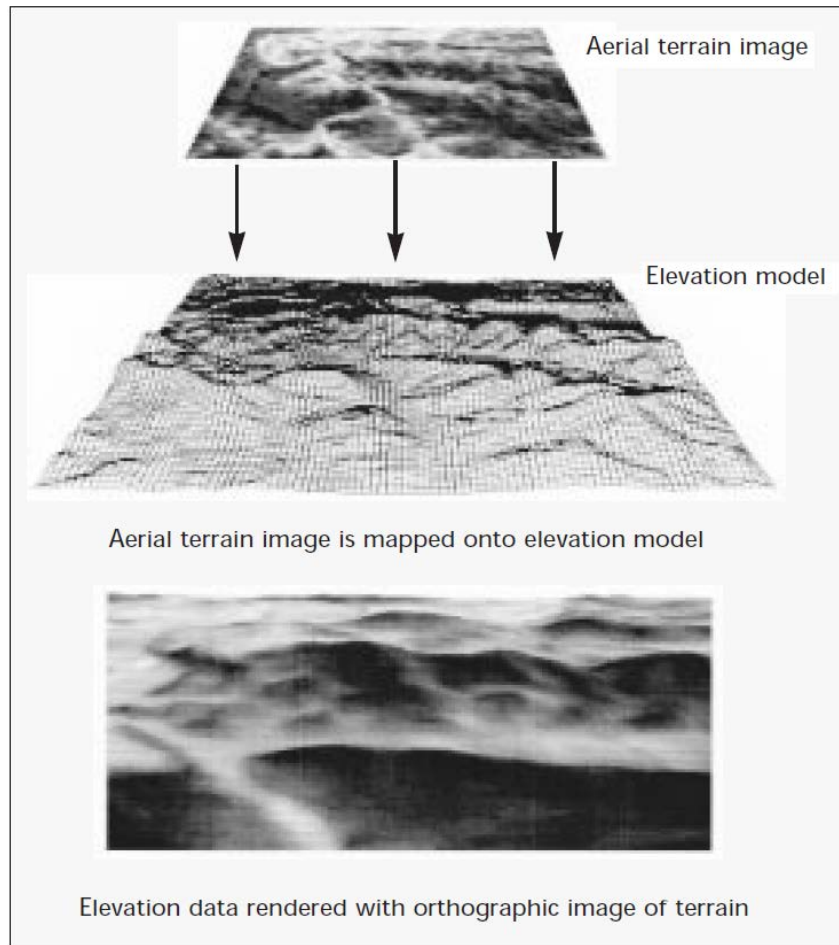


FIG.4C

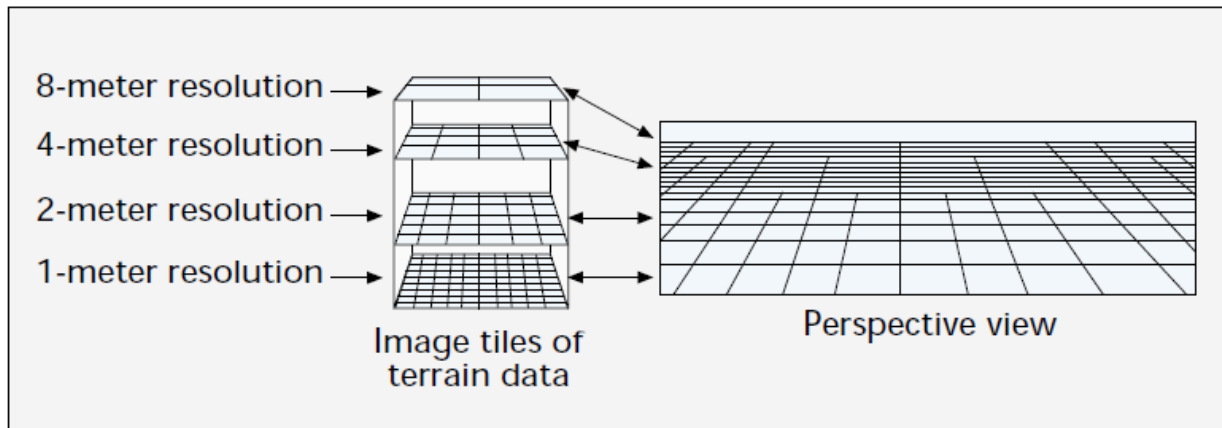
75. Migdal teaches that Fig. 4C illustrates that the clip-map “contains sufficient texel data to cover larger minified areas in the background of a display where coarser texture detail is appropriate.” *Id.* at 10:3-5. In my opinion, this teaching of Migdal is representative of what was already well-known in the art long before the earliest priority date claimed by the 506 Patent: that 3D graphics applications could use “mipmaps” or similar level-of-detail pyramids, to render objects closer to the viewpoint at a higher resolution and objects more distant from the viewpoint at a lower resolution.

76. Fuller (App. E) also illustrates this principle of using mipmaps to display closer objects at higher resolution and more distant objects at lower resolution. Fuller describes an online system for 2D and 3D visualization of map data, which creates a 3D perspective representation of a landscape using a digital elevation model (DEM), then uses mipmaps of aerial images as the textures. App. E, Fig. 4:



■ Figure 4. *Mapping an ortho-image onto its digital elevation model.*
(Source: SRI International)

77. Fig. 3 of Fuller shows that higher resolution images are mapped onto portions of the terrain nearest the viewer, while lower resolution images are mapped onto more distant portions of the terrain:



■ Figure 3. *Relationship between tile resolutions and perspective view.*
(Source: SRI International)

H. Storage of image data

78. In practical systems, the visual information in the form of image tiles, image pyramids, mip-maps, 3D graphics, meshes, and so on, is stored on a computer from which users can access the visual information.

79. The 506 Patent describes that a network image server system stores the source image data. Ex. 1002 at 5:60-66. The source image data is typically a high-resolution bit-map raster map and/or satellite imagery of geographic regions. *Id.* The 506 Patent describes that this source image data is pre-processed to obtain a series of derivative images, and is also subdivided into a regular array such that each resulting image parcel has a same resolution. The 506 Patent then describes that each image parcel is “preferably” stored in a file such that any image parcel can be located by specification of a KD, X and Y value.

80. Many formats for storing images, and the trade-offs for storage and retrieval efficiency, were well-known for several years before the filing of the 506 Patents. The formats and trade-offs related to the color space used for image representation, the tile size used in case of multi-resolution images, whether different portions of an image are stored in one file or multiple files, the number and details of headers used to indicate information about the image, and so on.

81. For example, GeoTIFF is a file format (Appendix GG) for storing geographical image files using the well-known Tagged Image File Format (TIFF). The GeoTIFF spec defines a set of TIFF tags provided to describe all "Cartographic" information associated with TIFF imagery that originates from satellite imaging systems, scanned aerial photography, scanned maps, digital elevation models, or as a result of geographic analyses. Its aim is to allow means for tying a raster image to a known model space or map projection, and for describing those projections. The 1.0 release version of the GeoTIFF file format builds upon the then available revision 6.0 of TIFF.

82. TIFF (Appendix HH) defines a file format for storing image data. In particular, TIFF 6.0 provides for storage of multiple images in a single TIFF file, each image corresponding to a subfile. These images could be, e. g., a full resolution image and its reduced resolution representations.

83. FlashPix by Kodak was another well-known format for storing images (Appendix II). A FlashPix file stored image data in a hierarchy of resolutions from the highest available for an image, down to the lowest defined format (§2.1). Thus, depending upon an implementer's choice, a FlashPix file either contained a single resolution image or an entire multi-resolution hierarchy (§2.2, FIG. 2.3). For convenient access, each resolution image was organized in tiles, which represented rectangular pixel arrays of the image. (§2.3. FIG. 2.4).

84. These techniques, and other obvious variations, for storing image data, including multiple resolution versions of an image, provided multiple implementation choices to an implementer. Depending on design criteria such as the amount of storage available, tolerance to header overhead, transmission latency when the images are to be transferred over a network, portability and compatibility with client viewing programs, an implementer would be able to select a storage method for storing images in a single file, with each tile in its separate file, or with all tiles of one resolution in a single file, and so on.

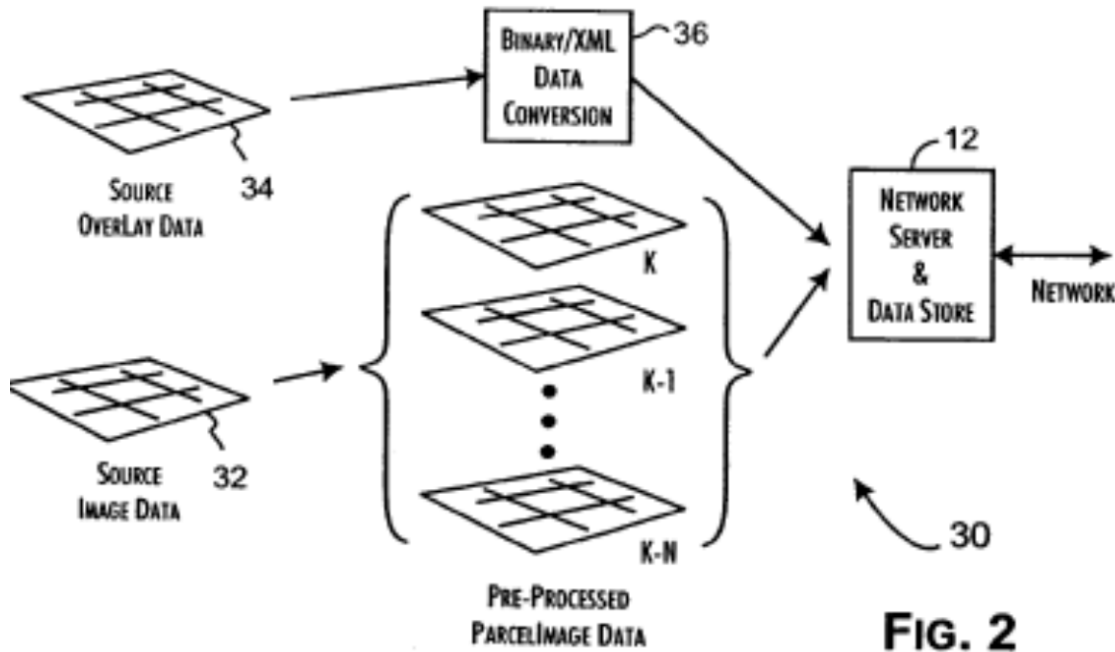
VII. OVERVIEW OF THE 506 PATENT

85. The 506 Patent describes a system in which “[l]arge-scale images are retrieved over network communications channels for display on a client device by selecting an update image parcel relative to an operator controlled image viewpoint to display via the client device.” Ex. 1001 at Abstract.

86. The “Background” section of the 506 Patent describes a “well recognized problem” of how to reduce the latency for transmitting full resolution images over the Internet on an “as needed” basis, particularly for “complex images” such as “geographic, topographic, and other highly detailed maps.” Ex. 1001 at 1:41-59. The 506 Patent states that solutions already in existence included “transmitting the image in highly compressed formats that support progressive resolution build-up of the image within the current client field of view.” *Id.* at 1:59-65. The 506 Patent also states that such “conventional” solutions, like the ones described in U.S. Pat. Nos. 4,698,689 (Tzou) and 6,182,114 (Yap), usually “presume that client systems have an excess of computing performance, memory and storage” and are “generally unworkable for smaller, often dedicated or embedded” clients. *Id.* at 1:59-3:12. According to the 506 Patent, the conventional solutions do not work well under “limited network bandwidth” situations. *Id.* at 3:12-25.

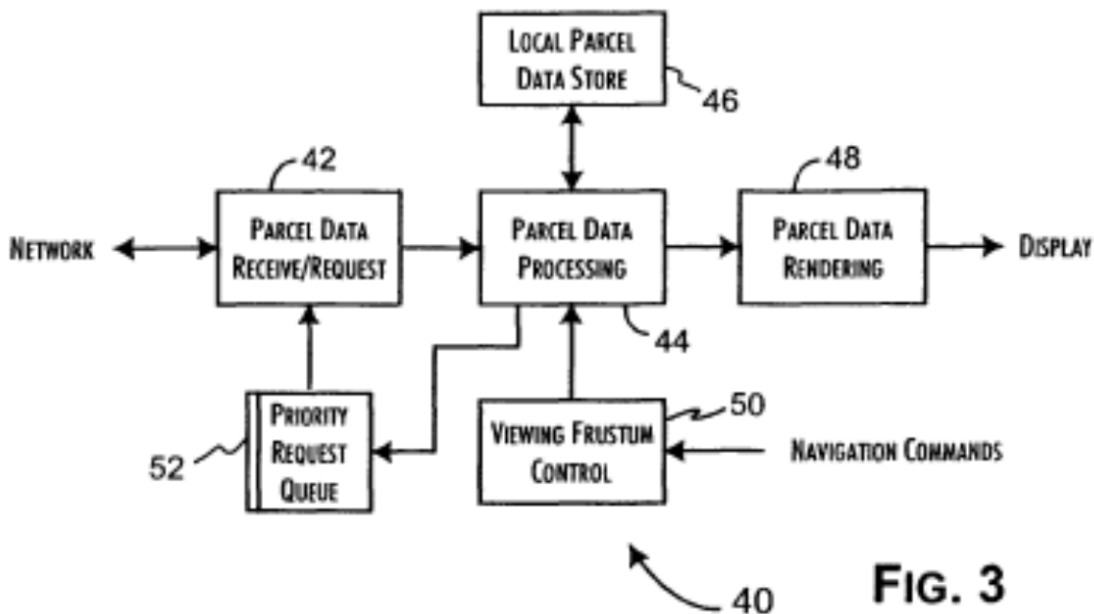
87. To address these perceived issues in the existing art, the 506 Patent discloses a system capable of “optimally presenting image data on client systems with potentially limited processing performance, resources, and communications bandwidth.” *Id.* at 3:46-49.

88. Specifically, the 506 Patent describes an image distribution system having a network image server and a client system, where a client can input navigational command to adjust a 3D viewing frustum for the image displayed on the client system. *Id.* at 5:30-59. High-resolution source image data is pre-processed by the image server into a series K_{1-N} of derivative images of progressively lower image resolution. *Id.* at 6:7-22, Fig. 2. The source image is also subdivided into a regular array of 64 by 64 pixel resolution image parcels



(a.k.a. image tiles), and each image parcel may be compressed to fit into a single TCP/IP packet for faster transmission. *Id.* at 6:12-28; 8:14-32.

89. The client system in the 506 Patent has a “parcel request” subsystem to request image parcels from the server, a “control block” that directs the transfer of received image parcels and overlay data to a local parcel data store. *Id.* at 7:10-32. The control block also decompresses the image parcels and directs a “rendering engine” to render them. *Id.* at 7:34-36; Fig. 3.



90. When the viewing point is changed in response to a user navigation command, the control block “determines the ordered priority of image parcels to be requested from the server . . . to support the progressive rendering of the displayed image.” *Id.* at 7:59-62. A number of image parcel requests are then placed in a request queue, to be issued by the parcel request subsystem according to each request’s assigned request priority. *Id.* at 7:62-64; 9:7-19. Although various factors may affect the priority assigned to a parcel request, e.g., the “resolution of the

client display” (9:37-54) or whether the image parcel is “outside of the viewing frustum” (10:9-12), generally speaking, “image parcels with lower resolution levels will accumulate greater priority values,” so “a complete image of at least low resolution will be available for rendering” in a fast manner (10:59-67). In addition, the control parameter for calculating the priority can be set in a way that gives “higher priority for parcels covering areas near the focal point of the viewer” to make sure that image parcels are requested “based on the relative contribution of the image parcel data to the total display quality of the image.” *Id.* at 11:1-19.

91. In the 506 Patent, after the needed parcels are requested and received, an algorithm is used to select the image parcel for rendering and display. *Id.* at 9:20-25. Overlay data may also be added to the display if its image coordinates matches the current image parcel location. *Id.* at 9:29-34. The 506 Patent discloses that two-dimensional image parcels are displayed in a three-dimensional space using projection transform. *Id.* at 5:50-59; 7:50-58; 9:20-24; 11:1-7; 11:43-47. In my opinion, there is no disclosure in the specification of the 506 Patent that teaches or suggests that the images displayed are mapped onto an elevation model, as there is neither an independent description of such a method or any reference to the existing prior art that already had this feature, as I discussed above. The 506 Patent specification suggests that an overlay may include 3D objects (*Id.* at 5:66-6:6), but this disclosure by itself does not support any teaching that the “satellite imagery”

discussed is warped to fit on a three-dimensional elevation model. The specification of the 506 Patent effectively discloses a view that is “three-dimensional” in the sense that it generates a viewing perspective that contains position, rotation, and height components, but operating over a flat plane of terrain imagery. *Id.* at 5:53-56.

92. The 506 Patent states that its disclosed technology can achieve faster image transfer by (1) dividing the source image into parcels/tiles, (*id.* at 6:7-22), (2) processing the parcels/tiles into a series of progressively lower resolution parcels/tiles, (*id.*) and (3) requesting and transmitting the parcels/tiles needed for a particular viewpoint in a priority order, generally lower-resolution tiles first. *Id.* 3:46-4:47.

VIII. IDENTIFICATION OF THE PRIOR ART AND SUMMARY OF OPINIONS

93. As explained below, it is my opinion that the prior art references cited in this Declaration disclose all technical features in Claims 1-21 of the 506 Patent, thus rendering them unpatentable.

94. Based on my review of the prior art references, claims 1-21 of the 506 Patent are rendered obvious by Reddy in view of Hornbacker.

A. Reddy

95. “TerraVision II: Visualizing Massive Terrain Databases in VRML” by Martin Reddy et al. (“Reddy”) was published in the March/April 1999 issue of IEEE Computer Graphics and Applications. From my 25+ year experience as an IEEE member, I am very familiar with the IEEE, which is the world’s largest association of technical professionals and publishes a number of well-respected peer-reviewed periodical journals. The IEEE Computer Graphics and Applications was a well-known publication. In my opinion, persons of ordinary skill in the art in the field of computer graphics would be familiar with the IEEE and its publications, consider IEEE publications to be established, reliable sources of information accessible to those of skill in the art, and would rely on the publication and copyright dates indicated on the face of an article in an IEEE publication as a reliable indication of the actual publication date of the article. Reddy bears such publication dates on each page and a 1999 copyright date on the first page of the article (page 30 of the journal). In my opinion, the publication dates indicated on the face of Reddy are a type of information that persons in the field of computer graphics would reasonably rely upon. Reddy includes a footnote “0272-1716/99/\$10.00 © 1999 IEEE.” Based on my experience, this footnote means that Reddy was published sometime in the year 1999. Further, it was common practice to release IEEE journals to public in the beginning of the stated period of the

journal, For example, based on my experience receiving numerous IEEE journals over the past 25 years, the March-April 1999 issue would have been sent to subscribers in late February or early March 1999. Therefore, Reddy was published more than one year prior to December 27, 2000.

B. Hornbacker

96. WO 99/41675 to Cecil V. Hornbacker, III (“Hornbacker”) is a patent publication that was filed on Feb. 12, 1998 and published on Aug. 19, 1999.

IX. CLAIM CONSTRUCTION

97. In conducting my analyses of the asserted claims of the 506 Patent, I have applied the legal understandings I set out below regarding claim constructions consistent with the “broadest reasonable interpretation” (BRI) standard described above, and offer them only for this *Inter Partes* Review. The claim constructions do not necessary reflect the appropriate claim constructions to be used in litigation proceedings, such as litigation in a district court, where a different standard applies.

98. I understand that, under the BRI claim construction, claim terms are given their ordinary and customary meaning as would be understood by one of ordinary skill in the art in the context of the entire disclosure. An inventor may rebut that presumption by providing a definition of the term in the specification with reasonable clarity, deliberateness, and precision. In the absence of such a definition, limitations are not to be read from the specification into the claims.

A. “Data Parcel”

99. The proposed BRI claim construction for the term “data parcel” in all claims of the 506 Patent is “data that corresponds to an element of a source image array.” Support for this construction is found in the specification, for example, at 6:7-17 and Fig. 2.

B. “A mesh” in claim 13

100. The proposed BRI construction for the term “mesh” is “a collection of polygons.” The 506 Patent only uses the term “mesh” in the claims, rather than in the specification, so I analyze this term in light of the knowledge of a person of ordinary skill in the art and the related teachings that appear in the 506 patent. As I discussed above in paragraphs 56-62, a “mesh” in the field of three-dimensional imaging typically refers to a set of polygons representing the three-dimensional object (such as a view of terrain elevation) and this representation is used as the base on which textures are mapped. The 506 Patent teaches at 9:55-10:15 that “the image display space is progressively split [] by four to one reductions into polygons.” The quad-tree structure which holds the image data is then traced to “establish a correspondence with the polygon map.” After the correspondence has been established between the polygon map and the image data, the image data is mapped in video memory onto the appropriate coordinates of the corresponding

polygon as an image texture. In my opinion, these teachings are the closest disclosure in the 506 Patent to a “mesh” as the term is commonly used in the art.

C. Other claim terms

101. In my opinion, all other terms in claims 1-21 of the 506 Patent should be construed according to their plain and ordinary meaning in view of the specification of the 506 Patent.

X. UNPATENTABILITY OF THE 506 PATENT CLAIMS

A. GROUND 1: CLAIMS 1-21 ARE UNPATENTABLE UNDER 35 U.S.C. § 103(a) AS BEING OBVIOUS OVER REDDY IN VIEW OF HORNBACKER

102. In my opinion, claims 1-21 are disclosed and rendered obvious by Reddy (Ex. 1004) in view of Hornbacker (Ex. 1003). These references collectively teach all features of claims 1-21, and a person of ordinary skill in the art would be motivated to combine the teachings of the references for the reasons discussed below.

103. Reddy and Hornbacker provide related, contemporary teachings about the state of the art in 2D and 3D visualization of large image data sets such as maps that give examples of reasons why the claims of the 506 Patent cover technology that was already well-known to those of skill in the art. Reddy describes in detail a versatile software system for retrieving geographic information over the Internet or WWW and displaying it to a user in a dynamic, three-dimensional manner using

well-known level-of-detail management techniques to optimize the use of limited bandwidth to retrieve geographic data.

104. Reddy teaches a system for disseminating and viewing massive terrain databases and 3D maps over the Web using a VRML browser or a customized browser or a plug-in to a web browser such as Netscape Navigator or Internet Explorer. The 3D map images are stored in a hierarchical data structure made up of a multiresolution hierarchy of images in which an original image and its downsampled versions at multiple resolutions are stored.

105. Hornbacker teaches a system for retrieving portions of very large documents over a network on a client device, utilizing power-of-two tile pyramids with a specific structure for requesting tiles based on resolution and location. Hornbacker further teaches several well-known techniques for implementing such a system on a common client device with a limited bandwidth connection, such as fixed compression ratios and progressive resolution enhancement. I discuss each of these references in more detail below.

106. In my opinion, Reddy and Hornbacker are each directed to substantially similar problems to the 506 Patent and to each other, i.e. downloading and displaying large images on devices over limited bandwidth connections, and the teachings of these two references illustrate that claims 1-21 of the 506 Patent

do no more than articulate the use of known features in the art according to their known purposes.

1. Overview of Reddy and Hornbacker

107. Reddy is an IEEE publication that was published in the March/April 1999 issue of IEEE Computer Graphics and Applications journal. The authors worked at SRI International, where they developed a system called TerraVision II for visualization of massive terrain databases using the VRML language. Ex. 1004 at Title. TerraVision II was designed to allow users to access large terrain databases via a network connection such as the World Wide Web or the Internet. *Id.* at ¶ 1.

108. The TerraVision II system built on earlier work over several years by SRI on an earlier version of the TerraVision system, which was designed to operate in connection with a project called MAGIC (“Multidimensional Applications and Gigabit Internetwork Consortium”) for the visualization of large amounts of three-dimensional data over a high-speed ATM network. I discussed a 1996 paper regarding earlier work on the MAGIC project, Fuller *et al.* (App. E) previously in this declaration. TerraVision, MAGIC, and TerraVision II were all funded partially by Defense Advanced Research Projects Agency (DARPA). *See* Ex. 1004, p. 37 (Acknowledgements).

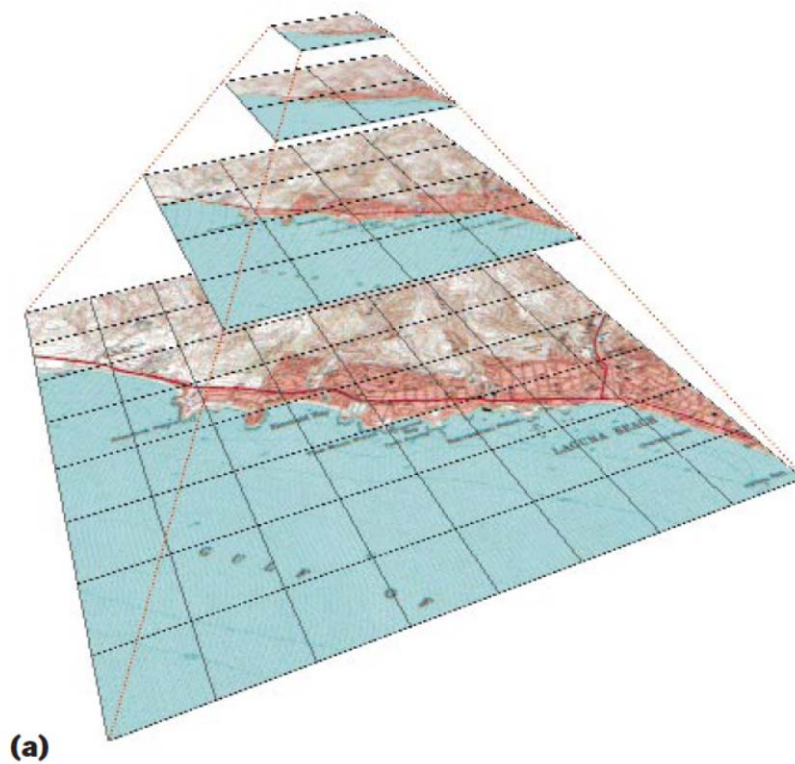
109. As Reddy and his co-authors describe, TerraVision II extends the work performed in connection with TerraVision and MAGIC to a software program that could access data over a conventional WWW connection (not just a high-speed connection) and could operate on a variety of devices such as PCs and laptops, by working in connection with standard Internet browsers.

110. Reddy discloses that there was an increasing interest and need for researchers, including geographers, cartographers, geologists, and computer scientists, to access 3D maps and spatial data over the World Wide Web or even 28.8K modem connections. *Id.* at inset on p. 30. Reddy notes that, however, traditional single-resolution VRML images did not scale well for such use. *Id.* at ¶ 1 For example, Reddy notes that some terrain models could run into hundreds of gigabytes in size. *Id.* at ¶ 2.

111. To allow user access to map details, e.g., being able to view a particular building in a particular city starting from a satellite image of the earth Reddy discloses a technique in which data is progressively downloaded to the user as the user performs rotations and zooms. *Id.* at ¶ 3. In the TerraVision II system, the progressive downloading functionality was implemented on user devices in multiple ways including a VRML browser, a VRML plug-in to a web browser such as Netscape Communicator or Microsoft Internet Explorer or a custom TerraVision II browser. *Id.* at ¶ 31.

112. A key feature of Reddy involved generating a pyramid of multi-resolution images of an original image. *Id.* at ¶ 15. For example, if the original image is 1024 x 1024 pixels, then the pyramid might contain the original image along with down-sampled versions different levels of detail (LOD) at resolutions of 512 x 512 pixels, 256 x 256 pixels, 128 x 128 pixels, and so on. *Id.* at ¶ 15.

113. As Figure 1a shows, each pyramid image is then segmented into rectangular tiles, where all tiles have the same pixel dimensions.



114. Because each tile has the same pixel dimensions, the resolution and the number of tiles is reduced by a factor of two at each down-sampled level. The

following chart summarizes the example disclosed by Reddy in Fig. 1 and the accompanying text at ¶¶15-17 (referring to the “source” layer as level 0):

Layer	Total dimensions of layer (pixels)	Tiles in layer	Tile dimensions (pixels per tile)	Each pixel at this layer corresponds to ____ pixels in the source layer
0 (source)	1024x1024	8x8	128x128	1
1	512x512	4x4	128x128	4 (2x2)
2	256x256	2x2	128x128	16 (4x4)
3	128x128	1	128x128	64 (8x8)

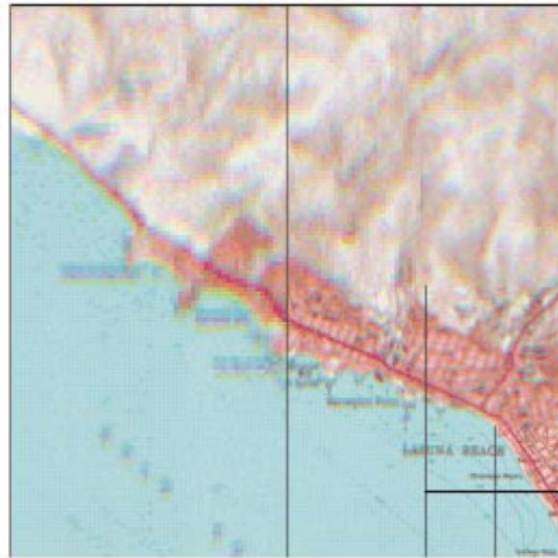
115. Using Figure 1b, Reddy further gives an example of how the multi-resolution pyramid can be used to download tiles at the appropriate resolutions to render a perspective view. *Id.* at ¶¶ 15-17. Reddy describes this as follows: “Figure 1b shows the lower-right corner in high resolution with the surrounding regions displayed in progressively lower resolution. Assuming a tile size of 128 x 128 pixels, this example requires downloading and rendering only 491 Kbytes (10 tiles) instead of the entire 3.1-Mbyte high-resolution image.² If the user’s location is the

² The definitions of “kilobyte” and “megabyte” can be confusing because of an informal custom that has become prevalent in the computer industry. According to the standard definition, the terms “kilobyte” and “megabyte” refer to 1,000 and 1,000,000 bytes, respectively. However, because 2¹⁰ bytes is 1024 bytes, which is very nearly 1000, and 2²⁰ bytes is 1,048,576 bytes, which is very nearly 1,000,000,

bottom-right corner, then distant imagery is rendered at lower resolution than near imagery and we have achieved distance-based LOD.” *Id.* at ¶ 16.³

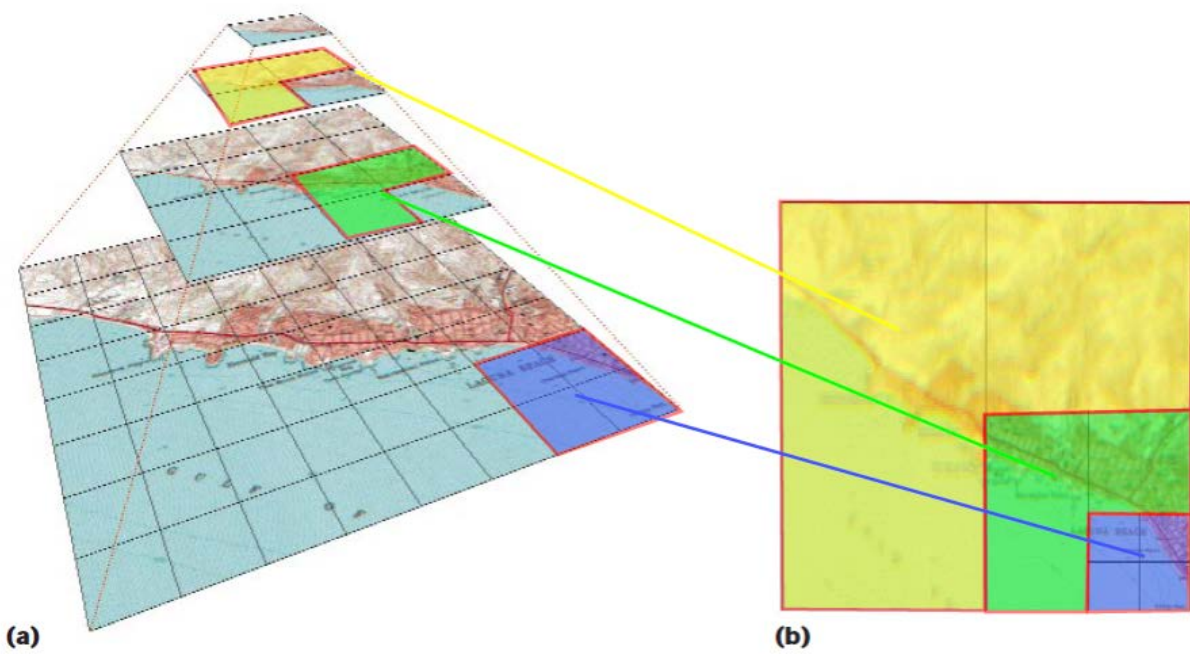
it has become commonplace in the computer industry to refer to 1024 bytes and 1,048,576 bytes as 1 kilobyte and 1 megabyte, respectively, and in fact this alternative interpretation has become incorporated in some standards such as the JEDEC family of memory standards. The reasons for the distinction are described in a National Institute of Standards and Technology webpage available at <http://physics.nist.gov/cuu/Units/binary.html>. I note the difference for purposes of this analysis because the figures given in Reddy for tile sizes suggest that Reddy uses the standard decimal (1000 bytes/kilobyte) notation while the 506 Patent and Hornbacker use the binary (1024 bytes/kilobyte) notation.

³ The numbers in this example work as follows. The original image has a size of 1,024 x 1,024 pixels, at 24 bits per pixel (e.g., 8 bits for each color GB). This adds up to 1024 x 1024 pixels x 24 bits/pixel = 25165824 bits = 3145728 bytes = 3.145728 MB (assuming 1Mbyte = 1 million bytes). Similarly, 10 tiles at 128 x 128 resolution each add up to 128 x 128 x 24 bits/pixel x 10 = 3932160 bits = 491420 bytes = 491.42 KB (assuming 1 KB = 1,000 bytes).



(b)

116. The graphic below, based on Fig. 1 of Reddy, illustrates how ten of the tiles from the different levels in Fig. 1(a) correspond to Fig. 1(b), which shows the tiles that would be used to render a perspective view:



(a)

(b)

117. Another important aspect of Reddy is that the image data is downloaded from the image database based on a user's viewpoint. *Id.* at ¶ 3. For example, if a user's location is the bottom-right corner, then distant imagery is rendered at lower resolution than near imagery and we have achieved distance-based LOD. *Id.* at ¶ 16. Thus, Reddy's technique only needs to fetch and display data for the region that the user is viewing and only at a sufficient resolution for the user's viewpoint. *Id.* at ¶ 17. As the user zooms into an image, the program downloads imagery at higher and higher resolutions. *Id.* at ¶ 3.

2. Overview of Hornbacker

118. Hornbacker teaches a method and a system for displaying portions of very large images (such as digital documents) retrieved over a network from a server. The system described in Hornbacker includes a web server networked to client workstations, which use a web browser on the workstation to request image views by means of Uniform Resource Locator (URL) code using the HTTP protocol. Ex. 1003 at 5:3-8, 5:16-25. Hornbacker provides several relevant details that further explain the implementation of the process of requesting and transmitting tiles. Reddy teaches that map data and images are divided into 128 x 128 pixel view tiles, which are further organized into a hierarchy of tiles at differing resolutions spaced by factors of two.

119. Hornbacker explains that tiles may be compressed at a ratio of 4:1 using GIF compression, which yields a data size for a view tile of approximately 512 bytes. *Id.* at 6:26-28. Hornbacker teaches the use of the TCP/IP protocol, which relies on packet transmission, and further explains a reason why a person of ordinary skill in the art would choose to design a system of this type such that individual view tiles are transmitted as individual packets. Specifically, Hornbacker teaches that the view tile size is optimized to avoid the increase of “network transaction overhead... with smaller view tiles since each of the view tiles requires a network transaction.” *Id.*, 8:13-14. In my opinion, this disclosure suggests that the “network transactions” referenced are packet transmissions, and this teaching would suggest to a person of ordinary skill in the art that splitting tile information into too many packets would excessively increase the “overhead,” or that amount of bandwidth taken up by the header information for packets.

120. Hornbacker further explains that individual tiles may be located using a URL code which specifically identifies the tiles based on factors including the source image, scale, location of the tile in X and Y dimensions, view rotation angle, etc. *Id.* at 8:30-10:2.

3. Motivations to Combine Reddy and Hornbacker.

121. It is my opinion that a person of ordinary skill in the art would be motivated to combine teachings in Reddy and teachings in Hornbacker for several

reasons. First, both references are analogous art because they both address the common technical issues in visualizing large amounts of data obtained over a data network as such the Web via HTTP protocol, using a client viewing device with much smaller memory than the database which stores the imagery data. For example, Reddy describes how the user can visualize part of a multi-gigabit underlying terrain data set by downloading just the few tiles necessary to render the current view. Both Reddy and Hornbacker address similar or the same technical problems in rendering the images on the client device from image data received over a data network (e.g. optimizing bandwidth, prioritizing use of bandwidth, determining which portions of a larger set of image data to request, etc.). The problem addressed by both references, at a very high level, is how to view portions of a very large remote image data set using a client device and a connection with limited bandwidth, or on a connection where the client device has to share bandwidth with other devices, where the user only needs to see some portion of the data at a given time but the resolution at which the user may need to see the data varies. Both references reach solutions to this problem which include the transmission of large images (e.g. maps) which are divided into arrays of tiles at various resolutions within a power-of-two pyramid. Accordingly, in my opinion, a person of ordinary skill in the art in late 2000 would have viewed Reddy and Hornbacker as analogous art. Both references teach similar solutions to bandwidth

issues, including dividing large references into tiles, further processing those tiles into a “pyramid” structure containing representations of those tiles at differing resolutions, caching algorithms to optimize which tiles are requested and stored, and compression of tiles. Therefore, a person of ordinary skill in the art faced with the technical difficulties presented by transmitting large images over a narrow communication bandwidth in the late 1990s would view Hornbacker as relevant art to the technical challenges described by Reddy, and would therefore be motivated to consider the solutions taught by Hornbacker to the similar technical challenges when designing map visualization system like that taught by Reddy.

122. In addition, Reddy and Hornbacker are analogous art because both of them are applicable to mapping related applications. Reddy specifically discloses uses of its technology in terrain visualization applications and map applications. The teachings of Hornbacker are readily applicable to online mapping references too because online maps represent a scenario in which a much larger amount of geographically organized imagery must be stored on a server than can be stored at one time on a client.

123. In particular, a person of ordinary skill in the art would recognize that one challenge faced by Reddy is optimizing the use of limited bandwidth and limited computing resources on a client device when accessing a large geographic database. For example, Reddy teaches in ¶48 that the software may be

implemented on a PC or a laptop in a “distributed, time-critical” environment such as military mission planning, battle damage assessments, and emergency relief efforts. In my opinion, a person of ordinary skill in the art would recognize that such scenarios would be very likely to encounter limited bandwidth conditions, particularly in the field (e.g. in a field expedient command center with corresponding field expedient data connectivity), and would be motivated to design such a system to be robust enough to be capable of operating under such limited bandwidth conditions. Additionally, Reddy specifically teaches that its system can be used on a PC (personal computer) or a laptop computer, which indicates that a primary purpose of the TerraVision II system is to expand on the previous TerraVision system by expanding it to be operable on a wide range of devices, including the type of devices suitable for mobile or field use in the scenarios described in Reddy. These teachings of Reddy would motivate a person of ordinary skill in the art to consider teachings relevant to how to (1) make the software operate effectively in on a portable device capable of operating in an environment such as emergency relief, and (2) retrieve large amounts of data under conditions of limited bandwidth. In my opinion, based on my review of Reddy and Hornbacker as well as my knowledge of the art of computer graphics and networking, it is my opinion that a person of ordinary skill in the art would consider Hornbacker relevant to both questions.

124. For example, Hornbacker teaches that the system using common web browser software on a client workstation (Ex. 1003, 6:16-21) can operate in a limited bandwidth environment 28.8 kilobaud modem network (*id.*, 13:28-30), and is “available on all common workstation types as well [*sic*] other devices such as notebook computers, palm-to computers, Network Computers and Web television adapters to provide a widely available solution.” *Id.*, 14:26-28. Therefore, in my opinion, a person of ordinary skill in the art would be motivated by the teachings of Reddy, in light of Hornbacker, to operate the geographic browser software of Reddy on a variety of common workstation types, including the types of devices taught by Hornbacker, in conditions of limited bandwidth, and to have a reasonable expectation of success doing so.

125. Additionally, it is my opinion that Reddy’s teaching of the need to optimize the use of bandwidth (e.g. retrieving only needed tiles) to operate under conditions where bandwidth is scarce would also lead a person of ordinary skill in the art to consider and use compression in order to reduce the bandwidth used per tile, and therefore the total bandwidth. It is my opinion based on my personal knowledge of the field that a person of ordinary skill in the art would consider a wide variety of image compression techniques to be available as of 1999. Specifically, Reddy mentions the VRML97 specification (*see, e.g.* Ex. 1004, ¶¶21, 24) which, as I mentioned above at paragraph 69 of this declaration, supported the

use of various types of image compression. Therefore, a person of ordinary skill in the art would recognize that the system taught by Reddy is ready for improvement using image compression. Utilizing compression as taught by Hornbacker in the geographic visualization system of Reddy would be a simple matter of using known techniques with predictable (and desirable) results.

126. A person of ordinary skill in the art would also recognize that the system of Reddy requires a way to identify desired tiles based on their location and LOD, and request only those tiles specifically needed. For example, Fig. 1(a) and (b) and the corresponding text at ¶¶15-17 teach that the system retrieves the specific files shown in Fig. 1(b) to render a perspective view, but not the remainder of the 3.1 Mbyte high-resolution image. Because Reddy teaches that the system is implemented in conjunction with a common web browser, in my opinion, a person of ordinary skill in the art would recognize that the most likely way to request particular tiles is using URLs, which are mentioned in ¶26. A person of ordinary skill in the art would also recognize that using URLs to request tiles from a large terrain database would preferably require an organized system to uniquely identify each tile. Hornbacker teaches such a system, which composes URLs to request particular tiles based on a tile's zoom level (level of detail) and position (row and column) within an image.

127. In the following sections, I will discuss other specific reasons or motivations that would cause a person of ordinary skill in the art to combine Reddy and Hornbacker when analyzing these references in connection with claim limitations of claims 1-21 of the 506 Patent.

4. Claim 1 is Rendered Obvious by Reddy in view of Hornbacker

128. **Element 1.Preamble** A method of retrieving large-scale images over network communications channels for display on a limited communication bandwidth computer device, said method comprising: In my opinion, the preamble of claim 1 is taught by both Reddy and Hornbacker, which collectively describe systems that perform the claimed method and the use of those systems in the claimed manner.

129. The term “Limited communication bandwidth” is not explicitly defined in the specification of the 506 Patent. However, the 506 Patent describes that limited bandwidth conditions may exist due to either the direct technological constraints dictated by the use of a low bandwidth data channel or indirect constraints imposed on relatively high-bandwidth channels by high concurrent user loads. Ex. 1002 at 3:12-22. In my opinion, Reddy in view of Hornbacker teaches communication over a “limited communication bandwidth” channel under either interpretation.

130. Reddy teaches a system for retrieving “massive” terrain data sets including satellite and aerial imagery (which are typically very large-scale images, as Reddy teaches in ¶1, which suggests that photorealistic terrain models may be on the order of “hundreds of gigabytes”) over the Internet or World Wide Web (“WWW”) (network communications channels). *See, e.g.* Ex. 1004 at p. 30 (subtitle), ¶¶1, 5, 9, 10, 12, 31. Reddy teaches a system that can be implemented as a plug-in in connection with a standard web browser (Ex. 1004 at ¶¶ 31, 32), and that the system may be implemented on, for example, a PC connected to the Internet or a laptop machine (limited communication bandwidth computer device), which makes the system particularly useful in “military mission planning... emergency relief efforts, and other distributed time-critical conditions.” *Id.*, ¶48. The benefits of using a portable device (such as a laptop) in a military or emergency response scenario (where users of the system might be, for example, operating out of a mobile or field expedient command center with similarly field expedient network connections) should be readily apparent to the user. Reddy further teaches that the tiling employed by its system enables a user to visualize a scene utilizing a much smaller amount of downloaded data than the full-resolution underlying image. *Id.*, ¶16. A person of ordinary skill in the art would recognize that a goal of Reddy is enabling access to large databases of mapping information

under conditions of limited bandwidth, like those that would be encountered in a military or emergency response scenario.

131. In my opinion, a person of ordinary skill in the art would further recognize that Reddy's use of commonly known web browsers for accessing map information over the Internet would enable any device that is connected to the Internet, and which uses one of these known browsers, to access map data regardless of the speed at which that Internet connection operates. This is so because, as I discussed previously in section VI.A, the TCP/IP protocol provides a layer of abstraction for the data being sent; in other words, as long as there is a means to transport TCP/IP packets, virtually any type of digital data can be sent using TCP/IP packets.

132. It is also my opinion that a person of ordinary skill in the art would recognize the need to reduce bandwidth consumption in situations where multiple users need to share the use of a common communications channel. In my opinion, the military and disaster relief scenarios taught by Reddy are exemplary of situations where a person of ordinary skill in the art could expect a large group of users (e.g. the staff of a command center) to have to share access to a common network connection under field expedient conditions. Under such conditions, a person of ordinary skill in the art would recognize that the bandwidth available to each user would be limited by the need to share the connection, as well as the need

to minimize each user's use of bandwidth to avoid adversely impacting others.

Similar constraints could arise even in typical consumer or office use scenarios, e.g. when a large number of users of an Internet gateway approach the bandwidth limits of the gateway.

133. I previously explained the basis of my opinion that Hornbacker is in a related field to Reddy. Hornbacker teaches a system for using graphical web browsers on client systems to view large images divided into tiles. Ex. 1003, Abstract, 6:20-7:1, 13:28-14:11, 14:26-28. Hornbacker specifically teaches that individual view tiles can be compressed from an uncompressed size of 2048 bytes (based on a monochrome image) to as little as 512 bytes (6:20-7:1), and viewed using a "low bandwidth 28.8 kilobaud modem network" with "much lower demand on the network connection." *Id.*, 13:28-14:11.

134. Reddy and Hornbacker both address overlapping issues with downloading large sets of imagery over a bandwidth constrained system and reach similar solutions using arrays of tiles. In my opinion, a person of ordinary skill in the art considering the problems articulated by Reddy (i.e. viewing large data sets on a bandwidth-constrained device such as a laptop in a distributed computing environment) would therefore look to the similar teachings of Hornbacker to solve those problems using solutions such as the use of compression to reduce byte size of tiles. In other words, both references involve requesting and delivering 128 x

128 pixel tiles of image data from a server to a client. A person of ordinary skill in the art would recognize that Hornbacker's methods for reducing the size of those tiles would be provide the same benefit (reducing bandwidth) in the system of Reddy.

135. Additionally, Reddy teaches that the browsing system may be implemented on a laptop computer (mobile computer system) (Ex. 1004, ¶48), while Hornbacker teaches that the browsing system may be implemented on a notebook computer, palm-top computer, or television (Ex. 1003, ¶¶14:16-28). I note that the 506 Patent itself includes each of a mobile computer, handheld computer, palm-top computer, and a television as examples of a "limited communication bandwidth computer device." In my opinion, Reddy's teaching of a laptop computer corresponds to the "mobile computer" claimed by Bradium in the 506 Patent as an example of a "limited communication bandwidth computer device," and therefore Reddy itself teaches this element. However, it is also my opinion that a person of ordinary skill in the art would be motivated by Reddy's teaching of a system capable of operating in distributed, time-critical environments to implement the appropriate software on a handheld computer, palm-top computer, or television as taught by Hornbacker in order to enable the user to access map data in a field environment.

136. **Element 1.A** issuing, from a limited communication bandwidth computer device to a remote computer, a request for an update data parcel; In my opinion, this claim element is disclosed by both Reddy and Hornbacker.

137. I previously explained in regard to the preamble why both Reddy and Hornbacker teach or suggest “limited communication bandwidth computer devices.” Reddy teaches that geographic data (such as satellite or aerial imagery) is divided into tiles, each of which is an element of a source image array. Ex. 1004, ¶¶ 14-18, 22, 24, 41-46, Figs. 1, 3. Because Reddy teaches downloading individual tiles (*i.e.*, data representing an element of a source image array), it is my opinion that Reddy teaches an “update data parcel.” Hornbacker likewise teaches that image data is divided into tiles, which are elements of a source image array. Ex. 1003 at Abstract, 3:10-27, 6:13-19, 7:26-8:6, 8:30-9:28, 10:24-28, 12:24-13:10. These teachings of tiles in Reddy and Hornbacker correspond to the teachings of a source image divided into an array in the 506 Patent. Ex. 1002 at 6:12-22. Therefore, the data representing the tiles in both references is data that corresponds to one element of a source image array.

138. Reddy further teaches that the tiles are retrieved based on the user’s selected view (Ex. 1004, ¶¶ 3, 10, 14-17, 31, 34-38, 42, 44-46, Figs. 1, 5). As Reddy and co-authors describe in the introduction, when a user zooms into a target region (e.g. “flying” from a view of the whole earth down to a particular region

and then a specific point such as a building), higher resolution data is progressively downloaded and displayed. Ex. 1004, ¶3. Reddy teaches that the system need only fetch and display data for the region that the user is viewing, because the full data set representing the underlying image data may be massive and therefore impractical to download and store on a client device. Ex. 1004, ¶17. Thus, Reddy discloses that a user's computer downloads, via the Internet or WWW (from "a remote computer"), visual data such as higher resolution data, or data for region that the user is viewing ("data parcel") that is progressively downloaded based on user's viewpoint ("update" data parcel).

139. Reddy discloses that the tiles may be retrieved by Universal Resource Locator (URL). Id., ¶¶ 21, 26, 52. As I discussed above regarding claim 1, preamble, Reddy teaches that the system is implemented in conjunction with a web browser. The normal operation of Reddy's system utilizes "geotiles," which contain links to terrain tiles such as satellite, aerial, and map imagery. Ex. 1004, ¶22, Fig. 3. The conventional web browsers taught by Reddy are used to access web pages. Fundamental to the operation of the World Wide Web is the notion of a browser making hypertext transfer protocol, or HTTP, requests to an HTTP server. A POSITA would understand that by convention, these world wide web requests are made by sending requests for content to an HTTP server computer on TCP/IP port 80 (normal http requests) or on TCP/IP port 443 (secure https

requests). Requests from an HTTP client application, such as a known web browser such as those cited by Reddy and Hornbacker, are then responded to by the HTTP server, again using TCP/IP on ports 80 or 443.⁴

140. Therefore, in my opinion, a person of ordinary skill in the art would understand these teachings of Reddy, in view of the knowledge of a person of ordinary skill in the art and the background art of web browser software and the VRML, to teach that the geographic browser of Reddy sends requests to a server (remote computer) to retrieve geotiles containing links to imagery files, which are then requested by URL.

141. These teachings are analogous to the 506 Patent's use of HTML requests. Ex. 1002, 7:16-20.

142. Hornbacker further teaches that tiles within a large database may be located by requests to a server which use URLs to identify the specific tile by characteristics such as resolution, location, view angle, etc. See, e.g. Ex. 1003, Abstract, 3:10-27, 5:16-25, 6:13-19, 7:26-8:6, 8:30-9:28, 10:24-28, 12:24-13:10. In my opinion, a person of ordinary skill in the art would recognize that the teachings of the two references solving similar problems in closely related fields could be considered in combination when designing a display system addressing a similar problem.

⁴ While it is possible to use other TCP/IP ports to perform these transactions, by 1994 ports 80 and 443 had been adopted for http and https respectively. See RFC1700 pages 20 and 34 respectively.

143. A person of ordinary skill in the art facing the problem identified by Reddy, which includes how to identify the tiles desired to render a particular geographic view, would look to the solution taught by the analogous Hornbacker reference of identifying tiles by URL based on their coordinates and other viewing characteristics in order to provide an efficient way to identify needed tiles. Reddy and Hornbacker identify a common problem with the display of portions of very large images over a network, which is the latency and bandwidth consumption associated with downloading an entire image, and arrive at similar solutions in the form of requests for specific tiles in a priority order based on how soon the tiles need to be displayed. In my opinion, a person of ordinary skill in the art would consider the teachings of both references when attempting to design a system addressing these problems. Hornbacker provides further implementation details (such as fixed ratio compression and a URL syntax for identifying specific needed tiles) that complement and could be used to further refine the tile request process described by Reddy. Therefore, in my opinion, this element would be obvious to a person of ordinary skill in the art.

144. **Element 1.B** wherein the update data parcel is selected based on an operator controlled image viewpoint on the computer device relative to a predetermined image; In my opinion, Reddy teaches this element. Reddy teaches a system designed to enable a user to view geographic information or imagery

downloaded over the web using a 2D pan-and-zoom display or three-dimensional simulated viewpoint chosen or navigated to by the user/operator. Ex. 1004, ¶¶ 2-3, 13-17, 21, 38, 42, Fig. 5. Tiles are selected based on the user's proximity to the tile in question. *Id.*, ¶¶ 12-17, 19-22, 29, 42-46, Fig. 1, 4, 5. For example, Fig. 1(b) and accompanying text describe how higher resolution tiles are downloaded for the area nearest the user's viewpoint, while paragraph 21 teaches that "when the user approaches a region of terrain, more detail is progressively loaded and displayed in a coarse-to-fine fashion. In my opinion, a person of ordinary skill in the art would recognize these teachings to disclose that the update data parcel (terrain tiles) are selected based on an operator controlled image viewpoint (user perspective) relative to a predetermined image (the source imagery/map data that the user is viewing). Ex. 1004, ¶3.

145. As discussed in regard to claim element 1.A, the teachings in Reddy of a system that selects tiles based on user viewpoint in relation to a coordinate system would lead a person of ordinary skill in the art to consider Hornbacker's teaching that URLs containing the zoom level and coordinates of a tile are an advantageous way to locate and identify tiles at the appropriate resolution within the quad-tree structure taught by Reddy. *Id.*, ¶21.

146. **Element 1.C** the update data parcel contains data that is used to generate a display on the limited communication bandwidth computer device;

Reddy teaches a system for “visualizing” the map data (e.g. imagery) obtained by the TerraVision II system on the client device. I previously discussed regarding claim 1, preamble how the client device taught by Reddy may be a limited communication bandwidth computer device.

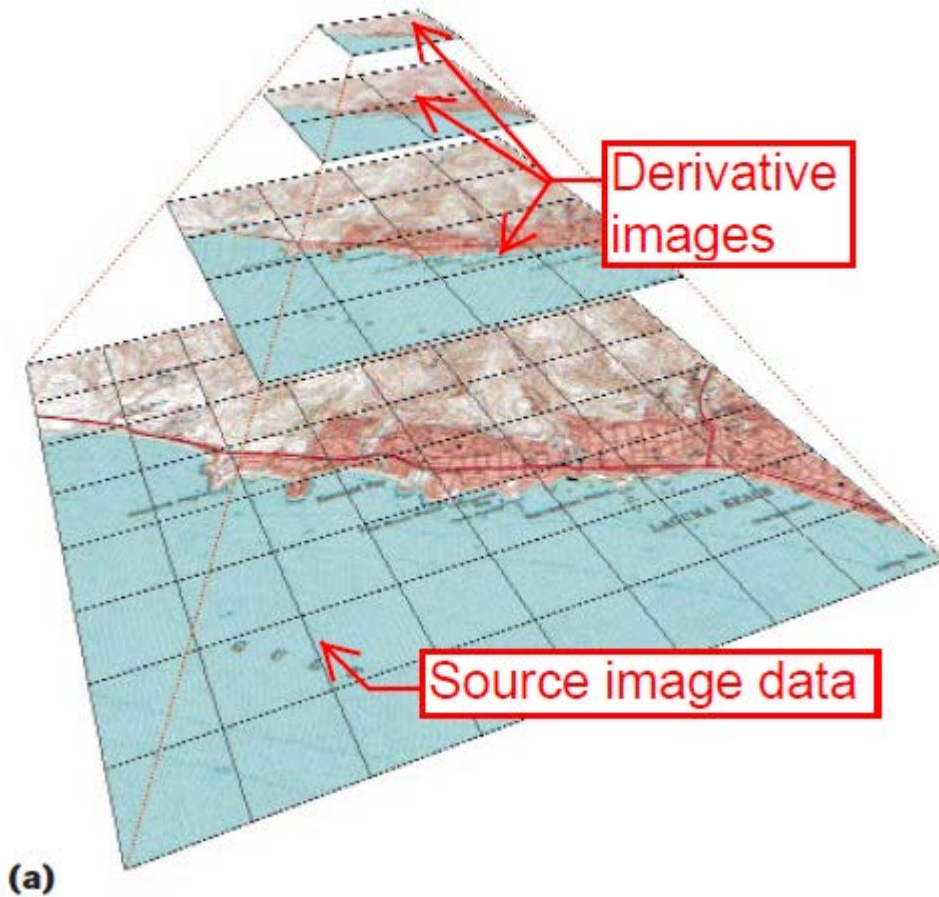
147. As I discussed above regarding claim element 1.B, the displayed image is generated using terrain tiles, including aerial, satellite, or map imagery that is downloaded and displayed to the user. Ex. 1004 ¶3. Therefore, the terrain tiles (update data parcels) comprise data that is used to put together the visualization of the terrain (generates a display) on the client (limited communication bandwidth computer device).

148. Hornbacker also teaches a system that uses image data tiles retrieved over the Internet to form an image on the browser at the client device. Ex. 1003 at Abstract.

149. **Element 1.D** processing, on the remote computer, source image data to obtain a series K_{1-N} of derivative images of progressively lower image resolution; In my opinion, both Reddy and Hornbacker teach this element because both references teach that the server systems for providing imagery may process source image data into a grid of tiles organized into a hierarchy of different resolution levels. As to the “processing” on the “remote computer,” Reddy discloses teaches two alternatives. First, the required terrain data may be pre-

computed offline. Ex. 1004, ¶52. Second, terrain data may be generated “on the fly” by parsing the URL path name, using a script on the server, to generate the necessary VRML data. *Id.* In other words, in this second alternative, if processed data is not already available (which is a situation that might come up in a time-critical scenario, e.g. military or disaster response, where new imagery is being made available for a first time) the script can parse the URL request, decide what data needs to be made available, and then conduct the necessary processing in order to make that data available.

150. I understand the phrase “series K_{1-N} of derivative images of progressively lower resolution” to mean that there is a source image, followed by a series of layers of tiles in which the source image has been divided into a derivative layer of tiles at lower resolution. For example, at layer K_{N+1} , there will be 1/4 as many tiles as there are at layer K_N , each tile having half the resolution of layer K_N . In other words, layer K_{N+1} is a derivative of layer K_N in which every tile in layer K_{N+1} is the result of combining an array of four adjacent K_N tiles at lower resolution. This system corresponds with the teachings of Reddy and Hornbacher, and is illustrated by Fig. 1 of Reddy:



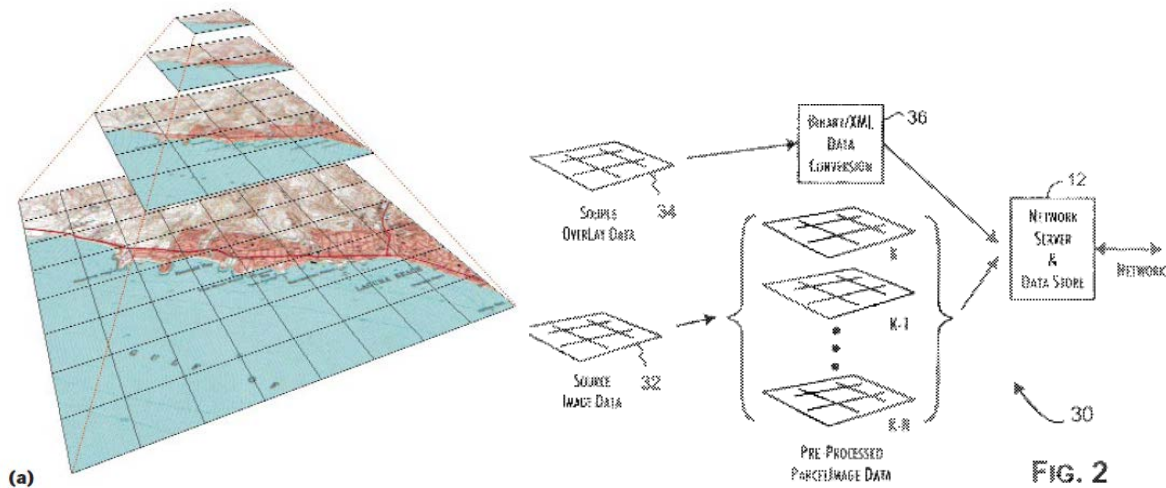
151. In my opinion, a person of ordinary skill in the art would understand the teaching of Reddy to mean that a remote computer generates the series of derivative images.

152. Hornbacker also discloses that view tiles are generated at the server (remote computer) by an image tiling routine that divides a given image into a grid of smaller images, which are further computed for distinct resolutions. The view tiles may either be pre-processed at the server (pre-cached) or newly computed in response to a request. Ex. 1003 at 3:22-27, 5:3-8, 5:16-6:19, 6:20-7:25, 8:30-9:28,

10:3-10, 11:19-28, 12:21-13:10, 13:26-14:6. In the “pre-caching” feature (note that Hornbacker teaches caching on both client and server sides; for this portion of my analysis I am referring specifically to the server side), Hornbacker processes tiles that are likely to be requested in the near future. In my opinion, both references address similar technical issues with a similar technical approach of dividing large high-resolution source image data sets into a pyramid of tiles having different resolutions at each level. A person of ordinary skill in the art would recognize that Reddy teaches the desirability of having a script prepare tiles “on the fly” at the server. In my opinion, Hornbacker provides a detailed explanation of how the processing pipeline to generate tiles either in advance or “on the fly” could operate, and therefore a person of ordinary skill in the art would look to the teachings of Hornbacker in order to successfully implement the “on the fly” tile generation of Reddy. In particular, Hornbacker’s tiling pipeline is designed to prepare a series of derivative images at different resolution levels, which makes it particularly well-suited to preparing geographically-linked images in the multi-resolution pyramid described by Reddy.

153. **Element 1.E** wherein series image K_0 being subdivided into a regular array; In my opinion, Reddy teaches this element. Reddy teaches that each layer of the pyramid, including the highest resolution layer (K_0) is “segmented into” an array of tiles, so that each tile at a given level maps onto four tiles at the

next higher level. Ex. 1004, ¶¶12-16, Fig. 1. For example, in Fig. 1, the bottom layer is an 8x8 array of tiles. The rest of the “image pyramid” contains four different resolutions of the original image, each formed into an array ranging from 8x8 tiles at the highest resolution to a single tile at the lowest, with 4x4 and 2x2 arrays in between. This teaching is substantially identical to the 506 Patent’s disclosure in Fig. 2 and at 6:6-20 of the division of source image data into derivative images of progressively lower image resolution, as shown by a comparison of Fig. 1 of Reddy and Fig. 2 of the 506 Patent:



154. **Element 1.F** wherein each resulting image parcel of the array has a predetermined pixel resolution; In my opinion, Reddy teaches this limitation. Reddy teaches that within each pyramid image, “all tiles have the same pixel dimensions,” and repeatedly discusses an example in which each tile is 128x128 pixels. Ex. 1004, ¶¶ 15-16, Fig. 1. This teaching is similar to the support for this

limitation in the 506 Patent at 6:12-17, which says that “each resulting image parcel of the array has for example at 64 x 64 resolution.” Hornbacker contains similar teachings that tiles are preferably fixed as 128 x 128 pixel image files, and that fixed size tiles (i.e., tiles that have the same pixel dimensions) allow more efficient use of the caching mechanism. Ex. 1003 at 6:20-7:25, 8:30-9:28, 10:3-10, 11:19-28, 12:21-13:10, 13:26-14:6. Although it is my opinion that Reddy itself teaches this limitation, I note the similarity with Hornbacker because it demonstrates the similarity of the two reasons and why a person of ordinary skill in the art would have a reasonable probability of success adapting features from the system of Hornbacker to improve the system of Reddy.

155. **Element 1.G** wherein image data has a color or bit per pixel depth representing a data parcel size of a predetermined number of bytes. In my opinion, this element is obvious over Reddy.

156. In my opinion, there are at least two ways that a person of ordinary skill in the art reading Reddy in light of the existing knowledge in the art would recognize that Reddy teaches that the use of image data having a fixed color or bit per pixel depth. First, Reddy teaches the use of known imagery formats such as Portable Bitmap (PBM) and Land Analysis System (LAS) bitmaps. Both of these formats have a fixed bit depth. PBM (see

http://oceancolor.gsfc.nasa.gov/staff/norman/seawifs_image_cookbook/faux_shuttl

[e/pbm.html](#)) is known in the art to be a monochrome format (1 bit/pixel), while LAS also uses bitmaps. In my opinion, a person of ordinary skill in the art would recognize both formats as having a fixed bit per pixel depth. Ex. 1004, p. 31 (sidebar).

157. Additionally, a person of ordinary skill in the art would be able to confirm that Reddy teaches tiles having a fixed bit per pixel depth based on the calculations given in ¶¶15-17. The size of the data representing an uncompressed tile is simply the product of the bit depth (bits/pixel) multiplied by the pixel dimensions. For example, a 128x128 pixel tile (16,384 pixels) with 8-bit RGB color (e.g., one byte for each of the three colors) would occupy approximately 49 Kbytes (1 byte is 8 bits) on disk.

158. Reddy discloses at ¶16 that the example in ¶15 and Fig. 1 takes up 491 Kbytes for 10 tiles and 3.1Mbytes for the full high-resolution (1024x1024 pixel) image. A person of ordinary skill in the art would understand from this teaching that the data parcel size for each tile is the same because each tile has a bit depth of 24 bits per pixel, or 8-bit RGB (red, green, blue) color, yielding the total sizes in ¶16.

159. This teaching is similar to the support for this claim element in the 506 Patent at 6:12-22, which teaches that a 64x64 pixel parcel with a 16-bit color

depth has a resulting data parcel size of approximately 8 Kbytes (using the 1024 bytes/KB convention).

160. Similarly, Hornbacker teaches the use of GIF compression with a fixed size (for monochrome tiles before compression) of 2 KB (using the 1024 bytes/KB convention). A person of ordinary skill in the art would recognize that this teaching also reflects a fixed data parcel size that is dependent on the number of bits per pixel (color depth). Ex. 1003, 6:20-7:3.

161. A person of ordinary skill in the art would recognize that the same principles would apply to color and grayscale images; e.g., while an 8-bit grayscale image would take up 8 times the space of a similarly proportioned monochrome (1 bit) image and a color image might be 16 or 24 bits (among other possibilities depending on the exact format used) in each case the use of constant pixel resolution and constant bit depth results in constant size on disk for the data parcel.
[start here]

162. **Element 1.H** resolution of the series K_{1-N} of derivative images being related to that of the source image data or predecessor image in the series by a factor of two; and said array subdivision being related by a factor of two such that each image parcel being of a fixed byte size; Reddy teaches that image data is stored as a pyramid, which is a multiresolution hierarchy, with source images down-sampled according to powers of two. Ex. 1004, ¶¶14-15. The image data that

makes up Reddy’s pyramid thus includes images that are derived from a top-level source image (“derivative images being related to that of the source image data or predecessor image in the series”). For example, a 1024x1024 original image gets down-sampled to 512x512 pixels, then 256x256 pixels, and so on. Ex. 1004, ¶¶14-15, Fig. 1. Because “all tiles have the same pixel dimensions,” when a source image array is down-sampled to produce the next lower resolution level in the pyramid, the next level will have half as many tiles in each direction (1/4 the total number of tiles) (array subdivision being related by a factor of two) and each tile will have half the pixel resolution as the predecessor image in each direction (resolution... being related to that of the source image data or predecessor image in the series by a factor of two). In other words, an 8x8 array of tiles covering a 1024x1024 pixel area is down-sampled to a 4x4 array of tiles covering the same area at half the resolution. Reddy expresses this relationship by stating that “a tile at a given pyramid level will thus map onto four tiles at the next higher level; that is, at each higher resolution area, the tiles cover half the geographical area of the previous level.” Ex. 1004, ¶15. The following chart (repeated for convenience) summarizes the example disclosed by Reddy in Fig. 1 and the accompanying text at ¶¶15-17 (referring to the “source” layer as level 0):

Layer	Total dimensions of layer (pixels)	Tiles in layer	Tile dimensions (pixels per	Each pixel at this layer corresponds

			tile)	to ____ pixels in the source layer
0 (source)	1024x1024	8x8	128x128	1
1	512x512	4x4	128x128	4 (2x2)
2	256x256	2x2	128x128	16 (4x4)
3	128x128	1	128x128	64 (8x8)

163. **Element 1.I** receiving said update data parcel from the data parcel stored in the remote computer over a communications channel; As I previously discussed in regard to the preamble of claim 1, as well as claim element 1.A, Reddy teaches a system for browsing geographic data over the WWW (communications channel), which means that image tiles (update data parcels) are received from a server (remote computer).

164. **Element 1.J** displaying on the limited communication bandwidth computer device using the update data parcel that is a part of said predetermined image, an image wherein said update data parcel uniquely forms a discrete portion of said predetermined image. Reddy teaches that the data viewed by the user (i.e. displayed on the client device) includes satellite or aerial imagery which is specific to a particular location, and therefore forms a unique, discrete portion of the predetermined image. Ex. 1004, ¶¶3, 15-16, 18, 24, Fig.

5. Claims 2-7 are Rendered Obvious by Reddy in view of Hornbacker

165. **Claim 2.** The method of claim 1, wherein processing the source image data further comprises one of pre-processing the source image data on the remote computer and processing the source image data in real-time on-demand based on the request for the updated image parcel. In my opinion, and according to the plain language of this claim, “one of” pre-processing source image data and processing the source image data in real-time means that either form of processing satisfies the element. In my opinion, in order to provide tiles processed according to the method of claim 1, the tiles must either have been appropriately processed before receiving the request (*i.e.* pre-processed, so that the tile is already ready to deliver when the server receives the request) or the tile must be processed in response to the server receiving the request (real-time processing). Even without any further explicit disclosure, it is my opinion that a person of ordinary skill in the art would recognize that one of these two steps would have to occur because there are a finite number of ways to deliver a processed parcel of data in response to a request. Even though it is my opinion that this claim element would already be obvious over the teachings discussed in regard to claim 1 for this reason, it is also my opinion that Reddy teaches both options. Reddy teaches in ¶52 that the system may either “statically generate” the data offline (which is the same as pre-

processing, because the data is generated before the server receives the request for it) or generate the data on the fly using a server script that interprets the path name of the received URL and uses that URL to generate the required data from an underlying database (generating in real-time). Reddy's discussion of "distributed, time-critical environments" represents a scenario where real-time processing of data may be preferable because new data could rapidly be becoming available (e.g. military intelligence or battle damage assessment imagery) and need to be analyzed in a distributed manner.

166. Hornbacker also discloses that view tiles are generated at the server by an image tiling routine that divides a given image into a grid of smaller images, which are further computed for distinct resolutions. The view tiles may be pre-processed at the server and pre-cached so that they are ready before they are requested. Ex. 1003 at Fig. 1, 7:26-8:6, 8:30-9:4, 10:13-16, 11:9-12:23. However, Hornbacker also teaches a view composer on the server which may process tiles in real-time in response to requests. *Id.* at Fig. 1, 3:22-27, 5:3-8, 5:16-6:19, 7:19-20, 10:11-23, 11:19-28.

167. In my opinion, the teaching in Reddy to utilize received URLs to generate the necessary geographic tiles "on the fly" would motivate a person of ordinary skill in the art to look to the more extensive teachings in Hornbacker regarding a tile generation pipeline which (like the teaching in Reddy) interprets

the URLs contained in the HTTP requests to identify which tiles to generate, in order to successfully implement real-time processing of image data on demand. Therefore, in my opinion, the additional elements of this claim are taught by Reddy alone as well as Reddy in view of Hornbacker.

168. **Claim 3.** The method of claim 2, wherein receiving the update data parcel over a communications channel further comprises streaming the update data parcel over a communications channel to the limited communication bandwidth computer device. Reddy teaches that the TerraVision II system uses a “progressive coarse-to-fine algorithm to load and display new data,” allowing the user to “continue to interact with the terrain” while higher resolution tiles (update data parcels) are arriving. Ex. 1004, ¶44. Reddy refers to this teaching as a form of “streaming.” *Id.* Additionally, in my opinion, a person of ordinary skill in the art would recognize Reddy’s teachings as a form of “streaming” over the communications channel, because the user is able to visualize and interact with some content while additional content is still being received (analogous to watching part of an online video while the video is still being received).

169. **Claim 4.** The method of claim 1, wherein the limited communication bandwidth computer device further comprises one of a mobile computer system, a cellular computer system, an embedded computer system, a handheld computer system, a personal digital assistants and an internet-capable

digital phone and a television. Reddy teaches that its system may be implemented on a browser on a laptop machine, which is a mobile computer system. Ex. 1004, ¶48. Reddy further teaches the need for a flexible system capable of operating in connection with commonly used web browsers in a “distributed, time-critical environment” such as a disaster response. *Id.* Hornbacker’s teaching of a web browser plug-in capable of operating on a “palm-top computer,” which a person of ordinary skill in the art would recognize as synonymous with a “handheld computer system,” or “personal digital assistant,” meets the need for a device capable of operating in such an environment. Ex. 1003, 14:26-28.

170. **Claim 5.** The method of claim 1, wherein a size of the data parcel on the remote computer is different from the update data parcel on the limited communication bandwidth computer device. This claim element does not specify which size of the data parcel on the remote computer must be different from the size of the parcel on the client, only that *a* size must be. Therefore, different byte sizes satisfy the claim limitation. Hornbacker teaches that tiles may be compressed at a 4:1 compression ratio, so that the client receives tiles at 1/4 the size they were stored on the server. Ex. 1003 at 6:26-28. Therefore, the size of the compressed tile received by the client (update data parcel) is different from the size of the uncompressed tile (data parcel) on the server. Reddy teaches the need to reduce bandwidth use to make the system capable of operating on a distributed network

with limited bandwidth, e.g. a laptop computer in an emergency response situation. *See, e.g.* Ex. 1004, ¶48. In my opinion, a person of ordinary skill in the art would recognize that compression as taught in Hornbacker would meet the goal of reduced bandwidth taught by Reddy. Because Reddy teaches a VRML-based system compatible with compression techniques, a person of ordinary skill in the art would have a reasonable expectation of success implementing compression such as the GIF compression taught by Hornbacker in the three-dimensional geographic data browser of Reddy.

171. **Claim 6.** The method of claim 1, wherein processing the source image data further comprises queuing the update data parcels on the remote computer based on an importance of the update data parcel as determined by the remote computer. Reddy teaches that “TerraVision attempts to predict user’s future moves by a simple extrapolation of their current flight path” and then “pre-fetches” tiles so that they are immediately available. Ex. 1004, ¶ 46. Hornbacker contains a similar teaching, except that it is the *server* that “pre-compute[s] view tiles that may be required by the next view request.” Ex. 1003, 7:26-8:6. Reddy teaches both a desire to predict the user’s future moves in order to make tiles available more readily and that data may be generated “on the fly.” Ex. 1004, ¶ 52. In view of these goals taught by Reddy, a person of ordinary skill in the art would look to the tile pre-computation process taught by Hornbacker as a solution to meet those

goals by enabling the server to respond quickly to new requests for tiles. In other words, Hornbacker teaches a process that would enable the on-the-fly computation of image tiles taught by Reddy. In my opinion, a person of ordinary skill in the art would recognize that the most efficient use of pre-computation of tiles (as taught by Hornbacker) is made when the server prepares the tiles most likely to be requested next by the client, and in the order that they are likely to be requested by the client, so that those tiles can be delivered immediately without additional delay at the server while the server prepares the tiles. In other words, in order to most effectively pre-compute the tiles most likely to be requested, the server should use an algorithm for predicting which tiles will be requested that resembles the algorithm that the client software will use to request tiles.

172. Reddy teaches a “progressive, coarse-to-fine” algorithm based on the resolution of the tiles and the user’s proximity to the tiles. Ex. 1004, ¶¶21, 44, 46. A person of ordinary skill in the art would recognize that in order to make the best use of such a server pre-computing function as disclosed in Hornbacker, in combination with the “progressive, coarse-to-fine” algorithm and pre-fetching of Reddy, the server would optimally calculate the pre-computed tiles in the sequence that they are most likely to be requested by Reddy’s client. In other words, a person of ordinary skill in the art would recognize that the best “pre-computation” algorithm would queue the update data parcels for processing in a way that mirrors

the pre-fetching of Reddy, placing a higher priority on tiles closer to the user position and on lower resolution tiles before higher resolution tiles.

173. In my opinion, a person of ordinary skill in the art would also recognize that there are limited ways to accomplish such a calculation, and a queue would be the most likely way to accomplish this function. Queues are very widely known data structures in the field of computer science, and even an undergraduate computer science would be familiar with the concept of a queue of tasks. A queue is simply a data structure that contains a list of tasks to be performed in a particular order. When a series of tasks needs to be done in a particular order, as in this case where the server is attempting to compute the tiles most likely to be requested next, a person of ordinary skill in the art would recognize a queue as an ideal way to capture the sequence in which new tiles should be generated. Therefore, in my opinion, it would be obvious to a person of ordinary skill in the art over the combined teachings of Reddy and Hornbacker to queue the pre-computed tiles (update data parcels) on the server (remote computer) based on the importance of the tiles (likelihood to be requested by the client) as determined by the server.

174. **Claim 7.A** The method of claim 1, wherein the processing further comprises compressing each data parcel; Reddy does not explicitly teach the use of compression. However, Reddy builds on and references the VRML97 specification

which allows the use of compression, as I discussed previously in section VI.F.3.

Ex. 1004, ¶¶ 1, 21, 24, 28, 34, 39, p. 31 (sidebar).

175. Hornbacker teaches that GIF files used for image tiles may be compressed at a 4:1 ratio so that, for example, a monochrome tile with an uncompressed size of 2,048 bytes reduces to a compressed tile of 512 bytes. Ex. 1003, 6:20-7:3. Reddy teaches the need to reduce bandwidth use to make the system capable of operating on a distributed network with limited bandwidth, e.g. a laptop computer in an emergency response situation. See, e.g. Ex. 1004, ¶48. A person of ordinary skill in the art would recognize that compression as taught in Hornbacker would meet the goal of reduced bandwidth taught by Reddy.

176. Because Reddy teaches a VRML-based system compatible with compression techniques, a person of ordinary skill in the art would have a reasonable expectation of success implementing compression such as the GIF compression taught by Hornbacker in the three-dimensional geographic data browser of Reddy, as the use of GIF files is part of the VRML specification that Reddy relies on.

177. **Claim 7.B** storing each data parcel on the remote computer in a file of defined configuration such that a data parcel can be located by specification of a K_D , X, Y value that represents the data set resolution index D and corresponding image array coordinate. This claim element relates to a way in which the data

resulting from the processing operation is stored on a remote computer for retrieval by a user device. In my opinion, the broadest reasonable interpretation of the phrase “storing each data parcel... in a file of defined configuration” includes storing each data parcel in an individual file corresponding to that data parcel. By way of analogy, consider the following sentences:

- “Each child wore a raincoat to school.”
- “Each child received a Christmas gift.”
- “Each employee was given a bonus.”
- “Each end of the boat was tied up with a rope.”

Since the word “each” connotes individuality, we read these sentences to mean that each child receives their own raincoat, Christmas gift, etc., not that all of the children wear the same raincoat, receive the same gift, etc. Therefore, as a matter of basic grammar, this term includes each data parcel individually being associated with a separate file of defined configuration.

178. I previously discussed with respect to claim element 1.D how Reddy teaches processing of source image data into an image tree. Reddy teaches that the multiresolution images in the image pyramid (described in Ex. 1004, Fig. 1 and ¶¶12-17) can be navigated using a standard quad-tree structure which utilizes “tree” files, described as the “glue” that holds together the “geotiles.” Id., ¶¶19-21. The geotiles in turn contain links to terrain tile files, each of which “contains the

actual elevation and image texture data for a given image, geographical area, and LOD. *Id.*, ¶ 18; see also ¶¶22-24. Reddy teaches that this structure is used to “simplify the task of maintaining and adding new data sets” (*id.*, ¶23) and allows a common navigation structure to link to multiple types of data sets within the hierarchy, including multiresolution image pyramids as well as data representing, e.g., elevation and “features” such as buildings, roads and annotations. *Id.*, ¶¶22, 24-26, Fig. 3. This structure of quad-tree files, which are files of defined configuration used to locate imagery tiles, is very similar to the structure taught by the specification of the 506 Patent, which teaches that “image data parcels are stored in conventional quad-tree data structures, where tree nodes of depth D correspond to the stored image parcels of a derivative image of resolution KD.” *Ex. 1001*, 6:56-59, see also 6:22-26, 7:54-57.

179. Reddy further teaches that “a terrain tile file contains the actual elevation and image texture data for a given image, geographical area, and LOD” (*Ex. 1004*, ¶18), and that tiles are selected based on the proximity to the user’s view point, using the quad-tree structure to select tiles at the appropriate resolution based on their position in the LOD hierarchy. *Id.*, ¶¶12-17, 21, Fig. 1. Tiles are related to geographic area using a VRML geocentric coordinate system. *Id.*, ¶¶27-30. Therefore, a person of ordinary skill in the art would interpret the teachings of Reddy to mean that the TerraVision II geographic browser obtains the appropriate

imagery tiles utilizing the quad-tree structure by specifying the appropriate level of detail and the coordinates of the image. Since Reddy teaches progressively downloading and viewing data based on zoom level (related to the resolution level), and also user's region of interest or proximity (related to the user's X and Y location), a person of ordinary skill in the art would have understood that the images stored in Reddy's database were locatable based on zoom level and coordinate. Ex. 1004 at ¶¶3, 36.

180. Hornbacker further teaches that tiles may be located and requested using a specially formatted URL specifying the "scale and region," corresponding to the claimed "data set resolution index" (scale) and region (coordinates). Ex. 1003, 3:17-21, 5:16-24, 8:16-23, 8:30-10:6, 10:24-28. For example, URLs generated using the system taught by Hornbacker specify both a scale of the requested tile and the tile number, which itself incorporates the row and column (x,y coordinates) of the tile. Id., 8:30-9:19. Because Reddy teaches that imagery tiles are specified by coordinates and reached through links in the quad-tree structure, a person of ordinary skill in the art would recognize that that the URL syntax format of Hornbacker would provide an advantageous means to locate tiles by coordinates and level-of-detail as required by Reddy.

181. While I do not interpret this claim element to mean that all of the image tiles are stored in the *same* file, it is my opinion that it would be obvious to a

person of ordinary skill in the art to do so. A person of ordinary skill in the art would have been aware of various choices in which to store image data. For example, FlashPix by Kodak was a well-known format for storing images (Appendix II). A FlashPix file stored image data in a hierarchy of resolutions from the highest available for an image, down to the lowest defined format (§2.1). Thus, a FlashPix file either contained a single resolution image or an entire multi-resolution hierarchy (§2.2, FIG. 2.3). For convenient access, each resolution image was organized in tiles, which represented rectangular pixel arrays of the image. (§2.3. FIG. 2.4).

182. GeoTIFF is a file format (Appendix GG) for storing geographical image files using the well-known Tagged Image File Format (TIFF). The GeoTIFF spec defines a set of TIFF tags provided to describe all "Cartographic" information associated with TIFF imagery that originates from satellite imaging systems, scanned aerial photography, scanned maps, digital elevation models, or as a result of geographic analyses. Its aim is to allow means for tying a raster image to a known model space or map projection, and for describing those projections. The 1.0 release version of the GeoTIFF file format builds upon the then available revision 6.0 of TIFF.

183. TIFF (Appendix HH) defines a file format for storing image data. In particular, TIFF 6.0 provides for storage of multiple images in a single TIFF file,

each image corresponding to a subfile. These images could be, e. g., a full resolution image and its reduced resolution representations.

184. All these file formats were well-known industry standards and provided for multiple obvious variations by which a person of ordinary skill in the art building an online database (such as the geographic information database described in Reddy) could store image files. For example, when stored as a single file, the resulting storage may be compact due to reduced overhead, while when stored in multiple files, the resulting storage may be less efficient for storage but easier for caching performed in the user device side (e.g., as described in Reddy). In my opinion, a person of ordinary skill in the art facing the design need to store large amounts of multi-resolution imagery would recognize that there were a finite number of identified, predictable solutions (in this case, commonly used file formats that could be used to support an online imagery database), and this would lead a person of ordinary skill in the art to pursue the known options of various file types, including file types that can support storing multiple resolutions in the same file.

185. Based on this, it is my opinion that this claim element is disclosed by Reddy under the correct interpretation based on Reddy's teaching that terrain tile files may be located using the quad-tree structure. Alternatively, this claim element is obvious over Reddy in view of Hornbacker based on Hornbacker's teaching of

specially structured URLs to locate individual tiles based on zoom level and location within a larger image. Additionally, to the extent that this term is construed to require all update data parcels to be stored in one file, although it is my opinion that this construction is incorrect, this term is obvious over Reddy in view of the image file storage techniques known to a person of ordinary skill in the art.

6. Claim 8 is Rendered Obvious by Reddy in view of Hornbacker

186. **Claim 8.Preamble:** A display system for displaying a large-scale image retrieved over a limited bandwidth communications channel, said display system comprising: In my opinion, to the extent that the preamble is limiting, it is taught by Reddy and Hornbacker, which collectively teach systems for performing this function. For example, both Reddy and Hornbacker teach systems designed to operate in conjunction with a browser and display map imagery on a display. The relevant teachings that I previously discussed in regard to claim 1 apply here as well.

187. **Claim 8.A:** a display of defined screen resolution for displaying a defined image; Reddy teaches that the LOD techniques used to render a 3D image use selection criteria such as projected screen size. Ex. 1004, ¶¶13, 42, Fig. 5. Therefore, the software must operate on a display of defined screen resolution.

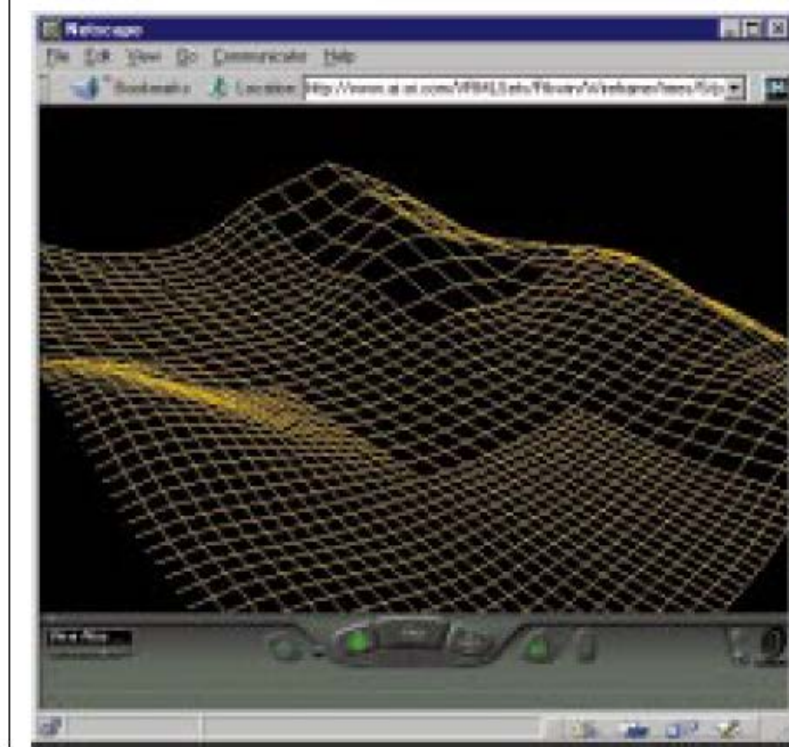
188. Hornbacker teaches that the system may be preferably optimized for viewing tiles at a particular resolution, e.g. 200 pixels-per-inch (7:4-25) and that the window may have a typical view size of 896 by 512 pixels. 14:2-6, see also 11:19-28, 13:4-10. Additionally, it would be obvious to a person of ordinary skill in the art that any digital display, including the devices taught by Reddy and Hornbacker, would necessarily have a finite resolution and that the rendering of images on that display would be limited by the resolution of the screen or the defined area of the screen.

189. **Claim 8.B:** a memory providing for the storage of a plurality of image parcels; Reddy teaches that “TerraVision maintains a tile cache,” which corresponds to the claimed “memory.” Ex. 1004, ¶45. Hornbacker likewise teaches that retrieved tiles may be stored in a local cache. Ex. 1003, 6:1-19, 8:1-6, 8:16-23, 10:13-28, 11:29-12:9, 12:17-23, 13:19-23. Reddy and Hornbacker both teach that a local cache on a client display device is advantageous to enable rapid display of pre-cached tiles and minimize delays associated with downloading. A person of ordinary skill in the art would recognize that a memory is required for these caching functions and both references teach similar solutions to similar problems.

190. **Claim 8.C:** displayable over respective portions of a mesh corresponding to said defined image; Reddy discloses the use of a mesh to model

the geometry of the terrain surface. Fig. 2 depicts a mesh of polygons that models the varying elevation for a geographic area. Ex. 1004, Fig.2, ¶ 17.

2 Using a tiled pyramid structure to represent terrain geometry. Closer terrain is represented in higher fidelity (more polygons) than distant terrain.



191. The caption notes that “closer terrain is represented in higher fidelity (more polygons) than distant terrain.” See also id. ¶¶ 12-14 (discussing the “polygonal representation” of a digital elevation model). This is apparent in Figure 2 as the bottom of the screen in the figure (which depicts points closer to the user) has a finer mesh than points in the middle or top of the screen (which depicts points furthest from the user). A POSITA would understand that the finer mesh represents higher fidelity than the coarser mesh.

192. Reddy further teaches that image data may be displayed over respective portions of a polygon mesh (e.g., a digital elevation model). Both the terrain geometry (i.e., mesh) and the corresponding imagery that gets texture mapped onto that geometry are stored together in terrain tiles. Ex. 1004, ¶¶17, 18, 24. For example, ¶24 notes that “terrain tile files” include the “elevation geometry and the texture map imagery for the specific terrain tile.” Reddy also refers repeatedly to images being “texture-mapped” on top of the terrain geometry, or that imagery such as satellite or aerial imagery can be viewed “on top of” the elevation geometry. *See, e.g.* Ex. 1004, ¶¶3 (“This model is **texture-mapped with satellite imagery** of 100 kilometers resolution—that is, each pixel in the texture map represents a region on the planet’s surface covering 100 km²... As she zooms into the region, higher resolution data, such as elevation and imagery, are progressively downloaded and displayed until she is “flying” over mountains with imagery down to one-meter resolution. Over certain parts of the terrain, alternative imageries are available, such as aerial photographs; the user can select any image to view **on top of the terrain geometry**”); 32 (“Letting users switch **the texture map for a single polygon** in VRML is a trivial exercise) (emphasis added).

193. Reddy’s quad-tree structure links the terrain tiles to their respective GeoTiles so that the correct terrain and imagery for a geographic area and resolution can be identified. *Id.*, Fig. 3, ¶¶19-24. In the 506 Patent, image data

parcels are similarly stored in “conventional quad-tree data structures.” Ex. 1002, 7:27-30. The result of this texture-mapping process (mapping imagery onto 3D polygons derived from the elevation model) is shown in the viewpoint perspective representation of Fort Irwin, CA in Fig. 5:

5 Screenshot of the TerraVision system.



194. Therefore, in my opinion, Reddy teaches a system based on well-known graphics principles which uses a polygonal mesh and maps the corresponding image data onto those polygons in order to create the desired “photorealistic” view of the terrain. Id., ¶¶2-3. Therefore, Reddy teaches this claim element.

195. **Claim 8.D:** a communications channel interface supporting the retrieval of a defined data parcel over a limited bandwidth communications channel; Reddy teaches a client system for dynamic visualization of image data

(through a terrain rendering program) which requests image data in the form of image tiles from a network server. Ex. 1002, Abstract. Hornbacker teaches a client system for dynamic visualization of image data (through an online image viewing program) which requests image data from a network server containing image tiles. Ex. 1003, Abstract. It would be obvious to a person of ordinary skill in the art that the client devices disclosed by both references would require a communications channel to receive tiles. Ex. 1009, ¶ 141. The “limited communication bandwidth communications channel” is discussed previously in regard to the preambles of claims 1 and 8.

196. **Claim 8.E:** a processor coupled between said display, memory and communications channel interface, Reddy teaches a navigation system operable on a computer device such as a laptop computer. Ex. 1004, ¶48. Hornbacker similarly teaches that the software is executed on a client workstation in connection with a web browser. Ex. 1003, 5:1-8, 16-21. It would be obvious to a person of ordinary skill in the art that in order to perform the functions claimed, including retrieving, caching, and displaying tiles, the device would require a processor coupled between the display, memory, and communications channel interface.

197. **Claim 8.F:** said processor operative to select said defined data parcel, The process of selecting a defined data parcel is described above in regard to claim element 1.B, and it would be obvious to a person of ordinary skill in the

art that the processor performs the selection of specific identified tiles since the selection in both Reddy and Hornbacker is performed by an algorithm executed by a processor.

198. **Claim 8.G:** retrieve said defined data parcel via said limited bandwidth communications channel interface for storage in said memory; Pertinent teachings regarding the retrieval of tiles over the limited bandwidth communications tile are discussed above in regard to claim element 1.A, while teachings relevant to the use of a local cache (“said memory”) are discussed above in regard to claim element 8.B. Reddy and Hornbacker both teach that a local cache on a client display device is advantageous to enable rapid display of pre-cached tiles and minimize delays associated with downloading. See, e.g. Ex. 1004, ¶ 45; Ex. 1003, 6:2-7, 6:17-20, 7:14-15, 8:1-6. A person of ordinary skill in the art would understand that both references teach that tiles (data parcels) are cached after being retrieved and recognize that both references teach similar, compatible solutions to similar problems.

199. Reddy and Hornbacker both teach that a local cache on a client display device is advantageous to enable rapid display of pre-cached tiles and minimize delays associated with downloading. A person of ordinary skill in the art would recognize that both references teach similar solutions (caching previously received tiles) to similar problems of enhancing display speed and avoid

downloading tiles that the client device already has. A person of ordinary skill in the art would also recognize that caching was well-recognized, common, and advantageous in the use of a typical Internet browser, as there is no guarantee using a TCP/IP protocol that data packets will arrive in the same order that they are requested because of the asynchronous nature of the protocol, and it would be necessary at least to temporarily store data packets until they are actually needed. The use of caches was widely known in several different contexts in common use on Internet browsers. Typical Internet browsers of the time would utilize caches in order to avoid consuming additional bandwidth in order to download the same information on a webpage that the user had already visited. Caches were (and still are) frequently used in, for example, a streaming audio or video context, in order to save data likely to be needed in the near future. Therefore, in my opinion, this element is both disclosed by and obvious over Reddy in view of Hornbacker.

200. **Claim 8.H:** render said defined data parcel over a discrete portion of said mesh to provide for a progressive resolution enhancement of said defined image on said display; Pertinent teachings regarding the quadtree data structure (mesh) taught by Reddy are discussed above in regard to claim element 8.C.

Reddy further teaches that the quadtree structure is used to implement a “progressive coarse-to-fine algorithm” to retrieve and display tiles, in which “when a user approaches a region of terrain, more detail is progressively loaded and

displayed in a coarse-to-fine fashion” so that “the user can still interact with the scene while higher resolution imagery and elevation loads” and “if some high-resolution tiles have yet to arrive over the network, TerraVision simply uses the highest resolution data it has so that the user can continue to interact with the terrain.” See, e.g. Ex. 1004 ¶¶ 21, 44. The reasons that the downloaded image tiles form a discrete portion of the image are discussed above in regard to claim element 1.J. A person of ordinary skill in the art would understand in view of these teachings that image tiles (data parcels) are rendered over a discrete portion of the geographic quadtree structure (mesh) to provide for a coarse-to-fine progressive resolution enhancement of the terrain image displayed on the client device.

201. Hornbacker teaches that tiles within a large database may be located by requests over a server which use URLs to identify the specific tile by incorporating identifying characteristics such as resolution, location, view angle, etc. See, e.g. Ex. 1003, 3:10-27, 5:16-25, 6:13-19, 8:30-9:28. Hornbacker teaches that tiles may be requested in a priority order in order to provide progressive resolution enhancement by providing lower resolution tiles first. See, e.g., *id.* at Abstract, 12:24-13:10. Hornbacker also teaches that view tiles that may be required by the next view request may be “pre-computed” based on the anticipated view. *Id.*, 7:26-8:6.

202. It is my opinion that these disclosures in Hornbacker collectively disclose or suggest rendering tiles to provide for progressive resolution enhancement of tiles within the visible display area.

203. Reddy and Hornbacker both teach or suggest methods of requesting image tiles over a network according to a priority order to provide progressive regional resolution enhancement. Reddy and Hornbacker both identify a common problem with the display of portions of very large images over a network, which is the latency and bandwidth consumption associated with downloading an entire image, and arrive at similar solutions in the form of requests for specific tiles in a priority order based on how soon the tiles need to be displayed. In my opinion, a person of ordinary skill in the art would recognize that the teachings of the two references solving similar problems in closely related fields could be considered in combination when designing a display system addressing a similar problem -- displaying images quickly in a visually pleasing manner.

204. Further, in my opinion, a person of ordinary skill in the art would recognize that the specific system for requesting tiles by URL, as disclosed in Hornbacker, could be advantageously utilized in the tile request process of Reddy, which includes identifying tiles likely to be needed in the near future based on their geographic coordinates, because both Reddy and Hornbacker concern web-based

systems, and URLs are the most common way for identifying and requesting resources in such web-based systems.

205. Additionally, both Reddy and Hornbacker teach that tiles are retrieved so that they can be rendered on the screen (i.e. viewed in the browser window), and that the tiles correspond to discrete portions of the image (e.g. map data).

206. **Claim 8.I:** wherein a remote computer coupled to the limited bandwidth communications channel delivers the defined data parcel; As discussed above in regard to claim element 1.I, the tiles (data parcels) are delivered to the client by a server (remote computer) coupled to a limited communication bandwidth communications channel (see discussions of claim 1 and 8 preambles regarding the channel).

207. **Claim 8.J:** wherein delivering the defined data parcel further comprises processing source image data to obtain a series K_{1-N} of derivative images of progressively lower image resolution; This claim element is substantially similar to claim 1.D of the 506 Patent. I previously discussed the teachings of Reddy and Hornbacker relevant to this element in regard to claim 1.D, and in my opinion, the same teachings apply to this element.

208. **Claim 8.K:** wherein series image K_0 being subdivided into a regular array; This claim element is substantially similar to claim 1.E of the 506 Patent. I previously discussed the teachings of Reddy and Hornbacker relevant to this

element in regard to claim 1.E, and in my opinion, the same teachings apply to this element.

209. **Claim 8.L:** wherein each resulting image parcel of the array has a predetermined pixel resolution; This claim element is substantially similar to claim 1.F of the 506 Patent. I previously discussed the teachings of Reddy and Hornbacker relevant to this element in regard to claim 1.F, and in my opinion, the same teachings apply to this element.

210. **Claim 8.M:** wherein image data has a color or bit per pixel depth representing a data parcel size of a predetermined number of bytes, This claim element is substantially similar to claim 1.G of the 506 Patent. I previously discussed the teachings of Reddy and Hornbacker relevant to this element in regard to claim 1.G, and in my opinion, the same teachings apply to this element.

211. **Claim 8.N:** resolution of the series K_{1-N} of derivative images being related to that of the source image data or predecessor image in the series by a factor of two, said array subdivision being related by a factor of two, such that each image parcel being of a fixed byte size; This claim element is substantially similar to claim 1.H of the 506 Patent. I previously discussed the teachings of Reddy and Hornbacker relevant to this element in regard to claim 1.H, and in my opinion, the same teachings apply to this element.

7. Claims 9-14 are Rendered Obvious by Reddy in view of Hornbacker

212. **Claim 9.** The display system of claim 8, wherein processing the source image data further comprises one of pre-processing the source image data on the remote computer and processing the source image data in real-time on-demand based on the request for the updated image parcel. This claim is substantially similar to claim 2 of the 506 Patent; therefore, in my opinion, the same teachings that I previously discussed in regard to that claim element apply.

213. **Claim 10:** The display system of claim 9, wherein receiving the update data parcel over a communications channel further comprises streaming the update data parcel over a communications channel to the limited communication bandwidth computer device. This claim is substantially similar to claim 3 of the 506 Patent; therefore, in my opinion, the same teachings that I previously discussed in regard to that claim element apply.

214. **Claim 11:** The display system of claim 8, wherein the limited communication bandwidth computer device further comprises one of a mobile computer system, a cellular computer system, an embedded computer system, a handheld computer system, a personal digital assistants and an internet-capable digital phone and a television. This claim is substantially similar to claim 4 of the

506 Patent; therefore, in my opinion, the same teachings that I previously discussed in regard to that claim element apply.

215. **Claim 12** The display system of claim 8, wherein a size of the data parcel on the remote computer is different from the update data parcel on the limited communication bandwidth computer device. This claim is substantially similar to claim 5 of the 506 Patent; therefore, in my opinion, the same teachings that I previously discussed in regard to that claim element apply.

216. **Claim 13.** The display system of claim 8, wherein processing the source image data further comprises queuing the update data parcels on the remote computer based on an importance of the update data parcel as determined by the remote computer. This claim is substantially similar to claim 6 of the 506 Patent; therefore, in my opinion, the same teachings that I previously discussed in regard to that claim element apply.

217. **Claim 14.A** The display system of claim 8, wherein the processing may further comprises compressing each data parcel; This claim element is substantially similar to claim element 7.A of the 506 Patent; therefore, in my opinion, the same teachings that I previously discussed in regard to that claim element apply.

218. **Claim 14.B** storing each data parcel on the remote computer in a file of defined configuration such that a data parcel can be located by specification of a

K_D, X, Y value that represents the data set resolution index D and corresponding image array coordinate. This claim element is substantially similar to claim element 7.B of the 506 Patent; therefore, in my opinion, the same teachings that I previously discussed in regard to that claim element apply.

8. Claim 15 is Rendered Obvious Over Reddy in view of Hornbacker

219. **Claim 15.Preamble** A remote computer for delivering large-scale images over network communications channels for display on a limited communication bandwidth computer device that has a display system for displaying a large-scale image retrieved over a limited bandwidth communications channel, To the extent that the preamble is limiting, it is my opinion that the preamble of claim 15 is taught by the same teachings of Reddy and Hornbacker that I previously discussed in regard to claim element 8.I, as to the remote server computers that deliver large-scale images divided into tiles, and the preamble of claim 8 as to the client computer devices with accompanying display systems.

220. **Claim 15.A:** a display of defined screen resolution for displaying a defined image, This claim element is substantially similar to claim element 8.A of the 506 Patent; therefore, in my opinion, the same teachings that I previously discussed in regard to that claim element apply.

221. **Claim 15.B:** a memory providing for the storage of a plurality of image parcels; This claim element is substantially similar to claim element 8.B of the 506 Patent; therefore, in my opinion, the same teachings that I previously discussed in regard to that claim element apply.

222. **Claim 15.C:** displayable over respective portions of a mesh corresponding to said defined image, This claim element is substantially similar to claim element 8.C of the 506 Patent; therefore, in my opinion, the same teachings that I previously discussed in regard to that claim element apply.

223. **Claim 15.D:** a communications channel interface supporting the retrieval of a defined data parcel over a limited bandwidth communications channel; This claim element is substantially similar to claim element 8.D of the 506 Patent; therefore, in my opinion, the same teachings that I previously discussed in regard to that claim element apply.

224. **Claim 15.E:** a processor coupled between said display, memory and communications channel interface, This claim element is substantially similar to claim element 8.E of the 506 Patent; therefore, in my opinion, the same teachings that I previously discussed in regard to that claim element apply.

225. **Claim 15.F:** said processor operative to select said defined data parcel, This claim element is substantially similar to claim element 8.F of the 506

Patent; therefore, in my opinion, the same teachings that I previously discussed in regard to that claim element apply.

226. **Claim 15.G:** retrieve said defined data parcel via said limited bandwidth communications channel interface for storage in said memory; This claim element is substantially similar to claim element 8.G of the 506 Patent; therefore, in my opinion, the same teachings that I previously discussed in regard to that claim element apply.

227. **Claim 15.H:** render said defined data parcel over a discrete portion of said mesh to provide for a progressive resolution enhancement of said defined image on said display, the remote computer comprises; This claim element is substantially similar to claim element 8.H of the 506 Patent; therefore, in my opinion, the same teachings that I previously discussed in regard to that claim element apply.

228. **Claim 15.I:** a parcel processing unit that processes a piece of source image data; As I previously discussed in regard to claims 2, 6, and 7, Reddy and Hornbacker teach systems that include servers which process source image data into an array of tiles, and perform operations such as sampling into tiles, reformatting, resampling, compression, and tile queueing. In my opinion, a person of ordinary skill in the art would recognize that these functions need to be performed by a processor on the server computer, and that the processor would

accordingly include some hardware or software component for executing these functions. Hornbacker further teaches a specific example of processing unit that processes source image data, i.e. the foreground view composer. Ex. 1003 at Fig. 1, 5:3-8, 6:1-19, 10:3-23,

229. In my opinion, the specification of the 506 Patent contains little if any support for this element, as the specification does not disclose any details concerning either the hardware or software architecture of the server. There is some cursory discussion of the function of the server in the 506 Patent, for example, at Ex. 1002 at Fig. 2, 6:60-63, and 6:7-61. However, none of this discussion identifies any discrete component of the server that comprises “a parcel processing unit that processes a piece of source image data.” Reddy additionally identifies a software structure, *i.e.* a Common Gateway Interface script, for performing processing in the “on-the-fly” scenario. Ex. 1004, ¶52. Therefore, in my opinion, because Reddy and Hornbacker describe this element in as much or more detail as the specification of the 506 Patent, this limitation is obvious to a person of ordinary skill in the art over Reddy, Hornbacker, or both.

230. **Claim 15.J:** delivers the defined data parcel to the limited communication bandwidth computer device; This claim element is substantially similar to claim element 8.I of the 506 Patent; therefore, in my opinion, the same teachings that I previously discussed in regard to that claim element apply.

231. **Claim 15.K:** wherein the parcel processing unit further comprises a parcel processing control that processes source image data to obtain a series K_{1-N} of derivative images of progressively lower image resolution; This claim element is substantially similar to claim element 1.D of the 506 Patent; therefore, in my opinion, the same teachings that I previously discussed in regard to that claim element apply. As I previously discussed in regard to claim element 15.I, the specification of the 506 Patent contains no description of a “parcel processing unit,” much less how the parcel processing unit “further comprises a parcel processing control.” However, for the same reasons that I discussed above regarding claim element 15.I, Reddy and Hornbacker each teach a server having a processor that performs the claimed function.

232. **Claim 15.L:** wherein series image K_0 being subdivided into a regular array; This claim element is substantially similar to claim element 1.E of the 506 Patent; therefore, in my opinion, the same teachings that I previously discussed in regard to that claim element apply.

233. **Claim 15.M:** wherein each resulting image parcel of the array has a predetermined pixel resolution; This claim element is substantially similar to claim element 1.F of the 506 Patent; therefore, in my opinion, the same teachings that I previously discussed in regard to that claim element apply.

234. **Claim 15.N:** wherein image data has a color or bit per pixel depth representing a data parcel size of a predetermined number of bytes. This claim element is substantially similar to claim element 1.G of the 506 Patent; therefore, in my opinion, the same teachings that I previously discussed in regard to that claim element apply.

235. **Claim 15.O:** resolution of the series K_{1-N} of derivative images being related to that of the source image data or predecessor image in the series by a factor of two, said array subdivision being related by a factor of two, such that each image parcel being of a fixed byte size; This claim element is substantially similar to claim element 1.H of the 506 Patent; therefore, in my opinion, the same teachings that I previously discussed in regard to that claim element apply.

9. Claims 16-21 are rendered obvious over Reddy in view of Hornbacker

236. **Claim 16:** The remote computer of claim 15, wherein processing the source image data further comprises one of pre-processing the source image data on the remote computer and processing the source image data in real-time on-demand based on the request for the updated image parcel. This claim is substantially similar to claim 2 of the 506 Patent; therefore, in my opinion, the same teachings that I previously discussed in regard to that claim apply.

237. **Claim 17:** The remote computer of claim 16, wherein receiving the update data parcel over a communications channel further comprises streaming the update data parcel over a communications channel to the limited communication bandwidth computer device. This claim is substantially similar to claim 3 of the 506 Patent; therefore, in my opinion, the same teachings that I previously discussed in regard to that claim apply.

238. **Claim 18:** The remote computer of claim 15, wherein the limited communication bandwidth computer device further comprises one of a mobile computer system, a cellular computer system, an embedded computer system, a handheld computer system, a personal digital assistants and an internet-capable digital phone and a television. This claim is substantially similar to claim 4 of the 506 Patent; therefore, in my opinion, the same teachings that I previously discussed in regard to that claim apply.

239. **Claim 19:** The remote computer of claim 15, wherein a size of the data parcel on the remote computer is different from the update data parcel on the limited communication bandwidth computer device. This claim is substantially similar to claim 5 of the 506 Patent; therefore, in my opinion, the same teachings that I previously discussed in regard to that claim apply.

240. **Claim 20:** The remote computer of claim 15, wherein processing the source image data further comprises queuing the update data parcels on the remote

computer based on an importance of the update data parcel as determined by the remote computer. This claim is substantially similar to claim 6 of the 506 Patent; therefore, in my opinion, the same teachings that I previously discussed in regard to that claim apply.

241. **Claim 21:** The remote computer of claim 15, wherein processing further comprises compressing each data parcel and storing each data parcel on the remote computer in a file of defined configuration such that a data parcel can be located by specification of a K_D , X, Y value that represents the data set resolution index D and corresponding image array coordinate. This claim is substantially similar to claim 7 of the 506 Patent; therefore, in my opinion, the same teachings that I previously discussed in regard to that claim apply.

242. This is the END of my Declaration.

**Curriculum Vitae for
William R. Michalson**

Research Associates, LLC
26 West Main Street, STE 2
Dudley, MA 01571
Email: wrm@wmichalson.com
Tel: (508) 461-6242
Cell: (508) 331-4134

1. Personal:

1.1 Education

Ph.D. in Electrical Engineering, 1989, Worcester Polytechnic Institute, Worcester, Massachusetts.

Dissertation: *A Parallel Computer Architecture for Real-Time Decision Making*. The dissertation develops a hierarchical, multiple processor, computer architecture for executing artificial intelligence programs in real-time. Dissertation Directors: Dr. Peter E. Green and Dr. R. James Duckworth.

Minor Areas: Minor sequences completed in Mathematics and Physics.

Specialties: Area examinations passed in the fields of Computer Architecture, Probabilistic Systems Analysis, and State Space Analysis.

M.S. in Electrical Engineering, 1985, Worcester Polytechnic Institute, Worcester, Massachusetts.

Specialties: The courses taken stressed Computer Architecture, Communications Systems, and Solid-State Physics.

B.S. in Electrical Engineering, 1981, Syracuse University, Syracuse, New York.

1.2 Work experiences - Academic.

1991-Present Worcester Polytechnic Institute

Professor of Electrical and Computer Engineering; Professor of Computer Science.

Effective July 1, 2005 Promoted to the rank of Full Professor (Professor of Electrical and Professor of Computer Science)

November 17, 2004 Appointed dual professorship, adding the title of Associate Professor of Computer Science.

July 1, 1998 Granted tenure and promoted to the rank of Associate Professor.

August 1, 1992 Assistant Professor of Electrical Engineering (tenure-track).

August 1, 1991 Visiting Assistant Professor of Electrical Engineering.

January 1, 1990 Adjunct Assistant Professor of Electrical Engineering.

1.3 Work experiences other than teaching (chronological).

2012-2014 Grid Roots, LLC

Grid Roots, LLC is a company which was formed in 2012 for the purpose of commercializing a navigation and tracking device for use by children and the elderly to allow caregivers to non-intrusively monitor their activities. The system under development integrates GPS, inertial and beacon-based navigation technologies to develop a system for users to track deployed devices. My responsibilities within Grid Roots, LLC relate to hardware and software engineering, as well as the development of IP related to tracking individuals.

1995-Present Research Associates, LLC

Research Associates, LLC is a company I formed in which I perform engineering and consulting in the areas of computer systems, communications and navigation. All of my litigation-related and other consulting activities are performed through Research Associates, LLC.

1988-1991 Raytheon Company

Subsequent to receiving my Ph. D., I returned to the Equipment Division of the Raytheon Company. Shortly after I returned, I was promoted to a title of Engineer, Design and Development which was the highest title I could hold based on my level of education and years of experience. Within a year, I was selected to sit on the engineering staff of a newly formed System Engineering Department of the Division's Computer and Displays Laboratory. In this department I acted primarily as a consultant to other departments within the laboratory. My responsibilities ranged from leading the hardware/software development of supercomputer-class computer systems to performing applied research into the exploitation of new technology. My role was similar to that of a Principle Investigator in an academic setting as I was responsible for securing funding and personnel, leading research efforts, interacting with the research sponsor, and reporting results. At the time of my departure I was involved with the following projects:

Fault-Tolerant Multiprocessor

The development of a highly fault tolerant, highly reliable, real-time computer system intended for long-duration spaceborne applications. This system is designed to produce in excess of one gigaoperation per second of raw processing power.

Optimal Task Allocation

A program of applied research into the use of Genetic Algorithms for deriving optimal mappings of software tasks to the hardware processing elements in distributed systems.

Performance Modeling and Scaling

This project focused on the development of simulation models for characterizing the performance of a large scale multiple processor system. These models formed a basis for predicting system performance for several different hardware configurations to ensure compliance with system specifications.

High Clutter Signal Detection

A program of applied research into the use of Neural Networks to detect the presence of targets in extremely high clutter environments.

Power Efficient Computing

A program of applied research into an Integrated Optical computer structure that is designed to maximize the number of computations that can be performed per unit of power.

1985-1988 Raytheon Company (Leave of Absence)

In 1985 I became one of two people in the Equipment Division to receive Aldo Miccioli Fellowships. This Fellowship was awarded to allow me to pursue full-time study towards the Ph.D. degree. I returned to Raytheon during the summer of 1986, but otherwise remained on leave of absence to dedicate my time to my studies.

1982-1985 Raytheon Company

Engineer in the VLSI Design Department of the Computer and Displays Laboratory within Raytheon's Equipment Division. I was lead engineer for the design of several semi-custom VLSI circuits for both signal and data processing applications.

1981-1982 Raytheon Company

Engineer in the Cursive Displays Department of the Computer and Displays Laboratory. I designed and debugged circuit assemblies which were used in vector displays for air traffic control applications.

1.4 Consulting experiences.

1.4.1 Law-Related

Locata LBS LLC v. YellowPages.com LLC,

Retained by Baker Botts on behalf of defendant YellowPages.com. Case before the Central District of California (2:13-cv-07664). See also IPR2015-00151. Retained 9/14 to present.

M/A-COM Technology Solutions Holdings, Inc. v. Laird Technologies, Inc.

Retained by Erise IP on behalf of Laird Technologies, Inc., for invalidity consulting regarding U.S. Patent No. 6,272,349. Retained 6/14 to present.

Certusview Technologies, LLC v. S&N Locating Services, LLC and S&N Communications, Inc.,

Retained by Baker & McKenzie on behalf of defendant S&N. Patents-in-suit are U.S. Patents 8,265,344, 8,290,204, 8,340,359, 8,407,001, and 8,532,341. Case before the Eastern District of Virginia, (2:13-cv-346). Deposed 11/8/14; Retained 6/14 to present.

adidas AG and adidas America, Inc. v. Under Armour, Inc. and MapMyFitness, Inc.

Retained by Kilpatrick Townsend on behalf of plaintiff adidas. Case before the District of Delaware, (1:14-cv-00130). Retained 5/14 to present.

GeoTag, Inc., v. AT&T Mobility LLC and AT&T Services, Inc.,

Retained by Baker Botts as an expert on behalf of defendant AT&T. Patent-in-suit is U.S. Patent 5,930,474. Case before the Northern District of Texas, Dallas Division, (3:13-cv-00169). Deposed 5/29/14; Retained 1/14 to 9/14. Matter settled.

Nokia Corp v. HTC Corp.

Retained by Quinn Emanuel as an expert on behalf of defendant HTC. Case being litigated in Germany. Patent number EP0766811B1. Retained 12/13 to 2/14. Matter settled.

Porto Technology, Co., Ltd. et al. v. Cellco Partnership d/b/a Verizon Wireless

Retained by Wiley-Rein as an expert on behalf of defendant Verizon. Case before the United States District Court for the Eastern District of Virginia (Case No. 3:13-cv-00265). Retained 10/13 to 2/14. Matter dismissed.

Nokia Corp v. HTC Corp.

Retained by McDermott Will and Emery and White & Case as an expert on behalf of defendant HTC. Case before the United States District Court for the District of Delaware (Case No. Case No. 1:12-cv-00550-UNA and Case No. 1:12-cv-551-UNA). Retained 6/13 to 2/14. Matter settled.

NXP B.V. v. Research In Motion, Ltd., et al.

Retained by Fish and Richardson as an expert on behalf of defendant Research In Motion. Patent-in-suit is U.S. Patent 6,501,420. Case before the United States District Court for the

Middle District of Florida (Case No. 6:12-cv-00498). Deposed 9/18/13; Testified in Court: 4/1/14 and 4/2/14; Retained 5/13 to 4/14.

Vehicle IP LLC v. Wal-Mart Stores Inc., et al.

Retained by Polsinelli-Shugart, as an expert on behalf of defendant Werner Enterprises. Patent-in-suit is U.S. Patent 5,694,322. Case before the United States District Court for the District of Delaware (Case No. 1:10-cv-00503). Deposed 7/15/13 and 9/20/13; Testified in Court: 9/27/13 and 9/30/13; Retained 4/13 to 9/13.

TracBeam, LLC v. Google Inc.

Retained by Quinn-Emanuel as an expert on behalf of defendant Google. Patents-in-suit are U.S. Patents 7,764,231 and 7,525,484. Case before the United States District Court for the Eastern District of Texas (6:11-cv-00093). Deposed 2/5/14; Retained 3/13 to 6/14.

Microsoft Corporation and Google Inc., v. GeoTag, Inc.

Retained by Perkins-Coie as an expert on behalf of plaintiff Microsoft Corporation. Patent-in-suit is U.S. Patents 5,930,474. Case before the United States District Court for the District of Delaware (1:11-cv-00175). Retained 1/13 to present.

GeoTag, Inc., v. Frontier Communications Corp., et al.,

Retained by multiple firms as an expert on behalf of defendants. Patent-in-suit is U.S. Patents 5,930,474. Case before the Eastern District of Texas, Marshall Division, (2:10-cv-00265; other defendants are listed in case numbers 2:10-cv-00265, 2:10-cv-00272, 2:10-cv-00437, 2:10-cv-00569, 2:10-cv-00570, 2:10-cv-00571, 2:10-cv-00572, 2:10-cv-00573, 2:10-cv-00574, 2:10-cv-00575, 2:10-cv-00587, 2:11-cv-00175, 2:11-cv-00404, 2:11-cv-00421, 2:11-cv-00424, 2:11-cv-00425, 2:11-cv-00570, 2:12-cv-00043, 2:12-cv-00051, 2:12-cv-00436, 2:12-cv-00438, 2:12-cv-00439, 2:12-cv-00441, 2:12-cv-00442, 2:12-cv-00444, 2:12-cv-00445, 2:12-cv-00446, 2:12-cv-00447, 2:12-cv-00448, 2:12-cv-00449, 2:12-cv-00450, 2:12-cv-00452, 2:12-cv-00454, 2:12-cv-00456, 2:12-cv-00459, 2:12-cv-00460, 2:12-cv-00462, 2:12-cv-00464, 2:12-cv-00466, 2:12-cv-00468, 2:12-cv-00469, 2:12-cv-00470, 2:12-cv-00471, 2:12-cv-00473, 2:12-cv-00474, 2:12-cv-00475, 2:12-cv-00476, 2:12-cv-00476, 2:12-cv-00477, 2:12-cv-00480, 2:12-cv-00481, 2:12-cv-00482, 2:12-cv-00482, 2:12-cv-00483, 2:12-cv-00486, 2:12-cv-00487, 2:12-cv-00520, 2:12-cv-00521, 2:12-cv-00523, 2:12-cv-00524, 2:12-cv-00525, 2:12-cv-00527, 2:12-cv-00528, 2:12-cv-00530, 2:12-cv-00532, 2:12-cv-00534, 2:12-cv-00535, 2:12-cv-00536, 2:12-cv-00537, 2:12-cv-00542, 2:12-cv-00543, 2:12-cv-00545, 2:12-cv-00547, 2:12-cv-00548, 2:12-cv-00549, 2:12-cv-00550, 2:12-cv-00551, 2:12-cv-00552, 2:12-cv-00555, 2:12-cv-00556, 2:12-cv-00570, 2:12-cv-00572, 2:12-cv-00573, 2:12-cv-00575, 2:12-cv-00587, 3:13-cv-00217). Deposed 5/29/14; Retained 1/13 to 9/14.

MOSAID Technologies Inc., v. Realtek Semiconductor Corporation

Retained by Sidley Austin, LLP as an expert on behalf of defendant Realtek Semiconductor Corporation. Patents-in-suit are U.S. Patents 5,131,006; 5,151,920; 5,422,887; 5,706,428; 6,563,786; and 6,992,972. Case before the United States District Court for the Eastern District of Texas (Tyler Division) (Case No. 2:11-cv-00179). Retained 12/12 to 12/12. Matter settled.

Hoyt A. Flemming v. Cobra Electronics Corporation

Retained by Sidley Austin, LLP as an expert on behalf of defendant Cobra Electronics Corporation. Patents-in-suit are U.S. Patents RE39038, RE40653 and RE41905. Case before the United States District Court for the District of Idaho (Case No. 1:12-cv-00392). Retained 11/12 to 06/13. Matter settled.

LBS Innovations LLC v. Aaron Bros., Inc., et al.

Retained as an expert on behalf of defendants Whole Foods Marketplace, Comerica, Hotels.com, Academy, Ltd., and Homestyle Dining. Patent-in-suit is U.S. Patent 6,091,956. Case before the Eastern District of Texas, Marshall Division, (Case No. 2:11-cv-00142-MHS-CMC. Deposited 10/5/12; Retained 7/12 to 12/12. Plaintiff moved to dismiss.

Advanced Media Networks, L.L.C. v. Gogo LLC et al.

Retained by Sidley Austin, LLP as an expert on behalf of defendant Gogo. Patent-in-suit is U.S. Patent 5,960,074. Case before the United States District Court for the Central District of California (Case No. 11-cv-10474). Deposited 2/6/13. Retained 7/12 to 8/13.

Walker Digital, LLC v. Google Inc.

Retained by O'Melveny & Meyers, LLP as an expert on behalf of defendant Google Inc. Patents-in-suit are U.S. Patents 6,199,014. Case before the United States District Court for the District of Delaware (Case No. 1:11-cv-00309-SLR). Deposited 2/27/13 - 2/28/13. Retained 6/12 to present (case stayed as of 8/13).

Silver State Intellectual Technologies, Inc. v. Garmin International, Inc., et al.

Retained by Erise IP, P.A. as an expert on behalf of defendants Garmin International, Inc. and Garmin USA, Inc. Patents-in-suit are U.S. Patents 6,525,768; 6,529,824; 6,542,812; 7,343,165; 7,522,992; 7,593,812; 7,650,234; 7,702,455 and 7,739,039. Case before the United States District Court for the District of Nevada (Case No. 2:11-cv-1578). Deposited 2/19/14; Retained 4/12 to present.

Beacon Navigation GmbH v. Toyota Motor Corporation, et al.

Retained by Kirkland & Ellis, LLP on behalf of defendants Toyota Motor Corporation; Toyota Motor North America, Inc.; Toyota Motor Sales, U.S.A. Inc.; Toyota Motor Engineering & Manufacturing North America, Inc.; Toyota Motor Manufacturing, Indiana, Inc.; Toyota Motor Manufacturing, Kentucky, Inc.; Toyota Motor Manufacturing Mississippi, Inc.; Mazda Motor Corporation; Mazda Motor of America, Inc.; Fuji Heavy Industries, Ltd.; Fuji Heavy Industries U.S.A. Inc.; Subaru of America, Inc.; Jaguar Land Rover North America, LLC; Jaguar Cars Limited; Land Rover; Volvo Car Corporation; and Volvo Cars of North America, LLC; Adduci Mastriani & Schaumberg, LLP on behalf of defendants Suzuki and Garmin; Crowell-Moring on behalf of General Motors; Dickstein Shapiro on behalf of Chrysler Group, LLC; Finnegan, Henderson, Farabow, Garrett & Dunner on behalf of Bayerische Motoren Werke AG, BMW of North America, LLC, and BMW Manufacturing Co. LLC; Fish & Richardson on behalf of Honda Motor Co., Ltd., Honda North America, Inc., American Honda Motor Co., Inc., Honda Manufacturing of Alabama, LLC, Honda Manufacturing of Indiana, LLC, and Honda of America, Mfg., Inc.; Frommer Lawrence and Haug, LLP on behalf of Dr. Ing. h.c.F. Porsche AG and Porsche Cars North America, Inc.; Hogan Lovells on behalf of Daimler AG, Mercedes-Benz USA, LLC, or Mercedes-Benz U.S. International, Inc.; Quinn-Emanuel on behalf of Nissan and

Ford. Case before the US International Trade Commission, Washington D.C., in the matter of: “Certain Automotive Navigation Systems, Components Thereof, and Products Containing Same, Inv. No. 337-TA-814. Case withdrawn by Plaintiff. Retained 1/12 – 4/12.

Beacon Navigation GmbH v. Toyota Motor Corporation, et al.

Retained by Kirkland & Ellis, LLP on behalf of defendants Toyota Motor Corporation; Toyota Motor North America, Inc.; Toyota Motor Sales, U.S.A. Inc.; Toyota Motor Engineering & Manufacturing North America, Inc.; Toyota Motor Manufacturing, Indiana, Inc.; Toyota Motor Manufacturing, Kentucky, Inc.; Toyota Motor Manufacturing Mississippi, Inc.; Mazda Motor Corporation; Mazda Motor of America, Inc.; Fuji Heavy Industries, Ltd.; Fuji Heavy Industries U.S.A. Inc.; Subaru of America, Inc.; Jaguar Land Rover North America, LLC; Jaguar Cars Limited; Land Rover; Volvo Car Corporation; and Volvo Cars of North America, LLC. Multiple cases before the United States District Court for the District of Delaware. Case numbers 1:11-cv-00942-UNA, 1:11-cv-00941-UNA, 1:11-cv-00951-UNA, 1:11-cv-00952-UNA, 1:11-cv-00936-UNA, 1:11-cv-00937-UNA, 1:11-cv-00955-UNA, 1:11-cv-00959-UNA, and 1:11-cv-00960-UNA. Currently stayed. Retained 1/12 to Present.

Beacon Wireless Solutions, Inc., et al., v. Garmin International, Inc., et al.

Retained by Shook, Hardy and Bacon, LLP as an expert on behalf of defendant Garmin. Matter involves alleged trade secret misappropriation. Case before the United States District Court for the Western District of Virginia, Harrisonburg Division (Case No. 5:11-cv-00025). Testified in Court: 5/25/12. Retained 12/11 to 5/25/12.

Tramontane IP, LLC v. Garmin Int’l, Inc., et al.

Retained by Shook, Hardy and Bacon, LLP as an expert on behalf of defendant Garmin. Patents-in-suit are U.S. Patents 6,526,268 and 7,133,775. Case before the United States District Court for the Eastern District of Virginia (Case No. 1:2011-cv-00918). Case Settled. Retained 11/11 to 12/11.

Sourceprose, Inc. v. AT&T, Inc., MetroPCS Communications, Inc., et al.

Retained by Kilpatrick Townsend as an expert on behalf of defendant AT&T. Patents-in-suit are US Patent Nos. 7,142,217 and 7,161,604. Case before the United States District Court for the Western District of Texas, Austin Division. Case number 1:11-cv-00117. Retained 11/11 to present.

Furuno Electric Co., Ltd. and Furuno U.S.A., Inc. v. Honeywell International, Inc.

Retained by Quinn-Emanuel as an expert on behalf of complainant Furuno. Case before the US International Trade Commission, Washington D.C., in the matter of: “Certain GPS Navigation Products, Components Thereof, and Related Software,” Investigation number 337-TA-810. Patents-in-suit are U.S. Patent Nos. 6,084,565; 7,095,367; 7,089,094; and 7,161,561. Case settled. Retained 8/11 – 12/11.

Honeywell International, Inc. v. Furuno Electric Co., Ltd. and Furuno U.S.A., Inc.

Retained by Quinn-Emanuel as an expert on behalf of respondent Furuno. Case before the US International Trade Commission, Washington D.C., in the matter of: “Certain GPS Navigation

Products, Components Thereof, and Related Software,” Investigation number 337-TA-783. Patents-in-suit are U.S. Patent Nos. 7,209,070; 6,865,452; 5,461,388; and 6,088,653. Case Settled. Retained 8/11 – 12/11.

Triangle Software, Inc. v. Garmin International, Inc.

Retained by Weil, Gotshal & Manges, LLP., as an expert on behalf of defendant Garmin. Patents-in-suit are US Patents 7,557,730, 7,221,287, 7,375,649, 7,508,321 and 7,702,452. Case before the United States District Court, Eastern District of Virginia, Case No. 1:10-cv-1457 CMH/TCB. Deposed 7/28/11; Testified in Court: 11/3/11 (jury trial). Retained 4/11 to 11/11.

Garmin International, Inc. v. Pioneer Corporation and Pioneer Electronics (USA), Inc.

Retained by Shook, Hardy and Bacon, LLP as an expert on behalf of plaintiff Garmin. Patents-in-suit are U.S. Patents 5,365,448; 5,424,951; and 6,122,592. Case before the United States District Court for the District of Kansas. Case No. 10-CV-2080 JWL/GLR. Declarative Judgment action stayed. Retained 3/11 to 11/11.

Visteon Global Technologies, Inc. And Visteon Technologies, LLC v. Garmin International, Inc.

Retained by Shook, Hardy and Bacon as an expert on behalf of defendant Garmin. Patents-in-suit are US Patents 5,544,060, 5,654,892, 5,832,408, 5,987,375 and 6,097,316. Case before the United States District Court, Eastern District of Michigan, Case No. 2:10-cv-10578--PDB-MAR. Deposed 10/9/12; Retained 12/10 to present.

Thomson Licensing SAS and Thomson Licensing, LLC. v. Realtek Semiconductor Corporation

Retained by Sidley-Austin as an expert on behalf of respondent Realtek Semiconductor. Case before the US International Trade Commission, Washington D.C., in the matter of: “Certain Liquid Crystal Display Devices, Including Monitors, Televisions, Modules, And Components Thereof,” Investigation number 337-TA-741. Patent-in-suit is US Patent 6,121,941. Deposed 6/29/11; Testified in Court: 9/15/11 and 9/16/11. Retained 11/10 – 9/11.

Ambato Media, LLC. v. Clarion Co., LTD., et al.

Retained by Traurig-Greenberg as an expert on behalf of defendant Garmin. Patent-in-suit is US Patent 5,432,542. Case before the United States District Court for the Eastern District Of Texas, Marshall Division. Case number 2:09-CV-242. Deposed 4/26/12 and 5/10/12; Testified in Court: 7/11/12. Retained 10/10 to present.

Gabriel Technologies Corporation and Trace Technologies, LLC, v. Qualcomm Incorporated, Snaptrack, Inc. and Norman Krasner

Retained by Cooley-Godward as an expert on behalf of defendants Qualcomm, Snaptrack, and Krasner. Trade secret misappropriation case related to US Patents 6,377,209, 6,583,757, 6,661,372, 6,799,050, 6,861,980, 6,895,249, 7,254,402, 7,289,786, 7,319,876, 7,421,277, 7,446,655, 7,570,958, 7,574,195, and 7,660,588. Case before the Southern District of California San Diego Division, Case No. 08-cv-1992 MMA POR. Retained 6/10 to 7/12.

SiRF/CSR v. Global Locate/Broadcom Corporation

Retained by Wilmer-Hale as an expert on behalf of defendant Global Locate / Broadcom. Patents-in-suit are US Patents 5,663,735, 6,480,150, 6,519,466, 6,650,879, 6,882,827, 6,934,322, 7,412,157, 7,236,883, and 7,573,422. Case before the Central District of California, Case No. 8:06-cv-01216 and Case No. 8:10-cv-01281. Retained 9/10 to 1/11.

Pioneer Electronics v. Garmin Corporation

Retained by Shook, Hardy and Bacon, LLP as an expert on behalf of respondent Garmin In the matter of Certain Multimedia Display and Navigation Devices and Systems, Components Thereof, and Products Containing Same; Inv. No. 337-TA-694. Patents-in-suit are U.S. Patents 5,365,448; 5,424,951; and 6,122,592. Case before the U.S. International Trade Commission. Deposed on 7/29/10; Testified at technology tutorial (8/27/20) and in the evidentiary hearing (9/20/10). Retained 4/10 to 9/10.

EMSAT Advanced Geo-Location Technology, LLC and Location Based Services LLC, v. AT&T Mobility, LLC

Retained by Baker-Botts as an expert on behalf of defendant AT&T Mobility, LLC. Patents-in-suit are U.S. Patents 7,289,763; 5,946,611; 6,324,404; and 6,847,822. Case before the U.S. District Court for the Northern District of Ohio, Eastern Division, Civil Action No. 4:08 CV 822. Deposed 5/4/10; Testified in Court: 5/10/10 (Markman hearing). Retained 12/09 to 3/11.

Tendler Cellular of Texas, LLC v. AT&T Mobility, LLC, et al.

Retained by Baker-Botts as an expert on behalf of defendants AT&T Mobility, LLC, et al. Patents-in-suit are U.S. Patents 7,447,508; 7,305,243; 7,050,818; and 6,519,463. Case before the U.S. District Court for the Eastern District of Texas (Tyler), Civil Action No. 6:09-CV-00115. Retained 8/09 to 7/10.

Ambit Corporation v. Delta Air Lines, Inc., and Aircell LLC.

Retained by Sidley Austin, LLP., as an expert on behalf of defendants Delta Airlines, Inc., and Aircell, LLC. Patent-in-suit is US patent 7,400,858. Case before the US District Court, District of Massachusetts, Boston, Civil Action No. 1:09-CV-10217-WGY. Deposed 12/4/09; Testified in Court: 12/7/09 (evidentiary hearing), 7/10 (jury trial). Retained 8/09 to 7/10.

GPS Industries, Inc. and Optimal I.P. Holdings, L.P. v. Altex Corporation, et. al.

Retained by Hitchcock-Evert as an expert on behalf of defendants Altex Corporation, Deca International Corporation, Golflogix, Inc. and L1 Technologies. Patent-in-suit is US patent 5,364,093. Case before the US District Court, Northern District of Texas, Dallas Division, Civil Action No. 3-07-CV0831-K. Deposed 6/30/09. Retained 5/08 through 7/09.

Satellite Tracking of People, LLC v. Omnilink Systems, Inc.

Retained by DLA Piper as an expert on behalf of defendant Omnilink Systems, Inc.. Patent-in-suit is US patent RE39,909. Case before the US District Court, Eastern District of Texas, Marshall Division, Civil Action No. 2-08CV-116. 12/08 – 1/11.

SiRF Technology, Inc. v. Global Locate, Inc.

Retained by DLA Piper/WilmerHale as an expert on behalf of Global Locate. Patents-in-suit include US patents 6,304,216; 6,417,801; 6,606,346; 6,651,000; 6,704,651; 6,937,187;

7,043,363; 7,091,904; 7,132,980 and 7,158,080. Case before the US International Trade Commission, Washington D.C., in the matters of: “Certain GPS Devices and Products Containing Same,” Investigation number 337-TA-602 (Global Locate, plaintiff) and “Certain GPS Chips, Associated Software and Systems, and Products Containing Same,” Investigation number 337-TA-596 (SiRF Technologies, plaintiff). My work focused on the 7,043,363 and 7,091,904 patents in defense of Global Locate/Broadcom from June 2007 through March 2008. Deposed 1/18-1/19/08; testified at trial 3/18-3/19/08.

Intellectual Science and Technology

Retained Dykema Gossett, PLLC as a technical expert on patent infringement issues related to “suspend-to-RAM” technologies in personal computers. Pre-litigation work.

Intellectual Science and Technology, Inc., v. Sony, JVC and Panasonic

Retained Dykema Gossett, PLLC as a technical expert on patent infringement issues related to US Patent 5,748,575, US Patent 6,222,799, US Patent 6,785,198, US Patent 6,662,239 and US Patent 6,717,890. Sony Electronics Inc., case number 2:06-CV-10406, JVC Americas Corp., case number 2:06-CV-10409 and Panasonic Corporation of North America case number 2:06-CV-10412. Cases heard in United States District Court, Eastern District of Michigan, Southern Division. Expert for Intellectual Science and Technology, Inc. Dec 2006 – 2008..

Kirsch Technologies v. Xerox, Canon

Retained Dykema Gossett, PLLC as an Expert Witness on patent infringement issues related to US Patent 4,816,911, Canon case number CA 00-72775, Xerox case number CA 00-72778, cases heard in United States District Court, Eastern District of Michigan, Southern Division. Expert for Kirsch Technologies. Nov 2006 – 2008..

American Video Graphics v. ATI Technologies

Retained by Sidley, Austin, Brown & Wood, Dallas, TX as a technical expert on patent infringement issues related to US Patents 5,132,670, 5,109,520, 5,084,830, 4,761,642, 4,742,474, 4,734,690, 4,730,185 and 4,694,286. Hewlett-Packard Co., et al., defendants, case number CA 6:04-CV-379-LED and Sony Corporation of America et al., defendants, case number CA 6:04-CV-399-LED. Cases heard in United States District Court, Eastern District of Texas, Tyler Division. Expert for ATI Technologies, intervener. Jan 2005 – Sep 2005.

Microsoft v. EMC

Dewey Ballantine, LLP, Washington, D.C., as a technical expert on patent infringement issues related to US Patents 5,588,147; 5,689,700; 6,393,466; 6,424,151; 6,490,594; and 6,632,248. Wrote a declaration on behalf of Microsoft. Oct 2004 – Jan 2005.

Optimum Return v. Meier Brothers

Retained by Sidley, Austin, Brown & Wood, Dallas, TX as a technical expert on Copyright infringement allegations related to software owned by Optimum Return, LLC. Cyberkatz Consulting, Inc., Handsquare, LLC, Meier Brothers, et al., defendants, case number CA 3-03CV1064-D. Case heard in United States District Court, Northern District of Texas, Dallas Division. Expert for Meier Brothers. July 2004.

Parental Guide of Texas, Inc. v. Funai Corp., et. al.

Technical expert for defendants JVC and Panasonic in their dispute over non-infringement of US Patent number 4,605,964.

Elonex I.P. Holdings, LTD. and Elonex PLC, Phase II

Expert witness for defendants Chuntex, Acer, Tatung, Lite-On, Daewoo and Envision in their dispute with Elonex non-infringement and validity for US Patent numbers 5,389,952; 5,648,799; and 5,880,719.

Storage Computer Corporation vs. Veritas Software

Technical expert for the plaintiff in matters involving Patents US 5,257,367; US 5,893,919; and US 6,098,128.

Storage Computer Corporation vs. Seagate Technology LLC

Technical expert for the defendant in matters involving US Patent RE 34,100.

Elonex I.P. Holdings, LTD. and Elonex PLC, vs. Packard Bell et. al., CA 98-689-GMS

Expert witness for defendants ViewSonic Corp., Dell Computer Corporation, MAG Technology USA, Princeton Graphic Systems, Inc., Micron Electronics, Sony Electronics and Capetronic Computer USA in their dispute with Elonex non-infringement and validity for US Patent numbers 5,389,952; 5,648,799; and 5,880,719.

1.4.2 Engineering Consulting

Offspring Media Inc.

Technical consultant for the development of real-time auralization algorithms for integration into a consumer electronics product. Sep 2004.

Raytheon Company, Sudbury, MA

Development of techniques and requirements for implementing a fault tolerant computer system using software implemented fault tolerance (SIFT) techniques on commercial off-the-shelf processing hardware. The resultant system is to be used for highly reliable radar data and signal processing.

TVM Techno Venture Management

Provided consulting services to assist in assessing the technical claims of a company pursuing venture capital investment for a hardware implemented RAID 5 system .

Keyhold Engineering Inc., Northboro, MA.

Development of a prototype system for automatically calibrating multiple channel audio systems.

American Navigation Systems Inc., Milbury, MA.

Consulting on the development of the hand-held personal navigation and mapping system.

Lincoln Laboratory, Bedford, MA.

Simulated, tested and evaluated a GPS integrity monitoring algorithm developed at Lincoln Laboratories.

1.5 Licenses and Certifications

1.5.1 Commercial

General Radiotelephone Operator License
Ship RADAR endorsement

1.5.2 Amateur

Amateur Extra class radio operator license.

2. Courses Taught at WPI

2.1 Course Descriptions

Short descriptions of the courses taught are as follows:

EE572N Advanced System Architecture

This course focuses on the architectural techniques used to achieve high-performance in SISD and SIMD computer systems. In this course the interaction between the software application and hardware architecture and the effect of this interaction on achievable performance is stressed. The course begins by covering the basic architectural tricks used to enhance system performance and ends with a series of case studies that analyze specific architectures such as the CRAY and CDC vector supercomputers, the MasPar, the Connection Machine, the ICL DAP, and others.

ECE505 Computer Architecture

This course is an introductory graduate course in computer architecture. Most aspects of CPU architecture are covered using a combined hardware/software approach. Specific topics include datapath design, memory systems, microprogramming, memory management, operating systems, and instruction set design.

ECE579M Real-Time System Design

This course focuses on the design of computer systems for which the timeliness of producing results is a critical factor for establishing the correctness of the system design. Topics covered include hardware specification, real-time operating systems and programming, scheduling, communications, and validation/verification. Issues and choices arising for single processor and distributed systems are also covered. Both hard and soft real-time system issues and the interactions between real applications and real systems is stressed.

ECE2010 Introduction to Electrical and Computer Engineering.

The objective of this course is to introduce students to the broad field of electrical and computer engineering within the context of real world applications. This course is designed for first-year students who are considering ECE as a possible major or for non-ECE students fulfilling an out-of-major degree requirement. The course will introduce basic electrical circuit theory as well as analog and digital signal processing methods currently used to solve a variety of engineering design problems in areas such as entertainment and networking media, robotics, renewable energy and biomedical applications. Laboratory experiments based on these applications are used to reinforce basic concepts and develop laboratory skills, as well as to provide system-level understanding. Circuit and system simulation analysis tools are also introduced and emphasized.

ECE2022 Introduction to Digital Circuits and Computer Engineering.

The objective of this course is to expose students (including first year students) to basic electrical and mathematical concepts that underlie computer engineering while continuing an introduction to basic concepts of circuits and systems in a hands-on environment. Experiments representing practical devices introduce basic electrical engineering concepts and skills which typify the study and practice of electrical and computer engineering. In the laboratory, the students construct, troubleshoot, and test analog and digital circuits that they have designed. They will also be introduced to the nature of the interface between hardware and software in a typical microprocessor based computer.

ECE2801 Foundations of Embedded Systems

This course teaches the principles of programming microprocessors and microcontrollers for real-time applications. Students are introduced to software engineering principles and are taught how to translate product specifications into engineering solutions.

ECE2799 ECE Design

This is a new course added to the curriculum that teaches sophomore Electrical Engineering students the basic principles of design. Topics are covered which range from project planning and management through manufacturing and implementation. Students are exposed to external factors influencing design such as safety, liability, cost, and other constraints.

ECE3801 Logic Circuits

This is an introductory course in logic circuit design. Topics covered include Boolean Logic, Algebraic minimization of logic equations, Karnaugh Maps, sequential machine design and timing analysis.

ECE3803 Introduction to Microprocessor Systems

This is an undergraduate-level first-course in microprocessor design. Topics covered include timing analysis, address decoding, memory system design, assembly language programming, programmed I/O, and digital/analog interfacing. Experiments are run using ISA-bus interfaces to standard PCs.

ECE3810 Advanced Digital System Design

This course addresses the design of advanced digital logic systems using VHDL to design, synthesize and model digital circuits, and to implement these circuits using Xilinx FPGA devices. The course emphasizes understanding functional design, designing for speed and power objectives, and testing. The course ends with students designing moderately complicated “system on a chip” (SOC) based systems

ECE4815 Computer Architecture (crosslisted as CS4515)

A first course in computer architecture. Essential aspects of CPU architecture are covered using a combined hardware/software approach. Students learn how a CPU interprets and processes instructions. Issues associated with interfacing hardware with software are covered in detail as are the hardware/software tradeoffs associated with performance optimization.

ECE4801 Microprocessor System Design

Microprocessor System Design is the second course in the microprocessor sequence. In this course, students learn the advanced concepts used in modern microprocessor systems. Topics such as system organization, dynamic and cache memory systems, communications, mixed language programming, and device driver design are covered.

ECE430X Fundamentals of Navigation Systems

This course introduces students to the fundamentals of navigation using electronic systems. The course covers types of navigation systems, how to interpret sensor data and sources of navigation system error. Topics include: types of navigation systems (dead reckoning, inertial, radio based systems), sensors and error sources, coordinate frames and transformations, system dynamics and measurement processing. Case studies explore the use of accelerometers, gyroscopes, GPS (including, differential and assisted GPS) as well as other types of navigation systems.

RBE 3001 Unified Robotics III

Third of a four-course sequence introducing foundational theory and practice of robotics engineering from the fields of computer science, electrical engineering and mechanical engineering. The focus of this course is actuator design, embedded computing and complex response processes. Concepts of dynamic response as relates to vibration and motion planning will be presented. The principles of operation and interface methods various actuators will be discussed, including pneumatic, magnetic, piezoelectric, linear, stepper, etc. Complex feedback

mechanisms will be implemented using software executing in an embedded system. The necessary concepts for real-time processor programming, re-entrant code and interrupt signaling will be introduced. Laboratory sessions will culminate in the construction of a multi-module robotic system that exemplifies methods introduced during this course.

RBE 3002 Unified Robotics IV

Fourth of a four-course sequence introducing foundational theory and practice of robotics engineering from the fields of computer science, electrical engineering and mechanical engineering. The focus of this course is navigation, position estimation and communications. Concepts of dead reckoning, landmark updates, inertial sensors, vision and radio location will be explored. Control systems as applied to navigation will be presented. Communication, remote control and remote sensing for mobile robots and tele-robotic systems will be introduced. Wireless communications including wireless networks and typical local and wide area networking protocols will be discussed. Considerations will be discussed regarding operation in difficult environments such as underwater, aerospace, hazardous, etc. Laboratory sessions will be directed towards the solution of an open-ended problem over the course of the entire term.

RBE 400x Robot System Engineering and Design

The designers of robotic systems start with a system requirement to define the mechanical, electrical and software systems which must work together to achieve the system goals. Typically, parallel teams of engineers will work concurrently to create the requirements document as well as model various aspects of the system to verify operational capabilities and the ability to meet time and budget constraints. For complex systems, the development of such teams can itself be a complex problem since the project has to be organized in such a way that parallel teams can work independently, yet have excellent communication channels and information passing to insure project success.

This course explores the tools and techniques used to develop complex systems. The topics covered include: requirements development; system architecture and partitioning; requirements flowdown; functional and interface specifications; trade studies; system modeling and simulation; system integration; as well as design verification and validation.

RBE500 Foundations of Robotics

Foundations and principles of processing sensor information in robotic systems. Topics include an introduction to probabilistic concepts related to sensors, sensor signal processing, multi-sensor control systems and optimal estimation. The material presented will focus on the types of control problems encountered when a robot must operate in an environment where sensor noise and/or tracking errors are significant. Techniques for assessing the stability, controllability and expected accuracy of multi-sensor control and tracking systems will be presented. Lab projects will involve processing live and synthetic data, robot simulation and projects involving the control of robot platforms.

3. List of Publications:

3.1 Journal Papers

- [1] T. Padir, W.R. Michalson, et. al., "Implementation of an Undergraduate Robotics Engineering Curriculum," *Computers in Education Journal*, vol. 1, no. 3, pp. 92-101, 2010.
- [2] W. R. Michalson, A. Navalekar and H. Parikh, "Error mechanisms in indoor positioning systems without support from GNSS," *The Journal of Navigation*, Cambridge University Press, vol. 62, no. 2, pp. 239-49, 2009.
- [3] H.K. Parikh and W. R. Michalson, "Impulse Radio - UWB or Multicarrier UWB for Non-GPS based Indoor Precise Positioning Systems," *NAVIGATION J. Inst. Nav.*, Vol. 55, no. 1, 2008.
- [4] I. F. Proгри, W. R. Michalson, J. Wang and M.C. Bromberg, "Indoor Geolocation Using FCDMA Pseudolites: Signal Structure and Performance Analysis", *NAVIGATION J. Inst. Nav.*, Vol. 54, No. 3, Fall 2007.
- [5] I. F. Proгри, M. C. Bromberg and W. R. Michalson, "Maximum-Likelihood GPS Parameter Estimation", *NAVIGATION J. Inst. Nav.*, vol. 52, no. 4, pp. 229-238, Winter, 2005-2006.
- [6] I.F. Proгри, W.R. Michalson, and D. Cyganski, "An OFDM/FDMA Indoor Geolocation System," *NAVIGATION J. Inst. Nav.*, vol. 51, no. 2, pp. 133-142, Summer, 2004.
- [7] I.F. Proгри, G. Bogdanov, V.C. Ramanna, and W.R. Michalson, "An Investigation Of A GPS Adaptive Temporal Selective Attenuator," *NAVIGATION J. Inst. Nav.*, Vol. 49 No. 3, pp. 137-147, 2002.
- [8] G. Bogdanov, W. R. Michalson, and R. Ludwig, "A new apparatus for non-destructive evaluation of green-state powder metal compacts using the electrical resistivity method," *Measurement Science and Technology*, IOP Publishing, vol. 11, pp. 157-166, January 2000.
- [9] B. Findlen, E. Reuter, R. Campbell, and W. R. Michalson, "Effects of time domain response on the sonic characteristics of microphones," *Journal of the Acoustical Society of America*, vol. 104, no. 3, pt. 2 pp. 1764, September 1998.
- [10] J. Sedgwick, W. R. Michalson, and R. Ludwig, "Design of a Digital Gauss Meter for Precision Magnetic Field Measurements," *IEEE Transactions on Instrumentation and Measurement*, vol. 47, no. 4, pp. 972-977, August 1998.
- [11] J. Stander, J Plunkett, W. Michalson, J. McNeill, and R. Ludwig, "A Novel Multi-Probe Resistivity Approach to Inspect Green-State Powdered Metallurgy Compacts," *Journal of Non-Destructive Evaluation*, vol. 16, no. 4, pp. 205-214, 1997.
- [12] Woltereck, R. Ludwig, and W. Michalson, "A Quantitative Analysis of the Separation of Aluminum Cans Out of a Waste Stream Based on Eddy Current Induced Levitation," *IEEE Transactions on Magnetics*, vol. 33, no. 1, pp. 772-781, January 1997.
- [13] W. R. Michalson, "Auralization on a Laptop PC," abstract appears in *The Journal of the Acoustical Society of America* , vol. 100 No. 4, Pt 2, pp. 2608, October 1996.

- [14] R. H. Campbell, S. K. Martin, I. Schneider, and W. R. Michalson, "Analysis of Mosquito Wingbeat Sound," *The Journal of the Acoustical Society of America* vol. 100 No. 4, Pt 2, pp. 2710, October 1996.
- [15] W. R. Michalson, "Ensuring GPS Navigation Integrity using Receiver Autonomous Integrity Monitoring," *IEEE Aerospace and Electronic Systems Magazine*. vol. 10, no. 10, pp. 31-34, October 1995.
- [16] S. Clayton, R. J. Duckworth, W. Michalson, A. Wilson, "Determining Update Latency Bounds in Galactica Net," *Concurrency: Practice, and Experience*, vol. 7, no. 7, pp. 595-611, October 1994.

3.2 Conference Papers

- [1] K.C. Seals, W.R. Michalson, R.J. Hartnett and P.F. Swaszek, "Using Both GPS L1 C/A and L1C: Strategies to Improve Acquisition Sensitivity," Proc. 26th International Technical Meeting of the Satellite Division of the Institute of Navigation (ION GNSS 2013), Sep. 16-20, 2013.
- [2] K.C. Seals, W.R. Michalson, R.J. Hartnett and P.F. Swaszek, "Semi-Coherent and Differentially Coherent Integration for L1C Acquisition," Proc. 2013 International Technical Meeting of the Institute of Navigation (ION ITM 2013), Honolulu, HI, Apr. 22-25, 2013.
- [3] J.M. Barrett, M.G. Gennert, W. R. Michalson, et. al., "Development of a Low-Cost, Self-Contained, Combined Vision and Inertial Navigation System," 2013 International Conference on Technologies for Practical Robotic Applications, Boston, MA Apr. 22-23, 2012.
- [4] K.C. Seals, W.R. Michalson, R.J. Hartnett and P.F. Swaszek, "Analysis of L1C Acquisition by Combining Pilot and Data Components Over Multiple Code Periods," Proc. 2013 International Technical Meeting of the Institute of Navigation (ION ITM 2013), San Diego, CA, Jan. 28-30, 2013.
- [5] K.C. Seals, W.R. Michalson, R.J. Hartnett and P.F. Swaszek, "Analysis of Coherent Combining for L1C Acquisition," Proc. 25th International Technical Meeting of the Satellite Division of the Institute of Navigation (ION GNSS 2012), pp.384-393, Nashville, TN, Sep. 17-21, 2012.
- [6] J.M. Barrett, M.G. Gennert, W. R. Michalson and J.L. Center, "Analyzing and Modeling an IMU for Use in a Low-Cost Combined Vision and Inertial Navigation System," 2012 International Conference on Technologies for Practical Robotic Applications, Boston, MA Apr. 23-24, 2012.
- [7] G. Fischer, W. R. Michalson, T. Padir and G. Pollice, "Development of a Laboratory Kit for Robotics Engineering Education" AAAI 2010 Spring Symposium on Educational Robotics and Beyond: Design and Evaluation, Mar. 22-24, Palo Alto, CA, 2010.
- [8] W.R. Michalson, and F. J. Looft, "Designing Robotic Systems: Preparation for an Interdisciplinary Capstone Experience," American Society of Engineering Educators 2010 Annual Conference, Louisville, KY, Jun 20-23, 2010.

- [9] W.R. Michalson, S.J. Bitar and R.C.Labonte, "The Technical, Process, and Business Considerations For Engineering Design – A 10 Year Retrospective," American Society of Engineering Educators 2010 Annual Conference, Louisville, KY, Jun 20-23, 2010.
- [10] M. Gennert, M. Demietrio and W. R. Michalson, "A Robotics Engineering M.S. Degree," American Society of Engineering Educators 2010 Annual Conference, Louisville, KY, Jun 20-23, 2010.
- [11] R. Beach, W. R. Michalson, et. al., "Robotics Innovations Competition and Conference (RICC): Building Community Between Academia and Industry Through a University-Level Student Competition," American Society of Engineering Educators 2010 Annual Conference, Louisville, KY, Jun 20-23, 2010.
- [12] G. Tryggvason, W.R. Michalson, et. al., Teaching Multidisciplinary Design to Engineering Students: Robotics Capstone," American Society of Engineering Educators 2010 Annual Conference, Louisville, KY, Jun 20-23, 2010.
- [13] A.C.Navalekar, W.R. Michalson, "Asymmetric throughput problem due to Push-to-talk (PTT) delays in CSMA/CA based heterogeneous Land Mobile Radio (LMR) networks", accepted for publication, Milcom 2009.
- [14] A. Navalekar and W.R. Michalson, "Effects of Unintentional Denial of Service (DOS) due to PTT delays on performance of CSMA/CA based Adhoc Land Mobile Radio (LMR) networks", accepted for publication, ICST Adhocnet 2009.
- [15] W. R. Michalson, et. al., "Unified Robotics: Balancing Breadth and Depth in Engineering Education", American Society of Engineering Educators 2009 Annual Conference, AC 2009-1681, Austin, TX, Jun 14-17, 2009.
- [16] G. Tryggvason, W.R. Michalson, et. al., "Robotics Engineering: A New Discipline for a New Century", American Society of Engineering Educators 2009 Annual Conference, AC 2009-997, Austin, TX, Jun 14-17, 2009.
- [17] M. DiBlasi, W. R. Michalson, et. al., "Social Networking in the FIRST Robotics Competition Community," ASEE Northeast Section Conference, University of Bridgeport, Apr 3-4, Bridgeport, CT, 2009.
- [18] A. Navalekar, H.K Parikh and William R. Michalson, "Error Mechanisms in Indoor Positioning Systems without Support from GNSS", RIN NAV 08.
- [19] A. Navelekar, W.R. Michalson, et. al., "Effects of Push-To-Talk (PTT) delays on Throughput Performance of CSMA/CA based Distributed Digital Radios (DDR) for Land Mobile Radio (LMR) Networks," 37th International Conference on Parallel Processing (ICPP-08), Portland, OR, Sep. 8-12, 2008.
- [20] M. Ciraldi, W. Michalson, et. al., "The New Robotics Engineering BS Program at WPI," American Society of Engineering Educators 2008 Annual Conference, AC 2008-1048, Pittsburgh, PA, Jun 22-25, 2008.
- [21] A.C.Navalekar, W.R. Michalson, "A New Approach to Improve BER Performance of a High Peak-to-Average Ratio (PAR) OFDM signal over FM based Land Mobile Radios (LMR)", IEEE WTS 08, Pomona, CA.

- [22] A. Navelekar and W.R. Michalson, "Effects of Push-To-Talk (PTT) delays on CSMA based Capacity Limited Land Mobile Radio (LMR) Networks," *Proc. IEEE Intl. Symp. Wireless Pervasive Computing 2008 (ISWPC08)*, Santorini, Greece, May 7-9, 2008.
- [23] H.K. Parikh and W.R. Michalson, "Error Mechanisms In An Rf-Based Indoor Positioning System," *Proc. 2008 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '08)*, Las Vegas, NV Mar 30 – Apr 4, 2008.
- [24] B. Woodacre, W. Michalson, et. al., "WPI Precision Personnel Locator System - Automatic Antenna Geometry Estimation," to appear, *Proc. ION-NTM 2008*, San Diego, CA, Jan 27-29, 2008.
- [25] H.K. Parikh, A. Navelekar and W.R. Michalson, "Issues in Achieving Precise Positioning Indoors Without Support from GNSS," *Proc. ION-NTM 2008*, San Diego, CA, Jan 27-29, 2008.
- [26] D. Cyganski, W. Michalson, et. al., "WPI Precision Personnel Locator System - Evaluation by First Responders," *Proc. Institute of Navigation GNSS 2007*, Fort Worth, TX, September 25-28, 2007.
- [27] I.F. Proгри, W. R. Michalson, et. al., "Maximum Likelihood OFDMA Parameter Estimation," *Proc. Institute of Navigation GNSS 2007*, Fort Worth, TX, September 25-28, 2007.
- [28] C.W. Kelley, W. R. Michalson, et. al., "Discrete vs. Continuous Carrier Tracking Loop Theory, Implementation, and Testing with Large BnT," *Proc. Institute of Navigation GNSS 2007*, Fort Worth, TX, September 25-28, 2007.
- [29] W. R. Michalson and J. W. Matthews, "Distributed Digital Radios and WLAN Interoperability," 2007 IEEE Conference on Technologies for Homeland Security, May 16-17, Woburn, MA, 2007.
- [30] I.F. Proгри, W.R. Michalson, J. Wang, M.C. Bromberg, and R.J. Duckworth, "Requirements of a C-CDMA Pseudolite Indoor Geolocation System," *Proc. ION-AM 2007*, Cambridge, MA, pp. 654-658, Apr. 2007.
- [31] D. Cyganski, R. J. Duckworth, S. Makarov, W. R. Michalson, et.al., "WPI Precision Personnel Locator System," *Proc. Institute of Navigation National Technical Meeting (NTM 2007)*, Catamaran Resort Hotel, San Diego, CA, January 22-24, 2007.
- [32] I.F. Proгри, M.C. Bromberg, W.R. Michalson, and J. Wang, "A Theoretical Survey of the New GPS Signals (L1C, L2C, and L5)," *Proc. Institute of Navigation National Technical Meeting (NTM 2007)*, Catamaran Resort Hotel, San Diego, CA, January 22-24, 2007.
- [33] I.F. Proгри, M.C. Bromberg, W.R. Michalson, and J. Wang, "Field Measurement Data on Support of a Unified Indoor Geolocation Channel Model," *Proc. Institute of Navigation National Technical Meeting (NTM 2007)*, Catamaran Resort Hotel, San Diego, CA, January 22-24, 2007.
- [34] I. F. Proгри, J. Maynard, W.R. Michalson, and J. Wang, "The Performance and Simulation of a C-CDMA Pseudolite Indoor Geolocation System," *Proc. Institute of Navigation GNSS 2006*, Fort Worth, TX, September 26-29, 2006.

- [35] I. F. Proгри, W. Ortiz, W.R. Michalson, and J. Wang, "The Performance and Simulation of an OFDMA Pseudolite Indoor Geolocation System," *Proc. Institute of Navigation GNSS 2006*, Fort Worth, TX, September 26-29, 2006.
- [36] J.W. Coyne, R. J. Duckworth, W. R. Michalson and H.K. Parikh, "2-D Radio Navigation System Using MC-UWB," *Proc. NAV 2005 – Pushing the boundaries*, Royal Institute of Navigation, London, Nov 1-3, 2005.
- [37] H.K. Parikh, W.R. Michalson and R.J. Duckworth, "MC-UWB Precise Positioning System –Field Tests, Results and Effect of Multipath," *Proc. Institute of Navigation GNSS 2005*, Long Beach Convention Center, Long Beach, CA, September, 2005.
- [38] R.J. Duckworth, H.K. Parikh and W.R. Michalson, "Radio Design and Performance Analysis of a Multi-carrier Ultrawideband (MC-UWB) Positioning System," *Proc. Institute of Navigation National Technical Meeting (NTM 2005)*, Catamaran Resort Hotel, San Diego, CA, January 24-26, 2005.
- [39] H.K. Parikh and W.R. Michalson, "Performance Evaluation of the Receiver RF Front-End of a Precision Positioning System," *Proc. Institute of Navigation GNSS 2004*, Long Beach Convention Center, Long Beach, CA, September 21-24, 2004.
- [40] W.R. Michalson and H. Ahlehagh, "A 3D Location Discovery Algorithm for Ad Hoc Networks," *ICWN-04, 2004 International Conference on Wireless Networks*, Las Vegas, pp. 567-572, Jun 21-24, 2004.
- [41] I.F. Proгри, W.R. Michalson and M.C. Bromberg, "Accurate Synchronization Using a Full Duplex DSSS Channel," *Proc. IEEE PLANS 2004*, Monterrey, CA, pp. 220-226, Apr 26-29, 2004.
- [42] I.F. Proгри and W.R. Michalson, "An Investigation of a DSSS-OFDM-CDMA-FDMA Indoor Geolocation SYSTEM," *Proc. IEEE PLANS 2004*, 2004, Monterrey, CA, pp. 662-670, Apr 26-29, 2004.
- [43] D. Cyganski, J.A. Orr, W.R. Michalson, "Performance of a Precision Indoor Positioning System Using a Multi-Carrier Approach," *Proc. Institute of Navigation 2004 National Technical Meeting*, San Diego, CA, pp. 175-180, Jan 26-28, 2004.
- [44] I.F. Proгри, W.R. Michalson and M.C. Bromberg, "An Enhanced Acquisition Process of a Maximum Likelihood GPS Receiver," *Proc. Institute of Navigation 2004 National Technical Meeting*, San Diego, CA, pp. 390-398, Jan 26-28, 2004.
- [45] D. Cyganski, J. Orr, W. Michalson, "A Multi-Carrier Technique for Precision Geolocation for Indoor/Multipath Environments", *Proc. Institute of Navigation GPS 2003*, Portland, OR, pp. 1069-1073, Sep 9-12, 2003.
- [46] I.F. Proгри, M.C. Bromberg and W.R. Michalson, "The Acquisition Process of a Maximum Likelihood GPS Receiver," *Proc. Institute of Navigation GPS 2003*, Portland, OR, pp. 2533-2542, Sep 9-12, 2003.
- [47] W.R. Michalson, H. Ahlehagh and I.F. Proгри, "Dynamic Node Location in an Ad Hoc Indoor Communications and Positioning Network," *Proc. Institute of Navigation GPS 2003*, Portland, OR, pp. 1185-1191, Sep 9-12, 2003.

- [48] J. DeChiaro, C. Strus and W.R. Michalson, "A Personal Navigation Test Platform Based on Low-Cost Inertial Components," *Proc. Institute of Navigation GPS 2003*, Portland, OR, pp. 2869-2877, Sep 9-12, 2003.
- [49] H. AhleHagh, W. R. Michalson and D. Finkel, "Statistical Characteristics of Wireless Network Traffic and its Impact on Ad Hoc Network Performance," *Proc. Applied Telecommunication Symposium 2003 Advanced Simulation Technologies Conference*, Orlando, FL, pp. 66-71, Mar 30 – Apr 3, 2003.
- [50] I.F. Proгри, W.R. Michalson and D. Cyganski, "An OFDM/FDMA Indoor Geolocation System," *Proc. ION National Technical Meeting*, Anaheim, CA, pp. 272-281, Jan 22-24, 2003.
- [51] I.F. Proгри and W.R. Michalson, "Synchronization of Measurements of Power System Harmonics By Means of the Global Positioning System," *Proc. ION Annual Meeting*, Albuquerque, NM, pp. 372-382, Jun 24-26, 2002.
- [52] I.F. Proгри, W.R. Michalson, and M.C. Bromberg, "A recursive solution to the generalized eigenvalue problem," *Proc. ION Annual Meeting*, Albuquerque, NM, pp. 154-162, Jun 24-26, 2002.
- [53] I.F. Proгри and W.R. Michalson, "A Combined GPS Satellite/Pseudolite System for Category III Precision Landing," *Proc. IEEE PLANS*, pp. 212-218, Apr 15-17, 2002.
- [54] I.F. Proгри, W.R. Michalson and M. Bromberg, "A Comparison Between the Recursive Cholesky and MGSO Algorithms," *Proc. Institute of Navigation 2002 National Technical Meeting*, San Diego, CA, pp. 655-665, Jan 28-30, 2002.
- [55] I.F. Proгри, W.R. Michalson and M. Bromberg, "A Study of a Blind Adaptive Algorithm in the Time and Frequency Domain," *Proc. Institute of Navigation 2002 National Technical Meeting*, San Diego, CA, pp. 439-447, Jan 28-30, 2002.
- [56] L. Polizzotto and W. R. Michalson, "The Technical, Process, and Business Considerations for Engineering Design," *Frontiers in Education 2001*, Reno, NV, pp. F1G-19-F1G-24, Oct 10-13, 2001.
- [57] I.F. Proгри and W.R. Michalson, "An Improved Adaptive Spatial Temporal Selective Attenuator," *Proc. Institute of Navigation GPS 2001*, Salt Lake City, UT, pp. 932-938, Sep 11-14, 2001.
- [58] I.F. Proгри and W.R. Michalson, "An Alternative Approach to Multipath and the Near-Far Problem for Indoor Geolocation Systems," *Proc. Institute of Navigation GPS 2001*, Salt Lake City, UT, pp. 1434-143, Sep 11-14, 2001.
- [59] W.R. Michalson and I.F. Proгри, "An Investigation of the Adaptive Spatial Temporal Selective Attenuator," *Institute of Navigation GPS 2001*, Salt Lake City, UT, pp 1985-1996, Sep 11-14, 2001.
- [60] I.F. Proгри and W.R. Michalson, "The Impact of Proposed Pseudolite's Signal Structure on the Receiver's Phase and Code Error," *Proc. ION Annual Meeting*, Albuquerque, NM, pp. 414-422, Jun 11-13, 2001.

- [61] I.F. Proгри, J.M. Hill and W.R. Michalson, "An Investigation of the Pseudolite's Signal Structure for Indoor Positioning," *Proc. ION Annual Meeting*, Albuquerque, NM, pp. 453-462, Jun 11-13, 2001.
- [62] J. M. Hill and W. R. Michalson, "Real Time Verification of Bit-Cell Alignment for C/A Code Only Receivers," *Proc. ION Annual Meeting*, pp. 463-468, Jun 11-13, 2001.
- [63] I. Proгри , J. M. Hill and W. R. Michalson, "The Impact of the Proposed Pseudolite's Signal Structure on the Receiver's Phase and Code Error," *Proc. ION Annual Meeting*, pp. 414-422, Jun 11-13, 2001.
- [64] I.F. Proгри, W.R. Michalson and R. Chassaing "Fast and efficient filter design and implementation on the TMS320C6711 digital signal processor," *Proc. Student Forum ICASSP*, Salt Lake City UT, May 2001.
- [65] I. Proгри and W. R. Michalson, "An Innovative Navigation Algorithm Using a System of Fixed Pseudolites," *Institute of Navigation National Technical Meeting*, pp. 619-627, Long Beach, CA, Jan 22-24, 2001.
- [66] I. Proгри, J.M. Hill and W. R. Michalson, "A Doppler-based navigation algorithm," *Institute of Navigation National Technical Meeting*, pp. 482-490, Long Beach, CA, Jan 22-24, 2001.
- [67] I. Proгри , J. M. Hill and W. R. Michalson, "Assessing the Accuracy of Navigation Algorithms Using a Combined System of GPS Satellites and Pseudolites," *Institute of Navigation National Technical Meeting*, Long Beach, CA, pp 473-481, Jan 22-24, 2001.
- [68] J. M. Hill and W. R. Michalson, "Design of a Stable Discrete Time Costas Loop," *Institute of Navigation National Technical Meeting*, pp. 228-234, Jan 22-24, 2001.
- [69] J. M. Hill and W. R. Michalson, "Techniques for Reducing the Near-Far Problem in Indoor Geolocation Systems," *Institute of Navigation National Technical Meeting*, pp. 860-865, Jan 22-24, 2001.
- [70] I. Proгри and W. R. Michalson, "Adaptive Spatial and Temporal Selective Attenuator in the Presence of Mutual Coupling and Channel Errors," *Institute of Navigation GPS 2000*, Salt Lake City, UT, pp. 462-470, Sep 19-22, 2000.
- [71] W. R. Michalson, J. A. Orr and D. Cyganski, "A System for Tracking and Locating Emergency Personnel Inside Buildings," *Institute of Navigation GPS 2000*, Salt Lake City, UT, pp. 560-568, Sep 19-22, 2000.
- [72] W. R. Michalson and I. Proгри, "Assessing the Accuracy of Underground Positioning Using Pseudolites," *Institute of Navigation GPS 2000*, Salt Lake City, UT, pp. 1007-1015, Sep 19-22, 2000.
- [73] I. Proгри and W. R. Michalson, "Performance Evaluation of Category III Precision Landing Using Airport Pseudolites," *IEEE Conf. Position, Location, and Navigations Systems (PLANS)*, Mar, pp. 262-269, 2000.
- [74] R. Ludwig, G. Bogdanov, W. Michalson, and D. Apelian, "Instrumentation Development for Crack Detection of Surface and Subsurface Defects in Green-State P/M Compacts through Multi-Probe Electric Resistivity Testing," *Review of Progress in NDE*, Snowbird Ski and Summer Resort, Snowbird UT, Jul 19-24, 1998.

- [75] W. R. Michalson, W. Cidela, et. al., "A GPS-Based Hazard Detection and Warning System," in review, " *ION GPS-96, 9th International Meeting of the Satellite Division of the Institute of Navigation*, pp. 167-175, Kansas City, MO, Sep 17-20, 1996.
- [76] W. R. Michalson and R. L. Labonte, "Capstone Design in the ECE Curriculum: Assessing the Quality of Undergraduate Projects at WPI," *American Society of Engineering Educators 1996 Annual Conference (CD-ROM)*, session 1232 Washington, D.C., Jun, 1996.
- [77] W. R. Michalson, J. Single, M. Spadazzi, and M. Wehr, "A Real-Time GPS System for Monitoring Forestry Operations," *Sixth Biennial Forest Service Remote Sensing Applications Conference*, pp. 264-269, Denver, CO, May 1-3, 1996.
- [78] W. R. Michalson and D. Nicoletti, "Computers in Introductory and Upper-Level ECE Courses," *American Society of Engineering Educators 1995 Annual Conference*, Anaheim, CA, Jun 25-28, 1995, pp. 2795-99.
- [79] W. R. Michalson, D. B. Cox, and H. Hua, "GPS Carrier-Phase RAIM," *ION GPS-95, 8th International Meeting of the Satellite Division of the Institute of Navigation*, Palm Springs, CA., pp. 1975-1984, Sep 12-15, 1995.
- [80] J. Bernick and W. R. Michalson, "UDSRAIM: An Innovative Approach to Increasing RAIM Availability," *ION GPS-95, 8th International Meeting of the Satellite Division of the Institute of Navigation*, Sep 12-15, Palm Springs, CA., pp. 1965-1973, 1995.
- [81] C. Easton and W. R. Michalson, "Effects of Worst Case Geometries on RAIM Testing," *ION GPS-95, 8th International Meeting of the Satellite Division of the Institute of Navigation*, Sep 12-15, Palm Springs, CA., pp. 2015-2022, 1995.
- [82] D. B. Cox and W. R. Michalson, "Use of Uncorrected GPS Carrier Phase Measurements for Incremental RAIM with WAAS," *ION 51st Annual Meeting*, Jun 5-7, Colorado Springs, CO., pp. 515-520, 1995.
- [83] W. Michalson, et. al., "RAIM Availability for Augmented GPS-Based Navigation Systems," *ION GPS-94, 7th International Meeting of the Satellite Division of the Institute of Navigation*, pp. 587-95, Sep 20-23, 1994.
- [84] V. G. Virball, W. Michalson, et. al., "A GPS Integrity Channel Based Fault Detection and Exclusion Algorithm Using Maximum Solution Separation," *Proceedings of the 1994 IEEE Position Location and Navigation Symposium (PLANS-94)*, pp. 747-54, Las Vegas, Apr 11-15, 1994.
- [85] W. Michalson and C. Easton, "Experiences Implementing the Supplemental MOPS Off-Line Test Procedure," *ION GPS-93, 6th International Meeting of the Satellite Division of the Institute of Navigation*, pp. 519-27, Sep 22-24, 1993.
- [86] D. Beatovic, P.L. Levin, A. Meroth, M. Spasojevic, W. R. Michalson, A. Unstundag, "Iterative Matrix Solvers for Large Full Systems," *Eighth International Symposium on High Voltage Engineering*, Aug 23-27, Yokohama, Japan, 1993.
- [87] E. Schnieder, W. R. Michalson, "A Comparison of Large Guided-Wave Interconnection Networks for Optical Computation Systems," *SPIE Conf. on Optoelectronic Interconnects OE/LASE 93*.

- [88] E. Schnieder, W. R. Michalson, "Integrated Guided-Wave Crossbar Interconnection of SEED Arrays," SPIE Conf. on Optoelectronic Interconnects OE/LASE 93.
- [89] W. R. Michalson, "The Application of Neural Networks to Nonlinear Filtering," Proc. SPIE Cooperative Intelligent Robotics in Space III, pp. 219-228, Boston, 1992.
- [90] W. R. Michalson and C. S. Tocci, "Exploiting Programmable Devices in High-Performance System Design: Current Trends," *Proc. Electro 92*, vol 2, pp. 104-109, May 12-14, 1992.
- [91] S. Clayton, R. J. Duckworth, W. Michalson, A. Wilson, "Determining Update Latency Bounds in Galactica Net," *Proc. IEEE Conf. on High-Performance Distributed Computing*, pp. 104-111, Sep 9-11, 1992.
- [92] R. J. Duckworth, J. E. Lavalley, W. R. Michalson, L. Becker, and P. Green, "The Development of Intelligent Real-Time Systems Using Ada," *12th IEEE Symposium on Real-Time Systems*, Dec 3-6, 1991.
- [93] W. Jessop, W. Michalson, and R. Some, "Fault Injection for Verifying Fault Tolerant System Behavior," *Workshop on Experimental Evaluation*, El Segundo, CA, May 1-3, 1990.
- [94] W. R. Michalson and P. Heldt, "A Hybrid Architecture for the ART 2 Neural Model," *Proc. International Joint Conference on Neural Networks*, pp. 167-70, Washington D.C., Jan 15-19, 1990.
- [95] W. R. Michalson, "A Review of the Current State of Logic Synthesis," *2nd Annual IEEE ASIC Seminar and Exhibit*, Rochester, NY, Sep 25- 28, 1989.
- [96] P. E. Green and W. R. Michalson, "Real-Time Evidential Reasoning and Network Based Processing," *Proc. IEEE 1st International Conference on Neural Networks*, pp. 359-365, San Diego, CA, Jun 21-24, 1987.
- [97] P. E. Green, R. J. Juels, and W. R. Michalson, "Real Time Artificial Intelligence Architecture," *Proc. Workshop on Future Directions in Computer Architecture and Software*, pp. 328-330, Charleston, SC, May 5-7, 1986.

3.3 Book Chapters

- [1] W. Michalson and E. Schnieder, "An Approach for Implementing a Reconfigurable Optical Interconnection Network for Massively Parallel Computers," in *Optical Interconnection - Foundations and Approaches*, C. Tocci and H. J. Caulfield Eds., Artech House, proposed release January 1994.

3.4 Patents

Precision location methods and systems

United States Patent 8,928,459, Issued January 6, 2015

The invention describes systems and methods for determining the location of a transmitter by jointly and collectively processing the full sampled signal data from a plurality of receivers to form a single solution.

Apparatus and methods for addressable communication using voice-grade radios

United States Patent 8,284,711, Issued 9 October 2012

The invention relates to methods and apparatus for conducting directed communication using voice-grade radios. The methods and apparatus can be used to form a packet-switched wireless network using legacy analog transceivers, providing, e.g., both data and voice-over-Internet Protocol communication.

Multi-channel electrophysiologic signal data acquisition system on an integrated circuit.

United States Patent 7896807, issued 3 March 2011.

A physiologic data acquisition system includes an analog input, a sigma-delta front end signal conditioning circuit adapted to subtract out DC and low frequency interfering signals from and amplify the analog input before analog to digital conversion. The system can be programmed to acquire a selected physiologic signal, e.g., a physiologic signal characteristic of or originating from a particular biological tissue. The physiologic data acquisition system may include a network interface modulating a plurality of subcarriers with respective portions of an acquired physiologic signal. A receiver coupled to the network interface can receive physiologic data from, and send control signals and provide power to the physiologic data acquisition system over a single pair of wires. The network interface can modulate an RF carrier with the plurality of modulated subcarriers and transmit the resulting signal to the receiver across a wireless network. An integrated circuit may include the physiologic data acquisition system. Also included are methods for acquiring physiologic data comprising the step of selectively controlling an acquisition circuit to acquire the physiologic signal.

Methods and apparatus for high resolution positioning. United States Patent 7292189. The invention relates to a method of signal analysis that determines the location of a transmitter and to devices that implement the method. The method includes receiving by at least three receivers, from a transmitter, a first continuous-time signal having a first channel. The first channel includes a first plurality of signal carriers having known relative initial phases and having known frequencies which are periodically spaced and which are orthogonal to one another within a first frequency range. The signal analysis method also includes determining the phase shifts of the carriers of the first channel resulting from the distance the carriers traveled in reaching the first receiver. Analysis of the phase shifts yields time difference of arrival information amongst the receivers, which is further processed to determine the location of the transmitter. 6 Nov 2007.

A Reconfigurable Indoor Geolocation System, US Patent Number 7,079,025. A portable reconfigurable geolocation system is provided. The system includes a portable user node and one or more portable pseudolite nodes in communication one another and with the user node. Each of the user nodes and pseudolite nodes includes a transmitter that generates a signal on one or more carrier frequencies. Each signal is modulated with digital signals necessary to establish distances between the nodes and to convey data between the nodes. Each node also includes a receiver for receiving and demodulating the signals transmitted between the nodes, and a processor for receiving the demodulated signals, extracting data values and derived values from the

demodulated signals and determining a three-dimensional position of each node in the system. Issued 18 Jul 2006.

Auto-Calibrating Surround System, United States Patent 7158643. A multi-channel surround sound system and method is described that allows automatic and independent calibration and adjustment of the frequency, amplitude and time response of each channel of the surround sound system. The disclosed auto-calibrating surround sound (ACSS) system includes a processor that generates a test signal represented by a temporal maximum length sequence (MLS) and supplies the test signal as part of an electric input signal to a loudspeaker. A microphone coupled to the processor receives the signal in a listening environment. The processor correlates the received sound signal with the test signal in the time domain and determines from the correlated signals a whitened response of the audio channel in the listening environment. Issued 2 Jan 2007.

Hand-held GPS-mapping device, US. Patent Number 5,987,380. A hand-held navigation, mapping and positioning device contains a GPS receiver, a database capable of storing vector or bit mapped graphics, a viewing port, an embedded processor, a simplified user interface, a data compression algorithm, and other supporting electronics. The viewport is configured such that the data presented in the viewport is clearly visible in any ambient light condition. The database stores compressed image data which might include topographical map data, user annotations, building plans, or any other image. The system includes an interface to a personal computer which may be used to annotate or edit graphic information externally to the device for later upload. In addition, the device contains a simple menu-driven user interface which allows panning and zooming the image data, marking locations of interest, and other such functions. The device may be operated from an internal rechargeable battery, or powered externally. , Issued 16 Nov 1999.

Hand-held GPS-mapping device, US. Patent Number 5,902,347. A hand-held navigation, mapping and positioning device contains a GPS receiver, a database capable of storing vector or bit mapped graphics, a viewing port, an embedded processor, a simplified user interface, a data compression algorithm, and other supporting electronics. The viewpoint is configured such that the data presented in the viewport is clearly visible in any ambient light condition. The database stores compressed image data which might include topographical map data, user annotations, building plans, or any other image. The system includes an interface to a personal computer which may be used to annotate or edit graphic information externally to the device for later upload. In addition, the device contains a simple menu-driven user interface which allows panning and zooming the image data, marking locations of interest, and other such functions. The device may be operated from an internal rechargeable battery, or powered externally. Issued 11 May 19/99.

3.5 Professional Presentations

American Ambulance Association Annual Meeting: Low-cost VHF/UHF Interoperability for digital telemetry, Las Vegas, NV, Dec. 2005.

California Ambulance Association, Keynote address: Alternatives to solving interoperability problems in Land Mobile Radios, Lake Tahoe, NV, July 2005.

Museum of Science Lecture Series: The Next Generation of Information and Communications Technologies-What does the Future Hold?, William R. Michalson and Brian King, January 14, 2004.

Agilent Wireless Technology Summit: Dynamic Node Location in an Ad Hoc Indoor Communications and Positioning Network, William R. Michalson, January 27, 2004

Security & Technology Online (SATO) Security Leadership Council. Panel discussion on Smart Surveillance, Command and Control, Oct 28-29, 2004.

“Worcester Polytechnic Institute Barcelona Summit: The Future of Information Technology,” delivered presentation entitled “Personal Navigation in the Information Age,” Apr 2001.

4. Projects advised (undergraduate).

4.1 Major Qualifying Projects (current)

[1] *Voice Release System, B. Waldron, WZM-MQP-1M10, in process.*

4.2 Major Qualifying Projects (completed)

[2] *Aeacus, N. Anderson, D. Praetorius and C. Roddy, co-advised with S. Nestinger, 2011.*

[3] *Realization of Performance Advancements for WPI’s UGV-Prometheus, M. Akmanalp, R. Doherty, J. Gorges, P. Kalasuskas, E. Peterson and F. Polido, co-advised with T. Padir, S. Nestinger, M. Ciraldi, K. Stafford, 2011.*

[4] *Autonomous Underwater Vehicle, J. Baker, C. Frumento, J. Grzyb and T. North, co-advised w/I. Hussein, 2011.*

[1] *Tactical Vest, V. Brisian, J. Fernando, A. Khandaker and J. Zorrilla DeLos Santos, 2011.*

[2] *Marsupial AUV, N. Smith, B. Berard and C. Pietre, Lincoln Laboratory Project Center, co-advised with G. Heiniman, 2010.*

[3] *Voice Release System, J. Low, WZM-MQP-1M10, 2010.*

[4] *Design and Realization of an Intelligent Unmanned Ground Vehicle, J. Barrett, B. Roy and D. Sacco, Co-Advised w/T. Padir, 2010.*

[5] *Accurate Real-Time Audio Circuit Simulation, B. Gleason, WZM-MB09, 2010.*

[6] *Optimization and Control Design of an Autonomous Underwater Vehicle, D. Moussette, A. Palooparambil, and J. Raymond, AE- IIH-0003, co-advised w/I. Hussein, 2010.*

[7] *Design of Autonomous Underwater Vehicle and Optimization of Hydrodynamic Properties and Control, R. David, WZM-3A08, 2009.*

[8] *Robotic Bass Player, B. Kosherick, M. Brown, and A. Teti, WZM-RB08, 2008.*

- [9] *Public Safety Radio System*, P. Lucia, I. Levin and M. Barone, WZM-1A08, 2008.
- [10] *Aircraft Lasercom Terminal Compact Optical Module*, B. Scoville and S. Rose, Lincoln Laboratory Project Center, WZM-2A08, 2008.
- [11] *GPS Attitude Determination System*, J.P. Salmon, Michael LaBossiere and Mark Minotaur, 2005.
- [12] *FPGA-Based VHF Modem With Integrated Testability*, Andrew Dupont and Jack Coyne, 2005.
- [13] *TMR Computer System*, Maulin Patel, Omar Moussa and Matthew Kwiatkowski, 2005.
- [14] *GPS Signal Generator*, Tim Coffey, 2005.
- [15] *Dipole Antenna Placement in a Falcon-20 Aircraft*, Emily Anesta and David Plourde, Lincoln Laboratories Project Center, A-Term, 2004.
- [16] *GPS Attitude Determination System*, Joshua Holwell, Himanshu Agrawal and Andrew Coonradt, 2004.
- [17] *TMR Computer System*, Ryan Angilly, Mitch Lauer and Dan DeBiasio, 2004.
- [18] *Personal Inertial Navigation System*, Jason DeChiaro and Christopher Strus, 2003.
- [19] WZM-MQP-4A02: *PC I/O in High Stress Environments*, John Niesz and James Kent, 2003.
- [20] WZM-MQP-2A02: *Vacuum Tube Amplifier*, Joseph Kambourakis and Gregory Molnar, 2003.
- [21] *Container Tracking System*, Victoria Chaplick, 2003.
- [22] WZM-MQP-2A03: *Heat Management System for PCs*, Ernest Cardin, Kevin Candiloro and Stephen Leavey, 2002.
- [23] WZM-MQP-1313: *Digital Image Enhancement*, Julie Bolduc, Joeseph Perry, Wei Fu, 2002.
- [24] WZM-MQP-2A01: *Synchronized Audio Sample Looper*, Joel Gottshalk, Robert Conrad and Sanford Freedman, 2002.
- [25] WZM-MQP-1A01: *Springboard Digital Multimeter*, Pavel Loven and Andrew Young, 2002.
- [26] *Ballistic Missile Defense Analysis Toolkit*, Winfield Peterson, Doug Tilkin and Benjamin Wilson, Lincoln Laboratories Project Center, 2002.
- [27] HU-FB-CS01, *C Sound Synthesizer*, Peter W. DeBonte (co-advised).
- [28] WZM-MQP-1A00: *StrongArm-Based Computer System*, Bradford Snow, 2001.
- [29] WZM-MQP-1C01: *PC Controlled Laser Light Show Device*, Joel Smith, 2001.
- [30] WZM-MQP-1E00: *PIC-based MIDI Sequencer* Malcolm Beaulieu, 2001.
- [31] WZM-MQP-2A00: *Automotive PC Development Platform* Travis Pouliot and David Philips, 2001.

- [32] *The RoadCom Automotive Computing System*, Benjamin Kennedy and John Pong, 2000.
- [33] WZM-MQP-1A99: *Compressed Sample Wavetable Synthesizer*, Justin Brzozoski, 2000.
- [34] WZM-MQP-3499, *Automatically Equalizing Monitor*, Fernando Braghin, Tenzin Lama, Rahul Bhan and Dion Soetadi, 2000.
- [35] 99D163M: *Railroad Communication*, Benjamin Richards, 2000.
- [36] CS-MXC-IE00: *PIC Real-Time Sequencer*, Alexander Goodrich, 2000.
- [37] 99D514M: *Design of a Microphone Preamplifier*, Eric Reuter, 2000.
- [38] WZM-MQP-1A99, *MPEG Audio Deck*, Justin Brzozoski, 2000.
- [39] *Digital Image Enhancement*, Julie Bolduc, Wei Fu and Joseph Perry, 2000.
- [40] WZM-MQP-4A98, *Railroad Communications System*, Matthew Lug, 1999.
- [41] 99D078M: *Modular Effects Processor II*, Erik Neyland, 1999.
- [42] 99D176M: *Portable Digital Audio Recorder* Eric Toledo and Duc Truong, 1999.
- [43] EE-WZM-1A97, *C Sound Synthesizer*, Ross E. Borgeson, Michael W. Hamel and Matthew S. Walsh, 1998.
- [44] EE-WZM-4A97, *Firewire Audio Device*, Daniel R. Stutzbach, 1998.
- [45] EE-WZM-2A97, *Modular Effects Processor*, Michael J. Dellisanti, 1998.
- [46] EE-WZM-3A97, *GPS Personal Navigation*, Jeffery A. Alderson and Helder Machado, 1998.
- [47] HU-FB-CS01, *C Sound Synthesizer*, Peter W. DeBonte (co-advised).
- [48] EE-WZM-1E97, *PM Measurement System*, Yevgeniy Bogdanov.
- [49] EE-REL-C008, *Design and Development of a Microprocessor-Based Gaussmeter*, David M. Burnham.
- [50] EE-WZM-RC01, *Acoustic Guitar Amplifier*, Christopher Thomas.
- [51] EE-WZM-GSD1, *Guitar Sustaining Device*, Paul D'Ambra.
- [52] EE-RXV-5260, *Audio Feedback Elimination System*, Ross D. Pease and John R. Pelliccio.
- [53] EE-RJD-M963, *Embedded Systems Design*, Christopher A. Briggs and Anthony J. Viapiano.
- [54] EE-WHE-9601, *GPS Hazard Detector*, Michael Roberts, William Cidela, and Chris Mangiarelli.
- [55] EE-WZM-2C96, *Flexible Synthesis*, Noah T. Vawter and Luke Demoracski.
- [56] EE-WZM-1A96, *Tap Dancer MIDI Interface*, Thomas Trela and William Dowell.
- [57] EE-WZM-2A96, *GPS Hazard Detector II*, Will Brothers, Jon Day, and John Zaghi.
- [58] EE-WZM-3A96, *Loudspeaker Data Acquisition System*, Adam Gross.
- [59] EE-WZM-4A96, *Audio Morphing Processor*, William Butterfield and Ted Phipps.

- [60] EE-WZM-5A96, *Acoustic Modeling*, Peter DeBonte.
- [61] EE-WZM-1B96, *Distributed Audio Controller*, Stephen S. Richardson.
- [62] EE-WZM-1A95, *Acoustic Hazard Meter*, Ronald D. Slack.
- [63] EE-WZM-2A95, *Forest Service DGPS*, Joshua J. Single and Michael T. Spadazzi.
- [64] EE-WZM-1C95, *Audio to MIDI Converter*, Jennifer R. Principe.
- [65] EE-WZM-1D95, *Low Cost Auralizer*, Jason R. Hills and Mark R. Paulson.
- [66] EE-WZM-3A95, *Passive Radiator Design*, Kevin R. Weldon.
- [67] EE-WZM-1C94, *Char Model Generation*, Colin J. Florendo.
- [68] EE-WZM-1D94, *Wide Area DGPS Simulator*, Daniel Cohen and Robert Schroter.
- [69] EE-WZM-2D94, *Digital Soundcard*, Timothy Alsberg (Russian Project Center).
- [70] EE-WZM-3D94, *Digital Univibe*, Andrew Willis and Daniel Toohey.
- [71] EE-WZM-1A94, *DSP Based Real-Time Audio Feedback Eliminator*, Kevin M. Eddy.
- [72] EE-WZM-2A94, *Digital LCD Oscilloscope*, William F. Brown and John F. Ebersole.
- [73] 93D236M, *MIDI Mapper*, Jonathan Kemble and Brian Candiloro.
- [74] EE-WZM-1C93, *Fault-Tolerant Computer*, Frederick N. Parmenter.
- [75] EE-WZM-1A93, *Wireless MIDI Controller*, Sanjay Raja, Charles Cimalore, Ty Panagoplos.
- [76] EE-WZM-2A93, *Multiple Pitch Detector*, Jeanne A. Sawtelle.
- [77] EE-WZM-3A93, *Multiprocessor Cache Coherence*, Lauren C. Lind and Norman E. Rhodes.
- [78] EE-WZM-1C92, *A Simulation of the DLX Architecture*, Lisa Harlow.
- [79] EE-WZM-2C92, *A New Microprocessor Development System*, Gregory B. Burlingame, David J. Fortin, Kevin S. Pearson.
- [80] EE-WZM-1A92, *Digital Audio Sampler*, Roger D. Gagnon and James M. Lach.
- [81] EE-WZM-2A92, *Intelligent Harmonizer*, Prabhjot S. Anand and Aftab M. Yusuf.
- [82] EE-WZM-3A92, *Computerized Audio Mixer*, Richard J. Wood.
- [83] EE-WZM-1B92, *Real-Time Harmonizer*, Mohiuddin M. Kahn.
- [84] EE-WZM-1A91, *Residue Number System Processor*, Ravdeep S. Anand and Christine A. Easton.
- [85] EE-WZM-2A91, *SCSI Bus Analyzer*, Brian Costello, George Delouriero, Matthew Maguire, and Keith Nevins.

4.3 Graduate Theses Advised and Co-Advised

4.3.1 MS Theses (current)

- [1] No Current MS Students

4.3.2 MS Theses (completed)

- [1] Morin, Russell, "A Novel Localization System For Experimental Autonomous Underwater Vehicles," MS Thesis, co-advisor, Worcester Polytechnic Institute, 2010.
- [2] Navalekar, Abhijit, "Design of an OFDM-Based VHF Modem," MS Thesis, primary advisor, Worcester Polytechnic Institute.
- [3] Ahlehagh, Hasti, "Techniques for Communications and Geolocation Using Wireless Ad Hoc Networks," MS Thesis, primary advisor, Worcester Polytechnic Institute, 2004.
- [4] Sebastian, Dalys, "Development of a Field-Deployable Ultrasound Scanner System," MS Thesis, co-advisor, Worcester Polytechnic Institute, 2004.
- [5] Tobgay, Sonam, "Novel Concepts for RF Surface Coils with Integrated Receivers," MS Thesis, co-advisor, Worcester Polytechnic Institute, 2004.
- [6] Breen, Daniel, "Characterization of Multi-Carrier Locator Performance," MS Thesis, co-advisor, 2004.
- [7] Aghogho, Obi, "A Novel Radio Frequency Coil Design for Breast Cancer Screening in a Magnetic Resonance Imaging System," MS Thesis, co-advisor, Worcester Polytechnic Institute, 2003.
- [8] Fei, Ming, "Electromagnetic Detection, Infrared Visualization and Image Processing Techniques for Non-Metallic Inclusions in Molten Aluminum," MS Thesis, co-advisor, 2002.
- [9] Lavoie, Bruce, "Design and Implementation of an N-Channel Self Calibrating Audio System," MS Thesis, primary advisor, Worcester Polytechnic Institute, 2000.
- [10] Bogdonov, Gene, "Theoretical and Practical Implementation of Electrical Impedance Material Inspection of Powder Metallurgy Compacts," MS Thesis, co-advisor, Worcester Polytechnic Institute, 1999.
- [11] Messier, Andrew, "Modeling the Effects of Terrain Masking on GPS Accuracy and Integrity," MS Thesis, primary advisor, Worcester Polytechnic Institute, 1998.
- [12] Antonescu, Bogdan, "Elliptic Curve Cryptosystems on Embedded Microprocessors," Bogdan Antonescu, MS Thesis, co-advisor, Worcester Polytechnic Institute, 1998.
- [13] Lai, Qiang, "Ground-Penetrating Radar Data Processing System," MS Thesis, co-advisor, Worcester Polytechnic Institute, 1998.
- [14] Soria-Rodríguez, Pedro, "Multicast-Based Interactive-Group Object-Replication For Fault Tolerance," MS Thesis, co-advisor, Worcester Polytechnic Institute, 1998.

- [15] Hoy, William, "Audio Signal Denoising Using Wavelets," MS Thesis, primary advisor, Worcester Polytechnic Institute, 1997.
- [16] Progri, Ilir, "Harmonic Flow Monitoring by means of Global Positioning System," MS Thesis, primary advisor, Worcester Polytechnic Institute, 1997.
- [17] Bretchko, Pavel, "Pulsed Hysteresis Graph System," MS Thesis, co-advisor, Worcester Polytechnic Institute, 1997.
- [18] Repkin, Dmitry V., "A Hierarchical Neural Network Based Data Processing System for Ground Penetrating Radar," MS Thesis, co-advisor, Worcester Polytechnic Institute, 1997.
- [19] Metsis, Sophocles, "Design of a Real-Time Capable, Fault-Tolerant, Distributed System," MS Thesis, primary advisor, Worcester Polytechnic Institute, 1996.
- [20] Hill, Jonathan, "Efficient Implementation of Mesh Generation and FDTD Simulation of Electromagnetic Fields," MS Thesis, primary advisor, Worcester Polytechnic Institute, 1996.
- [21] Dunkelberg, John, "FEM Mesh Mapping to a SIMD Machine Using Genetic Algorithms," MS Thesis, co-advisor, Worcester Polytechnic Institute, 1996.
- [22] Leuenberger, Georg, "Design and Development of a Microprocessor Based Gauss Meter," MS Thesis, co-advisor, Worcester Polytechnic Institute, 1995.
- [23] Valentino, Ralph, "DISC: A Dynamic Instruction Set Coprocessor," MS Thesis, primary advisor, Worcester Polytechnic Institute, 1995.
- [24] Muley, Aalok, "A Fault Tolerant Network for a Real-Time Environment," MS Thesis, primary advisor, Worcester Polytechnic Institute, 1994.
- [25] Mohan, Surrender, "Automatic Surface Mesh Generation for 3D Solid Models Using Delaunay Algorithm," MS Thesis, co-advisor, Worcester Polytechnic Institute, 1994.
- [26] Petrangelo, John, "Experimental Preconditioners for Large Dense Systems," MS Thesis, co-advisor, Worcester Polytechnic Institute, 1994.
- [27] Schneider, Eric, "Design, Simulation, and Analysis of a 3D Integrated Optical Computer," MS Thesis, primary advisor, Worcester Polytechnic Institute, 1993.
- [28] Palmer, Bradley, "A Comparison of Three Protocols Supporting Time-Dependent and Time-Independent Communications," MS Thesis, primary advisor, Worcester Polytechnic Institute, 1992.
- [29] Clayton, Shawn, "An Analysis of the Real-Time Behavior of Galactica Net," MS Thesis, primary advisor, Worcester Polytechnic Institute, 1992.
- [30] Levergood, Thomas, "An Experimental Evaluation of Split User/Supervisor Cache Memories," MS Thesis, primary advisor, Worcester Polytechnic Institute, 1992.
- [31] Lavalee, James, "The Design and Development of Real-Time Systems Using Ada and the Activation Framework," MS Thesis, co-advisor, Worcester Polytechnic Institute, 1992.
- [32] Velazques, Javier, "The Development of a Real-Time Environment Using the Activation Framework," MS Thesis, co-advisor, Worcester Polytechnic Institute, 1992.

4.3.3 Ph. D. Dissertations (current)

- [1] Jitesh, “Ad-Hoc Networking for Bandwidth Limited LMR Systems,” primary advisor.

4.3.4 Ph. D. Dissertations (completed)

- [1] Iyer, Vishwanath, “Broadband Impedance Matching of Antenna Radiators,” Ph.D. Dissertation, co-advisor, Worcester Polytechnic Institute, 2010.
- [2] Navalekar, Abhijit, “Distributed Digital Radios For Land Mobile Radio Applications,” Ph.D. Dissertation, primary advisor, Worcester Polytechnic Institute, 2009.
- [3] Parikh, Hemish, “Design of an OFDM Transmitter and Receiver for Precision Personnel Location,” primary advisor.
- [4] Proгри, Ilir, “An Assessment of Indoor Geolocation Systems,” Ph.D. Dissertation, primary advisor, Worcester Polytechnic Institute, 2003.
- [5] Li, Xinrong, “Super-Resolution TOA Estimation with Diversity Techniques for Indoor Applications,” Ph.D. Dissertation, co-advisor, Worcester Polytechnic Institute, 2003.
- [6] Leuenberger, Gerog H. W., “Electrostatic Density Measurements in Green-State PM Parts,” Ph.D. Dissertation, co-advisor, Worcester Polytechnic Institute, 2003.
- [7] Bogdanov, Gene, “Radio-Frequency Coil Design for High Field Magnetic Resonance Imaging,” Ph.D. Dissertation, co-advisor, Worcester Polytechnic Institute, 2002.
- [8] Elbirt, Adam J., “Reconfigurable Computing for Symmetric-Key Algorithms,” Ph.D. Dissertation, co-advisor, Worcester Polytechnic Institute, 2002.
- [9] Bretchko, Pavel, “Design and Development of Ultra-wideband DC-Coupled Amplifier,” Ph.D. Dissertation, co-advisor, Worcester Polytechnic Institute, 2001.
- [10] Hill, Jonathan, “Development of an Experimental Global Positioning System (GPS) Receiver Platform for Navigation Algorithm Evaluation,” Ph.D. Dissertation, primary advisor, 2001.
- [11] Spasojević, Mirko, “Creation of Sparse Boundary Element Matricies for 2-D and Axi-symmetric Electrostatic Problems Using a Bi-orthogonal Wavelet,” Ph.D. Dissertation, co-advisor, Worcester Polytechnic Institute, 1997.
- [12] Shi, Funan, “Optimal Designs of Gradient and RF Coils for Magnetic Resonance Imaging (MRI) Instrument,” Ph.D. Dissertation, co-advisor, Worcester Polytechnic Institute, 1996.

5. Proposals and Funding (past 5 years):**5.1 In Review**

- \$ 199,996 A National Model Robotics Curriculum, NSF (PI: Dr. M. Gennert, Co-PIs: Drs. T. Padir, W.R. Michalson, G. Fischer and C. Demetry), May 2009.
- \$ 199,052 A National Model Robotics Capstone, NSF (PI: Dr. W.R. Michalson, Co-PIs: Drs. T. Padir, C. Demetry, G. Tryggvason and F. Looft), May 2009.

\$ 399,791 Modular System for Teaching Robotics Engineering (MySTRE), NSF (PI: Dr. G. Fischer, Co-PIs: Drs. W.R. Michalson and T. Padir), March 2009.

5.2 Funding Received

\$ 50,524 PCGO Broadband Modem, Powerwave Technologies, Inc., (PI: Dr. W.R. Michalson), May 2009.

\$ 1,245,000 Real-Time Troop Status Monitoring System, US Army Telemedicine and Advanced Technology Research Center. (PI: Dr. Peder Pedersen, Co-PIs: Drs. William R. Michalson and Yitzhak Mendelson). Third year of funding. Projected funding period: Oct 1, 2004 to Sep 30, 2005.

\$ 148,422 Precision Personnel Locator System, National Institute of Justice (PI: Dr. John Orr, Co-PIs: Drs. David Cyganski and William R. Michalson). Second year of funding. Projected funding period: Sep 1, 2004 to Oct 31, 2005. Grant code 219240.

\$ 74,048 High-Speed VHF Modem, US Army Telemedicine and Advanced Technology Research Center (PI: Dr. William R. Michalson). Funding period: Mar 1, 2004 to Dec 31, 2004. Grant code 214370.

\$ 81,499 WPI Nanosat Program, Air Force Office of Scientific Research (PI: Dr. Fred Looft, Co-PIs: Drs. William R. Michalson and Diran Apelian). Funding period: Apr 1, 2003 to Mar 31, 2005. Grant code 214400.

\$ 996,000 Precision Personnel Locator System, National Institute of Justice (PI: Dr. David Cyganski, Co-PIs: Drs. William R. Michalson and John Orr). Second year of funding. Projected funding period: Sep 1, 2004 to Aug 31, 2005. Grant code 219240.

\$ 813,141 Real-Time Troop Status Monitoring System, US Army Telemedicine and Advanced Technology Research Center, (PI: Dr. William R. Michalson, Co-PIs: Drs. Peder Pedersen and Yitzhak Mendelson). Second year of funding. Funding period Oct 1, 2003 to Sep 30, 2004. Grant code 214370.

6. Honors, Awards, and Recognitions:

Elected Senior Member of the IEEE.

Joseph Samuel Satin Distinguished Fellowship awarded for the 1994-1995 academic year.

Aldo Miccioli Fellowship recipient from Raytheon Equipment Division.

ION Best Paper Award - GPS-96 for W. R. Michalson, W. Cidela, et. al., “A GPS-Based Hazard Detection and Warning System,” in review, " *ION GPS-96, 9th International Meeting of the Satellite Division of the Institute of Navigation*, pp. 167-175, Kansas City, MO, Sep 17-20, 1996

2nd Place - 2004 ECE Department MQP Award / Provost's MQP Award for *GPS-Based Orbit and Attitude Determination System for PANSAT*, Joshua Holwell, Andrew Coonradt and Himanshu Agrawal.

1st Place - 2003 ECE Department MQP Award / Provost's MQP Award for *Personal Inertial Navigation System*, Jason DeChiaro and Chris Struus.

1st Place - 2002 ECE Department MQP Award / Provost's MQP Award for *Handspring Digital Voltmeter*, Andrew Young and Pavel Loven.

3rd Place - 1998 ECE Department MQP Award / Provost's MQP Award for *Design of a Personal Handheld GPS Receiver*, Jeffery Alderson and Helder Machado.

2nd Place - 1997 ECE Department MQP Award / Provost's MQP Award for *Distributed Audio Controller*, EE-WZM-1B96, Stephen S. Richardson.

3rd Place - 1997 ECE Department MQP Award / Provost's MQP Award for *GPS Hazard Detector II*, EE-WZM-2A96, Will Brothers, Jon Day, and John Zaghi.

1st Place - 1996 ECE Department MQP Award / Provost's MQP Award for *GPS Hazard Detector*, EE-WHE-9601, Michael Roberts, William Cidela, and Chris Mangiarelli.

6.1 Memberships and offices held in professional society

Institute of Electrical and Electronic Engineers, Senior Member

Institute of Navigation

Royal Institute of Navigation

American Society of Engineering Educators

6.2 Professional Service

Massachusetts Board of Bar Overseers Hearing Committee Member, 2010-Present.

Steering Committee – 2009 First Annual Robotics Innovations Competition and Conference (RICC '09), Nov 7-8, Worcester, MA, 2009.

Conference Technical Co-Chair – 2009 IEEE International Conference on Technologies of Practical Robot Applications (TePRA 2009), Nov. 9-11, Woburn, MA, 2009.

Reviewer – Proposal number CRDPJ 379622-08, Natural Sciences and Engineering Research Council of Canada (NSERC), Mar. 2009

Co-Chair – Urban and Indoor Geolocation, Institute of Navigation International Technical Meeting (ITM2009), Anaheim CA, Jan. 2009.

Reviewer – Proposal number CRDPJ 379622-08, Natural Sciences and Engineering Research Council of Canada (NSERC), Mar. 2009

The Design and Analysis of Spatial Data Structures

The Design and Analysis of Spatial Data Structures

Hanan Samet

UNIVERSITY OF MARYLAND



ADDISON - WESLEY PUBLISHING COMPANY, INC.
Reading, Massachusetts • Menlo Park, California • New York
Don Mills, Ontario • Wokingham, England • Amsterdam
Bonn • Sydney • Singapore • Tokyo • Madrid • San Juan

This book is in the Addison-Wesley Series in Computer Science
 Michael A. Harrison: Consulting Editor

Many of the designations used by manufacturers and sellers to distinguish their products are claimed as trademarks. Where those designations appear in this book, and Addison-Wesley was aware of a trademark claim, the designations have been printed in initial caps or all caps.

The programs and applications presented in this book have been included for their instructional value. They have been tested with care, but are not guaranteed for any particular purpose. The publisher does not offer any warranties or representations, nor does it accept any liabilities with respect to the programs or applications.

Library of Congress Cataloging-in-Publication Data

Samet, Hanan.

The Design and analysis of spatial data structures/by Hanan Samet.

p. cm.

Bibliography: p.

Includes index.

ISBN 0-201-50255-0

1. Data structures (Computer science) 2. Computer graphics.

I. Title.

QA76.9.D35S26 1989

89-30382

005.73—dc19

CIP

Reprinted with corrections January, 1994

Copyright © 1990 by Addison-Wesley Publishing Company, Inc.

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording, or otherwise, without the prior written permission of the publisher. Printed in the United States of America. Published simultaneously in Canada.

4 5 6 7 8 9 10 11 12 13 14-MA-97 96 95 94

Credits:

Thor Bestul created the cover art.

Gyuri Fekete generated Figure 1.16; Daniel DeMenthon, Figures 1.20, 1.21, and 1.23; Jiang-Hsing Chu, Figures 2.48 and 2.52; and Walid Aref, Figures 4.38 through 4.40.

Figures 1.1, 4.9, and 4.10 are from H. Samet and R. E. Webber, On encoding boundaries with quad-trees, *IEEE Transactions on Pattern Analysis and Machine Intelligence* 6, 3 (May 1984), 365-369. © 1984 IEEE. Reprinted by permission of IEEE.

Figures 1.2, 1.3, 1.5 through 1.10, 1.12, 1.14, 1.25, 1.26, 2.3, 2.4, 2.18, 2.20, 2.30, 2.32, 2.53, 2.54, 2.57, 2.58, 3.20, 3.21, 4.1 through 4.5, 4.7, 4.8, 4.11, and 5.2 are from H. Samet, The quadtree and related hierarchical data structures, *ACM Computing Surveys* 16, 2 (June 1984), 187-260. Reprinted by permission of ACM.

Figures 1.4 and 5.6 are from H. Samet and R. E. Webber, Hierarchical data structures and algorithms for computer graphics. Part I. Fundamentals, *IEEE Computer Graphics and Applications* 8, 3 (May 1988), 48-68. © 1988 IEEE. Reprinted by permission of IEEE.

Figure 1.30 is from M. Li, W. I. Grosky, and R. Jain, Normalized quadtrees with respect to translations, *Computer Graphics and Image Processing* 20, 1 (September 1982), 72-81. Reprinted by permission of Academic Press.

Figures 2.7 and 2.10 through 2.15 are from H. Samet, Deletion in two-dimensional quad trees, *Communications of the ACM* 23, 12 (December 1980), 703-710. Reprinted by permission of ACM.

Figures 2.26 and 2.27 are from D. T. Lee and C. K. Wong, Worst-case analysis for region and partial region searches in multidimensional binary search trees and quad trees, *Acta Informatica* 9, 1 (1977), 23-29. Reprinted by permission of Springer Verlag.

Continued on p. 493

To my parents, Julius and Lotte

PREFACE

Spatial data consist of points, lines, rectangles, regions, surfaces, and volumes. The representation of such data is becoming increasingly important in applications in computer graphics, computer vision, database management systems, computer-aided design, solid modeling, robotics, geographic information systems (GIS), image processing, computational geometry, pattern recognition, and other areas. Once an application has been specified, it is common for the spatial data types to be more precise. For example, consider a geographic information system (GIS). In such a case, line data are differentiated on the basis of whether the lines are isolated (e.g., earthquake faults), elements of tree-like structures (e.g., rivers and their tributaries), or elements of networks (e.g., rail and highway systems). Similarly region data are often in the form of polygons that are isolated (e.g., lakes), adjacent (e.g., nations), or nested (e.g., contours). Clearly the variations are large.

Many of the data structures currently used to represent spatial data are hierarchical. They are based on the principle of recursive decomposition (similar to *divide and conquer* methods [Aho74]). One such data structure is the quadtree (octree in three dimensions). As we shall see, the term *quadtree* has taken on a generic meaning. In this book, it is my goal to show how a number of hierarchical data structures used in different domains are related to each other and to quadtrees. My presentation concentrates on these different representations and illustrates how a number of basic operations that use them are performed.

Hierarchical data structures are useful because of their ability to focus on the interesting subsets of the data. This focusing results in an efficient representation and in improved execution times. Thus they are particularly convenient for performing set operations. Many of the operations described can often be performed as efficiently, or more so, with other data structures. Nevertheless hierarchical data structures are attractive because of their conceptual clarity and ease of implementation. In addition, the use of some of them provides a spatial index. This is very useful in applications involving spatial databases.

As an example of the type of problems to which the techniques described in this book are applicable, consider a cartographic database consisting of a number of maps and some typical queries. The database contains a contour map, say at 50-foot elevation intervals, and a land use map classifying areas according to crop growth. Our goal is to determine all regions between 400- and 600-foot elevation levels where wheat is grown. This will require an intersection operation on the two maps. Such an analysis could be rather costly, depending on the way the maps are represented. For example, since areas where corn is grown are of no interest, we wish to spend a minimal amount of effort searching such regions. Yet traditional region representations such as the boundary code [Free74] are very local in application, making it difficult to avoid examining a corn-growing area that meets the desired elevation criterion. In contrast, hierarchical representations such as the region quadtree are more global in nature and enable the elimination of larger areas from consideration.

Another query might be to determine whether two roads intersect within a given area. We could check them point by point; however, a more efficient method of analysis would be to represent them by a hierarchical sequence of enclosing rectangles and to discover whether in fact the rectangles do overlap. If they do not, the search is terminated. If an intersection is possible, more work may have to be done, depending on which method of representation is used.

A similar query can be constructed for point data—for example, to determine all cities within 50 miles of St. Louis that have a population in excess of 20,000. Again we could check each city individually. However, using a representation that decomposes the United States into square areas having sides of length 100 miles would mean that at most four squares need to be examined. Thus California and its adjacent states can be safely ignored.

Finally, suppose we wish to integrate our queries over a database containing many different types of data (e.g., points, lines, areas). A typical query might be, “Find all cities with a population in excess of 5,000 people in wheat-growing regions within 20 miles of the Mississippi River.” In this book we will present a number of different ways of representing data so that such queries and other operations can be efficiently processed.

This book is organized as follows. There is one chapter for each spatial data type, in which I present a number of different data structures. The aim is to gain the ability to evaluate them and to determine their applicability. Two problems are treated in great detail: the rectangle intersection problem, discussed in the context of the representation of collections of small rectangles (Chapter 3), and the point location problem, discussed in the context of the representation of curvilinear data (Chapter 4). A comprehensive treatment of the use of quadtrees and octrees in other applications in computer graphics, image processing, and geographic information systems (GIS) can be found in [Same90b].

Chapter 1 gives a general introduction to the principle of recursive decomposition with a concentration on two-dimensional regions. Key properties, as well as a historical overview, are presented.

Chapter 2 discusses hierarchical representations of multidimensional point data. These data structures are particularly useful in applications in database management systems because they are designed to facilitate responses to search queries.

Chapter 3 examines the hierarchical representation of collections of small rectangles. Such data arise in applications in computational geometry, very large-scale integrations (VLSI), cartography, and database management. Examples from these fields (e.g., the rectangle intersection problem) are used to illustrate their differences. Many of the representations are closely related to those used for point data. This chapter is an expansion of [Same88a].

Chapter 4 treats the hierarchical representation of curvilinear data. The primary focus is on the representation of polygonal maps. The goal is to be able to solve the point location problem. Quadtree-like solutions are compared with those from computational geometry such as the K-structure [Kirk83] and the layered dag [Edel86a].

Chapter 5 looks at the representation of three-dimensional region data. In this case, a number of octree variants are examined, as well as constructive solid geometry (CSG) and the boundary model (BRep). Algorithms are discussed for converting between some of these representations. The representation of surfaces (i.e., 2.5-dimensional data) is also briefly discussed in this chapter.

There are a number of topics for which justice requires a considerably more detailed treatment. However, due to space limitations, I have omitted a detailed discussion of them and instead refer interested readers to the appropriate literature. For example, surface representations are discussed briefly with three-dimensional data in Chapter 5 (also see Chapter 7 of [Same90b]). The notion of a pyramid is presented only at a cursory level in Chapter 1 so that it can be contrasted with the quadtree. In particular, the pyramid is a multiresolution representation, whereas the quadtree is a variable resolution representation. Readers are referred to Tanimoto and Klinger [Tani80] and the collection of papers edited by Rosenfeld [Rose83a] for a more comprehensive exposition on pyramids.

Results from computational geometry, although related to many of the topics covered in this book, are discussed only in the context of representations for collections of small rectangles (Chapter 3) and curvilinear data (Chapter 4). For more details on early work involving some of these and related topics, interested readers should consult the surveys by Bentley and Friedman [Bent79b], Overmars [Over88a], Edelsbrunner [Edel84], Nagy and Wagle [Nagy79], Peuquet [Peuq84], Requicha [Requ80], Srihari [Srih81], Samet and Rosenfeld [Same80d], Samet [Same84b, Same88a], Samet and Webber [Same88c, Same88d], and Toussaint [Tous80].

There are also a number of excellent texts containing material related to the topics that I cover. Rosenfeld and Kak [Rose82a] should be consulted for an encyclopedic treatment of image processing. Mäntylä [Mänt87] has written a comprehensive introduction to solid modeling. Burrough [Burr86] provides a survey of geographic information systems (GIS). Overmars [Over83] has produced a particularly good treatment of multidimensional point data. In a similar vein, see Mehlhorn's [Mehl84] unified treatment of multidimensional searching and computational geometry. For thorough introductions to computational geometry, see Preparata and

k-structure and the layered dag in Section 4.3 are relevant to computational geometry. Bucket methods such as linear hashing, spiral hashing, grid file, and EXCELL, in Section 2.8, and R-trees in Section 3.5.3 are important in the study of database management systems. Methods for multidimensional searching that are discussed include k-d trees in Section 2.4, range trees and priority search trees in Section 2.5, and point-based rectangle representations in Section 3.4. The discussions of the representation of two-dimensional regions in Chapter 1, polygonal representations in Chapter 4, and use of point methods for focussing the Hough Transform are relevant to image processing. Finally the rectangle-representation methods and plane-sweep methods of Chapter 3 are important in the field of VLSI design.

The natural home for courses that use this book is in a computer science department, but the book could also be used in a curriculum in geographic information systems (GIS). Such a course is offered in geography departments. The emphasis for a course in this area would be on the use of quadtree-like methods for representing spatial data.

Shamos [Prep85] and Edelsbrunner [Edel87] (also see [Prep83, ORou88]). A broader view of the literature can be found in related bibliographies such as the ongoing collective effort coordinated by Edelsbrunner [Edel83c, Edel88], and Rosenfeld's annual collection of references in the journal *Computer Vision, Graphics, and Image Processing* (e.g., [Rose88]).

Nevertheless, given the broad and rapidly expanding nature of the field, I am bound to have omitted significant concepts and references. In addition at times I devote a disproportionate amount of attention to some concepts at the expense of others. This is principally for expository purposes; I feel that it is better to understand some structures well rather than to give readers a quick runthrough of buzzwords. For these indiscretions, I beg your pardon and hope you nevertheless bear with me.

My approach is an algorithmic one. Whenever possible, I have tried to motivate critical steps in the algorithms by a liberal use of examples. I feel that it is of paramount importance for readers to see the ease with which the representations can be implemented and used. In each chapter, except for the introduction (Chapter 1), I give at least one detailed algorithm using pseudo-code so that readers can see how the ideas can be applied. The pseudo-code is a variant of the ALGOL [Naur60] programming language that has a data structuring facility incorporating pointers and record structures. Recursion is used heavily. This language has similarities to C [Kern78], PASCAL [Jens74], SAIL [Reis76], and ALGOL W [Baue68]. Its basic features are described in the Appendix. However, the actual code is not crucial to understanding the techniques, and it may be skipped on a first reading. The index indicates the page numbers where the code for each algorithm is found.

In many cases I also give an analysis of the space and time requirements of different data structures and algorithms. The analysis is usually of an asymptotic nature and is in terms of *big O* and Ω notation [Knut76]. The *big O* notation denotes an upper bound. For example, if an algorithm takes $O(\log_2 N)$ time, then its worst-case behavior is never any worse than $\log_2 N$. The Ω notation denotes a lower bound. As an example of its use, consider the problem of sorting N numbers. When we say that sorting is $\Omega(N \cdot \log_2 N)$ we mean that given any algorithm for sorting, there is some set of N input values for which the algorithm will require at least this much time.

At times I also describe implementations of some of the data structures for the purpose of comparison. In such cases counts, such as the number of fields in a record, are often given. These numbers are meant only to amplify the discussion. They are not to be taken literally, as improvements are always possible once a specific application is analyzed more carefully.

Each chapter contains a substantial number of exercises. Many of the exercises develop further the material in the text as a means of testing the reader's understanding, as well as suggesting future directions. When the exercise or its solution is not my own, I have preceded it with the name of its originator. The exercises have not been graded by difficulty. They rarely require any mathematical skills beyond the undergraduate level for their solution. However, while some of the exercises are quite straightforward, others require some ingenuity. Solutions, or references to papers that

contain the solution, are provided for a substantial number of the exercises that do not require programming. Readers are cautioned to try to solve the exercises before turning to the solutions. It is my belief that much can be learned this way (for the student and, even more so, for the author). The motivation for undertaking this task was my wonderful experience on my first encounter with the rich work on data structures by Knuth [Knut73a, Knut73b].

An extensive bibliography is provided. It contains entries for both this book and the companion text [Same90b]. Not all of the references that appear in the bibliography are cited in the two texts. They are retained for the purpose of giving readers the ability to access the entire body of literature relevant to the topics discussed in them. Each reference is annotated with a key word(s) and a list of the numbers of the sections in which it is cited in either of the texts (including exercises and solutions). In addition, a name and credit index is provided that indicates the page numbers in this book on which each author's work is cited or a credit is made.

ACKNOWLEDGMENTS

Over the years I have received help from many people, and I am extremely grateful to them. In particular Robert E. Webber, Markku Tamminen, and Michael B. Dillencourt have generously given me much of their time and have gone over critical parts of the book. I have drawn heavily on their knowledge of some of the topics covered here. I have also been extremely fortunate to work with Azriel Rosenfeld over the past ten years. His dedication and scholarship have been a true inspiration to me. I deeply cherish our association.

I was introduced to the field of spatial data structures by Gary D. Knott who asked "how to delete in point quadtrees." Azriel Rosenfeld and Charles R. Dyer provided much interaction in the initial phase of my research. Those discussions led to the discovery of the neighbor-finding principle. It is during that time that many of the basic conversion algorithms between quadtrees and other image representations were developed as well. I learned much about image processing and computer vision from them. Robert E. Webber taught me computer graphics, Markku Tamminen taught me solid modeling and representations for multiattribute data, and Michael B. Dillencourt taught me about computational geometry.

During the time that this book was written, my research was supported, in part, by the National Science Foundation, the Defense Mapping Agency, the Harry Diamond Laboratory, and the Bureau of the Census. In particular I would like to thank Richard Antony, Y. T. Chien, Su-shing Chen, Hank Cook, Phil Emmerman, Joe Rastatter, Alan Saalfeld, and Larry Tokarcik. I am appreciative of their support.

Many people helped me in the process of preparing the book for publication. Acknowledgments are due to Rene McDonald for coordinating the day-to-day matters

of getting the book out and copyediting; to Scott Carson, Emery Jou, and Jim Purtilo for TROFF assistance beyond the call of duty; to Marisa Antoy and Sergio Antoy for designing and implementing the algorithm formatter used to typeset the algorithms; to Barbara Burnett, Michael B. Dillencourt, and Sandra German for help with the index; to Jay Weber for setting up the TROFF macrofiles so that I can keep track of symbolic names and thus be able to move text around without worrying about the numbering of exercises, sections, and chapters; to Liz Allen for early TROFF help; to Nono Kusuma, Mark Stanley, and Joan Wright Hamilton for drawing the figures; to Richard Muntz and Gerald Estrin for providing temporary office space and computer access at UCLA; to Sandy German, Gwen Nelson, and Janet Salzman for help in initial typing of the manuscript; to S. S. Iyengar, Duane Marble, George Nagy, and Terry Smith who reviewed the book; and to Peter Gordon, John Remington, and Keith Wollman at Addison-Wesley Publishing Company for their encouragement and confidence in this project.

Aside from the individuals named above, I have also benefited from discussions with many other people over the past years. They have commented on various parts of the book and include Chuan-Heng Ang, Walid Aref, James Arvo, Harvey H. Atkinson, Thor Bestul, Sharat Chandran, Chiun-Hong Chien, Jiang-Hsing Chu, Leila De Florian, Roger Eastman, Herbert Edelsbrunner, Claudio Esperanca, Christos Faloutsos, George (Gyuri) Fekete, Kikuo Fujimura, John Gannon, John Goldak, Erik Hoel, Liuqing Huang, Frederik W. Jansen, Ajay Kela, David Kirk, Per Åke Larson, Dani Lischinski, Don Meagher, David Mount, Randal C. Nelson, Glenn Pearson, Ron Sacks-Davis, Timos Sellis, Clifford A. Shaffer, Deepak Sherlekar, Li Tong, Brian Von Herzen, Peter Widmayer, and David Wise. I deeply appreciate their help.

A GUIDE TO THE INSTRUCTOR

This book can be used in a second data structures course, one with emphasis on the representation of spatial data. The focus is on the use of the principle of divide-and-conquer for which hierarchical data structures provide a good demonstration. Throughout the book both worst-case optimal methods and methods that work well in practice are emphasized in conformance with my view that the well-rounded computer scientist should be conversant with both types of algorithms. This material is more than can be covered in one semester; but the instructor can reduce it as necessary. For example, the detailed examples can be skipped or used as a basis of a term project or programming assignments.

The book can also be used to organize a course to be prerequisite to courses in computer graphics and solid modeling, computational geometry, database management systems, multidimensional searching, image processing, and VLSI design. The discussions of the representations of two-dimensional regions in Chapter 1, polygonal representations in Chapter 4, and most of Chapter 5 are relevant to computer graphics and solid modeling. The discussions of plane-sweep methods and their associated data structures such as segment trees, interval trees, and priority search trees in Sections 3.2 and 3.3 and point location and associated data structures such as the

CONTENTS

Preface	vii
1 INTRODUCTION	1
1.1 Basic Definitions	1
1.2 Overview of Quadtrees and Octrees	2
1.3 History of the Use of Quadtrees and Octrees	10
1.4 Space Decomposition Methods	16
1.4.1 Polygonal Tilings	17
1.4.2 Nonpolygonal Tilings	26
1.5 Space Requirements	32
2 POINT DATA	43
2.1 Introduction	44
2.2 Nonhierarchical Data Structures	46
2.3 Point Quadtrees	48
2.3.1 Insertion	49
2.3.2 Deletion	54
2.3.3 Search	64
2.4 k-d Trees	66
2.4.1 Insertion	68
2.4.2 Deletion	73
2.4.3 Search	77
2.4.4 Comparison with Point Quadtrees	80
2.5 Range Trees and Priority Search Trees	80
2.6 Region-based Quadtrees	85
2.6.1 MX Quadtrees	86
2.6.2 PR Quadtrees	92

2.6.3	Comparison of Point and Region-based Quadtrees	104
2.7	Bit Interleaving	105
2.8	Bucket Methods	110
2.8.1	Hierarchical Bucket Methods	111
2.8.2	Nonhierarchical Bucket Methods	116
2.8.2.1	Linear Hashing	117
2.8.2.2	Spiral Hashing	125
2.8.2.3	Grid File	135
2.8.2.4	EXCELL	141
2.9	Conclusion	147
3	COLLECTIONS OF SMALL RECTANGLES	153
3.1	Introduction	155
3.2	Plane-Sweep Methods and the Rectangle Intersection Problem	158
3.2.1	Segment Trees	160
3.2.2	Interval Trees	165
3.2.3	Priority Search Trees	171
3.2.4	Alternative Solutions and Related Problems	174
3.3	Plane-Sweep Methods and the Measure Problem	178
3.4	Point-based Methods	186
3.5	Area-based Methods	199
3.5.1	MX-CIF Quadtrees	200
3.5.1.1	Insertion	202
3.5.1.2	Deletion	206
3.5.1.3	Search	209
3.5.2	Multiple Quadtree Block Representations	213
3.5.3	R-trees	219
4	CURVILINEAR DATA	227
4.1	Strip Trees, Arc Trees, and BSPR	228
4.2	Methods Based on the Region Quadtree	235
4.2.1	Edge Quadtrees	235
4.2.2	Line Quadtrees	237
4.2.3	PM Quadtrees	239
4.2.3.1	The PM_1 Quadtree	240
4.2.3.2	The PM_2 Quadtree	257
4.2.3.3	The PM_3 Quadtree	261
4.2.3.4	PMR Quadtrees	264
4.2.3.5	Fragments	269
4.2.3.6	Maintaining Labels of Regions	275
4.2.4	Empirical Comparisons of the Different Representations	278
4.3	Methods Rooted in Computational Geometry	286
4.3.1	The K-structure	287

4.3.2	Separating Chains and Layered Dags	293
4.3.3	Comparison with PM Quadrees	306
4.4	Conclusion	312
5	VOLUME DATA	315
5.1	Solid Modeling	316
5.2	Region Octrees	318
5.3	PM Octrees	326
5.4	Boundary Model (BRep)	331
5.5	Constructive Solid Geometry (CSG)	338
5.5.1	CSG Evaluation by Bintree Conversion	340
5.5.1.1	Algorithm for a Single Halfspace	341
5.5.1.2	Algorithm for a CSG Tree	346
5.5.1.3	Incorporation of the Time Dimension	355
5.5.2	PM-CSG Trees	360
5.6	Surface-based Object Representations	365
5.7	Prism Trees	370
5.8	Cone Trees	374
	Solutions to Exercises	377
	Appendix: Description of Pseudo-Code Language	411
	References	415
	Name and Credit Index	465
	Subject Index	477

INTRODUCTION

1

There are numerous hierarchical data structuring techniques in use for representing spatial data. One commonly used technique is the quadtree, which has evolved from work in different fields. Thus it is natural that a number of adaptations of it exist for each spatial data type. Its development has been motivated to a large extent by a desire to save storage by aggregating data having identical or similar values. We will see, however, that this is not always the case. In fact, the savings in execution time that arise from this aggregation are often of equal or greater importance.

In this chapter we start with a historical overview of quadtrees, including definitions. Since the primary focus in this book is on the representation of regions, what follows is a discussion of region representation in the context of different space decomposition methods. This is done by examining polygonal and nonpolygonal tilings of the plane. The emphasis is on justifying the use of a decomposition into squares. We conclude with a detailed analysis of the space requirements of the quadtree representation.

Most of the presentation in this chapter is in the context of two-dimensional regions. The extension of the topics in this chapter, and remaining chapters, to three-dimensional region data, and higher, is straightforward and, aside from definitions, is often left to the exercises. Nevertheless, the concept of an octree, a quadtree-like representation of three-dimensional regions, is defined and a brief explanation is given of how some of the results described here are applicable to higher-dimensional data.

1.1 BASIC DEFINITIONS

First, we define a few terms with respect to two-dimensional data. Assume the existence of an array of picture elements (termed *pixels*) in two dimensions. We use the term *image* to refer to the original array of pixels. If its elements are black or

white, then it is said to be *binary*. If shades of gray are possible (i.e., gray levels), the image is said to be a *gray-scale* image. In the discussion, we are primarily concerned with binary images. Assume that the image is on an infinite background of white pixels. The *border* of the image is the outer boundary of the square corresponding to the array.

Two pixels are said to be *4-adjacent* if they are adjacent to each other in the horizontal or vertical direction. If the concept of adjacency also includes adjacency at a corner (i.e., diagonal adjacencies), then the pixels are said to be *8-adjacent*. A set S is said to be *four-connected* (*eight-connected*) if for any pixels p, q in S there exists a sequence of pixels $p = p_0, p_1, \dots, p_n = q$ in S , such that p_{i+1} is 4-adjacent (8-adjacent) to p_i , $0 \leq i < n$.

A black *region*, or black four-connected *component*, is a maximal four-connected set of black pixels. The process of assigning the same label to all 4-adjacent black pixels is called *connected component labeling* (see Chapter 5 of [Same90b]). A white *region* is a maximal *eight-connected* set of white pixels defined analogously. The complement of a black region consists of a union of eight-connected white regions. Exactly one of these white regions contains the infinite background of white pixels. All the other white regions, if any, are called *holes* in the black region. The black region, say R , is surrounded by the infinite white region and R surrounds the other white regions, if any.

A pixel is said to have four edges, each of which is of unit length. The *boundary* of a black region consists of the set of edges of its constituent pixels that also serve as edges of white pixels. Similar definitions can be formulated in terms of rectangular blocks, all of whose pixels are identically colored. For example, two disjoint blocks, P and Q , are said to be *4-adjacent* if there exists a pixel p in P and a pixel q in Q such that p and q are 4-adjacent. Eight-adjacency for blocks (as well as connected component labeling) is defined analogously.

1.2 OVERVIEW OF QUADTREES AND OCTREES

The term *quadtree* is used to describe a class of hierarchical data structures whose common property is that they are based on the principle of recursive decomposition of space. They can be differentiated on the following bases:

1. The type of data they are used to represent
2. The principle guiding the decomposition process
3. The resolution (variable or not)

Currently they are used for point data, areas, curves, surfaces, and volumes. The decomposition may be into equal parts on each level (i.e., regular polygons and termed a *regular decomposition*), or it may be governed by the input. In computer graphics this distinction is often phrased in terms of image-space hierarchies versus object-space hierarchies, respectively [Suth74]. The resolution of the decomposition

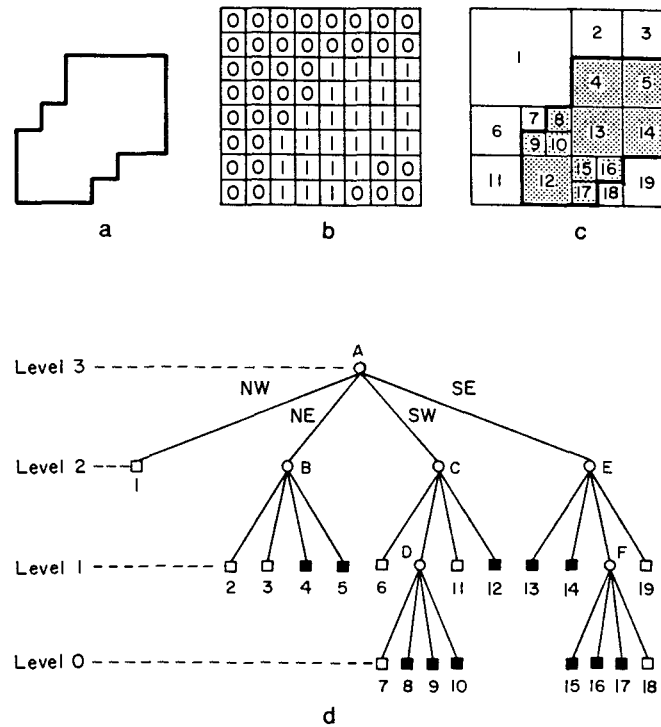


Figure 1.1 An example of (a) a region, (b) its binary array, (c) its maximal blocks (blocks in the region are shaded), and (d) the corresponding quadtree

(i.e., the number of times that the decomposition process is applied) may be fixed beforehand, or it may be governed by properties of the input data. For some applications we can also differentiate the data structures on the basis of whether they specify the boundaries of regions (e.g., curves and surfaces) or organize their interiors (e.g., areas and volumes).

The first example of a quadtree representation of data is concerned with the representation of two-dimensional binary region data. The most studied quadtree approach to region representation, called a *region quadtree* (but often termed a *quadtree* in the rest of this chapter), is based on the successive subdivision of a bounded image array into four equal-sized quadrants. If the array does not consist entirely of 1s or entirely of 0s (i.e., the region does not cover the entire array), then it is subdivided into quadrants, subquadrants, and so on, until blocks are obtained that consist entirely of 1s or entirely of 0s; that is, each block is entirely contained in the region or entirely disjoint from it. The region quadtree can be characterized as a variable resolution data structure.

As an example of the region quadtree, consider the region shown in Figure 1.1a represented by the $2^3 \times 2^3$ binary array in Figure 1.1b. Observe that the 1s correspond to picture elements (i.e., pixels) in the region, and the 0s correspond to picture elements outside the region. The resulting blocks for the array of Figure 1.1b are shown in Figure 1.1c. This process is represented by a tree of degree 4 (i.e., each nonleaf node has four sons).

In the tree representation, the root node corresponds to the entire array. Each son of a node represents a quadrant (labeled in order NW, NE, SW, SE) of the region represented by that node. The leaf nodes of the tree correspond to those blocks for which no further subdivision is necessary. A leaf node is said to be black or white depending on whether its corresponding block is entirely inside (it contains only 1s) or entirely outside the represented region (it contains no 1s). All nonleaf nodes are said to be gray (i.e., its block contains 0s and 1s). Given a $2^n \times 2^n$ image, the root node is said to be at level n while a node at level 0 corresponds to a single pixel in the image.¹ The region quadtree representation for Figure 1.1c is shown in Figure 1.1d. The leaf nodes are labeled with numbers, while the nonleaf nodes are labeled with letters. The levels of the tree are also marked.

Our definition of the region quadtree implies that it is constructed by a top-down process. In practice, the process is bottom-up, and one usually uses one of two approaches. The first approach [Same80b] is applicable when the image array is not too large. In such a case, the elements of the array are inspected in the order given by the labels on the array in Figure 1.2 (which corresponds to the image of Figure 1.1a). This order is also known as a Morton order [Mort66] (discussed in Section 1.3). By using such a method, a leaf node is never created until it is known to be maximal. An equivalent statement is that the situation does not arise in which four leaf nodes of the same color necessitate the changing of the color of their parent from gray to black or white as is appropriate. (For more details, see Section 4.1 of [Same90b].)

The second approach [Same81a] is applicable to large images. In this case, the elements of the image are processed one row at a time—for example, in the order given by the labels on the array in Figure 1.3 (which corresponds to the image of Figure 1.1a). This order is also known as a row or raster-scan order (discussed in Section 1.3). A quadtree is built by adding pixel-sized nodes one by one in the order in which they appear in the file. (For more details, see Section 4.2.1 of [Same90b].) This process can be time-consuming due to the many merging and node insertion operations that need to take place.

The above method has been improved by using a predictive method [Shaf86a, Shaf87a], which only makes a single insertion for each node in the final quadtree and performs no merge operations. It is based on processing the image in row order (top to bottom, left to right), always inserting the largest node (i.e., block) for which the current pixel is the first (upper leftmost) pixel. Such a policy avoids the necessity of merging since the upper leftmost pixel of any block is inserted before any other pixel of that block. Therefore it is impossible for four sibling nodes to be of the same color. This method makes use of an auxiliary array of size $O(2^n)$ for a $2^n \times 2^n$ image. (For more details, see Section 4.2.3 of [Same90b].)

The region quadtree is easily extended to represent three-dimensional binary region data and the resulting data structure is called a *region octree* (termed an *octree*

¹ Alternatively we can say that the root node is at depth 0 while a node at depth n corresponds to a single pixel in the image. In this book both concepts of level and depth are used to describe the relative position of nodes. The one that is chosen is context dependent.

1	2	5	6	17	18	21	22
3	4	7	8	19	20	23	24
9	10	13	14	25	26	29	30
11	12	15	16	27	28	31	32
33	34	37	38	49	50	53	54
35	36	39	40	51	52	55	56
41	42	45	46	57	58	61	62
43	44	47	48	59	60	63	64

Figure 1.2 Morton order for the pixels of Figure 1.1

in the rest of this chapter). We start with a $2^n \times 2^n \times 2^n$ object array of unit cubes (termed *voxels* or *obels*). The octree is based on the successive subdivision of an object array into octants. If the array does not consist entirely of 1s or entirely of 0s, it is subdivided into octants, suboctants, and so on until cubes (possibly single voxels) are obtained that consist of 1s or of 0s; that is, they are entirely contained in the region or entirely disjoint from it.

This subdivision process is represented by a tree of degree 8 in which the root node represents the entire object and the leaf nodes correspond to those cubes of the array for which no further subdivision is necessary. Leaf nodes are said to be black or white (alternatively, full or void) depending on whether their corresponding cubes are entirely within or outside the object, respectively. All nonleaf nodes are said to be gray. Figure 1.4a is an example of a simple three-dimensional object, in the form of a staircase, whose octree block decomposition is given in Figure 1.4b and whose tree representation is given in Figure 1.4c.

The region quadtree is a member of a class of representations characterized as being a collection of maximal (according to an appropriate definition) blocks, each of which is contained in a given region and whose union is the entire region. The simplest such representation is the runlength code, where the blocks are restricted to $1 \times m$ rectangles [Ruto68]. A more general representation treats the region as a union of maximal square blocks (or blocks of any other desired shape) that may possibly overlap. Usually the blocks are specified by their centers and radii. This representation is called the *medial axis transformation (MAT)* [Blum67, Rose66]. Of course, other approaches are also possible (e.g., rectangular coding [Kim83, Kim86], TID [Scot85, Scot86]).

1	2	3	4	5	6	7	8
9	10	11	12	13	14	15	16
17	18	19	20	21	22	23	24
25	26	27	28	29	30	31	32
33	34	35	36	37	38	39	40
41	42	43	44	45	46	47	48
49	50	51	52	53	54	55	56
57	58	59	60	61	62	63	64

Figure 1.3 Raster-scan order for the pixels of Figure 1.1

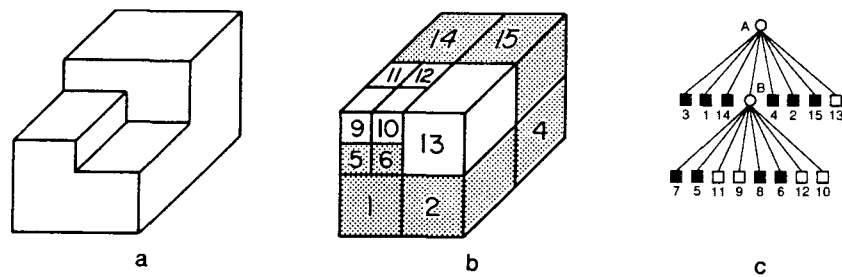


Figure 1.4 (a) Example three-dimensional object; (b) its octree block decomposition; (c) its tree representation

The region quadtree is a variant on the maximal block representation. It requires the blocks to be disjoint and to have standard sizes (i.e., sides of lengths that are powers of two) and standard locations. The motivation for its development is a desire to obtain a systematic way to represent homogeneous parts of an image. Thus to transform the data into a region quadtree, a criterion must be chosen for deciding that an image is homogeneous (i.e., uniform).

One such criterion is that the standard deviation of its gray levels is below a given threshold t . Using this criterion, the image array is successively subdivided into quadrants, subquadrants, and so on until homogeneous blocks are obtained. This process leads to a regular decomposition. If one associates with each leaf node the mean gray level of its block, the resulting region quadtree will then completely specify a piecewise approximation to the image where each homogeneous block is represented by its mean. The case where $t=0$ (i.e., a block is not homogeneous unless its gray level is constant) is of particular interest since it permits an exact reconstruction of the image from its quadtree.

Note that the blocks of the region quadtree do not necessarily correspond to maximal homogeneous regions in the image. Most likely there exist unions of the blocks that are still homogeneous. To obtain a segmentation of the image into maximal homogeneous regions, we must allow merging of adjacent blocks (or unions of blocks) as long as the resulting region remains homogeneous. This is achieved by a 'split-and-merge' algorithm [Horo76]. However, the resulting partition will no longer be represented by a quadtree; instead the final representation is in the form of an adjacency graph. Thus the region quadtree is used as an initial step in the segmentation process.

For example, Figure 1.5b–d demonstrates the results of the application, in sequence, of merging, splitting, and grouping to the initial image decomposition of Figure 1.5a. In this case, the image is initially decomposed into 16 equal-sized square blocks. Next the 'merge' step attempts to form larger blocks by recursively merging groups of four homogeneous 'brothers' (the four blocks in the NW and SE quadrants of Figure 1.5b). The 'split' step recursively decomposes blocks that are not homogeneous (the NE and SW quadrants of Figure 1.5c) until a particular homogeneity criterion is satisfied or a given level is encountered. Finally the 'grouping' step aggregates all homogeneous 4-adjacent black blocks into one region apiece;

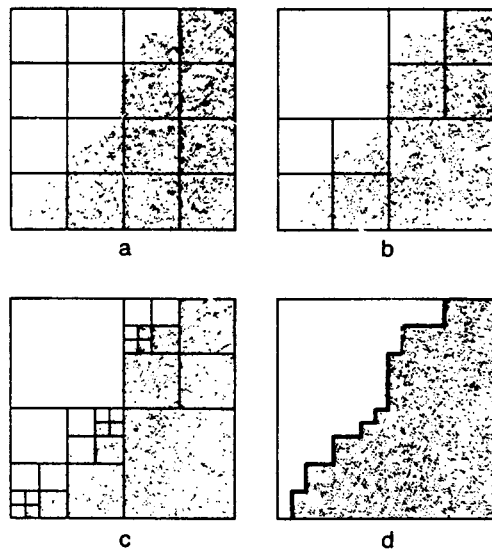


Figure 1.5 Example illustrating the 'split-and-merge' segmentation procedure: (a) start, (b) merge, (c) split, (d) grouping

the 8-adjacent white blocks are similarly aggregated into white regions (Figure 1.5d).

An alternative to the region quadtree representation is to use a decomposition method that is not regular (i.e., rectangles of arbitrary size rather than squares). This alternative has the potential of requiring less space. Its drawback is that the determination of optimal partition points may be computationally expensive (see Exercise 1.10). A closely related problem, decomposing a region into a minimum number of rectangles, is known to be NP-complete² [Gare79] if the region is permitted to contain holes [Ling82].

The homogeneity criterion ultimately chosen to guide the subdivision process depends on the type of region data represented. In the remainder of this chapter we shall assume that the domain is a $2^n \times 2^n$ binary image with 1, or black, corresponding to foreground and 0, or white, corresponding to background (e.g., Figure 1.1).

² A problem is in NP if it can be solved nondeterministically in polynomial time. A nondeterministic solution process proceeds by 'guessing' a solution and then verifying that the solution is correct. Assume that n is the size of the problem (e.g., for sorting, n is the number of records to be sorted). Intuitively, then, a problem is in NP if there is a polynomial $P(n)$ such that if one guesses a solution, it can be verified in $O(P(n))$ time, whether the guess is indeed a correct solution. Thus the verification process is the key to determining whether a problem is in NP, not the actual solution of the problem.

A problem is NP-complete if it is 'at least as hard' as any other problem in NP. Somewhat more formally, a problem P_1 in NP is NP-complete if the following property holds: for all other problems P_i in NP, if P_i can be solved deterministically in $O(f(n))$ time, then P_1 can be solved in $O(P(f(n)))$ time for some polynomial P . It has been conjectured that no NP-complete problem can be solved deterministically in polynomial time, but this is not known for sure. The theory of NP-completeness is discussed in detail in [Gare79].

Nevertheless the quadtree and octree can be used to represent multicolored data (e.g., a landuse class map associating colors with crops [Same87a]).

It is interesting to note that Kawaguchi, Endo, and Matsunaga [Kawa83] use a sequence of m binary-valued quadtrees to encode image data of 2^m gray levels, where the various gray levels are encoded by use of Gray codes (see, e.g., [McCl65]). This should lead to compaction (i.e., larger-sized blocks) since the Gray code guarantees that the binary representation of the codes of adjacent gray level values differ by only one binary digit.³ Note, though, that if the primary interest is in image compression, there exist even better methods (see, e.g., [Prat78]); however, they are beyond the scope of this book (but see Chapter 8 of [Same90b]). In another context, Kawaguchi, Endo, and Yokota [Kawa80b] point out that a sequence of related images (e.g., in an animation application) can be stored compactly as a sequence of quadtrees such that the i^{th} element is the result of exclusive oring the first i images (see Exercise 1.7).

Unfortunately the term *quadtree* has taken on more than one meaning. The region quadtree, as described earlier, is a partition of space into a set of squares whose sides are all a power of two long. This formulation is due to Klinger [Klin71] and Klinger and Dyer, who used the term *Q-tree* [Klin76], whereas Hunter [Hunt78] was the first to use the term *quadtree* in such a context. Actually a more precise term would be *quadtrie*, as it is really a trie structure [Fred60] in two dimensions.⁴ A similar partition of space into rectangular quadrants, also termed a *quadtree*, was used by Finkel and Bentley [Fink74]. It is an adaptation of the binary search tree [Knut73b] to two dimensions (which can be easily extended to an arbitrary number of dimensions). It is primarily used to represent multidimensional point data, and we shall refer to it as a *point quadtree* where confusion with a region quadtree is possible.

As an example of a point quadtree, consider Figure 1.6, which is built for the sequence Chicago, Mobile, Toronto, Buffalo, Denver, Omaha, Atlanta, and Miami⁵

³ The Gray code is motivated by a desire to reduce errors in transitions between successive gray level values. Its one bit difference guarantee is achieved by the following encoding. Consider the binary representation of the integers from 0 to $2^m - 1$. This representation can be obtained by constructing a binary tree, say T , of height m where each left branch is labeled 0 while each right branch is labeled 1. Each leaf node, say p , is given the label formed by concatenating the labels of the branches taken by the path from the root to p . Enumerating the leaf nodes from left to right yields the binary integers 0 to $2^m - 1$. The Gray codes of the integers are obtained by constructing a new binary tree, say T' , such that the labels of some of the branches in T' are the reverse of what they were in T . The algorithm is as follows. Initially, T' is a copy of T . Next, traverse T in preorder (i.e., visit the root node, followed by the left and right subtrees). For each branch in T labeled 1, exchange the labels of the two descendant branches of its corresponding branch in T' . No action is taken for descendants of branches in T labeled 0. Enumerating the leaf nodes in T' from left to right yields the Gray codes of the integers 0 to $2^m - 1$. For example, for 8 gray levels (i.e., $m = 3$), we have 000, 001, 011, 010, 110, 111, 101, 100.

⁴ In a one-dimensional *trie* structure, each data item or key is treated as a sequence of characters where each character has M possible values. A node at depth i in the trie represents an M -way branch depending on the i^{th} character. The data are stored in the leaf nodes, and the shape of the trie is independent of the order in which the data are processed. Such a structure is also known as a *digital tree* [Knut73b].

⁵ The correspondence between coordinate values and city names is not geographically correct. This liberty has been taken so that the same example can be used throughout the text to illustrate a variety of concepts.

1.2 OVERVIEW OF QUADTREES AND OCTREES || 9

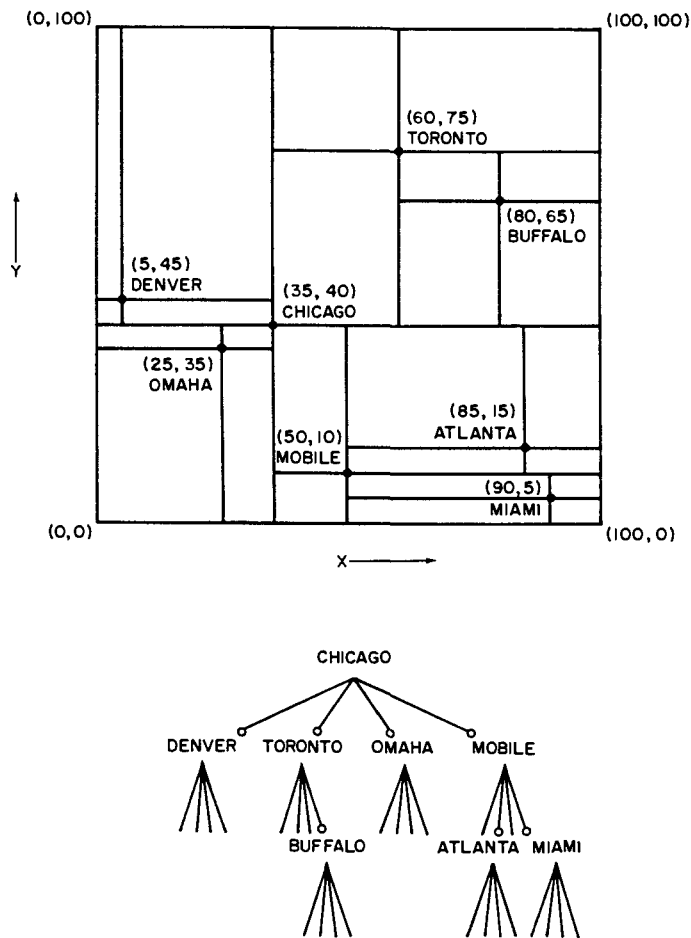


Figure 1.6 A point quadtree and the records it represents

in the order in which they are listed here.⁶ Its shape is highly dependent on the order in which the points are added to it. Of course, trie-based point representations also exist (see Sections 2.6.1 and 2.6.2).

Exercises

- 1.1. The region quadtree is an alternative to an image representation that is based on the use of an array or even a list. Each of these image representations may be biased in favor of the computation of a particular adjacency relation. Discuss these biases for the array, list, and quadtree representations.
- 1.2. Given the array representation of a binary image, write an algorithm to construct the corresponding region quadtree.

⁶ Refer to Figure 2.5 to see how the point quadtree is constructed in an incremental fashion for Chicago, Mobile, Toronto, and Buffalo.

- 1.3. Given an image represented by a region quadtree with B black and w white nodes, how many additional nodes are necessary for the nonleaf nodes?
- 1.4. Given an image represented by a region octree with B black and w white nodes, how many additional nodes are necessary for the nonleaf nodes?
- 1.5. Suppose that an octree is used to represent a collection of disjoint spheres. What would you use as a leaf criterion?
- 1.6. The quadtree can be generalized to represent data in arbitrary dimensions. As we saw, the octree is its three-dimensional analog. The renowned artist Escher [Coxe86] is noted for etchings of unusual interpretations of geometric objects such as staircases. How would you represent one of Escher's staircases?
- 1.7. Let \oplus denote an exclusive or operation. Given a sequence of related images, $\langle P_n, P_{n-1}, \dots, P_0 \rangle$, define another sequence $\langle Q_n, Q_{n-1}, \dots, Q_0 \rangle$ such that $Q_0 = P_0$ and $Q_i = P_i \oplus Q_{i-1}$ for $i > 0$. Show that when the sequences P and Q are represented as quadtrees, replacing sequence P by sequence Q results in fewer nodes.
- 1.8. Prove that in Exercise 1.7 the sequence P can be reconstructed from the sequence Q . In particular, given Q_i and Q_{i-1} , determine P_i .
- 1.9. Write an algorithm to construct the Gray codes of the integers 0 to $2^n - 1$.
- 1.10. Find a polynomial-time algorithm to decompose a region optimally so that its quadtree representation uses a minimum amount of space (i.e., a minimum number of nodes). In this case, you can assume that the decomposition lines can be placed in arbitrary positions so that the space requirement is reduced. In other words, the decomposition lines need not split the space into four squares of equal size. Thus the decomposition is similar to that induced by a point quadtree.

1.3 HISTORY OF THE USE OF QUADTREES AND OCTREES

The origin of the principle of recursive decomposition, upon which all quadtrees are based, is difficult to ascertain. Below, to give some indication of the uses of the region quadtree, some of its applications to geometric data are traced briefly. Most likely it was first seen as a way of aggregating blocks of zeros in sparse matrices. Indeed Hoare [Hoar72] attributes a one-level decomposition of a matrix into square blocks to Dijkstra. Morton [Mort66] used it as a means of indexing into a geographic database (i.e., it acts as a spatial index).

Warnock, in a pair of reports that serve as landmarks in computer graphics [Warn68, Warn69b], described the implementation of hidden-line and hidden-surface elimination algorithms using a recursive decomposition of the picture area. The picture area is repeatedly subdivided into rectangles that are successively smaller while searching for areas that are sufficiently simple to be displayed. Klinger [Klin71] and Klinger and Dyer [Klin76] applied these ideas to pattern recognition and image processing, while Hunter [Hunt78] used them for an animation application.

The SRI robot project [Nils69] used a three-level decomposition of space to represent a map of the robot's world. Eastman [East70] observes that recursive decomposition might be used for space planning in an architectural context and presents a simplified version of the SRI robot representation. A quadtree-like representation in the form of production rules called DF-expressions (denoting 'depth-first') is discussed by Kawaguchi and Endo [Kawa80a] and Kawaguchi, Endo, and Yokota

[Kawa80b] (see also Section 1.5). Tucker [Tuck84a] uses quadtree refinement as a control strategy for an expert vision system.

The three-dimensional variant of the region quadtree—the octree—was developed independently by a number of researchers. Hunter [Hunt78] mentioned it as a natural extension of the quadtree. Reddy and Rubin [Redd78] proposed the octree as one of three representations for solid objects. The second is a three-dimensional generalization of the point quadtree of Finkel and Bentley [Fink74]—that is, a decomposition into rectangular parallelepipeds (as opposed to cubes) with planes perpendicular to the x , y , and z axes. The third breaks the object into rectangular parallelepipeds that are not necessarily aligned with an axis. The parallelepipeds are of arbitrary sizes and orientations. Each parallelepiped is recursively subdivided into parallelepipeds in the coordinate space of the enclosing parallelepiped. Reddy and Rubin prefer the third approach for its ease of display.

Situated somewhere between the second and third approaches of Reddy and Rubin is the method of Brooks and Lozano-Perez [Broo83] (see also [Loza81]), who use a recursive decomposition of space into an arbitrary number of rectangular parallelepipeds, with planes perpendicular to the x , y , and z axes, to model space in solving the *findpath* or *piano movers* problem [Schw88] in robotics. This problem arises when planning the motion of a robot in an environment containing known obstacles and the desired solution is a collision-free path obtained by use of a search. Faverjon [Fave84] discusses an approach to this problem that uses an octree, as do Samet and Tamminen [Same85g] and Fujimura and Samet [Fuji89].

Jackins and Tanimoto [Jack80] adapted Hunter and Steiglitz's quadtree translation algorithm [Hunt78, Hunt79b] to objects represented by octrees. Meagher [Meag82a] developed numerous algorithms for performing solid modeling operations in an environment where the octree is the underlying representation. Yau and Srihari [Yau83] extended the octree to arbitrary dimensions in the process of developing algorithms to handle medical images.

Both quadtrees and octrees are frequently used in the construction of meshes for finite element analysis. The use of recursive decomposition for meshes was initially suggested by Rheinboldt and Mesztenyi [Rhei80]. Yerry and Shephard [Yerr83] adapted the quadtree and octree to generate meshes automatically for three-dimensional solids represented by a superquadric surface-based modeler. This has been extended by Kela, Voelcker, and Goldak [Kela84b] (see also [Kela86]) to mesh boundary regions directly, rather than through discrete approximations, and to facilitate incremental adaptive analysis by exploiting the spatial index nature of the quadtree and octree.

Parallel to the development of the quadtree and octree data structures, there has been related work by researchers in the field of image understanding. Kelly [Kell71] introduced the concept of a *plan*, which is a small picture whose pixels represent gray-scale averages over 8×8 blocks of a larger picture. Needless effort in edge detection is avoided by first determining edges in the plan and then using these edges to search selectively for edges in the larger picture. Generalizations of this idea motivated the development of multiresolution image representations—for example,

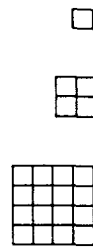


Figure 1.7 Structure of a pyramid having three levels

the recognition cone of Uhr [Uhr72], the preprocessing cone of Riseman and Arbib [Rise77], and the pyramid of Tanimoto and Pavlidis [Tani75]. Of these representations, the pyramid is the closest relative of the region quadtree.

Given a $2^n \times 2^n$ image array, say $A(n)$, a *pyramid* is a sequence of arrays $\{A(i)\}$ such that $A(i-1)$ is a version of $A(i)$ at half the scale of $A(i)$. $A(0)$ is a single pixel. Figure 1.7 shows the structure of a pyramid having three levels. It should be clear that a pyramid can also be defined in a more general way by permitting finer scales of resolution than the power of two scale.

At times, it is more convenient to define a pyramid in the form of a tree. Again, assuming a $2^n \times 2^n$ image, a recursive decomposition into quadrants is performed, just as in quadtree construction, except that we keep subdividing until we reach the individual pixels. The leaf nodes of the resulting tree represent the pixels, while the nodes immediately above the leaf nodes correspond to the array $A(n-1)$, which is of size $2^{n-1} \times 2^{n-1}$. The nonleaf nodes are assigned a value that is a function of the nodes below them (i.e., their sons) such as the average gray level. Thus we see that a pyramid is a multiresolution representation, whereas the region quadtree is a variable

1	2	3	4	5	6	7	8
9	10	11	12	13	14	15	16
17	18	19	20	21	22	23	24
25	26	27	28	29	30	31	32
33	34	35	36	37	38	39	40
41	42	43	44	45	46	47	48
49	50	51	52	53	54	55	56
57	58	59	60	61	62	63	64

Figure 1.8 Example pyramid $A(3)$

A	B	C	D
E	F	G	H
I	J	K	L
M	N	O	P

Figure 1.9 $A(2)$ corresponding to Figure 1.8

1	2	3	4	5	6	7	8
9	10	11	12	13	14	15	16
17	18	19	20	21	22	23	24
25	26	27	28	29	30	31	32
33	34	35	36	37	38	39	40
41	42	43	44	45	46	47	48
49	50	51	52	53	54	55	56
57	58	59	60	61	62	63	64

Figure 1.10 The overlapping blocks in which pixel 28 participates

resolution representation. Another analogy is that the pyramid is a complete quadtree [Knut73a].

The above definition of a pyramid is based on nonoverlapping 2×2 blocks of pixels. An alternative definition, termed an *overlapping pyramid*, uses overlapping blocks of pixels. One of the simplest schemes makes use of 4×4 blocks that overlap by 50% in both the horizontal and vertical directions [Burt81]. For example, Figure 1.8 is a $2^3 \times 2^3$ array, say $A(3)$, whose pixels are labeled 1-64. Figure 1.9 is $A(2)$ corresponding to Figure 1.8 with elements labeled A-P. The 4×4 neighborhood corresponding to element F in Figure 1.9 consists of pixels 10-13, 18-21, 26-29, and 34-37. This method implies that each block at a given level participates in four blocks at the immediately higher level. Thus the containment relations between blocks no longer form a tree. For example, pixel 28 participates in blocks F, G, J, and K in the next higher level (see Figure 1.10 where the four neighborhoods corresponding to F, G, J, and K are drawn as squares).

To avoid treating border cases differently, each level in the overlapped pyramid is assumed to be cyclically closed (i.e., the top row at each level is adjacent to the bottom row and similarly for the columns at the extreme left and right of each level). Once again we say that the value of a node is the average of the values of the nodes in its block on the immediately lower level. The overlapped pyramid may be compared with the Quadtree Medial Axis Transform (see Section 9.3.1 of [Same90b]) in the sense that both may result in nondisjoint decompositions of space.

Pyramids have been applied to the problems of feature detection and extraction since they can be used to limit the scope of the search. Once a piece of information of interest is found at a coarse level, the finer resolution levels can be searched. This approach was followed by Davis and Roussopoulos [Davi80] in approximate pattern matching. Pyramids can also be used for encoding information about edges, lines, and curves in an image [Shne81c, Krop86]. One note of caution: the reduction of resolution has an effect on the visual appearance of edges and small objects [Tani76]. In particular, at a coarser level of resolution, edges tend to get smeared, and region separation may disappear. Pyramids have also been used as the starting point for a 'split-and-merge' segmentation algorithm [Piet82].

Quadtree-like decompositions are useful as space-ordering methods. The purpose is to optimize the storage and processing sequences for two-dimensional data by mapping them into one dimension (i.e., linearizing them). This mapping should pre-

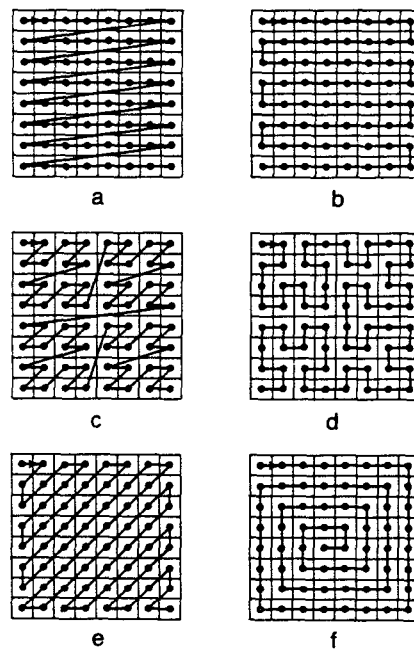


Figure 1.11 The result of applying a number of different space-ordering methods to an 8×8 image whose first element is in the upper left corner of the image: (a) row order, (b) row-prime order, (c) Morton order, (d) Peano-Hilbert order, (e) Cantor-diagonal order, (f) spiral order

serve the spatial locality of the original two-dimensional image in one dimension. The result of the mapping is also known as a *space-filling curve* [Gold81, Witt83] because it passes through every point in the image.

Goodchild and Grandfield [Good83] discuss a number of space-ordering methods, some of which are illustrated in Figure 1.11. Each has different characteristics. The row (Figure 1.11a), also known as raster-scan, and row-prime orders (Figure 1.11b) are similar in the same way as are the Morton [Mort66, Pean90] (Figure 1.11c) and the Peano-Hilbert [Hilb91] (Figure 1.11d) orders. The primary difference is that in both the row-prime and Peano-Hilbert orders every element is a 4-adjacent neighbor of the previous element in the sequence, and thus they have a slightly higher degree of locality than the row and Morton orders, respectively. Both the Morton and Peano-Hilbert orders exhaust a quadrant or subquadrant of a square image before exiting it. They are both related to quadrees; however, as we saw above, the Morton order does not traverse the image in a spatially contiguous manner (the result has the shape of the letter 'N' or 'Z' and is also known as N order [Whit82] and Z order [Oren84]).

For both the Morton and Peano-Hilbert orders, there is no need to know the maximum values of the coordinates. The Morton order is symmetric, while the Peano-Hilbert order is not. One advantage of the Morton order is that the position of each element in the ordering (termed its *key*) can be determined by interleaving the

bits of the x and y coordinates of the element; this is not easy for the Peano-Hilbert order. Another advantage of the Morton order is that the recursion necessary for its generation is quite easy to specify.

Other orders are the Cantor-diagonal order (Figure 1.11e) and the spiral order (Figure 1.11f). The Cantor-diagonal order proceeds outward from the origin and visits the elements in an order similar to row-prime with the difference that elements are visited in order of their increasing ‘Manhattan’ (or ‘city block’) distance.⁷ Thus it is good for ordering a space that is unbounded in the two directions emanating from the origin which has been relocated to the center of the image. On the other hand, the spiral order is attractive when ordering a space that is unbounded in the four directions emanating from the origin.

The most interesting orders, as far as we are concerned, are the Morton and Peano-Hilbert orders since they can also be used to order a space that has been aggregated into squares. Of these two orderings, the Morton order is by far the more frequently used as a result of the simplicity of the conversion process between the key and its corresponding element in the multidimensional space. In this book we are primarily interested in Morton orderings. (For further discussion of some of the properties of these two orderings, see [Patr68, Butz71, Alex79, Alex80, Laur85].)

Exercises

- 1.11. Write an algorithm to extract the x and y coordinates from a Peano-Hilbert order key.
- 1.12. Write an algorithm to construct the Peano-Hilbert key for a given point (x, y) . Try to make it optimal.
- 1.13. Suppose that you are given a $2^n \times 2^n$ array of points such that the horizontal and vertical distances between 4-adjacent points are 1. What is the average distance between successive points when the points are ordered according to the orders illustrated in Figure 1.11? What about a random order?
- 1.14. Suppose that you are given a $2^n \times 2^n$ image. Assume that the image is stored on disk in pages of size $2^m \times 2^m$ where n is much larger than m . What is the average cost of retrieving a pixel and its 4-adjacent neighbors when the image is ordered according to the orders illustrated in Figure 1.11?
- 1.15. The traveling salesman problem [Lawl85] is one where a set of points is given and it is desired to find the path of minimum distance such that each point is visited only once. This is an NP-complete problem [Gare79] and thus there is a considerable amount of work in formulating approximate solutions to it [Bent82]. For example, consider the following approximate solution. Assume that the points are uniformly distributed in the unit square. Let d be the expected Euclidean distance between two independent points. Now, sort the points using the row order and the Morton order. Laurini [Laur85] simulated the average Euclidean distance between successive points in these orders and found it to be $d/2$ for the row order and $d/3$ for the Morton order. Can you derive these averages analytically? What are the average values for the other orders illustrated in Figure 1.11? What about a random order?

⁷ The Manhattan distance between points (x_1, y_1) and (x_2, y_2) is $|x_1 - x_2| + |y_1 - y_2|$ (for more details, see Section 9.1 of [Same90b]).

- 1.16. Suppose that the traveling salesman problem is solved using a traversal of the points in Morton order as discussed in Exercise 1.15. In particular, assume that the set of points is decomposed in such a way that each square block contains just one point. This yields a point representation that is analogous to the region quadtree (termed a PR quadtree and discussed in Section 2.6.2). How close does such a solution come to optimality?

1.4 SPACE DECOMPOSITION METHODS

In general, any planar decomposition used as a basis for an image representation should possess the following two properties:

1. The partition should be an infinitely repetitive pattern so that it can be used for images of any size.
2. The partition should be infinitely decomposable into increasingly finer patterns (i.e., higher resolution).

In this section, the discussion is restricted to two-dimensional data. Thus we are dealing with planar space decompositions. Space decompositions can be classified into two categories, depending on the nature of the pattern. The pattern can consist of polygonal shapes or nonpolygonal shapes. The polygonal shapes are generally computationally simpler since their sides can be expressed in terms of linear relations (e.g., equations of lines). They are good for approximating the interior of a region. The nonpolygonal shapes are more flexible since they provide good approximations, in terms of measures, of the boundaries (e.g., perimeter) of regions as well as their interiors (e.g., area).⁸

Moreover, the normals to the boundaries of nonpolygonal shapes are not restricted to a fixed set of directions. For example, in the case of rectangular tiles, there is a 90 degree discontinuity between the normals to boundaries of adjacent tiles. This lack of continuity is a drawback in applications in fields such as computer graphics where such tasks as shading make use of the directions of the surface. However, working with nonpolygonal shapes generally requires use of floating point arithmetic, and hence it is usually more complex.

The remainder of this section expands on a number of polygonal decompositions and compares them. It also contains a brief discussion of one nonpolygonal decomposition that consists of a collection of sector-like objects whose arcs are not necessarily part of a circle. This method is based on polar coordinates where the arc joining two distinct points is formed by linear interpolation. The term *sector tree* is used to describe it. This discussion is of an advanced nature and can be skipped on an initial reading.

⁸ Recall the statement in Section 1.2 that hierarchical data structures are often differentiated on the basis of whether they specify the boundaries of regions or organize their interiors.

1.4.1 Polygonal Tilings

Bell, Diaz, Holroyd, and Jackson [Bell83] discuss a number of polygonal tilings of the plane (i.e., tessellations) that satisfy property 1. Figure 1.12 illustrates some of these tessellations. They also present a taxonomy of criteria to distinguish between the various tilings. The tilings, consisting of polygonal tiles, are described by use of a notation based on the degree of each vertex as the edges (i.e., sides) of the 'atomic' tile are visited in order, forming a cycle. For example, the tiling described by $[4.8^2]$ (Figure 1.12c) has the shape of a triangle where the first vertex has degree four while the remaining two vertices have degree eight apiece.

A tiling is said to be *regular* if the atomic tiles are composed of regular polygons (i.e., all sides are of equal length as are the interior angles). A *molecular tile* is an aggregation of atomic tiles to form a hierarchy. It is not necessarily constrained to have the same shape as the atomic tile. When a tile at level k (for all $k > 0$) has the same shape as a tile at level 0 (i.e., it is a scaled image of a tile at level 0), then the tiling is said to be *similar*.

Bell et al. focus on the isohedral tilings where a tiling is said to be *isohedral* if all the tiles are equivalent under the symmetry group of the tiling. A more intuitive

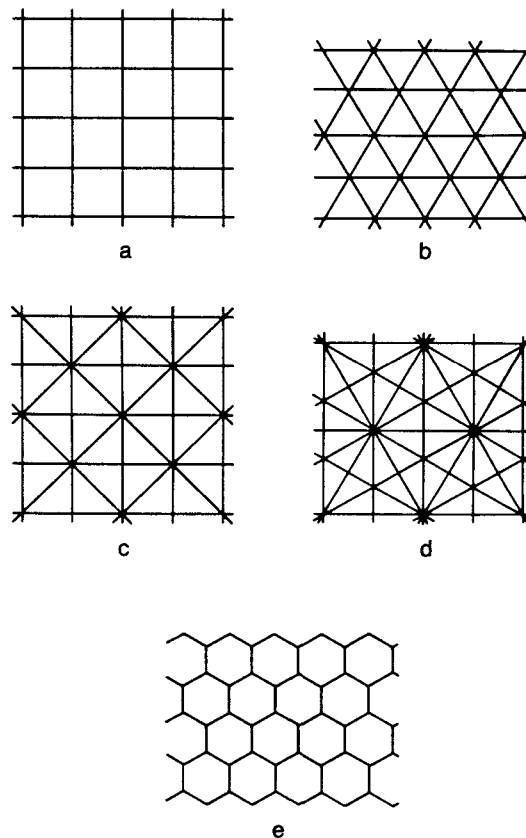


Figure 1.12 Sample tessellations: (a) $[4^4]$ square; (b) $[6^3]$ equilateral triangle; (c) $[4.8^2]$ isosceles triangle; (d) $[4.6.12]$ 30–60 right triangle; (e) $[3^6]$ hexagon

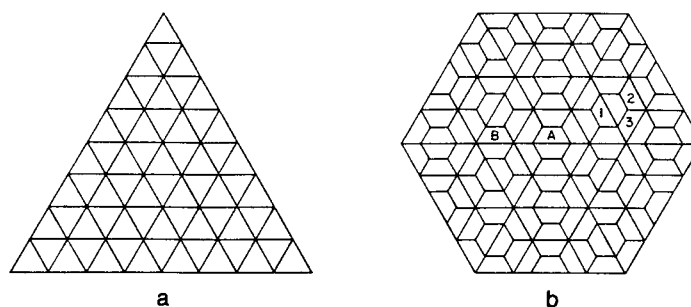


Figure 1.13 Examples of (a) isohedral and (b) nonisohedral tilings

way to conceptualize this definition is to assume the position of an observer who stands in the center of a tile having a given orientation and scans the surroundings. If the view is independent of the tile, the tiling is isohedral. For example, consider the two tilings in Figure 1.13 consisting of triangles (Figure 1.13a) and trapezoids (Figure 1.13b). The triangles are isohedral, whereas the trapezoids are not, as can be seen by the view from tiles A and B.

In the case of the trapezoidal tiling, the viewer from A is surrounded by an infinite number of concentric hexagons, whereas this is not the case for B. In other words, the trapezoidal tiling is not periodic. Also note that all of the tiles in Figure 1.13a are described by $[6^3]$, while those in Figure 1.13b are either $[3^2.4^2]$, $[3^2.6^2]$, or $[3.4.6^2]$ (i.e., tiles labeled 1, 2, and 3, respectively, in Figure 1.13b). When the isohedral tilings are classified by the action of their symmetry group, there are 81 different types [Grün77, Grün87]. When they are classified by their adjacency structure, as done here, there are 11 types.

The most relevant criterion to the discussion is the distinction between limited and unlimited hierarchies of tilings. A *limited* tiling is not similar. A tiling that satisfies property 2 is said to be *unlimited*. Equivalently, in a limited tiling, no change of scale lower than the limit tiling can be made without great difficulty. An alternative characterization of an unlimited tiling is that each edge of a tile lies on an infinite straight line composed entirely of edges. Interestingly the hexagonal tiling $[3^6]$ is limited. Bell et al. claim that only four tilings are unlimited. These are the tilings given in Figure 1.12a–d. Of these, $[4^4]$, consisting of square atomic tiles (Figure 1.12a), and $[6^3]$, consisting of equilateral triangle atomic tiles (Figure 1.12b), are well-known regular tessellations [Ahuj83]. For these two tilings we consider only the molecular tiles given in Figures 1.14a and 1.14b.

The tilings $[4^4]$ and $[6^3]$ can generate an infinite number of different molecular tiles where each molecular tile at the first level consists of n^2 atomic tiles ($n > 1$). The remaining nonregular unlimited triangular tilings, $[4.8^2]$ (Figure 1.12c) and $[4.6.12]$ (Figure 1.12d), are less well understood. One way of generating $[4.8^2]$ and $[4.6.12]$ is to join the centroids of the tiles of $[4^4]$ and $[6^3]$, respectively, to both their vertices and midpoints of their edges. Each of the tilings $[4.8^2]$ and $[4.6.12]$ has two

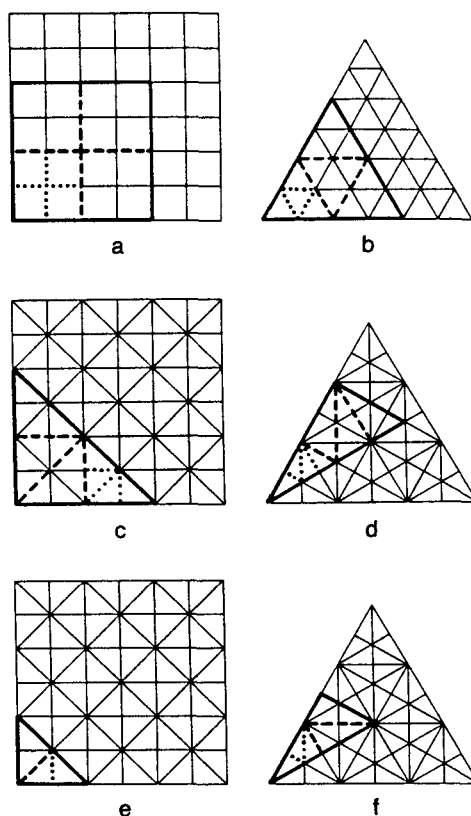


Figure 1.14 Examples illustrating unlimited tilings: (a) $[4^4]$ hierarchy, (b) $[6^3]$ hierarchy, (c) ordinary $[4.8^2]$ hierarchy, (d) ordinary $[4.6.12]$ hierarchy, (e) rotation $[4.8^2]$ hierarchy, (f) reflection $[4.6.12]$ hierarchy

types of hierarchy. $[4.8^2]$ has an ordinary (Figure 1.14c) and a rotation hierarchy (Figure 1.14e) requiring a rotation of 135 degrees between levels. $[4.6.12]$ has an ordinary (Figure 1.14d) and a reflection hierarchy (Figure 1.14f), which requires a reflection of the basic tile between levels.

The distinction between the two types of hierarchies for $[4.8^2]$ and $[4.6.12]$ is necessary because the tiling is not similar without a rotation or a reflection when the hierarchy is not ordinary. This can be seen by observing the use of dots in Figure 1.14 to delimit the atomic tiles in the first molecular tile. Similarly broken lines are used to delimit the components of tiles at the second level (assuming atomic tiles are at level 0). For the ordinary $[4.8^2]$ and $[4.6.12]$ hierarchies, each molecular tile at the first level consists of n^2 ($n > 1$) atomic tiles. In the reflection hierarchy of $[4.6.12]$, each molecular tile at the first level consists of $3 \cdot n^2$ ($n > 1$) atomic tiles, while for the rotation hierarchy of $[4.8^2]$, $2 \cdot n^2$ ($n > 1$) atomic tiles comprise a molecular tile at the first level.

To represent data in the Euclidean plane, any of the unlimited tilings could have been chosen. For a regular decomposition, the tilings $[4.8^2]$ and $[4.6.12]$ are ruled out. Comparing 'square' $[4^4]$ and 'triangular' $[6^3]$ quadtrees, we find that they differ in terms of adjacency and orientation. Let us say that two tiles are *neighbors* if they are

adjacent either along an edge or at a vertex. A tiling is *uniformly adjacent* if the distances between the centroid of one tile and the centroids of all its neighbors are the same. The adjacency number of a tiling is the number of different intercentroid distances between any one tile and its neighbors. In the case of $[4^4]$, there are only two adjacency distances, whereas for $[6^3]$ there are three adjacency distances.

A tiling is said to have *uniform orientation* if all tiles with the same orientation can be mapped into each other by translations of the plane that do not involve rotation or reflection. Tiling $[4^4]$ displays uniform orientation, while $[6^3]$ does not. Under the assumption that uniform orientation and a minimal adjacency distance is preferable, we say that $[4^4]$ is more useful than $[6^3]$. It is also very easy to implement. Nevertheless, $[6^3]$ has its uses. For example, Yamaguchi, Kunii, Fujimura, and Toriya [Yama84] use a triangular quadtree to generate an isometric view from an octree representation of an object (see Section 7.1.4 of [Same90b]).

Of the *limited* tilings, many types of hierarchies may be generated [Bell83]; however, in general, they cannot be decomposed beyond the atomic tiling without changing the basic tile shape. This is a serious deficiency of the hexagonal tessellation $[3^6]$ (Figure 1.12e) since the atomic hexagon can be decomposed only into triangles. Nevertheless the hexagonal tessellation is of considerable interest. It is regular, has a uniform orientation, and, most important, displays a uniform adjacency (i.e., each neighbor of a tile is at the same distance from it).

There are a number of different hexagonal hierarchies distinguished by classifying the shape of the first-level molecular tile on the basis of the number of hexagons that it contains. Three of these tiling hierarchies are given in Figure 1.15 and are called *n-shapes* where *n* denotes the number of atomic tiles in the first-level molecular tile. Of course, these n-shapes are not unique.

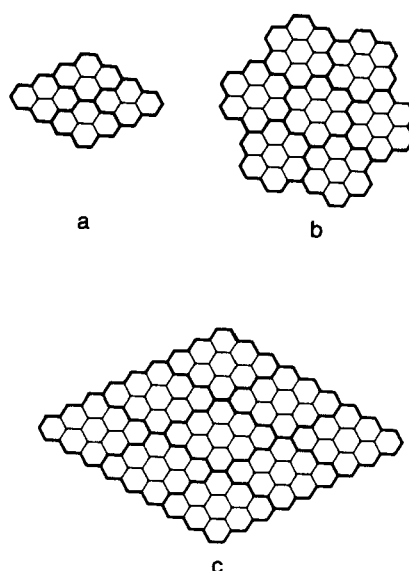


Figure 1.15 Three different hexagonal tiling hierarchies: (a) 4-shape, (b) 7-shape, (c) 9-shape

The 4-shape and the 9-shape have an unusual adjacency property in the sense that no matter how large the molecular tile becomes, contact with two of the tiles (i.e., the one above and the one below) is along only one edge of a hexagonal atomic tile, while contact with the remaining four molecular tiles is along nearly one-quarter of the perimeter of the corresponding molecular tile. The hexagonal pattern of the 4-shape and 9-shape molecular tiles has the shape of a rhombus. In contrast, a 7-shape molecular tile has a uniform contact with its six neighboring molecular tiles.

The type of quadtree used often depends on the grid formed by the image sampling process. Square quadtrees are appropriate for square grids and triangular quadtrees for triangular grids. In the case of a hexagonal grid [Burt80], the 7-shape hierarchy is frequently used since the shape of its molecular tile is more like a hexagon. It is usually described as *rosette*-like (i.e., a *septree*). Note that septrees have jagged edges as they are merged to form larger units (e.g., Figure 1.15b). The septree is used by Gibson and Lucas [Gibs82] (who call it a *generalized balanced ternary* or *GBT* for short) in the development of algorithms analogous to those existing for quadtrees.

Although the septree can be built up to yield large septrees, the smallest resolution in the septree must be decided upon in advance since its primitive components (i.e., hexagons) cannot later be decomposed into septrees. Therefore the septree yields only a partial hierarchical decomposition in the sense that the components can always be merged into larger units, but they cannot always be broken down. For region data, a pixel is generally an indivisible unit, and thus unlimited decomposition is not absolutely necessary. However, in the case of other data types such as points (see Chapter 2) and lines (see Chapter 4), we will see that the decomposition rules of some representations require that two entities be separated, which may lead to a level of decomposition not known in advance (e.g., a decomposition rule that restricts each square to contain at most one point). In this book the discussion is limited to square quadtrees and their variants.

When the data are spherical, a number of researchers have proposed the use of a representation based on an icosahedron (a 20-faced polyhedron whose faces are regular triangles) [Dutt84, Feke84]. The icosahedron is attractive because, in terms of the number of faces, it is the largest possible regular polyhedron. Each of the triangular faces can be further decomposed in a recursive manner into n^2 ($n > 1$) spherical triangles (the $[6^3]$ tiling).

Fekete and Davis [Feke84] let $n = 2$, which means that at each level of decomposition, three new vertices are generated by halving each side of the triangle; connecting them together yields four triangles. They use the term *property sphere* to describe their representation. The property sphere has been used in object recognition; it is also of potential use in mapping the globe because it can enable accurate modeling of regions around the poles. For example, see Figure 1.16, which is a property sphere representation of some spherical data. In contrast, planar quadtrees are less attractive the farther we get from the equator due to distortions in planarity caused by the earth's curvature. Of course, for true applicability for mapping, we need a closer approximation to a sphere than is provided by the 20 triangles of the icosahedron. Moreover, we want a way to distinguish between different elevations.

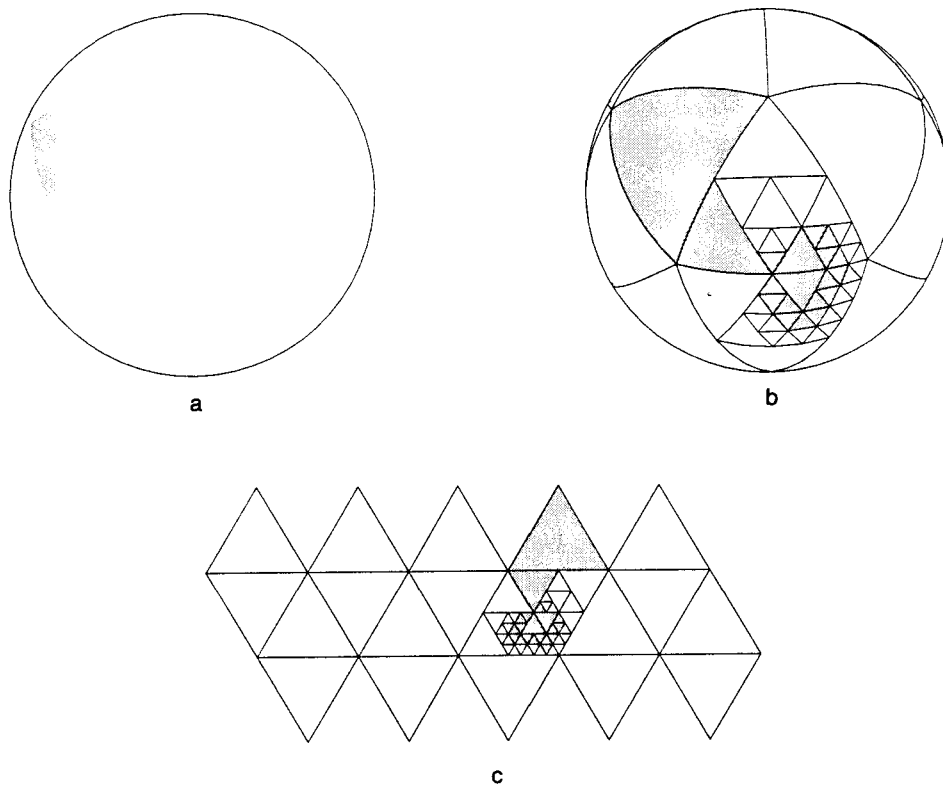


Figure 1.16 Property sphere representation of some spherical data: (a) data, (b) decomposition on a sphere, (c) decomposition on a plane

Dutton [Dutt84] lets $n = \sqrt{3}$, which means that at each level of decomposition, one new vertex is created by connecting the centroid of the triangle to its vertices. The result is an alternating sequence of triangles so that each level is fully contained in the level that was created two steps previously and has nine times as many triangles as that level. Dutton uses the term *triacon* to describe the resulting hierarchy. As an example, consider Figure 1.17, which illustrates four levels of a triacon decomposition. The initial and odd-numbered decompositions are shown with heavy lines, and the even-numbered decompositions are shown with broken and thin lines.

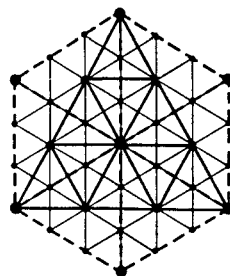


Figure 1.17 Example of a triacon hierarchy

The icosahedron is not the only regular polyhedron that can be used to model spherical data. Others include the tetrahedron, hexahedron, octahedron, and dodecahedron, which have 4, 6, 8, and 12 faces, respectively. Collectively these five polyhedra are known as the *Platonic solids* [Peuq84]. The faces of the tetrahedron and octahedron are equilateral triangles, while the faces of the hexahedron and dodecahedron are squares and regular pentagons, respectively.

The dodecahedron is not an appropriate primitive because the pentagonal faces cannot be further decomposed into pentagons or other similar shapes. The tetrahedron and hexahedron (the basis of the octree) have internal angles that are too small to model a sphere properly, thereby leading to shape distortions.

Dutton [Dutt84] points out that the octahedron is attractive for modeling spherical data such as the globe because it can be aligned so that the poles are at opposite vertices and the prime meridian and the equator intersect at another vertex. In addition, one subdivision line of each face is parallel to the equator. Of course, for all of the Platonic solids, only the vertices of the solids touch the sphere; the facets of the solids are interior to the sphere.

Other decompositions for spherical data are also possible. Tobler and Chen [Tobl86] point out the desirability of a close relationship to the commonly used system of latitude and longitude coordinates. In particular, any decomposition that is chosen should enable the use of meridians and parallels to refer to the data. An additional important goal is for the partition to be into units of equal area, which rules out the use of equally spaced lines of latitude (of course, the lines of longitude are equally spaced). In this case, the sphere is projected into a plane using Lambert's cylindrical projection [Adam49], which is locally area preserving. Authalic coordinates [Adam49], which partition the projection into rectangles of equal area, are then derived. (For more details, see [Tobl86].)

The quadtree decomposition has the property that at each subdivision stage, the image is subdivided into four equal-sized parts. When the original image is a square, the result is a collection of squares, each of which has a side whose length is a power of 2. The binary image tree (termed *bintree*) [Know80, Tamm84a, Same88b] is an alternative decomposition defined in a manner analogous to the region quadtree except that at each subdivision stage we subdivide the image into two equal-sized parts. In two dimensions, at odd stages, we partition along the x coordinate, and at even stages, along the y coordinate. The bintree is equivalent to the region quadtree if we replace all leaf nodes at odd stages of subdivision by two identically colored sons.

The bintree is related to the region quadtree in the same way as the k -d tree [Bent75b] (see Section 2.4) is related to the point quadtree [Fink74]. The difference is that region quadtrees and bintrees are used to represent region data with fixed subdivision points, while point quadtrees and k -d trees are used to represent point data where the values of the points determine the subdivision. For example, Figure 1.18 is the bintree representation corresponding to the image of Figure 1.1. We assume that for the x (y) partition, the left subtree corresponds to the west (south) half of the image and the right subtree corresponds to the east (north) half. Once again, as in Figure 1.1, all leaf nodes are labeled with numbers, and the nonleaf nodes are labeled with letters.

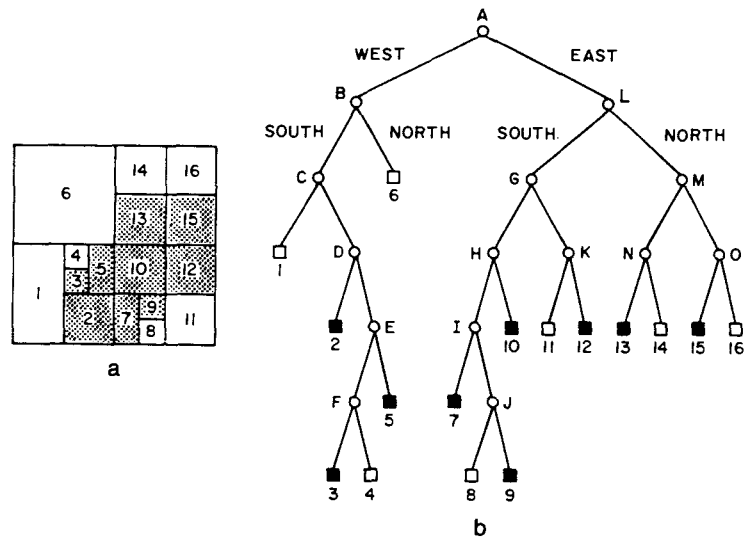


Figure 1.18 Bintree representation corresponding to Figure 1.1: (a) block decomposition, (b) bintree representation of blocks in (a)

The quadtree and bintree decompose a region into equal-sized parts. Kanatani [Kana85] suggests using splitting rules based on the Fibonacci sequence of numbers. The Fibonacci numbers consist of the sequence of numbers f_i that satisfy the relation $f_i = f_{i-1} + f_{i-2}$, with $f_0 = 1$ and $f_1 = 1$. We can try to devise both quadtree and bintree splitting rules based on such a sequence. Generally for a decomposition scheme to be useful in geometric applications, it must have pixel-sized squares (i.e., 1×1) as the primitive tiles. At first glance, it appears that the Fibonacci sequence gives quite a bit of leeway in deciding on a splitting sequence and on the sizes of the regions corresponding to the subtrees and the primitive tiles.

One possible quadtree splitting rule is to restrict all shapes to squares with sides whose lengths are Fibonacci numbers. Clearly not all the shapes can be squares since we cannot aggregate these squares into larger squares that obey this rule. Another possibility is to restrict the shapes to rectangles the length of whose sides are either equal Fibonacci numbers or are successive Fibonacci numbers (see Exercise 1.26). We term this condition the *2-d Fibonacci condition*.

In this discussion, we have assumed splitting rules that ensure that vertical subdivision lines at the same level are colinear as well as for horizontal lines at the same level. For example, when using a quadtree splitting rule, the vertical lines that subdivide the NW and SW quadrants are colinear, as well as for the horizontal lines that subdivide the NW and NE quadrants. An alternative is to relax the colinearity restriction; however, the sides of the shapes must still satisfy the 2-d Fibonacci condition (see Exercise 1.27).

As can be seen in Exercises 1.26 and 1.27, neither a quadtree nor a bintree can

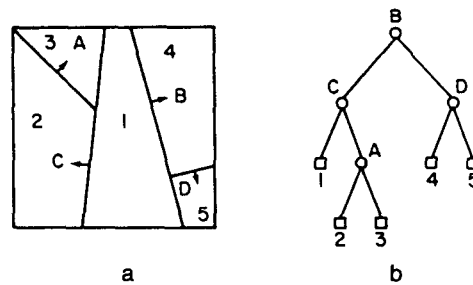


Figure 1.19 (a) An arbitrary space decomposition and (b) its BSP tree. The arrows indicate the direction of the positive halfspaces.

be used by itself as a basis for Fibonacci-based space decomposition; however, a combination of the two structures could be used. When the lengths of the sides of a rectangle are equal, the rectangle is split into four rectangles such that the lengths of the sides satisfy the 2-d Fibonacci condition. When the lengths of the sides of a rectangle are not equal, the rectangle is split into two rectangles with the split along a line (an axis) parallel to the shorter (longer) of the two sides. Interestingly the dimensions of the A-series of European paper are based on a Fibonacci sequence—that is, the elements of the series are of dimension $f_i \times f_{i-1}$ multiplied by an appropriate scale factor.

Another variation on the bintree idea, termed *adaptive hierarchical coding* (AHC), is proposed by Cohen, Landy, and Pavel [Coh85b]. In this case, the image is again split into two equal-sized parts at each stage, but there is no need to alternate between the x and y coordinates. The decision as to the coordinate on which to partition depends on the image. This technique may require some work to get the optimal partition from the point of view of a minimum number of nodes (see Exercise 1.29).

An even more general variation on the bintree is the *BSP tree* of Fuchs, Kedem, and Naylor [Fuch80, Fuch83]. Its variants are used in some hidden-surface elimination algorithms (see Section 7.1.5 of [Same90b]) and in some implementations of beam tracing (see Section 7.3 of [Same90b]). It is applicable to data of arbitrary dimension, although here it is explained in the context of two-dimensional data. At each subdivision stage, the image is subdivided into two parts of arbitrary size. Note that successive subdivision lines need be neither orthogonal nor parallel. Therefore the resulting decomposition consists of arbitrarily shaped convex polygons.

The BSP tree is a binary tree. To be able to assign regions to the left and right subtrees, we associate a direction with each subdivision line. In particular, the subdivision lines are treated as separators between two halfspaces.⁹ Let the line have the

⁹ A (linear) *halfspace* in d -space is defined by the inequality $\sum_{i=0}^d a_i \cdot x_i \geq 0$ on the $d+1$ homogeneous coordinates ($x_0 = 1$). The halfspace is represented by a column vector a . In vector notation, the inequality is written as $a \cdot x \geq 0$. In the case of equality, it defines a hyperplane with a as its normal. It is important to note that halfspaces are volume, not boundary, elements.

equation $a \cdot x + b \cdot y + c = 0$. We say that the right subtree is the ‘positive’ side and contains all subdivision lines formed by separators that satisfy $a \cdot x + b \cdot y + c \geq 0$. Similarly we say that the left subtree is ‘negative’ and contains all subdivision lines formed by separators that satisfy $a \cdot x + b \cdot y + c < 0$. As an example, consider Figure 1.19a, which is an arbitrary space decomposition whose BSP tree is given in Figure 1.19b. Notice the use of arrows to indicate the direction of the positive halfspaces.

Exercises

- 1.17. Given a $[6^3]$ tiling such that each side of an atomic tile has a unit length, compute the three adjacency distances from the centroid of an atomic tile.
- 1.18. Repeat Exercise 1.17 for $[3^6]$ and $[4^4]$, again assuming that each side of an atomic tile has a unit length.
- 1.19. Suppose that you are given an image in the form of a binary array of pixels. The result is a square grid. How can you view this grid as a hexagonal grid?
- 1.20. Show how the property sphere data structure can be used to model the earth. In particular, discuss how to represent landmass features, such as mountain ranges and crevices.
- 1.21. Suppose that you use an icosahedron to model spherical data. Initially there are 20 faces. How many faces are there after the first level of decomposition when $n = 2$? $n = \sqrt{3}$?
- 1.22. What is the ratio of leaf nodes to nonleaf nodes in a bintree for a d -dimensional image?
- 1.23. What is a lower bound on the ratio of leaf nodes in a bintree to that in a quadtree for a d -dimensional image? What is an upper bound? What is the average?
- 1.24. Is it true that the total number of nodes in a bintree is always less than that in the corresponding quadtree?
- 1.25. The Fibonacci numbers are defined by the relation $f_n = f_{n-1} + f_{n-2}$. Devise a two-dimensional analog of this relation to correspond to a splitting rule that would have to be satisfied in a Fibonacci-based space decomposition that yields four parts. Generalize this result to n dimensions.
- 1.26. Give a counterexample to the use of a quadtree splitting rule in a Fibonacci-based space decomposition.
- 1.27. Give a counterexample to the use of a bintree splitting rule in a Fibonacci-based space decomposition.
- 1.28. Suppose that you use the combination quadtree-bintree approach to a Fibonacci-based space decomposition. Prove that any image such that the lengths of its sides satisfy the 2-d Fibonacci condition can be decomposed into subimages whose sides obey this property and with a primitive tile of size 1×1 .
- 1.29. Suppose that you use the AHC method. How many different rectangles and positions must be examined in building such a structure for a $2^n \times 2^n$ image?

1.4.2 Nonpolygonal Tilings

In the previous section we focused on space decompositions based on polygonal tiles. This is the prevalent method in use today. For certain applications, however, the use of polygonal tiles can lead to problems. For example, suppose that we have a decomposition based on square tiles. In this case, as the resolution is increased, the area of the approximated region approaches the true value of the area; however, this is not

true for a boundary measure such as the perimeter. To see this, consider a quadtree approximation of an isosceles right triangle where the ratio of the approximated perimeter to the true perimeter is $4/(2 + \sqrt{2})$ (see Exercise 1.30). Other problems include the discontinuity of the normals to the boundaries of adjacent tiles.

There are a number of ways of attempting to overcome these problems. The *hierarchical probe model* of Chen [Chen85b] is an approach based on treating space as a polar plane and recursively decomposing it into sectors. We say that each sector consists of an origin, two sides (labeled 1 and 2 corresponding to the order in which they are encountered when proceeding in a counterclockwise direction), and an arc. The points at which the sides of the sector intersect (or touch) the object are called *contact points*. (ρ, θ) denotes a point in the polar plane. Let (ρ_i, θ_i) be the contact point with the maximum value of ρ in direction θ_i . Each sector represents a region bounded by the points $(0,0)$, (ρ_1, θ_1) , and (ρ_2, θ_2) , where $\theta_1 = 2k\pi/2^n$ and $\theta_2 = \theta_1 + 2\pi/2^n$ such that k and n are nonnegative integers ($k < 2^n$). The arc between the two nonorigin contact points (ρ_1, θ_1) and (ρ_2, θ_2) of a sector is approximated by the linear parametric equations ($0 \leq t \leq 1$):

$$\rho(t) = \rho_1 + (\rho_2 - \rho_1) \cdot t \quad \theta(t) = \theta_1 + (\theta_2 - \theta_1) \cdot t.$$

Note that the interpolation curves are arcs of spirals due to the linear relation between ρ and θ .

The *sector tree* is a binary tree that represents the result of recursively subdividing sectors in the polar plane into two sectors of equal angular intervals. Thus the recursive decomposition is only with respect to θ , not ρ . The decomposition stops whenever the approximation of a part of an object by a sector is deemed to be adequate. The computation of the stopping condition is implementation dependent. For example, it can be the maximum deviation in the value of ρ between a point on the boundary and the corresponding point (i.e., at the same value of θ) on the approximating arc. Initially the universe is the interval $[0, 2\pi)$.

In the presentation, we assume that the origin of the polar plane is contained within the object. See Exercise 1.36 for a discussion of how to represent an object that does not contain the origin of the polar plane. The simplest case arises when the object is convex. The result is a binary tree where each leaf node represents a sector and contains the contact points of its corresponding arc. For example, consider the object in Figure 1.20. The construction of its sector tree approximation is shown in



Figure 1.20 Example convex object

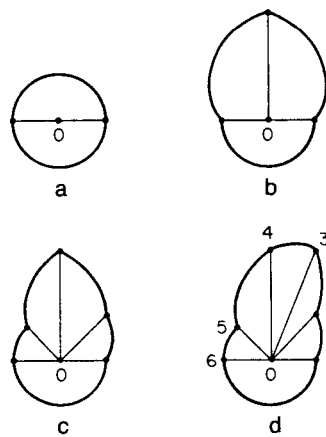


Figure 1.21 Successive sector tree approximations for the object of Figure 1.20: (a) π intervals, (b) $\pi/2$ intervals, (c) $\pi/4$ intervals, (d) $\pi/8$ intervals

Figures 1.21a–d. The final binary tree is given in Figure 1.22 with interval endpoints labeled according to Figure 1.21d.

The situation is more complex when the object is not convex. This means that each side of a sector may intersect the boundary of the object at an arbitrary, and possibly different, number of contact points. In the following, each sector will be seen to consist of a set of alternating regions within and outside the object. These regions are three-sided or four-sided and have at least one side that is colinear with a side of the sector. The discussion is illustrated with the object of Figure 1.23a whose sector tree decomposition is given in Figure 1.23b. The final binary tree is given in Figure 1.24. A better indication of the quality of the approximation can be seen by examining Figure 1.23c, which contains an overlay of Figures 1.23a and 1.23b.

When the boundary of the object intersects a sector at two successive contact points, say P and Q , that lie on the same side, say S , of the sector, then the region

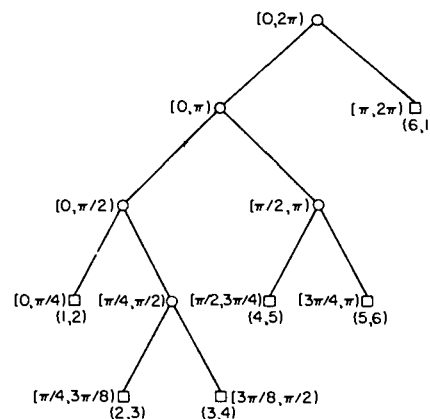


Figure 1.22 Binary tree representation of the sector tree of Figure 1.20

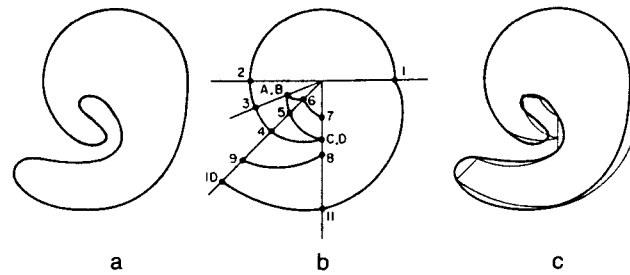


Figure 1.23 (a) Example object, (b) its sector tree description, and (c) a comparison of the sector tree approximation (thin lines) with the original object (thick lines). Note the creation of a hole corresponding to the region formed by points A, B, 6, 7, C, D, and 5

bounded by s and PQ must be approximated. Without loss of generality, assume that the region is inside the object. There are two choices. An inner approximation ignores the region by treating the segment of s between P and Q as part of the approximated boundary (e.g., the region between points 9 and 10 in sector $[9\pi/8, 5\pi/4]$ in Figure 1.23b).

An outer approximation inserts two identical contact points, say R and T , on the other side of the sector and then approximates the region by the three-sided region formed by the segment of s between P and Q and the spiral arc approximations of PR and QT . The value of R (and hence T) is equal to the average of the value of ρ at P and Q . For example, the region between points 4 and 5 in sector $[5\pi/4, 3\pi/2]$ in Figure 1.23b is approximated by the region formed with points C and D.

Of course, the same approximation process is applied to the part of the region outside the object. In Figure 1.23b, we have an inner approximation for the region between points 7 and 8 in sector $[3\pi/2, 2\pi)$, and an outer approximation for the region between points 5 and 6 in sector $[9\pi/8, 5\pi/4)$, by virtue of the introduction of points A and B.

One of the problems with the sector tree is that its use can lead to the creation of holes that do not exist in the original object. This situation arises when the decomposition is not carried out to a level of sufficient depth. For example, consider Figure 1.23b, which has a hole bounded by the arcs formed by points A, B, 6, 7, C, D, and 5. This is a result of the inner approximation for the region between points 7 and 8 in sector $[3\pi/2, 2\pi)$ and an outer approximation for the region between points 4 and 5 in sector $[5\pi/4, 3\pi/2)$. This situation can be resolved by further decomposition in either or both of sectors $[3\pi/2, 2\pi)$ and $[5\pi/4, 3\pi/2)$.

The result of the approximation process is that each sector consists of a collection of three-sided and four-sided regions that approximate the part of the object contained in the sector. This collection is stored in the leaf node of the sector tree as a list of pairs of points in the polar plane. It is interesting to observe that the boundaries of the interpolated regions are not stored explicitly in the tree. Instead each pair of points corresponds to the boundary of a region. Since the origin of the polar plane is within the object, an odd number of pairs of points is associated with each leaf node. For

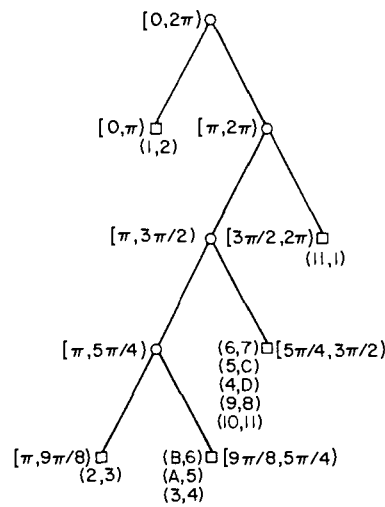


Figure 1.24 Binary tree representation of the sector tree of Figure 1.23

example, consider the leaf node in Figure 1.24 corresponding to the sector $[5\pi/4, 3\pi/2)$. The first pair, together with the origin, defines the first region (e.g., (6,7)). The next two pairs of points define the second region (e.g., (5,C) and (4,D)), with each successive two pairs of points defining the remaining regions.

The sector tree is a partial polar decomposition, as the subdivision process is based only on the value of θ . A total polar decomposition would partition the polar plane on the basis of both ρ and θ . The result is analogous to a quadtree, and it is termed a *polar quadtree*. There are a number of possible rules for the decomposition process (see Exercise 1.42). For example, consider a decomposition that recursively halves both ρ and θ at each level. In general, the polar quadtree is a variant of a maximal block representation. As in the sector tree, the blocks are disjoint. Unlike the sector tree, blocks in the polar quadtree do have standard sizes. In particular, all blocks in the polar quadtree are either three sided (i.e., sectors) or four sided (i.e., quadrilaterals, two of whose sides are arcs). Thus the sides of polar quadtree blocks are not based on interpolation.

The primary motivation for presenting the sector tree is to show that space decompositions could also be based on nonpolygonal tiles. In the rest of this book the primary concern is with space decompositions based on rectangles (especially squares) and showing how a number of operations can be performed when they serve as the underlying representation. The techniques are quite general and can be applied to most space decomposition methods. Thus the sector tree is not discussed further except in the context of its adaptation to the representation of three-dimensional data (see Section 5.6). Nevertheless, the following contains a brief mention of some of the operations to which the sector tree lends itself.

Set operations such as union and intersection are straightforward. Scaling is trivial as the sector tree need not be modified; all values of ρ are interpreted as scaled

by the appropriate scale factor. The number of nodes in a sector tree is dependent on its orientation—that is, on the points chosen as the origin and the contact point chosen to serve as $(\rho, 0)$. Rotation is not so simple; it cannot be implemented by simply rearranging pointers (but see Exercise 1.40). Translation is computationally expensive since the change in the relative position of the object with respect to the origin means that the entire sector tree must be reconstructed.

Exercises

- 1.30. Prove that for an isosceles right triangle represented by a region quadtree, the ratio of the approximated perimeter to the true perimeter is $4/(2 + \sqrt{2})$.
- 1.31. Repeat Exercise 1.30 for a circle (i.e., find the ratio).
- 1.32. When the objects have linear sides, polygonal tiles are superior. How would you use the sector tree decomposition method with polygonal tiles?
- 1.33. In the discussion of the situation arising when the boundary of the object intersects a sector at two successive contact points, say P and Q , that lie on the same side, say S , of the sector, we assumed that the region bounded by S and PQ was inside the object. Suppose that this region is outside the object. How does this affect the inner and outer approximations?
- 1.34. Can you traverse the boundary of an object represented by a sector tree by visiting each leaf node just once?
- 1.35. When using a sector tree, how would you handle the situation that the boundary of the object just touches the side of a sector without crossing it (i.e., a tangent if the boundary is differentiable)?
- 1.36. How would you use a sector tree to represent an object that does not contain the origin of the polar plane?
- 1.37. The outer approximation used in building a sector tree always yields a three-sided region. Two of the sides are arcs of spirals with respect to a common origin. This implies a sharp discontinuity of the derivative at the point at which they meet. Can you devise a way to smoothe this discontinuity?
- 1.38. Does the inner approximation used in building a sector tree always underestimate the area? Similarly does the outer approximation always overestimate the area?
- 1.39. Compare the inner and outer approximations used in building a sector tree. Is there ever a reason for the outer approximation to be preferred over the inner approximations (or vice-versa)?
- 1.40. Define a complete sector tree in an analogous manner to a complete binary tree—that is, all leaf nodes are at the same level, say n . Prove that a complete sector tree is invariant under rotation in multiples of $2\pi/2^n$.
- 1.41. Write an algorithm to trace the boundary of an object represented by a sector tree.
- 1.42. Suppose that it is desired to decompose space into nonpolygonal shapes. Develop a quadtree-like data structure based on polar coordinates (i.e., ρ and θ). Investigate different splitting rules for polar quadtrees. In particular, you do not need to alternate the splits—that is, you could split on ρ several times in a row, and so on. This technique is used in the adaptive k - d tree [Frie77] (see Section 2.4.1) by decomposing the quartering process into two splitting operations—one for the x coordinate and one for the y coordinate. What are the possible shapes for the quadrants of such trees (e.g., a torus, doughnut, wheels with spokes)?

1.5 SPACE REQUIREMENTS

The primary motivation for the development of the quadtree was the desire to reduce the amount of space necessary to store data through the use of aggregation of homogeneous blocks. As we will see in subsequent chapters, an important by-product of this aggregation is the reduction of the execution time of a number of operations (e.g., connected component labeling, component counting). However, a quadtree implementation does have overhead in terms of the nonleaf nodes. For an image with B and w black and white blocks, respectively, $4 \cdot (B + w)/3$ nodes are required. In contrast, a binary array representation of a $2^n \times 2^n$ image requires only 2^{2n} bits; however, this quantity grows quite quickly. Furthermore, if the amount of aggregation is minimal (e.g., a checkerboard image), the quadtree is not very efficient.

The overhead for the nonleaf nodes can be reduced at times by using a pointer-less representation. Pointer-less representations can be grouped into two categories. The first, termed a *DF-expression*, represents the quadtree as a traversal of its constituent nodes [Kawa80a]. For example, letting 'B', 'W', and 'G' correspond to black, white, and gray nodes, respectively, and assuming a traversal in the order NW, NE, SW, and SE, the quadtree of Figure 1.1 would be represented by GWGWWBGGWGW BBBWBGBBGBBBWW.

The second approach treats the quadtree as a collection of the leaf nodes comprising it. Each node is represented by a pair of numbers [Garg82c]. The first number is the level of the tree at which the node is located. The second number is termed a *locational code*. It is formed by a concatenation of base 4 digits corresponding to directional codes that locate the node along a path from the root of the quadtree. The directional codes take on the values 0, 1, 2, 3 corresponding to quadrants NW, NE, SW, SE, respectively. For example, node 15 in Figure 1.1 is represented by the pair of numbers (0,320), which is decoded as follows. The base 4 locational code is 320. The pair denotes a node at level 0 that is reached by a sequence of transitions, SE, SW, and NW, starting at the root. A quadtree representation based on the use of locational codes is called *linear quadtree* by Gargantini [Garg82a, Garg82c] (because the addresses are keys in a linear list of nodes). Pointer-less representations are discussed in greater detail in Chapter 2 of [Same90b].

The worst case for a quadtree of a given depth in terms of storage requirements occurs when the region corresponds to a checkerboard pattern as in Figure 1.25. The amount of space required is obviously a function of the resolution (i.e., the number of levels in the quadtree), the size of the image (i.e., its perimeter), and its positioning in the grid within which it is embedded. As a simple example, Dyer [Dyer82] has shown that arbitrarily placing a square of size $2^m \times 2^m$ at any position in a $2^n \times 2^n$ image requires an average of $O(2^{m+2} + n - m)$ quadtree nodes. An alternative characterization of this result is that the average amount of space necessary is $O(p+n)$ where p is the perimeter (in pixel widths) of the block.

Dyer's $O(p+n)$ result for a square image is merely an instance of the earlier work of Hunter and Steiglitz [Hunt78, Hunt79a] who proved some fundamental theorems on the space requirements of images represented by quadtrees. In their

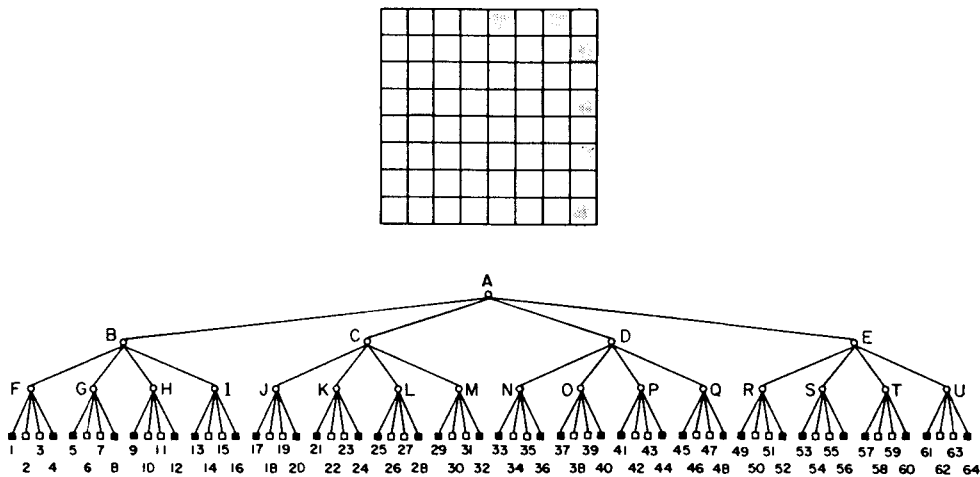


Figure 1.25 A checkerboard and its quadtree

studies, Hunter and Steiglitz used simple polygons (polygons with nonintersecting edges and without holes); however, these theorems have been observed to hold in arbitrary images (see [Rose82b] for empirical results in a cartographic environment).

In Hunter and Steiglitz's formulation, a polygon is represented by a three-color variant of the quadtree. In essence, there are three types of nodes: interior, boundary, and exterior. A node is said to be of type *boundary* if an edge of the polygon passes through it. *Interior* and *exterior* nodes correspond to areas within, and outside, respectively, the polygon and can be merged to yield larger nodes. The resulting quadtree is analogous to the MX quadtree representation of point data described below (for more details, see Section 2.6.1), and this term will be used to describe it. In particular, boundary nodes are analogous to black nodes, while interior and exterior nodes are analogous to white nodes.

Figure 1.26 illustrates a sample polygon and its MX quadtree. One disadvantage of the MX quadtree representation for polygonal lines is that a width is associated with them, whereas in a purely technical sense these lines have a width of zero. Also shifting operations may result in information loss. (For more appropriate representations of polygonal lines, see Chapter 4.)

An upper bound on the number of nodes in such a representation of a polygon can be obtained in the following manner. First, we observe that a curve of length $d + \epsilon$ ($\epsilon > 0$) can intersect at most six squares of side width d . Now consider a polygon, say G , having perimeter p , that is embedded in a grid of squares each of side width d . Mark the points at which G enters and exits each square. Choose one of these points, say P , as a starting point for a decomposition of G into a sequence of curves. Define the first curve in G to be the one extending from P until six squares have been intersected and a crossing is made into a different seventh square. This is the starting point for another curve in G that intersects six new squares, not counting those intersected by any previous curve.

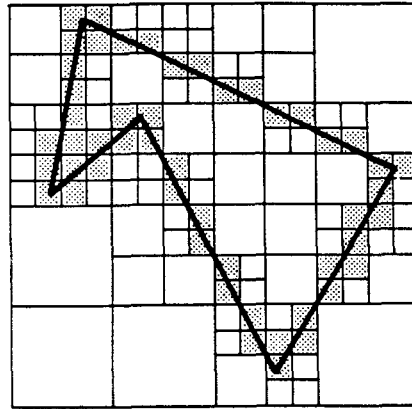


Figure 1.26 Hunter and Steiglitz's quadtree representation of a polygon

We now decompose G into a series of such curves. Since each curve adds at most six new squares and has length of at least d , we see that a polygon with perimeter p cannot intersect more than $6 \cdot \lceil p/d \rceil$ squares. Given a quadtree with a root at level n (i.e., the grid of squares is of width 2^n), at level i each square is of width 2^i . Therefore polygon G cannot intersect more than $B(i) = 6 \cdot \lceil p/2^i \rceil$ quadrants at level i . Recall that our goal is to derive an upper bound on the total number of nodes. This bound is attained when each boundary node at level i has three brother nodes that are not intersected. Of course, only boundary nodes can have sons, and thus no more than $B(i)$ nodes at level i have sons. Since each node at level i is a son of a node at level $i+1$, there are at most $4 \cdot B(i+1)$ nodes at level i . Summing up over n levels (accounting for a root node at level n and four sons), we find that the total number of nodes in the tree is bounded by

$$\begin{aligned}
 & 1 + 4 + \sum_{i=0}^{n-2} 4 \cdot B(i+1) \\
 & \leq 5 + 24 \cdot \sum_{i=0}^{n-2} \left\lceil \frac{p}{2^{i+1}} \right\rceil \\
 & \leq 5 + 24 \cdot \sum_{i=0}^{n-2} \left(1 + \frac{p}{2^{i+1}} \right) \\
 & \leq 5 + 24 \cdot (n-1) + 24 \cdot p \cdot \sum_{i=0}^{n-2} \frac{1}{2^{i+1}} \\
 & \leq 24 \cdot n - 19 + 24 \cdot p.
 \end{aligned}$$

Therefore, we have proved:

Theorem 1.1 The quadtree corresponding to a polygon with perimeter p embedded in a $2^n \times 2^n$ image has a maximum of $24 \cdot n - 19 + 24 \cdot p$ (i.e., $O(p+n)$) nodes. \square

The proof of Theorem 1.1 is based on a decomposition of the polygon into a sequence of curves, each of which intersects at most six squares. This bound can be tightened by examining patterns of squares to obtain minimum lengths and corresponding ratios of possible squares per unit length. For example, observe that once a curve intersects six squares, the next curve of length d in the sequence can intersect at most two new squares. In contrast, it is easy to construct a sequence of curves of length $d + \epsilon$ ($\epsilon > 0$) such that almost each curve intersects two squares of side length d . Such a construction leads to an upper bound of the form $a \cdot n + b + 8 \cdot p$ where a and b are constants (see Exercise 1.48). Hunter and Steiglitz use a slightly different construction to obtain a bound of $16 \cdot n - 11 + 16 \cdot p$ (see Exercise 1.49).

Nevertheless, the bound of Theorem 1.1 is attainable as demonstrated by the following examples. First, consider a square of side width 2 that consists of the central four squares in a $2^n \times 2^n$ image (see Figure 1.27). Its quadtree has $16 \cdot n - 11$ nodes (see Exercise 1.50). Second, consider a curve that follows a vertical line through the center of a $2^n \times 2^n$ image. Now, make it a bit longer by making it intersect all of the pixels on either side of the vertical line (see Figure 1.28). As n increases, the total number of nodes in the quadtree approaches $8 \cdot p$ where $p = 2^n$ (see Exercise 1.51). A polygon having a number of nodes approaching $8 \cdot p$ can be constructed in a similar manner by approximating a square in the center of the image whose side is one-fourth the side of the image (see Exercise 1.52). In fact, it has been shown by Hunter [Hunt78] that $O(p+n)$ is a least upper bound on the number of nodes in a quadtree corresponding to a polygon (see Exercise 1.53).

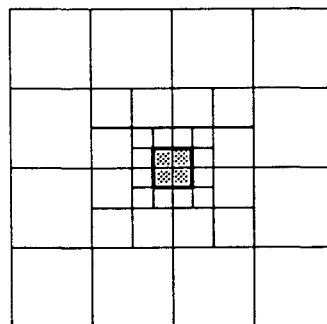


Figure 1.27 Example quadtree with $16 \cdot n - 11$ nodes

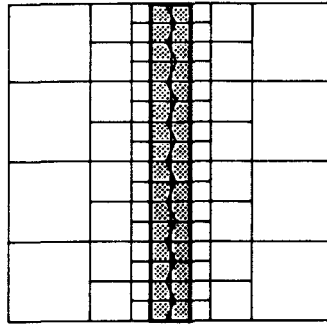


Figure 1.28 Example quadtree with approximately $8 \cdot p$ nodes

Theorem 1.1 can be recast by measuring the perimeter p in terms of the length of a side of the image in which the polygon is embedded—i.e., for a $2^n \times 2^n$ image $p = p' \cdot 2^n$. Thus the value of the perimeter no longer depends on the resolution of the image. Restating Theorem 1.1 in terms of p' results in a quadtree having $O(p' \cdot 2^n + n)$ nodes. This leads to the following important corollary:

Corollary 1.1 The maximum number of nodes in a quadtree corresponding to an image is directly proportional to the resolution of the image. \square

The significance of Corollary 1.1 is that when using quadtrees, increasing the image resolution leads to a linear growth in the number of nodes. This is in contrast to the binary array representation where doubling the resolution leads to a quadrupling of the number of pixels.

Since in most practical cases the perimeter, p , dominates the resolution, n , the results of Theorem 1.1 are usually interpreted as stating that the number of nodes in a quadtree is proportional to the perimeter of the regions contained therein.¹⁰ Meagher [Meag80] has shown that this theorem also holds for three-dimensional data (i.e., for polyhedra represented by octrees) when the perimeter is replaced by the surface area. The perimeter and the surface area correspond to the size of the boundary of the polygon and polyhedron—that is, in two and three dimensions, respectively. In d dimensions this result can be stated as follows:

Theorem 1.2: The size of a d -dimensional quadtree of a d -dimensional polyhedron is proportional to the sum of the resolution and the size of the boundary of the object. \square

¹⁰ Of course, the storage used by runlength codes is also proportional to the perimeter of the regions. However, runlength codes do not facilitate access to different parts of the regions (i.e., they have poor spatial indexing properties).

Aside from their implications on the storage requirements, Theorems 1.1 and 1.2 also directly affect the analysis of the execution time of algorithms. In particular, most algorithms that execute on a quadtree representation of an image instead of an array representation have an execution time proportional to the number of blocks in the image rather than the number of pixels. In its most general case, this means that the application of a quadtree algorithm to a problem in d -dimensional space executes in time proportional to the analogous array-based algorithm in the $(d-1)$ -dimensional space of the surface of the original d -dimensional image. Thus quadtrees are somewhat like dimension-reducing devices.

Theorem 1.2 assumes that the image consists of a polyhedron. Walsh [Wals85] lifts this restriction and obtains a weaker complexity bound. Assuming an image of resolution n and measuring the perimeter, say p , in terms of the number of border pixels, he proves that the total number of nodes in a d -dimensional quadtree is less than or equal to $4 \cdot n \cdot p$. Furthermore he shows that the number of black nodes is less than or equal to $(2^d - 1) \cdot n \cdot p / d$.

The complexity measures discussed above do not explicitly reflect the fact that the amount of space occupied by a quadtree corresponding to a region is extremely sensitive to its orientation (i.e., where it is partitioned). For example, in Dyer's experiment, the number of nodes required for the arbitrary placement of a square of size $2^m \times 2^m$ at any position in a $2^n \times 2^n$ image ranged between $4 \cdot (n-m) + 1$ and $4 \cdot p + 16 \cdot (n-m) - 27$, with the average being $O(p+n-m)$. Clearly shifting the image within the space in which it is embedded can reduce the total number of nodes. The problem of finding the optimal position for a quadtree can be decomposed into two parts. First, we must determine the optimal grid resolution and, second, the partition points.

Grosky and Jain [Gros83] have shown that for a region such that w is the maximum of its horizontal and vertical extent (measured in pixel widths) and $2^{n-1} < w \leq 2^n$, the optimal grid resolution is either n or $n+1$. In other words embedding the region in a larger area than $2^{n+1} \times 2^{n+1}$ and shifting it around will not result in fewer nodes. Using similar reasoning, it can be shown that translating a region by 2^k pixels in any direction does not change the number of black or white blocks of size less than $2^k \times 2^k$ [Li82].

Armed with the above results, Li, Grosky, and Jain [Li82] developed the following algorithm that treats the image as a binary array and finds the configuration of the region in the image so that its quadtree requires a minimum number of nodes. First, enlarge the image to be $2^{n+1} \times 2^{n+1}$, and place the region within it so that the region's northernmost and westernmost pixels are adjacent to the northern and western borders, respectively, of the image. Next apply successive translations to the image of magnitude power of two in the vertical, horizontal, and corner directions and keep count of the number of leaf nodes required. Initially 2^{2n+2} leaf nodes are necessary. The following is a more precise statement of the algorithm:

1. Attempt to translate the image by (x,y) where x and y correspond to unit translations in the horizontal and vertical directions, respectively. Each of x and y takes on the values 0 or 1.

2. For the result of each translation in step 1, construct a new array at one-half the resolution. Each entry in the new array corresponds to a 2×2 block in the translated array. For each entry in the new array that corresponds to a single color (not gray) 2×2 block in the translated array, decrement the leaf node count by 3.
3. Recursively apply steps 1 and 2 to each result of steps 1 and 2. This process stops when no single-color 2×2 block is found in step 2 (i.e., they are all gray) or if the new array is a 1×1 block. Record the total translation and the minimum leaf node count.

Step 2 makes use of the property that for a translation of 2^k , there is a need to check only if single-color blocks of size $2^k \times 2^k$ or more are formed. In fact, because of the recursion, at each step we check only for the formation of blocks of size $2^{k+1} \times 2^{k+1}$. Note that the algorithm tries every possible translation since any integer can be decomposed into a summation of powers of two (i.e., use its binary representation). In fact this is why a translation of (0,0) is part of step 1. Although the algorithm computes the positioning of the quadtree with the minimum number of leaf nodes, it is also the positioning of the quadtree with the minimum total number of nodes since the number of nonleaf nodes in a quadtree of T leaf nodes is $(T-1)/3$.

As an example of the algorithm, consider the region given in Figure 1.29a whose block decomposition is shown in Figure 1.29b. Its quadtree requires 52 leaf nodes. The first step is to enlarge the image, place the region in the upper left corner, and form the array (Figure 1.30). The optimal positioning is such that Figure 1.30 is shifted 7 units in the horizontal direction and 3 units in the vertical direction. This corresponds to a sequence of translations (1,1), (1,1), and (1,0). The intermediate translated arrays are shown in Figure 1.31. All gray nodes in the translated arrays are labeled with a 'G' while black nodes are shaded. The optimal quadtree contains 46 leaf nodes and is given in Figure 1.32.

Now let us trace the algorithm as it applies the optimal sequence of translations, in more detail. Initially the leaf node count is 256. A translation of (1,1) leads to Figure 1.31a where 58 of the array entries correspond to single-color 2×2 blocks in the translated array. The leaf node count is decremented by $58 \cdot 3 = 174$, resulting in

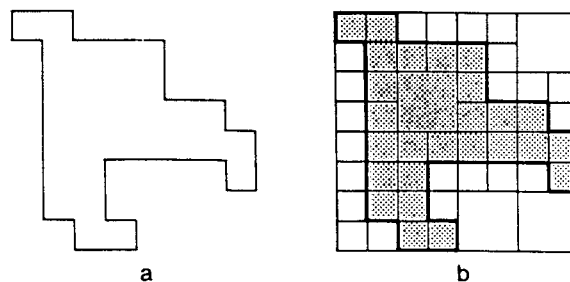


Figure 1.29 Example (a) image and (b) its block decomposition used to demonstrate the optimal positioning process

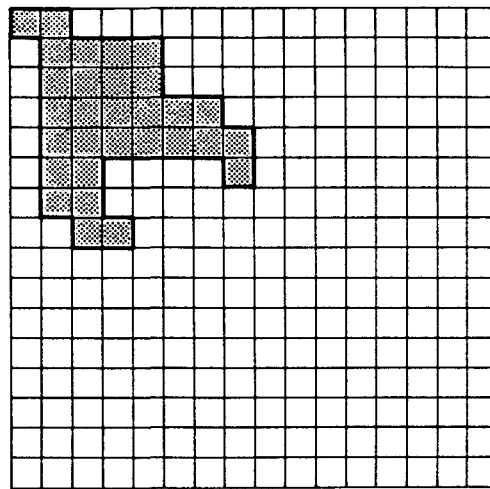


Figure 1.30 The array corresponding to the image in Figure 1.29 prior to the start of the optimal positioning process

82. The next translation of (1,1) leads to Figure 1.31b, where 11 of the array entries correspond to single-color 2×2 blocks. Therefore $11 \cdot 3 = 33$ is subtracted from 82, and the leaf node count is now 49. The final translation of (1,0) leads to Figure 1.31c, where only one of the array entries corresponds to a single-color 2×2 block in the translated array. Decrementing the leaf node count results in 46 nodes, and the process terminates. Of course, we have failed to describe the remaining $4^n - 3$ translations that were also attempted.

Despite trying all possible translations, the algorithm is quite efficient. The key is that for each translation, only the blocks whose motion can lead to space saving need to be considered. This is a direct consequence of the property that a translation of 2^k does not change the number of blocks of size less than $2^k \times 2^k$. For an image that has been enlarged to fit in a $2^{n+1} \times 2^{n+1}$ array, the algorithm will have a maximum depth of recursion of n . Since at each level of recursion we need an array at half the resolution of the previous level, the total amount of space required is $(4/3) \cdot 2^{2n+2}$.

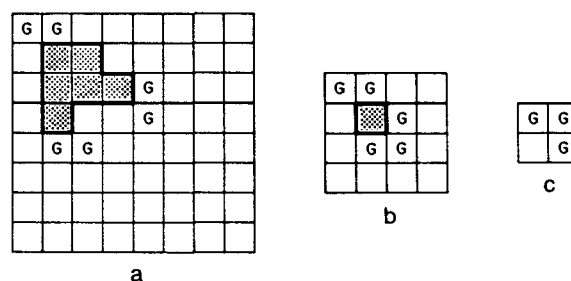


Figure 1.31 The successive translated arrays at half-resolution after application of (a) (1,1) and (b) (1,1), and (c) (1,0) to the original image array of Figure 1.30

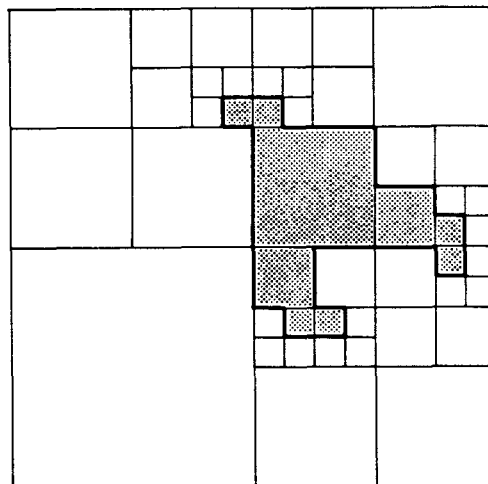


Figure 1.32 Optimal positioning of the quadtree of Figure 1.29

The basic computational task of the algorithm is to count 2×2 blocks of a single color. It can be shown that $4 \cdot n \cdot 2^{2n+2}$ array elements are examined in this process (see Exercise 1.63). Thus the algorithm uses $O(2^{2n})$ space and takes $O(n \cdot 2^{2n})$ time. Nevertheless experiments with typical images show that the algorithm has little effect (e.g., [Same84c]).

Exercises

- 1.43. Consider the arbitrary placement of a square of size $2^m \times 2^m$ at any position in a $2^n \times 2^n$ image. Prove that in the best case $4 \cdot (n-m) + 1$ nodes are required, while the worst case requires $4 \cdot p + 16 \cdot (n-m) - 27$ nodes. How many of these nodes are black and white, assuming that the square is black? Prove that on the average, the number of nodes that is required is $O(p+n-m)$.
- 1.44. What are the worst-case storage requirements of storing an arbitrary rectangle in a quadtree corresponding to a $2^n \times 2^n$ image? Give an example of the worst case and the number of nodes it requires.
- 1.45. Assume that the probability of a particular pixel's being black is one-half and likewise for being white. Given a $2^n \times 2^n$ image represented by a quadtree, what is the expected number of nodes, say $E(n)$, in the quadtree? Also compute the expected number of black, white, and gray nodes.
- 1.46. Suppose that instead of knowing the probability a particular pixel is black or white, we know the percentage of the total pixels in the image that are black. Given a $2^n \times 2^n$ image represented by a quadtree, what is the expected number of nodes in the quadtree?
- 1.47. The proof of Theorem 1.1 and the subsequent discussion raise the question of how N squares should be arranged so that each is intersected by a curve of minimum length extending to the outside of the squares on each end. Such a configuration leads to a minimal curve in the sense that it has a maximal ratio of squares to length. For which value of N is this ratio the smallest?
- 1.48. Try to prove that the upper bound of Theorem 1.1 can be tightened to be $a \cdot n + b + 8 \cdot p$ where a and b are constants.

- 1.49. Decompose the polygon used in the proof of Theorem 1.1 into a sequence of curves in the following manner. Mark the points where G enters and exits each square of side width d . Choose one of these points, say P , and define the first curve in G as extending from P until four squares have been intersected and a crossing is made into a different fifth square. This is the starting point for another curve in G that intersects four new squares, not counting those intersected by any previous curve. Prove that all of the curves, except for the last one, must be at least of length d . Using this result, prove that the upper bound on the number of nodes in the quadtree is $16 \cdot n - 11 + 16 \cdot p$.
- 1.50. Prove that the quadtree corresponding to a square of side width 2 consisting of the central four squares in a $2^n \times 2^n$ image has $16 \cdot n - 11$ nodes (see Figure 1.27).
- 1.51. Take a curve that follows a vertical line through the center of a $2^n \times 2^n$ image and lengthen it slightly by making it intersect all of the pixels on either side of the vertical line (see Figure 1.28). Prove that as n increases, the total number of nodes in the quadtree approaches $8 \cdot p$ where $p = 2^n$.
- 1.52. Using a technique analogous to that used in Exercise 1.51, construct a polygon of perimeter p by approximating a square in the center of the image whose side is one-fourth the side of the image. Prove that its quadtree has approximately $8 \cdot p$ nodes.
- 1.53. Prove that $O(p+n)$ is a least upper bound on the number of nodes in a quadtree corresponding to a polygon. Assume that $p \leq 2^{2^n}$ (i.e., the number of pixels in the image). Equivalently the polygon boundary can touch all of the pixels in the most trivial way but can be no longer. Decompose your proof into two parts depending on whether p is greater than $4 \cdot n$.
- 1.54. Can you prove that for an arbitrary quadtree (not necessarily a polygon), the number of nodes doubles as the resolution is doubled?
- 1.55. Derive a result analogous to Theorem 1.1 for a three-dimensional polyhedron represented as an octree. In this case the perimeter corresponds to the surface area.
- 1.56. Prove Theorem 1.2.
- 1.57. Assuming an image of resolution n and measuring the perimeter, say p , in terms of the number of border pixels, prove that the total number of nodes in a d -dimensional quadtree is less than or equal to $4 \cdot n \cdot p$.
- 1.58. Assuming an image of resolution n and measuring the perimeter, say p , in terms of the number of border pixels, prove that the total number of black nodes in a d -dimensional quadtree is less than or equal to $(2^d - 1) \cdot n \cdot p / d$.
- 1.59. How tight are the bounds obtained in Exercises 1.57 and 1.58 for the number of nodes in a d -dimensional quadtree for an arbitrary region? Are they realizable?
- 1.60. Prove that for a region such that w is the maximum of its horizontal and vertical extent (measured in pixel widths) and $2^{n-1} < w \leq 2^n$, the optimal grid resolution is either n or $n+1$.
- 1.61. Prove that translating a region by 2^k pixels in any direction does not change the number of black or white blocks of size less than $2^k \times 2^k$.
- 1.62. Can you formally prove that the method described in the text does indeed yield the optimal quadtree?
- 1.63. Prove that $4 \cdot n \cdot 2^{2n+2}$ array elements are examined in the process of constructing the optimal quadtree.
- 1.64. How would you find the optimal bintree?



US005263136A

United States Patent [19]
DeAguiar et al.

[11] **Patent Number:** **5,263,136**
 [45] **Date of Patent:** **Nov. 16, 1993**

- [54] **SYSTEM FOR MANAGING TILED IMAGES USING MULTIPLE RESOLUTIONS**
 5,020,003 5/1991 Moshenberg 395/164
 5,150,462 9/1992 Takeda et al. 395/166
- [75] **Inventors:** John R. DeAguiar, Sebastopol; Ross M. Larkin, Rollings Hills, both of Calif.
- [73] **Assignee:** Optographics Corporation, San Diego, Calif.
- [21] **Appl. No.:** 694,416
- [22] **Filed:** Apr. 30, 1991
- [51] **Int. Cl.⁵** G06F 15/20
- [52] **U.S. Cl.** 395/164; 345/201
- [58] **Field of Search** 395/162, 164, 166, 128-130; 340/798, 799; 358/452, 455

Primary Examiner—Dale M. Shaw
Assistant Examiner—Kee M. Tung
Attorney, Agent, or Firm—Knobbe, Martens, Olson & Bear

[57] **ABSTRACT**

An image memory management system for tiled images. The system defines an address space for a virtual memory that includes an image data cache and a disk. An image stack for each source image is stored as a full resolution image and a set of lower-resolution subimages. Each tile of an image may exist in one or more of five different states as follows: uncompressed and resident in the image data cache, compressed and resident in the image data cache, uncompressed and resident on disk, compressed and resident on disk and not loaded but re-creatable using data from higher-resolution image tiles.

- [56] **References Cited**
- U.S. PATENT DOCUMENTS**
- Re. 31,200 4/1983 Sukonick et al. 395/162
- 4,878,183 10/1989 Ewart 395/128 X
- 4,920,504 4/1990 Sawada et al. 395/166
- 4,951,230 8/1990 Dalrymple et al. 395/166

17 Claims, 39 Drawing Sheets

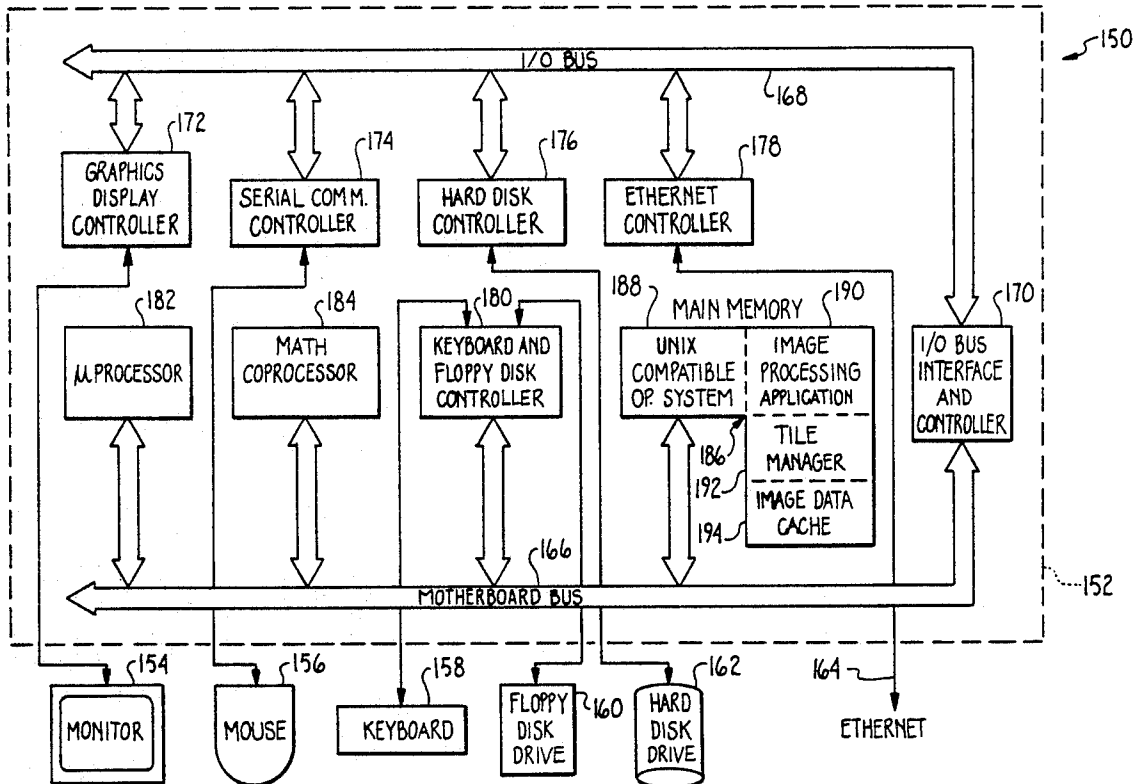
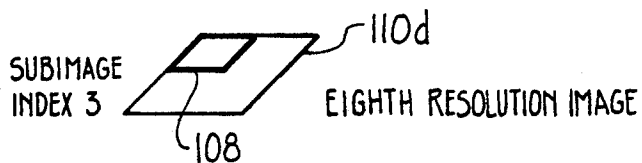
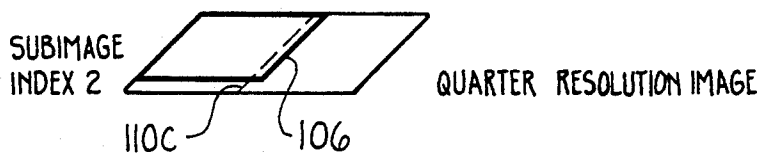
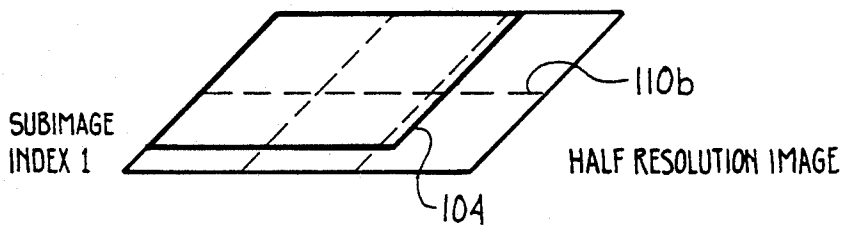
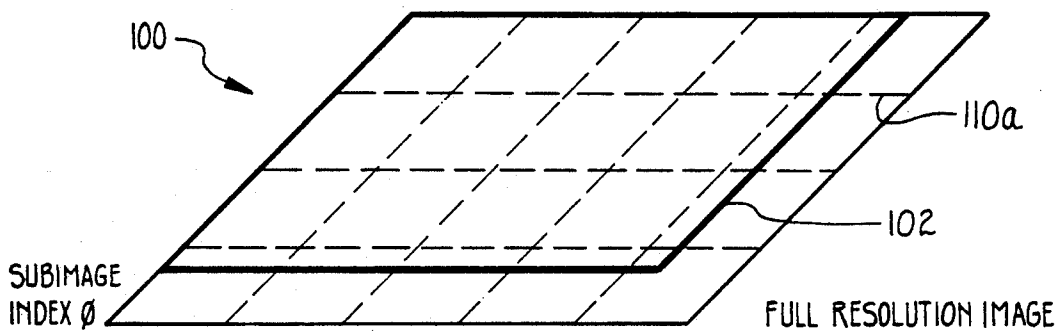
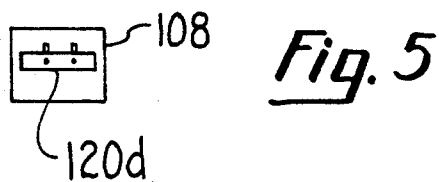
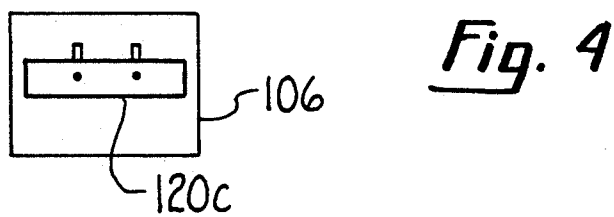
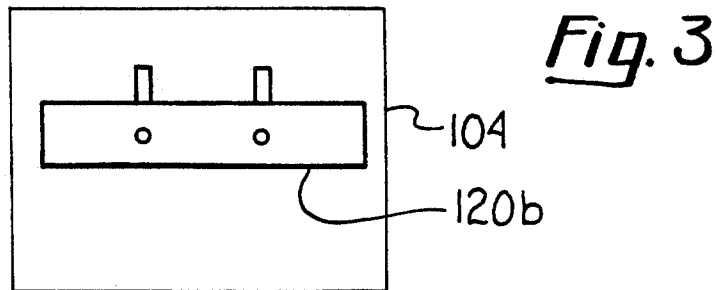
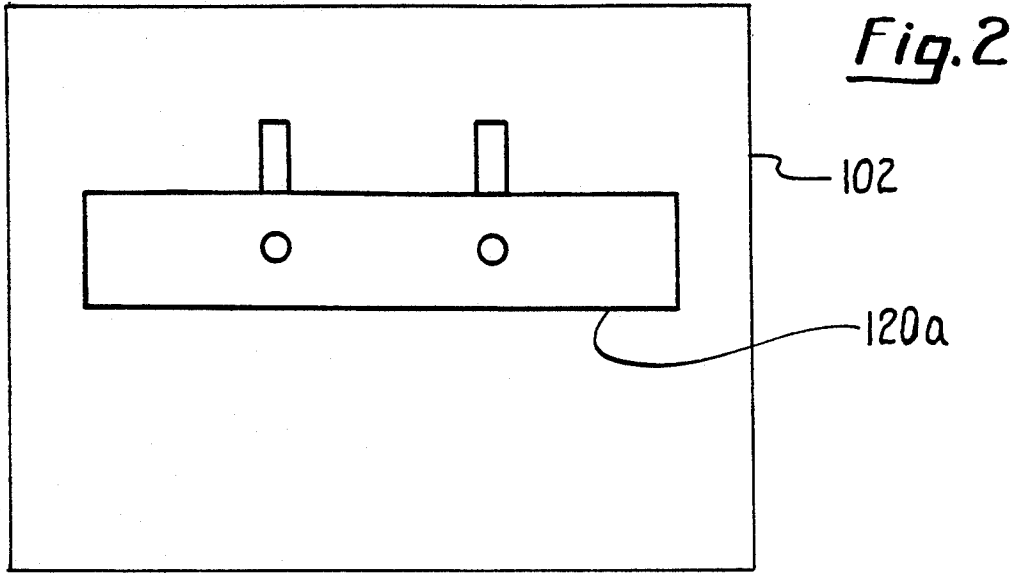


Image Stack

Fig. 1





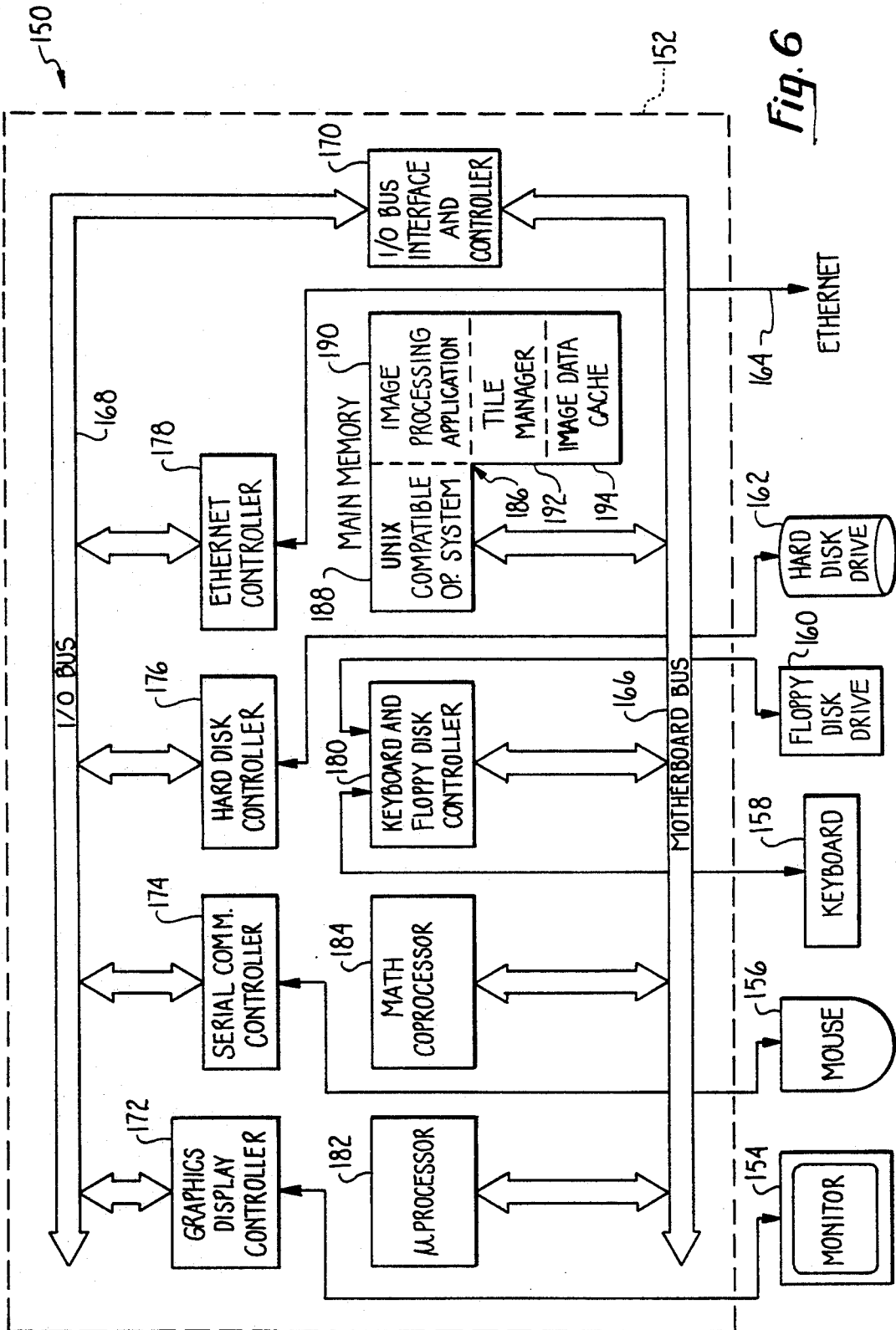


Fig. 6

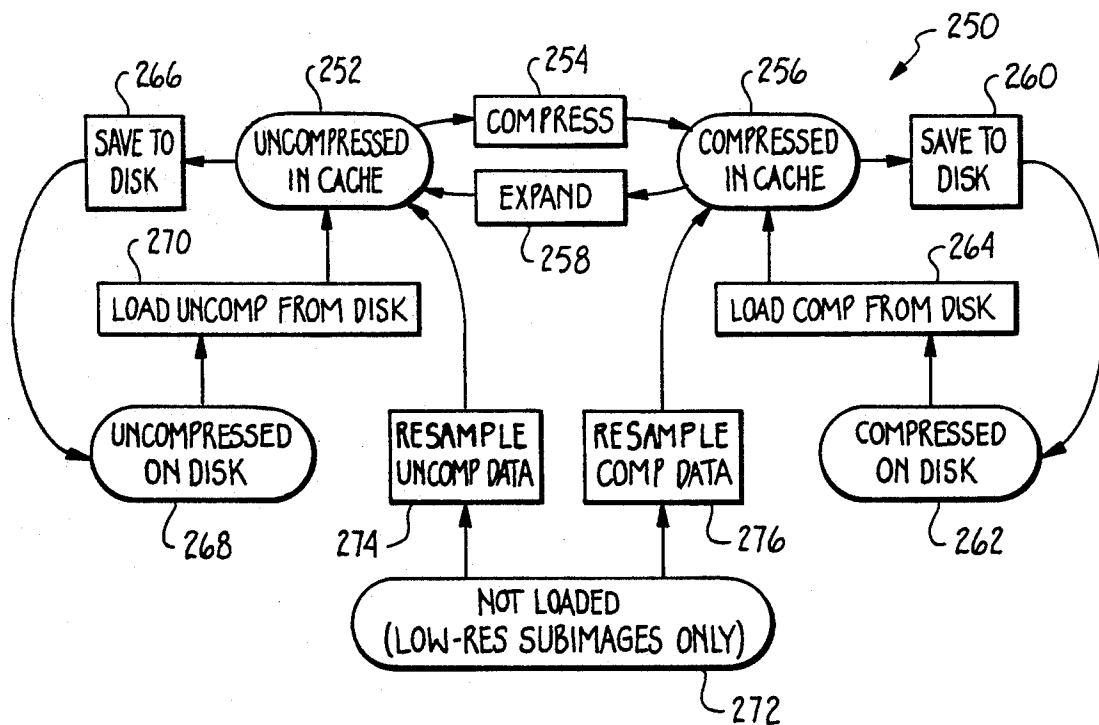
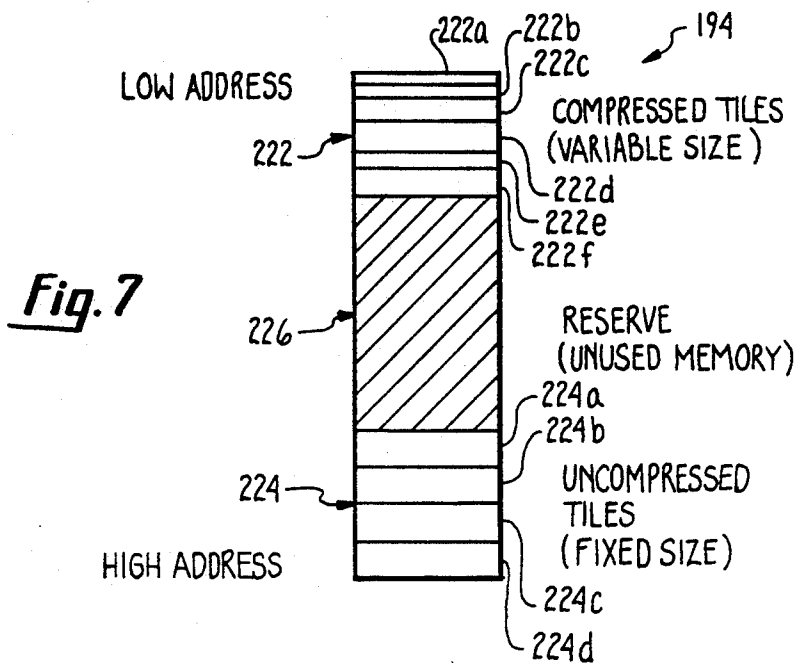


Fig. 9A Document Information Structure

↗ 300

SELF-REFERENCE TO DOCUMENT HANDLE	<u>302</u>	"OVERVIEWS INVALID" FLAG	<u>304</u>	CACHE IMAGE COMPRESSION ALGORITHM	<u>306</u>
IMAGE COLOR TYPE	<u>308</u>	BITS PER IMAGE PIXEL	<u>310</u>	TILE SIZE INFORMATION	<u>312</u>
NUMBER OF SUBIMAGES IN DOC	<u>314</u>	INPUT FILE INFO	<u>316</u>	OUTPUT FILE INFO	<u>318</u>
LIST OF SUBIMAGE HEADERS					<u>320</u>

Fig. 9B

↗ 321

POINTER TO TILE HEADERS	<u>322</u>	POINTER TO TILE DIRECTORY	<u>324</u>	SUBIMAGE WIDTH AND HEIGHT	<u>326</u>
NUMBER OF TILE ROWS & COLS IN SUBIMAGE	<u>328</u>	IMAGE STACK INDEX OF THIS SUBIMAGE	<u>330</u>	PIXEL RESOLUTION OF THIS SUBIMAGE	<u>332</u>
⋮					

Fig. 10 Tile Header

↗ 350

POINTER TO DOCUMENT CONTAINING THIS TILE	<u>352</u>	INDEX OF SUBIMAGE CONTAINING THIS TILE	<u>354</u>	ROW AND COLUMN INDICES OF TILE	<u>356</u>
STATUS INFORMATION		<u>358</u>	PRESERVE COUNT		<u>360</u>
LOCATION OF UNCOMPRESSED IMAGE DATA IN CACHE MEMORY		<u>362</u>	LOCATION OF COMPRESSED IMAGE DATA IN CACHE MEMORY		<u>364</u>
LOCATION OF UNCOMPRESSED IMAGE DATA ON DISK		<u>366</u>	LOCATION OF COMPRESSED IMAGE DATA ON DISK		<u>368</u>
LINK TO NEXT LESS RECENTLY USED TILE		<u>370</u>	LINK TO NEXT MORE RECENTLY USED TILE		<u>372</u>
NUMBER OF BYTES OF EXPANDED DATA IN TILE		<u>374</u>	NUMBER OF BYTES OF COMPRESSED DATA IN TILE		<u>376</u>

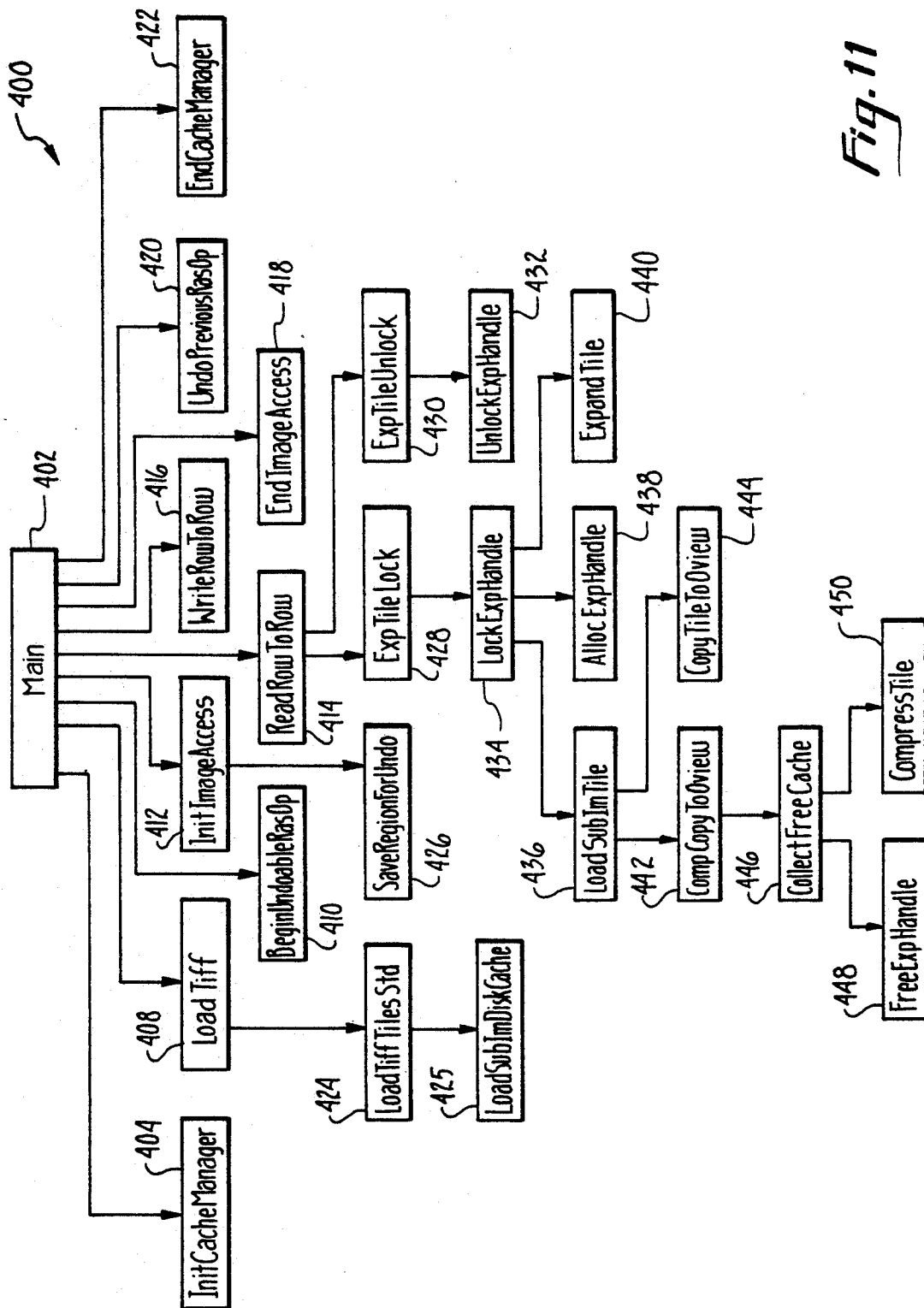
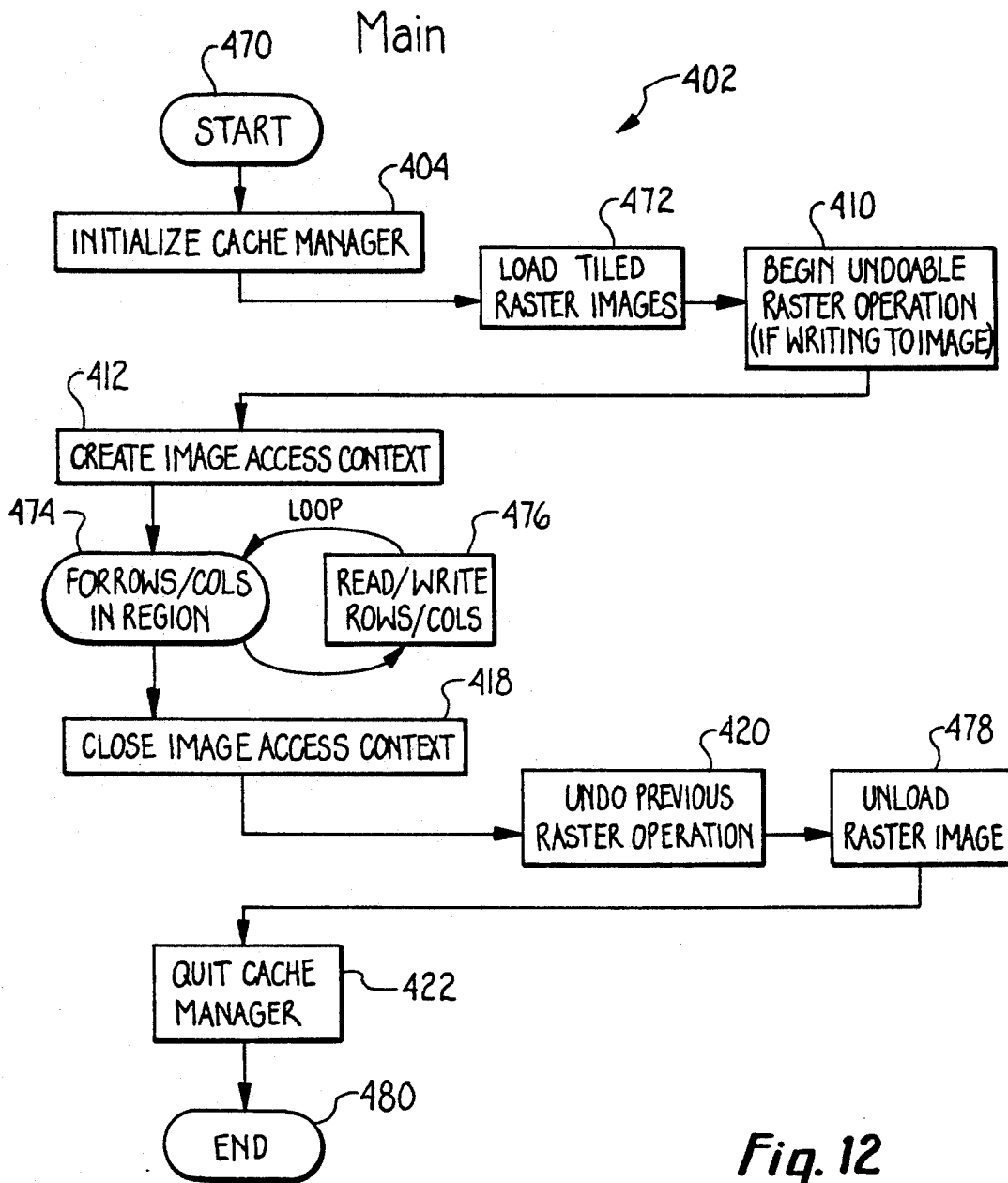
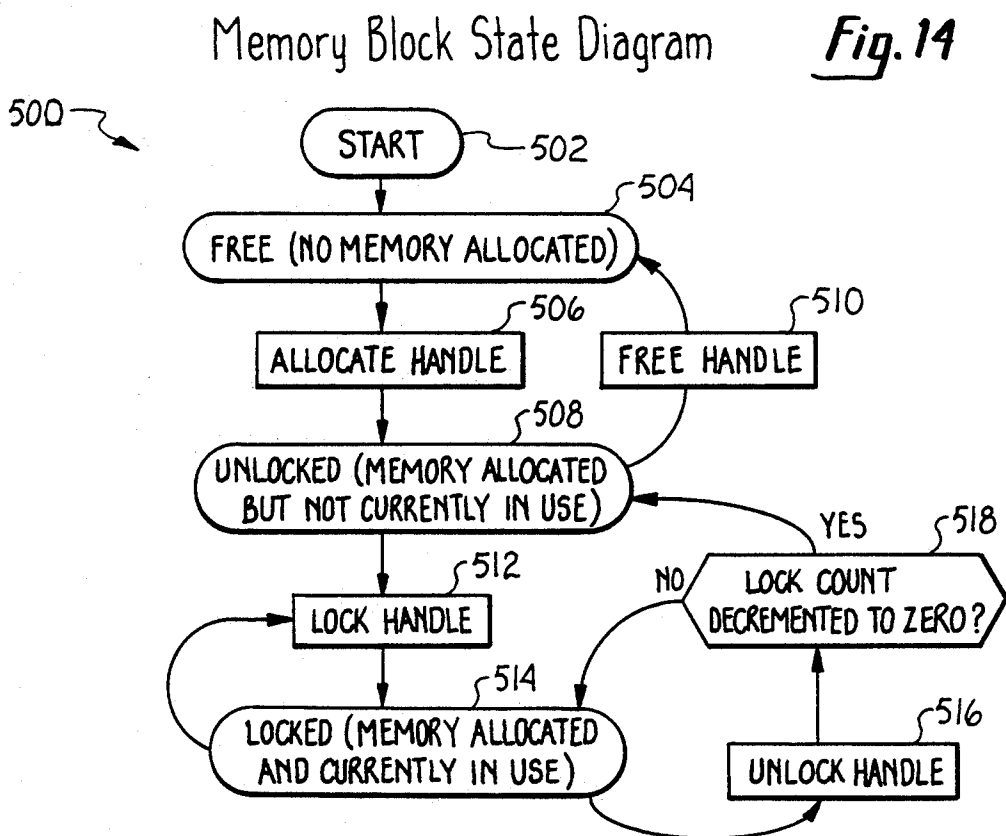
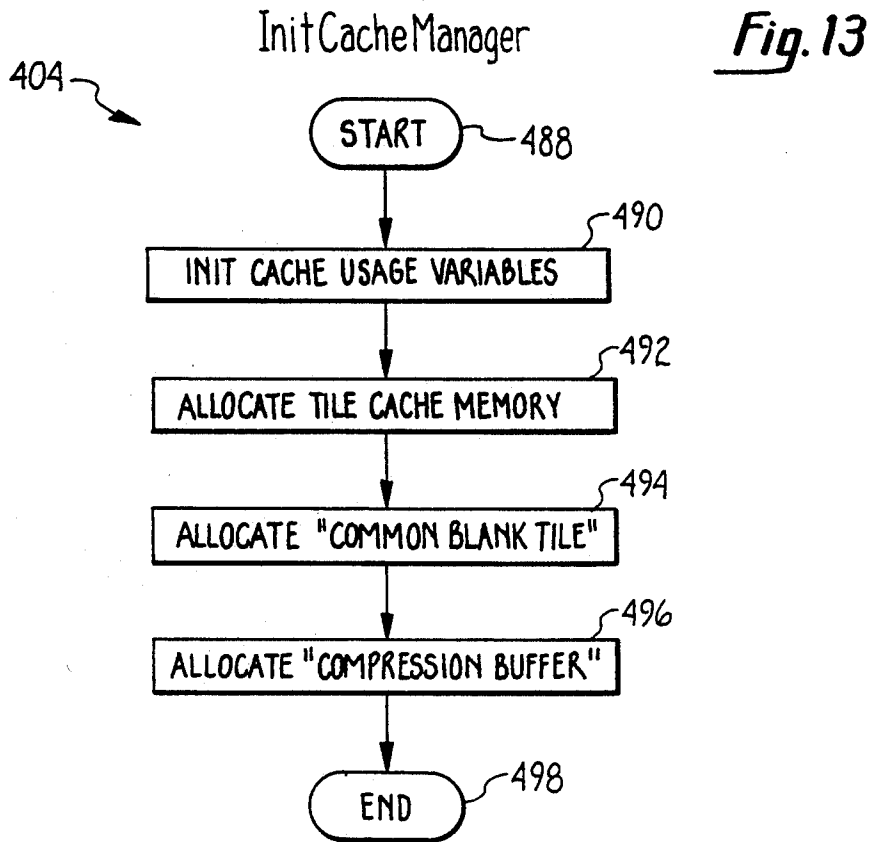


Fig. 11





Init Image Access

Fig. 15A

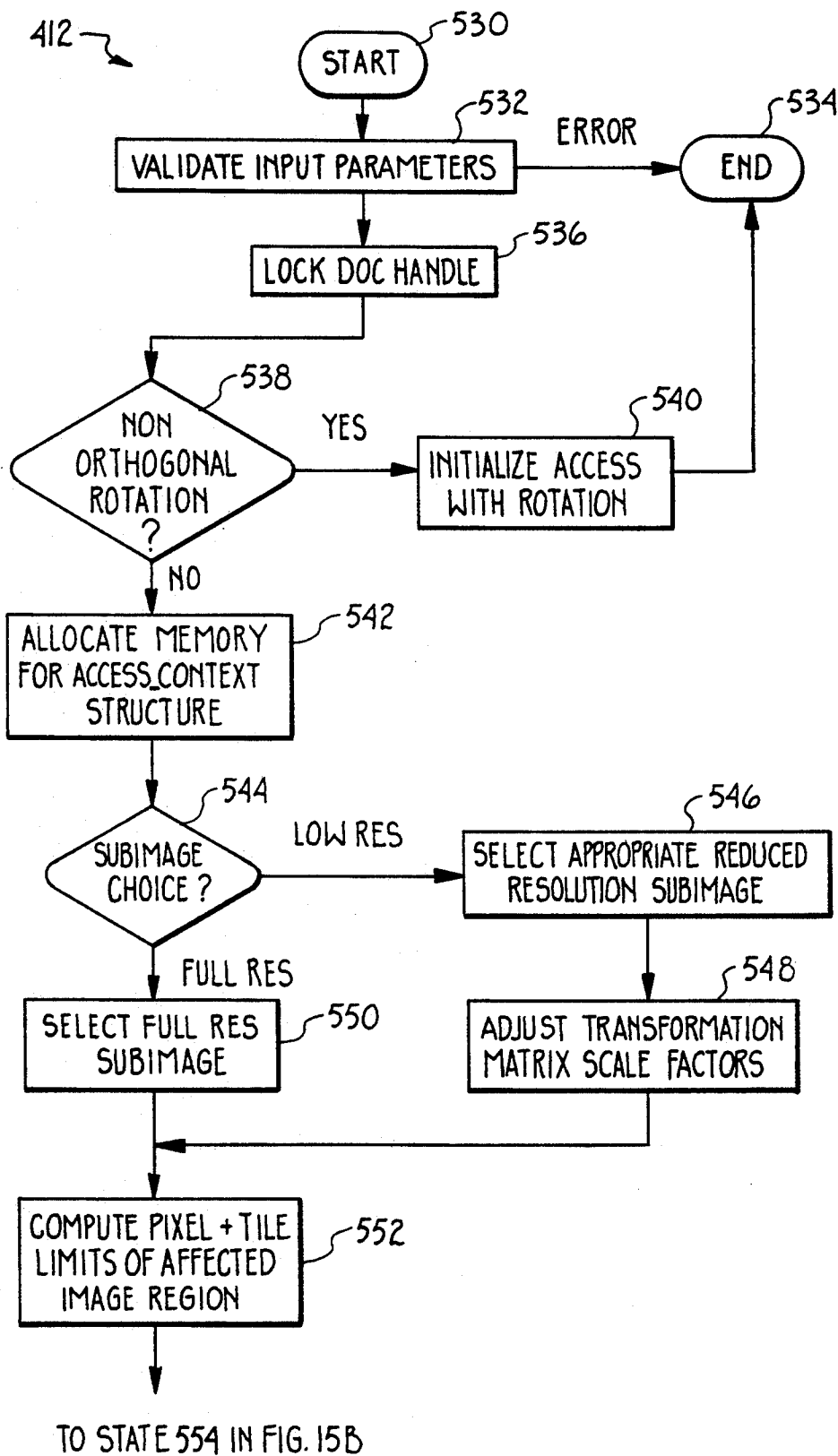


Fig. 15B

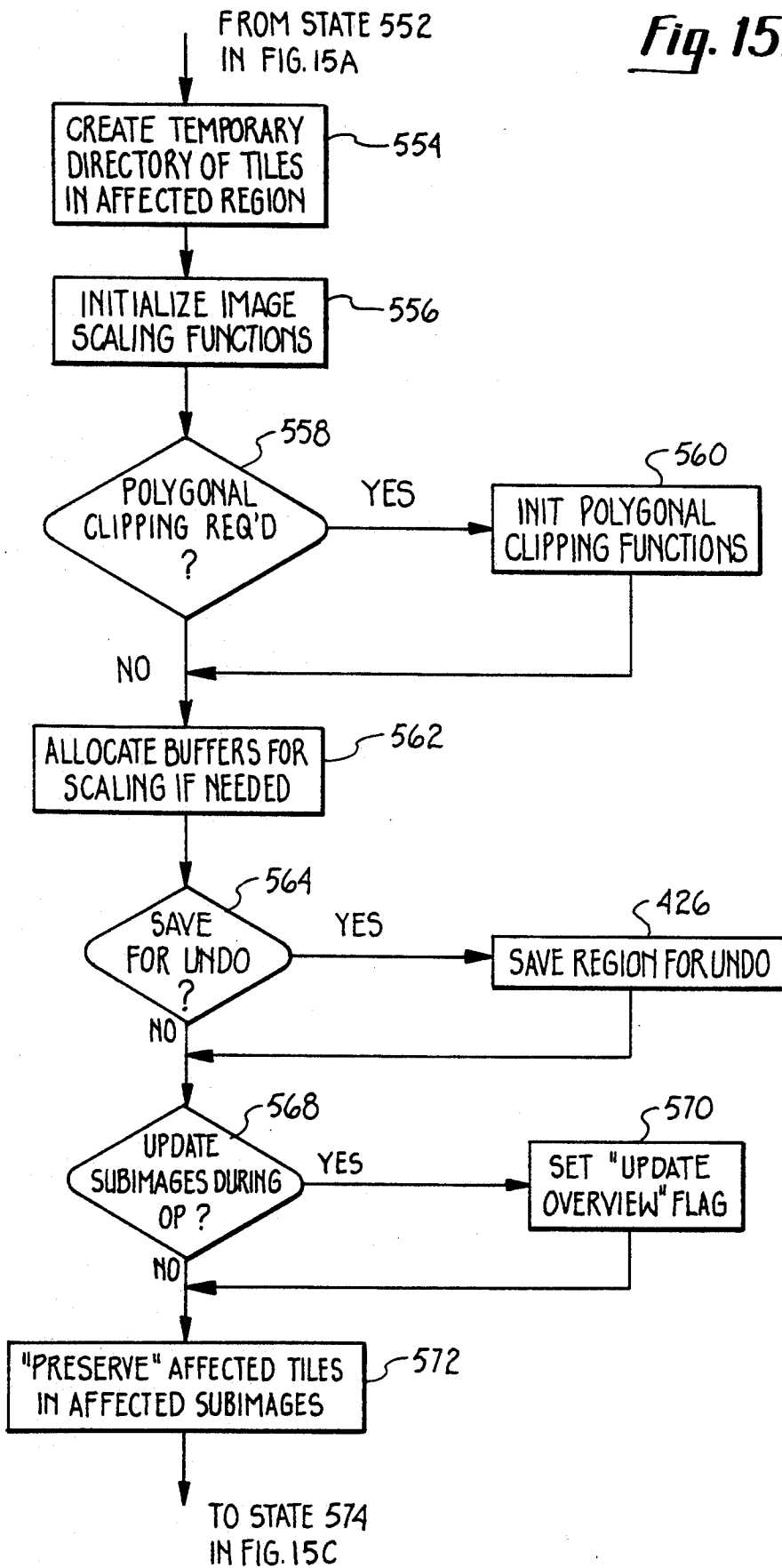


Fig. 15C

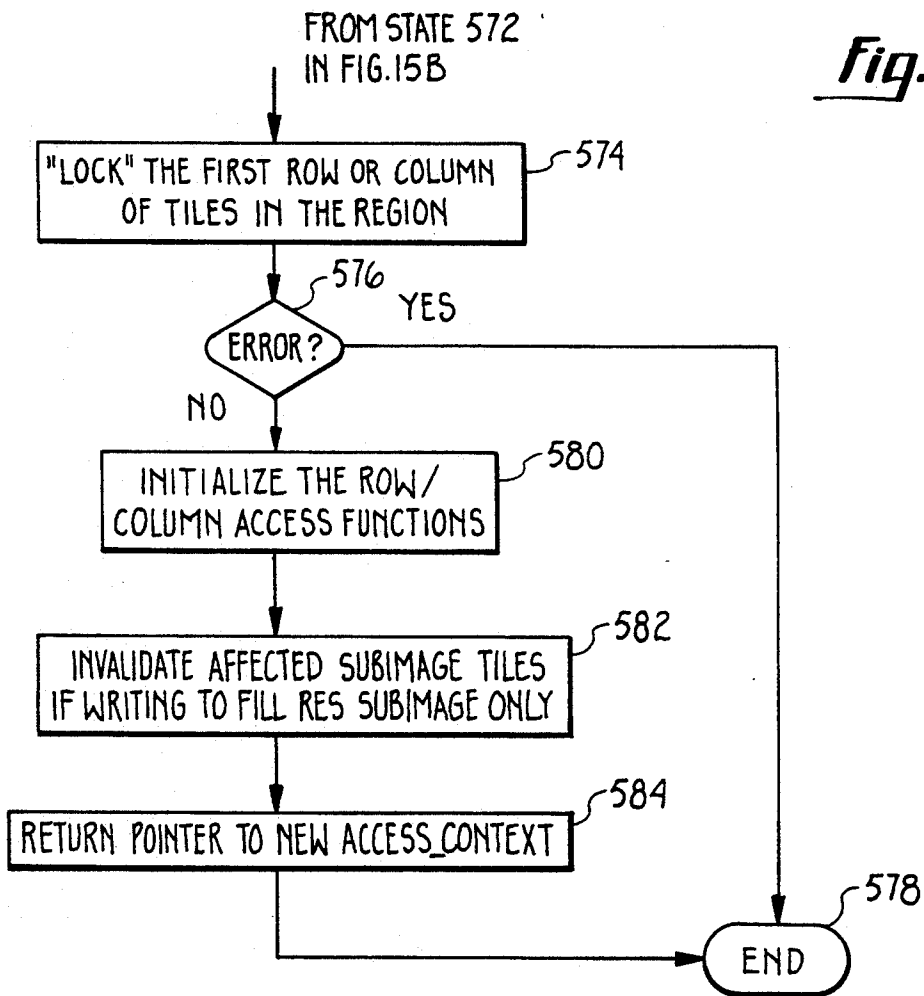


Fig. 16

600

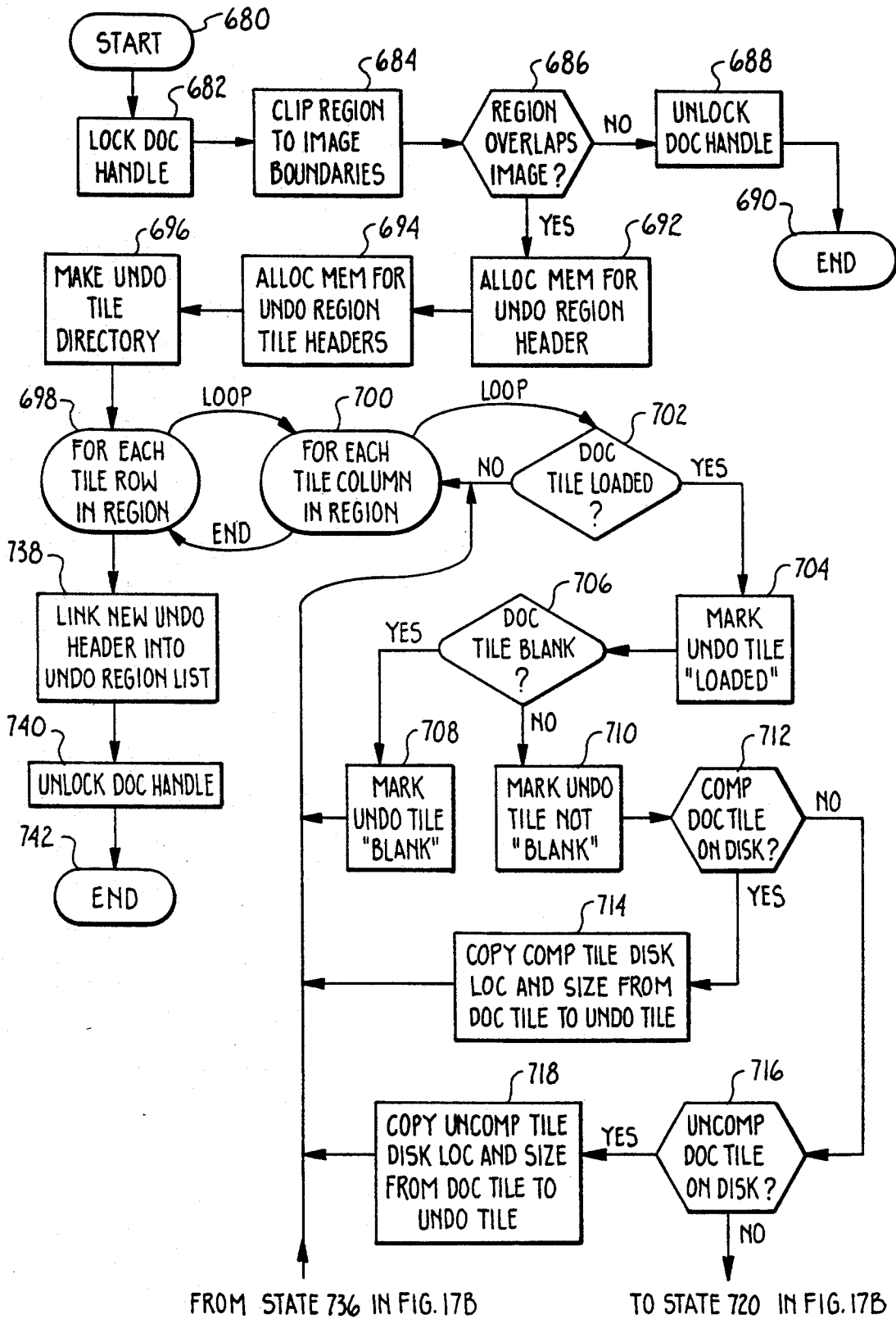
Access Context

POINTER TO <u>602</u> AFFECTED DOG	"SUBIMAGE CHOICE" OPTION VALUE <u>604</u>	INDEX OF <u>606</u> AFFECTED SUBIMAGE	ACCESS <u>608</u> QUANTUM	READ/WRITE <u>610</u> OPTION
BASIC ORTHOGONAL ROTATION VALUE <u>612</u>	PIXEL COMBINATION OPERATION <u>614</u>	SCALER TYPE OPERATION <u>616</u>	"UPDATE OVERVIEWS" FLAG <u>618</u>	
I/O BUFFER WIDTH & HEIGHT <u>620</u>	I/O BUFFER PITCH (BYTES/ROW) <u>622</u>	I/O BUFFER BIT OFFSET TO START OF RUN <u>624</u>		
ROWS PER STRIP (FOR "AQ_STRIP" ACCESS QUANTUM) <u>626</u>	NUMBER OF I/O BUFFER ROWS YET TO BE PROCESSED <u>628</u>			
POINTER TO ACCESS FUNCTION USED IN "SeqBufImageAccess" <u>630</u>	STEPPING DIRECTIONS FOR IMAGE ROW AND COLUMN INDICIES <u>632</u>			
POINTER TO POLYGON CLIPPING INFORMATION <u>634</u>	POINTER TO RASTER SCALING INFORMATION <u>636</u>			
POINTER TO UNCOMPRESSED DATA IN CURRENTLY LOCKED TILES <u>638</u>	POINTER TO REGION TILE DIRECTORY <u>640</u>	NEXT IMAGE ROW & COLUMN TO BE ACCESSED <u>642</u>		
TERMINAL ROW & COLUMN OF ACCESS REGION <u>644</u>	UNCLIPPED EXTENT OF ACCESS REGION <u>646</u>	CLIPPED EXTENT OF ACCESS REGION <u>648</u>		
CLIPPED IMAGE BUFFER BIT OFFSET AND LENGTH <u>650</u>	NUMBER OF TILE ROWS & COLS IN ACCESS REGION <u>652</u>	ROW & COLUMN OF CURRENTLY LOCKED TILES <u>654</u>		
IMAGE ROW & COL AT ORIGIN OF FIRST TILE IN ACCESS REGION <u>656</u>	NUMBER OF I/O BUFFER ROWS HELD OVER FOR NEXT STRIP <u>658</u>			
POINTER TO IMAGE TILING/UNTILING BUFFER <u>660</u>	NUMBER OF BYTES IN TILING/UNTILING BUFFER <u>662</u>	BIT OFFSET FOR TILING/UNTILING BUFFER <u>664</u>		
ACCESS TRANSFORMATION MATRIX <u>666</u>				

426 ↗

SaveRegionForUndo

Fig. 17A



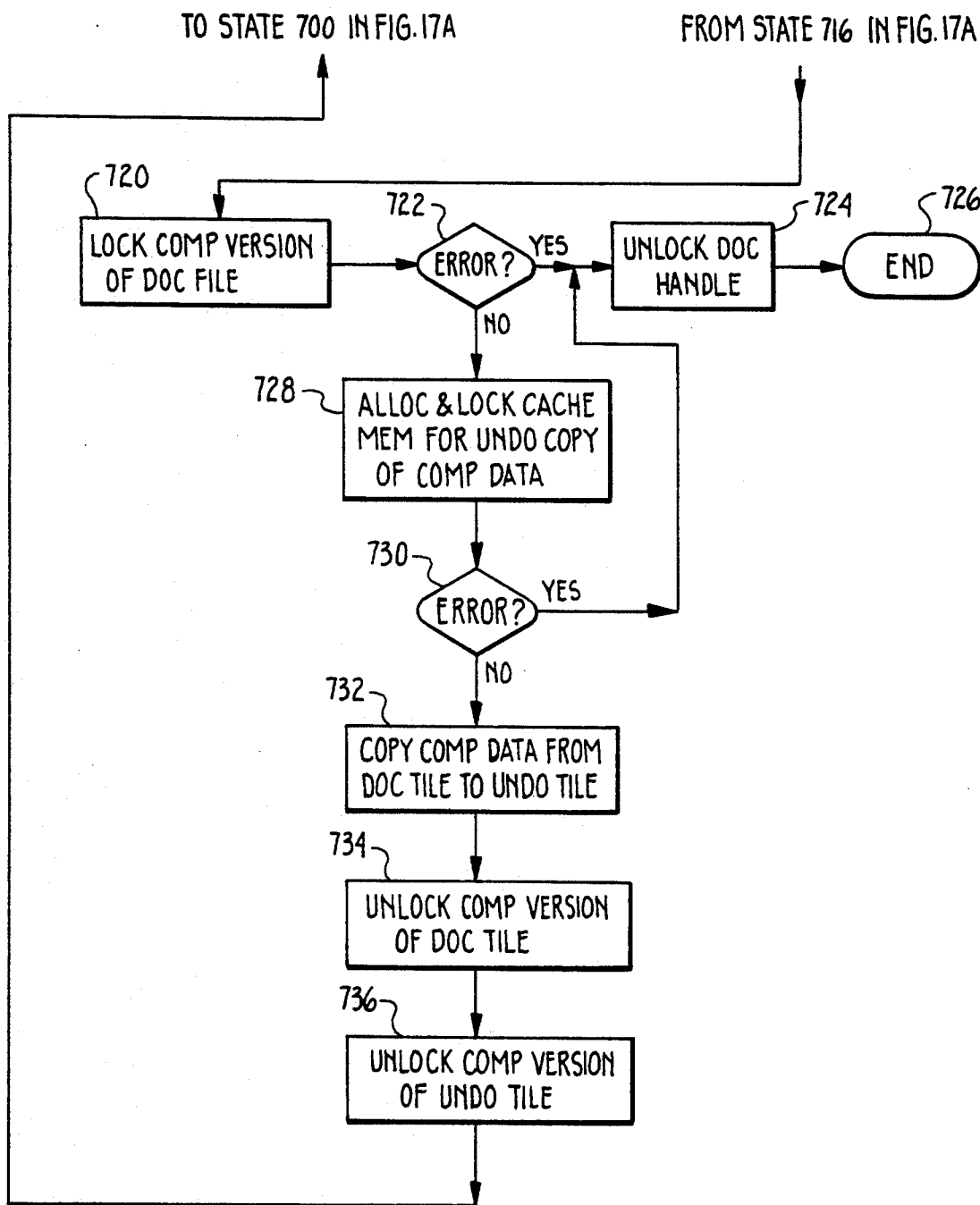
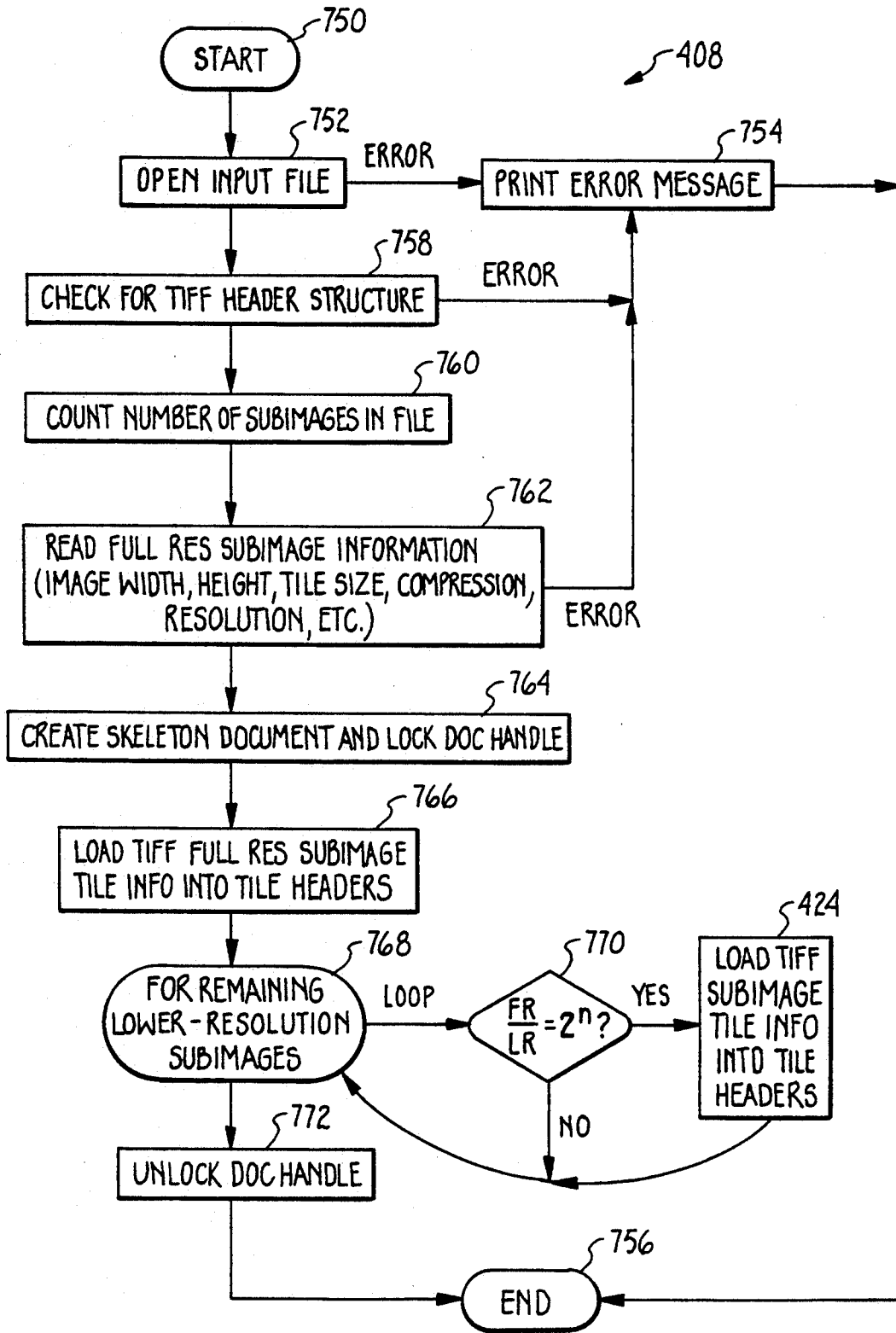


Fig. 17B

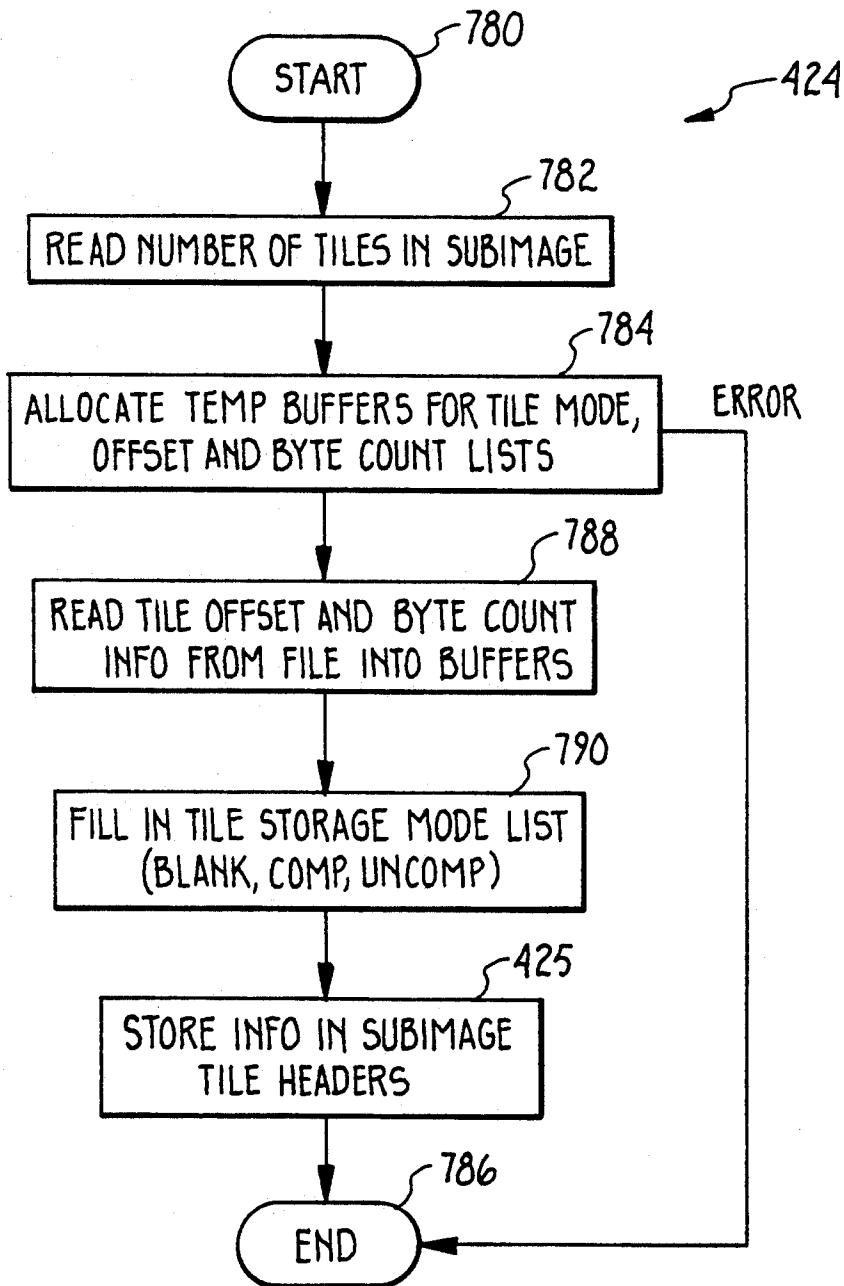
LoadTiff

Fig. 18



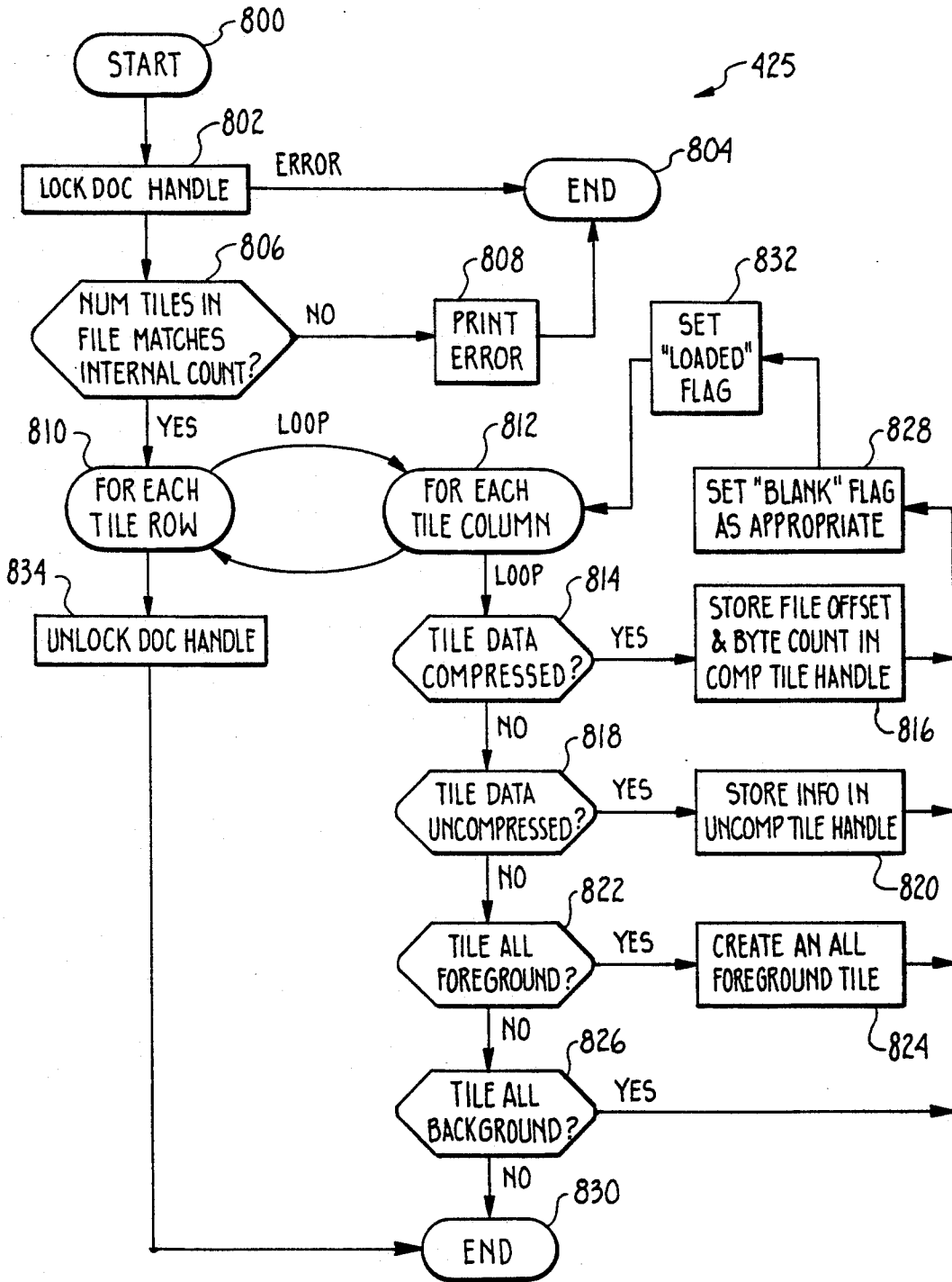
LoadTiff Tiles Std

Fig. 19



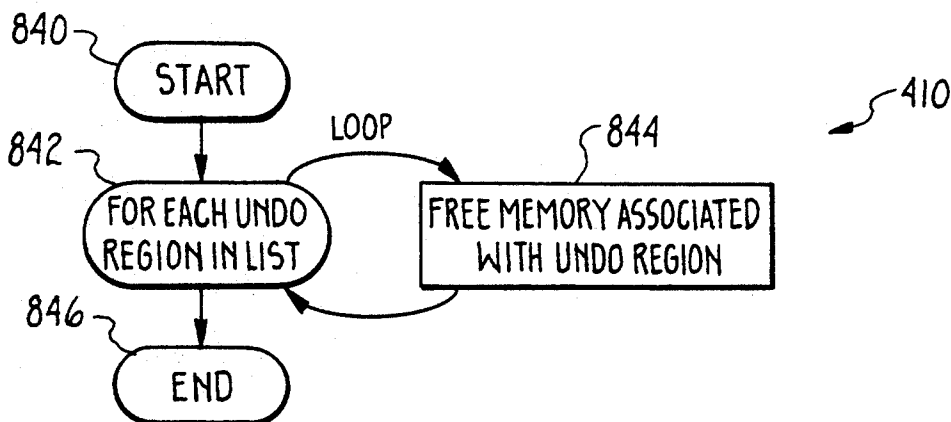
LoadSubImDiskCache

Fig. 20



BeginUndoableRasOp

Fig. 21



Read Row To Row

Fig. 22A

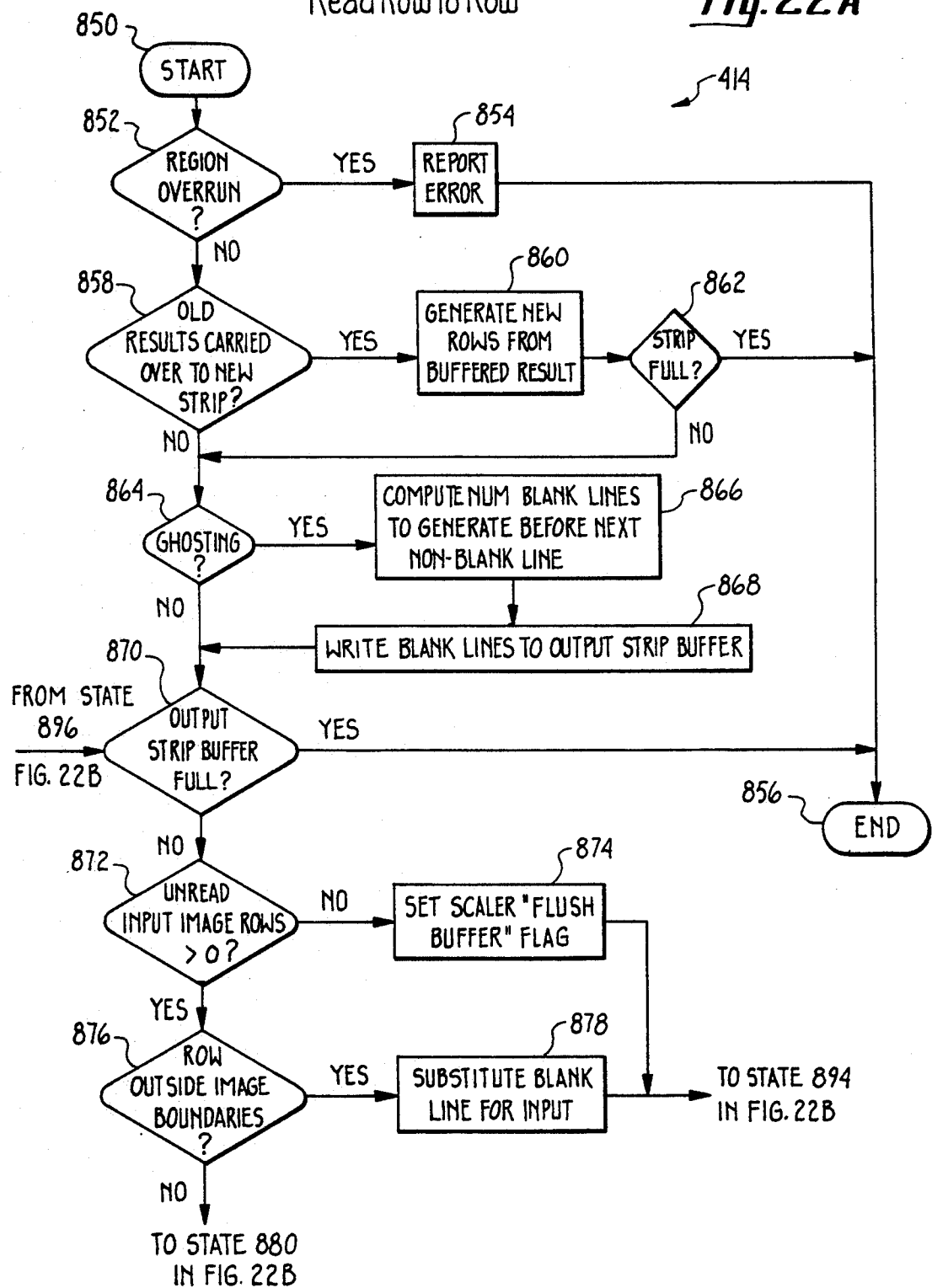
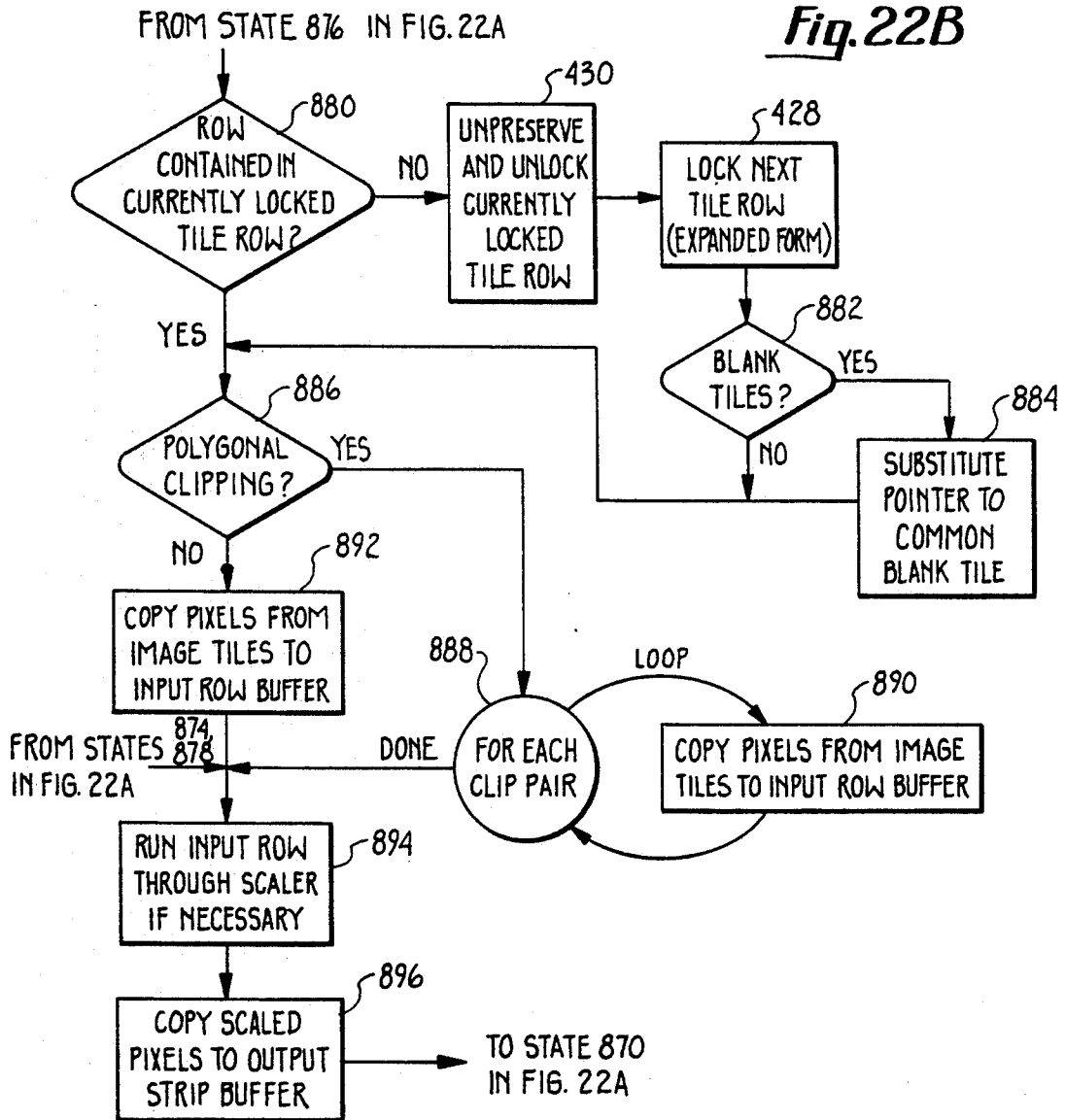


Fig. 22B



WriteRowToRow

Fig. 23A

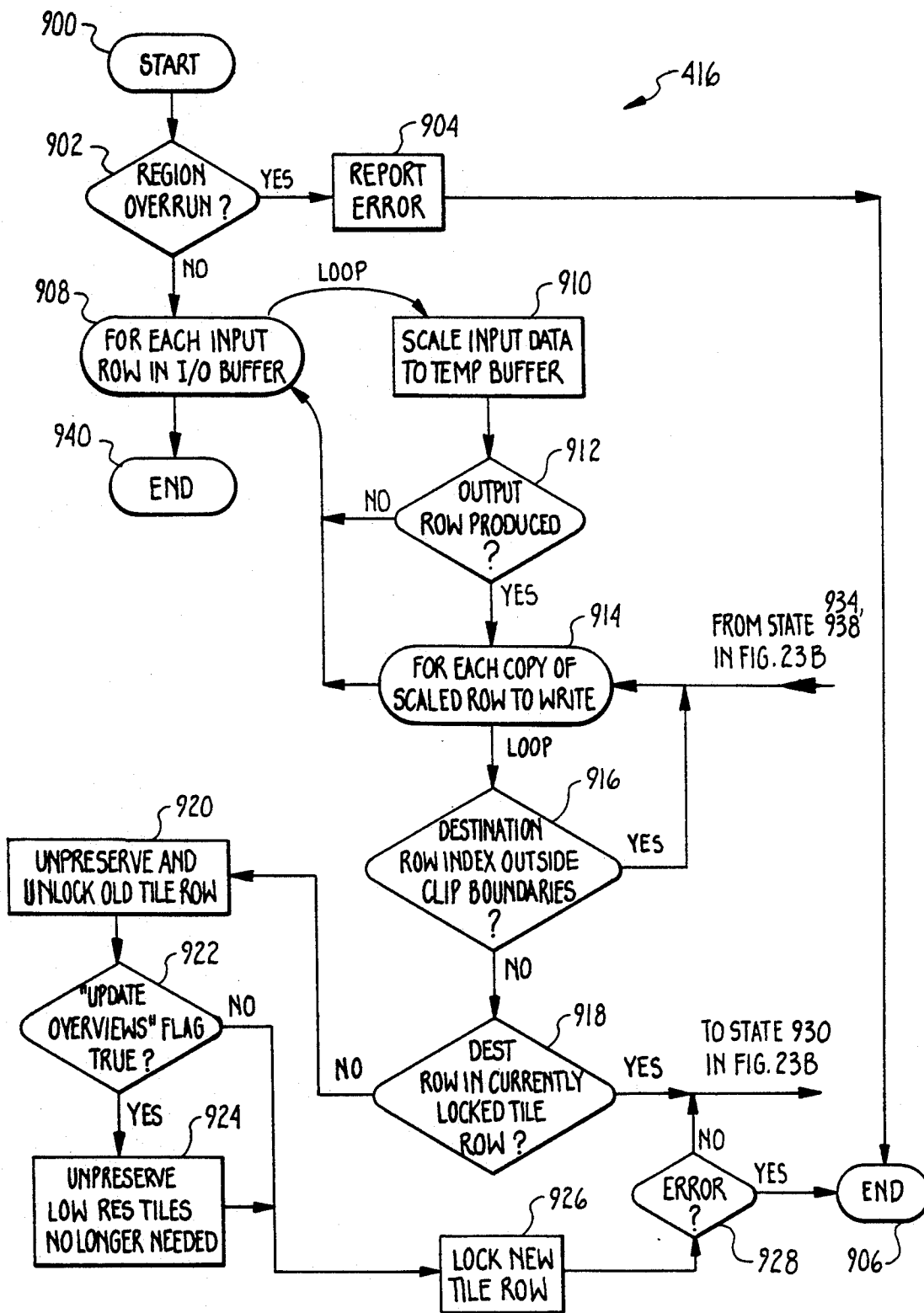
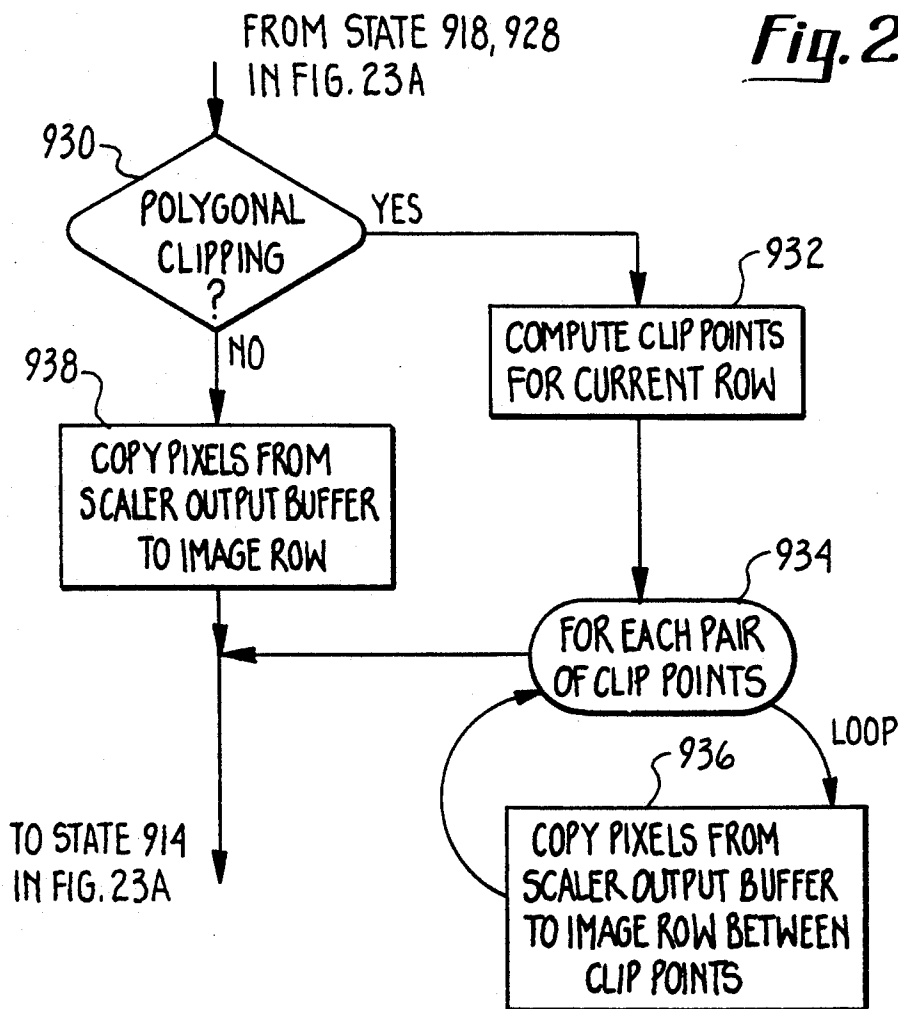
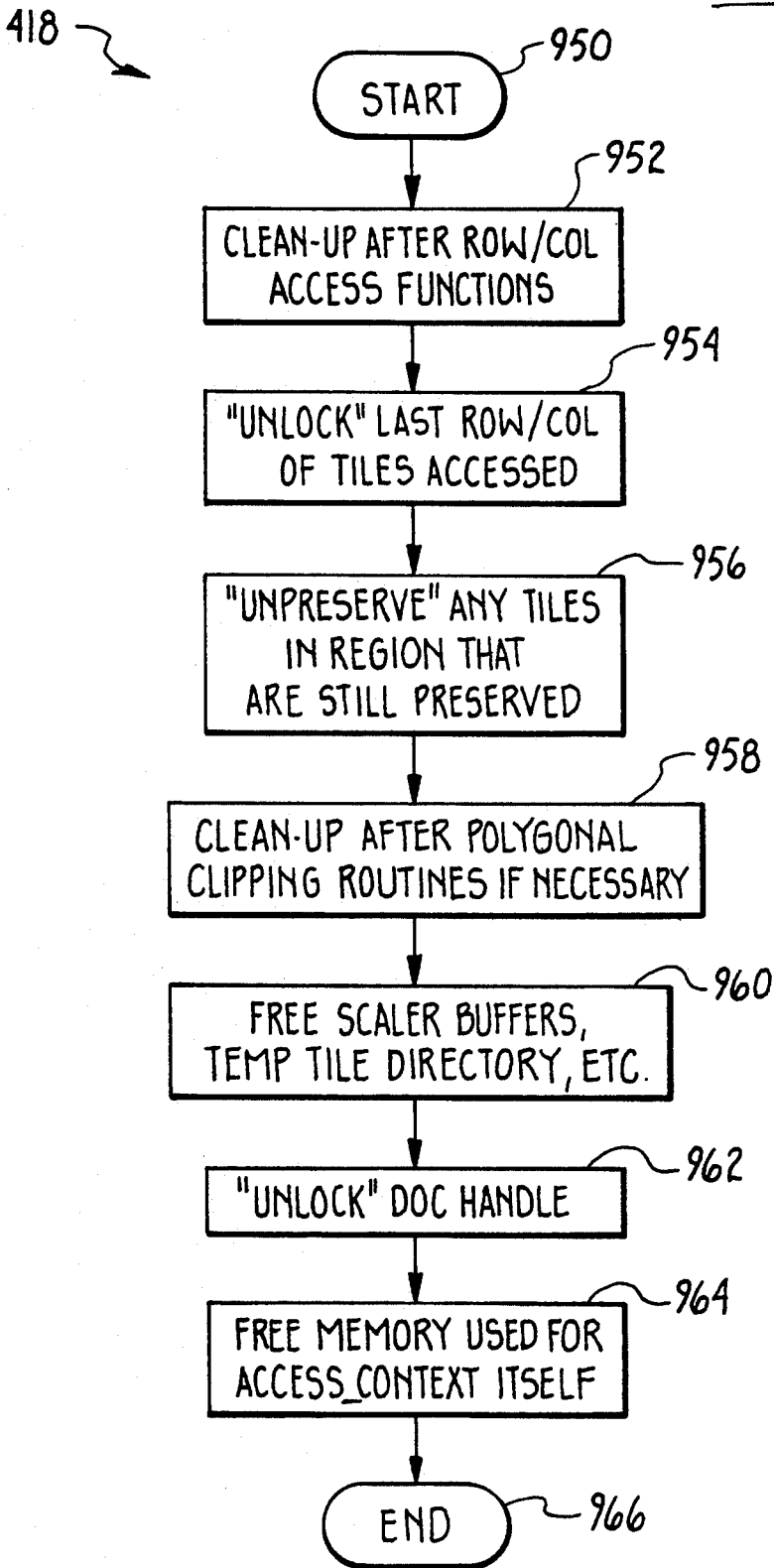


Fig. 23B



End Image Access

Fig. 24



UndoPreviousRasOp

Fig. 25A

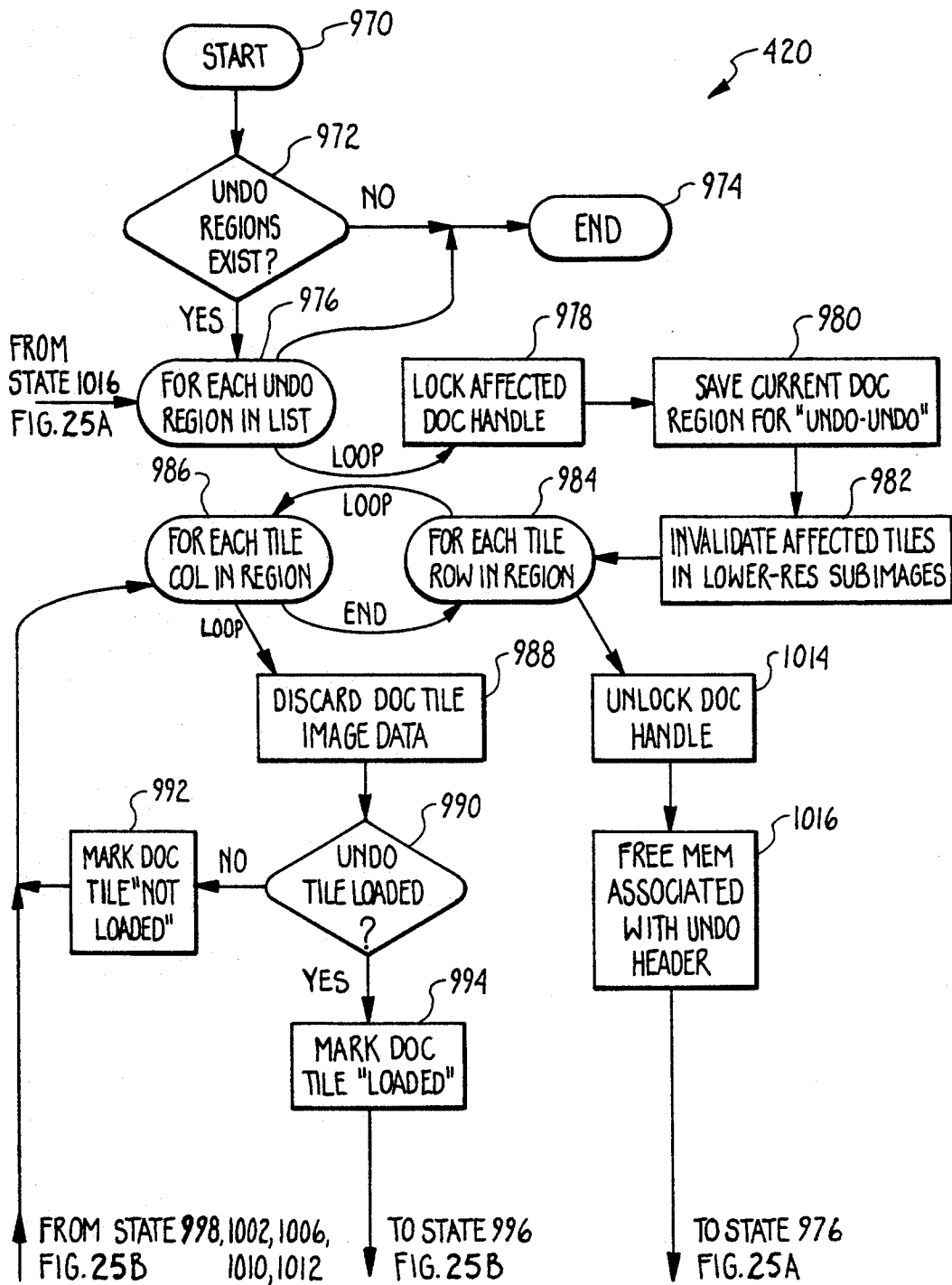
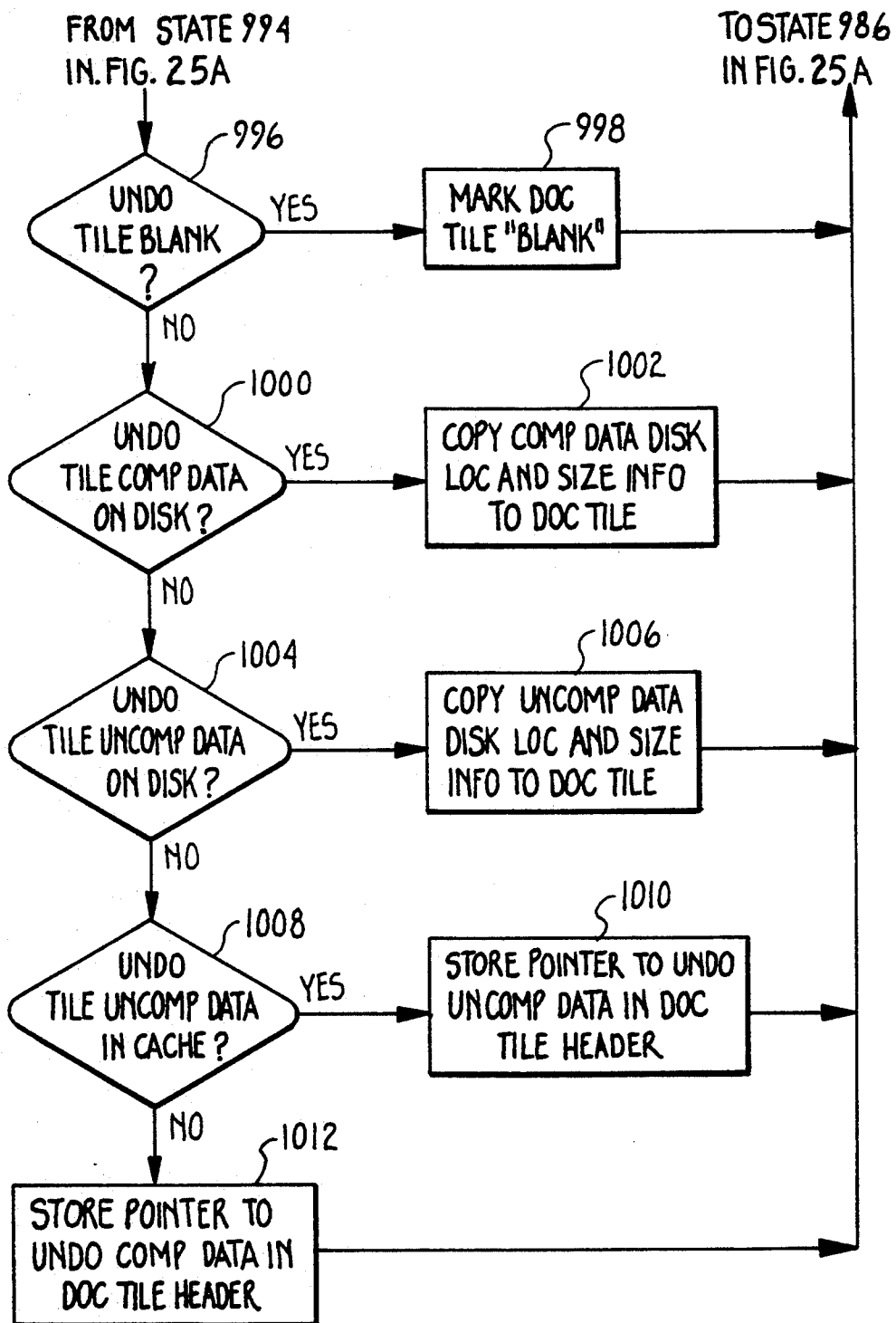
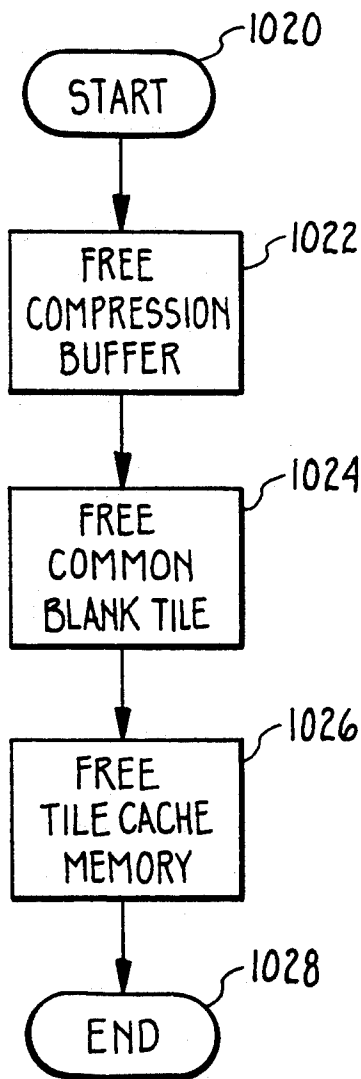


Fig. 25B



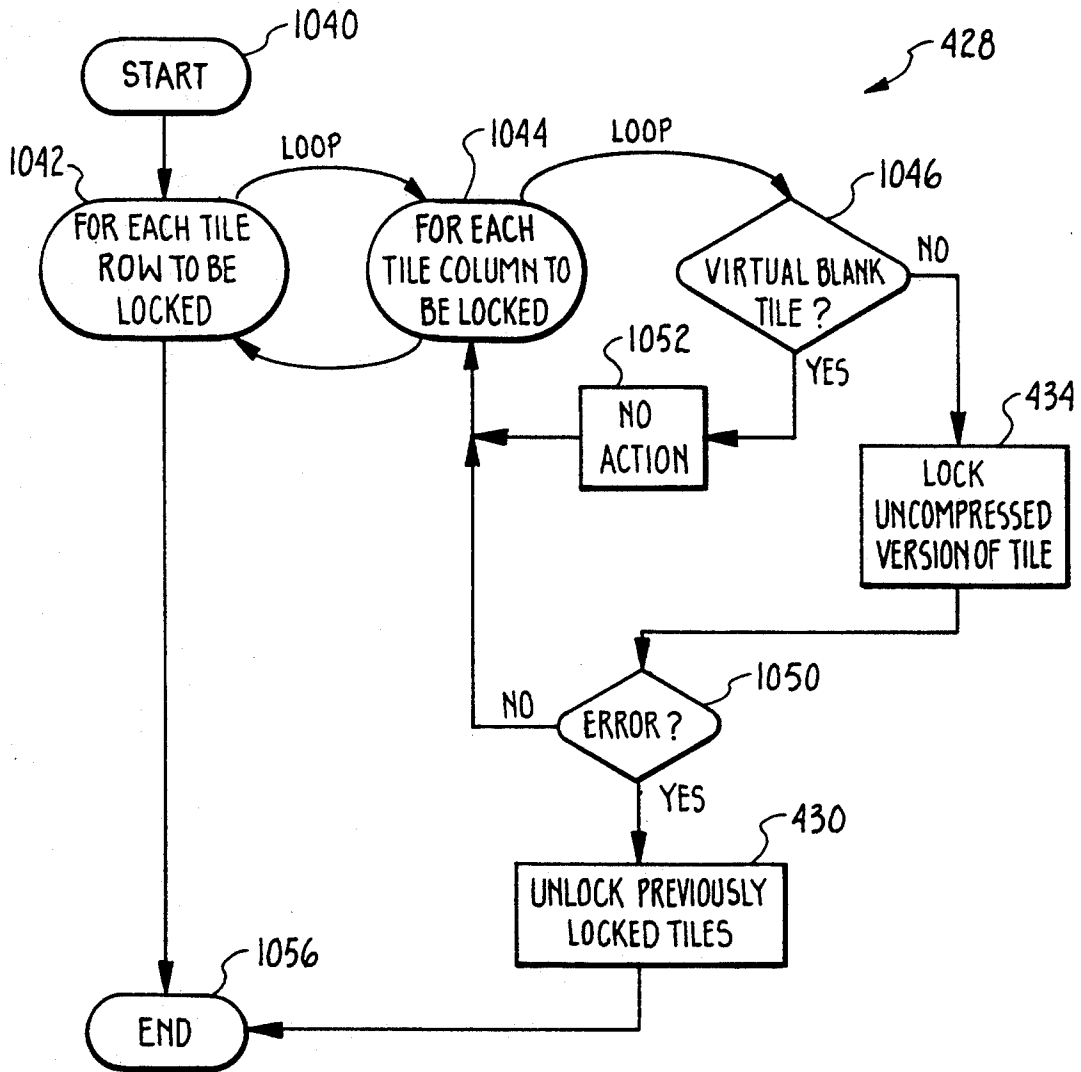
EndCacheManager

Fig. 26



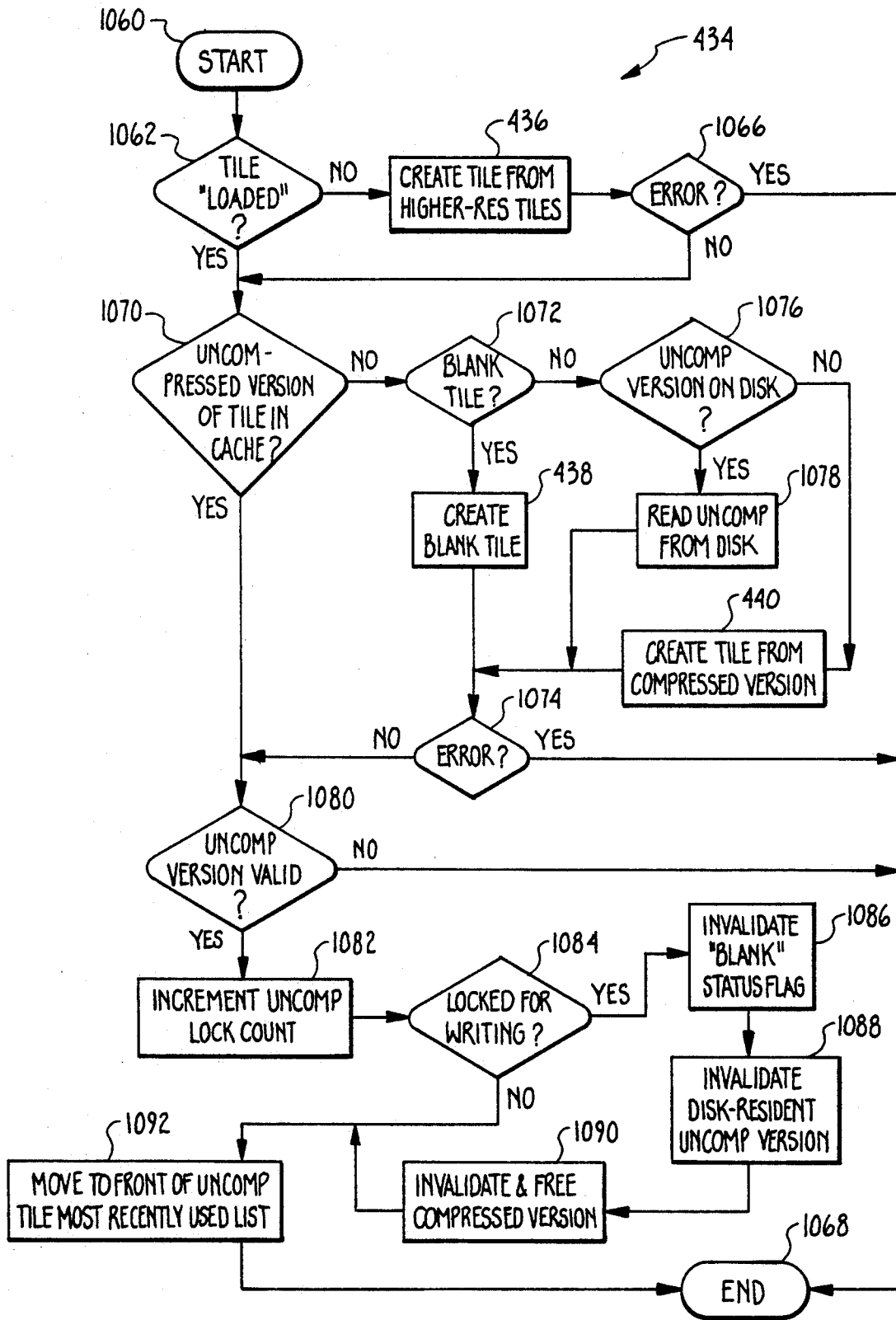
ExpTileLock

Fig. 27



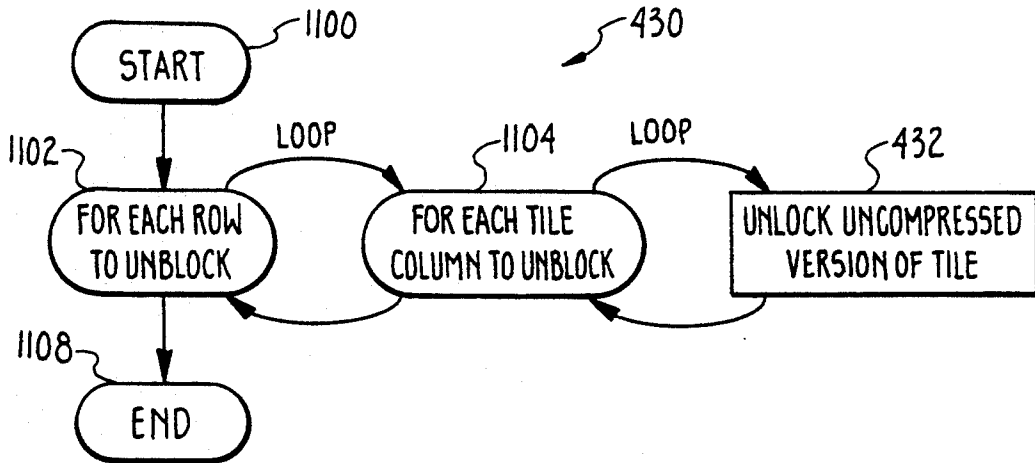
LockExpHandle

Fig. 28



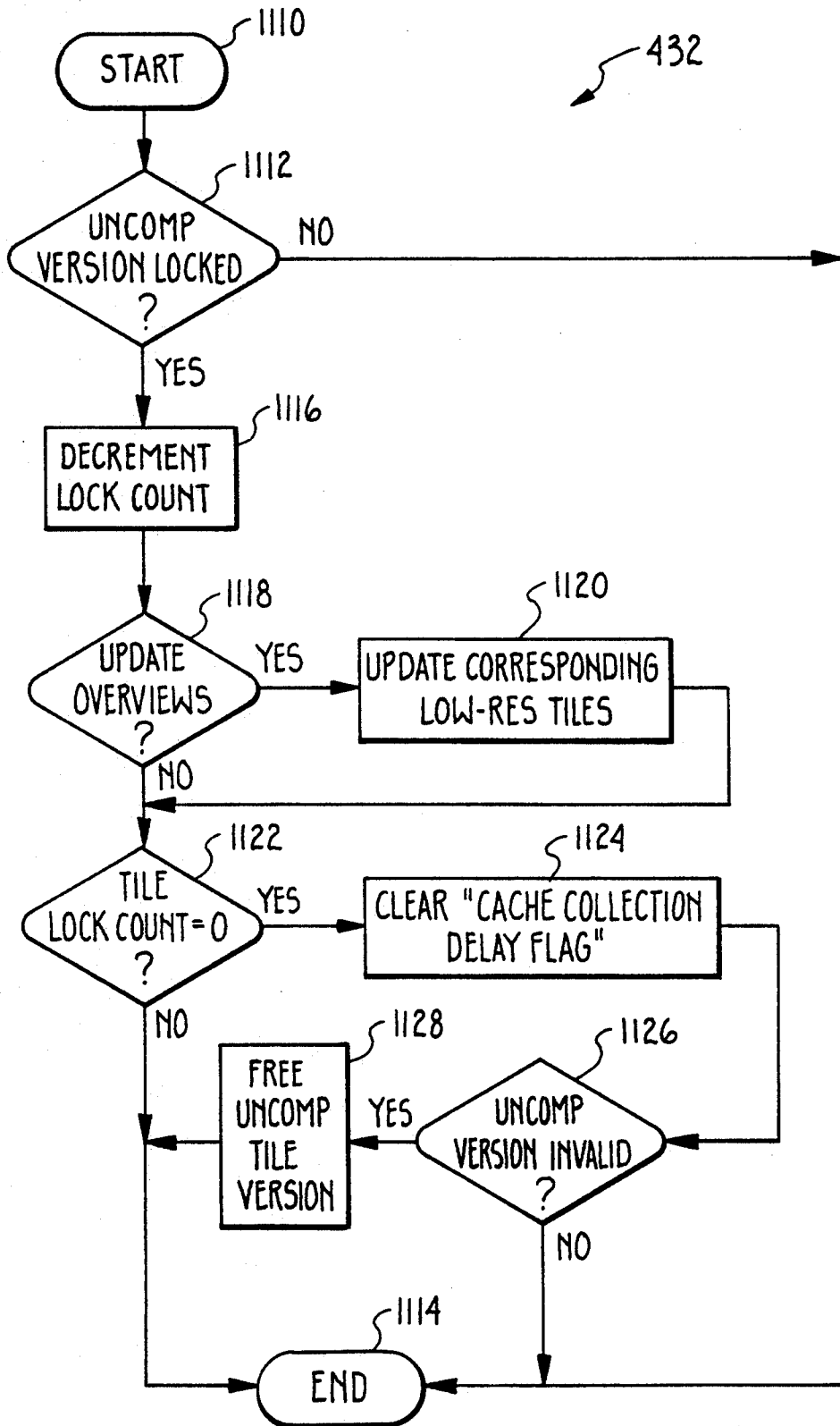
ExpTileUnlock

Fig. 29



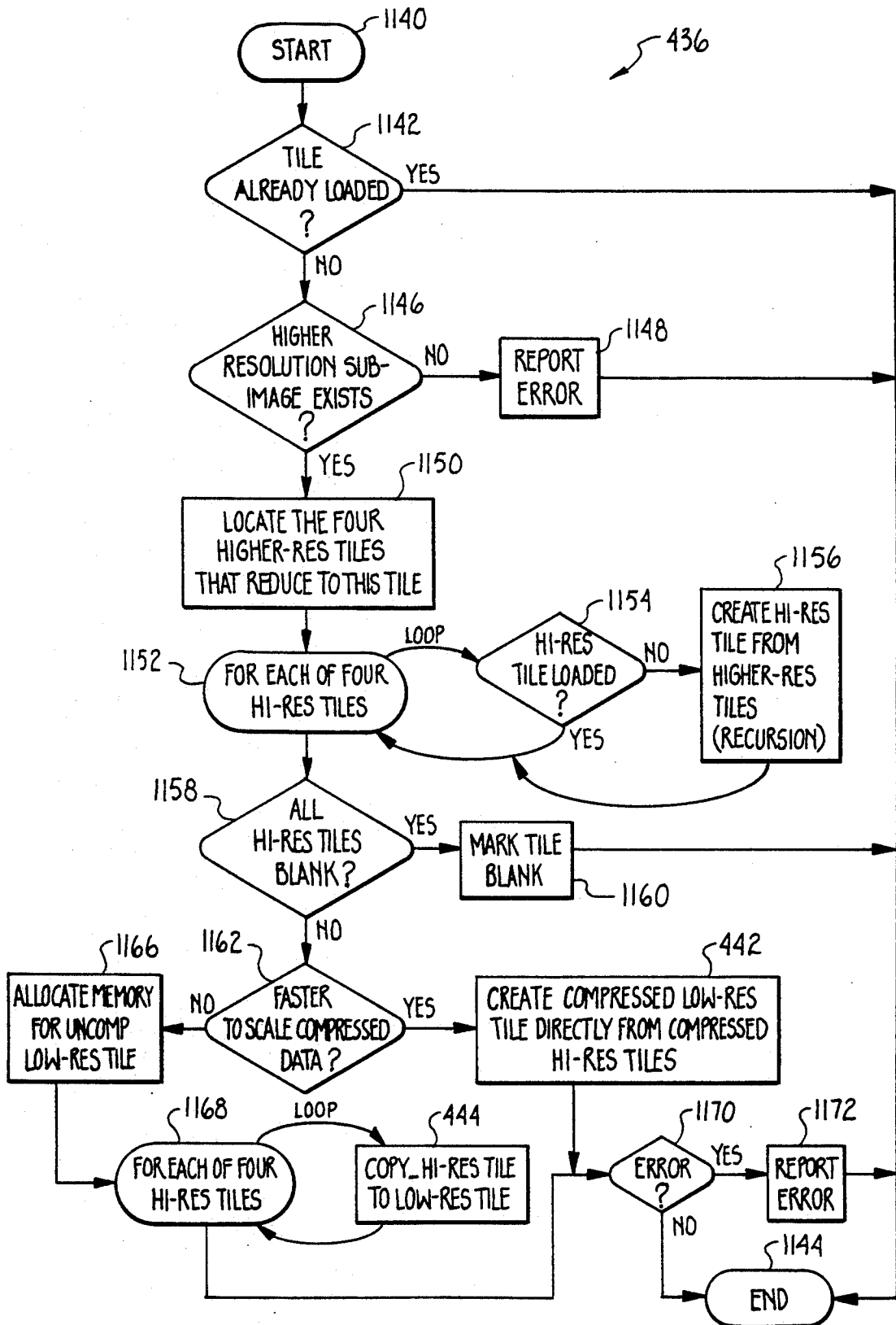
Unlock Exp Handle

Fig. 30



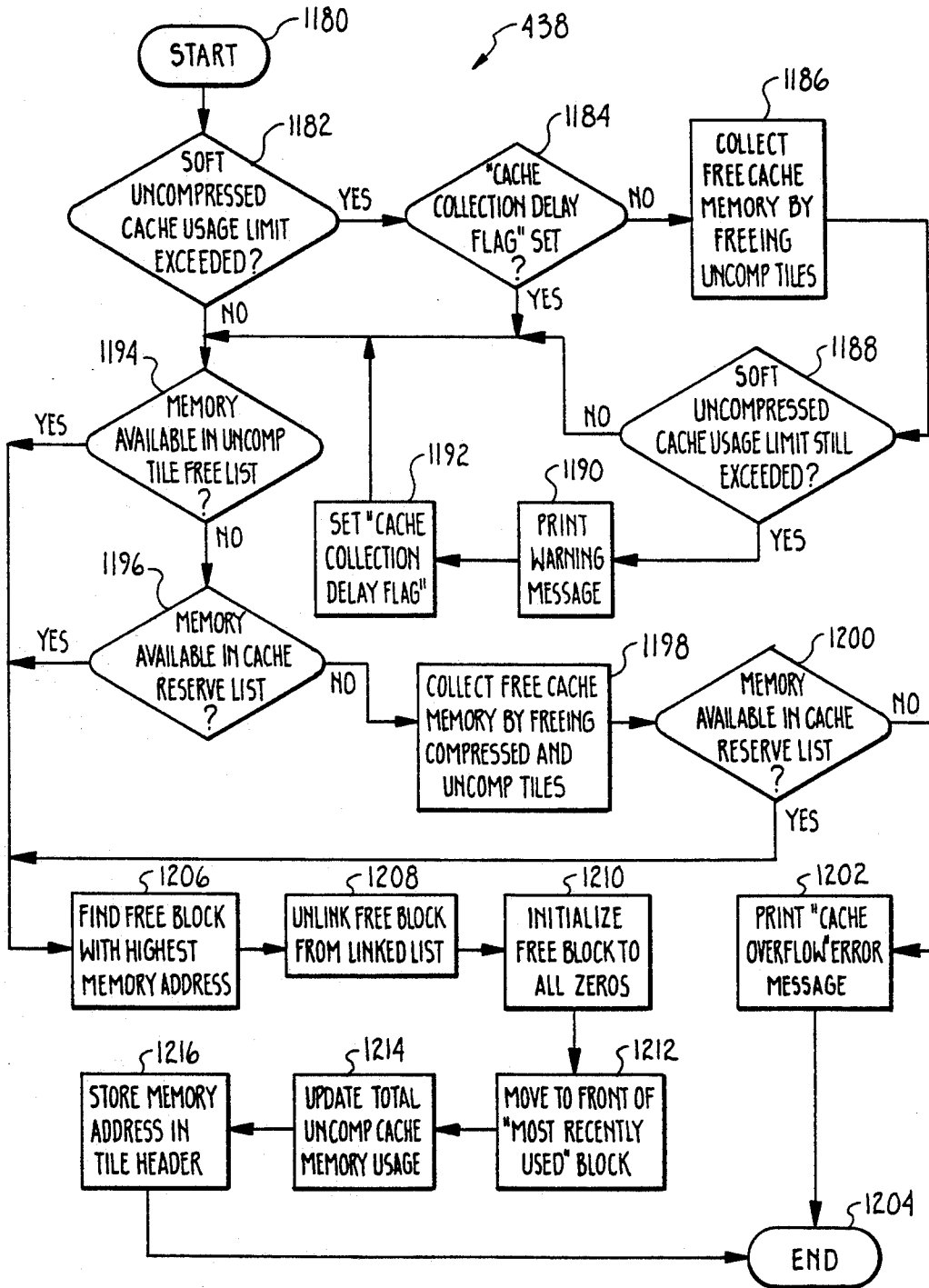
LoadSubImTile

Fig. 31



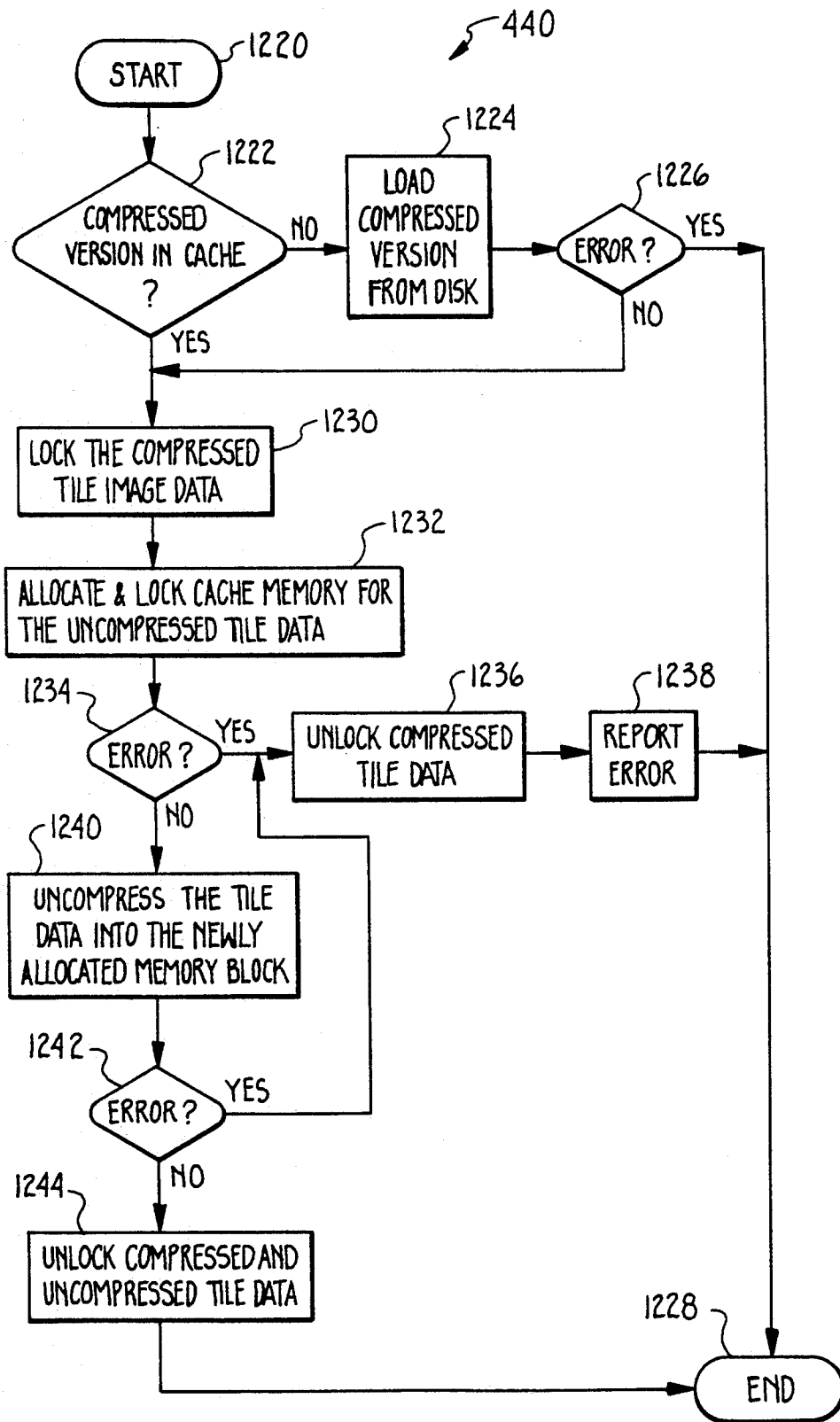
AllocExpHandle

Fig. 32



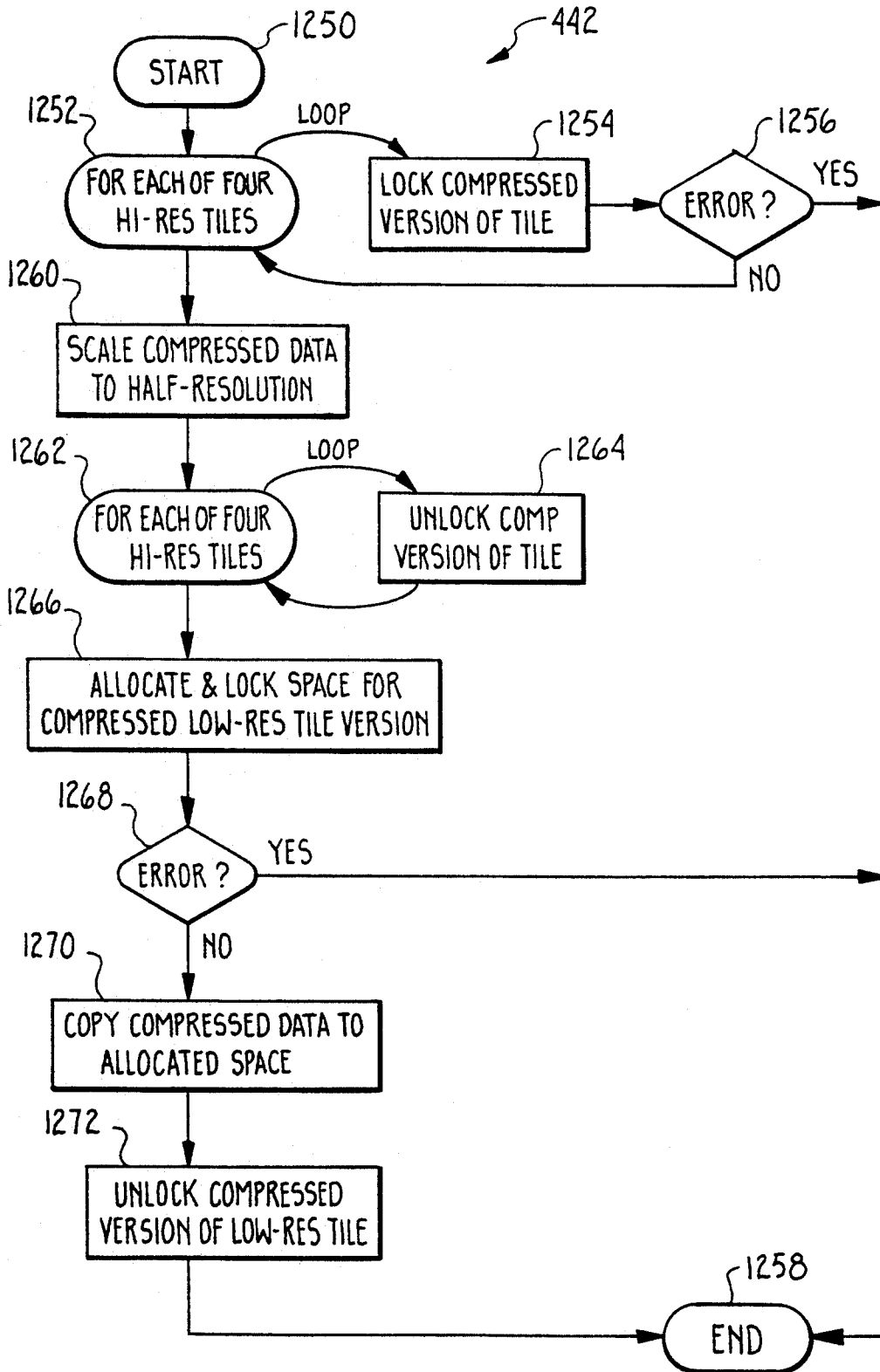
Expand Tile

Fig. 33



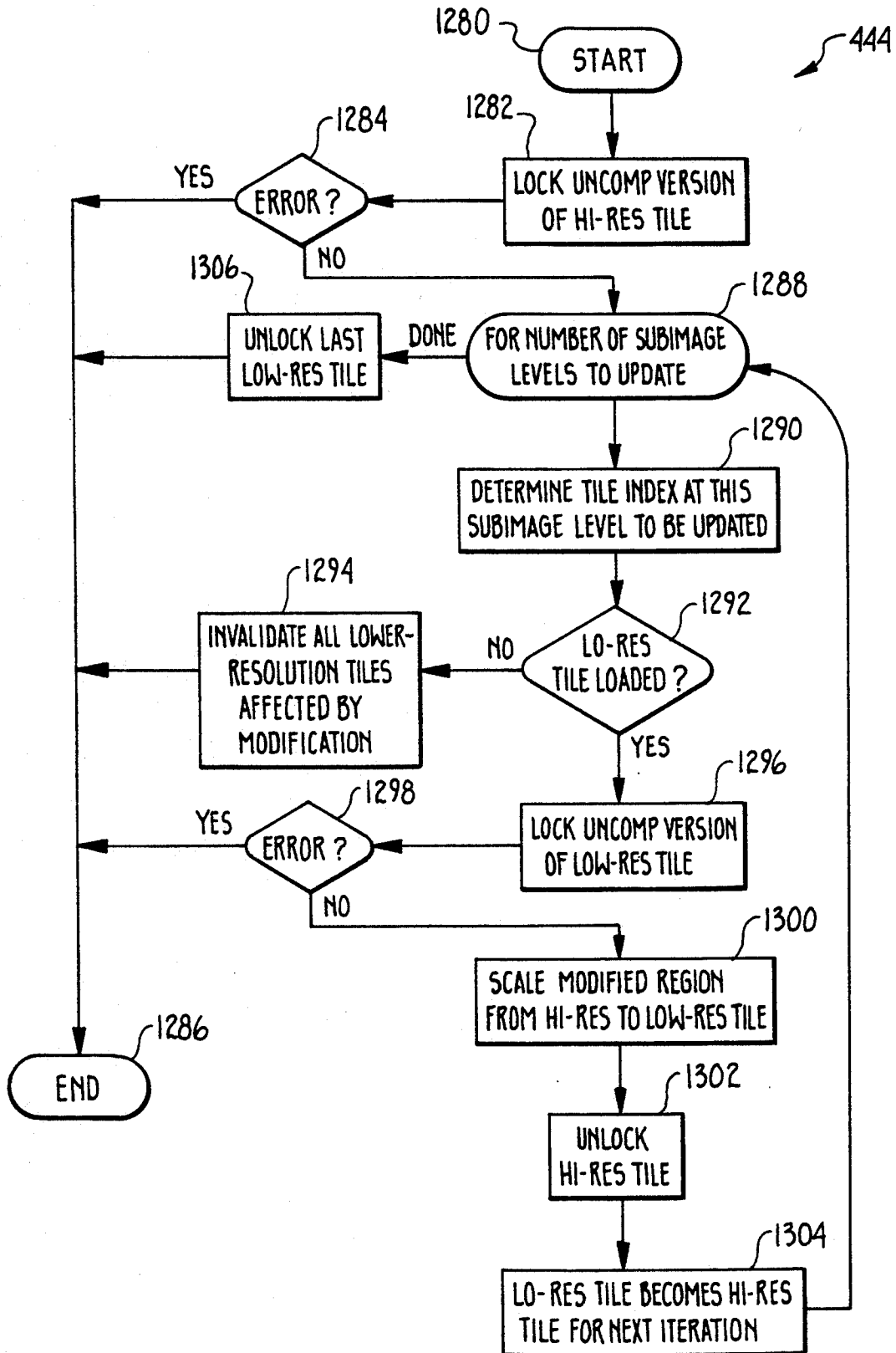
Comp Copy To Overview

Fig. 34



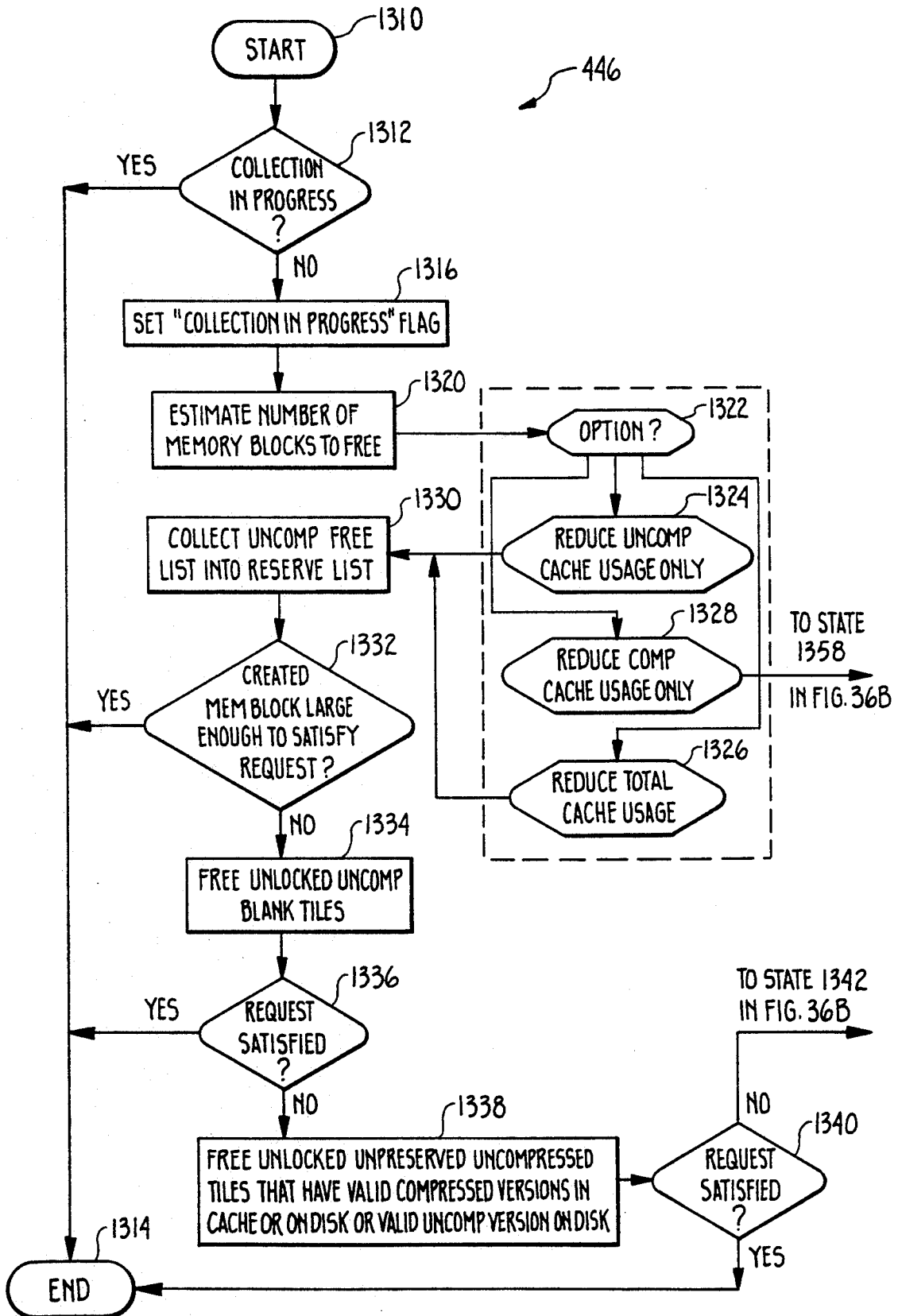
Copy Tile To Overview

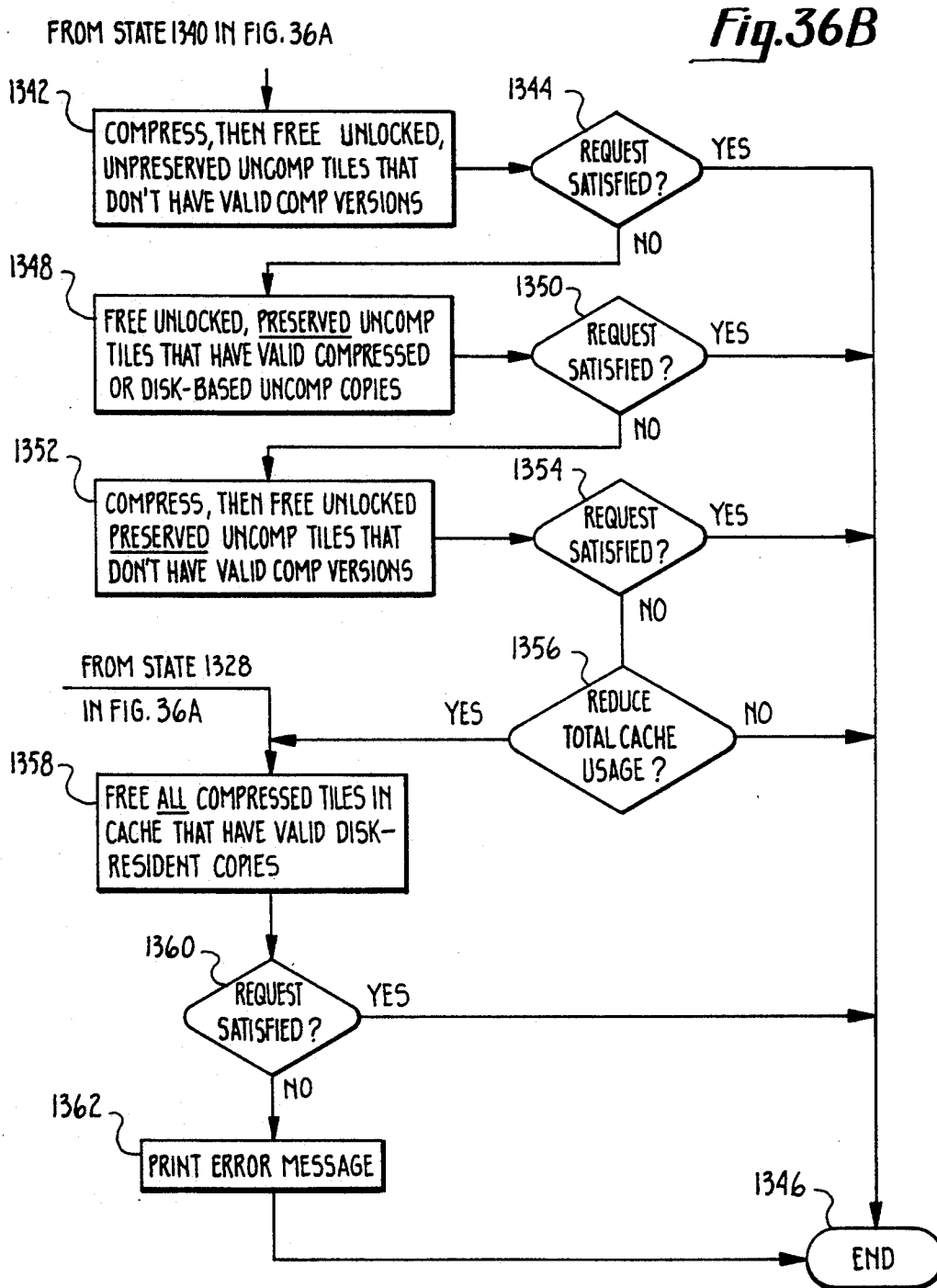
Fig. 35



Collect Free Cache

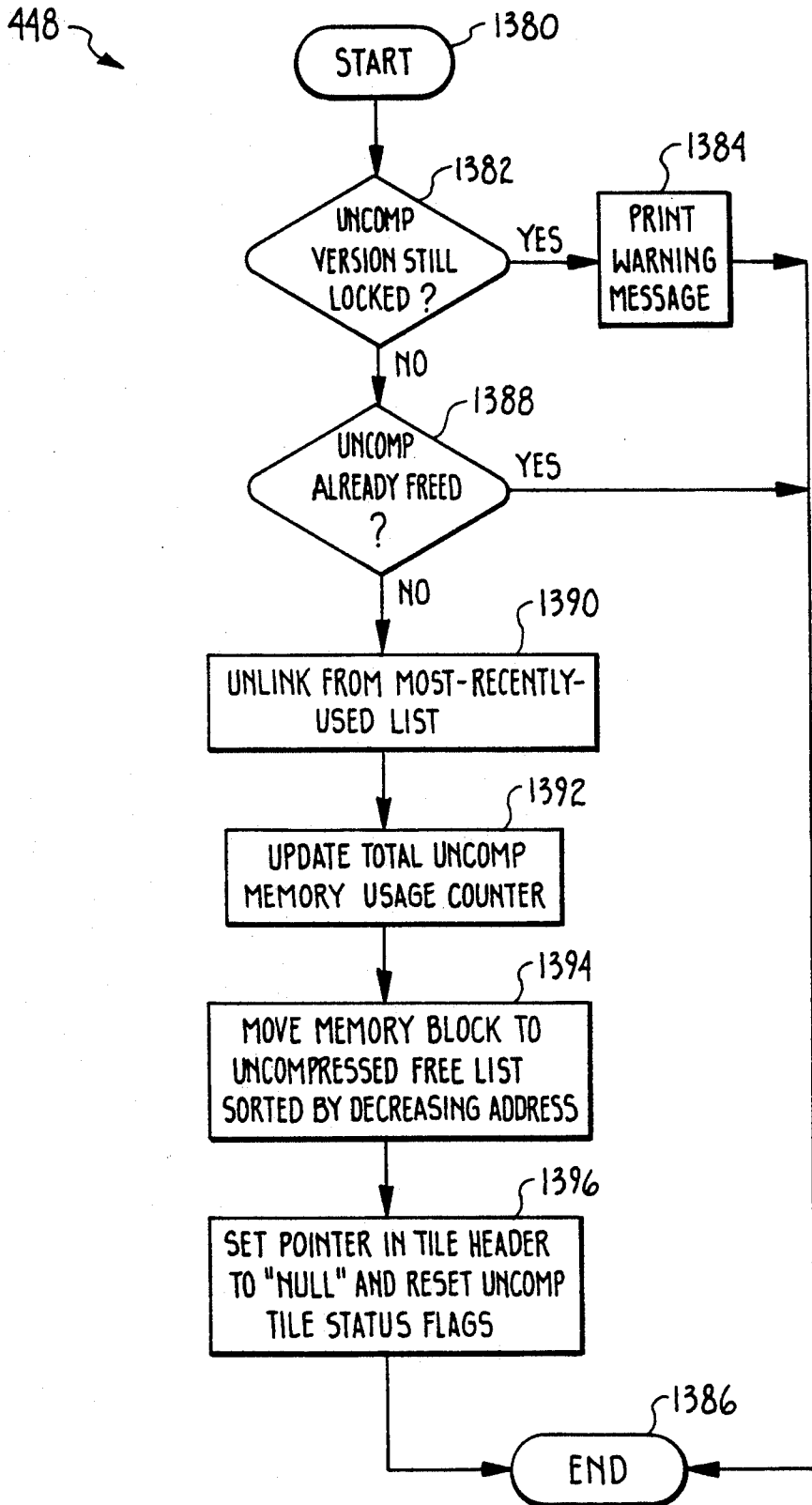
Fig. 36A





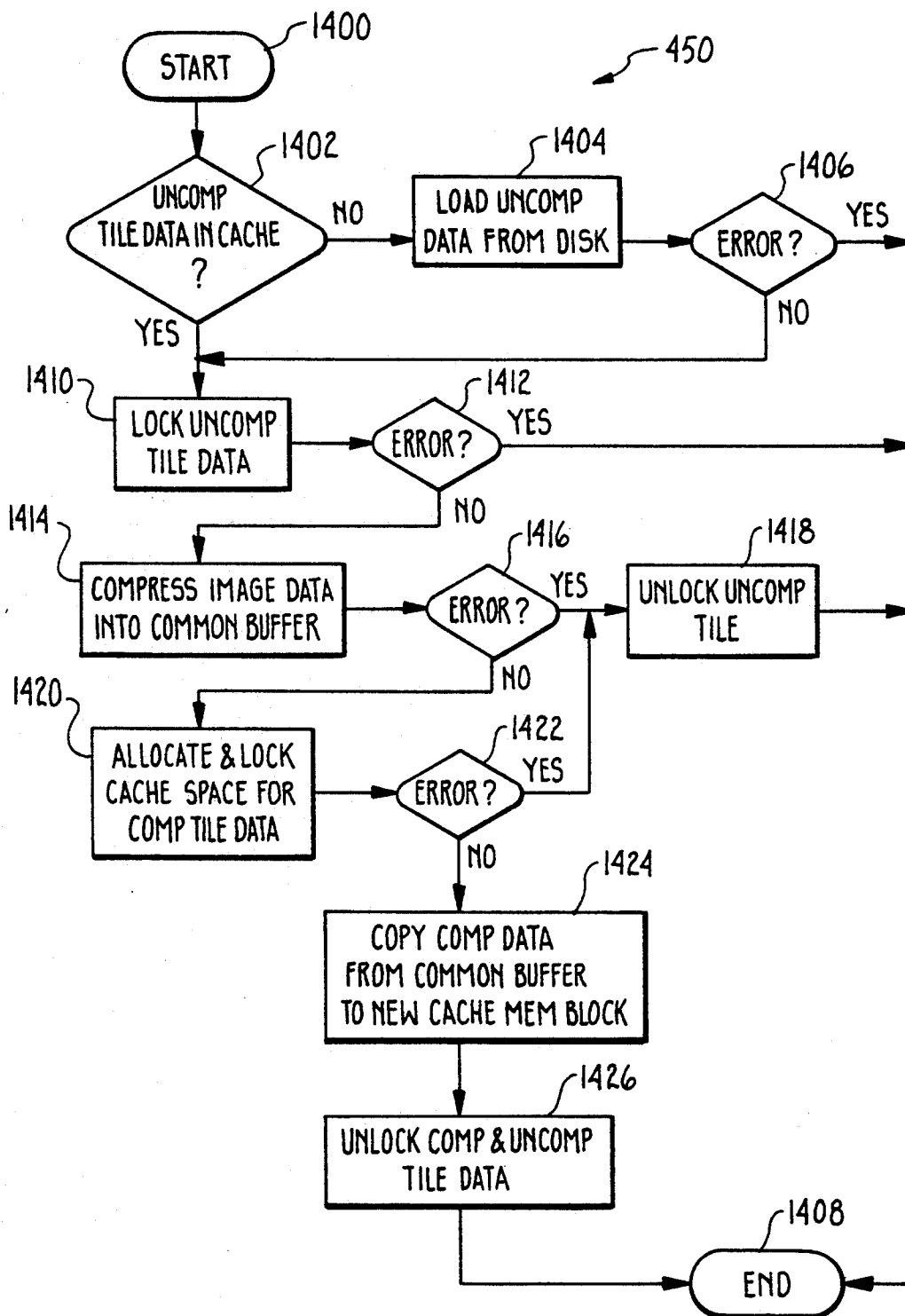
Free ExpHandle

Fig. 37



CompressTile

Fig. 38



SYSTEM FOR MANAGING TILED IMAGES USING MULTIPLE RESOLUTIONS

MICROFICHE APPENDIX

A microfiche appendix containing computer source code is attached. The microfiche appendix comprises one (1) sheet of microfiche having 74 frames.

BACKGROUND OF THE INVENTION

1. Field of the Invention

The present invention relates to memory management systems and, more particularly, to the memory management of large digital images.

2. Description of the Prior Art

The present invention comprises a memory management system for large digital images. These digital, or raster, images are made up of a matrix of individually addressable pixels, which are ultimately represented inside of a computer as bit-maps. Large digital images, such as those associated with engineering drawings, topographic maps, satellite images, and the like, are often manipulated by a computer for the purpose of viewing or editing by a user. The size of, such images are often on the order of tens and even hundreds of Megabytes. Given the current cost of semiconductor memory it is economically impracticable to dedicate a random access memory (RAM) to storing even a single large digital image (hereinafter just referred to as a "digital image"). Thus, the image is usually stored on a slower, secondary storage medium such as a magnetic disk, and only the sections being used are copied into main memory (also called RAM memory).

However, as is well known by users of computer aided design ("CAD") systems, a simplistic memory transfer scheme will cause degraded performance during many typical operations, including zooming or panning. Essentially, during such operations, the computer cannot transfer data between disk and main memory fast enough so that the user must wait for a video display to be refreshed. Clearly, these periods of waiting on memory transfers are wasteful of engineering time.

Presently, to enhance main memory storage of only relevant sections of a digital image, the image is logically segmented into rectangular regions called "tiles". Two currently preferred standards for segmenting an image into tiles are promulgated by the Computer Aided Logistics Support (CALs) organization of the United States government (termed the "CALs standard" herein) and by Aldus Corporation of Seattle, Washington, as defined in the Tagged Image Format File (TIFF) definition (e.g., "TIFF Specification, Revision 5.0, Appendix L). Among other tile sizes, both standards define a square tile having dimensions of 512x512 pixels. Thus, if each pixel requires one byte of storage, the storage of one such tile would require a minimum of 256 kilobytes of memory.

Others, such as Thayer, et al. (U.S. Pat. No. 4,965,751) and Sawada, et al. (U.S. Pat. No. 4,920,504) have discussed tiling or blocking a memory. However, such computer hardware is generally associated with a graphics board for improving the speed of pixel transfers between a frame buffer and a video display by addressing a group of pixels simultaneously. These systems have no relationship to tiling of the image itself and thus do not require knowledge of image size. Tiling has also been used to refer to polygon filling as in Dal-

rymple, et al. (U.S. Pat. No. 4,951,230), which is unrelated to the notion of tiling discussed herein.

The patent to Ewart (U.S. Pat. No. 4,878,183) discusses interlaced cells, each cell containing one or more pixels, for storing continuous tone images such as photographs. The variable size cells are used to vary the resolution of an image according to a distance which is to be perceived by a user. However, the Ewart disclosure does not discuss rasterized binary images containing line drawings, nor does Ewart discuss virtual memory management for modifying or editing images, as will be more fully discussed below.

Even when stored in a mass storage system, an image library, containing a number of digital images, will consume disk space very quickly. Furthermore, "raw" digital images are generally too large to transfer from mass storage to portable floppy disks, or between computer systems (by telephone, for example), in a timely and inexpensive manner unless some means is used to reduce the size of the image. Hence, users of binary images employ image compression techniques to improve storage and transfer efficiencies. One existing compression standard applicable to facsimile transmission, CCITT Group IV, or T.6 compression, is now being used for digital images. Like many other compression techniques, however, the CCITT standard uses statistical techniques to compress data and, hence, it does not always produce a compressed image that is smaller than the original, uncompressed image. That means that image libraries will often contain a mix of compressed and uncompressed binary images. Similar compression standards exist for color and gray-scale images such as those promulgated by the JPEG (Joint Photog. Exp. Group) Standards Committee of the CCITT as SGV III Draft Standard.

At the present time, digital images are typically viewed and modified with an image editor using an off-the-shelf computer workstation. These workstations usually come with a sophisticated operating system, such as UNIX, that employs a virtual memory to effectively manage memory accesses in secondary and main memories. In an operating system having virtual memory, the data that represents the executable instructions for a program or the variables used by that program do not need to reside entirely in main memory. Instead, the operating system brings portions of the program into main memory only as needed. (The data that is not stored in main memory being stored on magnetic disk or other like nonvolatile memory.) The address space that is available to any one application program is generally managed in blocks of convenient sizes called "pages" or "segments".

In general, a virtual memory system allows application programs to be written and executed without concern for the management of virtual memory carried out by the operating system. Thus, independence of the size of main memory is achieved by creating a "virtual" address space for the program. The operating system translates virtual addresses into physical addresses (in a main or cache memory) with the aid of an "address translation table". This table contains one entry per virtual memory segment of status information. For instance, segment status will commonly include information about whether a segment is currently in main memory, when a segment was last used, a disk address at which the disk copy of the segment resides, and a RAM address at which the segment resides (only valid if the segment is currently loaded in main memory).

When the program attempts to access data in a segment that is not currently resident in main memory, the operating system reads the segment from disk into main memory. The operating system may need to discard another segment to make room for the new one (by overwriting the area of main memory occupied by the old segment), so some method of determining which segment to discard is required. Usually the method is to discard the least recently used segment. If the discarded segment was modified then it must be written back to disk. The operating system completes the "swap" operation by updating the address translation table entries of the new and discarded segments.

In summary, the conventional memory management schemes consider data to be in one of two states: resident or not resident in main memory. Which segments are stored in main memory at any given time is generally determined only by past usage, with no way of predicting future memory demands. For instance, just because a segment is the least recently used does not mean that it will not be used at the very next memory access.

However, the management of virtual memory for images departs significantly from conventional virtual memory schemes because images and computer programs are accessed in very different ways. Computer programs tend to access one small neighborhood of virtual address, and then jump to some distant, essentially random, location. However, during normal image processing operations an image is accessed in one of a finite set of predictable patterns. It is not surprising then that conventional memory management systems can significantly degrade performance when used in image processing applications by applying inappropriate memory management rules. Rules which should be abided by a memory management system for large digital images are the following:

1. Image memory must be managed as rectangular image regions (called "tiles"), not as linear memory address ranges.

2. An image tile can exist in five forms: uncompressed memory-resident, compressed memory-resident, uncompressed disk-resident, compressed disk-resident and "can be derived from other available image tiles", in contrast to the two basic forms of memory-resident and disk-resident available in conventional virtual memory schemes.

3. The image region that will be affected by a particular image processing operation is known before the operation begins, and that information can be conveyed to the memory manager.

4. An image memory manager must be tunable to different system capabilities and image types. For example, many computers can decompress a tile of binary data much faster than they can retrieve the uncompressed version of the same tile from disk. On the other hand, some images cannot be compressed at all.

5. An image memory management system should support the capability to "undo" editing operations which is built into the memory manager for optimal performance and ease of use. Thus, the memory manager could easily save copies of the compressed tiles in the affected region, and quickly restore the image to the original state by simply modifying the tile directory entries to point to the old version.

Reader, et al., ("Address Generation and Memory Management for Memory Centered Image Processing Systems", SPIE, Vol. 757, Methods for Handling and

Processing Imagery, 1987) discuss a primitive memory management system for images. However, in that system, image tiles are only stored in memory and not on disk. Furthermore, in the Reader, et al., system, there is no capability to handle images in compressed form, nor is there any discussion of "undoing" editing operations.

Consequently, a need exists for an image memory management system that provides: linkages with a raster image editor which includes modify and undo operations, true virtual memory for large images specifying locations on disk and in memory, simultaneous handling of compressed and uncompressed images, and a method for rapidly constructing reduced resolution views of the image for display. The latter need is particularly important when viewing a large image reduced to fit on a video display.

SUMMARY OF THE INVENTION

The above-mentioned needs are satisfied by the present invention which includes a memory management system for tiled images. The memory management system includes a tile manager for maintaining a virtual memory comprising a main memory and a secondary memory such as a disk. The tiled images may include tiles in compressed or uncompressed form.

The tile manager selects the form of image tile that most appropriately matches a request. Each tile of an image may exist in one or more of five different forms, or states, as follows: uncompressed and resident in the image data cache, compressed and resident in the image data cache, uncompressed and resident on disk, compressed and resident on disk and not loaded but re-creatable using data from higher-resolution image tiles.

An image stack having successively lower-resolution subimages is constructed from a full resolution source image. The lower-resolution images in the image stack may be used to enhance such standard image accesses as zooming and panning where high speed image reduction is advantageous.

The image memory management system provides linkages with image processing applications that facilitate image modifications. The tile manager need only store compressed tiles that relate to so-called undoable operations.

These and other objects and features of the present invention will become more fully apparent from the following description and appended claims taken in conjunction with the accompanying drawings.

BRIEF DESCRIPTION OF THE DRAWINGS

FIG. 1 is a perspective view of an image stack comprising full, half, quarter and eighth resolution tiled images;

FIG. 2 is a full resolution image of a mechanical part; FIG. 3 is a half resolution image of the mechanical part shown in FIG. 2;

FIG. 4 is a quarter resolution image of the mechanical part shown in FIG. 2;

FIG. 5 is an eighth resolution image of the mechanical part shown in FIG. 2;

FIG. 6 is a block diagram showing one preferred embodiment of a computer system that includes the present invention;

FIG. 7 is a memory map showing the general arrangement of cache memory according to the present invention;

FIG. 8 is a state diagram defining the flow of tile data between different storage states according to the present invention;

FIGS. 9A and B are a diagram of one preferred data structure defining document information according to the present invention;

FIG. 10 is a diagram of one preferred data structure defining a tile header for maintaining the status of compressed or uncompressed tiles;

FIG. 11 is a diagram of a partial calling hierarchy for the various functions of the presently preferred embodiment of the tile manager of the present invention;

FIG. 12 is a flow diagram of one preferred embodiment of the tile manager;

FIG. 13 is a flow diagram defining the "initialize cache manager" function referred to in the flow diagram of FIG. 12;

FIG. 14 is a state diagram of the locking and unlocking of a memory, state, according to the present invention;

FIGS. 15A, 15B, and 15C are a flow diagram defining the "create image access context" function referred to in FIG. 12;

FIG. 16 is a diagram, of a data structure defining the access context referred to in FIGS. 15A,B;

FIGS. 17A and 17B are a flow diagram defining the "save region for undo" function referred to in FIG. 15B;

FIG. 18 is a flow diagram defining the "load tiled raster image" function referred to in FIG. 12;

FIG. 19 is a flow diagram defining the "load TIFF subimage tile information into tile headers" function referred to in FIG. 18;

FIG. 20 is a flow diagram defining a "store tile info in tile headers" function referred to in FIG. 12;

FIG. 21 is a flow diagram defining the "begin undoable raster operation" function referred to in FIG. 12;

FIGS. 22A and 22B are a flow diagram defining the "read rows from region" function referred to in FIG. 12;

FIGS. 23A and 23B are a flow diagram defining the "write rows to region" function referred to in FIG. 12;

FIG. 24 is a flow diagram defining the "close image access context" function referred to in FIG. 12;

FIGS. 25A and 25B are a flow diagram defining the "undo previous raster operations" function referred to in FIG. 12;

FIG. 26 is a flow diagram defining the "quit cache manager" function referred to in FIG. 12;

FIG. 27 is a flow diagram defining the "lock expanded image tile group" function referred to in FIG. 22A;

FIG. 28 is a flow diagram defining the "lock expanded tile" function referred to in FIG. 27;

FIG. 29 is a flow diagram defining the "unlock expanded image tile group" function referred to in FIG. 27;

FIG. 30 is a flow diagram defining the "unlock expanded tile" function referred to in FIG. 29;

FIG. 31 is a flow diagram defining the "create tile from higher-resolution tiles" function referred to in FIG. 28;

FIG. 32 is a flow diagram defining the "allocate space for uncompressed version of tile" function referred to in FIG. 28;

FIG. 33 is a flow diagram defining the "create uncompressed version of tile from compressed version" function referred to in FIG. 28;

FIG. 34 is a flow diagram defining the "create compressed low resolution tile from compressed higher-resolution tiles" function referred to in FIG. 31;

FIG. 35 is a flow diagram defining the "copy uncompressed high resolution tile to uncompressed low resolution tiles" function referred to in FIG. 31;

FIGS. 36A and 36B are a flow diagram defining the "collect freeable cache memory" function referred to in FIG. 32;

FIG. 37 is a flow diagram defining the "free uncompressed version of tile" function referred to in FIGS. 36A,B; and

FIG. 38 is a flow diagram defining the "create compressed version of tile from uncompressed version" function referred to in FIG. 17B.

DETAILED DESCRIPTION OF THE PREFERRED EMBODIMENTS

Reference is now made to the drawings wherein like parts are designated with like numerals throughout.

FIG. 1 illustrates an image stack, generally indicated at 100. The design of the image stack 100 is based on the idea that image memory can be managed as small square regions, called tiles, that are mostly independent of one another. In general, a tile may be either uncompressed (also termed expanded) or compressed. While the basic uncompressed tile size could be a variable, it is presently preferred to be fixed at 32 kilobytes, or 512 pixels by 512 pixels to conform with the Computer Aided Logistics Support (CALs) raster file format standard for binary images. (Note that the present invention allows binary and color images to coexist in a common image memory management system.)

In order to compensate for lower performance expected with a virtual memory management system for images, particularly when reducing large portions (by combining pixels) of the image for display, the present invention automatically maintains a series of reduced resolution copies, called subimages, of the full resolution image. Preferably, the resolution (i.e., pixels per inch) of each subimage is reduced by exactly half relative to the next higher-resolution subimage. Thus, the image stack 100 can be visualizing as an inverted pyramid, wherein the images can be stacked beginning with a full resolution subimage (or image) 102 at the top, followed by a half resolution subimage 104, then a quarter resolution subimage 106, and an eighth resolution subimage 108. (In FIG. 1, the subimages 102-108 are outlined by bolded lines.)

The subimages 102, 104, 106, 108 are superimposed on a set of tiled subimages 110a, 110b, 110c, 110d, respectively, defining sets of tiles. The extent of the image stack 100 ends at the resolution that allows the entire subimage to be stored within a single tile 108 (preferably 512×512 pixels square). Each lower-resolution subimage 104-108 is a faithful representation of the full resolution subimage 102 at all times, with the exception of certain times during operations that modify the appearance of the full resolution subimage 102.

FIG. 2 illustrates an 8½"×11", A-size mechanical drawing (to scale) as the full resolution subimage 102 showing a mechanical part 120a. Of course, other larger drawings such as, for example, D-size and E-size may be used by the present invention. Also, other image processing applications besides mechanical drawings may be used with the present invention including electrical schematics, topographical maps, satellite images, hea-

ting/ventilating/air conditioning (HVAC) drawings, and the like.

FIG. 3 illustrates the corresponding half resolution subimage 104 showing the half resolution part 120b. FIG. 4 illustrates the corresponding quarter resolution subimage 106 showing the quarter resolution part 120c. Lastly, FIG. 5 illustrates an eighth resolution subimage 108 showing the eighth resolution part 120d. In the preferred embodiment, reduced resolution subimages can be used any time that a reduction factor of 2:1 or higher would be used to scale a region of interest in the full resolution subimage 102 for display, plotting or copying.

The subimages 102-108 can be loaded from a source image file, if they exist, or they can be created on demand by the image memory management system of the present invention. The present invention includes editing capabilities that allow a user to trade off between "quick flash" pan/zoom performance and file size as measured by the number of reduced resolution subimages stored with each image. Depending on the application, the user will normally opt to store one or more reduced resolution subimages with each source image file.

The lower-resolution subimages, for example, subimages 104-108, are utilized by the image memory management system to produce the illusion of instant access to any region of the image at any scale factor (not just the scale factor of the overview subimage). Increasing the number of lower-resolution subimages gives a higher quality "first flash" image during panning and zooming and reduces the time to get the final version of the image to the screen.

FIG. 6 illustrates a computer workstation generally indicated at 150 which is representative of the type of computer that is used with the present invention. The workstation 150 comprises a computer 152, a color monitor 154, a mouse 156, a keyboard 158, a floppy disk drive 160, a hard disk drive 162 and an Ethernet communications port 164. The computer 152 includes a motherboard bus 166 and an I/O bus 168. The I/O bus 168, in one preferred embodiment, is an IBM PC/AT® bus, also known as an Industry Standard Architecture (ISA) bus. The two buses 166, 168 are electrically connected by an I/O bus interface and controller 170.

The I/O bus 168 provides an electromechanical communication path for a number of I/O circuits. For example, a graphics display controller 172 connects the monitor 154 to the I/O bus 168. In the presently preferred embodiment, the monitor 154 is a 19-inch color monitor having a 1,024 × 768 pixel resolution. A serial communications controller 174 connects the mouse 156 to the I/O bus 168. The mouse 156 is used to "pick" an image entity displayed on the monitor 154.

The I/O bus 168 also supports the hard disk drive 162, and the Ethernet communications port 164. A hard disk controller 176 connects the hard disk drive 162 to the I/O bus 168. The hard disk drive 162, in one possible configuration of the workstation generally indicated at 150, stores 60 megabytes of data. An Ethernet communications controller 178 connects an Ethernet communications port 164 with the I/O bus 168. The Ethernet communications controller 178 supports the industry standard communications protocol TCP/IP which includes FTP and Telnet functions. The Ethernet communications port 164 of the preferred embodiment allows the Workstation 150 to be connected to a network

which may include, among other things, a document scanner (not shown) and a print server (not shown).

The motherboard bus 166 also supports certain basic I/O peripherals. For example, the motherboard bus 166 is connected to a keyboard and floppy disk controller 180 which supports the keyboard 158 and the floppy disk drive 160. The floppy disk drive 160, in one present configuration, can access floppy disks which store up to 1.2 megabytes of data.

The fundamental processing components of the computer 152 are a microprocessor 182 such as, for example, an 80386 microprocessor manufactured by Intel, a math coprocessor 184 such as, for example, a 80387 math coprocessor also manufactured by Intel and a main memory generally indicated at 186 comprising, for example, 4 megabytes of random access memory (RAM). The main memory 186 is used to store certain computer software including a Unix compatible operating system 188 such as, for example, SCO Xenix licensed by Santa Cruz Operation of Santa Cruz, California, a subsidiary of Microsoft Corporation, an image processing application 190, a tile manager 192, and an image data cache 194. The image processing application 190 includes editing functions such as zoom and pan.

Another presently preferred computer workstation 150 having somewhat different processing components from those just described is available from Sun Microsystems, Inc. of Mountain View, California, under the tradename "SPARCstation 1". In such an embodiment, the UNIX compatible operating system would be licensed directly from Sun.

Although a representative workstation has been shown and described, one skilled in the applicable technology will understand that many other computer and workstation configurations are available to support the present invention.

FIG. 7 illustrates a representative configuration of the image data cache 194 some time after the tile manager 192 (FIG. 6) begins operation. A set of compressed tiles 222 are kept at the low addresses of the image data cache 194, and a set of uncompressed (or expanded) tiles 224 at the high addresses of the image data cache 194. The terms expanded or uncompressed are used interchangeably. In between the two sets of tiles 222, 224 is a reserved area 226 (free cache memory). As the operation of the tile manager 192 continues, the image data cache 194 becomes more unordered. As the cache requirement for compressed or uncompressed tiles increases, each set of tiles 222, 224 approach the reserve area 226 from each end. In fact, the reserve area 226 can become completely exhausted.

Since the memory management schemes that apply to compressed data allocation are very different from that of uncompressed data, it is desirable to keep the two sets of tiles 222, 224 separate. Compressed tiles are variable sized tiles (blocks of memory) 222a, b, c, d, e, f whereas the uncompressed tiles are all fixed sized tiles 224a, b, c, d and therefore the locations of the fixed sized tiles 224 are interchangeable. Linked lists of allocated memory are kept sorted according to size and address for compressed tiles. The number of linked lists is a variable number but presently there are about 64 different size categories for compressed tiles and only one size category for uncompressed tiles (for binary images).

To use the image data cache 194, the memory management functions begin by determining how much fast memory (RAM) and slow memory (disk or host memory) is available for image memory uses. When an image

is loaded, the system allocates memory for image information and related tile directory structures. Cache management parameters are modified as necessary to balance the requirements for expanded tile and compressed tile cache memory. The expanded tile cache memory pool and the compressed tile cache memory pool allow tiles from different images to intermingle. Expanded and compressed tiles are kept in separate areas as much as possible so that memory allocation can be optimized for each of two different situations (i.e., fixed allocation block size versus variable size). However, the storage ranges of compressed and expanded tiles are allowed to mingle so as to maximize the flexibility of the cache usage.

FIG. 8 is a state diagram illustrating the flow of image data or tiles between different storage states 250. A tile can contain data in one or more of five states or forms as illustrated by ovals in FIG. 8. The possible forms are: uncompressed and resident in cache memory (state 252); compressed and resident in cache memory (state 256); uncompressed and resident on disk (state 268); compressed and resident on disk (state 262); "not loaded" but re-creatable using information from higher-resolution image tiles (state 272).

For most image access operations, the image data must be uncompressed and resident in cache memory 252. However, that form consumes the most cache memory of any of the five forms. Therefore, a primary function of the tile manager 192 is to transform image tile data between state 252 and the other states which consume less (in the case of state 256) or no cache memory whatsoever (in the cases of states 268, 262 and 272).

The eight transformation operations, shown in square boxes in FIG. 8, constitute the main computational operations associated with managing image memory. The operation "load compressed tile image data from disk into cache memory" 264 is typically the first operation performed on a tile because most pre-scanned images are stored in compressed form in disk files. (A discussion of this "virtual loading" is provided hereinbelow.) The load operation 264 is performed by the LoadCompFromDisk function which simply copies data from the disk into cache memory. The disk location and number of bytes to read is stored in the tile header fields 368 and 376 shown in FIG. 10.

The function LoadCompFromDisk is normally used by the function LockCompHandle when the tile manager 192 needs to access the compressed form of data associated with a tile. LockCompHandle is analogous to LockExpHandle, described in FIG. 28. The LockCompHandle function is also included in source code form in the Microfiche Appendix, in the file tilealloc.c.

Compressed data in cache 256 can be written back to the disk by the operation 260. This is the reverse of the LoadCompFromDisk function. The present embodiment is capable of writing to disk in a wide variety of file formats. One skilled in the art can easily create a function to perform this task.

Compressed data in cache can be uncompressed (also termed "expanded") into another region of cache memory by the expand operation 258. The expand operation 258 is controlled by the "Expand Tile" function 440 which is described with respect to FIG. 33. The method of image compression varies according to image type (e.g. binary, 8-bit color, 24-bit color). Commonly used compression techniques include CCITT T.6 for binary images and CCITT SGVIII (draft standard) for color and gray-scale images. The ExpandTile function 440

selects the appropriate compression algorithm by referring to field 306 of the Document Information Structure shown in FIG. 9.

Uncompressed data in cache 252 can be compressed and written to a separate region of cache memory by the compress operation 254. The compress operation 254 is controlled by the CompressTile function 450 described with respect to FIG. 38. Like ExpandTile, the CompressTile function 450 uses an image compression algorithm appropriate to the image type.

Uncompressed data on disk 268 can also be read directly into cache memory by the load operation 270. The load operation 270 is performed by the LoadExpFromDisk function, which appears in source code form in the Microfiche Appendix, in file diskcach.c. The LoadExpFromDisk function is analogous to LoadCompFromDisk. The LoadExpFromDisk function refers to the fields 362 and 374 of the tile header 350 shown in FIG. 10, for the location and number of bytes of the expanded file data on the disk.

Uncompressed data in cache 252 can be written back to the disk by the save to disk operation 266. This operation is analogous to the save to disk operation 260 which operates on compressed data. The present embodiment can write compressed or uncompressed tile data to disk in a variety of formats. One skilled in the art can easily implement an equivalent function.

Image data for tiles in the "not loaded" state 272 must be constructed by resampling higher-resolution tiles. (During normal operation, only lower-resolution tiles can exist in this state—the full resolution subimage tiles are always "loaded".) The present embodiment provides two operations from the "not loaded" state 272 to the "loaded" state 252, 256. Uncompressed higher-resolution tile data is resampled to create uncompressed data in cache 252 by the resample operation 274. Similarly, in the resample operation 276, compressed data in cache 256 can be created from compressed higher-resolution tile data.

In both resampling operations, extensive advantage is taken of the fact that the resolutions of adjacent subimages in the subimage stack are related by a power of 2. This greatly simplifies and speeds the resampling operation. Basic resampling techniques are well-known (See, for example, A. Rosenfeld and A. C. Kab, *Digital Picture Processing*, Academic Press, 1976). The resampling operation 274 and 276 are controlled by the function LoadSubImTile 436 described with respect to FIG. 31.

In summary, FIG. 8 shows that a great part of the tile manager's utility derives from its ability to coordinate a variety of forms of image data in the course of complex image processing operations.

Generally, the way data starts out on the disk 162 is by loading a tiled image file into an application 190 via the tile manager 192. An image file, like a Tagged Image File Format (TIFF) or CALS tiled image file, for example, can be loaded instantaneously, in a virtual sense. In the tiled formats, there are tiled image data that is stored in the image file and at the beginning of the file there is a directory with entries that locate the tiles (for example, the disk file version of tile 0 in subimage 0, (0,0), is located at one address in the file and the disk file version of tile 1, subimage 0 (0,1) is located at another address in the file). When an image file is loaded, the tile manager 192 gets the tile offsets and stores them in the tile directory and does nothing else. Hence, the image file is basically loaded without copying any data from the disk 162 into the image data cache

194, and a directory is created that maps the tiles in the virtual image memory space onto the disk 162.

FIG. 9A illustrates a document information structure 300. Each image, or document, in the system is associated with (and described by) a document information structure (called "docinfo", defined in FIG. 9). The docinfo structure contains information about the image as a whole, such as color and pixel organization, etc. It also contains a list of subimages contained in the image. Each subimage entry in the docinfo structure contains information about that subimage, such as width and height, etc. The intention is to make this data visible only to cache management functions and low-level access functions. The overall docinfo data structure 300 contains the following information:

- 302 Self-reference to document handle. Handle value assigned to this document by the host procedure which created the document. This value is unique over the entire system.
- 304 "Overviews Invalid" flag. This flag is true if the document is in the middle of a write operation.
- 306 Cache image compression algorithm. Compression algorithm used by the memory manager for this image.
- 308 Image color type. How the image is displayed.
- 310 Bits per image pixel. Number of bits per image pixel.
- 312 Tile size information. Size of expanded tile in pixels. The tiles are assumed to be square.
- 314 Number of subimages in doc. Number of subimages maintained in this document. The minimum value is one (the full resolution subimage).
- 316 Input file info. Input raster file information.
- 318 Output file info. Output raster file information.
- 320 List of subimage headers. Array of pointers to subimage header structures 321. The first entry in the array is always the full resolution image. Each position thereafter corresponds to a 2x resolution reduction from the previous subimage.

The subimage header structure 321 is illustrated in FIG. 9B. Each subimage has its own entry with each field as follows:

- 312 Pointer to tile headers.
- 314 Pointer to tile directory. Pointer to array of pointers to tile header records. This two-dimensional table provides an easy way to access individual tile headers on a (row,col) basis.
- 326 Subimage width and height. The width (x extent) and height (y extent) of the document measured in pixels.
- 328 Number of tile rows & cols in subimage. Number of tile rows in the image and the number of tile columns (i.e., the number of tiles needed to span the height and width of the image).
- 330 Image stack index of this subimage. This is the position of the subimage in the docinfo structure subimage list. It can also be used to determine the factor by which the subimage resolution is reduced relative to the full resolution subimage.
- 332 Pixel resolution of this subimage. Scan resolution in pixels per millimeter.

FIG. 10 illustrates the tile header 350. The tile manager's analog to the conventional address translation table is the tile directory. The tile directory is a two-dimensional array of entries corresponding to the two-dimensional array of tiles that form the image. Each full and reduced resolution image has its own tile directory. The tile directory record contains a list of pointers to

lists of individual tile headers. The list in the tile directory record has one entry for each row of tiles. Each of those entries points to a tile header record list with as many elements as tile columns. Thus, there is one tile directory record per subimage and one tile header record per tile. The tile header record defines the current state of the tile and contains information used by the cache management functions. The tile header contains the following information:

- 352 Pointer to document containing this tile. Pointer to the document to which this tile belongs.
- 354 Index of subimage containing this tile. Index of the subimage (i.e., image stack layer) that contains this tile.
- 356 Row and column indices of tile. Tile row and column position of this tile within the subimage.
- 358 Status information. Defines the current state of the tile. This includes lock counts for expanded and compressed tiles.
- 360 Preserve count. Value greater than zero means the tile is desired for future operation, so the tile should be preserved in cache if possible.
- 362 Location of uncompressed image data in cache memory. Location of uncompressed (expanded) image data for this tile (if it exists). Status flag "ExpCached" will be true to indicate that the data is currently in expanded tile cache memory.
- 364 Location of compressed image data in cache memory. Location of compressed image data for this tile (if it exists). Status flag "CompCached" will be true to indicate that the data is currently in compressed tile cache memory.
- 366 Location of uncompressed image data on disk. Location of uncompressed (expanded) image data for this tile (if it exists). Status flag "ExpOnDisk" will be true to indicate that the data is currently on disk.
- 368 Location of compressed image data on disk. Location of compressed image data for this tile (if it exists). Status flag "CompOnDisk" will be true to indicate that the data is currently on disk.
- 370 Link to next less recently used tile. Pointer to next older (less recently used) tile, not necessarily a tile in this image.
- 372 Link to next more recently used tile. Pointer to next newer (more recently used) tile, not necessarily a tile in this image.
- 374 Number of bytes of expanded data in tile.
- 376 Number of bytes of compressed data in tile.

FIG. 11 illustrates a calling hierarchy 400 for the constituent functions. Further discussions relating to flow diagrams, herein, will include names which correspond to source code modules written in the "C" programming language. The object code is presently generated from the source code using a "C" compiler licensed by Sun Microsystems, Inc. However, one skilled in the technology will recognize that the steps of the accompanying flow diagrams can be implemented by using a number of different compilers and/or programming languages.

The top level in the program hierarchy is Main 402. Main initiates the functions calls to the lower level functions. Main embodies the top level control flow of the present invention.

The first function called by Main is Initialize Cache Manager 404 (InitCacheManager). InitCacheManager allocates the RAM and disk swap space needed for a

particular raster image. It must be called before attempting to load any image tiles into memory.

The next function Main may call is Load Tiled Raster Image 408 (LoadTIFF). LoadTIFF manages the loading of tiled images. This is the process where an existing image file on disk is mapped into memory.

Main will then call the function Begin Undoable Raster Operation 410 (BeginUndoableRasOp). BeginUndoableRasOp marks the beginning of a distinct, "undoable" raster image operation. This function does not save any region of image memory but only creates a new entry on the undo stack. The current version of the tiles in the affected region are saved by InitImageAccess.

The following function called by Main is Create Image Access Context 412 (InitImageAccess). InitImageAccess prepares the tile cache manager for upcoming accesses to a particular region of the specified image. This function creates a data structure called an "access context" (defined in FIG. 16) that is used by the sequential access functions.

Main optionally calls the function Read Rows From Region 414 (ReadRowToRow) next according to the operation performed by the user. ReadRowToRow causes one input/output buffer row or strip to be read and transformed from tiled image memory as specified in the associated InitImageAccess call and the resulting access context.

The next optional function called by Main is Write Rows To Region 416 (WriteRowToRow), again according to the operation performed by the user. WriteRowToRow causes one input/output buffer row or strip to be transformed and written to tiled image memory as specified in the associated InitImageAccess call and the resulting access context.

It should be understood that other access functions, such as random pixel accesses, may optionally be called by Main.

Main then calls the function Close Image Access Context 418 (EndImageAccess). EndImageAccess terminates and discards an image access context. The memory allocated for the access context structure is freed. The tile manager is informed that the specified region of image memory is no longer needed by this operator.

The next function, Undo Previous Raster Operations 420 (UndoPreviousRasOp), is optionally called by Main. UndoPreviousRasOp restores the specified region to its original state using information from the undo stack.

The last function Main calls is Quit Cache Manager 422 (EndCacheManager). EndCacheManager frees the RAM and disk swap space. This function basically reverses what InitCacheManager does.

The second level of functions on the calling hierarchy 400 is shown starting with Load TIFF Subimage Tile Information into Tile Headers 424 (LoadTiffTilesStd) which is called by function LoadTIFF 408. LoadTiffTilesStd manages the loading of TIFF images with strip structure.

The LoadTiffTilesStd function 424 calls a function Store Tile Information in Tile Headers 425 (LoadSubImDiskCache). LoadSubImDiskCache loads the tile directory of the specified subimage with information about the location, size and format of individual image tiles contained in a disk-resident tiled image file. It is the low-level interface for the "indirect file load"

capability. The tile headers are assumed to be completely zeroed when this function is called.

The InitImageAccess function 412 calls a function Save Region For Undo 426 (SaveRegionForUndo). SaveRegionForUndo saves the specified region on the undo stack. It is called from within InitImageAccess if the SaveForUndo flag is true. It can also be used for low level operations that do not go through InitImageAccess. SaveRegionForUndo can then be called multiple times for different documents and different regions within a document so that arbitrarily complex editing operations can be easily undone.

The ReadRowToRow function 414 calls a function Lock Expanded Image Tile Group 428 (ExpTileLock). ExpTileLock "locks" memory handles referring to expanded image tiles. (The notion of locking and unlocking memory blocks is further discussed below with reference to FIG. 14.) It also updates the associated tile header structure as appropriate for the operating system.

The ReadRowToRow function 414 also calls a function Unlock Expanded Image Tile Group 430 (ExpTileUnlock). ExpTileUnlock unlocks memory handles referring to expanded image tiles. It also updates the associated tile header structure as appropriate for the operating system.

The function ExpTileUnlock 430 calls a function Unlock Expanded Tile 432 (UnlockExpHandle). UnlockExpHandle unlocks an individual expanded tile handle. The lock count is decremented as appropriate. The tile is not actually swapped out of cache at this point but it becomes a candidate for swapping.

The function ExpTileLock 428 calls a function Lock Expanded Tile 434 (LockExpHandle). LockExpHandle locks an individual expanded tile handle. The lock count is incremented and the status flags are set as appropriate.

The LockExpHandle function calls a function Create Tile From Higher-Resolution Tiles 436 (LoadSubImTile). LoadSubImTile creates a valid expanded version of the specified tile by scaling down from the next higher-resolution subimage. This function is called recursively as necessary to get to a higher-resolution subimage where there is valid data. (Note: the tiles in the full-resolution subimage are always valid and loaded although not necessarily present in the cache memory.)

The function LockExpHandle 434 next calls a function Allocate Space for Uncompressed Version of Tile 438 (AllocExpHandle). AllocExpHandle allocates space in cache memory for a single expanded tile.

The function LockExpHandle 434 also calls a function Create Uncompressed Version of Tile From Compressed Version 440 (ExpandTile). ExpandTile uses a tile that exists in compressed form but not expanded form, allocates space for an expanded tile and decompresses the image data into that space.

The function LoadSubImTile 436 calls a function Create Compressed Lower-Resolution Tile From Compressed Higher-Resolution Tiles 442 (CompCopyToOview). CompCopyToOview creates a valid compressed version of the specified tile by scaling down from compressed or expanded version of the given higher-resolution subimage tiles. The function LoadSubImTile 436 also calls a function Copy Uncompressed High-Resolution Tiles to Uncompressed Low-Resolution Tile 444 (CopyTileToOview). CopyTileToOview updates the region of the next low-

er-resolution overview corresponding to the specified tile.

The Function `CompCopyToOview 442` calls a function `Collect Freeable Cache Memory 446` (`CollectFreeCache`). `CollectFreeCache` collects freed memory states or enlarges the cache file and adds the new memory capacity to the reserve list. This function is called when the cache manager usage exceeds preset limits. Therefore it makes sense to take time to free up as much memory as is convenient at this opportunity.

The function `CollectFreeCache` calls a function `Free Uncompressed Version of Tile 448` (`FreeExpHandle`). `FreeExpHandle` frees space used for storage of expanded image tiles.

The function `CollectFreeCache 446` also calls a function `Create Compressed Version of Tile From Uncompressed Version 450` (`CompressTile`). `CompressTile` uses a tile that exists in expanded form but not compressed form, allocates space for a compressed tile and compresses the image data into that space.

FIG. 12 is the top-level control flow for the tile manager 192 (also called "Main"). The tile manager 192 can be executed on a number of operating systems or without an operating system. However, the workstation 150 (FIG. 6) preferably includes the Unix compatible operating system 188. Another preferred operating system is Microsoft MS-DOS running with or without Microsoft Windows 3.0.

Moving from a start state 470 to an initialization state 404, the tile manager 192 performs an initialization of the image data cache 194 to determine the available memory space, or the amount of physical RAM and disk space available for a cache "file". At this point, the cache appears to the tile manager 192 as one contiguous range of physical addresses in memory. If the tile cache has already been initialized, this step is skipped. The possibility of multiple image access contexts (discussed below) allows multiple simultaneous requests.

The tile manager 192 has another parameter which is called the fast memory portion of the image data cache 194. This parameter is particularly relevant when working on top of another virtual operating system such as Unix. The fast memory limit specifies approximately how much of the image cache file is actually kept in RAM memory at any moment by the native operating system (e.g., Unix). The balance of data (the less recently used portion) is likely to have been swapped out to the disk. The tile manager attempts to limit the amount of cache space used to store expanded tiles to less than the fast memory limit, but the limit can be exceeded if necessary with some degradation in performance. However, the total cache size limit is never exceeded. In operating systems without virtual memory capabilities built in (e.g., MS-DOS), the fast memory limit is the same as the total cache size limit.

Then the tile manager 192 moves to a function 472 wherein the tile manager 192 loads a tiled raster image file. The function 472 (comprising the function 408, for example) loads any type of image file, and preferably a tiled image, into the memory address space configured by the tile manager 192. If the image to be modified is already loaded, this step is skipped. Then the tile manager 192 moves to a function 410 where the tile manager 192 marks the beginning of an undoable raster operation if the tile manager 192 is writing to the image. The function 410 is an optional state and it is only used if the user wants to be able to undo the operation that modifies the image.

Any time that a region of the image needs to be accessed (for reading or writing) an image access context is created. This image access context is used to define the region for use by the tile manager. The creation is performed automatically by the file manager without effort by the user. For example, an image access context is created when the user draws a line in a region of the image.

Referring back to FIG. 12, the tile manager 192 transitions to a function 412 to create the image access context. The image access context contains all of the state information about the access operation. It is possible to have multiple access contexts opened simultaneously with each access having stored state information contained in the access context. Thus, the tile manager 192 is re-entered and re-used by interleaved operations without confusion due to the unique access contexts of each image operation.

The tile manager 192 proceeds to a loop state 474 wherein the tile manager 192 begins a FOR-loop for all of the rows or columns in the region. The FOR-loop is executed multiple times if the operation specified by the user is a row or column strip oriented access. Strips are composed of one or more rows or one or more columns of data. For each of the strips, the tile manager 192 reads or writes the rows or columns of data in the strip in a function 476. The function 476 actually comprises a set of functions including `ReadRowtoRow 414` (FIG. 11) and `WriteRowtoRow 416`.

When the tile manager 192 has processed all the row and columns in the region, the tile manager 192 moves to a function (`EndImageAccess`) 418 where the tile manager 192 closes the image access context which frees all of the temporary buffers that were allocated for the image access context.

The tile manager 192 transitions to an undo previous raster operation function (`UndoPreviousRasOp`) 420. This causes a modified image to revert to its previous state. The image tiles that had been modified are replaced by their original versions. This again is an optional step that the user initiates, if a mistake is made.

If the raster image is required for future operations, the tile manager moves to state 422. Otherwise, moving to a state 478, the tile manager 192 unloads the raster image. Unloading the raster image simply frees the memory that had been associated with that particular raster image. This is not a save raster image operation which would be slightly more complicated, but a save operation could be executed here. Of course, the image processing application 190 supports loading and saving raster images.

If more operations will be performed the tile manager moves to state 480. Otherwise, from state 478, tile manager 192 moves to a quit cache manager function (`EndCacheManager`) 422. Herein, the tile manager 192 frees the image data cache 194 (FIG. 6). Presumably, all of the images have been unloaded as in the state 478 so that this operation frees the image data cache memory and prepares the system for shut down. Lastly, the tile manager 192 terminates at an end state 480.

FIG. 13 illustrates the initializing of the cache manager function 404. The function 404 is entered by the task manager 192 at a state 488. Then, moving to a state 490, the task manager 192 initializes the cache usage variables. Of course, in the beginning, all of the cache space is available for use, in what is called the free-memory reserve list. That is, no cache memory is being used for expanded or compressed image data.

At state 492, the task manager 192 allocates tile cache memory by requesting a portion of the address space from the memory space owned by the operating system. In a virtual memory system such as Unix, the request is handled by memory mapping a large file. The operating system does not allocate any memory, but it reserves an address space. Moving to a state 494, the task manager 192 allocates a common blank tile. When dealing with binary images, space is reserved for one blank tile, which is kept around at all times for common usage by any number of operations, or access contexts.

At state 496 a compression buffer is allocated to be used as a scratch buffer when compressing data since, in general, the size of the resulting compressed data is unknown before a tile of image data is compressed. Hence, compressed data blocks will be variable sized. The tile manager 192 then exits the InitCacheManager function 404 at an end state 498.

FIG. 14 illustrates a general memory state diagram with reference to a block of memory being "locked" or "unlocked". In the diagram, ovals are states and rectangular blocks are operations.

The state diagram is entered at a start state 502 by a new memory block. There are three basic states. "FREE" is a state 504 where there is no memory allocated. Actually, a block of memory is considered free if it is in one of the memory free lists, i.e., the "reserve free list", the "compressed free list" or the "expanded tile free list". It should be understood that the free list for the compressed tiles are actually composed of many lists based on the varying sizes of memory blocks.

Within a tile header (FIG. 10) the tile manager 192 controls a memory handle which is a structure that has a pointer to (or location of) image data in the cache and a lock count (not shown) for both compressed and expanded versions of a tile.

A memory block transitions from the free state to unlocked, but allocated is through a state 506 for allocating the memory handle, which moves the block out of the free list and into use by a tile. As opposed to free, unlocked means that the memory block contains valid data and that it is associated with a tile but not currently being accessed. That is, the block is not being read or written at the time.

Now, the tile is unlocked at a state 508 but it contains valid data. Therefore, the next step is to lock the block, or lock the memory handle at a state 512 and then it becomes a locked memory state at a state 514. That means it contains valid data and it is currently in use. The block can be locked more than once, each time just incrementing the lock count.

The lock count may be incremented multiple times, for example, when two access contexts (operations) are accessing the same region of memory. Hence, both contexts lock the block of memory or tile by incrementing the lock count. When the first access context is done it decrements the lock count. But the tile manager 192 knows that that tile is still in use by an access because the locked count is still non-zero.

The inverse operation is to unlock the handle at a state 516 and as long as the lock count is not decremented to zero at state 518, it stays locked. Once the lock count is decremented to zero, it becomes unlocked again at the state 508.

An unlocked tile is fair game for the tile manager 192 when the memory manager needs to find some space to lock a new tile. Therefore, when the tile manager 192 is

looking for space, unlocked memory blocks may be freed and returned to the free memory lists.

The way to go from the unlocked state 508 to the free state 504 is by freeing the handle in which case the memory block is moved onto the free memory list.

Referring now to FIG. 15, the flow diagram for the InitImageAccess function 412 shows the operation where the tile manager 192 creates the image access context starting at a state 530. At a state 532 the input parameters are validated. If there is an error with the input parameters, the function ends immediately at an end state 534.

Input parameters include a document handle indicating which image that the user wants to read or write from. Thus, the document handle must be validated. Another parameter is whether the user wants to read or write to the image. A transformation matrix, also input, basically directs how to scale, rotate, shear, etc., the image data.

If the input parameters are valid, the tile manager 192 locks the document handle at a state 536. The document handle locks and unlocks just like other structures and resources in the tile manager and it prevents one user of a particular document or image from modifying or deleting that image while another operation or another access context is still using that document.

Then, at a state 538, the tile manager 192 tests whether a non-orthogonal rotation has been specified. For example, a rotation of 30° causes the tile manager 192 go into a special operation that initializes the access with rotation. That also creates an access context but after a more involved process. Then the tile manager 192 ends the function 412 at a state 534 with a valid access context for rotations.

If an orthogonal rotation is specified then the tile manager 192, allocates a conventional access context at a state 542. Then the tile manager 192 continues to a decision state 544 wherein the subimage selection criterion is specified. For instance, the user may request the "low resolution" option which selects the lowest resolution subimage in the document's image stack. (In the context of an image editor, this may be the best solution during zooming or panning.) The user may also specify "most available"—i.e., whatever subimage has tiles currently in cache memory, regardless of the resolution. In either case, the tile manager 192 proceeds to a state 546 to select the reduced resolution subimage that is appropriate to that particular choice, i.e., either the one that has the resolution just greater than what was requested or a subimage whose tiles covering the access region are currently in cache. Now, at a state 548, the tile manager 192 adjusts the transformation matrix so as to now refer to the reduced resolution subimage rather than the full resolution subimage by adjusting scale factors.

Alternatively, if the state 544 determines that the full resolution subimage is selected then the transformation matrix is unchanged. Proceeding to a state 552, the pixel and tile limits of the affected image region are calculated. Knowing these limits, in a state 554, the tile manager 192 creates a temporary directory for the tiles in that region. This directory is a two-dimensional array that references the tiles that contains the affected pixels. Later on the tile manager 192 refers to the region tile directory because it is specific to tiles that are inside the affected region.

The tile manager 192 then initializes the image scaling functions in a state 556. Such scaling functions presently

used are the subject of applicant's concurrent application entitled "Process for High Speed Rescaling of Binary Images" (U.S. Ser. No. 08/014,085, filed Feb. 4, 1993, which is a continuation of Ser. No. 07/949,761 filed Sep. 23, 1992, now abandoned, which is a continuation of Ser. No. 07/693,010 filed Apr. 30, 1991 now abandoned.

Moving on, the tile manager 192 tests whether polygonal clipping is required at a state 558. For example, a request may be made to only read from within a specific polygonal region. If that is the case, the tile manager 192 initializes the polygonal region clipping functions in the tile manager 192 by passing in the boundary lists. The polygonal clipping function translates the boundary lists into edge lists that are used to very efficiently read out the rows or columns of data.

For example, suppose a "flood" request is made to turn all of the pixels black within an octagonal region. One way to accomplish the operation is to specify the points of the corners of the octagon in image coordinates and pass that in with the initialization of access context request, which would pass those vertices of the polygon into the polygonal clipping function set up function.

Then the tile manager 192 comes to a state 562, where the tile manager 192 allocates buffers for scaling, if necessary. This is the situation where intermediate copies of the rows or columns of data may need to be kept during the process of scaling. Then the tile manager 192 tests whether the user specified that the region needed to be saved for undoing, at a decision state 564.

An important feature of the present invention is an "undo" operation that is integrated with the image memory management so that only compressed tiles need to be saved after an undoable edit operation. In this way, a user can easily and quickly retract an edit operation that is no longer desired. For example, in mapping applications, e.g., USGS Quadrangle maps, the impression of a very large map is desired, but it is really composed of smaller map quadrants that were separately scanned, trimmed, adjusted and fit together. The smaller maps can be visually and logically joined into a single, large image. Using the present invention, a user can add a feature, such as a new sub-division, town, or road, that crosses a map boundary, specifying that the feature is undoable. Later, the user can remove the feature modification to the image by specifying the undo operation.

Now at a decision state 568, the question is whether to update the subimages during the operation. If this is a write operation the tile manager 192 always writes into the full resolution subimage and the changes "trickle down" into the low resolution subimages. But the tile manager 192 has an option as to whether the lower-resolution tiles are updated during the modification operation or later when the tiles are requested for viewing operations. There are advantages in doing them both ways.

For example, if the affected region is small, it is more efficient to update the subimages while progressing through the operation. In this mode, when the tile is unlocked, the manager 192 immediately copies the data down into the next lower subimage tile but only one of the corners of the tile is affected. Thus, only portions of the low resolution subimage tiles need to be modified.

If, however, the subimages are not updated during the operation, then as soon as the image access context is created all of the subimage tiles that overlap the af-

ected region are invalidated (they become "not loaded"). Hence, when the memory manager goes to access them again at some later time, it has to reconstruct them from the higher-resolution tiles. The advantage of that is that the memory requirement at any one moment is half of that of if the tile manager 192 was updating all of the tiles simultaneously. In this way, the tile manager 192 sets a flag at a state 570.

In state 572 the tile manager 192 "preserves" the affected tiles in the affected subimages. Again, it relates to whether the tile manager 192 is updating subimages or not. If the tile manager 192 is reading, then it preserves only the tiles in the region of the subimage that will be accessed.

The ability to "preserve", or preferentially retain tiles that will be accessed in the course of the operation, is an important feature of the present invention that can yield significantly higher performance in certain situations where memory capacity limitations are encountered. When a tile is "preserved" for a particular access operation, it's preserve count 360 is incremented. The cache manager treats tiles with non-zero preserve counts differently from tiles with zero preserve count. The cache manager will discard unlocked unpreserved tiles before discarding older preserved tiles. (The cache manager normally discards older or less recently used tiles before discarding newer or more recently used tiles.)

Then, within the creation of the access context, the tile manager 192 actually locks down the first row or column of tiles in the region to establish the cache memory requirement for this operation, at a state 574. If this succeeds, then the caller is assured that there will be sufficient cache space for the entire operation.

The tile manager 192 can perform row or column accesses. However, the following discussion only refers to a row access.

Then, at a decision state 576, if the tile manager 192 cannot satisfy the request to lock down that first row of tiles, the function 412 terminates at the end state 578. Otherwise, at state 580 the tile manager 192 initializes the row access functions.

Now, once the tile manager 192 has initialized the row access function in state 580 the tile manager 192 invalidates the affected subimage tiles if the tile manager 192 is writing to the full resolution subimage at a state 582. Finally, in a state 584 the tile manager 192 returns the handle or a pointer to this access context to the user. From then on the user just uses this pointer to the access context and pointers to input and output buffers to get the next row or column of data.

FIG. 16 illustrates the access context structure 600. The structure 600 operates on a high level to hide the low level operation from the user and contains book-keeping information along with some memory management information. The access context 600 contains the following information:

602 Pointer to affected doc. Pointer to the document being accessed.

604 "Subimage Choice" option value. Specifies how to choose which of the subimages will be read from or written to.

606 Index of affected subimage. Index of the specific subimage directly affected by this access context.

608 Access quantum. Specifies "granularity" of image access.

610 Read/write option. Specifies what type of image memory accesses to prepare for (e.g., read or write).

- 612** Basic orthogonal rotation value. Specifies the image rotation in terms of how the bits in each buffer row are read from or written to the image (e.g., write buffer row to image column with increasing "y" coordinate). 5
- 614** Pixel combination operation. Specifies the pixel operation performed when combining the buffer contents and image contents. The results of the operation are stored in the output buffer when reading. The results go into image memory when writing. 10
- 616** Scaler type operation. Specifies the type of scaler preferred. In other embodiments, this may include fast low-accuracy scaling and line width-preserving scaling. 15
- 618** "Update overviews" flag. True flag indicates overview subimages should be updated in the course of this modification of the full resolution image. This causes the overviews to be correct when the access is complete. 20
- 620** I/O buffer width & height. Width (i.e., row length), total number of rows to process and pitch in pixels of the input/output bitmap.
- 622** I/O buffer pitch (bytes/row). Pitch of the input/output buffer in bytes used for multi-row accesses. The input/output buffer is assumed to be a contiguous memory bitmap at least as large as the access quanta. It is always read or written in the natural order (by rows, low address to high). Flipping and rotation is always done on the image memory side. 30
- 624** I/O buffer bit offset to start of run. Indicates where the buffer's x=0 pixel lies within the first long word of the buffer's storage space. It must be between 0 and 31 inclusive. This parameter allows the caller to match up with arbitrary bit alignments. 35
- 626** Rows per strip (for AQ_STRIP access quantum). When operating in the AQ_STRIP mode, this specifies the maximum number of rows per input/output strip. Fewer rows may be written into the last strip if the end of the access region is hit before the strip is filled. 40
- 628** Number of I/O buffer rows yet to be processed. This variable is used in the access routines to keep track of the number of input/output rows remaining for the access operation. 45
- 630** Pointer to access function used in "Seq-BuflmageAccess". Pointer to the image access function that is tailored to the specific access mode requested. 50
- 632** Stepping directions for image row and column indices. The stepping increment each time the input/output buffer is advanced one row and one pixel. The allowed values are +1, 0, and -1.
- 634** Pointer to polygon clipping information. Refers to an edge table structure for controlling polygonal boundary clipping. 55
- 636** Pointer to raster scaling information. Tile level access information used by lower level modules in the course of the operation. 60
- 638** Pointer to uncompressed data in currently locked tiles. Pointer to an array of pointers directly into expanded tile image data. This list is used to accelerate sequential access into image memory. As each new tile row or column is encountered in a sequential access, this array is set to point directly into the affected tiles, which have been brought into cache memory and locked down. In other 65

- embodiments this could also be used to point to compressed tiles.
- 640** Pointer to region tile directory. Pointer to a 2-dimensional array of pointers to the tiles in the affected region of the subimage.
- 642** Next image row & column to be accessed. The index of the next image row and column to be accessed in sequential row and column operations.
- 644** Terminal row & column of access region. Stopping values for sequential row and column operations.
- 646** Unclipped extent of access region. Defines the image region that will be accessed over the course of the operation.
- 648** Clipped extent of access region. Defines the portion of the requested image region that actually falls within the boundaries of the image. Pixels outside of this rectangle are treated as background pixels.
- 650** Clipped image buffer bit offset and length. These values specify where, in the intermediate image row or column buffer, the first bit from the clipped image region is located and how many bits are to be read from or written to tiled image memory.
- 652** Number of tile rows & cols in access region. Number of tile columns and rows in the affected region.
- 654** Row & column of currently locked tiles. Column and/or row index of the currently locked tile or tiles.
- 656** Image row & col at origin of first tile in access region. Pixel coordinates of the upper-left pixel in the upper-left tile of the affected region.
- 658** Number of I/O buffer rows held over for next strip. Number of rows of output data that did not fit into the previous row and must be returned in the next and subsequent rows when expanding while reading image data.
- 660** Pointer to image tiling/untiling buffer. Points to a temporary buffer to hold data extracted from tiled memory prior to scaling when reading from image memory.
- 662** Number of bytes in tiling/untiling buffer. Size of buffer in bytes.
- 664** Bit offset for tiling/untiling buffer. Bit offset to the first valid pixel in tiling/untiling buffer.
- 666** Access transformation matrix. The transformation matrix mapping input/output buffer pixels onto the pixels of this subimage.
- FIG. 17** illustrates the flow diagram for the "Save Region for Undo" function 426 as referenced in FIG. 15. The tile manager 192 starts at a state 680, moves to 682 where the tile manager 192 locks the document handle of the affected document that contains the region to save for undo. The tile manager 192 can save multiple regions from multiple documents sequentially and then undo them all in one operation later. Thus, the application programmer is allowed to easily undo multiple-region operations with a single undo call at a later point.
- Moving to a state 684, the tile manager 192 clips the modified region to the image boundaries since there is no information to save outside of the image. Then the tile manager 192 moves to a decision state 686 wherein the tile manager 192 tests whether the affected region overlaps the image. If there is no overlap, that is to say, there is no image data to save, then the tile manager 192 moves to a state 688 where the tile manager 192 unlocks

the document handle and terminates the function 426 at an end state 690.

If, however at state 686, the modified region does overlap the image, the tile manager 192 moves to a state 692 wherein the tile manager 192 allocates memory for an "undo region header". The undo region header is similar to a document header, but reduced comparatively in the amount of data conveyed therein. The undo region header will be associated with tile header information, etc.

The tile manager 192 then moves to a state 694 where the tile manager 192 allocates memory for "undo region tile headers". These tile headers will be used to store copies of the original versions of the tiles in the affected region. The tile manager 192 then proceeds to a state 696 wherein the tile manager 192 makes an "undo tile directory".

Then the tile manager 192 moves to a loop state 698 where the tile manager 192 loops for each tile row in the region. The tile manager 192 then transitions to a loop state 700 wherein the tile manager 192 loops again for each tile column in the region (Thus, there is a two-dimensional loop.)

The tile manager 192 moves from the state 700 to a decision state 702 where the tile manager 192 checks to see if that particular tile in the document is loaded in the image cache memory. If the tile is not loaded, the tile manager 192 skips to the next tile in the region by returning to the loop state 700. Otherwise, if the tile is loaded, the tile manager 192 marks the undo copy of the tile as loaded in a state 704.

Note that there are two tiles. One is the original version of the tile that is still associated with the document and the second is the copy that the tile manager 192 is going to make and associate with the undo region header.

At a decision state 706, a test determines whether the document tile is blank. If the tile is blank (i.e., all background color), then the tile manager 192 moves to a state 708 and simply marks the undo tile as "blank" and returns to the FOR-loop at 700. If the document tile is not blank, then the tile manager 192 moves to a state 710 and the tile manager 192 marks the undo tile as "not blank" and moves to a state 712 wherein the tile manager 192 tests whether the document tile has a valid copy of compressed data on the disk.

If a valid copy of compressed data does reside on disk, the tile manager 192 moves to a state 714 and simply copies the compressed tile disk location and size information from the document tile header to the undo tile header. Note that it is possible for a particular tile to have multiple representations of the same data. That is, a compressed version and an expanded version of the tile may exist in cache simultaneously. And a tile may have a compressed version in cache as well as on the disk. For undo, the strategy is to store the most compact version possible. The most compact version with regard to cache memory usage is to have a copy of the compressed tile on the disk.

If there is no compressed copy of the tile on the disk, the tile manager 192 proceeds to a decision state 716 wherein the tile manager 192 determines whether an uncompressed copy of the document tile resides on the disk. If the test succeeds, the tile manager 192 enters a state 718 and copies the uncompressed tile disk location and size information from the document tile to the undo tile and then returns to the inner FOR-loop at a loop state 700.

If, at state 716, there is no uncompressed tile information on the disk, the tile manager 192 continues execution to a state 720 in FIG. 17B wherein the tile manager 192 locks the compressed version of the document tile. This locking of the compressed version of the document tile may cause an expanded version of the document tile to be compressed and a compressed version created. Therefore, there is a possibility of an error and that is checked at the decision state 722.

If there is an error than the tile manager 192 unlocks the document handle at a state 724 and terminates with an error condition at the end state 726. If there was no error in locking the compressed version of the tile then the tile manager 192 moves from the state 722 to a state 728 wherein the tile manager 192 allocates and locks down cache memory for a copy of the compressed data to be associated with the undo header. There is another error possibility at this point and the tile manager 192 checks for an error at a decision state 730. If there is an error then the tile manager 192 returns to a state 724 and thereafter terminates the function 426.

If there was no error in locking cache memory at the state 730, the tile manager 192 moves to a state 732 and copies the compressed data from the document tile to the undo tile. The tile manager 192 actually copies the data that is stored within the tile—i.e., the compressed image data is copied from the document version to the undo version. Then the tile manager 192 moves to a state 734 and unlocks the compressed version of the document tile. Now, at a state 736, the tile manager 192 unlocks the compressed version of the undo tile and the tile manager 192 returns to the inner FOR-loop at state 700 on FIG. 17A where the tile manager 192 loops back to continue the loop for all of the tiles in the affected region.

When the tile manager 192 is done with all of the tiles in the affected region, the tile manager 192 moves to a state 738 where the tile manager 192 links the new undo header into the undo region list. Thus, multiple regions can be saved in the undo list and then in one operation, by calling undo previous raster operation, all of the operations that had been accumulated, can be undone. Then the tile manager 192 moves to a state 742 wherein the tile manager 192 unlocks the document handle and terminates the function 426 normally.

FIG. 18 shows the load tile to raster image function (LoadTiff). FIG. 18 is a flow diagram for the part of LoadTiff that loads tiled images only. In reference to FIG. 18, the overall process may be understood whereby an existing file on the disk, i.e., an image file on disk, is mapped into memory. As described below, the overall process permits loading large images in a short time period relative to how long it would take to actually copy all of the image data into the computer's memory. In accordance with the present invention, the process shown in FIG. 18 is called the indirect loading capability. As shown in FIG. 18, the tile manager 192 begins the LoadTIFF function 408 at a start state 750 and moves to a state 752 where the tile manager 192 opens the input file that is on the disk. If there is an error on the disk, the tile manager 192 prints an error message at a state 754 and terminates at an end state 756. If no error exists, then the tile manager 192 moves to a state 758 and checks for the TIFF header structure that identifies that the input file is in fact a TIFF file. While the disclosure below discusses a TIFF file, it is to be understood that the process shown in FIG. 18 may be per-

formed on all types of tiled files, such as a MIL-R-28002A Type II file or an IBM IOCA tiled file.

Still referring to FIG. 18, if the tile manager 192 finds something other than TIFF header structure at state 758, the tile manager 192 moves to state 754 to indicate an error, and then exits at the end state 756. If the tile manager 192 finds a TIFF header structure while at state 758, the tile manager 192 move to a state 760, wherein the tile manager 192 counts the number of subimages in the TIFF file, one or more of which may exist in a TIFF file.

Next, the tile manager 192 moves to a state 762 and reads the full resolution subimage information which constitutes the basic information about the image, e.g., the image width and height, the size of the tiles, the compression format that is used, and the resolution. If the basic image information is not present and in proper form, the tile manager 192 moves to the state 754 to indicate an error. On the other hand, if no error is indicated at state 762, the tile manager 192 moves to state 764, wherein the tile manager 192 creates a skeleton document and locks that document. The skeleton document at this point contains no cache memory but only tile directory and tile headers that represent in a virtual sense the tiles that compose the image.

The tile manager 192 next moves to a state 766 where the TIFF full resolution subimage tile information is loaded into the tile headers for the full resolution subimage, as more fully disclosed below in reference to FIG. 19. Next, the tile manager 192 moves to a loop state 768 where there is a loop for each of the remaining lower resolution subimages. While in this loop, the tile manager 192 accesses a decision state 770, wherein the tile manager 192 determines whether

$$fr/lr=2^n \quad (1)$$

where

fr is the full resolution subimage resolution in pixels per inch; and
lr is the particular low resolution subimage resolution in pixels per inch.

If the ratio of fr to lr is a power of two, then a successful test is indicated, and the tile manager 192 moves to a function 424 and loads the TIFF subimage tile information into the tile headers for that particular subimage level. On the other hand, if the ratio of fr to lr is not a power of two, as indicated at the decision state 770, then the tile manager 192 ignores the particular subimage under test and returns to the state 768 until all of the subimages in the file are processed. When all subimages have been processed, the tile manager 192 moves to a state 772 and unlocks the document handle of the newly created document and terminates normally at an end state 756.

Now referring to FIG. 19, the function 424 whereby the tile manager 192 loads the TIFF subimage tile information into tile headers is shown. More particularly, the tile manager 192 begins at a start state 780 and moves to a state 782 wherein the tile manager 192 reads the number of tiles in the subimage. Then the tile manager 192 moves to a state 784 wherein the tile manager 192 allocates temporary buffers for the tile mode offset and byte count lists. These three lists have one entry each per tile in the subimage. If the tile manager 192 cannot properly allocate the temporary buffers, then the tile manager 192 exits with an error condition at an end state 786.

Upon successful allocation of the buffers, the tile manager 192 moves to a state 788 where the tile man-

ager 192 reads the tile offset and byte count information from the disk file into the allocated buffers. In the TIFF file standard, all tiles are stored in the same mode (e.g., compressed). However, other tiled file formats (e.g., MIL-R-28002A Type II) specify the storage mode for each tile. The tile mode simply states whether a particular tile is stored in compressed form, in uncompressed form, or whether the tile is all foreground or background color. The tile manager 192 next moves to a state 790 where the tile manager 192 fills in the tile storage mode list. At state 790, the tile manager 192 synthesizes the tile mode information that the TIFF file does not contain itself. Then the tile manager 192 moves to the function 425 wherein the tile manager 192 stores the information in the subimage tile headers (FIG. 10), and terminates at an end state 786.

Now referring to FIG. 20, the function 425 whereby the tile manager 192 stores file information in tile headers is shown. The tile manager 192 begins this process at a start state 800 and moves to a state 802 where the tile manager 192 locks the document handle of the document for which the tile manager 192 is loading the subimage for. This function is performed once per subimage in the file and there may be multiple subimages in the file. Consequently, the locking of the document handle function can be performed several times in the process of loading a single document.

As shown in FIG. 20, in the event that an error occurs in locking the document handle the tile manager 192 terminates at an end state 804. On the other hand, if the tile manager 192 successfully locks the document handle at state 802, the tile manager 192 moves to a state 806 where the tile manager 192 determines whether the number of tiles in the file matches the number of tiles expected for the particular subimage in the particular file or document. If a mismatch exists between the actual and expected number of tiles, the tile manager 192 moves to a state 808 to print an error message and then terminates at the end state 804. On the other hand, in the event that the number of actual tiles matches the number of expected tiles, the tile manager 192 moves to a loop state 810 where the tile manager 192 enters the first part of a FOR-loop for each tile row. Still referring to FIG. 20, the tile manager 192 moves from state 810 to state 812 for each tile column. Accordingly, it will be understood that the tile manager 192 is processing a two-dimensional array at the states 810, 812.

In accordance with the present invention, the tile manager 192 processes, at states 810, 812, all of the tiles required to cover the particular subimage. Next, the tile manager 192 moves to a decision state 814 wherein the tile manager checks the value in the tile mode entry to determine whether the tile data is compressed. If the tile data is compressed, the tile manager 192 moves to a state 816 and stores the file offset and byte count in the compressed tile handle. The compressed tile handle is a part of the tile header structure, and the file offset is the location of the compressed data for the particular tile within the file as measured by a byte offset from the start of the file. The byte count represents the number of bytes of compressed data associated with the particular tile starting at the offset that is provided at the tile. From state 816, the tile manager moves to state 828, wherein the tile manager sets a flag to indicate that the particular tile is not blank.

In the event that the tile manager determines at state 814 that the tile data is not compressed, the tile manager

192 moves to a decision state 818 where the tile manager 192 checks to see if the data is uncompressed. If the data is uncompressed on the disk, the tile manager 192 stores the file offset byte count information in the uncompressed tile handle in state 820. From state 818, the tile manager moves to state 828, wherein the tile manager sets a flag to indicate that the particular tile is not blank.

If the tile manager 192 determines at state 818 that the tile data is not uncompressed, then the tile manager 192 moves to state 822, wherein the tile manager 192 checks to see whether the tile is all foreground at a state 822. For example, in a black and white drawing engineering document, foreground color is black, so the tile manager 192 treats a foreground as a black tile. If the tile is determined to be a foreground tile, the tile manager 192 proceeds to state 824, wherein the tile manager 192 creates an all foreground tile, and then sets the flag as not blank at state 828. As an example, if the image being processed is a color image, the tile manager 192 could fill the tile with the foreground color at the state 824.

On the other hand, if the tile is not all foreground, the tile manager proceeds to state 826 to determine whether the tile is all background. As discussed above, binary images usually have background pixels which are white or zero value. If a particular tile is blank, the tile manager 192 moves to a state 828 where the tile manager 192 sets the blank flag to indicate that the tile is indeed a blank tile. If at the state 826 the tile manager 192 determines that the tile is not all background, the tile manager 192 terminates with an error at an end state 830. In other words, having determined at state 822 that the particular tile was not all foreground, the only possibility left at state 826 is that the tile is all background. Consequently, a determination at state 826 that the tile is not all background indicates an error.

From state 828, the tile manager 192 moves to a state 832 and sets the loaded flag to true indicating that a valid image information set has been associated with the particular tile. The tile manager 192 completes the loop described above for each tile. After having processed each tile in the particular image, the tile manager 192 exits the two FOR-loops and moves to a state 834 where the tile manager 192 unlocks the document handle and then terminates normally at the end state 830.

Now referring to FIG. 21, the tile manager 192 performs a function which for purposes of the present invention will be termed "Undoable Raster Operation". The function shown in FIG. 21 is performed by the tile manager 192 in the function "Begin Undoable Ras-Op", and is a relatively simple function, the purpose of which is to clear the undo region list. More particularly, in the process shown in FIG. 21, the tile manager 192 frees all of the undo regions associated with the previous operation to prepare for a new undo operation. Indeed, the present invention could be configured to have multiple level undo, i.e., the system of the present invention could undo two or three or more operations going into the past and also to be able to redo all of those operations at the user's choice. For example, the last three operations could be undone and then the oldest of those operations redone.

In specific reference to FIG. 21, the tile manager 192 begins at a start state 840 and then proceeds to loop state 842, in which the tile manager 192 executes a FOR-loop for each undo region in the current list. The tile manager 192 loops to a state 844 where the tile manager 192 frees all of the memory associated with that undo re-

gion. This may include freeing compressed data that is stored in cache or expanded data that is stored in cache and associated with the undo region. When the tile manager 192 finishes all of the regions, the tile manager 192 terminates at an end state 846.

Now referring to FIGS. 22A and 22B, there is shown the control flow for the ReadRowToRow function 414 which produces one or more rows of scaled image data each time it is performed. It is one of the basic image access functions. It should be understood that the tile manager 192 can also read columns of an image, etc., so as to produce a rotated output.

The tile manager 192 enters the function 414 by moving to a start state 850 and proceeds to a decision state 852 where the tile manager 192 checks for a region overrun. In other words, when the access context is created, the region that is going to be read in the course of the overall operation is specified, and in the event that the read row to row subfunction is accessed too many times, the region will be overrun. Any such overrun is detected by the tile manager 192 at state 852 and reported at state 854. In the event of an overrun, the tile manager 192 terminates at an end state 856.

If, on the other hand, no region overrun has occurred, the tile manager 192 moves to a decision state 858 where the tile manager 192 checks to see whether old results are carried over to the new strip. Such a carryover could occur when, for example, raster data is being enlarged by expanding one or more lines from the image. For example, when raster data is being enlarged by 4x, each line of input generates four (4) lines of output. Accordingly, three (3) output rows could be carried over for later strips. With this eventuality in mind, the tile manager 192 ascertains whether any data is being carried over and if so, the tile manager 192 uses the carried-over data before generating a new row. Consequently, if there is new data carried over, the tile manager 192 moves to a state 860 where new rows are generated from the carried over data.

Next, the tile manager 192 moves to a state 862 where the tile manager 192 checks to see if a particular strip is full. For purposes of the present invention, a strip is a collection of rows, i.e., a set of numbers arranged in rows. As indicated at state 862, if the strip is full, then the tile manager 192 ends at the end state 856.

If the strip is not full and the tile manager 192 has used up all the carried over data, then the tile manager 192 moves to a decision state 864 where the tile manager 192 checks for ghosting, i.e., the skipping of some rows of data in order to produce a low quality image while panning or zooming. If ghosting is in effect, the tile manager 192 moves to state 866, wherein the tile manager 192 calculates the number of blank lines to create. The system then moves to a state 868 where the tile manager 192 writes the blank lines to the output strip buffer.

From state 864, if no ghosting was detected, or state 868, if ghosting is not in effect, the system moves to state 870 where the tile manager 192 again checks to see if the strip buffer is full. If it is, the tile manager 192 exits at the end state 856. If it is not, the tile manager 192 checks to see that there are still input rows to read in a decision state 872. If there aren't, the tile manager 192 has reached the end of the specified image region, and proceeds to state 874 to obtain another row of output data by flushing the scaler buffers. In accordance with the present invention, in the state 874 the tile manager 192 sets a flag that is subsequently passed down to the

scaler functions to flush intermediate results from the scaler functions. This is the case when for reducing data, i.e., if a plurality of rows is being combined into one output row. That is how the last output row is produced.

From state 874, the system moves to state 894, shown in FIG. 22B. On the other hand, in the event that there are no unread image rows at state 872, the system moves to decision state 876, where the system determines whether the row is outside of the valid image boundaries. If yes, the system moves to a state 878, where the tile manager 192 substitutes blank lines for the input. The tile manager proceeds from state 878 to a state 894, shown in FIG. 22B. If the answer to the decision at state 876 is no, the system moves to a decision state 880, shown in FIG. 22B, to check whether the row is contained in the currently locked tile row.

At state 880, the tile manager 192 moves down the image, and the system sequentially passes through successive tile rows. Each tile contains, e.g., 512 rows, so when a particular tile row is locked it stays locked until all 512 image rows in that tile row have been read. Each time the system arrives at a new row it tests to see that the row is contained in the currently locked tile row. If it is not, the system moves to the state 430 (function ExpTileUnlock) to unlock the old tile row and lock down the new tile row (at state 428). In addition, the tile manager 192 has to unpreserve the row of tiles that was just unlocked. Unpreserving them tells the memory manager that those tiles are no longer needed for this access operation and it can do what it wishes with them.

Next, the system proceeds to a decision state 882 to determine whether any tiles are blank. If they are, the tile manager 192 substitutes a reference to a "common blank tile" and that common blank tile is used, as indicated at state 884. All tiles that are blank are mapped onto this common blank tile. Consequently, the tile manager 192 uses less image memory.

From state 884, 882, or 880, as appropriate, tile manager 192 proceeds to a decision state 886 to check for polygonal clipping. If the tile manager 192 is doing polygonal clipping then each input row of data is clipped as appropriate for that polygon in states 888 and 890. The loop allows multiple clipped regions within each row. If there is no clipping, then the tile manager 192 simply copies the entire input row from the image into the input row buffer in a state 892. Then the tile manager 192 move to a state 894 where the tile manager 192 passes these input rows through the scaler if the tile manager 192 is scaling the data. Finally, the tile manager 192 takes the results of the scalers and copies that information to the output strip buffer if necessary at a state 896. The tile manager 192 then returns to the state 870 (shown in FIG. 22A) where the tile manager 192 continues the process of retrieving input rows and scaling them until the tile manager 192 has filled the output strip buffer. The system then moves to the termination condition at the end state 856.

Now referring to FIG. 23A, a process which will be referred to as "Write Rows to Region" will be described. The tile manager 192 starts at state 900 and moves to state 902 where the tile manager 192 tests for region overrun. Region overrun can occur when the calling function attempts to write more rows to the image than was specified when the access context was created. If the region was overrun, the tile manager 192 reports an error at state 904 and terminates with an error at state 906. If there is no region overrun, the tile

manager 192 moves to the FOR-loop in state 908 where the tile manager 192 loops for each input row in the input buffer, which is the buffer that is passed in by the calling function. It contains the data that is to be processed and written to the image. The loop is executed for each row and moves to state 910 where the input data is passed through the scaler functions and put into a temporary buffer. If the scaler does not always produce an output row, as is the case when reducing the resolution, a plurality of input rows may have to be combined to produce a single output row. So, at the state 912, the tile manager 192 determines whether an output row was produced after the input row is scaled. If not, the tile manager 192 goes back to the loop at state 908 and continues the process as described. On the other hand, when the tile manager determines at state 912 that an output row was produced, the tile manager 192 moves to state 914 which is a FOR-loop for each copy of the scale row to write to the image. It may be the case that more than one copy of the scaled row needs to be written into image memory. This is the case when the tile manager 192 is expanding the input image data. It may be that one input row is replicated four times to get a 4x expansion factor.

Next, the tile manager 192 moves to state 916 where the tile manager 192 checks to see if the destination row index is outside of the image's clipping boundaries. If so, the tile manager 192 simply ignores it and moves back to state 914. If it is within the clip boundaries the tile manager 192 moves to state 918 where the tile manager 192 determines whether the destination row is in the currently locked tile row. If it is not, the tile manager 192 moves to state 920 where the tile manager 192 unpreserves and unlocks the old tile row that is currently locked. The tile manager 192 then moves to state 922 to determine whether the update overview flag is true. This is an option that is specified in the lo access context and it determines how lower-resolution tiles are updated when the full resolution subimage is modified. If the update overview flag is true, then the tile manager 192 moves to state 924 where the tile manager 192 unpreserves the low resolution tiles that will no longer be needed.

After the system has unpreserved the low resolution tiles that are no longer needed at state 924, the system moves to state 926 and locks down the new tile row. Only the full resolution tile row is locked at this level. The low resolution tiles are actually updated when the call to unlock the old tile row is made.

Next, the tile manager 192 moves to state 928 to determine whether an error was detected when the new tile row was locked. If so, the system terminates with an error condition at state 906. If there is no error or if in state 918 the tile manager 192 finds that the destination row is currently in the locked tile row, the tile manager 192 moves to state 930 in FIG. 23B. At state 930, the tile manager 192 determines whether polygonal clipping is activated. If it is, the tile manager 192 computes the clip points for the current image row, as indicated at state 932, which results in a list of clip point pairs.

The tile manager 192 then moves to state 934, wherein the tile manager 192 conducts a FOR-loop for each of the clip point pairs that the tile manager 192 computed in state 930. As shown in FIG. 23B, the tile manager 192 loops to state 936 where the tile manager 192 copies pixels from a scaler output buffer to the image row between each pair of clip points. When that loop terminates, the tile manager 192 returns to state

914 in FIG. 22A. On the other hand, if the tile manager determines at state 930 that polygonal clipping is not active, the tile manager 192 moves to state 938, wherein the tile manager 192 copies the scaler output buffer pixels to the image row without clipping. The tile manager 192 then proceeds to state 914.

Now referring to FIG. 24, the tile manager starts at state 950 in the end access function shown in FIG. 24 and proceeds to state 952. At state 952, the system cleans up after row or column access functions by freeing buffers used by the row or column access functions.

Next, at state 954, the tile manager 192 unlocks the last row or column of tiles accessed. Then, the system moves to state 956 where the tile manager 192 un-preserves any tiles in the region that are still preserved. The system may perform the functions at states 954, 956 when an operation was aborted in mid-progress and it cleans up after those partially completed operations.

At state 958, the tile manager 192 cleans up after the polygonal clipping function. If there was polygonal clipping involved in this access context the tile manager 192 has to free the buffers that contain the polygon edge information.

Next, the system moves to state 960, where the tile manager 192 frees scaler buffers, the temporary tile directory, etc.. From state 960, the system moves to state 962, wherein the tile manager 192 unlocks the document handle to indicate to the memory manager that the access context no longer is referring to the particular document associated with the document handle.

The tile manager 192 next moves to state 964 where the memory that was used to store the data for the access context is freed. Then, the system ends the clean up function at state 966.

Referring now to FIGS. 25A,B, a function is shown which, for purposes of the present invention, will be termed the "Undo Previous Raster Operations". The tile manager 192 starts at state 970 and moves to state 972, wherein the tile manager determines whether any undo regions exist in the list or if the list is empty. If no regions exist then the tile manager 192 moves to end state 974 and terminates normally.

If the tile manager 192 determines at state 972 that "undo" regions do exist, the tile manager 192 moves to state 976, where the tile manager 192 enters a loop for each undo region in the list. In this loop, the tile manager 192 moves to state 978 where the tile manager 192 locks the affected document handle. The document handle that is locked is the one that was stored in the undo region header that tells where that particular undo region came from. The tile manager 192 moves from state 978 to state 980 where the tile manager 192 saves the current document region to support redo (i.e. an "undo" operation following by another "undo" operation). Then the tile manager 192 moves to state 982 to invalidate the affected tiles in the lower-resolution subimages. The strategy represented by states 980, 982 in FIG. 25A is to save the minimum amount of information that is needed to reconstruct the image, which means the tile manager 192 saves only the affected tiles in the full res subimage.

Next, the system moves to a loop indicated by the states 984, 986. In this loop, for each tile, the tile manager 192 moves to state 988, discarding the document tile image data. Then the tile manager 192 moves to state 990 to determine whether the undo tile is loaded. If it is not loaded, the tile manager 192 moves to state 992

where the tile manager 192 marks the document tile as "not loaded". If the tile is determined to be loaded at state 990, the tile manager 192 moves to state 994 to mark the document tile as "loaded". From state 994, the system moves to state 996 in FIG. 25B.

At state 996, shown in FIG. 25B, the tile manager 192 determines whether the undo tile is marked as blank. If it is, the tile manager 192 moves to state 998, wherein the tile manager marks the document tile as blank, and then the system loops back to state 986. If the undo tile is determined to be not blank at state 996, the tile manager 192 move to state 1000. At state 1000, the tile manager 192 checks to see if the undo tile points to compressed data on the disk. If it does, the tile manager 192 moves to state 1002 and copies the disk location and size information about the compressed data into the document tile header and loops back around. If there is no compressed data on the disk, then the tile manager 192 moves from state 1000 to state 1004, wherein the tile manager 192 determines whether uncompressed data exists on the disk associated with the undo tile.

If so, the tile manager 192 moves to state 1006, wherein the file manager 192 copies the disk location and size information about the uncompressed data into the document tile header and loops back to state 986. If the system determines at state 1004 that there is no uncompressed data on the disk, the tile manager 192 proceeds to state 1008, wherein the tile manager 192 determines whether the undo tile "points" to uncompressed data in cache memory. If it does, the tile manager 192 moves to state 1010, wherein the tile manager 192 copies the pointer to the uncompressed data from the undo header to the document tile header.

From state 1010, the system returns to state 986. If no uncompressed data exists in the cache, however, as determined in state 1008, the tile manager 192 stores a pointer to the compressed data in cache in the document tile header and returns to state 986.

Referring back to FIG. 25A, when the tile manager 192 has completed the loop described above, the system moves to state 1014, unlocking the document handle. From state 1014, the tile manager 192 proceeds to state 1016, wherein the tile manager 192 frees the memory associated with the undo header. The tile manager 192 then moves to state 976. Thus, the system returns to state 976 for each undo region in the list. As intended by the present invention, the tile manager 192 continues the loop for all of the regions in the list. The undo regions are restored in "last-in-first-out" order. At the completion of the looping process described above, the system moves to state 974.

Now referring to FIG. 26, when the tile manager 192 ends the cache management, the tile manager 192 starts the process shown in FIG. 26 at state 1020 and proceeds to state 1022 wherein the system frees the compression buffer. From state 1022, the system proceeds to state 1024, wherein the system frees the common blank tile. Next, the system moves to state 1026 to free the tile cache memory. The system then ends the process shown in FIG. 26 at state 1028.

FIG. 27 provides an explanation of the function exp tile lock. The tile manager 192 starts at state 1040 and moves to state 1042 where the tile manager 192 enters a FOR-loop for each tile row to be locked. In accordance with the present invention, the system in the exp tile lock function is capable of locking down all the tiles in a two dimensional region.

For each tile in the specified region, the system moves to state 1046, wherein the tile manager 192 determines whether the particular tile is blank. To make this determination, the system examines flags in the tile header itself or checks the image data for that tile to determine if there are any non-background pixels. If it is not a blank tile, the tile manager 192 moves to state 434 where the tile manager 192 locks the uncompressed version of the tile. Then the tile manager 192 proceeds to state 1050, wherein the tile manager 192 determines whether an error had occurred in the process of creating the uncompressed version of the tile. If no error is found at state 1050, the tile manager 192 continues to loop to the next tile in the region by returning to state 1044. If an error did occur, as determined at state 1050, the system proceeds to state 430 to unlock previously locked tiles, and then ends at state 1056.

In the event that the tile manager 192 at state 1046 detected that the particular tile was a virtual blank tile, i.e., a tile that exists only by virtue of the fact that there is a tile directory entry for that tile, the tile manager 192 take no action, other than to loop back to state 1044 for further processing.

FIG. 28 illustrates the control flow for the "lock expanded tile" function 434 wherein the tile manager 192 takes a single tile and locks the expanded version of the tile in the image data cache 194. The tile manager 192 enters the function 434 at a start state 1060, and proceeds to a decision state 1062 wherein the tile manager 192 tests whether the tile is marked as "loaded". As already mentioned, a loaded tile is one that either contains or references valid image data, is either uncompressed or compressed image data, and it either resides in cache memory or on the disk. If the tile is not loaded, the tile manager 192 moves to a function 436 wherein the tile must be created from higher resolution tiles which are loaded. Afterwards, the tile manager 192 determines if there was an error in a decision state 1066. If there was an error, the tile manager 192 terminates the function 434 at an end state 1068 and reports the error condition. Otherwise, if there was no error in creating the tile, the tile manager 192 continues, moving from the state 1066 to a decision state 1070.

The tile to be locked is now loaded so the tile manager 192 tests whether the uncompressed version of the tile is in cache memory. The objective of the function 434 is to guarantee that there is an uncompressed version of the tile in cache memory. Now, if the uncompressed version is not in the cache, the tile manager 192 proceeds to a decision state 1072 to determine whether the selected tile is a blank tile.

If the tile is blank, the tile manager 192 proceeds to a state 438 to create a blank tile. Note here that the function ExpTileLock 428 (FIG. 27) will detect a blank tile before calling the function 434 if it can take advantage of using a common blank tile at a higher level. In other words, if the tiles are locked for reading only, i.e., the image data will not be modified in any way, then all blank tiles can refer to the same section of blank memory. However, if the tiles are locked for writing, all tiles must have their own memory because different image data can be written to the different tiles.

At this point, state 438, memory has presumably been allocated for a blank tile. Moving to a state 1074, the tile manager 192 tests whether there was an error and moves to the end state 1068 if there was an error.

Returning in the discussion to the decision state 1072, if the tile is not blank, then the tile manager 192 transi-

tions to a decision state 1076 and tests whether there is a uncompressed version of that tile on the disk. If the uncompressed version is on disk, then the tile manager 192 reads that uncompressed version from the disk into cache memory at a state 1078. Then the tile manager 192 moves to the state 1074 to test for errors.

If, at the state 1076, there is not an uncompressed version on the disk, the tile manager 192 moves to the function 440 so as to create the tile from the compressed version. The compressed version can be either in cache memory or on the disk, and this is handled by the function 440. Again, the tile manager 192 checks for an error at the state 1074.

Now, assuming that there was no error found at the state 1074, the result is that the tile manager 192 has an uncompressed version of the tile in cache. Therefore, the tile manager 192 proceeds to a decision state 1080 to verify that the uncompressed version is valid. It is sometimes the case that the uncompressed version of a tile is locked by one access context and then for some reason it is invalidated by another access context. This happens when the first access context is reading an uncompressed version of a tile from a lower resolution image, and another access context is actively modifying the full resolution subimage with a particular setting of parameters. If the tile not valid, the function 434 is terminated at the end state 1068.

Alternatively, a valid tile that was determined at the state 1080 causes the tile manager 192 to increment the uncompressed data lock count for that tile at a state 1082. The lock count starts out at zero for an unlocked tile and can increment as high as necessary. However, the lock count will be decremented once for each unlocking operation. It is important to match the number of times a tile is locked with the number of times the tile is unlocked. Otherwise, the tile would end up in a permanently allocated (unfreeable), locked state.

Proceeding to a decision state 1084, the tile manager 192 tests whether the tile is locked for writing or for reading. If the tile manager 192 locked the tile for writing, the execution of the function 434 continues to a state 1086 wherein the "blank" status flag is invalidated. The blank status flag is actually a combination of two flags. One that says that the tile is blank or not blank and the second flag that says if the first flag is valid or not. The reason for two flags is that the way to detect that a tile is blank is by searching through all the pixels in that tile. To do so every time the file is accessed would be wasteful so occasionally, truly blank tiles won't be handled as blank tiles. Hence, there is a second flag that is set, in the state 1086, when the first flag is invalid. The second flag indicates that the tile must later be examined to determine whether it is still blank.

The tile manager 192 next moves to a state 1088 to invalidate the disk-resident, uncompressed version of the tile, if one exists. This is because the tile manager 192 will modify the cache-resident version of the tile. To synchronize the cache-resident and disk-resident versions, the disk-resident version is invalidated. Then, at a state 1090, the tile manager 192 invalidates and frees the compressed versions if they exist.

A compressed version of the tile may be in cache or on the disk and, at the state 1090, the tile manager 192 cleans both out of memory. Thus, at the end of the "lock for writing" operation, the only valid version of the tile is the expanded version in cache, which at this point is locked. Then the tile manager 192 continues to a state 1092 to move the newly locked, expanded ver-

sion of the tile to the front of the "most recently used (MRU)" list of uncompressed tiles.

The MRU list is a doubly-linked list wherein, starting at the beginning, the tile is found that was most recently used, then the next most recently used, and so on, the last tile was used the longest time ago. That list is used by the cache manager to determine which tiles are least likely to be used again as a second level of criteria.

Finally, the tile manager 192 terminates the LockExpHandle at the end state 1068.

FIG. 29 illustrates the control flow for the "unlocking expanded image tile group" function 430. The function 430 is just the reverse of lock expanded image tile group. In other words, there is a region of locked tiles which must be unlocked because the access to the tiles is complete. Generally, the two functions, ExpTileLock and ExpTileUnlock are called for a row or column of image data rather than a region but an entire region lock/unlock is possible.

The tile manager 192 enters the function 430 at a start state 1110. The loop states 1102 and 1104 represent the beginning of nested FOR-loops. That is, the outer loop, beginning at the state 1102, unlocks a row of tiles, and the inner loop, beginning at the state 1104 unlocks a column of tiles. Moving from the state 1102, to the state 1104, and then to the function 432, the tile manager 192 unlocks the uncompressed version of the tile. When all the tiles in the region are unlocked, the tile manager 192 terminates the function 430 at an end state 1108.

Now referring to FIG. 30, the tile manager 192 enters the UnlockExpHandle function 432, referred to in FIG. 29, at a start state 1110. The tile manager 192 proceeds to a decision state 1112 to test whether the uncompressed version of the currently selected tile is in fact locked, i.e., whether the lock count is non-zero. If the tile is not locked, the tile manager 192 exits the function 432 at an end state 1114.

If, at the state 1112, the tile is found to be locked, the tile manager 192 moves to a state 1116 to decrement the lock count. Thereafter, the execution continues to a decision state 1118 wherein the tile manager 192 tests whether the "update overview" flag is set true. If the flag is set, the tile manager 192 moves to a state 1120 to update the corresponding lower-resolution tiles. In the process of modifying tiles, the tile manager 192 locks a tile down in the image data cache to write to it. When the tile is unlocked, that is a signal to the memory manager to update the lower resolution tiles that correspond to the higher resolution tile. Thus, the image data in the high resolution tile being unlocked is copied down into the lower resolution tiles, all the way down to the bottom of the image stack.

Once the lower resolution images are modified, or if the overviews are not being updated, the tile manager 192 proceeds to a decision state 1122 to test whether the lock count is exactly zero. If the lock count is not zero, the tile manager 192 terminates the function 432 at the end state 1114.

Otherwise, the tile manager 192 moves to a state 1124 to clear the "cache collection delay" flag. The cache collection delay flag is set by the tile manager after unsuccessfully trying to reduce the expanded memory usage of the cache file. It is cleared in the function 432 because there is now the possibility of freeing the tile that was just unlocked. In other words, the tile can be removed from the cache to create some space. This flag prevents the tile manager or the cache manager from making repeated, unsuccessful attempts to create space.

After the tile manager 192 clears the flag, execution proceeds to a decision state 1126 to determine whether the uncompressed version of the tile is invalid. As explained hereinabove, it is possible for one access context to have the expanded version of the tile locked down and another access context to invalidate the data in that tile. The tile must remain in memory until the first access context unlocks the tile. Once it is unlocked and the lock count is decremented to zero, if the tile is invalid, the tile manager 192 moves to a state 1128 to free the uncompressed tile version, or remove the tile from the image data cache. In either case, the tile manager 192 terminates the function 432 at the end state 1114.

FIG. 31 illustrates the control flow for the "create tile from higher-resolution tiles" function 436 referred to in FIG. 28. The tile manager 192 begins the function 436 at a start state 1140 and proceeds to a decision state 1142 to determine whether the tile is in fact already loaded, in which case no further processing is needed and the tile manager 192 terminates the function 436 at an end state 1144. Assuming that the tile is not loaded, the tile manager 192 moves to a decision state 1146 to test whether a higher resolution subimage exists.

This function is called only for lower resolution subimages where the tile manager 192 can create the lower-resolution tiles from higher-resolution tiles. Hence, higher-resolution subimages must exist for the function to succeed. If no higher-resolution subimages exist, the tile manager 192 reports the error and terminates the function 436 at the end state 1144.

If the higher-resolution subimage does exist, the tile manager 192 proceeds to a state 1150 to calculate the indices of, or locate, the four higher-resolution tiles that reduce to this tile. There are four tiles involved because the preferred resolution step between subimage levels is two in the presently preferred embodiment. Thus, since there are two dimensions, four higher-resolution tiles are required to produce each next lower resolution tile.

Thereafter, the tile manager 192 enters a FOR-loop at a loop state 1152. For each of the four higher-resolution tiles, the tile manager 192 tests whether the tile is loaded in the image data cache, at a decision state 1154. If the tile is not loaded, then the tile manager 192 moves to a state 1156 wherein a recursive call is made to the "load subimage tile" function to create the corresponding higher-resolution tile from yet higher-resolution tiles. This case occurs if a the tile is a few layers down in the image stack and the tiles in all but the full resolution subimage had been invalidated. Therefore, the function 436 invokes itself to work all the way back up to the top level, recreate the higher-resolution tiles and then work back down to the tile of interest. Only higher-resolution tiles that map to the particular lower-resolution tile need be loaded.

Assuming that all the higher-resolution tiles have been loaded, the FOR-loop terminates and the tile manager 192 proceeds to test whether all of the higher-resolution tiles are blank. If all four of the high resolution tiles mapped to this low resolution are blank, the tile manager 192 transitions to a state 1160 to mark the low resolution tile as blank. The tile manager 192 does not create any image data for the blank, lower-resolution tile. The tile manager 192 and terminates the function 436 at the end state 1144.

If, however, one or more of the higher-resolution tiles is not blank, the tile manager 192 moves to a state 1162 to make a determination as to whether it is faster to create the lower-resolution tile by scaling the com-

pressed version of the higher-resolution tiles or the expanded version of the higher-resolution tiles. An algorithm is used at the state 1162 to decide which is faster and depends on the machine that the program is running on, and other considerations. If it is faster to scale the compressed data the tile manager 192 moves to the function 442 to create the compressed, lower-resolution tile directly from the compressed higher-resolution tiles.

Now, if it is determined that it is faster to scale the expanded version of the data, the tile manager 192 moves from the state 1162 to a state 1166 to allocate memory for the uncompressed version of the lower-resolution tile. From the state 1166, the tile manager 192 moves to the beginning of a FOR-loop at a loop state 1168 wherein for each of the higher-resolution tiles the tile manager 192 scales the expanded version of the higher-resolution tile directly into the proper position in the lower-resolution tile using the function 444. When the tile manager 192 has scaled each of the four high resolution tiles, the tile manager 192 has completed the creation of the expanded version of the low resolution tile.

The tile manager 192 then proceeds, from either of the states 1168 or 442 to a decision state 256 wherein the tile manager 192 determines if an error was incurred in that process. If there was an error, the tile manager 192 moves to a state 1172 to report the error. From either of the states 1170 (if no error) or 1172, the tile manager terminates the function 436 at the end state 1144.

FIG. 32 contains the flow diagram for the "allocate space for uncompressed version of tile" function 438 referred to in FIG. 28. The tile manager 192 enters the function 438 at a start state 1180 and moves to a decision state 1182 to test whether the "soft" uncompressed cache usage limit is exceeded. The soft uncompressed cache limit is a number that is cast into the tile manager 192 during initialization and it basically sets a guideline for how much of the image data cache is to be devoted to uncompressed image data. If the cache manager gets a request for uncompressed cache space and finds that this soft limit has been exceeded, it attempts to reduce the amount of expanded image data that is held in cache either by compressing expanded tiles or by discarding expanded tiles that have valid compressed versions or some other way to recreate them.

If the tile manager 192 finds that the soft limit is exceeded, the tile manager 192 moves to a state 1184 to first check whether the "cache collection delay" flag is set. This flag is set after an unsuccessful attempt to reduce cache memory usage and prevents repeated unsuccessful calls to collect free cache at a state 1186.

Thus, the tile manager 192 will not try to reduce the expanded memory usage until the flag is cleared in the "unlock expanded tile handle" function 432 (FIG. 30).

If the cache collection delay flag is not set, the tile manager moves to a state 1186 to collect free cache memory by freeing uncompressed tiles. After that, the tile manager 192 moves to a decision state 1188 to test whether the soft uncompressed cache usage limit is still exceeded after an attempt to reduce the memory usage. If the usage is still exceeded, the tile manager 192 prints a warning message on the video display 154 (FIG. 6) at a state 1190 and then sets the cache collection delay flag at a state 1192.

Returning in the discussion to the state 1182, if the soft limit was not exceeded, or if it was not exceeded at the state 1188, the tile manager 192 moves to a decision

state 1194 to determine whether there is memory available in the uncompressed tile free list. If there is not memory available in the uncompressed tile free list, then the tile manager 192 moves to a decision state 1196 to determine whether there is memory available in the cache reserve list. If there is no memory available there, the tile manager 192 moves to a state 329 wherein the tile manager 192 again tries to collect free cache space by unlocking or freeing both uncompressed and compressed tiles. At this point, the tile manager 192 must free space in order to allocate space for this uncompressed tile. The tile manager 192 moves to a state 1200 to determine whether memory is now available in the cache reserve list. In the state 1198, when the cache memory space is freed, it is placed into the cache reserve list. If memory is not available, then the tile manager 192 moves to a state 1202 and prints a "cache overflow" error message and terminates the function 438 with an error condition at the end state 1204.

Now, taking an alternate path from the states 1194, 1196 and 1200, if the tile manager 192 can successfully get space for the uncompressed tile data, then the tile manager 192 moves to a state 1206 where the tile manager 192 finds the free block with the highest memory address. If there is a choice between two or more free memory blocks, the tile manager 192 chooses the one with the highest address to try to keep all of the expanded image data at the high address end of the cache file. Once the tile manager 192 finds the highest address block, it moves to a state 1208 to unlink the free block from the free memory link list.

There are actually two possibilities for the free memory link list when the tile manager 192 is looking for expanded memory. One is the uncompressed tile free list and the other is the cache reserve list. In either case, the tile manager 192 unlinks the block of memory that the tile manager 192 is interested in from the free list and relinks the remaining memory blocks of the affected free list.

The tile manager 192 then transitions to a state 1210 to initialize the newly allocated block to all background color. Then the tile manager 192 moves to a state 1212 to move the description of the memory block (a pointer to the tile header) to the front of the most recently used tile list. Moving to a state 1214, the tile manager 192 updates the soft uncompressed cache memory usage counter that was checked at the state 1182. The tile manager 192 continues to a state 1216 to store the memory address in the tile header. The memory block that the tile manager 192 has just allocated is a pointer that is stored in the tile header data structure. That is how the memory block is associated with the tile. Then the tile manager 192 terminates normally from the function 438 at the end state 1204.

FIG. 33 illustrates the process by which the present invention expands the compressed version of a tile to create an uncompressed version. Specifically, as shown in FIG. 33, the tile manager 192 starts at a start state 1220 and moves to a test function at state 1222, where the tile manager 192 determines whether the compressed version of the tile, or the compressed tile data, is in cache memory. If it is not, then the tile manager 192 moves to state 1224, wherein the system loads the necessary data from the disk. If there is an error detected at state 1224, the tile manager 192 moves to state 1228 to terminate the process.

From state 1226, if compressed data was successfully loaded from the disk or from state 1222 if it was in cache

to begin with, the tile manager 192 moves to state 1230, wherein the tile manager 192 locks the compressed tile image data. This step simply increments the lock count on the compressed memory state. From state 1230, the system moves to state 1232, wherein the tile manager 192 allocates and locks the uncompressed tile memory block. The system then moves to state 1234 to determine whether an error occurred at state 1232. If so, the tile manager 192 moves to state 1236 and unlocks the compressed tile data. From state 1236, the system moves to state 1238 to report the error. The system then terminates at end state 1228.

On the other hand, if no error existed as determined at state 1234, the system moves to state 1240, wherein the tile manager 192 uncompresses the compressed data. Next, the tile manager 192 moves to state 1242 to determine whether an error occurred at state 1240. If an error occurred at state 1240, the tile manager 192 moves to state 1236 and functions as described previously. Otherwise, the tile manager 192 moves to state 1244 to unlock the compressed and uncompressed data, and then terminates at end state 1228.

FIG. 34 illustrates a process for creating compressed low resolution tiles from compressed higher resolution tiles. The tile manager 192 starts at start state 1250 and proceeds to state 1252, wherein the system enters a loop which is followed by the system for each of the four high resolution tiles required to produce a single low resolution tile. More specifically, at state 1252 the tile manager 192 locks the compressed version of the high resolution tile. The system then proceeds to state 1256, wherein the tile manager 192 determines whether an error occurred at state 1254. In the event that an error occurred, the tile manager proceeds to end state 1258 and terminates. If no error occurred, the tile manager 192 returns to state 1252 and continues the loop described above for each of the four high resolution tiles.

After processing all four high resolution tiles as described, the system proceeds to state 1260 where the tile manager 192 scales the compressed data to half resolution. The process performed at state 1260 results in a compressed version of the low resolution tile. Then the tile manager 192 moves to a loop represented by states 1262, 1264, wherein for each of the high resolution tiles the tile manager 192 unlocks the compressed version of the tile.

Next, the tile manager 192 moves to state 1266 where the tile manager 192 allocates and locks memory for the compressed version of the low resolution tile. At state 1266, the tile manager 192 actually puts the compressed version of the low resolution tile in a general, common buffer that is large enough to hold the maximum possible size of the compressed results. The actual valid data is usually much less than that than the maximum possible size, so the tile manager 192 only saves the valid amount of data.

From state 1266, the system moves to state 1268 to determine whether an error occurred at state 1266. If an error occurred, the system moves to end state 1258 and terminates. Otherwise, the system moves to state 1270 where the tile manager 192 copies the compressed data out of the temporary compressed data buffer into the newly allocated space in the cache. Then the tile manager 192 moves to state 1272 where the tile manager 192 unlocks the compressed version of the low resolution tile that now contains valid data. The system then terminates normally at state 1258.

Now referring to FIG. 35, a process is shown whereby the system resamples uncompressed high resolution tiles to an uncompressed low resolution tile. The tile manager 192 starts at start state 1280 and moves to state 1282, wherein the tile manager 192 locks the uncompressed version of a single high resolution tile. This function scales a single high resolution tile to update one quarter of a tile in the half-resolution subimage. That quarter tile is rescaled to update one-sixteenth of a tile in the quarter-resolution subimage. This continues to the lowest resolution subimage. Next, the tile manager 192 proceeds to state 1284 to determine whether an error occurred in locking the uncompressed version of the high resolution tile. If there was an error, then the tile manager 192 proceeds to state 1286 and terminates with an error condition. Otherwise, the tile manager 192 moves to state 1288 where the tile manager 192 determines how many levels of the subimage are to be updated. This function can be used to update a subset of subimages or the entire image stack in the case where a single tile is modified in the full resolution subimage. It will propagate that change all the way down to the lowest-resolution subimage in the image stack.

Next, the tile manager 192 proceeds to state 1290 where the tile manager 192 determines the tile index that is to be updated. In accordance with the present invention, when a change is propagated from the higher resolution down to the low resolution of tiles, the system calculates which tile corresponds to the affected area. Then the tile manager 192 moves to state 1290 where the tile manager 192 determines whether the low resolution tile that the tile manager 192 is about to update is marked as loaded or not. This step is intended for the situation in which not all of the low resolution sub-states are populated during the loading of a raster image.

If the system determines that one or more low resolution tiles are not loaded, the system proceeds to state 1294, wherein the tile manager 192 invalidates all of the low resolution tiles that would otherwise be affected by the change. The system then exits normally at end state 1286. If the low resolution tile is about to be modified is loaded, as determined at state 1292, the tile manager 192 moves to state 1296, wherein the system locks the uncompressed version of the low resolution tile. The tile manager 192 then moves to state 1298 to determine whether an error occurred at state 1296 and, if so, the system moves to end state 1286 to terminate. Otherwise, the system moves to state 1300, wherein the tile manager 192 scales the raster data from the high resolution tile down to the low resolution tile. Then the tile manager 192 moves to state 1302 where the tile manager 192 unlocks the high resolution tile.

Next, the system moves to state 1304, wherein the tile manager 192 recursively modifies the loop variables such that the low resolution tiles that the tile manager 192 just finished updating become the high resolution tiles for the next succeeding iteration. Once all the subimages have been updated as described, the system exits at end state 1286.

Now referring to FIGS. 36A and 36B, a process to collect free cache is shown. This process can be called from several other processes. The tile manager 192 begins at start state 1310 in FIG. 36A and moves to state 1312 to determine whether a cache collection operation is in process. If so, the system exits at end state 1314. This prevents recursive calls to collect free cache which might otherwise occur. If the system at state

1312 determines that no collection is in progress, then the tile manager 192 moves to state 1316 where the tile manager 192 sets a flag indicating that a collection is in progress.

From state 1316, the system moves to state 1320, where the tile manager 192 estimates the number of memory blocks to free in this operation. The reason for freeing a number of blocks instead of just one block is to reduce the computational overhead associated with the cache collection operations. The tile manager 192 typically estimates the amount of memory required to equal the number of tiles in a single row of the full resolution subimage of the document associated with the most recently used tile.

Once this estimate has been made, the system proceeds to state 1322 wherein the tile manager 192 considers the options that the tile manager 192 passed into this function. There are three options. One, as indicated at state 1324, is to reduce the uncompressed cache usage only while not affecting the compressed data that is currently held in cache. The second option, indicated at state 1328, is to reduce the compressed cache memory usage only. The third option, indicated at state 1326, is to reduce the total cache memory usage including both compressed and uncompressed data.

From state 1324 or state 1326, the tile manager 192 moves to state 1330, where the tile manager 192 stores all of the free states currently in the uncompressed free list into the cache reserve list. As the tile manager 192 performs the process in state 1330, the tile manager 192 attempts to consolidate the memory blocks. That is, if there are two free blocks that are adjacent to one another, the system automatically turns them into a single, larger contiguous block. From state 1328, on the other hand, the system moves to state 1358, shown in FIG. 36B and discussed below.

From state 1330, the tile manager 192 moves to state 1332, wherein the tile manager 192 determines whether the tile manager 192 has created a memory block large enough to satisfy the initial request. If so, the tile manager 192 terminates normally at end state 1314. Otherwise, the tile manager 192 moves to state 1334 where the tile manager 192 frees any unlocked, uncompressed tiles which are blank. The tile manager 192 then moves to state 1336 where the tile manager 192 determines whether the tile manager 192 has free sufficient memory. If so, the tile manager 192 exits at end state 1314. Otherwise, the tile manager 192 moves to state 1338 where the tile manager 192 frees unlocked, unpreserved uncompressed tiles that have valid compressed versions in cache or are on a disk, or that have valid, uncompressed versions on the disk beginning with the least recently used tile. After having freed that particular class of tiles, if the tile manager 192 determines, at state 1340, that the memory request has been satisfied, the tile manager 192 moves to state 1314 and terminates. Otherwise, the tile manager 192 moves to state 1342, shown in FIG. 36B.

Now referring to FIG. 36B, the tile manager 192 begins at state 1342, wherein the tile manager 192 compresses the free unlocked, unpreserved uncompressed tiles that don't have a valid compressed version or other source from which the tile can be recreated. To do this the tile manager 192 processes expanded tile data through a compression algorithm. The tile manager 192 then creates a compressed version of that tile so that the uncompressed version of the tile can be discarded.

Next, the tile manager 192 moves to state 1344, wherein the system determines whether the request made at state 1342 has been satisfied. If so, the system terminates at end state 1346. Otherwise, the system moves to state 1348, wherein the tile manager 192 frees unlocked, but preserved uncompressed tiles that have valid compressed or uncompressed copies. The tile manager 192 preferentially frees the oldest such tiles.

From the state 1348, the tile manager 192 proceeds to a decision state 1350 to test whether the request made at the state 1348 was satisfied. If so, the function 446 is terminated at the end state 1346. Otherwise, the tile manager 192 moves to a state 1352 to compress and then free unlocked, but preserved, uncompressed tiles that do not have valid compressed versions.

Next, the tile manager 192 moves to state 1354, wherein the system determines whether the request made at state 1352 has been satisfied. If so, the system terminates at end state 1346. Otherwise, the system moves to state 1356, wherein the tile manager 192 determines whether to free data memory blocks. If not, the system terminates at state 1346. Otherwise, the system moves to state 1358, to free unlocked preserved, uncompressed tiles that don't have valid compressed versions already.

The system next moves to state 1360 to determine whether the request has been satisfied. If so, the system terminates at state 1346. Otherwise, the system moves to state 1362 to print an error message, and then terminate at state 1346.

Now referring to FIG. 37, the tile manager 192 starts at state 1380 and moves to state 1382 where the tile manager 192 determines whether the uncompressed version is in fact still locked—that is if the lock count for uncompressed version of that tile is non-zero. If the tile is still locked then the tile manager 192 moves to state 1384 and prints a warning message. Then the tile manager 192 terminates at end state 1386.

If, at state 1382, the system determined that the uncompressed version is not locked, then the tile manager 192 moves to state 1388 where the tile manager 192 determines whether the uncompressed data has already been freed. If it has then the tile manager 192 terminates at end state 1386. Otherwise, the tile manager 192 moves to state 1390 where the tile manager 192 unlinks the uncompressed memory state from the most recently used list.

From state 1390, the tile manager 192 moves to state 1392 where the tile manager 192 updates and decrements the total uncompressed memory usage counter by the appropriate amount. The tile manager 192 then moves to state 1394 where the tile manager 192 moves the memory block to the uncompressed memory free list. In accordance with the present invention, the tile manager 192 keeps the list sorted by decreasing address. Consequently, when the tile manager 192 allocates expanded memory blocks, the tile manager 192 tends to choose the preferred blocks that have higher addresses because they are at the front of the free list.

Next, the tile manager 192 moves to state 1396, wherein the tile manager 192 sets a pointer in the tile header to null and the tile manager 192 sets the uncompressed tile status flags. This ensures that the tile header reflects the fact that it no longer has an uncompressed data associated with it. Then the tile manager 192 terminates at end state 1386.

Now referring to FIG. 38, a process by which the system compresses a tile is shown. The system begins at

start state 1400, and moves to state 1402, wherein the tile manager 192 determines whether the uncompressed tile data is in cache memory. If it is not, the tile manager 192 moves to state 1404 and loads the uncompressed data into cache memory from the disk. The system then moves to state 1406, to determine whether an error occurred at state 1404. If so, the system terminates at end state 1408. Otherwise, the system proceeds to state 1410.

At state 1410, the tile manager 192 locks the uncompressed tile data, and then moves to state 1412, to determine whether an error occurred at state 1410. If an error occurred, the system terminates at end state 1408. Otherwise, the system moves to state 1414, wherein the tile manager 192 compresses the image data into a common buffer. For binary images of text and line drawings, the tile manager 192 uses a CCITT group 4 encoding.

From state 1414, the tile manager 192 moves to state 1416 to determine whether an error occurred at state 1414. If an error indeed occurred, the system moves to state 1418 to unlock the uncompressed tiles, and then exits at end state 1408. Otherwise, the system proceeds to state 1420, wherein the tile manager 192 allocates and locks cache memory space for the compressed tile data.

From state 1420, the system proceeds to state 1422 to determine whether an error occurred at state 1420. If an error occurred, the system moves to state 1418 and proceeds as described above. Otherwise, the system moves to state 1424, wherein the tile manager 192 copies the compressed data from the common buffer into the newly allocated cache memory state. The system moves from state 1424 to state 1426, wherein the tile manager 192 unlocks the compressed and uncompressed tile data and then terminates at end state 1408.

While the above detailed description has shown, described and pointed out the fundamental novel features of the invention as applied to various embodiments, it will be understood that various omissions and substitutions and changes in the form and details of the device illustrated may be made by those skilled in the art, without departing from the spirit of the invention.

What is claimed is:

1. An image memory management system, comprising:
 - a computer having a processor and an image memory, the image memory comprising a main memory and a secondary memory;
 - an image stack, located in the image memory, comprising a plurality of similar digital images, each digital image having a plurality of pixels grouped into at least one tile, and each digital image having a resolution different from the other digital images;
 - means for accessing a selected one of the tiles in the image stack;
 - first means for transferring a selected one of the tiles from the secondary memory to the main memory when the tile is accessed by the accessing means and the tile is absent from the main memory; and
 - second means for transferring a selected one of the tiles from the main memory to the secondary memory when the main memory is full.
2. The system defined in claim 1, additionally comprising means for modifying a selected one of the tiles.

3. The system defined in claim 2, wherein the second transferring means only transfer tiles that have been modified by the modifying means.

4. The system defined in claim 1, wherein the main memory is semiconductor memory.

5. The system defined in claim 1, wherein the secondary memory is a magnetic disk.

6. The system defined in claim 1, wherein each tile is square.

7. The system defined in claim 1, wherein a lowest resolution digital image comprises one tile.

8. The system defined in claim 1, wherein a preselected digital image in the image stack is resampled to obtain another digital image in the image stack.

9. The system defined in claim 1, wherein at least one of the digital images is compressed.

10. The system defined in claim 1, wherein the accessing means is responsive to an image access operation selected by a user.

11. The system defined in claim 10, wherein the image access operation is zooming or panning the image.

12. The system defined in claim 10, wherein the image access operation is reversible.

13. A method of managing images in a computer having a processor and an image memory comprising a slower access memory and a faster access memory, comprising the steps of:

- creating a digital image;
- resampling the digital image so as to form an image stack comprising the digital image and one or more lower resolution digital images;
- dividing each image into equal sized, rectangular tiles; and
- evaluating a location in the image memory of tiles in each digital image of the image stack in a given region of interest.

14. The method defined in claim 13, additionally comprising updating modified regions of all images when an edit operation is completed.

15. The method defined in claim 13, wherein the evaluating step includes the following order of decreasing availability:

- exists in the faster access memory in uncompressed form;
- exists in the slower access memory in uncompressed form;
- exists in the faster access memory in compressed form;
- exists in the slower access memory in compressed form; and
- must be constructed from higher resolution tiles.

16. The method defined in claim 13, wherein the evaluating step includes the following order of decreasing availability:

- exists in the faster access memory in uncompressed form;
- exists in the slower access memory in uncompressed form;
- exists in the slower access memory in compressed form; and
- must be constructed from higher resolution tiles.

17. The method defined in claim 13, wherein the evaluating step includes selecting the digital image with the lowest resolution higher than a requested resolution at a given view scale.

* * * * *

United States Patent [19]

[11] Patent Number: 4,972,319

Delorme

[45] Date of Patent: Nov. 20, 1990

[54] ELECTRONIC GLOBAL MAP GENERATING SYSTEM

[76] Inventor: David M. Delorme, 356 Range Rd., Cumberland, Me. 04021

[21] Appl. No.: 101,315

[22] Filed: Sep. 25, 1987

[51] Int. Cl.⁵ G09B 29/00

[52] U.S. Cl. 364/419; 434/150; 340/990

[58] Field of Search 364/419, 449; 434/150, 434/130; 340/990

[56] References Cited

U.S. PATENT DOCUMENTS

400,642	4/1889	Beaumont	283/34
751,226	10/1899	Van Der Grinten	283/34
752,957	2/1904	Colas	283/34
1,050,596	1/1913	Bacon	283/34
1,610,413	12/1924	Balch	283/34
2,094,543	9/1937	Lackey et al.	353/11
2,354,785	8/1944	Von Rohl	434/150
2,431,847	12/1947	Dusen	353/11
2,650,517	9/1953	Falk	355/77
3,248,806	5/1966	Schrader	434/150
3,724,079	4/1973	Jasperson et al.	33/15 B
4,315,747	2/1982	McBryde	434/150
4,673,197	6/1987	Stipelman et al.	434/150
4,689,747	8/1987	Krouse et al.	364/449
4,737,927	4/1988	Hanabusa et al.	340/990

OTHER PUBLICATIONS

"Equal-Area Projections for World Statistical Maps",

McBryde and Thomas, U.S. Dept. of Commerce, Coast and Geodetic Survey, Spec. Pub. 245, 1949.

"The Quadtree and Related Hierarchical Data Structures", Hanan Samet, Computer Surveys, vol. 16, No. 2, Jun. 1984.

Primary Examiner—Jerry Smith
 Assistant Examiner—Kim T. Bui
 Attorney, Agent, or Firm—Sughrue, Mion, Zinn, Macpeak & Seas

[57] ABSTRACT

A global mapping system which organizes mapping data into a hierarchy of successive magnitudes or levels for presentation of the mapping data with variable resolution, starting from a first or highest magnitude with lowest resolution and progressing to a last or lowest magnitude with highest resolution. The idea of this hierarchical structure can be likened to a pyramid with fewer stones or "tiles" at the top, and where each successive descending horizontal level or magnitude contains four times as many "tiles" as the level or magnitude directly above it. The top or first level of the pyramid contains 4 tiles, the second level contains 16 tiles, the third contains 64 tiles and so on, such that the base of a 16 magnitude or level pyramid would contain 4 to the 16th power or 4,294,967,296 tiles. This total includes "hyperspace" which is later clipped or ignored. Digital data corresponding to each of the separate data base tiles is stored in the database under a unique filename.

33 Claims, 9 Drawing Sheets

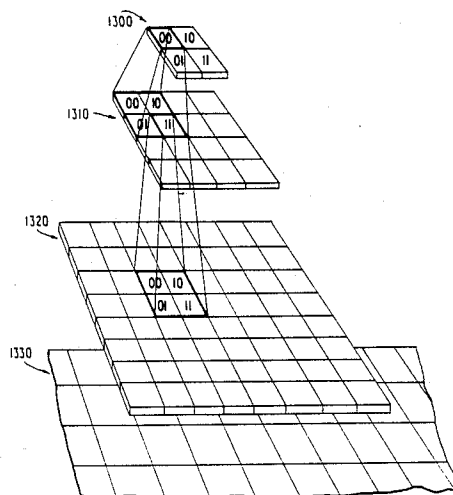


FIG. 1

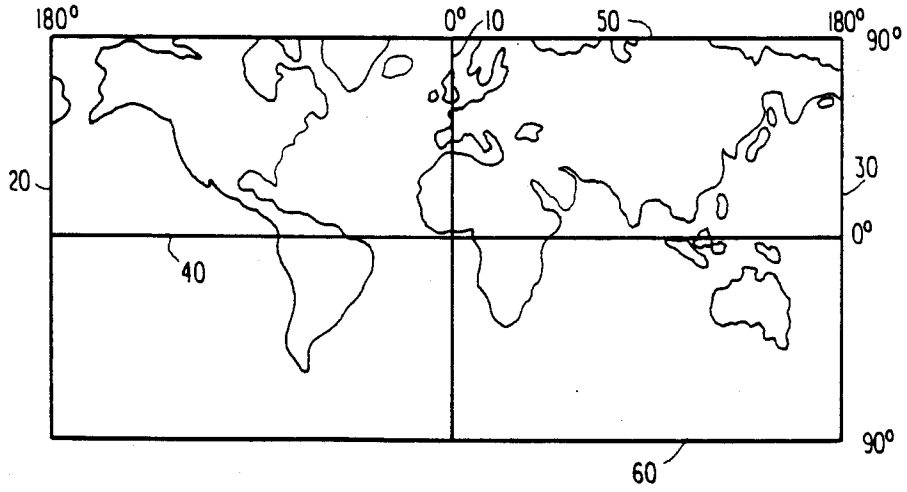


FIG. 2

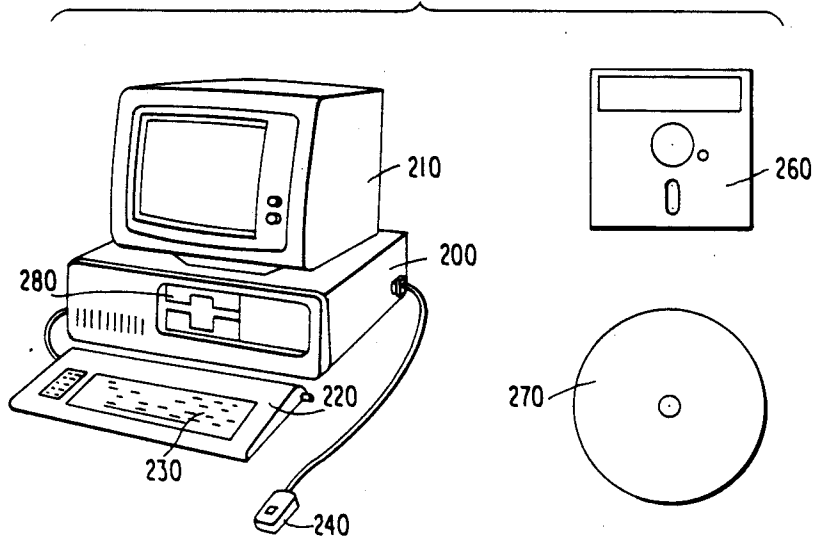


FIG. 3A

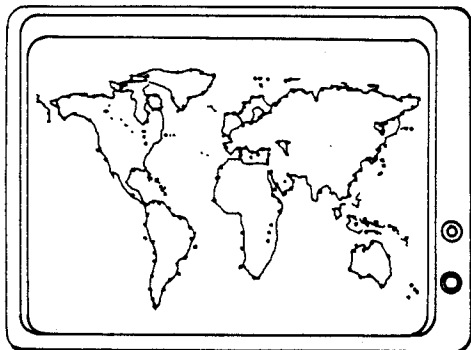


FIG. 3B

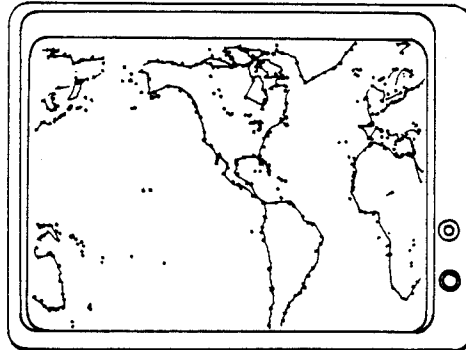


FIG. 3C

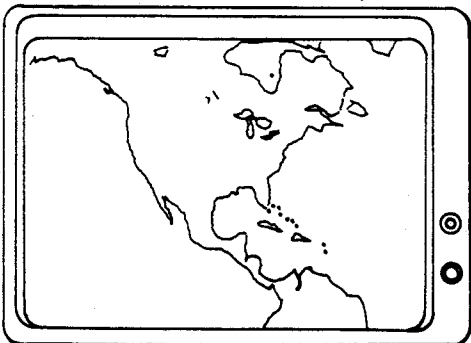


FIG. 3D

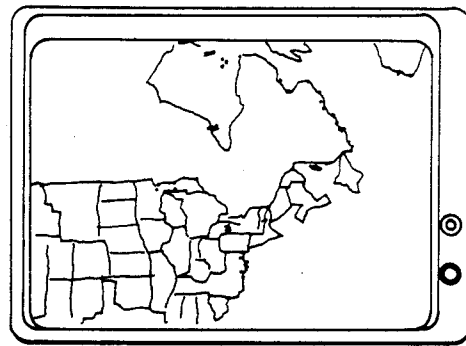


FIG. 3E

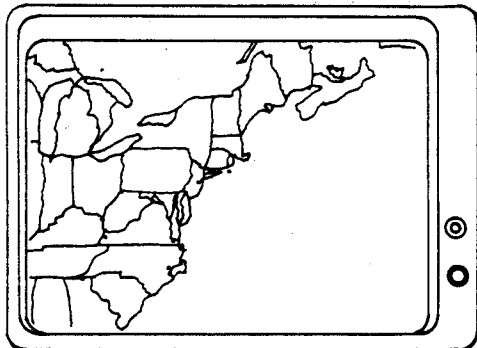


FIG. 3F

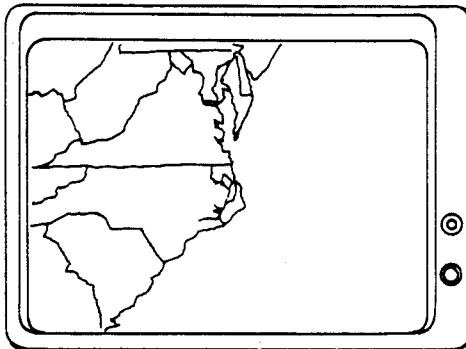


FIG. 4

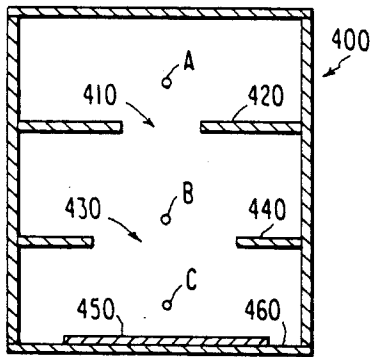


FIG. 5A

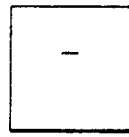


FIG. 5B

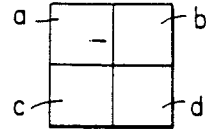


FIG. 6

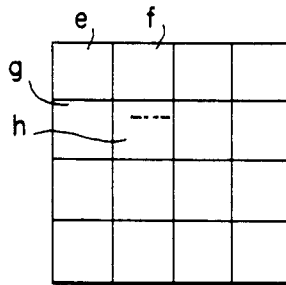


FIG. 7

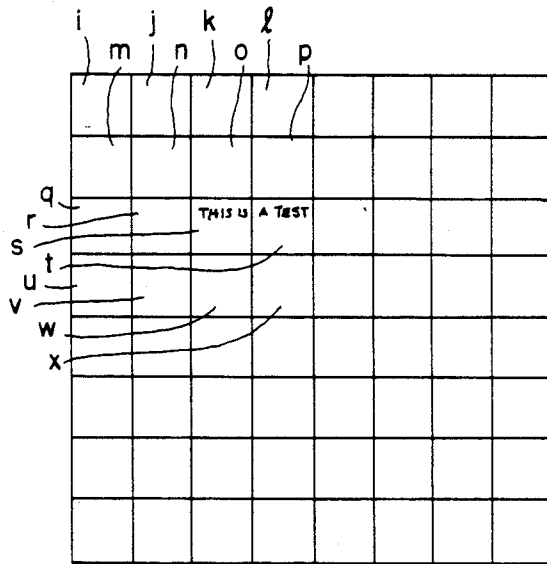


FIG. 8

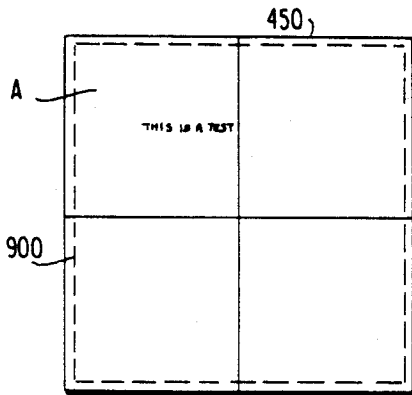
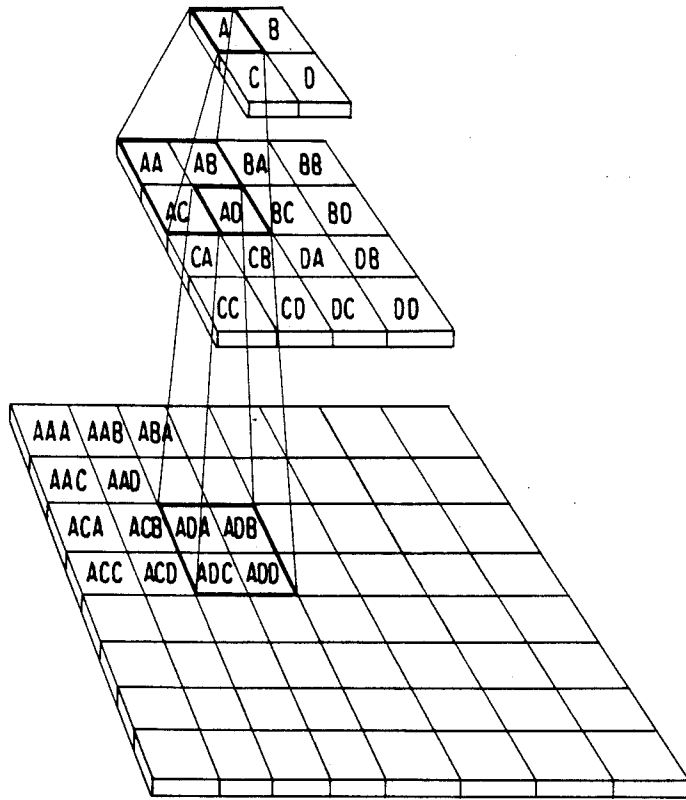


FIG. 9A

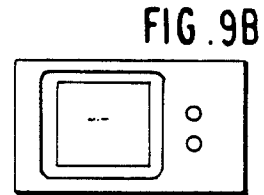


FIG. 9B

FIG. 10B

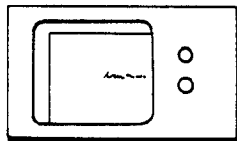
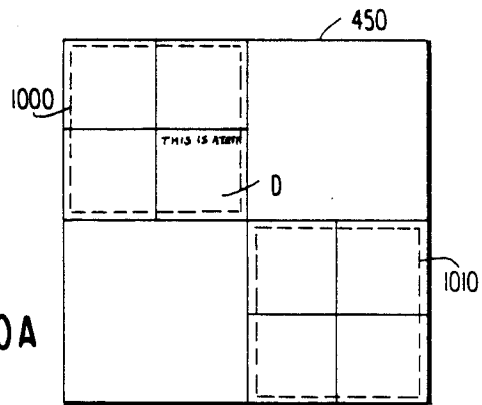


FIG. 10A



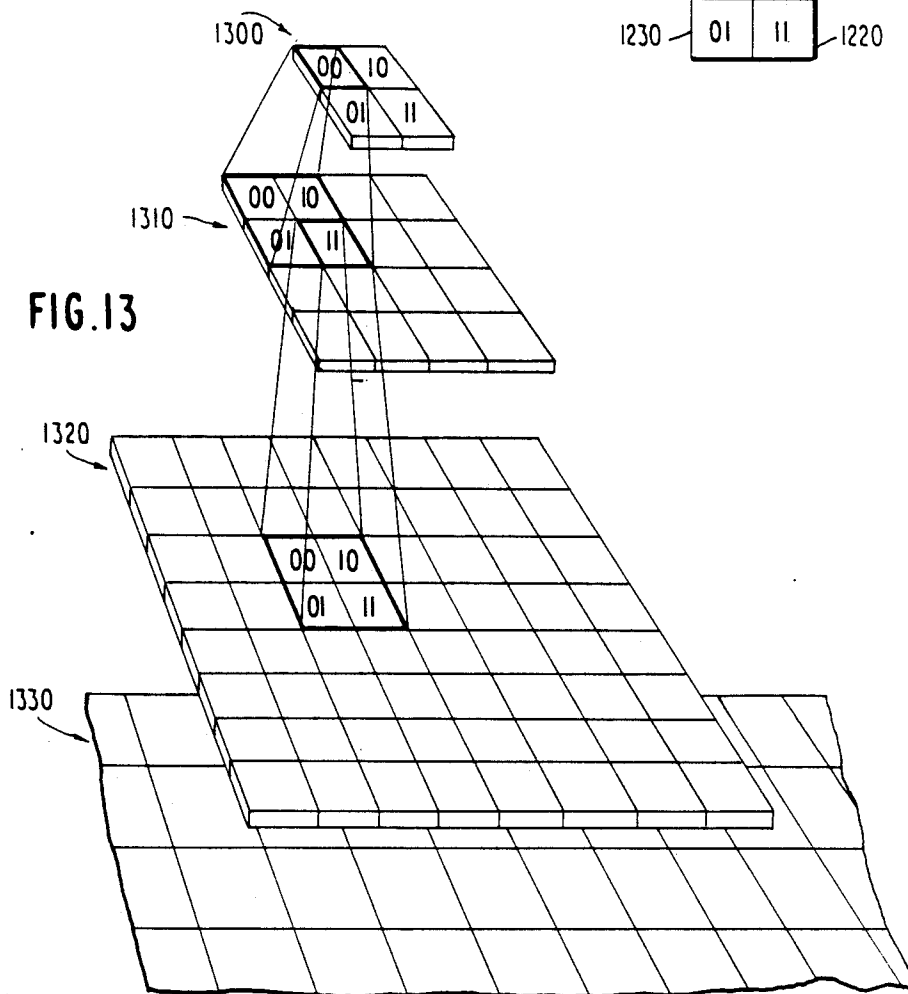
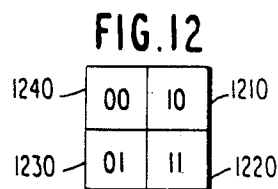
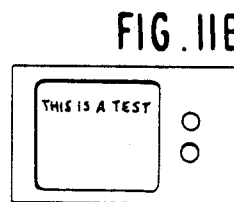
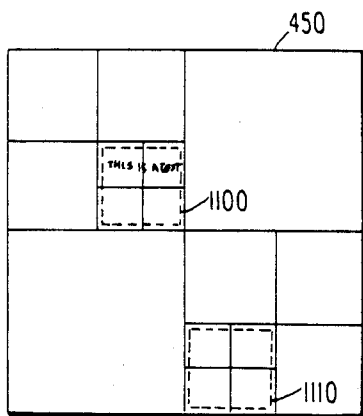


FIG. 14

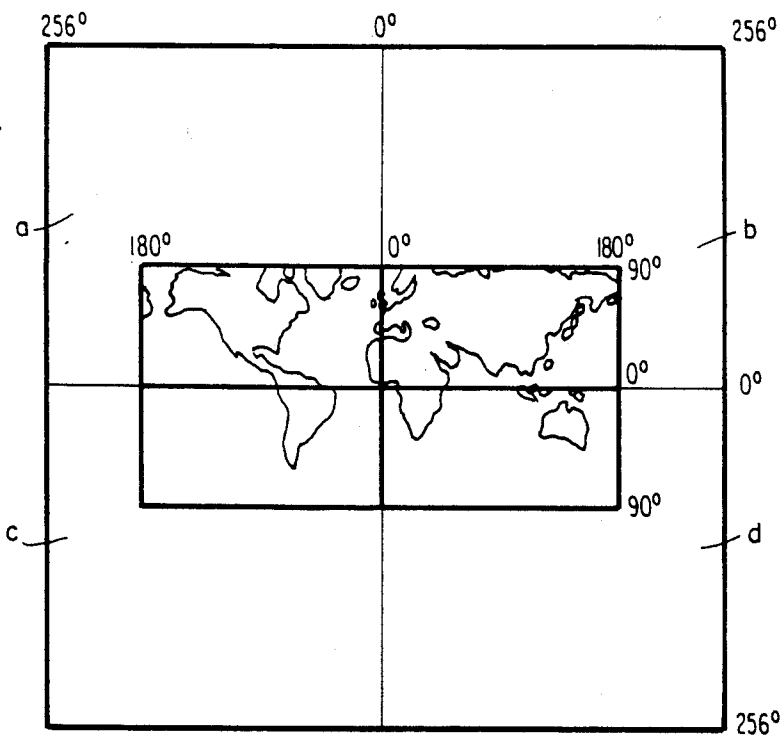


FIG. 15

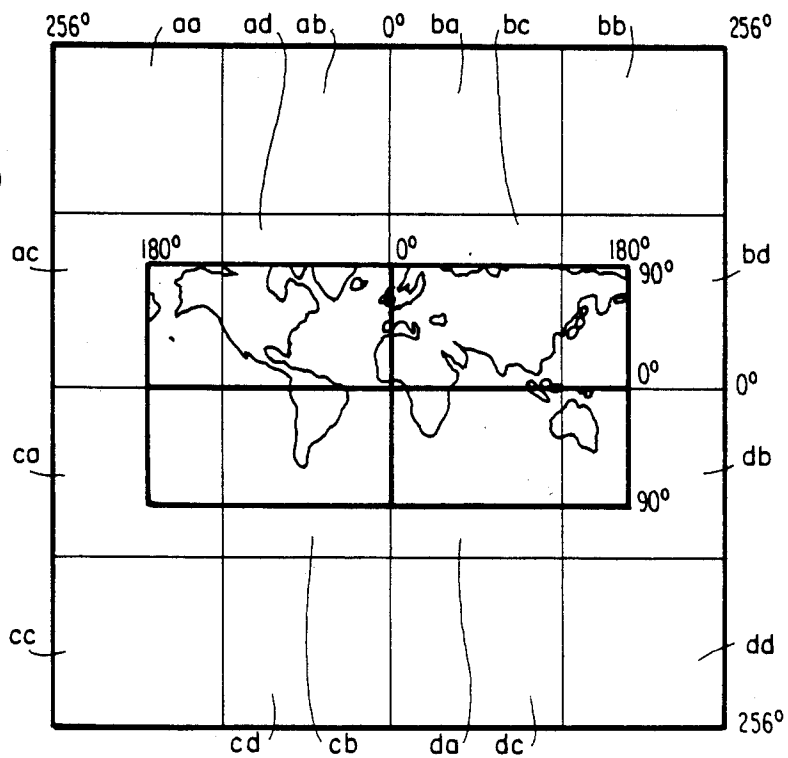


FIG. 16

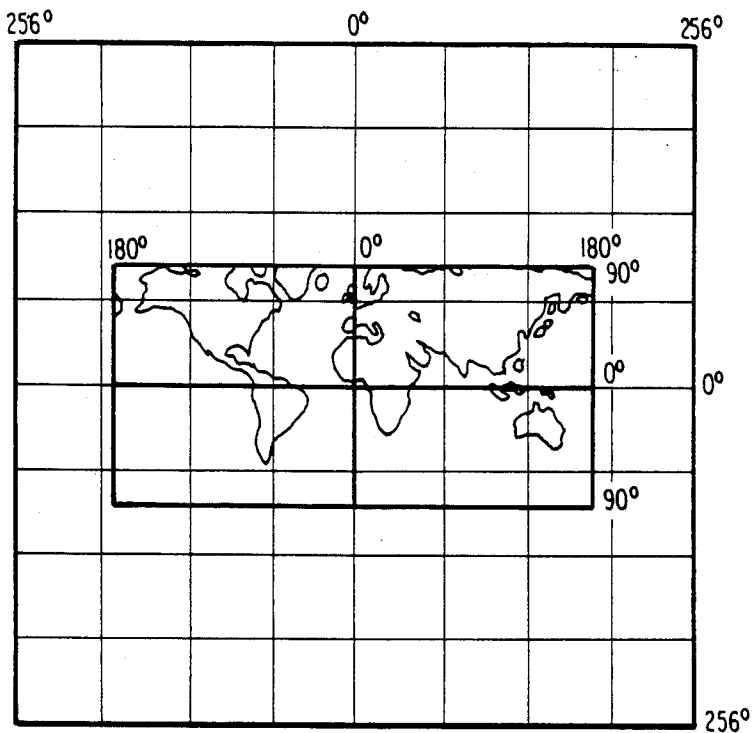


FIG. 17

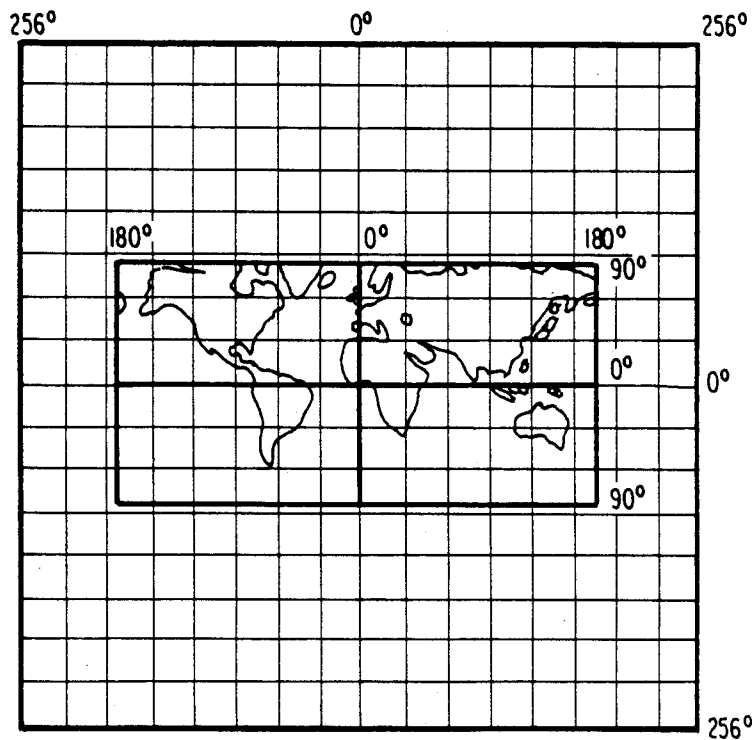


FIG. 18

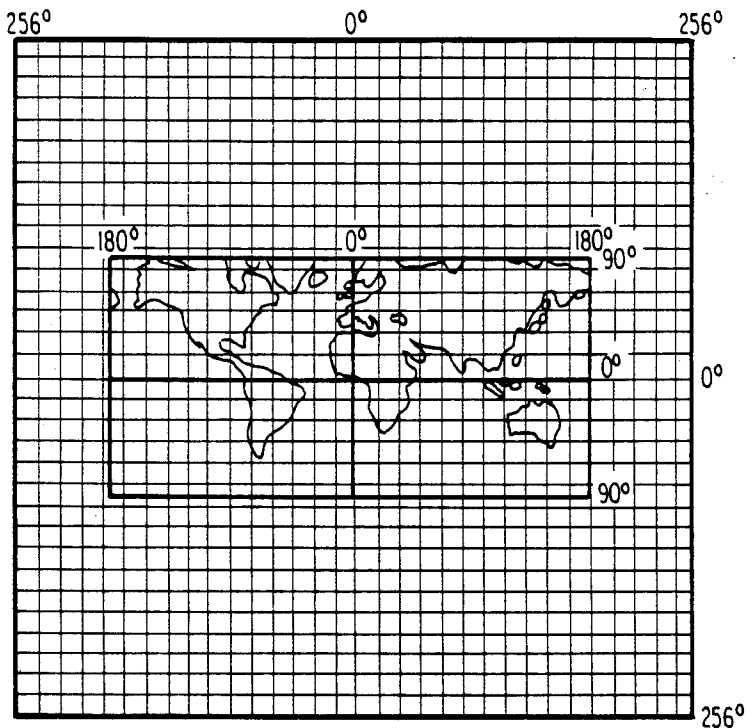


FIG. 19

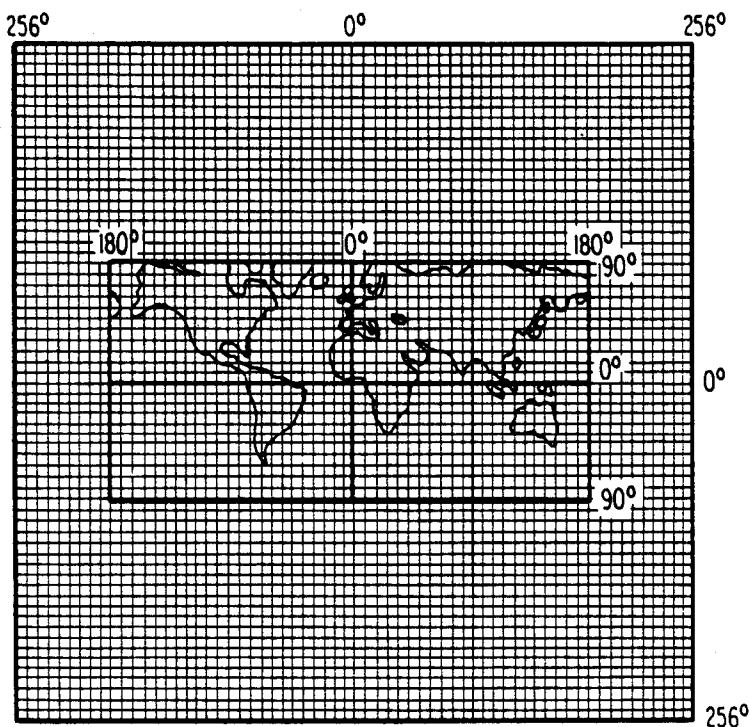
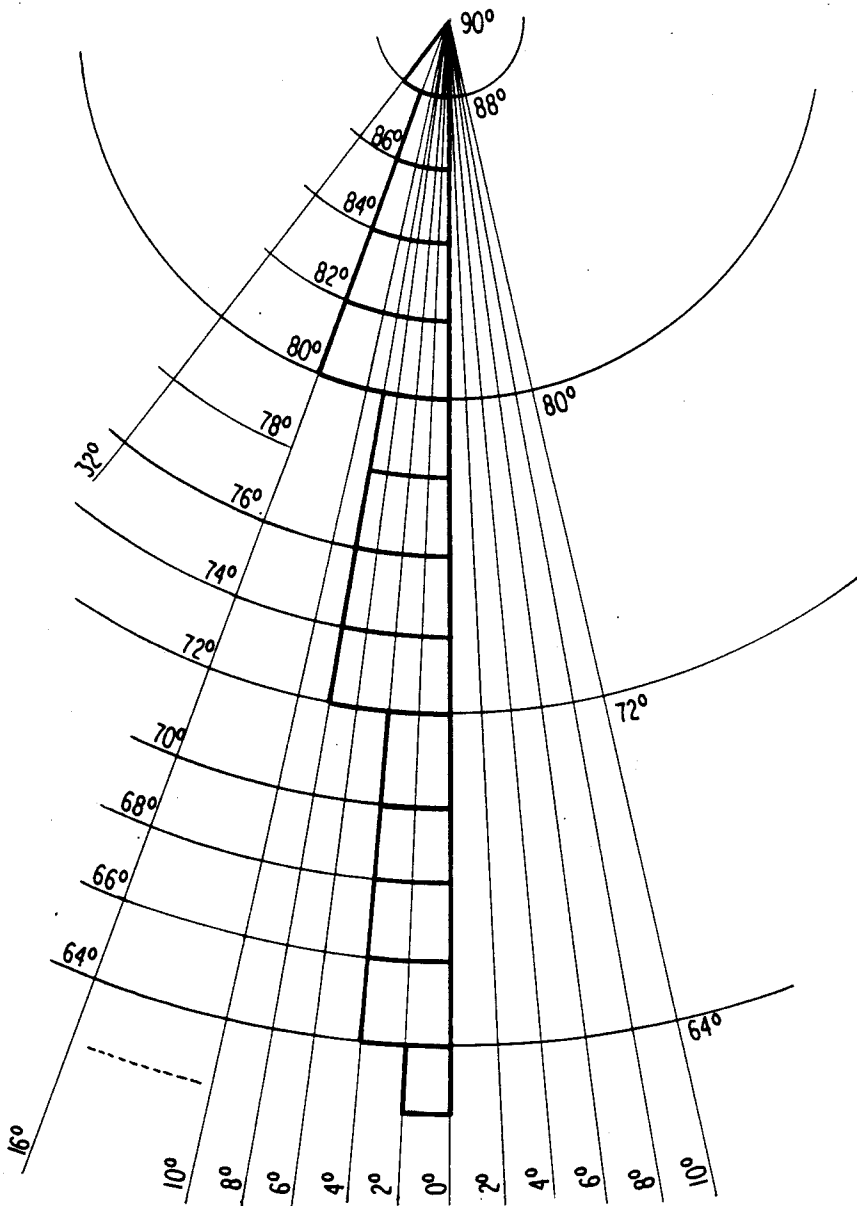


FIG. 20
ILLUSTRATION OF POLAR COMPRESSION
AT THE 8th MAGNITUDE



ELECTRONIC GLOBAL MAP GENERATING SYSTEM

BACKGROUND OF THE INVENTION

1. Technical Field

This invention relates to a new variable resolution global map generating system for structuring digital mapping data in a new data base structure, managing and controlling the digital mapping data according to new mapping data access strategies, and displaying the mapping data in a new map projection of the earth.

2. Background Art

Numerous approaches have been forwarded to provide improved geographical maps, for example:

U.S. Pat. No. 4,315,747, issued to McBryde on Feb. 16, 1982, describes a new map "projection" and intersecting array of coordinate lines known as the "graticule", which is a composite of two previously known forms of projection. In particular, the equatorial portions of the world are represented by a fusiform equal area projection in which the meridian curves, if extended, would meet at points at the respective poles, referred to as "pointed poles". In contrast, the polar regions of the world map are represented by a flat polar equal area projection in which the poles are depicted as straight horizontal lines with the meridians intersecting along its length. Thus, in a flat polar projection the meridian curves converge toward the poles but do not meet at a point and, instead, intersect a horizontal linear pole. The two component portions of the flat world map are joined where the parallels are of equal length. The composite is said to be "homolinear" because all of the meridian curves are similar curves, for example, sine, cosine or tangent curves, which merge where the two forms of projection are joined where the respective parallels are equal. The flat polar projections in the polar portions of the map provide a compromise with the Mercator cylinder projections, thereby greatly reducing distortion.

U.S. Pat. No. 1,050,596, issued to Bacon on Jan. 14, 1913, describes another composite projection for world maps and charts which uses a Mercator or cylindrical projection for the central latitudes of the earth and a convergent projection at the respective poles. In the central latitudes, the grids of the Mercator projection net or graticule are rectangular. In the polar regions, the converging meridians may be either straight or curved.

U.S. Pat. No 1,620,413, issued to Balch on Dec. 14, 1926, discusses gnomonic projections from a conformal sphere to a tangent plane and Mercator or cylindrical projections from the conformal sphere to a tangent cylinder. Balch is concerned with taking into account the non-spherical shape of the earth, and therefore, devises the so-called "conformal sphere" which represents the coordinates from the earth whose shape is actually that of a spheroid or ellipsoid of revolution, without material distortion.

U.S. Pat. No. 752,957, issued to Colas on Feb. 23, 1904, describes a map projection in which a map of the entire world is plotted or transcribed on an oval constructed from two adjacent side by side circles with arcs joining the two circles. The meridians are smooth curves equally spaced at the equator, while the latitude lines are non-parallel curves.

U.S. Pat. No. 400,642 issued to Beaumont on Apr. 2, 1889, describes a map of the earth on two intersecting

spheres, on which the coordinate lines of latitude and longitude are all arcs of circles.

U.S. Pat. No. 751,226, issued to Grinten on Feb. 2, 1904, represents the whole world upon the plane surface of a single circle with twice the diameter of the corresponding globe, the circle being delineated by a graticule of coordinates of latitude and longitude which are also arcs of circles.

U.S. Pat. No. 3,248,806, issued to Schrader on May 3, 1966, discloses a subdivision of the earth into a system of pivotally mounted flat maps, each map segment representing only a portion of the earth's surface in spherical projection on an equilateral spherical triangle to minimize distortion.

U.S. Pat. No. 2,094,543, issued to Lackey et al on Sept. 28, 1937, describes a projector for optically producing a variety of different map projections, including orthographic, stereographic and globular projections onto flat translucent screens and a variety of other projections on shaped screens.

U.S. Pat. No. 2,650,517, issued to Falk on Sept. 1, 1953, describes a photographic method for making geographical maps.

U.S. Pat. No. 2,354,785, issued to Rohl on Aug. 1, 1944, discloses two circular maps which are mounted side by side, and an arrangement for rotating the two maps in unison so that corresponding portions of the earth's surface are at all times in proper relationship.

U.S. Pat. No. 3,724,079, issued to Jaspersen et al on Apr. 3, 1973, discloses a navigational chart display device which is adapted to display a portion of a map and enable a pilot to fix his position, to plot courses and to measure distances.

U.S. Pat. No. 2,431,847 issued to Van Dusen on Dec. 2, 1947, discloses a projection arrangement, in which a portion of the surface of a spherical or curved map may be projected in exact scale and in exact proportional relationship.

McBryde and Thomas, *Equal Area Projections for World Statistical Maps*, Special Publication No. 245, Coast & Geodetic Survey 1949.

In addition to the above further teachings as to geographical mapping can be found in the *Elements of Cartography*, 4th edition which was written by Arthur Robinson, Randall Sale and Joel Morrison, and published by John Wiley & Sons (1978).

The present invention seeks to provide a low cost and efficient mapping system which allows the quick and easy manipulation of and access to an extraordinary amount of mapping information, i.e., a mapping system which allows a user to quickly and easily access a detailed map of any geographical area of the world.

Map information can be stored using at least three different approaches, i.e., paper, analog storage and digital storage, each approach having its own advantages and disadvantages as detailed below.

The paper mapping approach has been around since papyrus and will probably exist for the next thousand years.

Advantages of paper storage:
inexpensive.

once printed, no further processing is required to access the map information, so not subject to processing breakdown.

Disadvantages of paper storage:

can become bulky and unwieldy when dealing with a large geographical area, or a large amount of maps.

paper does not have the processing capabilities or "intelligence" of computers, and therefore does not support automated search or data processing capabilities.

cannot be updated cheaply and easily.

The analog mapping approach is used to provide what is commonly known as videodisc maps. The information is stored as still frames under N.T.S.C. (National Television Standards Committee) conventions. To make maps, a television camera moves across a paper map lying on a workbench. Every few inches a frame is recorded on videotape. After one row of the map is completely recorded, the camera is moved down to the next row of frames to be recorded. This process is repeated until frames representing a checkerboard pattern of the entire map are recorded. The recorded videotape could be used to view the map: however, access time to scan to different areas of the recorded map is usually excessive. As a result, a videodisc, with its quicker access time, is typically used as the medium for analog map storage. The recorded videotape is sent to a production house which "stamps" out 8 inch or 12 inch diameter, videodiscs.

Advantages of the analog storage approach:

one side of a 12 inch videodisc can hold 54,000 "frames" of a paper map. A frame is typically equal to $2\frac{1}{2} \times 3$ inches of the paper map.

access time to any frame can be fast usually under 5 seconds.

once located on the videodisc, the recorded analog map information will be used to control the raster scan of a monitor and to produce a reproduction of the map in 1/30th of a second.

through additional hardware and software, mapping symbols, text and/or patterns can be overlaid on top of the recorded frame.

Disadvantages of the analog storage approach:

the "frames" are photographed from paper maps, which, as mentioned above, cannot be updated cheaply or easily.

due to paper map projections, mechanical camera movements, lens distortions and analog recording electronics, the videodisc image which is reproduced is not as accurate as the original paper map.

as a result of the immediately above phenomena, latitude and longitude information which is extracted from the reproduced image cannot be fully trusted.

if a major error is made in recording any one of the 54,000 frames, it usually requires redoing and re-stamping.

since frames cannot be scrolled, most implementations employ a 50% overlap technique. This allows the viewer to jump around the database with a degree of visual continuity: however, this is at a sacrifice of storage capacity. If the frame originally covered $2\frac{1}{2} \times 3$ inches or approximately 8 square inches of the paper map, the redundant overlap information is 6 square inches, leaving only 2 square inches of new information in the centroid of each frame.

as a result of the immediately above deficiency, a 2×3 foot map containing 864 square inches would require 432 frames; thus, only 125 paper maps could be stored on one side of a 12 inch videodisc.

must take hundreds of video screen dumps to make a hard copy of a map area of interest and, even then, the screens do not immediately splice together because of the overlap areas.

the biggest disadvantage is that, since frames have to be arranged in a checkerboard fashion, there is no way to jump in directions other than north, south, east or west and maintain visual continuity. As an example, the visual discontinuity in viewing a "great circle" route from Alaska to New York would be unbearable for all but the most hearty.

The digital mapping approach has been around for at least 20 years and is much more frequently used than the analog approach. Digital data bases are stored in computers in a format similar to text of other databases. Unlike map information on a videodisc, the outstanding map features are stored as a list of objects to be drawn, each object being defined by a plurality of vector "dot" coordinates which define the crude outline of the object. As one example, a road is drawn by connecting a series of dots which were chosen to define the path (i.e., the "outline") of the road. Once drawn, further data and processing can be used to smooth the crude outline of the object, place text, such as the name or description of the object in a manner similar to what happens when drawing on a paper map.

Advantages of the digital approach:

digital maps are the purest form of geographical mapping data: from them, paper and analog maps can be produced.

digital maps can be quickly and easily updated in near real-time, and this updating can be in response to data input from external sources (e.g., geographical monitoring devices such as satellite photography).

digital maps can be easily modified to effect desirable mapping treatments such as uncluttering, enhancing, coloring, etc.

digital maps can be easily and accurately scaled, rotated and drawn at any perspective view point.

digital maps can be caused to reproduce maps in 3-D.

digital maps can drive pen-plotters (for easy paper reproductions), robots, etc.

digital maps can be stored on any mass storage device.

Disadvantages of the digital approach:

digital maps require the use or creation of a digital database: this is a very time-consuming and expensive process, but once it is made, the data base can be very easily copied and used for many different projects.

The digital approach is utilized with the present invention, as this approach provides overwhelming advantages over the above-described paper and analog approaches.

In designing any mapping system, several features are highly desirable:

First, it is highly desirable that the mapping system be of low cost.

Second, and probably most important, is access time. Not only is it generally desirable that the desired map section be accessible and displayed within a reasonable amount of time, but in some instances, this access time is critical.

In addition to the above, the present invention (as mentioned above), seeks to provide a third important feature,—a mapping system which allows the manipulation of and access to an extraordinary amount of mapping information, i.e., a mapping system which allows a user to quickly and easily access a detailed map of any geographical area of the world.

A tremendous barrier is encountered in any attempt to provide this third feature. In utilizing the digital approach to map a large geographical area in detail

(e.g., the earth), one should be able to appreciate that the storage of mapping data sufficient to accurately define all the geographical features would represent a tremendous data base.

While there have been digital mapping implementations which have successfully been able to manipulate a tremendous data base, these implementations involve tremendous cost (i.e., for the operation and maintenance of massive mainframe computer and data storage facilities). Furthermore, there is much room for improvement in terms of access time as these mainframe implementations result in access times which are only as quick as 20 seconds. Thus, there still exists a need for a low-cost digital mapping system which can allow the storage, manipulation and quick (i.e., "real time") access and visual display of a desired map section from a tremendous mapping data base.

There are several additional mapping system features which are attractive.

It is highly desirable that a mapping system be sensitive to and compensate for distortions caused by mapping curved geographical (i.e., earth) surfaces onto a flat, two-dimensional representation. While prior art approaches have provided numerous methods with varying degrees of success, there is a need for further improvements which are particularly applicable to the digital mapping system of the present invention.

It is additionally attractive for a mapping system to easily allow a user to change his/her "relative viewing position", and that in changing this relative position, the change in the map display should reflect a feeling of continuity. Note that the "relative viewing position should be able to be changed in a number of different ways. First, the mapping system should allow a user to selectively cause the map display to scroll or "fly" along the geographical map to view a different (i.e., "lateral") position of the geographical map while maintaining the same degree of resolution as the starting position. Second, the mapping system should allow a user to selectively vary the size of the geographical area being displayed (i.e., "zoom") while still maintaining an appropriate degree of resolution, i.e., allow a user to selectively zoom to a higher "relative viewing position" to view a larger geographical area with lower resolution regarding geographical, political and cultural characteristics, or zoom to a lower "relative viewing position" to view a smaller geographical area with higher resolution. (Note that maintaining the appropriate amount of resolution is important to avoid map displays which are effectively barren or are cluttered with geographical, political and cultural features.) Again, while prior art approaches have provided numerous methods with varying degrees of success, there is a need for further improvements which are particularly applicable to the digital mapping system of the present invention.

The final feature concerns compatibility with existing mapping formats. As mentioned above, the creation of a digital database is a very tedious, time-consuming and expensive process. Tremendous bodies of mapping data are available from many important mapping authorities, for example, the U.S. Geological Survey (USGS), Defense Mapping Agency (DMA), National Aeronautics and Space Administration (NASA), etc. In terms of both being able to easily utilize the mapping data produced by these agencies, and represent an attractive mapping system to these mapping agencies, it would be highly desirable for a mapping system to be compatible with all of the mapping formats used by these respective

agencies. Prior art mapping systems have been deficient in this regard; hence, there still exists a need for such a mapping system.

SUMMARY OF THE INVENTION

The present invention provides a digital mapping method and system of a unique implementation to satisfy the aforementioned needs.

The present invention provides a computer implemented method and system for manipulating and accessing digital mapping data in a tremendous data base, and for the reproduction and display of electronic display maps which are representative of the geographical, political and cultural features of a selected geographical area. The system includes a digital computer, a mass storage device (optical or magnetic), a graphics monitor, a graphics controller, a pointing device, such as a mouse, and a unique approach for structuring, managing, controlling and displaying the digital map data.

The global map generating system organizes the mapping data into a hierarchy of successive magnitudes or levels for presentation of the mapping data with variable resolution, starting from a first or highest magnitude with lowest resolution and progressing to a last or lowest magnitude with highest resolution. The idea of this hierarchical structure can be likened to a pyramid with fewer stones or "tiles" at the top, and where each successive descending horizontal level or magnitude contains four times as many "tiles" as the level or magnitude directly above it. The top or first level of the pyramid contains 4 tiles, the second level contains 16 tiles, the third contains 64 tiles and so on, such that the base of a 16 magnitude or level pyramid would contain 4 to the 16th power or 4,294,967,296 tiles. This total includes "hyperspace" which is later clipped or ignored. Hyperspace is that excess imaginary space left over from mapping of 360 deg, space to a zero magnitude virtual or imaginary space of 512 deg, square.

A first object of the present invention is to provide a digital mapping method and system which are of low cost.

A second and more important object of the present invention is to provide a unique digital mapping method and system which allow access to a display of the geographical, political and cultural features of a selected geographical area within a minimum amount of time.

A third object of the present invention is to provide a digital mapping method and system which allow the manipulation of and access to an extraordinary amount of mapping information, i.e., a mapping method and system which allow a user to quickly and easily access a detailed map of any geographical area of the world.

Another object of the present invention is to provide a digital mapping method and system which recognize and compensate for distortion introduced by the representation of curved (i.e., earth) surfaces onto a flat two-dimensional display.

Still a further object of the present invention is to provide a digital mapping method and system which allow a user to selectively change his/her "relative viewing position", i.e., to cause the display monitor to scroll or "fly" to display a different "lateral" mapping position of the same resolution, and to cause the display monitor to "zoom" to a higher or lower position to display a greater or smaller geographical area, with an appropriate degree of resolution.

A fifth object of the present invention is to provide a digital mapping method and system utilizing a unique

mapping graticule system which allows mapping data to be compatibly adopted from several widely utilized mapping graticule systems.

BRIEF DESCRIPTION OF THE DRAWINGS

The foregoing and other objects, structures and features of the present invention will become more apparent from the following detailed description of the preferred mode for carrying out the invention; in the description to follow, reference will be made to the accompanying drawings in which:

FIG. 1 is an illustration corresponding to a flat projection of the earth's surface.

FIG. 2 is an illustration of a digital computer and mass storage devices which can be utilized in implementing the present invention.

FIGS. 3A-3F are illustrations of monitor displays showing the ability of the present invention to display varying sizes of geographical areas at varying degrees of resolution.

FIG. 4 is a cross-sectional diagram of a simple building example explaining the operation of the present invention.

FIG. 5A and B are plan view representations of a paper 450 as it is viewed from the relative viewing position A shown in FIG. 4.

FIG. 6 is a plan view representation of a paper 450 as it is viewed from the relative viewing position B shown in FIG. 4.

FIG. 7 is a plan view representation of a paper 450 as it is viewed from the relative viewing position C shown in FIG. 4.

FIG. 8 is a pyramidal hierarchy of the data base file structure showing an example of the ancestry which exists between files.

FIG. 9A is a plan view representation of a paper 450, with the paper being divided into a first level of quadrant areas.

FIG. 9B is an illustration of a monitor displaying a digital map of the area enclosed by the dashed portions in FIG. 9A.

FIG. 10A is a plan view representation of a paper 450, with the upper-left and lower-right paper quadrant areas being further divided into quadrants.

FIG. 10B is an illustration of a monitor displaying a digital map of the area enclosed by the upper-left dashed portion in FIG. 10A.

FIG. 11A is a plan view representation of a paper 450, with several sections of the second level of quadrants being further divided into additional quadrants.

FIG. 11B is a higher resolution display of the area enclosed within the dashed portion in FIG. 11A.

FIG. 12 is a plan view illustration of a quadrant area division, with a two-bit naming protocol being assigned to each of the quadrant areas.

FIG. 13 is a pyramidal hierarchy of the data base files using the two-bit naming protocol of FIG. 12, and showing an example of the ancestry which exists between files.

FIG. 14 is a plan view illustration of a $360^\circ \times 180^\circ$ flat projection of the earth being impressed in the $512^\circ \times 512^\circ$ mapping area of the present invention, with a first quadrant division dividing the mapping area into four equal $250^\circ \times 256^\circ$ mapping areas.

FIG. 15 is the same plan view illustration of FIG. 14, with a second quadrant division dividing the mapping area into 16 equal $126^\circ \times 128^\circ$ mapping areas.

FIG. 16 is the same plan view illustration of FIG. 15, with a third quadrant division dividing the mapping area into 64 equal $64^\circ \times 64^\circ$ mapping areas.

FIG. 17 is the same plan view illustration of FIG. 16, with a fourth quadrant division dividing the mapping area into 256 equal $32^\circ \times 32^\circ$ mapping areas.

FIG. 18 is the same plan view illustration of FIG. 17, with a fifth quadrant division dividing the mapping area into 1024 equal $16^\circ \times 16^\circ$ mapping areas.

FIG. 19 is the same plan view illustration of FIG. 18, with a sixth quadrant division dividing the mapping area into 4096 equal $8^\circ \times 8^\circ$ mapping areas.

FIG. 20 is an illustration showing the application of polar compression at the 8th level or magnitude of resolution.

DETAILED DESCRIPTION OF THE PREFERRED EMBODIMENTS OF THE INVENTION

Before turning to the detailed description of the preferred embodiments of the invention, it should be noted that the map illustrations used throughout the drawings are only crude approximations which are only being used to illustrate important features and aspects and the operation of the present invention; therefore, the geographical political and cultural outlines may very well differ from actual outlines.

FIG. 1 is a crude representation of what the earth's surface would look like if it were laid flat and viewed from a "relative viewing position" which is a great distance in space. Shown as vertical lines are: 10, corresponding to the 0° meridian extending through Greenwich, England; 20, corresponding to the 180° west meridian; and, 30, corresponding to the 180° east meridian. Shown as horizontal lines are: 40, corresponding to the equator; 50, corresponding to 90° north (i.e., the north pole); and 60, corresponding to 90° south (i.e., the south pole).

Note that at this "relative viewing position", not much detail as to cultural features is seen; i.e., all that is seen is the general outline of the main geographical masses of the continents.

The present invention seeks to provide a low cost and efficient computer-based mapping method and system having a unique approach for arranging and accessing a digital mapping database of unlimited size, i.e., a mapping method and system which can manipulate and access a data base having sufficient data to allow the mapping system to reproduce digital maps of any geographical area with different degrees of resolution. This can be most easily understood by viewing FIG. 2 and FIGS. 3A-F.

Because of the overwhelming advantages over the paper and analog mapping approaches, the digital mapping approach is utilized with the present invention; thus, there is shown in FIG. 2, a digital computer 200, having a disk or hard drive 280, a monitor 210, a keyboard 220 (having a cursor control portion 230), and a mouse device 240. As mentioned previously, in a digital mapping approach, mapping information is stored in a format similar to the text of other databases, i.e., the outstanding map features are stored as a list of objects to be drawn, each object being defined by a plurality of vector "dot" coordinates which define the crude outline of the object. (Note: the reproduction of a digital map from a list of objects and "dot" vectors is well known the art, and is not the subject matter of the present invention; instead, the invention relates to a unique

method and system for storing and accessing the list of objects and "dot" vectors contained in a tremendous digital data base.)

Once a geographical map has been "digitized",—i.e., converted to a list of objects to be drawn and a plurality of vector "dot" coordinates which define the crude outline of the object —, the mapping database must be stored in the memory of a mass storage device. Thus, the digital computer 200, which is to be used with the mapping method and system of the present invention, is shown associated with the magnetic disk 260 (which represents any well-known magnetic mass storage medium, e.g., floppy disks, hard disks, magnetic tape, etc.), and the CD-ROM 270 (which represents any well-known optical storage medium, e.g. a laser-read compact disk). Alternatively, the digital mapping database can be stored on, and the digital computer can be associated with any well known electronic mass storage memory medium (e.g., ROM, RAM, etc.). Because of every increasing availability, reductions in cost, and tremendous storage capacities, the preferred memory mass storage medium is the CD-ROM, i.e., a laser-read compact disk.

The discussion now turns to FIGS. 3A–F, showing illustrations of monitor displays which provide a brief illustration of the operation of the present invention. Although the digital nature of the maps of FIGS. 3A–3F can easily be detected due to the jagged outlines, it should be understood that these geographical outlines could easily be smoothed using any of a number of "smoothing" techniques which are well-known to those skilled in the digital mapping art.

In FIG. 3A, the digital computer has retrieved relevant mapping information from the digital mapping database, and has produced a monitor display of a digital map substantially corresponding to the flat projection of the earth's surface which was shown in FIG. 1. In FIG. 3A, the monitor display reflects a "relative viewing position" which is a great distance in space, and hence, only the crude geographical outline of the continents is shown with sparse detail.

Suppose a user wishes to view a map of the states of Virginia and Maryland in greater detail. By entering the appropriate commands using the keyboard 220 or the mouse device 240, a user can cause the monitor display to "zoom" to a lower "relative viewing position", such that the monitor displays a digital map of a smaller geographical area which is shown at a higher degree of resolution. Thus, in FIG. 3B the a digital map of the continents of the western hemisphere is displayed in greater detail.

By entering additional commands, a user can cause the monitor display to further "zoom" to the following displays: FIG. 3C showing North America in greater detail; FIG. 3D showing the eastern half of the United States in greater detail; FIG. 3E showing the east coast of the United States in greater detail; and, FIG. 3F showing Virginia and Maryland in greater detail.

Although in this example, the monitor display was caused to "zoom" to Virginia and Maryland, it should, be appreciated that the present invention allowed a user to selectively zoom into any geographical area of the earth, and once a user has reached the desired degree of mapping resolution, the mapping system of the present invention also allows the user to "scroll" or "fly" to a different lateral position on the map.

Furthermore, although the drawings illustrate the monitor display zooming to display state boundaries,

and features, it should be further appreciated that the present invention is by no means limited to this degree of resolution. In fact, the degree of resolution capable with the present invention will be shown to be limited only by the operating system of the digital computer 200 with which the present invention is used. In one demonstration, the monitor display has been shown to be able to zoom to resolution where the outlines of streets were displayed. Even further degrees of resolution are possible as will be more fully understood after the discussions below.

In digitally mapping a large geographical area (e.g., the earth) in detail, —especially in the degree of resolution mentioned above —, one should be able to appreciate that the storage of digital mapping data sufficient to accurately define all the geographical, political and cultural features would represent a tremendous digital mapping database. In order to provide a low cost mapping system having quick access time and allowing a high degree of resolution, what is needed is a mapping system having an effective approach for arranging an accessing the digital database. Prior art mapping systems have been deficient in this regard.

The mapping system of the present invention utilizes a new and extremely effective approach, which can be most easily understood using the following simplified example.

In FIG. 4, there is shown the cross-section of a building 400, with a square hole 410 (shown in cross-section) cut through the third level floor 420, with a larger square hole 430 (shown in cross-section) cut in the second level floor 440, and with a large square piece of paper 450 (shown in cross-section) laid out on the first level floor 460. Suppose it was desired to build up a digital data base which could be used to reproduce a digital map of the paper 450 with varying degrees of resolution.

First, one would take the "relative viewing position" A, and view the paper 450 through the square hole 410 in the third level floor 420. At this level, the paper 450 appears small (FIG. 5A), and the degree of resolution is such that the message appears only as a series of dots. In order to build up a digital mapping database, the visual perception (FIG. 5A) is imagined to be divided into four equal quadrants a, b, c, d (FIG. 5B), and visual features appearing in each respective area is digitized and stored in a separate database file. Thus, four separate database files can be utilized to reproduce a digital map of the paper 450 as viewed from position A (FIG. 4).

In order to digitize and record data corresponding to a second (or higher) degree of resolution, the next "relative viewing position" B (FIG. 4) is taken to view the paper 450 through the square hole 430. At this level, the paper 450 appears larger (FIG. 6), and the degree of resolution is such that the message now appears as a series of lines. At this second level, the map is imagined as being divided into four times as many areas as the first imaginary division, and then, the visual information contained within each area is digitized and stored in a separate database file. Thus, 16 files can be used to reproduce a digital map of the paper 450, as viewed from the relative viewing position B (FIG. 4).

In order to digitize and record data corresponding to a third (or higher) degree of resolution, the next "relative viewing position" C (FIG. 4) is taken to view the paper 450. At this level, paper 450 now appears larger (FIG. 7) and has visual features of higher resolution.

The paper 450 is imagined as being divided into four times as many areas as the second imaginary division, and the visual information is digitized and stored. Thus, 64 files could be used to reproduce a digital map of the paper 450, as viewed from the relative viewing position C (FIG. 4).

Once digital data has been entered for the above three "relative viewing positions" A, B, C (FIG. 4), the digital mapping database contains $4+16+64$ or 84 files which can be conceptually envisioned as being arranged in a pyramid structure as shown in FIG. 8. In order to allow a user to selectively display any desired map section at the desired degree of resolution, the digital computer 200 must be able to know which of the 84 files to access such that the appropriate mapping data can be obtained. The present invention accomplishes this by conceptually arranging the files in a pyramidal structure, and assigning a file name to each file which is related both to the file's position and ancestry within the pyramidal structure. This can be more specifically described as follows:

A file's ancestry can be explained using the illustrations of FIGS. 5B, 6 and 7. In FIG. 5B, the paper 450, as viewed from "relative viewing position" A (FIG. 4), is subjected to an imaginary division into four quadrants a, b, c, and d. Quadrants a, b, c, d are related to one another in the sense that it takes all four areas to represent the paper 450; hence quadrants a, b, c, d can be termed as brothers and sisters.

FIG. 6 is an illustration of the paper 450 as it appears from the relative viewing position B (FIG. 4). with the paper 450 being subjected to an imaginary division into 16 areas. Note that the areas e, f, g, h (FIG. 6) represent the same area of paper 450 as the quadrant a (FIG. 5B). In effect, quadrant a has been enlarged (to show a higher degree of resolution) and divided into quadrants e, f, g, h. Thus, it can be said that quadrant a (FIG. 5B) is the parent, and that quadrants e, f, g, h (FIG. 6) are brothers and sisters and the offspring of ancestor a. Similar discussions can be made for quadrants b, c and d and the remaining area of FIG. 6.

FIG. 7 is an illustration of the paper 450 as it appears from the relative viewing position C (FIG. 4). with the paper 450 being subjected to an imaginary division into 64 areas. In a manner similar to the discussion above, note that areas s, t, w, x (FIG. 7) represent the same area of paper 450 as the quadrant h (FIG. 6). In effect, quadrant h has been enlarged (to show a higher degree of resolution) and divided into quadrants s, t, w, x. Thus, it can be said that quadrant a (FIG. 5B) is the grandparent, quadrant h (FIG. 6) is the parent, and quadrants s, t, w, x (FIG. 7) are the brothers and sisters and offspring of ancestors a and h.

As described previously, once FIGS. 5B, 6 and 7 are subjected to the imaginary divisions, the visual information in each area (or quadrant) is digitized and stored in a separate file. The 84 resulting files can be conceptually envisioned as the pyramidal structure shown in FIG. 8. In FIG. 8, dashed lines are utilized to show the lineage of the files just discussed.

FIG. 8 is further exemplary of one file naming operation which can be utilized with the present invention.

At the top of the pyramidal structure (FIG. 8), each of the four quadrant files is arbitrarily assigned a different character. A, B, C, D, (Note: The characters assigned are not critical with regard to the invention and hence it should be noted that any characters can be assigned, e.g., 0,1,2,3, etc.)

In moving down one level in the pyramidal structure, the filenames for each of the respective files on the second level is increased to two characters.

In calculating the filenames, it is convenient to first divide the second level files into groups of four, according to parentage. To maintain a record of ancestry, the ancestor filename of each file is maintained as the first part of the filename. In determining the second part, the naming protocol which was utilized to name the quadrant files of the top level, is also utilized in naming the respective quadrant files on the second level. Thus, parent file A is shown as being related to descendent (i.e., brother and sister) files AA, AB, AC, AD. Similar discussion can be made for the remaining files along these two level.

A similar process can be utilized in providing the unique filenames to the third level files. At this level, the filenames consist of three characters. Again, the ancestor filename of each file would be maintained as a first filename part, in order to maintain a record of ancestry. In the example illustrated (FIG. 8), parent file AD is shown as being related to descendent (i.e., brother and sister) files ADA, ADB, ADC, ADD. Similar discussions can be made for the remaining files along these two levels, and furthermore, similar discussions can be made each time a pyramidal level is added.

From the above discussion, one should be able to realize that the above-described naming convention is particularly useful in programming a digital computer to move through the pyramidal file structure to access the appropriate data corresponding to varying degrees of resolution. More particularly, one should be able to realize that, since file names increase one character in length each time there is a downward movement through the pyramidal structure and the protocol for naming descendent files is known, the digital computer can be programmed to quickly and easily access the appropriate files for a smaller mapping area with a greater degree of resolution. Similarly, one should be able to realize that, since the filenames decrease one character in length each time there is an upward movement through the pyramidal structure, the digital computer can be programmed to quickly and easily access the appropriate files for a greater mapping area with a smaller degree of resolution.

The following example is believed to provide an increase in the understanding of the present invention.

In the example, it is assumed that the digital database corresponding to the three resolutions of the paper 450 (as shown in FIGS. 4, 5A-B, 6, 7) have been loaded to be accessible from the memory mass storage device, and furthermore, it is assumed that the mapping system is programmed to initially access and display a digital map corresponding to the digital mapping data in the files A, B, C, D (FIG. 8). Thus, the monitor (FIG. 9B) would display (in low resolution) the entire area enclosed within dashed portion 900 illustrated on the paper 450 (FIG. 9A). (Note: The reproduction of a digital map from digital data from several different files or sources is well-known in the art and is not the subject matter of the present invention.)

Suppose the user notices the dotted area on the low resolution map and wishes to investigate this area further. By using the appropriate keys (e.g., \backslash , $/$, \wedge , \searrow) and/or a mouse device, a user can give the mapping system an indication that he/she wishes to see the smaller area (i.e., quadrant A) at a higher degree of resolution. Upon receiving this preference, the mapping

system can use its knowledge of the file naming operations to quickly determine the names of the files which must be accessed. More specifically, using A as the parent file name and following the existing quadrant naming protocol the mapping system is quickly and easily able to calculate that it is files AA, AB, AC, AD which it needs to access. Once these files are accessed, the monitor in FIG. 10B displays (in higher resolution) the area enclosed within the dashed portion 1000 as illustrated on the paper 450 (FIG. 10A).

If a user is still not satisfied with the degree of mapping resolution, the user can again use the appropriate keys or mouse device to indicate that he/she wishes to see the smaller area (e.g., quadrant D; FIG. 10A) in a higher degree of resolution. In using AD as the parent filename and following the existing quadrant naming protocol, the mapping system is quickly and easily able to calculate that it is files ADA, ADB, ADC, ADD which it needs to access. Once these files are accessed, the monitor (FIG. 11B) displays (in higher resolution), the area enclosed within the dashed portion 1100 as illustrated on the paper 450 (FIG. 11A).

One skilled in the digital mapping and computer programming art should recognize that "scrolling" or "flying" to different lateral "relative viewing positions" to display a different lateral portion of the map is also provided by the present invention. Instead of adding or removing filename characters as in a change of resolution, in this instance, the mapping system must be programmed to keep track of the filenames of the current position and also, the orderly arrangement of filenames so that the appropriate filenames corresponding to the desired lateral position can be determined. As an example if the user desired to scroll to the right border of the paper 450, the mapping system would respond by accessing and causing the monitor to display the digital maps corresponding to the following sequence of files: (Note: In this example, it is assumed that it takes 4 files to provide sufficient digital data to display a full digital map on a monitor) ADA, ADB, ADC, ADD; ADB, ADD, BCA, BCC; BCA, BCB, BCC, BCD; BCB, BCD, BDA, BDC; and BDA, BDB, BDC, BDD. If the user, then desired to scroll to the bottom (right corner) of the paper 450, the mapping system would respond by accessing and causing the monitor to display the digital maps corresponding to the following files: BDA, BDB, BDC, BDD; BDC, BDD, DBA, DBB; DBA, DBB, DBC, DBD; DBC, DBD, DDH, DDB; DDA, DDB, DDC, DDD. In effect as all of the files in the above example correspond to the same level of resolution all these files (and any group of files which exist on the same level of resolution) can be taken as being related as cousins.

FIGS. 9A, 10A, 11A can also be used to illustrate the operation of moving toward the display of a larger mapping area with a lower degree of resolution.

Assume that after lateral "scrolling" or "flying", that the monitor is now displaying (not shown) a digital map corresponding to the enclosed area 1110 shown in FIG. 11A. (Note: at this position the mapping system is accessing and display a digital map corresponding to the digital data in the files DCA, DCB, DCC, DCD). Suppose the user now wishes to cause the "relative viewing position" to zoom upward, such that the monitor will display a larger portion of the paper 450 at a lower degree of resolution. By using the appropriate keys or a mouse device, the user indicates his/her preference to the mapping system. Upon receiving this preference,

the mapping system is programmed to quickly determine the names of the files which must be accessed. More specifically, the mapping system is able to look at the first portion of the filenames currently being used (i.e., DCA, DCB, DCC, DCD), to immediately determine that these files have the ancestry DC, i.e., have a grandfather D and a parent DC. The mapping system then immediately determines brother and sister files of parent file DC as being DA, DB and DD. The mapping system then accesses these files and causes the monitor to display a digital map (not shown) corresponding to the enclosed portion 1010 (FIG. 10A) of the paper 450.

Suppose the user again indicate a preference to cause the "relative viewing position" zoom upward. Upon receiving this preference, the mapping system again goes through a process similar to that discussed immediately above. However, this time the mapping system looks at the filenames currently being used (i.e., DA, DB, DC, DD) and determines that parent file D has brother and sister files A, B and C. The mapping system then immediately accesses these files and causes the monitor to display a digital map (FIG. 9B) corresponding to the enclosed portion 900 (FIG. 9A) of the paper 450.

The text now turns to a description of the operation for assigning unique filenames in the currently preferred embodiment, i.e., in a digital mapping system which is implemented in a DOS operating system.

As anyone skilled in the computer art will know, every computer operating system has its own unique set of rules which must be followed. In an implementation of the present invention in a DOS operating system, the DOS rules must be followed. Since a critical feature of the present invention is the division of the digital mapping database into a plurality of files (each having a unique filename), of particular concern with the present invention is the DOS rules regarding the naming of filenames.

A DOS filename may be up to eight (8) characters long, and furthermore, may contain three (3) additional trailing characters which can represent a file specification. Thus, a valid DOS filename can be represented by the following form:

where "-" can be replaced by any ASCII character (including blanks), except for the following ASCII characters:

./\ [] : | < > + , ;

and ASCII characters below 20H. The currently preferred embodiment stays within these DOS filename rules by using the file naming operations which are detailed below.

Because the assigned filenames will be seen to be related to hexadecimals, a useful chart containing the hexadecimal base and also a conversion list (which will be shown to be convenient ahead), is reproduced below:

Column 1	Column 2	Column 3
0000	0	G
0001	1	H
0010	2	I
0011	3	J
0100	4	K
0101	5	L
0110	6	M

-continued

Column 1	Column 2	Column 3
0111	7	N
1000	8	O
1001	9	P
1010	A	Q
1011	B	R
1100	C	S
1101	D	T
1110	E	U
1111	F	V

The first column contains a list of all the possible 4-bit binary combinations: the second column contains the hexadecimal equivalent of these binary numbers: and the third column concerns a "mutant-hex" conversions which will be shown to be important in the discussion to follow. In the operations to assign unique filenames for use in a DOS operating system, the present invention looks at each of the eight DOS filename characters as hexadecimal characters rather than ASCII characters. Hence, while the following discussion will center around determining unique filenames using hexadecimal (and "mutant-hexadecimal") characters, it should be understood in an actual DOS implementation, the hexadecimal filenames must be further converted into the equivalent ASCII characters such that the appropriate DOS file naming rules are followed.

At this point, it is also useful to note that the file naming operation of the preferred embodiment is not concerned with the trailing three character filename extension. However, it should be further noted that this three character filename extension may prove useful in specifying data from different sources, and allowing the different types of data to reside in the same database. As examples, the filename extension ".spm" might specify data from scanned paper maps, the filename extension ".si" might specify data from satellite imagery, the filename extension ".ged" might specify gridded elevation data, etc.

As a result of the foregoing and following discussions, it will be seen that the naming operation of the preferred embodiment is concerned only with a filename of the following form:

where each "-" represents a character which is a hexadecimal character within the character set of "0-9" and "A-F", or is a "mutant-hexadecimal" character within the character set of "G-V".

Several more important file naming details should be discussed.

First, it should be pointed out that the first four (4) filename characters is designated as corresponding to the "x" coordinate characters, and the last four (4) filename characters are designated as corresponding to the "y" coordinate characters.

Second, during the file naming operations, often it is necessary to convert the filename characters into the equivalent binary representation. As each hexadecimal character can be converted into a four bit binary number, it can be seen that the first four (4) filename characters (designated as "x" coordinate characters) can be converted into sixteen (16) binary bits designated as "x" bits, and similarly, that the last four (4) filename characters (designated as "y" coordinate characters) can be converted into sixteen (16) binary bits designated as "y" bits. As will become more apparent ahead, each of these

sixteen (16) "x" and "y" bits corresponds to a filename bit which can be manipulated when assigning filenames at a corresponding magnitude or level of mapping resolution, e.g., the first "x" and first "y" bits correspond to filename bits which can be manipulated when assigning unique filenames at the first magnitude, the second "x" and second "y" bits correspond to filename bits which can be manipulated when assigning unique filenames at the second magnitude, etc.

Third, FIG. 12 corresponds to the naming protocols which are utilized to modify and relate a parent filename to four (4) quadrant filenames. Note that there is a two-bit naming protocol in each of the quadrant files. As will become more clear ahead, the first bit of each protocol determines whether the current "x" filename bit will be modified (i.e., if the first protocol bit is a "1", the current "x" filename bit is changed to a "1", and if first protocol bit is a "0", the current "x" filename bit is maintained as a "0"), and the second bit determines whether the current "y" filename bit will be modified (in a similar manner).

The text now turns to a file naming example which is believed to provide further teachings and clarity to the currently preferred file naming operation.

FIG. 13 is an illustration of a portion of the preferred digital data base, with the plurality of files (partially shown) being arranged in a conceptual pyramidal manner in a manner similar to that which was described with reference to FIG. 8. More specifically, there are shown four files 1300 having digital data corresponding to a first level or magnitude of mapping resolution, sixteen files 1310 having digital data corresponding to a second level or magnitude of mapping resolution, sixty-four files 1320 having digital data corresponding to a third level or magnitude of mapping resolution, and a partial cut-away of a plurality of files 1330 having data corresponding to a fourth level or magnitude of mapping resolution. Although not shown, it is to be understood that, in the preferred embodiment, additional pyramidal structure corresponding to levels magnitudes five through sixteen similarly exist. As examples of the file naming operation, filenames will now be calculated for the files which essentially occupy the same positions as the files which were outlined in FIG. 8.

We begin with the initializing eight (8) character filename:

0 0 0 0 0 0 0 0

which can be converted to the binary equivalent:

0000 0000 0000 0000 0000 0000 0000 0000

This binary representation is the basic foundation which will be used to calculate all of the filenames for the files on the first level (1300). Note, that the first and last four filename characters, and the first and last sixteen bits are slightly separated in order to conveniently distinguish the "x" and "y" coordinate characters and bits. Both the first (leftmost) "x" bit and the first (leftmost) "y" bit are the bits which can be manipulated in assigning a unique filename to the files on the first level.

File naming begins with the first (upper-rightmost) file on the first level 1300. The naming protocol assigned to this quadrant file is the two-bit protocol "10".

As the first protocol bit is a "1", this means that the current "x" bit must be changed to a "1". As the second protocol bit is a "0", this means that the current "y" bit is maintained as a "0". As a result of the foregoing, the first (upper-rightmost) file is assigned the filename having the binary equivalent of:

1000 0000 0000 0000 0000 0000 0000 0000

which can be converted to the hex characters:

8 0 0 0 0 0 0 0

In proceeding clockwise, next is the second (lower-rightmost) file on the first level 1300. The naming protocol assigned to this quadrant file is the two-bit protocol "11". As the first protocol bit is a "1", the current "x" bit is changed to a "1": similarly, as the second protocol bit is a "1", the current "y" bit is changed to a "1". As a result of the foregoing, the second (lower-rightmost) file is assigned the filename having the binary equivalent of:

1000 0000 0000 0000 1000 0000 0000 0000

which can be converted to the hex characters:

8 0 0 0 8 0 0 0

Continuing clockwise, next is the third (lower-leftmost) file on the first level 1300. The naming protocol assigned to this quadrant file is the two-bit protocol "01". As the first protocol bit is a "0", the current "x" bit is maintained at 0. As the second protocol bit is a "1", the current "y" bit is changed to a "1". As a result of the foregoing, the third (lower-leftmost) file is assigned the filename having the binary equivalent of:

0000 0000 0000 0000 1000 0000 0000 0000

which can be converted to the hex characters:

0 0 0 0 8 0 0 0

Finally, there is the fourth (upper-leftmost) file on the first level 1300. The naming protocol assigned to this quadrant is the two-bit protocol "00". As neither of the protocol bits is a "1", it can be easily seen that neither of the current "x" and "y" bits changes, and hence, the fourth (upper-leftmost) file is assigned the filename having the binary equivalent of:

0000 0000 0000 0000 0000 0000 0000 0000

which can be converted to the hex characters:

0 0 0 0 0 0 0 0

In further discussions of the example, it is important to note that the initializing (8) character filename of 0000 0000 (which was utilized to calculate the filenames

of the files on the first level 1300) is not utilized in assigning filenames on subsequent levels. In naming files from the second level or magnitude downward, the binary equivalent of the parent file's name is utilized as the foundation from which the descendent file's name is derived. It is only coincidental that the filename of the parent file 00000000 (located in the user-left most corner of the first level 1300) is the same as the initializing filename. Use of the parent's filename to calculate the descendent's filename will become more readily apparent ahead in the example.

In continuing the file naming example, the fourth (upper-leftmost) file (having filename 00000000) in the first level 1300 can be viewed as being the parent file of the four (highlighted) quadrant files in the second level 1310. As stated above, the binary equivalent of parent file's 00000000 name is utilized as the foundation for calculating the descendent file's filenames. At this second level or magnitude, the second "x" and "y" bits from the left in the parent's binary filename are taken as the "current" bits which can be manipulated to provide a unique filename for the descendent files.

As the calculation of the filename for the fourth (upper-leftmost) file of the second level 1310 illustrates a very important modification in the file naming operation, the example will first continue with discussions corresponding to this file.

As the naming protocol assigned to the fourth (upper-leftmost) file of the second level 1310 is two-bit protocol "00", it can be seen that neither of the current "x" and "y" bit would be changed. Hence the parent's filename 00000000 is unchanged, and is attempted to be adopted as the descendent's filename. However, note that this is extremely undesirable as the operation of the present invention is based on assigning each data file a unique filename, and furthermore, a DOS operation system will not allow the same filename to be assigned to two different files. To avoid this clash, the preferred file naming operation of the present invention incorporates a further step which can be detailed as follows:

First calculate the filename as explained above. Once the binary filename is obtained, convert to the eight character hexadecimal equivalent.

Next, take the decimal number of the current level or magnitude and subtract one (1) to result in a decimal magnitude modifier. Convert the decimal magnitude modifier into a four-bit binary magnitude modifier, and line these four bits up with the four hexadecimal "x" filename characters. Whenever a "1" appears in the binary magnitude modifier, the corresponding aligned "x" filename character is converted to a "mutant-hexadecimal" character. i.e., a decimal 16 value is added to convert the aligned filename character into a one of the "mutant-hexadecimal" characters in the character set of "G-V".

Conversions from a hexadecimal character to a "mutant-hexadecimal" character can be most readily made using the chart detailed above. As an example, if decimal 16 is added to the hex character "0" (Column 2), there is a conversion to the "mutant-hexadecimal" character "G" (Column 3). Similarly, if decimal 16 is added to the hex character "1" (Column 2), there is a conversion to the "mutant-hexadecimal" character "H" (Column 3). Similar discussion can be made for the remaining hex and "mutant-hexadecimal" characters in the chart.

Once correspondingly aligned filename characters are converted to "mutant-hexadecimal", the resultant

19

eight (8) characters correspond to the file's unique filename.

The above processing will now be applied to the fourth (upper-rightmost) file of the second level 1310 (which was recently discussed above). The resultant binary filename:

0000 0000 0000 0000 0000 0000 0000 0000

is converted to the hex characters:

0 0 0 0 0 0 0 0.

The level or magnitude two (2) minus one (1) results in a decimal magnitude modifier of one (1). The decimal magnitude modifier is converted to the four-bit binary equivalent and is aligned with the "x" filename characters above, as follows:

0 0 0 1

Only the fourth bit of the binary magnitude modifier is a "1", so only the fourth "x" filename character needs to be converted to "mutant-hexadecimal". From the chart, the hexadecimal character "0" is shown to convert to a "mutant-hexadecimal" character "G". Thus, the unique filename which is assigned to the fourth (upper-leftmost) file of the second level 1310, is:

0 0 0 G 0 0 0 0.

In continuing the example to calculate the filename for the first (upper-right-quadrant) file of the second level 1310. it can be seen that this file is assigned the two-bit naming protocol "10". The first protocol bit is a "1" which indicates that the current (second from the left) "x" bit of the parent file's binary filename must be changed to a "1", In contrast, the second protocol bit is a "0", which indicates that the current (second from the left) "y" bit is maintained as "0" Thus the parent filename:

0000 0000 0000 0000 0000 0000 0000 0000

is converted to:

0100 0000 0000 0000 0000 0000 0000 0000

which results in the hex characters:

4 0 0 0 0 0 0 0.

The level or magnitude two (2) minus one (1) results in a decimal magnitude modifier of one (1). The decimal magnitude modifier is converted to the four-bit binary equivalent and is aligned with the "x" filename characters above, as follows:

0 0 0 1

20

Only the fourth bit of the binary magnitude modifier is a "1", so only the fourth "x" filename character needs to be converted to "mutant-hexadecimal". From the chart, the hexadecimal character "0" is shown to convert to a "mutant-hexadecimal" character "G". Thus, the unique filename which is assigned to the first (upper-right-quadrant) file of the second level 1310, is:

4 0 0 G 0 0 0 0.

Turning now to the second (lower-right-quadrant) file, this file is assigned the two-bit naming protocol "11". The first protocol bit is a "1" which indicates that the current (second from the left) "x" bit of the parent file's binary filename must be changed to a "1", and similarly, the second protocol bit is a "1", which indicates that the current (second from the left) "y" bit of the parent file's binary filename must be changed to a "1" Thus the parent filename:

0000 0000 0000 0000 0000 0000 0000 0000

is converted to:

0100 0000 0000 0000 0100 0000 0000 0000

which results in the hex characters:

4 0 0 0 4 0 0 0.

The level or magnitude two (2) minus one (1) results in a decimal magnitude modifier of one (1). The decimal magnitude modifier is converted to the four-bit binary equivalent and is aligned with the "x" filename characters above, as follows:

0 0 0 1

Only the fourth bit of the binary magnitude modifier is a "1", so only the fourth "x" filename character needs to be converted to "mutant-hexadecimal". From the chart, the hexadecimal character "0" is shown to convert to a "mutant-hexadecimal" character "G". Thus, the unique filename which is assigned to the second (lower-right quadrant) file of the second level 1310, is:

4 0 0 G 4 0 0 0.

In applying the above operations to the third (lower-left-quadrant) file of the second level 1310, it can be easily calculated that the resultant filename is:

0 0 0 G 4 0 0 0.

The example of the file naming operation is further extended to the third level or magnitude. as this example is illustrative of both the use of the parent file's binary filename to calculate the descendent's filename, and the removal of "mutant-hexadecimal" conversions before calculating the descendent's filename.

In FIG. 13. the third (lower-right-quadrant) file of the second level 1310 is shown as being the parent of the four (4) quadrant files highlighted in the third level or magnitude 1320.

The discussion centers on the calculation of the unique filename for the second (lower-right-quadrant) file in the third level 1320. Before the parent filename can be used as the foundation for calculating the descendent's filename. all "mutant-hexadecimal" conversions must be removed. Thus the parent filename:

4 0 0 G 4 0 0 0.

is converted back to:

4 0 0 0 4 0 0 0.

which is further converted to the binary equivalent:

0100 0000 0000 0000 0100 0000 0000 0000

In continuing the calculation, this second (lower-right-quadrant) file is assigned the two-bit naming protocol "11". The first protocol bit is a "1" which indicates that the current (third from the left) "x" bit of the parent file's binary filename must be changed to a "1", and similarly, the second protocol bit is a "1", which indicates that the current (third from the left) "y" bit of the parent file's binary filename must be changed to a "1". Thus the parent filename:

0100 0000 0000 0000 0100 0000 0000 0000

is converted to:

0110 0000 0000 0000 0110 0000 0000 0000

which results in the hex characters:

6 0 0 0 6 0 0 0.

The level or magnitude three (3) minus one (1) results in a decimal magnitude modifier of two (2). The decimal magnitude modifier is converted to the four-bit binary equivalent and is aligned with the "x" filename characters above, as follows:

0 0 1 0

Only the third bit of the binary magnitude modifier is a "1", so only the third "x" filename character needs to be converted to "mutant-hexadecimal". From the chart, the hexadecimal character "0" is shown to convert to a "mutant-hexadecimal" character "G". Thus, the unique filename which is assigned to the second (lower-right-quadrant) file of the third level 1320, is:

6 0 G 0 6 0 0 0.

The filenames for several additional third level files will be given to give the patent reader further practice.

In applying the above operations to the first (upper-right-quadrant) file of the third level 1320, it can be easily calculated that the resultant filename is:

6 0 G 0 4 0 0 0.

In applying the above operations to the third (lower-left-quadrant) file of the third level 1320, it can be easily calculated that the resultant filename is:

4 0 G 0 6 0 0 0.

Finally, in applying the above operations to the fourth (upper-left-quadrant) file of the third level 1320, it can be easily calculated that the resultant filename is:

4 0 G 0 4 0 0 0.

As a result of all of the foregoing teachings, one skilled in the art should now be able to calculate the filename of any other of the 1.4 billion files which would be required to provide digital maps corresponding to sixteen (16) resolutions of any geographical area on earth. Furthermore, once a file is being accessed, by understanding the rules and operations of the file naming operation one skilled in the art should be able to calculate any other related files, i.e., parent files. and brother/sister/cousin files.

While the unique approach for storing and accessing files in the pyramidal file structure has been particularly pointed out. further discussion is needed as to an additional advantageous feature of the present invention.

As mentioned previously, the creation of a digital database is a very tedious, time consuming and expensive process. Tremendous bodies of mapping data are available from many important mapping authorities, for example, the U.S. Geological Survey (USGS), Defense Mapping Agency (DMA), National Aeronautics and Space Administration (NASA), etc.

The maps and mapping information produced by the above recited agencies, is always based on well established mapping area divisions. As a few examples, the Defense Mapping Agency (DMA) produces maps and mapping information based on the following mapping areas: GNC maps which are 2°x2°; JNC maps which are 1°x1°; ONC maps which are 30'x30'; TPC maps which are 15'x15'; and JOG maps which are 7.5'x7.5'. As a further example, the U.S. Geological Survey (USGS) also produces maps and utilizes mapping information based on 15'x15' and 7.5'x7.5'.

In terms of both being able to easily utilize the mapping data produced by these agencies, and represent an attractive mapping system to these mapping agencies, it would be highly desirable for the mapping system of the present invention to be compatible with all of the mapping formats used by these respective agencies. Such is not the case when the mapping database is based on a graticule system corresponding to 360°

If one were to apply multiple quadrant divisions to the 360°x180° flat map projection of the earth (FIG. 1). one would result in the following mapping area subdivisions:

Level of quadrant div.:	Resultant mapping area:
1	(4) 180° × 90°
2	(16) 90° × 45°
3	(64) 45° × 22.5°
4	(256) 22.5° × 11.25°
5	(1024) 11.25° × 5.625°
etc.	

Note that these mapping area subdivisions are very awkward, and do not match any of the well settled mapping area subdivisions. (It should be further noted that no better results are obtained if the initial map projection is imagined as being a 360°×360° square instead of a rectangle.)

In order to avoid these awkward mapping subdivisions, and result in quadrant divisions which precisely match widely used mapping area subdivisions, the present invention utilizes a unique initial map projection.

More specifically, as can be seen in FIG. 14, the present invention initially begins with a unique 512°×512° initial map projection. Shown centered in the 512°×512° map projection is the now familiar 360°×180° flat projection of the surface of the earth. Although the 512°×512° projection initially appears awkward and a waste of map projection space, the great advantages which are resultant from the use of this projection will become more apparent in the discussions to follow.

To aid in this discussion, provided on the next page is a chart which details these important advantages as well as other useful information regarding the use of this map projection.

less complicated, the non-DOS file naming operation will be used in the discussion.

The digital mapping of the earth surfaces begins in FIG. 14. The visual perception of the earth surfaces is experienced as being centered, and occupying only a portion of the 512°×512° projection. A first quadrant division is applied to result in four equal 256°×256° mapping areas. The visual information in each of the areas is digitized, and stored in a separate file. Thus, it can be seen that one would have to access four files a, b, c, d in order to reproduce a digital map corresponding to the earth surfaces as viewed from this "relative viewing position."

One skilled in the art, might, at this point, wonder if the massive blank portions of the 512°×512° projections result in large blank portions on the digital map display. The preferred embodiment avoid this phenomena, through a simple watchdog operation, i.e., the computer is programmed to keep track of longitudinal and latitudinal movements from an initial position of 0° longitude and 0° latitude, and the computer does not allow scrolling of the monitor display beyond 90° north or south.

As to side to side movements, the computer allows scrolling beyond 180° east or west by patching the appropriate data files together to perform a "wrap around" operation. Note that, with the knowledge of the logical file naming operation, the computer can quickly and easily calculate the appropriate files to access.

Before moving to the next level or magnitude of mapping resolution, it is beneficial to note the correspondence between our findings and the entries in the

MAGNITUDE EQUIVALENCY CHART FOR DELORME PROJECTION
Chart assumes 69 statute miles per degree at equator

MAG-NITUDE	Window Size without overlap	Ht of window statute miles	Ht of window kilometers	# Windows per MAG	# Windows/MAG w/polar compression	Pixel resolution 480 monitor (ft)	Data resolution (ft) 1024-based window	Equivalent Paper Map Scales	Size of paper map image at equator (in)
1	256° × 256°	17664	28421	4	4		91080		
2	128° × 128°	8832	14211	8	8		45540		
3	64° × 64°	4416	7105	24	24	48576	22770	1:100 million	2.8 × 2.8
4	32° × 32°	2208	3553	72	72	24288	11385	1:50 million	2.8 × 2.8
5	16° × 16°	1104	1776	288	288	12144	5693	1:30 million	2.3 × 2.3
6	8° × 8°	552	888	1152	858	6072	2846	1:16 million	2.2 × 2.2
7	4° × 4°	276	444	4232	3432	3036	1423	1:10 million	1.7 × 1.7
8	2° × 2°	138	222	16200	12808	1518	712	1:5 million	1.7 × 1.7
9	1° × 1°	69	111	64800	51210	759	356	1:2 million	2.2 × 2.2
10	30' × 30'	34.5	55.5	259000	204840	380	178	1:1 million	2.2 × 2.2
11	15' × 15'	17.25	27.8	1036800	813600	190	89	1:500,000	2.2 × 2.2
12	7.5' × 7.5'	8.625	13.9	4147200	3277440	95	44	1:250,000	2.2 × 2.2
13	3.75' × 3.75'	4.312	6.9	16588800	13109760	47.4	22	1:125,000	2.2 × 2.2
								1:100,000	2.73 × 2.73
14	1.875' × 1.875'	2.156	3.5	66355200	52439040	23.7	11.1	1:80,000	3.4 × 3.4
								1:62,500	2.2 × 2.2
								1:50,000	2.73 × 2.73
15	0.9375' × 0.9375'	1.078	1.7	265420800	209756160	11.9	5.6	1:40,000	3.4 × 3.4
								1:24,000	2.8 × 2.8
								1:20,000	3.4 × 3.4
16	0.46875' × 0.46875'	0.539	0.9	1016683200	839024640	5.9	2.8	1:12,000	2.8 × 2.8

The best way to see the advantages of the 512°×512° mapping projection, is to use it with the previously, taught, quadrant division and pyramid file structure to show how this unique mapping projection can provide digital maps of any geographical areas of the earth, with 16 levels or magnitudes of resolution. As it is slightly

above-indicated chart.

In looking at the left-most column, and tracing down to magnitude 1, note that the 256°×256° window size exactly matches our determination. Furthermore, note that our findings is also in agreement with the number of widows i.e., 4. It is also interesting to note from the third column, that the height or "relative viewing posi-

tion" of this magnitude or level would be 17, 664 statute miles above the earth's surface.

Turning now to the second level or magnitude of resolution (FIG. 15), a further quadrant division is applied, resulting in sixteen (16) mapping areas of $128^\circ \times 128^\circ$. The respective filenames which are assigned to each of the mapping areas is shown. In viewing FIG. 15, note that there are eight (8) mapping areas which are not intersected by the earth's surface. In order to save valuable memory space, the preferred embodiment will ignore, and in fact will never create these files. Note that there is no use for these files as they do not contain any digital mapping data nor will they ever have any descendents which hold mapping data. In order to implement this "file selectivity", the preferred embodiment again utilizes a watchdog approach. More specifically, as the computer already knows the degree ($^\circ$) size of the earth's surface and the degree ($^\circ$) size of each of the mapping areas (i.e., at each level or magnitude of resolution), it can be seen that the computer can easily calculate the filenames which will not intersect the earth's surface.

Again it is useful to correspond our findings with the entries in the chart.

Our findings are substantiated, as, at a magnitude of 2, the window size is shown as being $128^\circ \times 128^\circ$, and there are shown to be eight (8) pertinent windows or files at this magnitude. Again, it is interesting to note that the height or "relative viewing position" of this window would be 8,832 statute miles above the earth's surface.

It is important to note that, although the "relative viewing position" of each level or magnitude is moving closer to the earth, the visual perception of the earth (as seen in FIGS. 14-19) is not illustrated as getting larger with a greater degree of detail. This is because of the paper size limitations.

In the third level or magnitude of resolution (FIG. 16), a further quadrant division is applied, resulting in sixty-four (64) mapping areas of $64^\circ \times 64^\circ$. As the projection is beginning to represent a large plurality of mapping areas, the filenames have been omitted. However, it should be understood that the filename assigned to a respective file in this and subsequent degrees of resolution, can easily be calculated by following the previously described file naming operation. In this projection, it can be seen that 40 mapping areas or files are not used, resulting in 24 files which contain the digital mapping data of this resolution. Note that the observed window, and used files again correlates to the entries in the chart. Furthermore, it can be seen that the height or "relative viewing position" is at 4,416 statute miles above the earth.

Further quadrant divisions and the corresponding data can be seen in the FIGS. 17-19 and the chart. From the foregoing discussions, prior teachings, and data from the chart, one skilled in the art should be able to quickly appreciate that a mapping system can be constructed which can provide digital maps corresponding to a plurality of resolutions, of any geographical area of the world.

The chart can now be used to observe the tremendous advantage provided by the $512^\circ \times 512^\circ$ projection. In the second column of the chart, one can view the sizes of the mapping area divisions which are produced as a result of the continued quadrant division of the $512^\circ \times 512^\circ$ projection. One skilled in the mapping art will be able to fully appreciate that the resultant map-

ping area divisions exactly correspond to well settled and widely used mapping area formats.

Having described all of the important operations of the present invention, the following further conclusions, comments and teachings can be made.

With the mapping system of the present invention, the mapping data are structured at each magnitude or level into windows, frames or tiles representing subdivisions or partitions of the surface area at the specified magnitude. The windows, frames or tiles of all magnitudes for whatever resolution are structured to receive substantially the same amount or quantity of mapping data for segmented visual presentation of the mapping data by window.

As a further improvement, the lapping system of the present invention can further store and organize mapping data into attributed or coded geographical and cultural features according to the classification and level or resolution or magnitude for presentation on the map display. Several examples of this was previously discussed with regard to the use of the filename extension. If this further improvement is used, the computer can be programmed and arranged for managing and accessing the mapping data, and excluding or including coded features in tiles of a particular magnitude according to the resolution and density of mapping data appropriate to the particular magnitude of the window. The selective display of attributed geographical and cultural features according to resolution maintains or limits the mapping data entered in each tile to no greater than a specified full complement of mapping data for whatever magnitude.

In reviewing the file naming operations which were described, one can see that the global map generating system data base structure relates tiles of the same magnitude by tile position coordinates that are keyed to the control corner of each tile and maintained in the name of the "tile-file". Continuity of same scale tiles is maintained during scrolling between adjacent or neighboring tiles in any direction. The new data base structure also relates tiles of different magnitudes by vertical lineage through successive magnitudes. Each tile of a higher magnitude and lower resolution is an "ancestor tile" encompassing a lineage of "descendant tiles" of lower magnitude and high resolution in the next lower magnitude. Thus the present invention permits accessing, displaying and presenting the structured mapping data by tile, by scrolling between adjacent or neighboring tiles of different magnitude in the same vertical lineage for varying the resolution.

In its simplest form the coordinate system is Cartesian, but the invention contemplates a variety of virtual tile manifestations of windowing the mapping data at each magnitude: for example: tilting the axes; scaling one axis relative to another; having one or both axes logarithmic; or rendering the coordinate space as non-Euclidean all together.

When dealing with vector or point information and gridded data, the most common method is to describe individual points as an x-y offset from the control corner of the tile. In this way the mapping data exist as pre-processed relative points on a spherical surface in a de-projected space. The mapping data can then be projected at the user interface with an application program. When projected, all data ultimately represent points of latitude and longitude. Tiles may also contain mapping data as variable offsets of arc in the x and y directions. The tile header may carry an internal descriptor defin-

ing what type of mapping data is contained. The application or display program may then decode and project the data to the appropriate latitude or longitude positions.

The map generating system contemplates storing analog mapping data in electronic mapping frames in which the raw analog data would be scanned and converted digitally to the tile structure and then later accessed and projected for the purpose of displaying continuous analog mapping data.

In the preferred example embodiment, the digital mapping data are structured by window or tile in a substantially rectangular configuration encompassing defined widths and heights in degrees of latitude and longitude for each magnitude. The mapping data representing each magnitude or level are stored in a de-projected format according to mapping on an imaginary cylindrical surface. For display of the maps, however, the data base manager accesses and presents the tiles in a projected form, according to the real configuration of the mapped surface, by varying the aspect ratio of latitude to longitude dimensions of the tiles according to the absolute position of the window on the surface area.

For example, for a spherical or spheroidal globe having an equator and poles, such as the earth, the mapping data are accessed and displayed by aspecting or narrowing the width in the west-east dimension of the tiles of the same magnitude, while scrolling from the equator to the poles. This is accomplished by altering the width of the tile relative to the height. In the graphics display of each window or tile on the monitor, the tiles are presented essentially as rectangles having an aspect ratio substantially equal to the center latitude encompassed by the tile. Thus, the width of the visual display windows is corrected in two respects. First, the overall width is corrected by aspecting to a narrower width, during scrolling in the direction of the poles, and to a wider width during scrolling in the direction of the equator. Second, the width of the tile is averaged to the center latitude width encompassed by the tile throughout the tile height to conserve the rectangular configuration. Alternatively, or in addition, further compensation may be provided by increasing the number of degrees of longitude encompassed by the tiles during scrolling from the equator to the poles to compensate for the compound curvature of the globe.

A feature and advantage of this new method and new system of map projection are that the dramatic and perverse distortion of the globe near the poles, introduced by the traditional and conventional Mercator projection is substantially eliminated. According to the invention, the compensating aspect ratio of latitudinal to longitudinal dimension of aspecting is a function of the distance from the equator, where the aspect ratio is one, to the poles where the aspect ratio approaches zero, all as described for example in *Elements of Cartography*, 4th edition. John Wiley & Sons (1978) by Arthur Robinson, Randall Sale and Joel Morrison.

The new system contemplates "polar compression" (FIG. 20) in the following manner. Starting at 64 degrees latitude, the width of each tile doubles for every eight degrees of latitude. From 72 degrees to 80 degrees latitude, there are 4 degrees of longitude for 1 degree of latitude. From 80 degrees to 88 degrees latitude, it becomes eight to one, and from 88 degrees to the pole (90 degrees) it becomes 16 to one (see illustration of polar compression). (FIG. 20)

Another feature and advantage of the way in which the new map system and new projection handle polar mapping data are in the speed required to access and display polar data. The new polar compression method drastically minimizes tile or window seeks and standard I/O time. Also, without compressing the poles, the Creation/Edit Software would have to work on increasingly narrow tiles as the aspect ratio approached zero at the poles.

The invention embodies an entirely new cartographic organization for an automated atlas of the earth or other generally spherical or spheroidal globe with 360 degrees of longitude and 180 degrees of latitude, an equator and poles. The digital mapping data for the earth is structured on an imaginary surface space having 512 degrees of latitude and longitude. The imaginary 512 degree square surface represents the zero magnitude or root node at the highest level above the earth for a hierarchical type quadtree data base structure. In fact, the 512 degree square plane at the zero magnitude encompasses the entire earth in a single tile. The map of the earth, of course, fills only a portion of the root node window of 512 degrees square, and the remainder may be deemed imaginary space or "hyperspace".

In the preferred example embodiment from a zero magnitude virtual or imaginary space 512 degrees square, the data base structure of the global map generating system descends to a first magnitude of mapping data in four tiles, windows or quadrants, each comprising 256 degrees of latitude and longitude. Each quadrant represents mapping data for one-quarter of the earth thereby mapping 180 degrees of longitude and 90 degrees of latitude in the imaginary surface of the tile or frame comprising 256 degrees square, leaving excess imaginary space or "hyperspace". In the second magnitude, the digital mapping data are virtually mapped and stored in an organization of 16 tiles or windows each comprising 128 degrees of latitude and longitude.

The map generating system supports two windowing formats, one based on the binary system of the 512 degree square zero magnitude root node with hyperspace and the other based on a system of a 360 degree square root node without hyperspace. A feature and advantage of the virtual 512 degree data base structure with hyperspace are that the tiles or windows to be displayed at respective magnitudes are consistent with conventional mapping scale divisions, for example, those followed by the U.S. Geological Survey (USGS), Defense Mapping Agency (DMA), National Aeronautics and Space Administration (NASA) and other government mapping agencies. Thus, typical mapping scale divisions of the USGS and military mapping agencies include scale divisions in the same range of 1 deg, 30 minutes, 15 minutes, 7.5 minutes of arc on the earth's surface. This common subdivision of mapping space does not exist in a data structure based on a 360 degree model without hyperspace (see chart).

Thus, according to the present invention, the world is represented in an assemblage of magnitudes, with each magnitude divided into adjacent tiles or windows on a virtual or imaginary two-dimensional plane or cylinder. At higher magnitudes the quadtree tiles of mapping data do not fill the imaginary projection space. However, from the seventh magnitude down, the mapping data fills a virtual closed cylinder, and no hyperspace exists at these levels.

In the preferred example embodiment the invention (running on a 16 bit computer) has sixteen magnitudes

or levels (with extensions to 20 levels) representing sixteen altitudes or distances above the surface of the earth. At the lowest (16th) magnitude of highest resolution and closest to the earth, the data base structure contains over one billion tiles or windows (excluding hyperspace), each encompassing a tile height of approximately one half statute mile. At this level of resolution, one pixel on a monitor of 480 pixels in height represents approximately 6 feet on the ground. Mapping data are positioned within each tile using a 0 to 1023 offset coordinate structure, resulting in a data resolution of approximately 3 feet at this level of magnitude (see chart). The contemplated 20th magnitude tile or window height is approximately 175 feet, which results in a pixel resolution of about 4 inches on a monitor of 480 pixels in height and a data resolution of about 2 inches, when utilizing the 0 to 1023 offset coordinate structure. Alternatively, the map-generating system contemplates an extended offset from 10 bits (0 to 1023) to an offset of 16 bits (0 to 65,535). In this case, the extended 20th magnitude results in a data resolution of 3 hundredths of an inch.

For still more resolution, the map generating system contemplates 32 magnitudes on a 32 bit computer and representing 32 altitudes or distances about the surface of the earth. Each level of magnitude may define mapping data within each tile using a 32 bit offset coordinate structure, thereby giving relative mathematical accuracy to a billionth of an inch. In all practicality, 20 separate magnitudes or levels are more than sufficient to carry the necessary levels of resolution and accuracy.

The new invention provides users with the ability graphically to view mapping data from any part of the world-wide data base graphically on a monitor, either by entering coordinates and a level of zoom (or magnitude) on the keyboard, or by "flying" to that location in the "step-zoom" mode using consecutive clicks of the mouse or other pointing device. Once a location has been chosen (this point becomes the user-defined screen center), the mapping software accesses all adjacent tiles needed to fill the entire view window of the monitor and, then, projects the data to the screen. Same scale scrolling is accomplished by simply choosing a new screen center and maintaining the same magnitude.

Vertical zooming up or down is accomplished by choosing another magnitude or level from the menu area with the pointing device or by directly entering location and magnitude on the keyboard. An advantage of this vertical lineage of tiles organized in a quadtree structure is that it affords the efficient and easily followed zooming continuity inherent in the present invention. Further discussion of such quadtree data organization is found in the article, "The Quadtree and Related Hierarchical Data Structures", by Hannan Samet, Computer Surveys, Volume 16, No. 2, (June 1984), Pages 187 et seq.

The map-generating system also supports many types of descriptive information such as that contained in tabular or relational data bases. This descriptive information can be linked to the mapping data with a latitude and longitude coordinate position but may need to be displayed in alternate ways. Descriptive information is better suited for storage in a relational format and can be linked to the map with a "spatial hook".

In summary, the present invention provides a new automated world atlas and global map generating system having a multi-level hierarchial quadtree data base structure and a data base manager or controller which

permits scrolling, through mapping tiles or windows of a particular magnitude, and zooming between magnitudes for varying resolution. While the data base organization is hierarchial between levels or magnitudes, it is relational within each level, resulting in a three dimensional network of mapping and descriptive information. The present invention also provides a new mapping projection that has similarities to the Mercator projection but eliminates drastic distortions near the poles for the purpose of presentation through a method of "aspecting" tile widths as a function of the latitudinal distance from the equator.

While the invention has been particularly shown and described with reference to the preferred embodiment thereof, it will be understood by those skilled in the art that various changes in form and details of the device and the method may be made therein without departing from the spirit and scope of the invention.

What is claimed is:

1. A computer implemented method for generating, displaying and presenting an electronic map from digital mapping data for a surface area having geographical and cultural features, said method comprising the steps of:

organizing the mapping data into a hierarchy of a plurality of successive magnitudes or levels for presentation of said mapping data with variable degrees of mapping resolution, each magnitude for presentation of said mapping data with a different degree of mapping resolution from a first or highest magnitude with lowest resolution to a last or lowest magnitude with highest resolution;

structuring said mapping data at each magnitude into a plurality of windows, frames or files representing subdivisions or partitions of said surface area, said windows of a respective magnitude including mapping data which are appropriate to a degree of mapping resolution being afforded at said magnitude while excluding mapping data which are not appropriate to said degree of mapping resolution, and at least a portion of said windows of each magnitude being structured to receive substantially a same predetermined amount or quantity of mapping data for segmented presentation of the mapping data by window;

organizing said mapping data into records of geographical or cultural features for presentation within said windows, and coding said features;

managing said mapping data for each window by excluding or including coded features appropriate to the degree of mapping resolution and density being afforded by said window, such that a quantity of mapping data entered in each window is no greater than said predetermined amount;

relating windows of a same magnitude by window position coordinates or names and structuring said windows with overlap or mapping data between adjacent or neighboring windows of a magnitude or achieve display continuity during generation, display and presentation of an electronic map;

relating windows of different magnitude by vertical lineage through successive magnitudes, each window of a higher magnitude and lower resolution being an ancestor window being related to a plurality of descendant windows of lower magnitude and higher resolution in a next lower magnitude;

accessing and displaying or presenting mapping data for different positions of a selected magnitude by

scrolling between adjacent or neighboring windows of a same magnitude in predetermined north, south, east and west directions;

and accessing and displaying or presenting mapping data for different selected magnitudes having different resolutions by zooming between windows of different magnitudes in a same vertical lineage.

2. The method of claim 1 further comprising:

organizing said mapping data of said surface area by degrees of latitude and longitude;

structuring each said window of mapping data to represent a substantially rectangular surface area configuration encompassing defined degrees of latitude and longitude for each magnitude, and storing the mapping data for each magnitude in a vertical Mercator projection format;

accessing and presenting said windows of mapping data in a corrected or compensated projection format departing from said Mercator projection format according to a real configuration of said surface area, by varying an aspect ratio of latitude to longitudinal dimensions of each window according to a coordinate position of said window with respect to a coordinate layout of said surface area.

3. The method of claim 2 wherein said surface area comprises a spherical or spheroidal globe having an equator and poles, said method comprising the further steps of:

accessing and presenting mapping data in a corrected projection format by aspecting or narrowing, in a direction from an equator to pole, the width or latitudinal dimension of windows, of a same magnitude, which encompass the same number of degrees of latitude and longitude;

and periodically increasing a number of degrees of longitude encompassed by said windows in said direction from equator to pole to compensate for compound curvature of said globe.

4. The method of claim 1 wherein said surface area comprises a generally spherical or spheroidal globe with 360 degrees of longitudinal, 180 degrees of latitude and an equator and poles, said method comprising the further steps of:

relating windows of different magnitudes by vertical lineage in a hierarchical quadtree database structure, by successively partitioning or subdividing ancestor windows of a vertical lineage into four descent windows or quadrants at a next lower magnitude or level, and incorporating additional records of features in said descendant windows to incorporate mapping data for a next higher resolution.

5. The method of claim 4 wherein said hierarchical quadtree database structure comprises at least sixteen degrees of magnitudes or levels.

6. The method of claim 4 comprising the further steps of:

mapping and storing mapping data for said globe in a virtual Mercator projection format representing an imaginary surface having 512 degrees of longitude and latitude comprising a zero magnitude or root node of said hierarchical quadtree database structure;

mapping and storing a first degree or highest magnitude of mapping data in four windows or quadrants each comprising 256 degrees of longitude and latitude, each window of said first degree of magnitude comprising mapping data for one quarter of

said globe thereby mapping 180 degrees of surface area longitude and 90 degrees of surface area latitude in said imaginary surface of 256 degrees of longitude and latitude and leaving excess imaginary space;

mapping and storing a second degree of magnitude of mapping data in sixteen windows each comprising 128 degrees of longitude and latitude of said imaginary surface, each window of said second degree of magnitude comprising mapping data for a further subdivision or partition of said globe;

and mapping and storing third through twelfth degrees of magnitude thereby forming additional levels of a hierarchical quadtree database structure so that an eleventh magnitude comprises windows encompassing 15 seconds of latitude and a twelfth magnitude comprises windows encompassing seven and a half seconds of latitude;

whereby, as a result of the foregoing, windows of said electronic map at respective magnitudes or levels are consistent with conventional mapping scale divisions.

7. The method of claim 6 wherein said hierarchical quadtree database structure comprises sixteen degree of magnitudes or levels including a sixteenth magnitude comprising over 1.4 billion windows, each encompassing approximately a fraction of a minute of a degree of latitude.

8. The method of claim 6 wherein each said window corresponds to a trapezoidal surface area configuration.

9. The method of claim 6 comprising the step of floating mapping data records of selected features from a window of one magnitude to a window of the same vertical lineage in another magnitude.

10. The method of claim 6 comprising the further steps of: generating analog mapping data, structuring said analog mapping data according to a same format as digital mapping data, and overlaying and presenting said digital mapping data and analog mapping data during generation, display and presentation of an electronic map.

11. The method of claim 6 comprising the further step of selectively filling said windows with mapping data so that some windows contain a full complement of mapping data appropriate to a degree of mapping resolution being afforded at said magnitude, and other windows, each of which correspond to a subdivision of surface area containing few or no geographical or cultural features, contain less than a full complement of mapping data.

12. The method of claim 6 comprising the further steps of:

accessing and presenting mapping data in a corrected projection format by aspecting or narrowing, in a direction from an equator to pole, a width or latitudinal dimension of windows, of a same magnitude, which encompass the same number of degrees of latitude and longitude;

and periodically increasing a number of degrees of longitude encompassed by said windows in said direction from equator to pole to compensate for a compound curvature of said globe.

13. The method of claim 12 comprising the further steps of accessing and presenting mapping data in corrected projection format, with each window having a width substantially equal to a center latitude width of said window throughout said window, so that said window is of rectangular configuration.

14. An electronic map generating system including a digital computer, a mass storage device, a display monitor, graphics controller, and system software for structuring, managing, controlling and displaying digital mapping data for a surface area having cultural and geographical features, said system comprising:

a database structure comprising a hierarchical database structure programmed and arranged for organizing said digital mapping data into a hierarchy of a plurality of successive magnitudes or levels for presentation of mapping data with variable resolution, each magnitude for presentation of mapping data with a different degree of mapping resolution from a first or highest magnitude of lowest resolution to a last or lowest magnitude of highest resolution, and for structuring said digital mapping data at each magnitude into a plurality of windows, frames or files representing subdivisions or partitions of said surface area, said windows of a respective magnitude including mapping data which are appropriate to a degree of mapping resolution being afforded at said magnitude while excluding mapping data which are not appropriate to said degree of mapping resolution, at least a portion of said windows of all magnitudes being structured to receive substantially a same predetermined amount of mapping data for segmented presentation of said mapping data by window, said mapping data being organized into coded records of geographical and cultural features within each window;

a database manager or controller programmed and arranged for managing said mapping data by magnitude or level by excluding or including coded records of features in each window of a particular magnitude according to a resolution and density of mapping data appropriate to the particular magnitude of said each window, and maintaining a quantity of mapping data entered in each window to no greater than a specified full complement whatever the magnitude of the window;

said database structure being programmed to relate windows of a same magnitude by position coordinates or names, and to structure windows of a same magnitude with overlap of mapping data between adjacent or neighboring windows of a magnitude to achieve display continuity during generation, display and presentation of an electronic map, and to relate windows of different magnitude by vertical lineage through successive magnitudes, each window of a higher magnitude and lower resolution being an ancestor window of a plurality of descendant windows of lower magnitude and higher resolution in a next lower magnitude;

said database manager being programmed to access and display or present mapping data for different positions of a selected magnitude by scrolling between adjacent or neighboring windows of a same magnitude in predetermined north, south, east and west directions, and being programmed to access and display or present mapping data for different magnitudes having different resolutions by zooming between windows of different magnitudes in a same vertical lineage.

15. The system of claim 14 wherein said hierarchical database structure is programmed to organize said mapping data by degrees of latitude and longitude and to structure each window of mapping data to represent a

substantially rectangular surface area configuration encompassing predetermined degrees of latitude and longitude, said windows for each magnitude being stored in virtual Mercator projection format, said database manager being programmed to access and present windows of mapping data in a corrected or compensated projection format departing from Mercator projection format according to a real configuration of said surface area by varying an aspect ratio of latitude and longitude dimensions of each window according to a coordinate position of said each window with respect to a coordinate layout of said surface area.

16. The system of claim 15 wherein said surface area comprises a spherical or spheroidal globe having an equator and poles, and wherein said database manager is programmed to access and present mapping data in a corrected projection format by aspecting or narrowing, in a direction from an equator to pole, the width or latitudinal dimension of windows, of a same magnitude, which encompass the same number of degrees of longitude, said database manager being further programmed to periodically increase a number of degrees of longitude encompassed by said windows in said direction from equator to pole to compensate for compound curvature of said globe.

17. The system of claim 16 wherein said hierarchical database structure comprises a hierarchical quadtree database structure successively partitioning or subdividing ancestor windows of a vertical lineage into four descendant windows or quadrants at a next lower magnitude or level, and incorporating additional coded records of features in said descendant windows to incorporate mapping data for a next higher resolution.

18. The system of claim 17 wherein said database structure is programmed and arranged to store the mapping data in a virtual Mercator projection representing an imaginary surface having 512 degrees of longitude and latitude comprising a zero magnitude or root node of said hierarchical quadtree database structure, wherein a first degree or first magnitude of mapping data comprises four windows, each window of said first magnitude comprising mapping data for one quarter of said globe on an imaginary surface area of 256 degrees of longitude and latitude, said hierarchical quadtree database structure comprising, in addition to first through tenth magnitudes each having windows which are predetermined subdivisions of said imaginary surface having 512 degrees of longitude and latitude, at least an eleventh magnitude having windows encompassing 15 minutes of latitude, and a twelfth magnitude having windows encompassing 7.5 minutes of latitude, so that windows of a resultant electronic map at respective said eleventh and twelfth magnitudes or levels are consistent with conventional mapping scale divisions.

19. The system of claim 18 wherein said hierarchical quadtree database structure comprises at least 16 degrees of magnitudes or levels, said sixteenth magnitude comprising over 1.4 billion windows, each encompassing degrees of latitude of approximately a fraction of a second of a degree.

20. The system of claim 19 further comprising a database of digital mapping data selectively entered in said database structure, such that some of said windows contain a full complement of mapping data appropriate to a degree of mapping resolution being afforded at said magnitude, and other windows, each of which correspond to a subdivision of surface area containing few or

no geographical or cultural features, contain less than a full complement of mapping data.

21. The system of claim 19 further comprising a database of analog data structured according to a same format as said digital data, and means for overlaying said digital and analog data for electronic map presentation.

22. An electronic map generating system for generating reproductions of a map with selectable degrees of mapping resolution, said map generating system comprising:

database means storing a plurality of computer files containing mapping data corresponding to respective surface areas of a mapping surface, wherein said plurality of computer files is organized into a plurality of successive magnitudes, each magnitude for presentation of said mapping data with a different degree of mapping resolution from a first or highest magnitude with lowest resolution to a last or lowest magnitude with highest resolution, files of a respective magnitude including mapping data which are appropriate to a degree of mapping resolution being afforded at said respective magnitude while excluding mapping data which are not appropriate to said degree of mapping resolution, and wherein a predetermined file naming procedure is utilized to assign, to each respective computer file, a unique filename which:

relates said respective computer file to all other computer files having mapping data corresponding to a same magnitude or degree of mapping resolution; and

relates said respective computer file to any computer file comprising mapping data corresponding to a same surface area of a mapping surface as said respective computer file; and

database manager means for accessing said plurality of computer files using said predetermined file naming procedure, to generate a reproduction of a selected area of a map at a selected degree of mapping resolution.

23. An electronic map generating system as claimed in claim 22,

wherein each said unique filename is represented by a value contained in a plurality of bits, and

wherein said predetermined file naming procedure: utilizes a first predetermined subset of said plurality of bits to relate said respective files having mapping data corresponding to a same magnitude or degree of mapping resolution; and

utilizes a second predetermined subset of said plurality of bits to relate said respective computer file to any computer file comprising mapping data corresponding to a same surface area of a mapping surface as said respective computer file.

24. An electronic map generating system as claimed in claim 23, wherein said unique filename also includes geographical information which can be used to relate a geographical coordinate position of a respective computer file with respect to a coordinate layout of surface areas of said mapping surface.

25. An electronic map generating system as claimed in claim 22,

wherein an assignment of said unique filenames using said predetermined file naming procedure results in said respective computer files of said plurality to be related in a quadtree database structure.

26. An electronic map generating system as claimed in claim 25, wherein the respective area of a mapping surface covered within the computer files of consecutive magnitudes or degrees of mapping resolution changes at a predetermined rate in that, when a computer file at a reference magnitude or degree of mapping resolution contains mapping data corresponding to an $N \times N$ area of a mapping surface (where N is a real number, and is associated with one of the conventional degree $^\circ$, minute $'$, or second $''$ mapping scale divisions), then a computer file at a next consecutive magnitude having a higher degree of mapping resolution contains mapping data corresponding to an $(N/2) \times (N/2)$ area of said mapping surface.

27. An electronic map generating system as claimed in claim 26, wherein the value of N at said reference magnitude or degree of mapping resolution, corresponds to one of the following values: 512° , 256° , 128° , 64° , 32° , 16° , 8° , 4° , 2° , 1° , $30'$, $15'$, $7.5'$, $3.75'$, $1.875'$, $0.9375'$ and $0.46875'$.

28. A method for providing an electronic map generating system for generating reproductions of a map with selectable degrees of mapping resolution, said method comprising the steps of:

storing a plurality of computer files containing mapping data corresponding to respective surface areas of a mapping surface, wherein said plurality of computer files is organized into a plurality of successive magnitudes, each magnitude for presentation of said mapping data with a different degree of mapping resolution from a first or highest magnitude with lowest resolution to a last or lowest magnitude with highest resolution, files of a respective magnitude including mapping data which are appropriate to a degree of mapping resolution being afforded at said respective magnitude while excluding mapping data which are not appropriate to said degree of mapping resolution, and wherein a predetermined file naming procedure is utilized to assign, to each respective computer file, a unique filename which:

relates said respective computer file to all other computer files having mapping data corresponding to a same magnitude or degree of mapping resolution; and

relates said respective computer file to any computer file comprising mapping data corresponding to a same surface area of a mapping surface as said respective computer file; and

accessing said plurality of computer files using said predetermined file naming procedure, to generate a reproduction of a selected area of a map at a selected degree of mapping resolution.

29. A method as claimed in claim 28, wherein each said unique filename is represented by a value contained in a plurality of bits, and

wherein said predetermined file naming procedure; utilizes a first predetermined subset of said plurality of bits to relate said respective computer file to all other computer files having mapping data corresponding to a same magnitude or degree of mapping resolution; and

utilizes a second predetermined subset of said plurality of bits to relate said respective computer file to any computer file comprising mapping data corresponding to a same surface area of a mapping surface as said respective computer file.

37

38

30. A method as claimed in claim 29, wherein said unique filename also includes geographical information which can be used to relate a geographical coordinate position of a respective computer file with respect to a coordinate layout of surface areas of said mapping surface.

31. A method as claimed in claim 28, wherein an assignment of said unique filenames using said predetermined file naming procedure results in said respective computer files of said plurality to be related in a quadtree database structure.

32. A method as claimed in claim 31, wherein the respective area of a mapping surface covered within the computer files of consecutive magnitudes or degrees of mapping resolution changes at a predetermined rate in

that, when a computer file at a reference magnitude or degree of mapping resolution contains mapping data corresponding to an $N \times N$ area of a mapping surface (where N is a real number, and is associated with one of the conventional degree °, minute ', or second '' mapping scale divisions), then a computer file at a next consecutive magnitude having a higher degree of mapping resolution contains mapping data corresponding to an $(N/2) \times (N/2)$ area of said mapping surface.

33. A method as claimed in claim 32, wherein the value of N at said reference magnitude or degree of mapping resolution, corresponds to one of the following values: 512° , 256° , 128° , 64° , 32° , 16° , 8° , 4° , 2° , 1° , $30'$, $15'$, $7.5'$, $3.75'$, $1.875'$, $0.9375'$ and $0.46875'$.

* * * * *

20

25

30

35

40

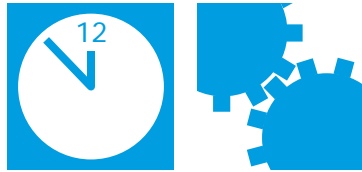
45

50

55

60

65



The MAGIC Project: From Vision to Reality

Barbara Fuller, Mitretek Systems
Ira Richer, Corporation for National Research Initiatives

Abstract

In the MAGIC project, three major components — an ATM internetwork, a distributed, network-based storage system, and a terrain visualization application — were designed, implemented, and integrated to create a testbed for demonstrating real-time, interactive exchange of data at high speeds among distributed resources. The testbed was developed as a system, with special consideration to how performance was affected by interactions among the components. This article presents an overview of the project, with emphasis on the challenges associated with implementing a complex distributed system, and with coordinating a multi-organization collaborative project that relied on distributed development. System-level design issues and performance measurements are described, as is a tool that was developed for analyzing performance and diagnosing problems in a distributed system. The management challenges that were encountered and some of the lessons learned during the course of the three-year project are discussed, and a brief summary of MAGIC-II, a recently initiated follow-on project, is given.

Gigabit-per-second networks offer the promise of a major advance in computing and communications: high-speed access to remote resources, including archives, time-critical data sources, and processing power. Over the past six years, there have been several efforts to develop gigabit networks and to demonstrate their utility, the most notable being the five testbeds that were supported by ARPA and National Science Foundation (NSF) funding: Aurora, BLANCA, CASA, Nectar, and VISTAnet [1]. Each of these testbeds comprised a mix of applications and networking technology, with some focusing more heavily on applications and others on networking. The groundbreaking work done in these testbeds had a significant impact on the development of high-speed networking technology and on the rapid progress in this area in the 1990s.

It became clear, however, that a new paradigm for application development was needed in order to realize the full benefits of gigabit networks. Specifically, network-based applications and their supporting resources, such as data servers, must be designed explicitly to operate effectively in a high-speed networking environment. For example, an interactive application working with remote storage devices must compensate for network delays. The MAGIC project, which is the subject of this article, is the first high-speed networking testbed that was implemented according to this paradigm. The major components of the testbed were considered to be interdependent parts of a system, and wherever possible they were designed to optimize end-

to-end system performance rather than individual component performance.

The objective of the MAGIC (which stands for “Multidimensional Applications and Gigabit Internetwork Consortium”) project was to build a testbed that could demonstrate real-time, interactive exchange of data at gigabit-per-second rates among multiple distributed resources. This objective was pursued through a multidisciplinary effort involving concurrent development and subsequent integration of three testbed components:

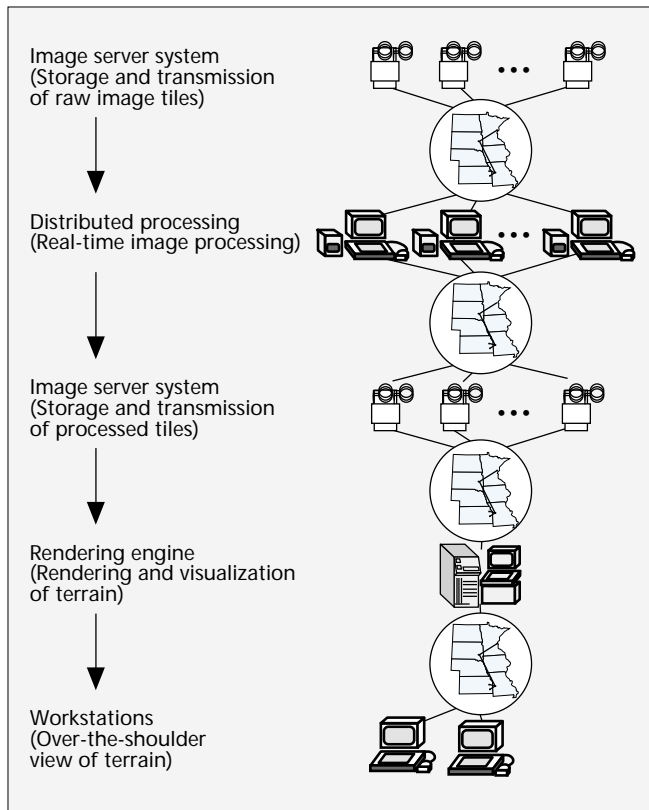
- An innovative terrain visualization application that requires massive amounts of remotely stored data
- A distributed image server system with performance sufficient to support the terrain visualization application
- A standards-based high-speed internetwork to link the computing resources required for real-time rendering of the terrain

The three-year project began in mid-1992 and involved the participation, support, and close cooperation of many diverse organizations from government, industry, and academia. These organizations had complementary skills and had the foresight to recognize the benefits of collaboration. The principal MAGIC research participants were:

- Earth Resources Observation System Data Center, U.S. Geological Survey (EDC)¹
- Lawrence Berkeley National Laboratory, U.S. Department of Energy (LBNL)¹
- Minnesota Supercomputer Center, Inc. (MSCI)¹
- MITRE Corporation¹
- Sprint
- SRI International (SRI)¹

The work reported here was performed while the authors were with the MITRE Corp. in Bedford, MA, and was supported by the Advanced Research Project Agency (ARPA) under contract F19628-94-D-001.

¹These organizations were funded by ARPA.



■ Figure 1. Planned functionality of the MAGIC testbed.

- University of Kansas (KU)¹
- U S WEST Communications, Inc.
- Other MAGIC participants that contributed equipment, facilities, and/or personnel to the effort were:
- Army High-Performance Computing Research Center (AHPCRC)
- Battle Command Battle Laboratory, U.S. Army Combined Arms Command (BCBL)
- Digital Equipment Corporation (DEC)
- Nortel, Inc./Bell Northern Research
- Southwestern Bell Telephone
- Splitrock Telecom

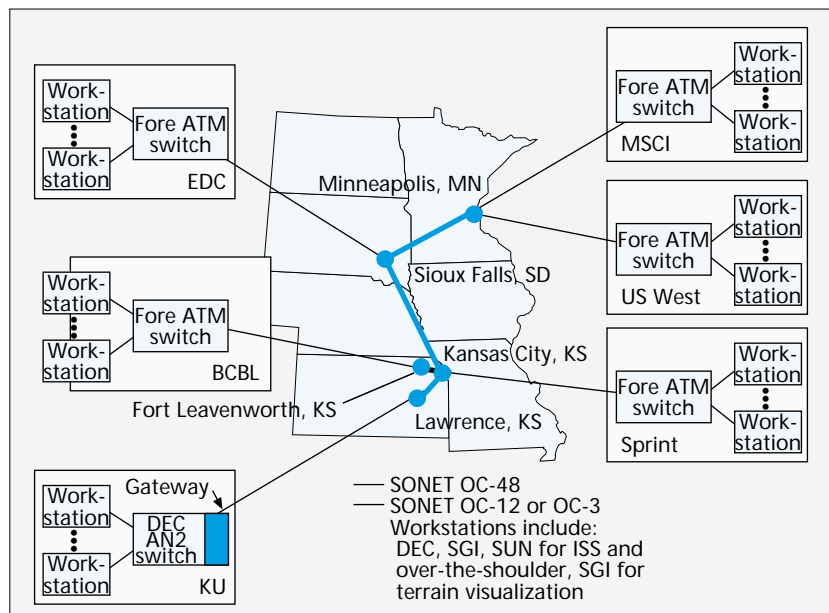
This article presents an overview of the MAGIC project with emphasis on the challenges associated with implementing a complex distributed system. Companion articles [2, 3] focus on a LAN/WAN gateway and a performance analysis tool that were developed for the MAGIC testbed. The article is organized as follows. The following section briefly describes the three major testbed components: the internetwork, the image server system, and the application. The third section discusses some of the system-level considerations that were addressed in designing these components, and the fourth section presents some high-level performance measurements. The fifth (affectionately entitled "Herding Cats") and sixth sections describe how this multi-organizational collaborative project was coordinated, and the technical and managerial lessons learned. Finally, the last section provides a brief summary of MAGIC-II, a follow-on project begun in early 1996.

Overview of the MAGIC Testbed

One of the primary goals of the MAGIC project was to create a testbed to demonstrate advanced capabilities that would not be possible without a very high-speed internetwork. MAGIC accomplished this goal by implementing an interactive terrain visualization application, TerraVision, that relies on a distributed image server system (ISS) to provide it with massive amounts of data in real time. The planned functionality of the MAGIC testbed is depicted in Fig. 1. Currently, TerraVision uses data processed off-line and stored on the ISS. In the future the application will be redesigned to enable real-time image processing as well as real-time terrain visualization (see the last section). Note that the workstations which house the application, the servers of the ISS, and the "over-the-shoulder" tool (see subsection entitled "The Terrain Visualization Application"), as well as those that will perform the on-line image processing, can reside anywhere on the network.

The MAGIC Internetwork

The MAGIC internetwork, depicted in Fig. 2, includes six high-speed local area networks (LANs) interconnected by a wide area network (WAN) backbone. The backbone, which spans a distance of approximately 600 miles, is based on synchronous optical network (SONET) technology and provides OC-48 (2.4 Gb/s) trunks, and OC-3 (155 Mb/s) and OC-12 (622 Mb/s) access ports. The LANs are based on asynchronous transfer mode (ATM) technology. Five of the LANs — those at BCBL in Fort Leavenworth, Kansas, EDC in Sioux Falls, South Dakota, MSCI in Minneapolis, Minnesota, Sprint in Overland Park, Kansas, and U S WEST in Minneapolis, Minnesota — use FORE Systems models ASX-100 and ASX-200 switches with OC-3c and 100 Mb/s TAXI interfaces. The ATM LAN at KU in Lawrence, Kansas, uses a DEC AN2 switch, a precursor to the DEC GigaSwitch/ATM, with OC-3c interfaces. The network uses permanent virtual circuits (PVCs) as well as switched virtual circuits (SVCs) based on both SPANS, a FORE Systems signaling protocol, and the ATM Forum User-Network Interface (UNI) 3.0 Q.2931 signaling stan-



■ Figure 2. Configuration of the MAGIC ATM internetwork.

dard. The workstations at the MAGIC sites include models from DEC, SGI, and Sun. As part of MAGIC, an AN2/SONET gateway with an OC-12c interface was developed to link the AN2 LAN at KU to the MAGIC backbone [2].

In addition to implementing the internetwork, a variety of advanced networking technologies were developed and studied under MAGIC. A high-performance parallel interface (HIPPI)/ATM gateway was developed to interface an existing HIPPI network at MSC I to the MAGIC backbone. The gateway is an IP router rather than a network-layer device such as a broadband integrated services digital network (B-ISDN) terminal adapter, and was implemented in software on a high-performance workstation (an SGI Challenge). This architecture provides a programmable platform that can be modified for network research, and in the future can readily take advantage of more powerful workstation hardware. In addition, the platform is general-purpose; that is, it is capable of supporting multiple HIPPI interfaces as well as other interfaces such as fiber distributed data interface (FDDI).

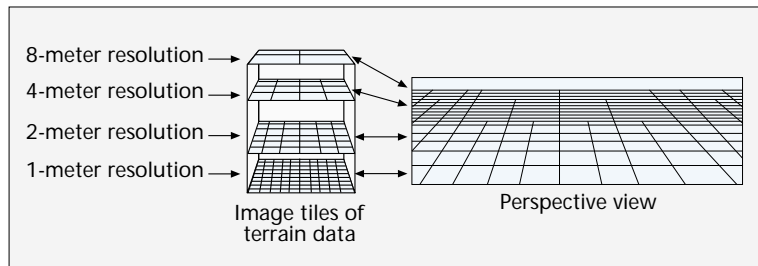
Software was developed to enable UNIX hosts to communicate using Internet Protocol (IP) over an ATM network. This IP/ATM software currently runs on SPARCstations under Sun OS 4.1 and includes a device driver for the FORE SBA series of ATM adapters. It supports PVCs, SPANS, and UNI 3.0 signaling, as well as the "classical" IP and Address Resolution Protocol (ARP) over ATM model [4]. The software should be extensible to other UNIX operating systems, ATM interfaces, and IP/ATM address-resolution and routing strategies, and will facilitate research on issues associated with the integration of ATM networks into IP internets.

In order to enhance network throughput, flow-control schemes were evaluated and applied, and IP/ATM host parameters were tuned. Experiments showed that throughput close to the maximum theoretically possible could be attained on OC-3 links over long distances. To achieve high throughput, both the maximum transmission unit (MTU) and the Transmission Control Protocol (TCP) window must be large, and flow control must be used to ensure fairness and to avoid cell loss if there are interacting traffic patterns [5, 6].

The Terrain Visualization Application

TerraVision allows a user to view and navigate through (i.e., "fly over") a representation of a landscape created from aerial or satellite imagery [7]. The data used by TerraVision are derived from raw imagery and elevation information which have been preprocessed by a companion application known as TerraForm. TerraVision requires very large amounts of data in real time, transferred at both very bursty and high steady rates. Steady traffic occurs when a user moves smoothly through the terrain, whereas bursty traffic occurs when the user jumps ("teleports") to a new position. TerraVision is designed to use imagery data that are located remotely and supplied to the application as needed by means of a high-speed network. This design enables TerraVision to provide high-quality, interactive visualization of very large data sets in real time. TerraVision is of direct interest to a variety of organizations, including the Department of Defense. For example, the ability of a military officer to see a battlefield and to share a common view with others can be very effective for command and control.

Terrain visualization with TerraVision involves two activities: generating the digital data set required by the appli-



■ Figure 3. Relationship between tile resolutions and perspective view. (Source: SRI International)

cation, and rendering the image. MAGIC's approach to accomplishing these activities is described below. Enhancements to the application that provide additional features and capabilities are also described.

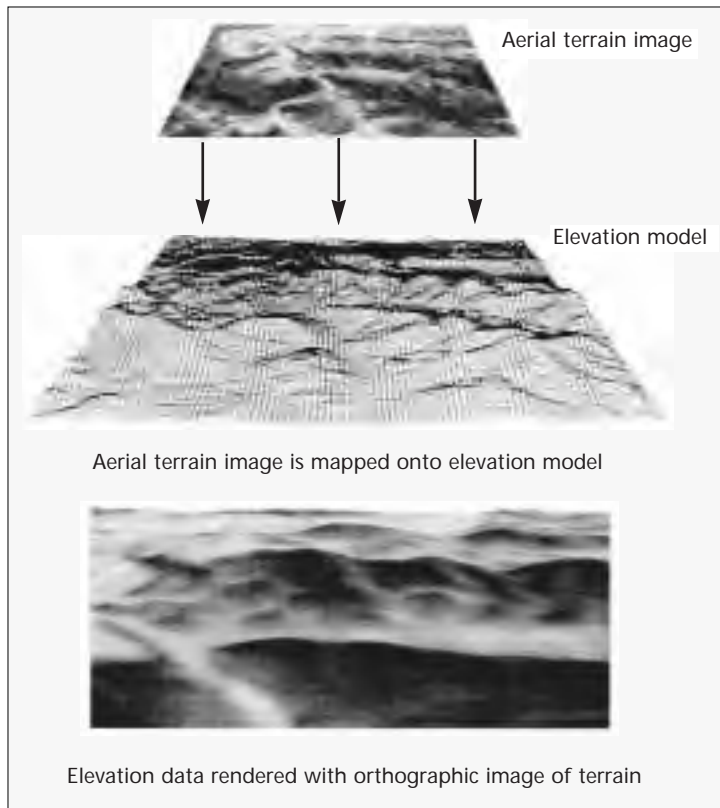
Data Preparation — In order to render an image, TerraVision requires a digital description of the shape and appearance of the subject terrain. The shape of the terrain is represented by a two-dimensional grid of elevation values known as a *digital elevation model* (DEM). The appearance of the terrain is represented by a set of aerial images, known as *orthographic projection images* (ortho-images), that have been specially processed (i.e., ortho-rectified) to eliminate the effects of perspective distortion, and are in precise alignment with the DEM. To facilitate processing, distributed storage, and high-speed retrieval over a network, the DEM and images are divided into small fixed-size units known as *tiles*.

Low-resolution tiles are required for terrain that is distant from the viewpoint, whereas high-resolution tiles are required for close-in terrain. In addition, multiple resolutions are required to achieve perspective. These requirements are addressed by preparing a hierarchy of increasingly lower-resolution representations of the DEM and ortho-image tiles in which each level is at half the resolution of the previous level. The tiled, multiresolution hierarchy and the use of multiple resolutions to achieve perspective are shown in Fig. 3.

Rendering of the terrain on the screen is accomplished by combining the DEM and ortho-image tiles for the selected area at the appropriate resolution. As the user travels over the terrain, the DEM tiles and their corresponding ortho-image tiles are projected onto the screen using a perspective transform whose parameters are determined by factors such as the user's viewpoint and field of view. The mapping of a transformed ortho-image to its DEM and the rendering of that image are shown in Fig. 4.

The data set currently used in MAGIC covers a 1200 km² exercise area of the National Training Center at Fort Irwin, California, and is about 1 Gpixel in size. It is derived from aerial photographs obtained from the National Aerial Photography Program archives and DEM data obtained from the U.S. Geological Survey. The images are at approximately 1 m resolution (i.e., the spacing between pixels in the image corresponds to 1 m on the ground). The DEM data are at approximately 30 m resolution (i.e., elevation values in meters are at 30 m intervals).

Software for producing the ortho-images and creating the multiresolution hierarchy of DEM and ortho-image tiles was developed as part of the MAGIC effort. These processes were performed "off-line" on a Thinking Machines Corporation Connection Machine (CM-5) supercomputer owned by the AHPCRC and located at MSC I. The tiles were then stored on the distributed servers of the ISS and used by terrain visualization software residing on rendering engines at several locations.



■ Figure 4. Mapping an ortho-image onto its digital elevation model. (Source: SRI International)

Image Rendering — TerraVision provides for two modes of visualization: two-dimensional (2-D) and three-dimensional (3-D). The 2-D mode allows the user to fly over the terrain, looking only straight down. The user controls the view by means of a 2-D input device such as a mouse. Since virtually no processing is required, the speed at which images are generated is limited by the throughput of the system comprising the ISS, the network, and the rendering engine.

In the 3-D mode, the user controls the visualization by means of an input device that allows six degrees of freedom in movement. The 3-D mode is computationally intensive, and satisfactory visualization requires both high frame rates (i.e., 15–30 frames/s) and low latencies (i.e., no more than 0.1 s between the time the user moves an input device and the time the new frame appears on the screen).

High frame rates are achieved by using a local very-high-speed rendering engine, an SGI Onyx, with a cache of tiles covering not only the area currently visible to the user, but also adjacent areas that are likely to be visible in the near future. A high-speed search algorithm is used to identify the tiles required to render a given view. For example, as noted above, perspective (i.e., 3-D) views require higher-resolution tiles in the foreground and lower-resolution tiles in the background. TerraVision requests the tiles from the ISS, places them in memory, and renders the view. Latency is minimized by separating image rendering from data input/output (I/O) so that the two activities can proceed simultaneously rather than sequentially (see the section entitled “Design Considerations”).

Additional Features and Capabilities — TerraVision includes two additional features: superposition of fixed and mobile objects on the terrain, and registration of the user’s viewpoint to a map. Both of these features are made

possible by precisely aligning the DEM and imagery data with a world coordinate system as well as with each other.

A number of buildings and vehicles have been created and stored on the rendering engine for display as an overlay on the terrain. The locations of vehicles can be updated periodically by transferring vehicle location data, acquired with a global positioning system receiver, to the rendering engine for integration into the terrain visualization displays. Registration of the user’s viewpoint to a map enables the user to specify the area he wishes to explore by pointing to it, and it aids the user in orienting himself.

In addition, an over-the-shoulder (OTS) tool was developed to allow a user at a remote workstation to view the terrain as it is rendered. The OTS tool is based on a client/server design and uses XWindow system calls. The user can view the entire image on the SGI screen at low resolution, and can also select a portion of the screen to view at higher resolution. The frame rate varies with the size and resolution of the viewed image, and with the throughput of the workstation.

The Image Server System

The ISS stores, organizes, and retrieves the processed imagery and elevation data required by TerraVision for interactive rendering of the terrain. The ISS consists of multiple coordinated workstation-based data servers that operate in parallel and are designed to be distributed around a WAN. This

architecture compensates for the performance limitations of current disk technology. A single disk can deliver data at a rate that is about an order of magnitude slower than that needed to support a high-performance application such as TerraVision. By using multiple workstations with multiple disks and a high-speed network, the ISS can deliver data at an aggregate rate sufficient to enable real-time rendering of the terrain. In addition, this architecture permits location-independent access to databases, allows for system scalability, and is low in cost. Although redundant arrays of inexpensive disks (RAID) systems can deliver higher throughput than traditional disks, unlike the ISS they are implemented in hardware and, as such, do not support multiple data layout strategies; furthermore, they are relatively expensive. Such systems are therefore not appropriate for distributed environments with numerous data repositories serving a variety of applications.

The ISS, as currently used in MAGIC, comprises four or five UNIX workstations (including Sun SPARCstations, DEC Alphas, and SGI Indigos), each with four to six fast SCSI disks on two to three SCSI host adapters. Each server is also equipped with either a SONET or a TAXI network interface. The servers, operating in parallel, access the tiles and send them over the network, which delivers the aggregate stream to the host. This process is illustrated in Fig. 5. More details about the design and operation of the ISS can be found in [8].

Design Considerations

In MAGIC, the single most perspicuous criterion of successful operation is that the end user observes satisfactory performance of the interactive TerraVision application. When the user flies over the terrain, the displayed scene must flow smoothly, and when he teleports to an entirely

different location, the new scene must appear promptly. Obtaining such performance might be relatively straightforward if the terrain data were collocated with the rendering engine. However, one of the original premises underlying the MAGIC project is that the data set and the application are not collocated. There are several reasons for this, the most important being that the data set could be extremely large, so it might not be feasible to transfer it to the user's site. Moreover, experience has shown that in many cases the "owner" of a data set is also its "curator" and may be reluctant to distribute it, preferring instead to keep the data locally to simplify maintenance and updates. Finally, it was anticipated that future versions of the application might work with a mobile user and with fused data from multiple sources, and neither of these capabilities would be practical with local data. Therefore, since the data will not be local, the MAGIC components must be designed to compensate for possible delays and other degradations in the end-to-end operation of the system.

In order to understand system-level design issues, it is necessary to outline the sequence of events that occurs when the user moves the input device, causing a new scene to be generated. TerraVision first produces a list of new tiles required for the scene. This list is sent to an ISS master, which performs a name translation, mapping the logical address of each tile (the tile identifier) to its physical address (server/disk/location on disk). The master then sends each server an ordered list of the tiles it must retrieve. The server discards the previous list (even if it has not retrieved all the tiles on that list) and begins retrieving the tiles on the new list. Thus, the design for the system comprising TerraVision, the ISS, and the internetwork must address the following questions:

- How can TerraVision compensate for tiles it needs for the next image but have not yet been received?
- How often should TerraVision request tiles from the ISS?
- Where should the ISS master be located?
- How should tiles be distributed among the ISS disks?
- How can cell loss be minimized near the rendering site where the tile traffic becomes aggregated and congestion may occur?

Missing Tiles

Network congestion, an overload at an ISS server, or a component failure could result in the late arrival or loss of tiles that are requested by the application. Several mechanisms were implemented to deal with this problem. First, although the entire set of high-resolution tiles cannot be collocated with the application, it is certainly feasible to store a complete set of lower-resolution tiles. For example, if the entire data set comprises 1 Tbyte of high-resolution tiles, then all of the tiles that are five or more levels coarser would occupy less than 1.5 Mbyte, a readily affordable amount of local storage. If a tile with resolution at, say, level 3 is requested but not delivered in time for the image to be rendered, then, until the missing level-3 tile arrives, the locally available coarser tile from level 5 would be

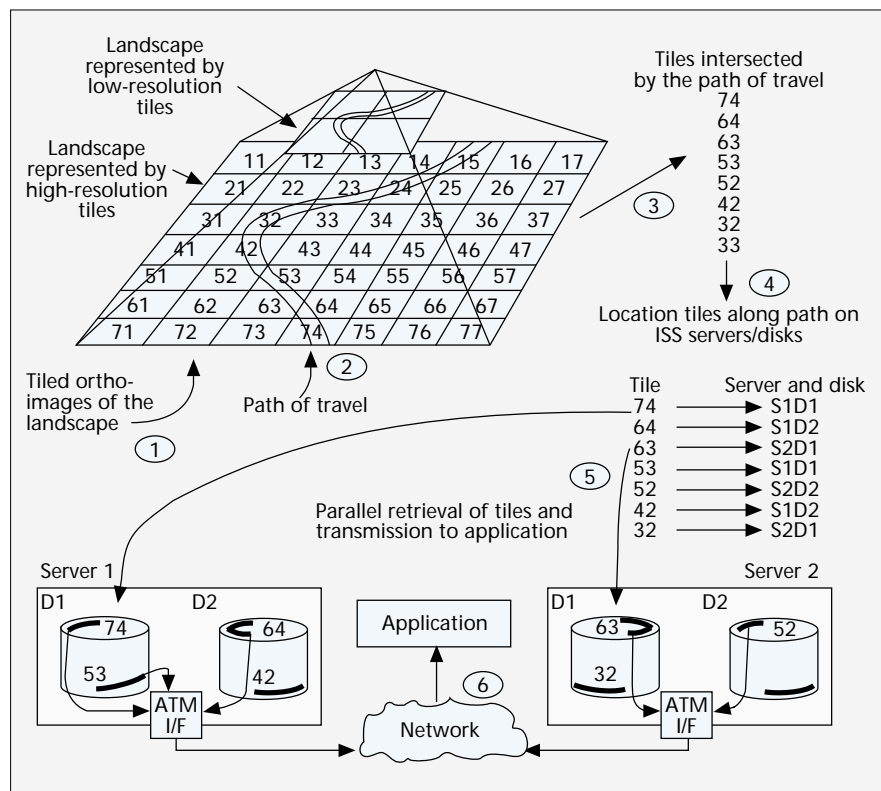


Figure 5. Schematic representation of the operation of the ISS. (Source: Lawrence Berkeley National Laboratory)

used in place of the 16 level-3 tiles. This substitution manifests itself by the affected portion of the rendered image appearing "fuzzy" for a brief period of time. Temporary substitution of low-resolution tiles for high-resolution tiles is particularly effective for teleporting because that operation requires a large number of new tiles, so it is more likely that one or more will be delayed.

Second, TerraVision attempts to predict the path the user will follow, requesting tiles that *might* soon be needed, and assigning one of three levels of priority to each tile requested. Priority-1 tiles are needed as soon as possible; the ISS retrieves and dispatches these first. This set of tiles is ordered by TerraVision, with the coarsest assigned the highest priority within the set. The reasons are:

- The rendering algorithm needs the coarse tiles before it needs the next-higher-resolution tiles.
- There are fewer tiles at the coarser resolutions, so it is less likely that they will be delayed.

The priority-2 tiles are those that the ISS should retrieve but should transmit only if there are no priority-1 tiles to be transmitted; that is, priority-2 tiles are put on a lower-priority transmit queue in the I/O buffer of each ISS server. (ATM switches would be allowed to drop the cells carrying these tiles.) Priority-3 tiles are those that should be retrieved and cached at the ISS server; these tiles are less likely to be needed by TerraVision. Note that there is a trade-off between "overpredicting" — requesting too many tiles — which would result in poor ISS performance and high network load, and "underpredicting," which would result in poor application performance.

Finally, a tile will continue to be included in TerraVision's request list if it is still needed and has not yet been delivered. Thus, tiles or tile requests that are dropped or otherwise "lost" in the network will likely be delivered in response to a subsequent request from the application.

Frequency of Requests

Another trade-off pertains to the frequency at which TerraVision sends its request list to the ISS. If the interval between requests is too large, then some tiles will not arrive when needed, resulting in a poor-quality display; in addition, the ISS will be idle and hence not used efficiently. On the other hand, if the interval is too short, then the request list might contain tiles that are currently in transit from servers to the application; this would result in poor ISS performance and redundant network traffic. For a typical MAGIC configuration, the interval between requests is currently set at 200 ms, a value that was found empirically to yield satisfactory performance. This value is based roughly on the measured latency of the ISS (about 100 ms) and on the estimated time required for a tile request to travel through the network from the TerraVision host to the ISS master and then to the most distant ISS server, plus the time for the tile itself to travel back to the host (perhaps a total of 50 ms). Additional measurements and analysis are needed to more precisely determine the appropriate request frequency as a function of the performance and location of system components and of network parameters.

Location of ISS Master

Since tile requests flow from TerraVision to the ISS master and thence to the servers themselves, the time for delivering the requests to the servers is minimized when the master is collocated with the TerraVision host. However, locating the master with the host is neither desirable nor practical for several reasons. The master is logically part of the ISS; therefore, its location should not be constrained by the application. Also, an ISS may be used with several applications concurrently, by multiple simultaneous users of a particular application, or by a user whose host may be unable to support any ISS functionality (e.g., a mobile user). Moreover, replication of the master would introduce problems associated with maintaining consistency among multiple masters when the ISS is in a read/write environment, as it would be when real-time data are being stored on the servers.

To first order, the delivery time of tile requests is limited by the time τ for a request to travel from TerraVision to the ISS server most distant from the TerraVision host. Hence, if the master is approximately on the path from the TerraVision host to that server, then τ will not be much greater than when the master and host are collocated. Furthermore, in the current MAGIC testbed, τ is much smaller than the sum of the disk latency and the network transit time. In other words, there is considerable freedom in choosing the location of the ISS master. Satisfactory system performance has been demonstrated, for example, with the TerraVision host in Kansas City, the ISS master in Sioux Falls, and servers in Minneapolis and Lawrence. Of course, this conclusion might change if faster servers reduce ISS latency considerably, or the geographic span of the network were substantially larger.

Distribution of Tiles on ISS Servers

The manner in which data are distributed among the servers determines the degree of parallelism and hence the

For a typical MAGIC configuration, the interval between requests is currently set at 200 ms, a value that was found empirically to yield satisfactory performance.

aggregate throughput which can be obtained from the ISS. The data placement strategy depends on the application and is a function of data type and access patterns. For example, the retrieval pattern for a database of video clips would be quite different from that for a database of images. A strategy was developed for a terrain visualization type of application that minimizes the retrieval time for a set of tiles: the tiles assigned to a given disk are as far apart as possible in the terrain in order to maximize parallelism by minimizing the probability that tiles on a request list are on the same disk; and on each disk, tiles that are near each other in the terrain are placed as close as possible to minimize retrieval time. Although this was shown to be an optimal strategy for terrain path-following as in TerraVision [9], it was subsequently shown that ISS performance with random placement of tiles was only slightly worse. This was partly because tile retrieval time is much less than the latency in the

ISS servers and network transit time, and is therefore not currently a significant factor in overall performance. Random placement is simpler to implement and is expected to be satisfactory for many other applications. However, as discussed for the location of the ISS master, this conclusion may have to be revisited if the performance or the geographic distribution of system components changes significantly.

Avoiding Cell Loss

When initially implemented, the MAGIC internetwork exhibited very low throughput in certain configurations. One cause of the low throughput was found to be mismatches between the burst rates of components in the communications path. Examples of such rate mismatches were:

- An OC-3 workstation interface transmitting cells at full rate across the network to a 100 Mb/s TAXI interface on another workstation
- Two or more OC-3 input ports at an ATM switch sending data to the same OC-3 output port

A mismatch, coupled with small buffers at the output ports of ATM switches, caused cells to be dropped, which in turn resulted in the retransmission of entire TCP packets, exacerbating the problem. In some cases the measured useful throughput was less than one percent of the capacity of the lower speed line.

Previously it was noted that in many cases a large MTU can increase throughput. However, once again there is a trade-off. As the MTU size is increased, the number of ATM cells needed to carry the MTU increases. The probability that one or more cells from the MTU will be dropped by the network therefore increases, which in turn increases the probability that the MTU will have to be retransmitted, thus possibly decreasing the effective throughput. Flow-control techniques together with large switch buffers and proper choice of protocol parameters did provide satisfactory performance. Nevertheless, the overriding conclusions are that the parameters of the entire end-to-end system, not just those of a single host or switch, must be tuned, each direction of the data path must be evaluated separately, and every component in each direction of the data path must be considered in the evaluation.

Performance

This section presents highlights of system-level performance issues and measurements of the MAGIC testbed. The input data rates that are needed to support the TerraVision application are calculated first, to provide the context for the subsequent discussion. Then the data rates that the network, the ISS, and the application host can actually support are described. Finally, a diagnostic tool that was developed to help analyze system performance is explained. More detailed information about all of the above topics can be found in [3, 5, 6, 10].

TerraVision is used in one of two modes, flyover or teleport, and the characteristics of the data flow for the two modes are quite different. Flyover requires a relatively steady flow over a relatively long period of time (many seconds), whereas teleport requires a large burst of data but occurs relatively infrequently. Quantitative requirements can be estimated as follows. A high-resolution full-screen display comprises about 100 tiles, each tile containing 128×128 pixels with 24 bits of color information, or approximately 0.4 Mb. If 10 new tiles are needed for a typical frame update during flyover, then at 30 frames/s, the average data rate is

$$(30 \text{ frames/s}) \times (10 \text{ tiles/frame}) \times (0.4 \text{ Mb/tile}) \\ \approx 120 \text{ Mb/s}$$

at the application level in the host. Protocol overhead might add approximately 15 percent to this value, resulting in a line rate of about 140 Mb/s. For a teleport, the burst rate is considerably higher because the entire screen must be repainted within, say, a quarter of a second after the user selects the new location. If the total latency between the instant the user enters the selection and the instant the first bit of the first tile arrives at the TerraVision host is 150 ms, then the full screen of data must be transferred in the remaining $(250 - 150) = 100$ ms, and the capacity needed to support the transfer is

$$(100 \text{ tiles}) \times (0.4 \text{ Mb/tile}) / (0.1 \text{ s}) \approx 400 \text{ Mb/s}$$

at the application level, or about 450 Mb/s on the transmission line.

The line capacity needed near the TerraVision host site can be determined from these required rates and from the end-to-end throughput that can be attained in the network. Measurements on the MAGIC network showed that if the MTU and the TCP window sizes were large enough, and if flow control were used, then end-to-end TCP rates corresponding to about 80 percent of the line rate could be sustained; this is about 120 Mb/s on an OC-3 line. Thus, a flyover would completely fill a single OC-3 line, so in practice two lines are needed to allow for possible degradations and for variations around the average rate derived above. Similarly, one OC-12 or four OC-3 lines are needed to support a low-response-time teleport. Lower line capacity on the path near the host would degrade the response time (although the degradation would be less than linear because of the additive factor of ISS latency). In summary, the equivalent of two OC-3 lines into the host should give satisfactory flyover performance and a teleport response time less than 0.5 s, but more capacity is needed to reduce the response time and to provide some cushion for contention near the host site.

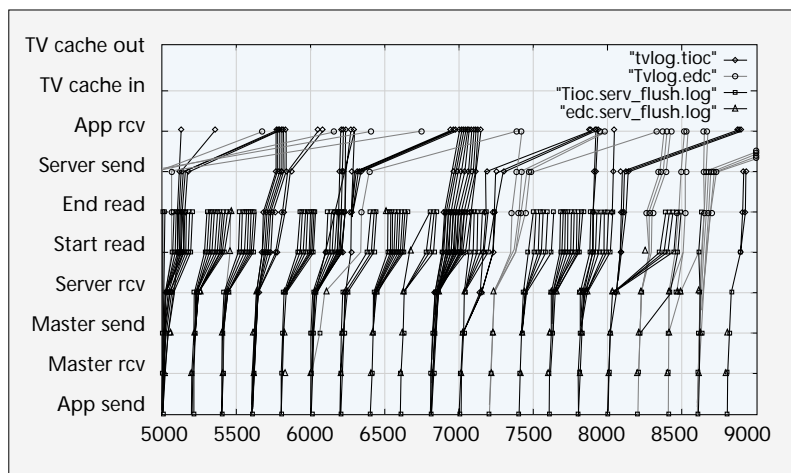


Figure 6. Timing data from a configuration with two ISS servers. (Source: Laurence Berkeley National Laboratory)

The next question is, “How many ISS servers are needed to support the application?” Early measurements of a variety of workstations configured as ISS servers showed that a typical SCSI disk delivered data at a steady rate of about 20 Mb/s; a single SCSI adapter with multiple disks could provide about 60 Mb/s; and a workstation with multiple adapters could deliver about 80 Mb/s. Additional disks or adapters did not increase throughput — the bottleneck apparently being memory bandwidth — but did increase the probability that the throughput could be sustained by ensuring that the server was not idle. These data indicate that tiles must be distributed over at least five servers to obtain the 400 Mb/s rate needed for good teleport performance.

The data streams from the ISS servers converge at the TerraVision host, and recent measurements showed that with four servers transmitting to a host with two OC-3 ATM ports, the aggregate application-level throughput was only about 100 Mb/s, and in fact was slightly less than the throughput with a single server. (The peak throughput was about 150 Mb/s, with two input streams.) Cells are apparently being dropped at the ATM interface. This is a serious bottleneck in overall system performance; the host and interface vendors are aware of the problem and are working on a solution.

Clearly, understanding the overall performance of a network-based distributed system such as MAGIC is an appreciably more complex undertaking than simply “concatenating” the standalone performance of the individual components because there are interactions among the components. It is important to be able to measure and correlate these interactions in order to understand and predict the performance of the system as a whole. Stated in concrete terms, a problem observed by a user could have a variety of causes. For example, in MAGIC it would be acceptable if low-resolution tiles are used occasionally in place of high-resolution tiles that are delayed or lost in transit (as described in the previous section), but it would be unacceptable if this occurred frequently. If such observable degradation did occur, the cause could be the application host dropping cells, ATM switches dropping cells, excessive delay somewhere in the ISS, low ISS throughput because of the way tiles are distributed among servers, processing limitations of the TerraVision host, or a combination of these and other phenomena.

To aid in pinpointing potential problems, accurately synchronized clocks were deployed at MAGIC sites, many components were instrumented to log traffic data, and a

tool was developed for collecting, processing, and displaying the logged data [3]. The tool's graphical portrayal of measured data gives a readily comprehensible view of the overall operation of the system, permits performance estimates to be calculated easily, and provides an indication of which components may be causing performance problems. This tool, which was developed toward the end of the MAGIC project, has proved to be extremely valuable in diagnosing problems and in providing insight into techniques for improving performance. The tool is applicable to many high-speed distributed systems. A brief description of its use is given below.

Figure 6 displays a representative sample of 4 s of data from a configuration with the application host in Kansas City ("tioc" in the legend), the ISS master in Sioux Falls ("edc"), and one server at each. (The host was not running TerraVision, but an application that emulates TerraVision by sending the identical tile request lists which were sent during a previously recorded TerraVision session.) The diagram traces a time history for each requested tile, showing:

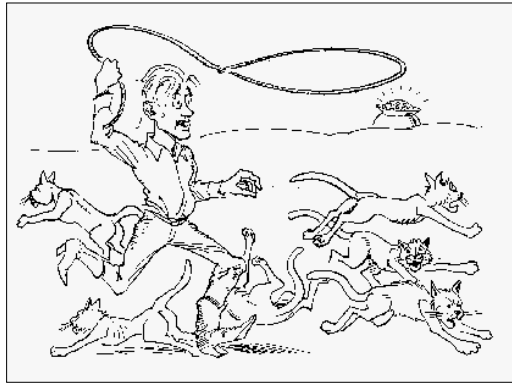
- When the application sends a request list (e.g., at 6800 ms) and when it is received by the ISS master (~ 6810)²
- When the master sends tile lists to the two servers (6820) and when they are received (~ 6840 and ~ 6850)
- When the servers start and complete their read operations
- When the tile data are transmitted by the servers and received by the host

In this example, the time between the request list leaving the application host and the first tile arriving at the host is 180 ms. The diagram shows that excluding the server time, the largest component of this 180 ms delay was queuing at the server, a result of TCP retransmissions of previously transmitted tiles that were, in effect, blocking transmission of subsequent tiles. (The shallow-sloped lines between "server send" and "app receive" represent tiles with TCP transmissions.) Rough calculations of throughput at each measurement point can readily be made by counting the number of tiles processed in a selected interval of time; for example, 15 tiles were received by the application between 6980 and 7130 ms, for a throughput of about 40 Mb/s.

Herding Cats

Although the MAGIC project was an ambitious undertaking, it nevertheless was able to achieve most of its goals. The success of the project seems all the more remarkable if one considers the degree of interorganizational collaboration that was required to design, develop, test, and integrate the individual testbed components and to ensure their interoperability. Indeed, fostering this collaboration was one of the most significant nontechnical challenges facing the project — and one of its noteworthy accomplishments.

More than a dozen diverse, geographically dispersed organizations participated in MAGIC, and many of the individuals involved in the project were experienced



■ Figure 7. *In the beginning, things looked difficult.*

researchers who were used to working independently. Although five of these organizations were funded by ARPA, each had its own contract and a statement of work that was complementary to the others but theoretically could be executed separately from the rest. In addition, the commercial carriers and other organizations that were expected to be major contributors were not externally funded and therefore were under no obligation to participate actively in the effort. Thus, the

situation at the outset was not unlike the metaphorical herding of cats (Fig. 7).

The authors of this article were funded by ARPA to oversee and coordinate the research and development (R & D) efforts of the five ARPA-funded research participants, and to help facilitate their collaboration with the carriers and with the other organizations contributing to the project. This was a challenging assignment because none of these organizations was contractually bound to answer to a third party, so voluntary compliance of all organizations was required. Considering the cast of players and the circumstances of their affiliation, it would have been imprudent to attempt to dictate direction or to impose preferences. Furthermore, to do so not only would have been ineffective but would have been counterproductive because a heavy-handed management style would have stifled the innovation that was critical to the success of the project. In other words, peremptory management might have led to passive obedience (Fig. 8), but the results would have been uninspired [11].

The challenge was to create an environment that facilitated progress and encouraged cooperation while at the same time promoting creativity and initiative. The approach used was to obtain mutual agreement on a common set of goals and related milestones which could not be achieved without the contributions of all of the participants. In this way, the focus of the work shifted from the pursuit of individual goals to the pursuit of common goals, and collaboration was implicitly understood to be essential for success. In retrospect, the reasons why this approach worked well seem obvious. Having a common set of goals engendered an esprit de corps among the participants which gave the sense of a "virtual" organization dedicated to the success of MAGIC.

However, participants soon recognized that while camaraderie and commitment were vital to success, team spirit alone was not sufficient to ensure that success. Differences in work styles, conflicting priorities, geographical dispersion of people and resources, and the sheer magnitude of the interdependencies underscored the need for centralized leadership and for "formal" procedures for coordinating activities. As a result, members of the MAGIC team willingly consented to, and complied with, a set of management practices that they perceived as facilitating the achievement of their technical objectives. The management style was collegial with the authors serving as facilitators for defining and prioritizing project activities, as mediators for resolving disputes, as liaison with the project sponsor (ARPA), and as catalysts for promoting the team interactions required to move forward. Thus, as indicated in Fig. 9, MAGIC took a hybrid approach to managing and coordinating its R & D, with

² These numerical values were obtained from a version of this diagram with an expanded timescale.

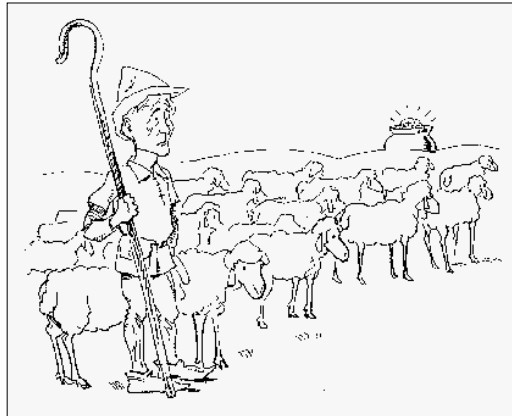
progress toward common goals achieved through high-level consensus-building among the participants.

Three practices stand out as being most critical to the success of the project: demonstrations, planning with flexibility, and ongoing communications.

Demonstrations — Although the components of the MAGIC testbed were designed to operate as parts of a system, they were developed independently by organizations that were not collocated. Therefore, interoperability testing and debugging were difficult. To deal with this problem, demonstrations for external observers were scheduled to mark the achievement of milestones. These demonstrations provided a strong incentive for overcoming the logistical obstacles to testing, and for uncovering and finding solutions to tough problems. At first glance, these events appeared to be distractions from the research and a drain on people and resources, and initially they were deemed antithetical to an R & D project. In actuality they were the single most important factor in accelerating progress. Often, it was in the typically frantic last hours before a crucial demonstration that creative solutions to unforeseen problems were conceived.

A number of major demonstrations were scheduled in conjunction with quarterly project meetings or technical symposia. The first, which took place approximately halfway into the three-year project duration, marked the completion of the first phase of the MAGIC testbed: initial versions of TerraVision and the ISS working together over a partially completed backbone. The second, which occurred about six months later, demonstrated improved versions of both TerraVision and the ISS working together over the full internetwork. This demonstration was attended by prospective end users of the system who provided valuable feedback, including suggestions for additional capabilities which were subsequently incorporated into TerraVision, substantially improving the utility of the application.

Planning ... with Flexibility — Researchers are notoriously reluctant to document their ideas and approaches in advance for fear of forfeiting their flexibility or limiting their options; however, failure to do so can spell disaster in a collaborative venture involving multiple organizations. Therefore, one of the first priorities of the MAGIC team was to develop a comprehensive research plan for the project. If truth be told, the process of planning was far more valuable than the plan itself. In creating the plan, each organization was forced to clearly define its tasks and milestones, to explore alternative approaches to accomplishing the work, and, most important, to identify interorganizational relationships and dependencies. It was understood that tasks and milestones, as well as technical approaches, would most likely change and evolve over the course of the three-year effort, and the plan was considered a working document to be revised and revisited as appropriate. However, at the conclusion of the project, it was gratifying to discover that the participants had accomplished most of the work they had intended to do within the allotted time and budget constraints.



■ Figure 8. *This wouldn't work either.*

Ongoing Communications — Ongoing communication was an important factor in maintaining cohesiveness among team members, and was essential for accomplishing the work. Regular interaction was achieved by holding weekly teleconferences and quarterly project meetings to discuss technical issues and interorganizational dependencies, to plan joint activities and events, and to identify and resolve problems. In addition, a variety of mechanisms for exchanging information were established, including

multiple mailers, and a project server for storing and retrieving documents such as project plans, papers, and reports. To facilitate collaboration on documents, a common desktop publishing package, which was available for multiple platforms, was adopted by the team very early in the project.

Lessons Learned

The previous section described the challenges of managing the MAGIC project, and discussed some of the factors that promoted cooperation and collaboration among the participants in this multidisciplinary, multi-organizational effort. Below are some additional lessons that were learned — sometimes with pain — during the course of the three-year project.

Technology for R&D Projects

R & D projects such as MAGIC depend on state-of-the-art technology to achieve their goals. There are two alternatives for obtaining this technology: develop it as part of the project, or procure it from vendors or other sources. Where possible, MAGIC opted for the latter alternative, and milestones were planned based on vendors' stated intentions regarding the capabilities of and projected delivery dates for critical hardware and software. As a consequence of this decision, MAGIC researchers learned two important lessons.

Be Prepared to Deal with the Limitations of Vendor Products — Some of the vendor-supplied state-of-the-art products required by MAGIC, for example, the SONET terminals and the ATM switches, were available on schedule and performed satisfactorily. Others, however, were either not available in the time frame expected (e.g., OC-12 cards for the ATM switches) or did not function as anticipated. Specifically, MAGIC researchers had to deal with three types of limitations:

- **Product (im)maturity:** Early production versions of products required a significant amount of tuning and debugging that would be unacceptable in a mature product. For example, some workstation operating systems initially had hard-coded upper limits on the TCP window size, limiting the achievable throughput across a network having a large bandwidth-delay product.
- **Standalone performance vs. system performance:** Products did not perform per their standalone specifications when incorporated into a system. For example, the measured rate of a disk on an ISS server was typically less than half the specified rate (perhaps caused by interactions with the SCSI adapter).
- **Single-component performance vs. multiple-component**

performance: When multiple components were made to operate in parallel, their performance did not scale linearly. For example, the rate at which the TerraVision host could absorb data increased only slightly as the number of ATM interfaces was increased.

Encourage the Active Involvement of Vendors in the R & D Effort — MAGIC depended on products that were under development or “on the horizon” when the project was initiated, and progress often hinged on timely access to early releases or upgrades. In some cases, market pressures on vendors took precedence over research needs, and the products were delayed, or anticipated features were postponed or eliminated. In other cases, products were released but were not robust, and vendor support was difficult to obtain. If equipment vendors had been more actively involved in the R & D effort, the other researchers, as well as ARPA and the carriers, would have been in a better position to influence vendor priorities and development schedules, and would have been more likely to gain the support and assistance they needed to correct shortcomings. Active vendor participation would have been beneficial to the vendors as well, providing them with insight into the strengths and limitations of their products, and helping them identify additional features and performance enhancements that might improve their competitive advantage.

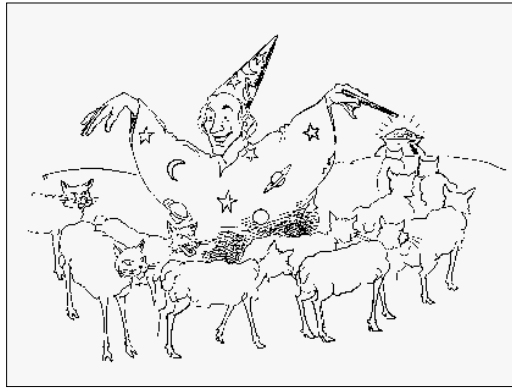
Despite the difficulties associated with relying on vendors for supporting technology, using vendor-supplied products was preferable to developing customized products as part of the project. Such development would have been time- and resource-intensive, and possibly a duplication of effort. In addition, customized technology is expensive to replicate and difficult to transfer to other domains.

As discussed previously, demonstrations were sometimes scheduled to coincide with major project events or milestones. In addition, requests to demonstrate the capabilities of the testbed were occasionally made by ARPA, by the management of the participating organizations, or by prospective end users. While there were significant benefits associated with holding these demonstrations, preparing for them was time-consuming because it was frequently necessary to reconfigure the network and to relocate and assemble the required hardware. The MAGIC team learned two lessons that helped facilitate the conduct of demonstrations during the later stages of the project.

Support for Demonstrations

As discussed previously, demonstrations were sometimes scheduled to coincide with major project events or milestones. In addition, requests to demonstrate the capabilities of the testbed were occasionally made by ARPA, by the management of the participating organizations, or by prospective end users. While there were significant benefits associated with holding these demonstrations, preparing for them was time-consuming because it was frequently necessary to reconfigure the network and to relocate and assemble the required hardware. The MAGIC team learned two lessons that helped facilitate the conduct of demonstrations during the later stages of the project.

Establish a Reliable Testbed Configuration to Support Demonstrations — Although demonstrations proved to be a significant stimulus to progress, they sometimes conflicted with planned experiments or with development and testing activities. This was particularly troubling when work-in-progress was interrupted or put on hold for a relatively long period of time in order to reconfigure the network (or to test modifications to TerraVision or the ISS) to support a scheduled event. This situation was remedied by implementing stable versions of TerraVision and the ISS and deploying them at selected locations. These versions were used to support demonstrations, performance measurements, and related activities. Updates to the demon-



■ Figure 9. *Heterogeneous collaborative interoperability.*

stration versions of TerraVision and the ISS were coordinated to ensure their compatibility.

Plan Equipment Logistics Carefully and in Advance — Another problem, pertinent to development as well as demonstrations, is the availability of equipment. Since budgets are finite, choices must be made regarding what equipment to purchase and where this equipment should be located. While it is impossible to foresee all contingencies, equipment needs should be determined as far in

advance as possible. Doing so will minimize the disruptions and stress associated with disassembling and transporting hardware over long distances, and acquiring essential components on short notice. It is especially important to develop strategies for supporting off-site demonstrations, particularly those that involve relocating large, cumbersome equipment or require expensive hardware which cannot easily be moved and is difficult to borrow or lease.

One way of helping to ensure that demonstrations can be accommodated without undue disruption is to purchase spares of inexpensive equipment. These spares would be available not only for demonstrations, but for development and experimentation in the event that an original malfunctioned. It is less feasible to duplicate expensive equipment; however, if vendors of critical components are actively involved in the project, they might be willing to support demonstrations by providing the necessary hardware.

Support for Development

The MAGIC testbed consists of components that were designed to interoperate but were developed independently by organizations that were geographically separated. In addition, the end users of the system were not research participants in the project. The following lessons were learned regarding how to work more effectively and efficiently under such conditions.

Build Tools to Enable Independent Development of Interoperable Components — Interoperability testing of a given component was challenging because it required that other components possess a level of functionality or performance that was not always available when the tests were ready to be conducted. One way to alleviate this problem was to implement component simulators that enabled interoperability testing. In MAGIC, the implementation of a TerraVision simulator hastened progress on the ISS, whereas the decision not to implement an ISS simulator increased the time needed to complete TerraVision.

Provide High-Speed Network Connections for All Major Participants — Proper testing of TerraVision and the ISS required high-speed interconnectivity. However, SRI and LBNL, the respective developers of these components, did not have such connectivity. As a result, interoperability testing could not be performed locally, and testing at remote sites was both burdensome and inefficient. In the MAGIC project, both of these organizations would have benefited from having high-speed links to the backbone.

Solicit Periodic Input from End Users — Getting input from end users helps to ensure that the final product has useful

features, satisfactory performance, and a well-designed interface. Input regarding desired capabilities should be solicited early in the effort and regularly thereafter as development progresses. Although, as noted in the previous section, MAGIC did benefit from such input, the project would have benefited even more if that input had been obtained earlier and more frequently.

Future Work

The MAGIC project has demonstrated a high-speed, wide-area IP/ATM internetwork that supports a real-time terrain visualization application and a distributed storage system. ARPA recently approved funding for a three-year follow-on effort, MAGIC-II, which will build on the technology developed in the original MAGIC project and on the existing MAGIC network facilities. There are two major inter-related goals in MAGIC-II:

- To enhance and upgrade the testbed to demonstrate the utility and capabilities of distributed processing and network-based storage, coupled with high-speed networks, to support a new generation of real-time applications.
- To create a very large internetwork with many end users that will be a realistic test environment for ATM technology and for the above type of application.

The MAGIC-II testbed will demonstrate the scalability of the distributed storage and distributed processing concepts by configuring systems that have a large number of servers and processors on many ATM networks spanning a large geographic area, and have multiple sets of data and multiple simultaneous users.

The MAGIC-II testbed is based on a very general paradigm in which high-performance computing, storage, and communications are used to provide rapid access to large amounts of distributed data, including real-time data that must be processed and delivered to an end user on demand. Applications that use this paradigm arise in a variety of situations, including military operations, intelligence imagery analysis, and natural disaster response. The exact type, location, and ownership of the data used by these applications may not be known in advance, and these data may require a large amount of processing to be transformed into useful information. In addition, the processed data may have to be delivered to end users with a range of communications speeds, link qualities, computational powers, and display capabilities. The data, as well as the computing and storage resources required to process them, may reside in multiple administrative domains that have different usage and access control policies.

Specific work to be done in MAGIC-II includes augmenting the MAGIC internetwork with wireless nodes and interconnecting it with other IP/ATM internetworks to create a nation-wide, high-speed, wide-area testbed. This testbed will be used for experimentation with protocols, with routing techniques, and with mobile access to backbone services. A new version of TerraVision that can perform on-the-fly rectification coupled with algorithms for

The MAGIC-II testbed will demonstrate the scalability of the distributed storage and distributed processing concepts by configuring systems that have a large number of servers and processors on many ATM networks spanning a large geographic area, and have multiple sets of data and multiple simultaneous users.

“in-transit” processing will permit near-real-time visualization of raw imagery, enabling data from sensors to be viewed within minutes (rather than hours) after being generated. (Fig. 1.) Data fusion techniques will allow disparate data types to be overlaid. The processing will be performed by sets of distributed devices that are constructed from resources owned by multiple administrative domains. Algorithms that dynamically determine the current state of the network will provide information to the application so that it can adapt to current system performance and to available system resources.

The MAGIC-II project will certainly benefit from the lessons learned in the original MAGIC project. Nevertheless, as with any research effort, new challenges will be encountered, and new lessons, both technical and organizational, will be learned in meeting these challenges. Stay tuned.

References

- [1] R. Binder, “Issues in Gigabit Networking,” *Proc. IEEE Globecom '92*, Orlando, FL, 12/92. Also see: <http://www.cnri.reston.va.us:4000/public/overview.html>
- [2] J. Evans *et al.*, “A 622 Mb/s LAN/WAN Gateway and Experiences with Wide Area ATM Networking,” *IEEE Network*, this issue.
- [3] B. Tierney *et al.*, “Performance Analysis in High-Speed Wide Area IP-over-ATM Networks: Top-to-Bottom End-to-End Monitoring,” *IEEE Network*, this issue.
- [4] M. Laubach, “Classical IP and ARP over ATM,” RFC 1577; 20, Jan. 1994.
- [5] B.J. Ewy *et al.*, “TCP/ATM Experiences in the MAGIC Testbed,” *Proc. 4th IEEE Symp. High Perf. Dist. Comp.*, Aug. 1995, pp. 87–93.
- [6] J. D. Cavanaugh, Minnesota Supercomputer Center, personal communication.
- [7] Y. G. Leclerc and S. Q. Lau Jr., “TerraVision: A Terrain Visualization System,” Tech. Note 540, SRI International, Menlo Park, CA, Apr. 1994.
- [8] B. Tierney *et al.*, “Distributed Parallel Data Storage Systems,” *Proc. ACM Multimedia '94*, Oct. 1994. Also available as <http://george.lbl.gov/ISS/papers/ISS-paper.ACM.final.html>
- [9] L. T. Chen and D. Rotem, “Declustering Objects for Visualization,” *Proc. 19th VLDB (Very Large Database) Conf.*, 1993.
- [10] B. Tierney *et al.*, “Using High Speed Networks to Enable Distributed Parallel Image Server Systems,” *Proc. Supercomputing '94*, Nov. 1994. Also available as: <http://george.lbl.gov/ISS/papers/ISS-paper.SC94.final.html>
- [11] S. Adams, “Dilbert,” *Boston Globe*, Jan. 14, 1996.

Biographies

BARBARA FULLER recently joined Mitretek Systems in McLean, Virginia, as lead staff in the Center for Information Technology Systems. Prior to joining Mitretek, she held a similar position with the MITRE Corporation in Bedford, Massachusetts, where she provided technical, systems engineering, and program planning, coordination, and analysis support to a variety of DoD-sponsored advanced information systems projects, including MAGIC. She also spent 12 years at MITRE managing multidisciplinary projects for U.S. Government agencies dealing with toxic chemicals in the environment. She received her B.A. degree in chemistry from Western Reserve University in Cleveland, Ohio, and her M.S. and Ph.D. degrees in chemistry from New York University.

IRA RICHER is director of networking research at the Corporation for National Research Initiatives (CNRI), Reston, Virginia. His current work includes coordinating the activities of about a dozen companies in a project involving trials of broadband services to residences, and managing the MAGIC project. Prior to joining CNRI, Dr. Richer was with MITRE, where he supervised a small group working on advanced networks and applications. From 1988 to 1991, Dr. Richer was program manager for high-performance networking at ARPA. He initiated ARPA's program in gigabit networks, and he launched ARPA's work in all-optical networking. Dr. Richer received a B.E.E. degree from Rensselaer Polytechnic Institute, and M.S. and Ph.D. degrees from the California Institute of Technology.



INTERNATIONAL TELECOMMUNICATION UNION

CCITT

THE INTERNATIONAL
TELEGRAPH AND TELEPHONE
CONSULTATIVE COMMITTEE

T.81

(09/92)

**TERMINAL EQUIPMENT AND PROTOCOLS
FOR TELEMATIC SERVICES**

**INFORMATION TECHNOLOGY –
DIGITAL COMPRESSION AND CODING
OF CONTINUOUS-TONE STILL IMAGES –
REQUIREMENTS AND GUIDELINES**



Recommendation T.81

Foreword

ITU (International Telecommunication Union) is the United Nations Specialized Agency in the field of telecommunications. The CCITT (the International Telegraph and Telephone Consultative Committee) is a permanent organ of the ITU. Some 166 member countries, 68 telecom operating entities, 163 scientific and industrial organizations and 39 international organizations participate in CCITT which is the body which sets world telecommunications standards (Recommendations).

The approval of Recommendations by the members of CCITT is covered by the procedure laid down in CCITT Resolution No. 2 (Melbourne, 1988). In addition, the Plenary Assembly of CCITT, which meets every four years, approves Recommendations submitted to it and establishes the study programme for the following period.

In some areas of information technology, which fall within CCITT's purview, the necessary standards are prepared on a collaborative basis with ISO and IEC. The text of CCITT Recommendation T.81 was approved on 18th September 1992. The identical text is also published as ISO/IEC International Standard 10918-1.

CCITT NOTE

In this Recommendation, the expression "Administration" is used for conciseness to indicate both a telecommunication administration and a recognized private operating agency.

© ITU 1993

All rights reserved. No part of this publication may be reproduced or utilized in any form or by any means, electronic or mechanical, including photocopying and microfilm, without permission in writing from the ITU.

Contents

	<i>Page</i>
Introduction.....	iii
1 Scope	1
2 Normative references.....	1
3 Definitions, abbreviations and symbols	1
4 General	12
5 Interchange format requirements	23
6 Encoder requirements	23
7 Decoder requirements	23
Annex A – Mathematical definitions.....	24
Annex B – Compressed data formats.....	31
Annex C – Huffman table specification.....	50
Annex D – Arithmetic coding	54
Annex E – Encoder and decoder control procedures.....	77
Annex F – Sequential DCT-based mode of operation.....	87
Annex G – Progressive DCT-based mode of operation.....	119
Annex H – Lossless mode of operation	132
Annex J – Hierarchical mode of operation.....	137
Annex K – Examples and guidelines.....	143
Annex L – Patents.....	179
Annex M – Bibliography.....	181

Introduction

This CCITT Recommendation | ISO/IEC International Standard was prepared by CCITT Study Group VIII and the Joint Photographic Experts Group (JPEG) of ISO/IEC JTC 1/SC 29/WG 10. This Experts Group was formed in 1986 to establish a standard for the sequential progressive encoding of continuous tone grayscale and colour images.

Digital Compression and Coding of Continuous-tone Still images, is published in two parts:

- Requirements and guidelines;
- Compliance testing.

This part, Part 1, sets out requirements and implementation guidelines for continuous-tone still image encoding and decoding processes, and for the coded representation of compressed image data for interchange between applications. These processes and representations are intended to be generic, that is, to be applicable to a broad range of applications for colour and grayscale still images within communications and computer systems. Part 2, sets out tests for determining whether implementations comply with the requirements for the various encoding and decoding processes specified in Part 1.

The user's attention is called to the possibility that – for some of the coding processes specified herein – compliance with this Recommendation | International Standard may require use of an invention covered by patent rights. See Annex L for further information.

The requirements which these processes must satisfy to be useful for specific image communications applications such as facsimile, Videotex and audiographic conferencing are defined in CCITT Recommendation T.80. The intent is that the generic processes of Recommendation T.80 will be incorporated into the various CCITT Recommendations for terminal equipment for these applications.

In addition to the applications addressed by the CCITT and ISO/IEC, the JPEG committee has developed a compression standard to meet the needs of other applications as well, including desktop publishing, graphic arts, medical imaging and scientific imaging.

Annexes A, B, C, D, E, F, G, H and J are normative, and thus form an integral part of this Specification. Annexes K, L and M are informative and thus do not form an integral part of this Specification.

This Specification aims to follow the guidelines of CCITT and ISO/IEC JTC 1 on *Rules for presentation of CCITT | ISO/IEC common text*.

INTERNATIONAL STANDARD

CCITT RECOMMENDATION

INFORMATION TECHNOLOGY – DIGITAL COMPRESSION AND CODING OF CONTINUOUS-TONE STILL IMAGES – REQUIREMENTS AND GUIDELINES

1 Scope

This CCITT Recommendation | International Standard is applicable to continuous-tone – grayscale or colour – digital still image data. It is applicable to a wide range of applications which require use of compressed images. It is not applicable to bi-level image data.

This Specification

- specifies processes for converting source image data to compressed image data;
- specifies processes for converting compressed image data to reconstructed image data;
- gives guidance on how to implement these processes in practice;
- specifies coded representations for compressed image data.

NOTE – This Specification does not specify a complete coded image representation. Such representations may include certain parameters, such as aspect ratio, component sample registration, and colour space designation, which are application-dependent.

2 Normative references

The following CCITT Recommendations and International Standards contain provisions which, through reference in this text, constitute provisions of this CCITT Recommendation | International Standard. At the time of publication, the editions indicated were valid. All Recommendations and Standards are subject to revision, and parties to agreements based on this CCITT Recommendation | International Standard are encouraged to investigate the possibility of applying the most recent edition of the Recommendations and Standards listed below. Members of IEC and ISO maintain registers of currently valid International Standards. The CCITT Secretariat maintains a list of currently valid CCITT Recommendations.

- *CCITT Recommendation T.80 (1992), Common components for image compression and communication – Basic principles.*

3 Definitions, abbreviations and symbols

3.1 Definitions and abbreviations

For the purposes of this Specification, the following definitions apply.

3.1.1 abbreviated format: A representation of compressed image data which is missing some or all of the table specifications required for decoding, or a representation of table-specification data without frame headers, scan headers, and entropy-coded segments.

3.1.2 AC coefficient: Any DCT coefficient for which the frequency is not zero in at least one dimension.

3.1.3 (adaptive) (binary) arithmetic decoding: An entropy decoding procedure which recovers the sequence of symbols from the sequence of bits produced by the arithmetic encoder.

3.1.4 (adaptive) (binary) arithmetic encoding: An entropy encoding procedure which codes by means of a recursive subdivision of the probability of the sequence of symbols coded up to that point.

3.1.5 application environment: The standards for data representation, communication, or storage which have been established for a particular application.

- 3.1.6 arithmetic decoder:** An embodiment of arithmetic decoding procedure.
- 3.1.7 arithmetic encoder:** An embodiment of arithmetic encoding procedure.
- 3.1.8 baseline (sequential):** A particular sequential DCT-based encoding and decoding process specified in this Specification, and which is required for all DCT-based decoding processes.
- 3.1.9 binary decision:** Choice between two alternatives.
- 3.1.10 bit stream:** Partially encoded or decoded sequence of bits comprising an entropy-coded segment.
- 3.1.11 block:** An 8×8 array of samples or an 8×8 array of DCT coefficient values of one component.
- 3.1.12 block-row:** A sequence of eight contiguous component lines which are partitioned into 8×8 blocks.
- 3.1.13 byte:** A group of 8 bits.
- 3.1.14 byte stuffing:** A procedure in which either the Huffman coder or the arithmetic coder inserts a zero byte into the entropy-coded segment following the generation of an encoded hexadecimal X'FF' byte.
- 3.1.15 carry bit:** A bit in the arithmetic encoder code register which is set if a carry-over in the code register overflows the eight bits reserved for the output byte.
- 3.1.16 ceiling function:** The mathematical procedure in which the greatest integer value of a real number is obtained by selecting the smallest integer value which is greater than or equal to the real number.
- 3.1.17 class (of coding process):** Lossy or lossless coding processes.
- 3.1.18 code register:** The arithmetic encoder register containing the least significant bits of the partially completed entropy-coded segment. Alternatively, the arithmetic decoder register containing the most significant bits of a partially decoded entropy-coded segment.
- 3.1.19 coder:** An embodiment of a coding process.
- 3.1.20 coding:** Encoding or decoding.
- 3.1.21 coding model:** A procedure used to convert input data into symbols to be coded.
- 3.1.22 (coding) process:** A general term for referring to an encoding process, a decoding process, or both.
- 3.1.23 colour image:** A continuous-tone image that has more than one component.
- 3.1.24 columns:** Samples per line in a component.
- 3.1.25 component:** One of the two-dimensional arrays which comprise an image.
- 3.1.26 compressed data:** Either compressed image data or table specification data or both.
- 3.1.27 compressed image data:** A coded representation of an image, as specified in this Specification.
- 3.1.28 compression:** Reduction in the number of bits used to represent source image data.
- 3.1.29 conditional exchange:** The interchange of MPS and LPS probability intervals whenever the size of the LPS interval is greater than the size of the MPS interval (in arithmetic coding).
- 3.1.30 (conditional) probability estimate:** The probability value assigned to the LPS by the probability estimation state machine (in arithmetic coding).
- 3.1.31 conditioning table:** The set of parameters which select one of the defined relationships between prior coding decisions and the conditional probability estimates used in arithmetic coding.
- 3.1.32 context:** The set of previously coded binary decisions which is used to create the index to the probability estimation state machine (in arithmetic coding).
- 3.1.33 continuous-tone image:** An image whose components have more than one bit per sample.
- 3.1.34 data unit:** An 8×8 block of samples of one component in DCT-based processes; a sample in lossless processes.

- 3.1.35 DC coefficient:** The DCT coefficient for which the frequency is zero in both dimensions.
- 3.1.36 DC prediction:** The procedure used by DCT-based encoders whereby the quantized DC coefficient from the previously encoded 8×8 block of the same component is subtracted from the current quantized DC coefficient.
- 3.1.37 (DCT) coefficient:** The amplitude of a specific cosine basis function – may refer to an original DCT coefficient, to a quantized DCT coefficient, or to a dequantized DCT coefficient.
- 3.1.38 decoder:** An embodiment of a decoding process.
- 3.1.39 decoding process:** A process which takes as its input compressed image data and outputs a continuous-tone image.
- 3.1.40 default conditioning:** The values defined for the arithmetic coding conditioning tables at the beginning of coding of an image.
- 3.1.41 dequantization:** The inverse procedure to quantization by which the decoder recovers a representation of the DCT coefficients.
- 3.1.42 differential component:** The difference between an input component derived from the source image and the corresponding reference component derived from the preceding frame for that component (in hierarchical mode coding).
- 3.1.43 differential frame:** A frame in a hierarchical process in which differential components are either encoded or decoded.
- 3.1.44 (digital) reconstructed image (data):** A continuous-tone image which is the output of any decoder defined in this Specification.
- 3.1.45 (digital) source image (data):** A continuous-tone image used as input to any encoder defined in this Specification.
- 3.1.46 (digital) (still) image:** A set of two-dimensional arrays of integer data.
- 3.1.47 discrete cosine transform; DCT:** Either the forward discrete cosine transform or the inverse discrete cosine transform.
- 3.1.48 downsampling (filter):** A procedure by which the spatial resolution of an image is reduced (in hierarchical mode coding).
- 3.1.49 encoder:** An embodiment of an encoding process.
- 3.1.50 encoding process:** A process which takes as its input a continuous-tone image and outputs compressed image data.
- 3.1.51 entropy-coded (data) segment:** An independently decodable sequence of entropy encoded bytes of compressed image data.
- 3.1.52 (entropy-coded segment) pointer:** The variable which points to the most recently placed (or fetched) byte in the entropy encoded segment.
- 3.1.53 entropy decoder:** An embodiment of an entropy decoding procedure.
- 3.1.54 entropy decoding:** A lossless procedure which recovers the sequence of symbols from the sequence of bits produced by the entropy encoder.
- 3.1.55 entropy encoder:** An embodiment of an entropy encoding procedure.
- 3.1.56 entropy encoding:** A lossless procedure which converts a sequence of input symbols into a sequence of bits such that the average number of bits per symbol approaches the entropy of the input symbols.
- 3.1.57 extended (DCT-based) process:** A descriptive term for DCT-based encoding and decoding processes in which additional capabilities are added to the baseline sequential process.
- 3.1.58 forward discrete cosine transform; FDCT:** A mathematical transformation using cosine basis functions which converts a block of samples into a corresponding block of original DCT coefficients.

- 3.1.59 frame:** A group of one or more scans (all using the same DCT-based or lossless process) through the data of one or more of the components in an image.
- 3.1.60 frame header:** A marker segment that contains a start-of-frame marker and associated frame parameters that are coded at the beginning of a frame.
- 3.1.61 frequency:** A two-dimensional index into the two-dimensional array of DCT coefficients.
- 3.1.62 (frequency) band:** A contiguous group of coefficients from the zig-zag sequence (in progressive mode coding).
- 3.1.63 full progression:** A process which uses both spectral selection and successive approximation (in progressive mode coding).
- 3.1.64 grayscale image:** A continuous-tone image that has only one component.
- 3.1.65 hierarchical:** A mode of operation for coding an image in which the first frame for a given component is followed by frames which code the differences between the source data and the reconstructed data from the previous frame for that component. Resolution changes are allowed between frames.
- 3.1.66 hierarchical decoder:** A sequence of decoder processes in which the first frame for each component is followed by frames which decode an array of differences for each component and adds it to the reconstructed data from the preceding frame for that component.
- 3.1.67 hierarchical encoder:** The mode of operation in which the first frame for each component is followed by frames which encode the array of differences between the source data and the reconstructed data from the preceding frame for that component.
- 3.1.68 horizontal sampling factor:** The relative number of horizontal data units of a particular component with respect to the number of horizontal data units in the other components.
- 3.1.69 Huffman decoder:** An embodiment of a Huffman decoding procedure.
- 3.1.70 Huffman decoding:** An entropy decoding procedure which recovers the symbol from each variable length code produced by the Huffman encoder.
- 3.1.71 Huffman encoder:** An embodiment of a Huffman encoding procedure.
- 3.1.72 Huffman encoding:** An entropy encoding procedure which assigns a variable length code to each input symbol.
- 3.1.73 Huffman table:** The set of variable length codes required in a Huffman encoder and Huffman decoder.
- 3.1.74 image data:** Either source image data or reconstructed image data.
- 3.1.75 interchange format:** The representation of compressed image data for exchange between application environments.
- 3.1.76 interleaved:** The descriptive term applied to the repetitive multiplexing of small groups of data units from each component in a scan in a specific order.
- 3.1.77 inverse discrete cosine transform; IDCT:** A mathematical transformation using cosine basis functions which converts a block of dequantized DCT coefficients into a corresponding block of samples.
- 3.1.78 Joint Photographic Experts Group; JPEG:** The informal name of the committee which created this Specification. The "joint" comes from the CCITT and ISO/IEC collaboration.
- 3.1.79 latent output:** Output of the arithmetic encoder which is held, pending resolution of carry-over (in arithmetic coding).
- 3.1.80 less probable symbol; LPS:** For a binary decision, the decision value which has the smaller probability.
- 3.1.81 level shift:** A procedure used by DCT-based encoders and decoders whereby each input sample is either converted from an unsigned representation to a two's complement representation or from a two's complement representation to an unsigned representation.

- 3.1.82 lossless:** A descriptive term for encoding and decoding processes and procedures in which the output of the decoding procedure(s) is identical to the input to the encoding procedure(s).
- 3.1.83 lossless coding:** The mode of operation which refers to any one of the coding processes defined in this Specification in which all of the procedures are lossless (see Annex H).
- 3.1.84 lossy:** A descriptive term for encoding and decoding processes which are not lossless.
- 3.1.85 marker:** A two-byte code in which the first byte is hexadecimal FF (X'FF') and the second byte is a value between 1 and hexadecimal FE (X'FE').
- 3.1.86 marker segment:** A marker and associated set of parameters.
- 3.1.87 MCU-row:** The smallest sequence of MCU which contains at least one line of samples or one block-row from every component in the scan.
- 3.1.88 minimum coded unit; MCU:** The smallest group of data units that is coded.
- 3.1.89 modes (of operation):** The four main categories of image coding processes defined in this Specification.
- 3.1.90 more probable symbol; MPS:** For a binary decision, the decision value which has the larger probability.
- 3.1.91 non-differential frame:** The first frame for any components in a hierarchical encoder or decoder. The components are encoded or decoded without subtraction from reference components. The term refers also to any frame in modes other than the hierarchical mode.
- 3.1.92 non-interleaved:** The descriptive term applied to the data unit processing sequence when the scan has only one component.
- 3.1.93 parameters:** Fixed length integers 4, 8 or 16 bits in length, used in the compressed data formats.
- 3.1.94 point transform:** Scaling of a sample or DCT coefficient.
- 3.1.95 precision:** Number of bits allocated to a particular sample or DCT coefficient.
- 3.1.96 predictor:** A linear combination of previously reconstructed values (in lossless mode coding).
- 3.1.97 probability estimation state machine:** An interlinked table of probability values and indices which is used to estimate the probability of the LPS (in arithmetic coding).
- 3.1.98 probability interval:** The probability of a particular sequence of binary decisions within the ordered set of all possible sequences (in arithmetic coding).
- 3.1.99 (probability) sub-interval:** A portion of a probability interval allocated to either of the two possible binary decision values (in arithmetic coding).
- 3.1.100 procedure:** A set of steps which accomplishes one of the tasks which comprise an encoding or decoding process.
- 3.1.101 process:** See coding process.
- 3.1.102 progressive (coding):** One of the DCT-based processes defined in this Specification in which each scan typically improves the quality of the reconstructed image.
- 3.1.103 progressive DCT-based:** The mode of operation which refers to any one of the processes defined in Annex G.
- 3.1.104 quantization table:** The set of 64 quantization values used to quantize the DCT coefficients.
- 3.1.105 quantization value:** An integer value used in the quantization procedure.
- 3.1.106 quantize:** The act of performing the quantization procedure for a DCT coefficient.
- 3.1.107 reference (reconstructed) component:** Reconstructed component data which is used in a subsequent frame of a hierarchical encoder or decoder process (in hierarchical mode coding).

- 3.1.108 renormalization:** The doubling of the probability interval and the code register value until the probability interval exceeds a fixed minimum value (in arithmetic coding).
- 3.1.109 restart interval:** The integer number of MCUs processed as an independent sequence within a scan.
- 3.1.110 restart marker:** The marker that separates two restart intervals in a scan.
- 3.1.111 run (length):** Number of consecutive symbols of the same value.
- 3.1.112 sample:** One element in the two-dimensional array which comprises a component.
- 3.1.113 sample-interleaved:** The descriptive term applied to the repetitive multiplexing of small groups of samples from each component in a scan in a specific order.
- 3.1.114 scan:** A single pass through the data for one or more of the components in an image.
- 3.1.115 scan header:** A marker segment that contains a start-of-scan marker and associated scan parameters that are coded at the beginning of a scan.
- 3.1.116 sequential (coding):** One of the lossless or DCT-based coding processes defined in this Specification in which each component of the image is encoded within a single scan.
- 3.1.117 sequential DCT-based:** The mode of operation which refers to any one of the processes defined in Annex F.
- 3.1.118 spectral selection:** A progressive coding process in which the zig-zag sequence is divided into bands of one or more contiguous coefficients, and each band is coded in one scan.
- 3.1.119 stack counter:** The count of X'FF' bytes which are held, pending resolution of carry-over in the arithmetic encoder.
- 3.1.120 statistical conditioning:** The selection, based on prior coding decisions, of one estimate out of a set of conditional probability estimates (in arithmetic coding).
- 3.1.121 statistical model:** The assignment of a particular conditional probability estimate to each of the binary arithmetic coding decisions.
- 3.1.122 statistics area:** The array of statistics bins required for a coding process which uses arithmetic coding.
- 3.1.123 statistics bin:** The storage location where an index is stored which identifies the value of the conditional probability estimate used for a particular arithmetic coding binary decision.
- 3.1.124 successive approximation:** A progressive coding process in which the coefficients are coded with reduced precision in the first scan, and precision is increased by one bit with each succeeding scan.
- 3.1.125 table specification data:** The coded representation from which the tables used in the encoder and decoder are generated and their destinations specified.
- 3.1.126 transcoder:** A procedure for converting compressed image data of one encoder process to compressed image data of another encoder process.
- 3.1.127 (uniform) quantization:** The procedure by which DCT coefficients are linearly scaled in order to achieve compression.
- 3.1.128 upsampling (filter):** A procedure by which the spatial resolution of an image is increased (in hierarchical mode coding).
- 3.1.129 vertical sampling factor:** The relative number of vertical data units of a particular component with respect to the number of vertical data units in the other components in the frame.
- 3.1.130 zero byte:** The X'00' byte.
- 3.1.131 zig-zag sequence:** A specific sequential ordering of the DCT coefficients from (approximately) lowest spatial frequency to highest.
- 3.1.132 3-sample predictor:** A linear combination of the three nearest neighbor reconstructed samples to the left and above (in lossless mode coding).

3.2 Symbols

The symbols used in this Specification are listed below.

A	probability interval
AC	AC DCT coefficient
AC _{ji}	AC coefficient predicted from DC values
A _h	successive approximation bit position, high
A _l	successive approximation bit position, low
A _{pi}	<i>i</i> th 8-bit parameter in APP _n segment
APP _n	marker reserved for application segments
B	current byte in compressed data
B2	next byte in compressed data when B = X'FF'
BE	counter for buffered correction bits for Huffman coding in the successive approximation process
BITS	16-byte list containing number of Huffman codes of each length
BP	pointer to compressed data
BPST	pointer to byte before start of entropy-coded segment
BR	counter for buffered correction bits for Huffman coding in the successive approximation process
B _x	byte modified by a carry-over
C	value of bit stream in code register
C _i	component identifier for frame
C _u	horizontal frequency dependent scaling factor in DCT
C _v	vertical frequency dependent scaling factor in DCT
CE	conditional exchange
C-low	low order 16 bits of the arithmetic decoder code register
C _{mi}	<i>i</i> th 8-bit parameter in COM segment
CNT	bit counter in NEXTBYTE procedure
CODE	Huffman code value
CODESIZE(V)	code size for symbol V
COM	comment marker
C _s	conditioning table value
C _{si}	component identifier for scan
CT	renormalization shift counter
C _x	high order 16 bits of arithmetic decoder code register
CX	conditional exchange
d _{ji}	data unit from horizontal position <i>i</i> , vertical position <i>j</i>
d _{ji} ^k	d _{ji} for component <i>k</i>
D	decision decoded

Da	in DC coding, the DC difference coded for the previous block from the same component; in lossless coding, the difference coded for the sample immediately to the left
DAC	define-arithmetic-coding-conditioning marker
Db	the difference coded for the sample immediately above
DC	DC DCT coefficient
DC _i	DC coefficient for <i>i</i> th block in component
DC _k	<i>k</i> th DC value used in prediction of AC coefficients
DHP	define hierarchical progression marker
DHT	define-Huffman-tables marker
DIFF	difference between quantized DC and prediction
DNL	define-number-of-lines marker
DQT	define-quantization-tables marker
DRI	define restart interval marker
E	exponent in magnitude category upper bound
EC	event counter
ECS	entropy-coded segment
ECS _i	<i>i</i> th entropy-coded segment
Eh	horizontal expansion parameter in EXP segment
EHUFCO	Huffman code table for encoder
EHUFSI	encoder table of Huffman code sizes
EOB	end-of-block for sequential; end-of-band for progressive
EOB _n	run length category for EOB runs
EOB _x	position of EOB in previous successive approximation scan
EOB ₀ , EOB ₁ , ..., EOB ₁₄	run length categories for EOB runs
EOI	end-of-image marker
Ev	vertical expansion parameter in EXP segment
EXP	expand reference components marker
FREQ(V)	frequency of occurrence of symbol V
H _i	horizontal sampling factor for <i>i</i> th component
H _{max}	largest horizontal sampling factor
HUFFCODE	list of Huffman codes corresponding to lengths in HUFFSIZE
HUFFSIZE	list of code lengths
HUFFVAL	list of values assigned to each Huffman code
<i>i</i>	subscript index
<i>I</i>	integer variable
Index(<i>S</i>)	index to probability estimation state machine table for context index <i>S</i>
<i>j</i>	subscript index
<i>J</i>	integer variable

JPG	marker reserved for JPEG extensions
JPG _n	marker reserved for JPEG extensions
k	subscript index
K	integer variable
Kmin	index of 1st AC coefficient in band (1 for sequential DCT)
Kx	conditioning parameter for AC arithmetic coding model
L	DC and lossless coding conditioning lower bound parameter
L _i	element in BITS list in DHT segment
L _i (t)	element in BITS list in the DHT segment for Huffman table t
La	length of parameters in APP _n segment
LASTK	largest value of K
Lc	length of parameters in COM segment
Ld	length of parameters in DNL segment
Le	length of parameters in EXP segment
Lf	length of frame header parameters
Lh	length of parameters in DHT segment
Lp	length of parameters in DAC segment
LPS	less probable symbol (in arithmetic coding)
Lq	length of parameters in DQT segment
Lr	length of parameters in DRI segment
Ls	length of scan header parameters
LSB	least significant bit
m	modulo 8 counter for RST _m marker
m _t	number of V _{i,j} parameters for Huffman table t
M	bit mask used in coding magnitude of V
M _n	<i>n</i> th statistics bin for coding magnitude bit pattern category
MAXCODE	table with maximum value of Huffman code for each code length
MCU	minimum coded unit
MCU _i	<i>i</i> th MCU
MCUR	number of MCU required to make up one MCU-row
MINCODE	table with minimum value of Huffman code for each code length
MPS	more probable symbol (in arithmetic coding)
MPS(S)	more probable symbol for context-index S
MSB	most significant bit
M2, M3, M4, ... , M15	designation of context-indices for coding of magnitude bits in the arithmetic coding models
n	integer variable
N	data unit counter for MCU coding
N/A	not applicable

Nb	number of data units in MCU
Next_Index_LPS	new value of Index(S) after a LPS renormalization
Next_Index_MPS	new value of Index(S) after a MPS renormalization
Nf	number of components in frame
NL	number of lines defined in DNL segment
Ns	number of components in scan
OTHERS(V)	index to next symbol in chain
P	sample precision
Pq	quantizer precision parameter in DQT segment
Pq(t)	quantizer precision parameter in DQT segment for quantization table t
PRED	quantized DC coefficient from the most recently coded block of the component
Pt	point transform parameter
Px	calculated value of sample
Q _{ji}	quantizer value for coefficient AC _{ji}
Q _{vu}	quantization value for DCT coefficient S _{vu}
Q ₀₀	quantizer value for DC coefficient
QAC _{ji}	quantized AC coefficient predicted from DC values
QDC _k	kth quantized DC value used in prediction of AC coefficients
Qe	LPS probability estimate
Qe(S)	LPS probability estimate for context index S
Qk	kth element of 64 quantization elements in DQT segment
r _{vu}	reconstructed image sample
R	length of run of zero amplitude AC coefficients
R _{vu}	dequantized DCT coefficient
Ra	reconstructed sample value
Rb	reconstructed sample value
Rc	reconstructed sample value
Rd	rounding in prediction calculation
RES	reserved markers
Ri	restart interval in DRI segment
RRRR	4-bit value of run length of zero AC coefficients
RS	composite value used in Huffman coding of AC coefficients
RST _m	restart marker number m
s _{yx}	reconstructed value from IDCT
S	context index
S _{vu}	DCT coefficient at horizontal frequency u, vertical frequency v

SC	context-index for coding of correction bit in successive approximation coding
Se	end of spectral selection band in zig-zag sequence
SE	context-index for coding of end-of-block or end-of-band
SI	Huffman code size
SIGN	1 if decoded sense of sign is negative and 0 if decoded sense of sign is positive
SIZE	length of a Huffman code
SLL	shift left logical operation
SLL α β	logical shift left of α by β bits
SN	context-index for coding of first magnitude category when V is negative
SOF ₀	baseline DCT process frame marker
SOF ₁	extended sequential DCT frame marker, Huffman coding
SOF ₂	progressive DCT frame marker, Huffman coding
SOF ₃	lossless process frame marker, Huffman coding
SOF ₅	differential sequential DCT frame marker, Huffman coding
SOF ₆	differential progressive DCT frame marker, Huffman coding
SOF ₇	differential lossless process frame marker, Huffman coding
SOF ₉	sequential DCT frame marker, arithmetic coding
SOF ₁₀	progressive DCT frame marker, arithmetic coding
SOF ₁₁	lossless process frame marker, arithmetic coding
SOF ₁₃	differential sequential DCT frame marker, arithmetic coding
SOF ₁₄	differential progressive DCT frame marker, arithmetic coding
SOF ₁₅	differential lossless process frame marker, arithmetic coding
SOI	start-of-image marker
SOS	start-of-scan marker
SP	context-index for coding of first magnitude category when V is positive
S _{qu}	quantized DCT coefficient
SRL	shift right logical operation
SRL α β	logical shift right of α by β bits
Ss	start of spectral selection band in zig-zag sequence
SS	context-index for coding of sign decision
SSSS	4-bit size category of DC difference or AC coefficient amplitude
ST	stack counter
Switch_MPS	parameter controlling inversion of sense of MPS
Sz	parameter used in coding magnitude of V
S0	context-index for coding of V = 0 decision
t	summation index for parameter limits computation
T	temporary variable

Ta_j	AC entropy table destination selector for j th component in scan
Tb	arithmetic conditioning table destination identifier
Tc	Huffman coding or arithmetic coding table class
Td_j	DC entropy table destination selector for j th component in scan
TEM	temporary marker
Th	Huffman table destination identifier in DHT segment
Tq	quantization table destination identifier in DQT segment
Tq_i	quantization table destination selector for i th component in frame
U	DC and lossless coding conditioning upper bound parameter
V	symbol or value being either encoded or decoded
V_i	vertical sampling factor for i th component
$V_{i,j}$	j th value for length i in HUFFVAL
V_{max}	largest vertical sampling factor
V_t	temporary variable
VALPTR	list of indices for first value in HUFFVAL for each code length
V1	symbol value
V2	symbol value
x_i	number of columns in i th component
X	number of samples per line in component with largest horizontal dimension
X_i	i th statistics bin for coding magnitude category decision
X1, X2, X3, ..., X15	designation of context-indices for coding of magnitude categories in the arithmetic coding models
XHUFCO	extended Huffman code table
XHUFSI	table of sizes of extended Huffman codes
X'values'	values within the quotes are hexadecimal
y_i	number of lines in i th component
Y	number of lines in component with largest vertical dimension
ZRL	value in HUFFVAL assigned to run of 16 zero coefficients
ZZ(K)	K th element in zig-zag sequence of quantized DCT coefficients
ZZ(0)	quantized DC coefficient in zig-zag sequence order

4 General

The purpose of this clause is to give an informative overview of the elements specified in this Specification. Another purpose is to introduce many of the terms which are defined in clause 3. These terms are printed in *italics* upon first usage in this clause.

4.1 Elements specified in this Specification

There are three elements specified in this Specification:

- a) An *encoder* is an embodiment of an *encoding process*. As shown in Figure 1, an encoder takes as input *digital source image data* and *table specifications*, and by means of a specified set of *procedures* generates as output *compressed image data*.
- b) A *decoder* is an embodiment of a *decoding process*. As shown in Figure 2, a decoder takes as input *compressed image data* and *table specifications*, and by means of a specified set of *procedures* generates as output *digital reconstructed image data*.
- c) The *interchange format*, shown in Figure 3, is a compressed image data representation which includes all table specifications used in the encoding process. The interchange format is for exchange between *application environments*.

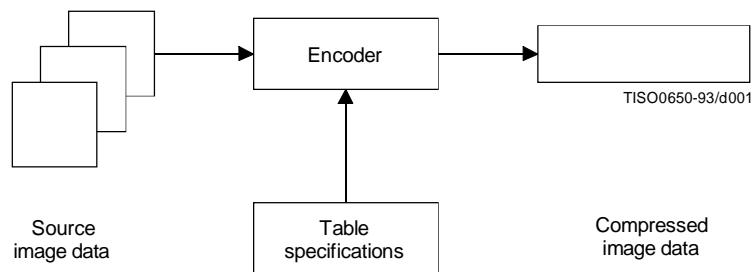


Figure 1 – Encoder

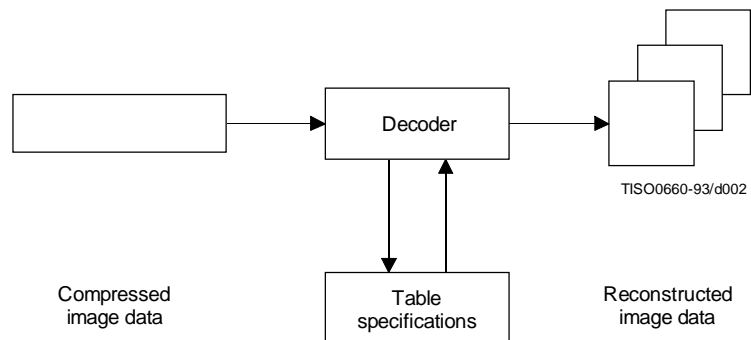


Figure 2 – Decoder

Figures 1 and 2 illustrate the general case for which the *continuous-tone* source and reconstructed image data consist of multiple *components*. (A *colour* image consists of multiple components; a *grayscale* image consists only of a single component.) A significant portion of this Specification is concerned with how to handle multiple-component images in a flexible, application-independent way.

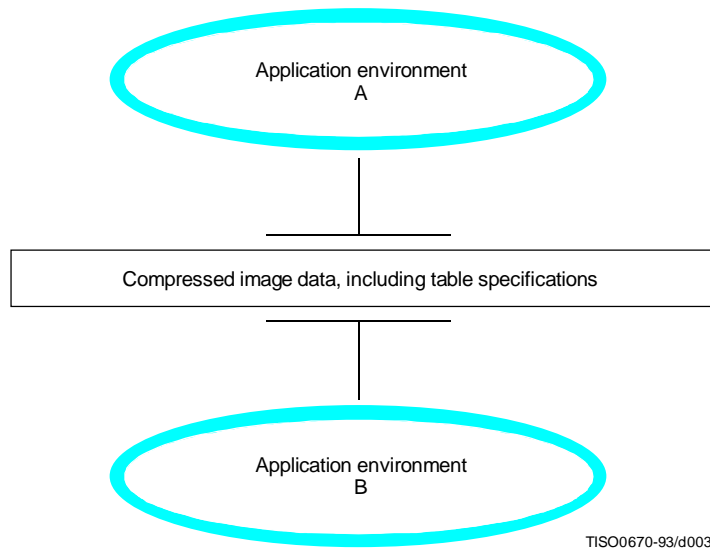


Figure 3 – Interchange format for compressed image data

These figures are also meant to show that the same tables specified for an encoder to use to compress a particular image must be provided to a decoder to reconstruct that image. However, this Specification does not specify how applications should associate tables with compressed image data, nor how they should represent source image data generally within their specific environments.

Consequently, this Specification also specifies the interchange format shown in Figure 3, in which table specifications are included within compressed image data. An image compressed with a specified encoding process within one application environment, A, is passed to a different environment, B, by means of the interchange format. The interchange format does not specify a complete coded image representation. Application-dependent information, e.g. colour space, is outside the scope of this Specification.

4.2 Lossy and lossless compression

This Specification specifies two *classes* of encoding and decoding processes, *lossy* and *lossless* processes. Those based on the *discrete cosine transform* (DCT) are lossy, thereby allowing substantial *compression* to be achieved while producing a reconstructed image with high visual fidelity to the encoder's source image.

The simplest DCT-based *coding process* is referred to as the *baseline sequential* process. It provides a capability which is sufficient for many applications. There are additional DCT-based processes which extend the baseline sequential process to a broader range of applications. In any decoder using *extended DCT-based decoding processes*, the baseline decoding process is required to be present in order to provide a default decoding capability.

The second class of coding processes is not based upon the DCT and is provided to meet the needs of applications requiring lossless compression. These lossless encoding and decoding processes are used independently of any of the DCT-based processes.

A table summarizing the relationship among these lossy and lossless coding processes is included in 4.11.

The amount of compression provided by any of the various processes is dependent on the characteristics of the particular image being compressed, as well as on the picture quality desired by the application and the desired speed of compression and decompression.

4.3 DCT-based coding

Figure 4 shows the main procedures for all encoding processes based on the DCT. It illustrates the special case of a single-component image; this is an appropriate simplification for overview purposes, because all processes specified in this Specification operate on each image component independently.

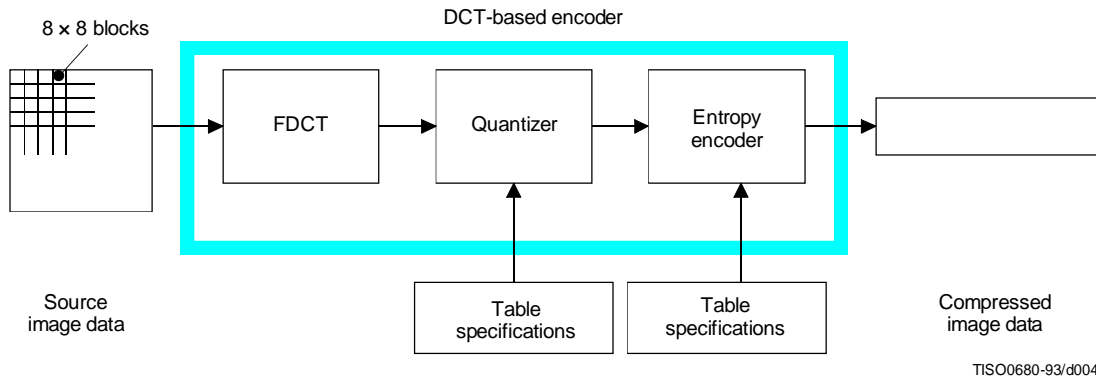


Figure 4 – DCT-based encoder simplified diagram

In the encoding process the input component's *samples* are grouped into 8×8 *blocks*, and each block is transformed by the *forward DCT* (FDCT) into a set of 64 values referred to as *DCT coefficients*. One of these values is referred to as the *DC coefficient* and the other 63 as the *AC coefficients*.

Each of the 64 coefficients is then *quantized* using one of 64 corresponding values from a *quantization table* (determined by one of the table specifications shown in Figure 4). No default values for quantization tables are specified in this Specification; applications may specify values which customize picture quality for their particular image characteristics, display devices, and viewing conditions.

After quantization, the DC coefficient and the 63 AC coefficients are prepared for *entropy encoding*, as shown in Figure 5. The previous quantized DC coefficient is used to predict the current quantized DC coefficient, and the difference is encoded. The 63 quantized AC coefficients undergo no such differential encoding, but are converted into a one-dimensional *zig-zag sequence*, as shown in Figure 5.

The quantized coefficients are then passed to an entropy encoding procedure which compresses the data further. One of two entropy coding procedures can be used, as described in 4.6. If *Huffman encoding* is used, *Huffman table specifications* must be provided to the encoder. If *arithmetic encoding* is used, *arithmetic coding conditioning table specifications* may be provided, otherwise the default conditioning table specifications shall be used.

Figure 6 shows the main procedures for all DCT-based decoding processes. Each step shown performs essentially the inverse of its corresponding main procedure within the encoder. The entropy decoder decodes the zig-zag sequence of quantized DCT coefficients. After *dequantization* the DCT coefficients are transformed to an 8×8 block of samples by the *inverse DCT* (IDCT).

4.4 Lossless coding

Figure 7 shows the main procedures for the lossless encoding processes. A *predictor* combines the reconstructed values of up to three neighbourhood samples at positions a, b, and c to form a prediction of the sample at position x as shown in Figure 8. This prediction is then subtracted from the actual value of the sample at position x, and the difference is losslessly entropy-coded by either Huffman or arithmetic coding.

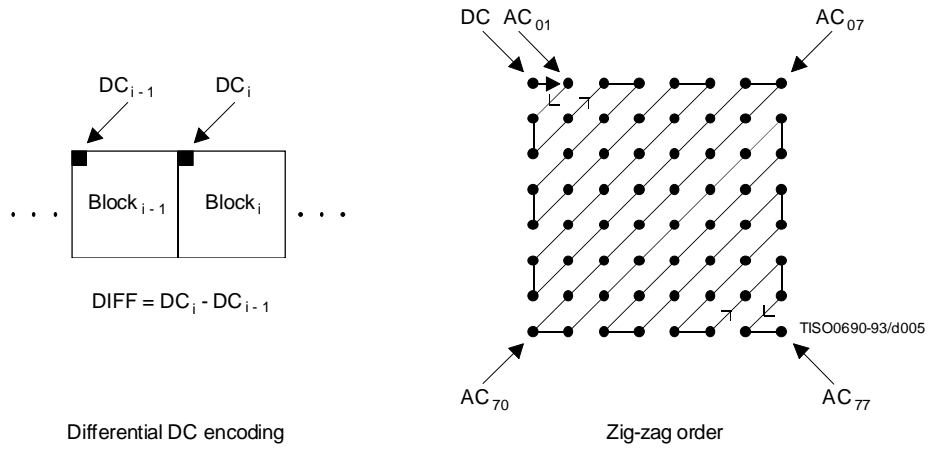


Figure 5 – Preparation of quantized coefficients for entropy encoding

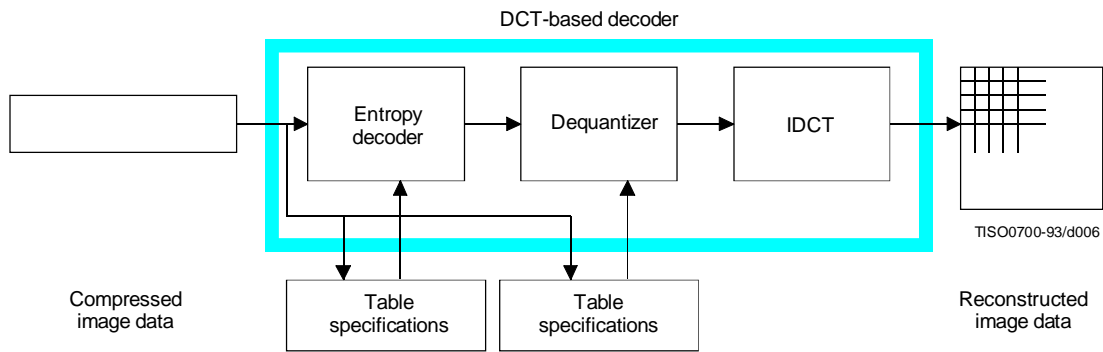


Figure 6 – DCT-based decoder simplified diagram

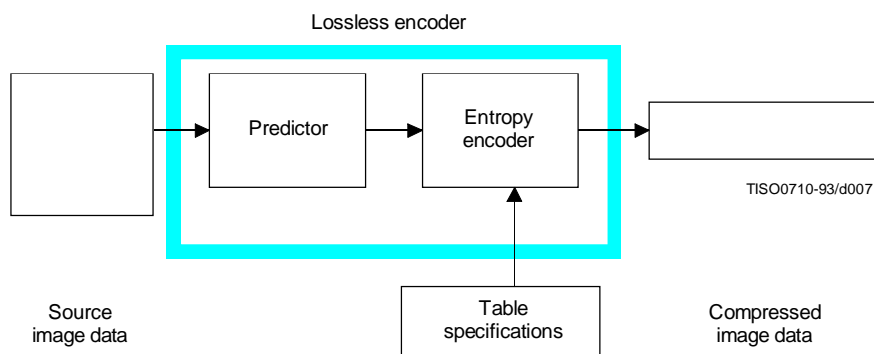


Figure 7 – Lossless encoder simplified diagram

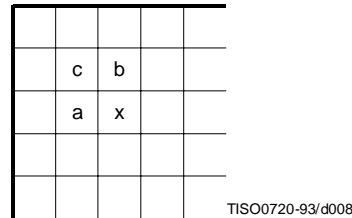


Figure 8 – 3-sample prediction neighbourhood

This encoding process may also be used in a slightly modified way, whereby the *precision* of the input samples is reduced by one or more bits prior to the lossless coding. This achieves higher compression than the lossless process (but lower compression than the DCT-based processes for equivalent visual fidelity), and limits the reconstructed image's worst-case sample error to the amount of input precision reduction.

4.5 Modes of operation

There are four distinct *modes of operation* under which the various coding processes are defined: *sequential DCT-based*, *progressive DCT-based*, *lossless*, and *hierarchical*. (Implementations are not required to provide all of these.) The lossless mode of operation was described in 4.4. The other modes of operation are compared as follows.

For the sequential DCT-based mode, 8×8 sample blocks are typically input block by block from left to right, and block-row by block-row from top to bottom. After a block has been transformed by the forward DCT, quantized and prepared for entropy encoding, all 64 of its quantized DCT coefficients can be immediately entropy encoded and output as part of the compressed image data (as was described in 4.3), thereby minimizing coefficient storage requirements.

For the progressive DCT-based mode, 8×8 blocks are also typically encoded in the same order, but in multiple *scans* through the image. This is accomplished by adding an image-sized coefficient memory buffer (not shown in Figure 4) between the quantizer and the entropy encoder. As each block is transformed by the forward DCT and quantized, its coefficients are stored in the buffer. The DCT coefficients in the buffer are then partially encoded in each of multiple scans. The typical sequence of image presentation at the output of the decoder for sequential versus progressive modes of operation is shown in Figure 9.

There are two procedures by which the quantized coefficients in the buffer may be partially encoded within a scan. First, only a specified *band* of coefficients from the zig-zag sequence need be encoded. This procedure is called *spectral selection*, because each band typically contains coefficients which occupy a lower or higher part of the *frequency* spectrum for that 8×8 block. Secondly, the coefficients within the current band need not be encoded to their full (quantized) accuracy within each scan. Upon a coefficient's first encoding, a specified number of most significant bits is encoded first. In subsequent scans, the less significant bits are then encoded. This procedure is called *successive approximation*. Either procedure may be used separately, or they may be mixed in flexible combinations.

In hierarchical mode, an image is encoded as a sequence of *frames*. These frames provide *reference reconstructed components* which are usually needed for prediction in subsequent frames. Except for the first frame for a given component, *differential frames* encode the difference between source components and reference reconstructed components. The coding of the differences may be done using only DCT-based processes, only lossless processes, or DCT-based processes with a final lossless process for each component. *Downsampling* and *upsampling filters* may be used to provide a pyramid of spatial resolutions as shown in Figure 10. Alternatively, the hierarchical mode can be used to improve the quality of the reconstructed components at a given spatial resolution.

Hierarchical mode offers a progressive presentation similar to the progressive DCT-based mode but is useful in environments which have multi-resolution requirements. Hierarchical mode also offers the capability of progressive coding to a final lossless stage.

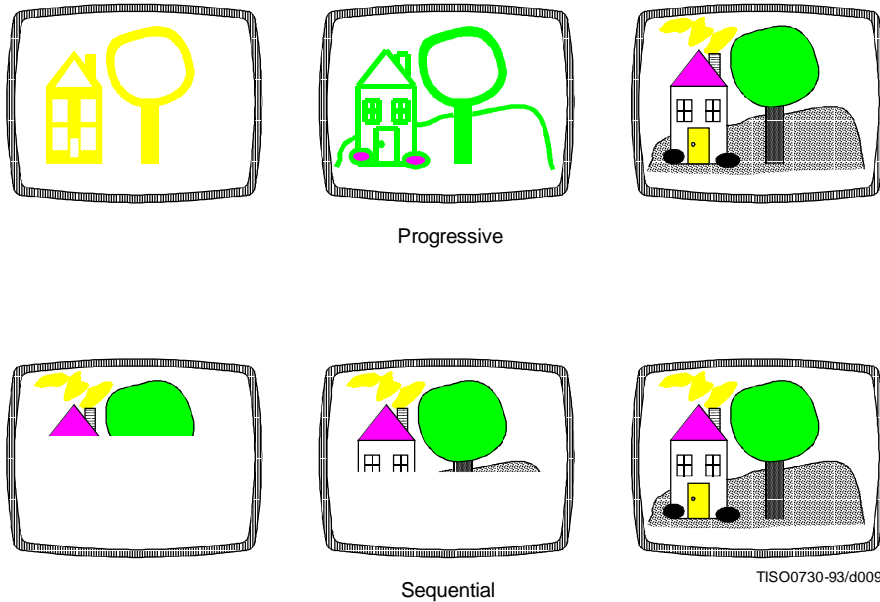


Figure 9 – Progressive versus sequential presentation

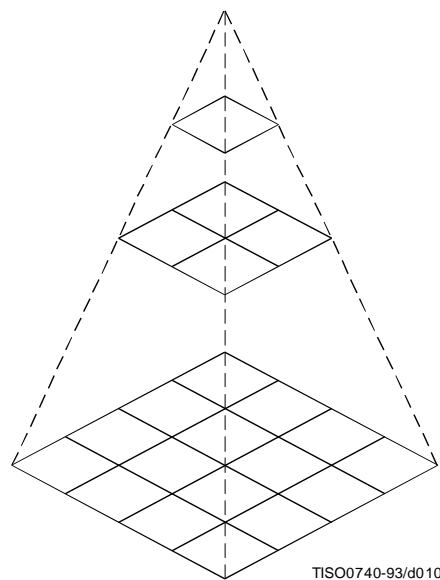


Figure 10 – Hierarchical multi-resolution encoding

4.6 Entropy coding alternatives

Two alternative entropy coding procedures are specified: Huffman coding and arithmetic coding. Huffman coding procedures use Huffman tables, determined by one of the table specifications shown in Figures 1 and 2. Arithmetic coding procedures use arithmetic coding conditioning tables, which may also be determined by a table specification. No default values for Huffman tables are specified, so that applications may choose tables appropriate for their own environments. Default tables are defined for the arithmetic coding conditioning.

The baseline sequential process uses Huffman coding, while the extended DCT-based and lossless processes may use either Huffman or arithmetic coding.

4.7 Sample precision

For DCT-based processes, two alternative sample precisions are specified: either 8 bits or 12 bits per sample. Applications which use samples with other precisions can use either 8-bit or 12-bit precision by shifting their source image samples appropriately. The baseline process uses only 8-bit precision. DCT-based implementations which handle 12-bit source image samples are likely to need greater computational resources than those which handle only 8-bit source images. Consequently in this Specification separate normative requirements are defined for 8-bit and 12-bit DCT-based processes.

For lossless processes the sample precision is specified to be from 2 to 16 bits.

4.8 Multiple-component control

Subclauses 4.3 and 4.4 give an overview of one major part of the encoding and decoding processes – those which operate on the sample values in order to achieve compression. There is another major part as well – the procedures which control the order in which the image data from multiple components are processed to create the compressed data, and which ensure that the proper set of table data is applied to the proper *data units* in the image. (A data unit is a sample for lossless processes and an 8×8 block of samples for DCT-based processes.)

4.8.1 Interleaving multiple components

Figure 11 shows an example of how an encoding process selects between multiple source image components as well as multiple sets of table data, when performing its encoding procedures. The source image in this example consists of the three components A, B and C, and there are two sets of table specifications. (This simplified view does not distinguish between the quantization tables and entropy coding tables.)

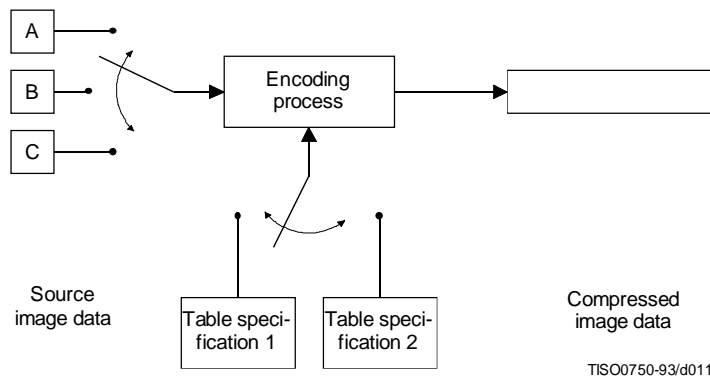


Figure 11 – Component-interleave and table-switching control

In sequential mode, encoding is *non-interleaved* if the encoder compresses all image data units in component A before beginning component B, and then in turn all of B before C. Encoding is *interleaved* if the encoder compresses a data unit from A, a data unit from B, a data unit from C, then back to A, etc. These alternatives are illustrated in Figure 12, which shows a case in which all three image components have identical dimensions: X columns by Y lines, for a total of n data units each.

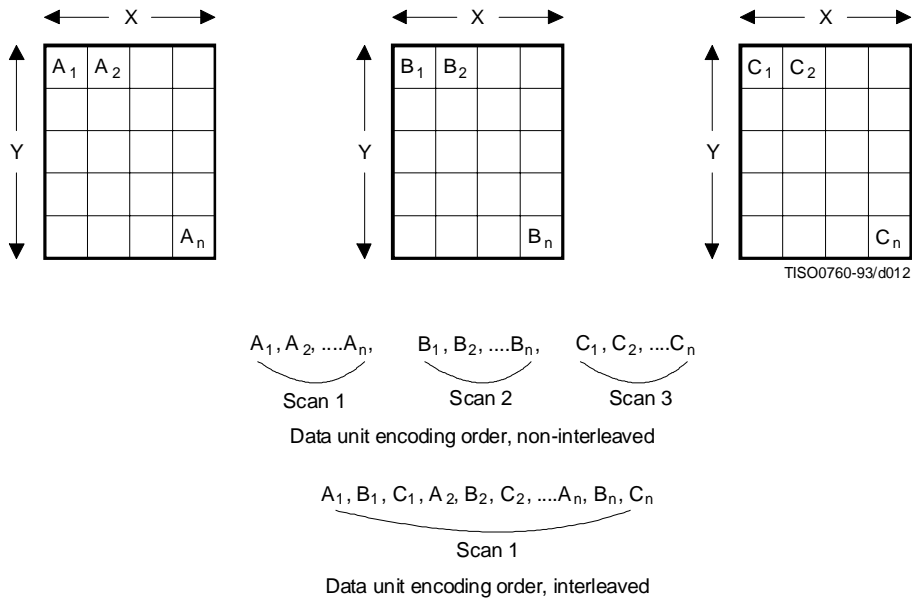


Figure 12 – Interleaved versus non-interleaved encoding order

These control procedures are also able to handle cases in which the source image components have different dimensions. Figure 13 shows a case in which two of the components, B and C, have half the number of horizontal samples relative to component A. In this case, two data units from A are interleaved with one each from B and C. Cases in which components of an image have more complex relationships, such as different horizontal and vertical dimensions, can be handled as well. (See Annex A.)

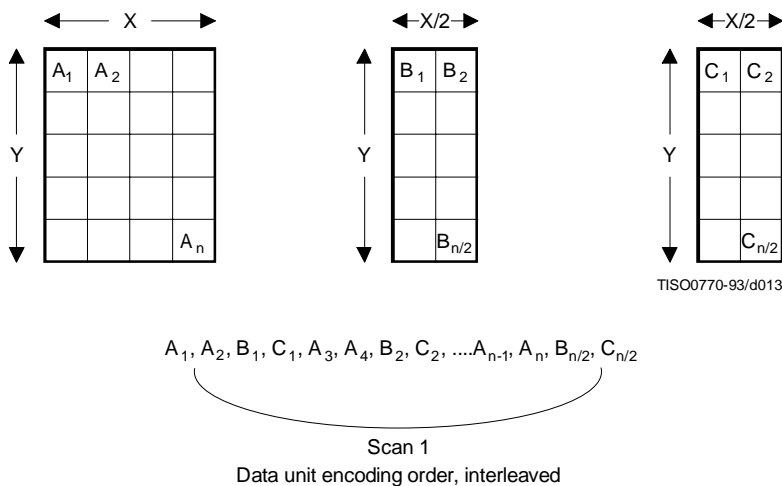


Figure 13 – Interleaved order for components with different dimensions

4.8.2 Minimum coded unit

Related to the concepts of multiple-component interleave is the *minimum coded unit* (MCU). If the compressed image data is non-interleaved, the MCU is defined to be one data unit. For example, in Figure 12 the MCU for the non-interleaved case is a single data unit. If the compressed data is interleaved, the MCU contains one or more data units from each component. For the interleaved case in Figure 12, the (first) MCU consists of the three interleaved data units A_1 , B_1 , C_1 . In the example of Figure 13, the (first) MCU consists of the four data units A_1 , A_2 , B_1 , C_1 .

4.9 Structure of compressed data

Figures 1, 2, and 3 all illustrate slightly different views of compressed image data. Figure 1 shows this data as the output of an encoding process, Figure 2 shows it as the input to a decoding process, and Figure 3 shows compressed image data in the interchange format, at the interface between applications.

Compressed image data are described by a uniform structure and set of *parameters* for both classes of encoding processes (lossy or lossless), and for all modes of operation (sequential, progressive, lossless, and hierarchical). The various parts of the compressed image data are identified by special two-byte codes called *markers*. Some markers are followed by particular sequences of parameters, as in the case of table specifications, *frame header*, or *scan header*. Others are used without parameters for functions such as marking the start-of-image and end-of-image. When a marker is associated with a particular sequence of parameters, the marker and its parameters comprise a *marker segment*.

The data created by the entropy encoder are also segmented, and one particular marker – *the restart marker* – is used to isolate *entropy-coded data segments*. The encoder outputs the restart markers, intermixed with the entropy-coded data, at regular *restart intervals* of the source image data. Restart markers can be identified without having to decode the compressed data to find them. Because they can be independently decoded, they have application-specific uses, such as parallel encoding or decoding, isolation of data corruptions, and semi-random access of entropy-coded segments.

There are three compressed data formats:

- a) the interchange format;
- b) the *abbreviated format* for compressed image data;
- c) the abbreviated format for table-specification data.

4.9.1 Interchange format

In addition to certain required marker segments and the entropy-coded segments, the interchange format shall include the marker segments for all quantization and entropy-coding table specifications needed by the decoding process. This guarantees that a compressed image can cross the boundary between application environments, regardless of how each environment internally associates tables with compressed image data.

4.9.2 Abbreviated format for compressed image data

The abbreviated format for compressed image data is identical to the interchange format, except that it does not include all tables required for decoding. (It may include some of them.) This format is intended for use within applications where alternative mechanisms are available for supplying some or all of the table-specification data needed for decoding.

4.9.3 Abbreviated format for table-specification data

This format contains only table-specification data. It is a means by which the application may install in the decoder the tables required to subsequently reconstruct one or more images.

4.10 Image, frame, and scan

Compressed image data consists of only one image. An image contains only one frame in the cases of sequential and progressive coding processes; an image contains multiple frames for the hierarchical mode.

A frame contains one or more scans. For sequential processes, a scan contains a complete encoding of one or more image components. In Figures 12 and 13, the frame consists of three scans when non-interleaved, and one scan if all three components are interleaved together. The frame could also consist of two scans: one with a non-interleaved component, the other with two components interleaved.

For progressive processes, a scan contains a partial encoding of all data units from one or more image components. Components shall not be interleaved in progressive mode, except for the DC coefficients in the first scan for each component of a progressive frame.

4.11 Summary of coding processes

Table 1 provides a summary of the essential characteristics of the various coding processes specified in this Specification. The full specification of these processes is contained in Annexes F, G, H, and J.

Table 1 – Summary: Essential characteristics of coding processes

Baseline process (required for all DCT-based decoders)
<ul style="list-style-type: none"> ● DCT-based process ● Source image: 8-bit samples within each component ● Sequential ● Huffman coding: 2 AC and 2 DC tables ● Decoders shall process scans with 1, 2, 3, and 4 components ● Interleaved and non-interleaved scans
Extended DCT-based processes
<ul style="list-style-type: none"> ● DCT-based process ● Source image: 8-bit or 12-bit samples ● Sequential or progressive ● Huffman or arithmetic coding: 4 AC and 4 DC tables ● Decoders shall process scans with 1, 2, 3, and 4 components ● Interleaved and non-interleaved scans
Lossless processes
<ul style="list-style-type: none"> ● Predictive process (not DCT-based) ● Source image: P-bit samples ($2 \leq P \leq 16$) ● Sequential ● Huffman or arithmetic coding: 4 DC tables ● Decoders shall process scans with 1, 2, 3, and 4 components ● Interleaved and non-interleaved scans
Hierarchical processes
<ul style="list-style-type: none"> ● Multiple frames (non-differential and differential) ● Uses extended DCT-based or lossless processes ● Decoders shall process scans with 1, 2, 3, and 4 components ● Interleaved and non-interleaved scans

5 Interchange format requirements

The interchange format is the coded representation of compressed image data for exchange between application environments.

The interchange format requirements are that any compressed image data represented in interchange format shall comply with the syntax and code assignments appropriate for the decoding process selected, as specified in Annex B.

Tests for whether compressed image data comply with these requirements are specified in Part 2 of this Specification.

6 Encoder requirements

An encoding process converts source image data to compressed image data. Each of Annexes F, G, H, and J specifies a number of distinct encoding processes for its particular mode of operation.

An encoder is an embodiment of one (or more) of the encoding processes specified in Annexes F, G, H, or J. In order to comply with this Specification, an encoder shall satisfy at least one of the following two requirements.

An encoder shall

- a) with appropriate accuracy, convert source image data to compressed image data which comply with the interchange format syntax specified in Annex B for the encoding process(es) embodied by the encoder;
- b) with appropriate accuracy, convert source image data to compressed image data which comply with the abbreviated format for compressed image data syntax specified in Annex B for the encoding process(es) embodied by the encoder.

For each of the encoding processes specified in Annexes F, G, H, and J, the compliance tests for the above requirements are specified in Part 2 of this Specification.

NOTE – There is **no requirement** in this Specification that any encoder which embodies one of the encoding processes specified in Annexes F, G, H, or J shall be able to operate for all ranges of the parameters which are allowed for that process. An encoder is only required to meet the compliance tests specified in Part 2, and to generate the compressed data format according to Annex B for those parameter values which it does use.

7 Decoder requirements

A decoding process converts compressed image data to reconstructed image data. Each of Annexes F, G, H, and J specifies a number of distinct decoding processes for its particular mode of operation.

A decoder is an embodiment of one (or more) of the decoding processes specified in Annexes F, G, H, or J. In order to comply with this Specification, a decoder shall satisfy all three of the following requirements.

A decoder shall

- a) with appropriate accuracy, convert to reconstructed image data any compressed image data with parameters within the range supported by the application, and which comply with the interchange format syntax specified in Annex B for the decoding process(es) embodied by the decoder;
- b) accept and properly store any table-specification data which comply with the abbreviated format for table-specification data syntax specified in Annex B for the decoding process(es) embodied by the decoder;
- c) with appropriate accuracy, convert to reconstructed image data any compressed image data which comply with the abbreviated format for compressed image data syntax specified in Annex B for the decoding process(es) embodied by the decoder, provided that the table-specification data required for decoding the compressed image data has previously been installed into the decoder.

Additionally, any DCT-based decoder, if it embodies any DCT-based decoding process other than baseline sequential, shall also embody the baseline sequential decoding process.

For each of the decoding processes specified in Annexes F, G, H, and J, the compliance tests for the above requirements are specified in Part 2 of this Specification.

Annex A

Mathematical definitions

(This annex forms an integral part of this Recommendation | International Standard)

A.1 Source image

Source images to which the encoding processes specified in this Specification can be applied are defined in this annex.

A.1.1 Dimensions and sampling factors

As shown in Figure A.1, a source image is defined to consist of N_f components. Each component, with unique identifier C_i , is defined to consist of a rectangular array of samples of x_i columns by y_i lines. The component dimensions are derived from two parameters, X and Y , where X is the maximum of the x_i values and Y is the maximum of the y_i values for all components in the frame. For each component, sampling factors H_i and V_i are defined relating component dimensions x_i and y_i to maximum dimensions X and Y , according to the following expressions:

$$x_i = \left\lceil X \times \frac{H_i}{H_{max}} \right\rceil \text{ and } y_i = \left\lceil Y \times \frac{V_i}{V_{max}} \right\rceil,$$

where H_{max} and V_{max} are the maximum sampling factors for all components in the frame, and $\lceil \cdot \rceil$ is the ceiling function.

As an example, consider an image having 3 components with maximum dimensions of 512 lines and 512 samples per line, and with the following sampling factors:

Component 0	$H_0 = 4, V_0 = 1$
Component 1	$H_1 = 2, V_1 = 2$
Component 2	$H_2 = 1, V_2 = 1$

Then $X = 512$, $Y = 512$, $H_{max} = 4$, $V_{max} = 2$, and x_i and y_i for each component are

Component 0	$x_0 = 512, y_0 = 256$
Component 1	$x_1 = 256, y_1 = 512$
Component 2	$x_2 = 128, y_2 = 256$

NOTE – The X , Y , H_i , and V_i parameters are contained in the frame header of the compressed image data (see B.2.2), whereas the individual component dimensions x_i and y_i are derived by the decoder. Source images with x_i and y_i dimensions which do not satisfy the expressions above cannot be properly reconstructed.

A.1.2 Sample precision

A sample is an integer with precision P bits, with any value in the range 0 through $2^P - 1$. All samples of all components within an image shall have the same precision P . Restrictions on the value of P depend on the mode of operation, as specified in B.2 to B.7.

A.1.3 Data unit

A data unit is a sample in lossless processes and an 8×8 block of contiguous samples in DCT-based processes. The left-most 8 samples of each of the top-most 8 rows in the component shall always be the top-left-most block. With this top-left-most block as the reference, the component is partitioned into contiguous data units to the right and to the bottom (as shown in Figure A.4).

A.1.4 Orientation

Figure A.1 indicates the orientation of an image component by the terms top, bottom, left, and right. The order by which the data units of an image component are input to the compression encoding procedures is defined to be left-to-right and top-to-bottom within the component. (This ordering is precisely defined in A.2.) Applications determine which edges of a source image are defined as top, bottom, left, and right.

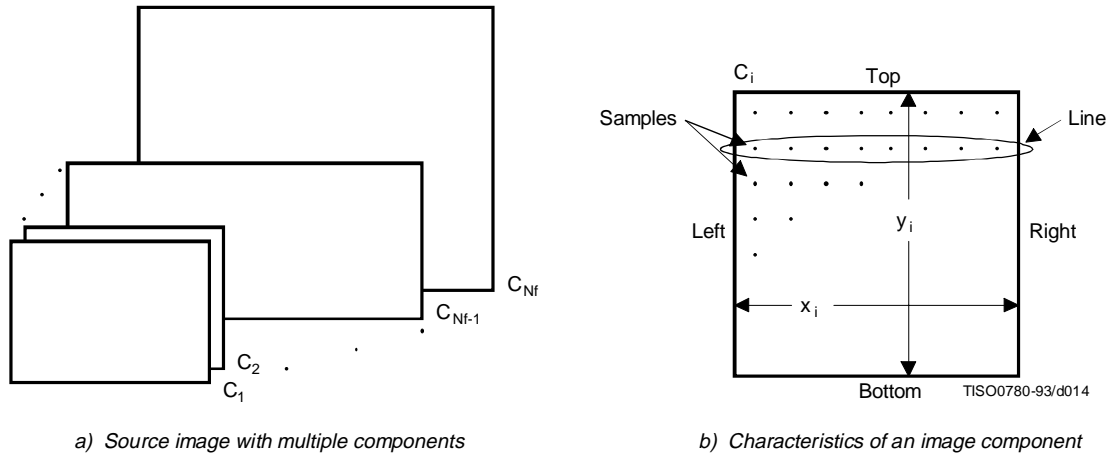


Figure A.1 – Source image characteristics

A.2 Order of source image data encoding

The scan header (see B.2.3) specifies the order by which source image data units shall be encoded and placed within the compressed image data. For a given scan, if the scan header parameter $N_s = 1$, then data from only one source component – the component specified by parameter C_{s_1} – shall be present within the scan. This data is non-interleaved by definition. If $N_s > 1$, then data from the N_s components C_{s_1} through $C_{s_{N_s}}$ shall be present within the scan. This data shall always be interleaved. The order of components in a scan shall be according to the order specified in the frame header.

The ordering of data units and the construction of minimum coded units (MCU) is defined as follows.

A.2.1 Minimum coded unit (MCU)

For non-interleaved data the MCU is one data unit. For interleaved data the MCU is the sequence of data units defined by the sampling factors of the components in the scan.

A.2.2 Non-interleaved order ($N_s = 1$)

When $N_s = 1$ (where N_s is the number of components in a scan), the order of data units within a scan shall be left-to-right and top-to-bottom, as shown in Figure A.2. This ordering applies whenever $N_s = 1$, regardless of the values of H_1 and V_1 .

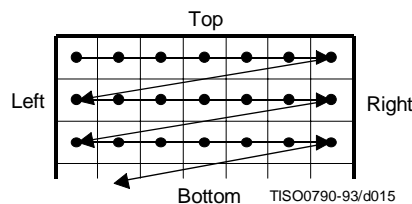


Figure A.2 – Non-interleaved data ordering

A.2.3 Interleaved order (Ns > 1)

When $N_s > 1$, each scan component Cs_i is partitioned into small rectangular arrays of H_k horizontal data units by V_k vertical data units. The subscripts k indicate that H_k and V_k are from the position in the frame header component-specification for which $C_k = Cs_i$. Within each H_k by V_k array, data units are ordered from left-to-right and top-to-bottom. The arrays in turn are ordered from left-to-right and top-to-bottom within each component.

As shown in the example of Figure A.3, $N_s = 4$, and MCU_1 consists of data units taken first from the top-left-most region of Cs_1 , followed by data units from the corresponding region of Cs_2 , then from Cs_3 and then from Cs_4 . MCU_2 follows the same ordering for data taken from the next region to the right for the four components.

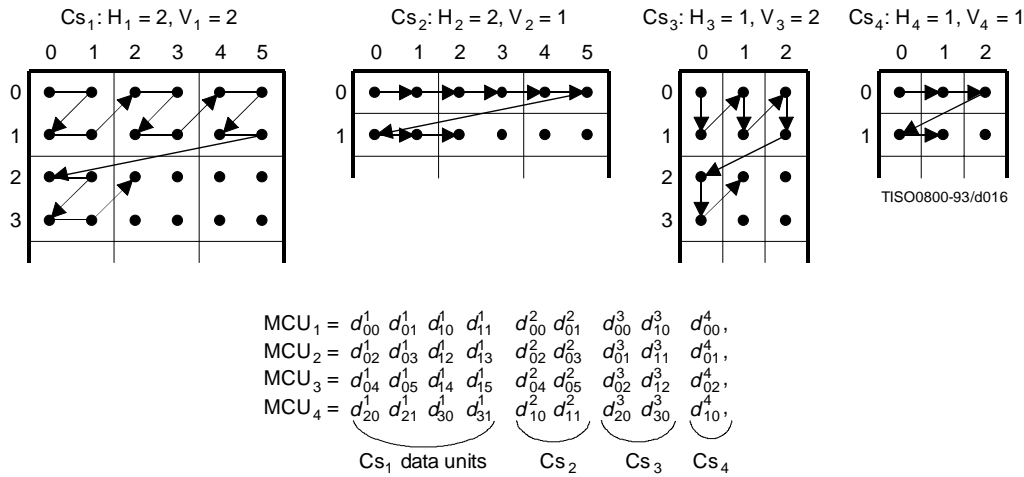


Figure A.3 – Interleaved data ordering example

A.2.4 Completion of partial MCU

For DCT-based processes the data unit is a block. If x_i is not a multiple of 8, the encoding process shall extend the number of columns to complete the right-most sample blocks. If the component is to be interleaved, the encoding process shall also extend the number of samples by one or more additional blocks, if necessary, so that the number of blocks is an integer multiple of H_i . Similarly, if y_i is not a multiple of 8, the encoding process shall extend the number of lines to complete the bottom-most block-row. If the component is to be interleaved, the encoding process shall also extend the number of lines by one or more additional block-rows, if necessary, so that the number of block-rows is an integer multiple of V_i .

NOTE – It is recommended that any incomplete MCUs be completed by replication of the right-most column and the bottom line of each component.

For lossless processes the data unit is a sample. If the component is to be interleaved, the encoding process shall extend the number of samples, if necessary, so that the number is a multiple of H_i . Similarly, the encoding process shall extend the number of lines, if necessary, so that the number of lines is a multiple of V_i .

Any sample added by an encoding process to complete partial MCUs shall be removed by the decoding process.

A.3 DCT compression

A.3.1 Level shift

Before a non-differential frame encoding process computes the FDCT for a block of source image samples, the samples shall be level shifted to a signed representation by subtracting 2^{P-1} , where P is the precision parameter specified in B.2.2. Thus, when $P = 8$, the level shift is by 128; when $P = 12$, the level shift is by 2048.

After a non-differential frame decoding process computes the IDCT and produces a block of reconstructed image samples, an inverse level shift shall restore the samples to the unsigned representation by adding 2^{P-1} and clamping the results to the range 0 to $2^P - 1$.

A.3.2 Orientation of samples for FDCT computation

Figure A.4 shows an image component which has been partitioned into 8×8 blocks for the FDCT computations. Figure A.4 also defines the orientation of the samples within a block by showing the indices used in the FDCT equation of A.3.3.

The definitions of block partitioning and sample orientation also apply to any DCT decoding process and the output reconstructed image. Any sample added by an encoding process to complete partial MCUs shall be removed by the decoding process.

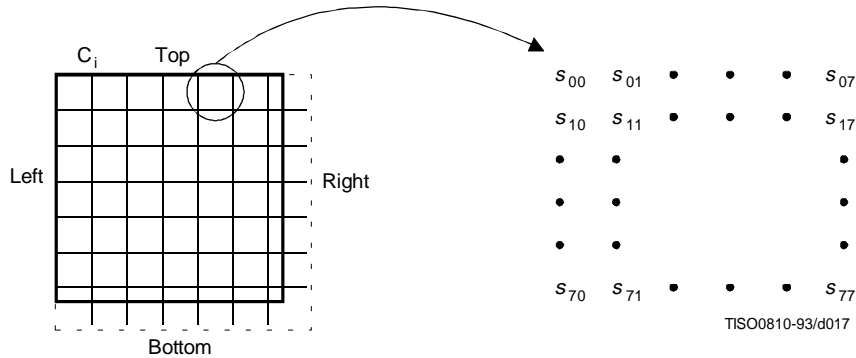


Figure A.4 – Partition and orientation of 8 x 8 sample blocks

A.3.3 FDCT and IDCT (informative)

The following equations specify the ideal functional definition of the FDCT and the IDCT.

NOTE – These equations contain terms which cannot be represented with perfect accuracy by any real implementation. The accuracy requirements for the combined FDCT and quantization procedures are specified in Part 2 of this Specification. The accuracy requirements for the combined dequantization and IDCT procedures are also specified in Part 2 of this Specification.

$$\text{FDCT: } S_{vu} = \frac{1}{4} C_u C_v \sum_{x=0}^7 \sum_{y=0}^7 s_{yx} \cos \frac{(2x+1)u\pi}{16} \cos \frac{(2y+1)v\pi}{16}$$

$$\text{IDCT: } s_{yx} = \frac{1}{4} \sum_{u=0}^7 \sum_{v=0}^7 C_u C_v S_{vu} \cos \frac{(2x+1)u\pi}{16} \cos \frac{(2y+1)v\pi}{16}$$

where

$$C_u, C_v = 1/\sqrt{2} \text{ for } u, v = 0$$

$$C_u, C_v = 1 \text{ otherwise}$$

otherwise.

A.3.4 DCT coefficient quantization (informative) and dequantization (normative)

After the FDCT is computed for a block, each of the 64 resulting DCT coefficients is quantized by a uniform quantizer. The quantizer step size for each coefficient S_{vu} is the value of the corresponding element Q_{vu} from the quantization table specified by the frame parameter Tq_i (see B.2.2).

The uniform quantizer is defined by the following equation. Rounding is to the nearest integer:

$$Sq_{vu} = \text{round} \left(\frac{S_{vu}}{Q_{vu}} \right)$$

Sq_{vu} is the quantized DCT coefficient, normalized by the quantizer step size.

NOTE – This equation contains a term which may not be represented with perfect accuracy by any real implementation. The accuracy requirements for the combined FDCT and quantization procedures are specified in Part 2 of this Specification.

At the decoder, this normalization is removed by the following equation, which defines dequantization:

$$R_{vu} = Sq_{vu} \times Q_{vu}$$

NOTE – Depending on the rounding used in quantization, it is possible that the dequantized coefficient may be outside the expected range.

The relationship among samples, DCT coefficients, and quantization is illustrated in Figure A.5.

A.3.5 Differential DC encoding

After quantization, and in preparation for entropy encoding, the quantized DC coefficient Sq_{00} is treated separately from the 63 quantized AC coefficients. The value that shall be encoded is the difference (DIFF) between the quantized DC coefficient of the current block (DC_i which is also designated as Sq_{00}) and that of the previous block of the same component (PRED):

$$DIFF = DC_i - PRED$$

A.3.6 Zig-zag sequence

After quantization, and in preparation for entropy encoding, the quantized AC coefficients are converted to the zig-zag sequence. The quantized DC coefficient (coefficient zero in the array) is treated separately, as defined in A.3.5. The zig-zag sequence is specified in Figure A.6.

A.4 Point transform

For various procedures data may be optionally divided by a power of 2 by a point transform prior to coding. There are three processes which require a point transform: lossless coding, lossless differential frame coding in the hierarchical mode, and successive approximation coding in the progressive DCT mode.

In the lossless mode of operation the point transform is applied to the input samples. In the difference coding of the hierarchical mode of operation the point transform is applied to the difference between the input component samples and the reference component samples. In both cases the point transform is an integer divide by 2^{Pt} , where Pt is the value of the point transform parameter (see B.2.3).

In successive approximation coding the point transform for the AC coefficients is an integer divide by 2^{Al} , where Al is the successive approximation bit position, low (see B.2.3). The point transform for the DC coefficients is an arithmetic-shift-right by Al bits. This is equivalent to dividing by 2^{Pt} before the level shift (see A.3.1).

The output of the decoder is rescaled by multiplying by 2^{Pt} . An example of the point transform is given in K.10.

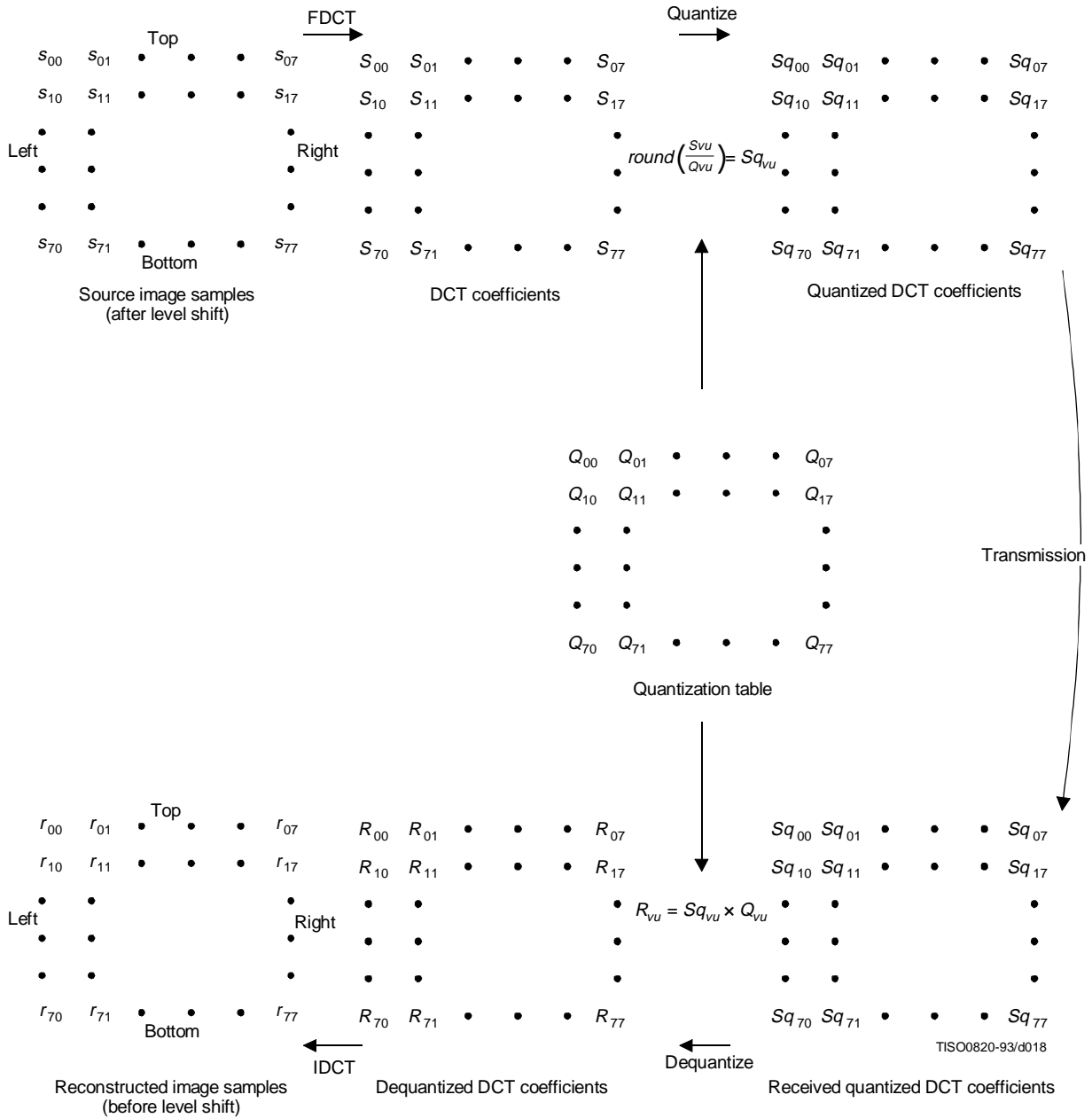


Figure A.5 – Relationship between 8 × 8-block samples and DCT coefficients

0	1	5	6	14	15	27	28
2	4	7	13	16	26	29	42
3	8	12	17	25	30	41	43
9	11	18	24	31	40	44	53
10	19	23	32	39	45	52	54
20	22	33	38	46	51	55	60
21	34	37	47	50	56	59	61
35	36	48	49	57	58	62	63

Figure A.6 – Zig-zag sequence of quantized DCT coefficients

A.5 Arithmetic procedures in lossless and hierarchical modes of operation

In the lossless mode of operation predictions are calculated with full precision and without clamping of either overflow or underflow beyond the range of values allowed by the precision of the input. However, the division by two which is part of some of the prediction calculations shall be approximated by an arithmetic-shift-right by one bit.

The two's complement differences which are coded in either the lossless mode of operation or the differential frame coding in the hierarchical mode of operation are calculated modulo 65 536, thereby restricting the precision of these differences to a maximum of 16 bits. The modulo values are calculated by performing the logical AND operation of the two's complement difference with X'FFFF'. For purposes of coding, the result is still interpreted as a 16 bit two's complement difference. Modulo 65 536 arithmetic is also used in the decoder in calculating the output from the sum of the prediction and this two's complement difference.

Annex B

Compressed data formats

(This annex forms an integral part of this Recommendation | International Standard)

This annex specifies three compressed data formats:

- a) the interchange format, specified in B.2 and B.3;
- b) the abbreviated format for compressed image data, specified in B.4;
- c) the abbreviated format for table-specification data, specified in B.5.

B.1 describes the constituent parts of these formats. B.1.3 and B.1.4 give the conventions for symbols and figures used in the format specifications.

B.1 General aspects of the compressed data format specifications

Structurally, the compressed data formats consist of an ordered collection of parameters, markers, and entropy-coded data segments. Parameters and markers in turn are often organized into marker segments. Because all of these constituent parts are represented with byte-aligned codes, each compressed data format consists of an ordered sequence of 8-bit bytes. For each byte, a most significant bit (MSB) and a least significant bit (LSB) are defined.

B.1.1 Constituent parts

This subclause gives a general description of each of the constituent parts of the compressed data format.

B.1.1.1 Parameters

Parameters are integers, with values specific to the encoding process, source image characteristics, and other features selectable by the application. Parameters are assigned either 4-bit, 1-byte, or 2-byte codes. Except for certain optional groups of parameters, parameters encode critical information without which the decoding process cannot properly reconstruct the image.

The code assignment for a parameter shall be an unsigned integer of the specified length in bits with the particular value of the parameter.

For parameters which are 2 bytes (16 bits) in length, the most significant byte shall come first in the compressed data's ordered sequence of bytes. Parameters which are 4 bits in length always come in pairs, and the pair shall always be encoded in a single byte. The first 4-bit parameter of the pair shall occupy the most significant 4 bits of the byte. Within any 16-, 8-, or 4-bit parameter, the MSB shall come first and LSB shall come last.

B.1.1.2 Markers

Markers serve to identify the various structural parts of the compressed data formats. Most markers start marker segments containing a related group of parameters; some markers stand alone. All markers are assigned two-byte codes: an X'FF' byte followed by a byte which is not equal to 0 or X'FF' (see Table B.1). Any marker may optionally be preceded by any number of fill bytes, which are bytes assigned code X'FF'.

NOTE – Because of this special code-assignment structure, markers make it possible for a decoder to parse the compressed data and locate its various parts without having to decode other segments of image data.

B.1.1.3 Marker assignments

All markers shall be assigned two-byte codes: a X'FF' byte followed by a second byte which is not equal to 0 or X'FF'. The second byte is specified in Table B.1 for each defined marker. An asterisk (*) indicates a marker which stands alone, that is, which is not the start of a marker segment.

Table B.1 – Marker code assignments

Code Assignment	Symbol	Description
Start Of Frame markers, non-differential, Huffman coding		
X'FFC0' X'FFC1' X'FFC2' X'FFC3'	SOF ₀ SOF ₁ SOF ₂ SOF ₃	Baseline DCT Extended sequential DCT Progressive DCT Lossless (sequential)
Start Of Frame markers, differential, Huffman coding		
X'FFC5' X'FFC6' X'FFC7'	SOF ₅ SOF ₆ SOF ₇	Differential sequential DCT Differential progressive DCT Differential lossless (sequential)
Start Of Frame markers, non-differential, arithmetic coding		
X'FFC8' X'FFC9' X'FFCA' X'FFCB'	JPG SOF ₉ SOF ₁₀ SOF ₁₁	Reserved for JPEG extensions Extended sequential DCT Progressive DCT Lossless (sequential)
Start Of Frame markers, differential, arithmetic coding		
X'FFCD' X'FFCE' X'FFCF'	SOF ₁₃ SOF ₁₄ SOF ₁₅	Differential sequential DCT Differential progressive DCT Differential lossless (sequential)
Huffman table specification		
X'FFC4'	DHT	Define Huffman table(s)
Arithmetic coding conditioning specification		
X'FFCC'	DAC	Define arithmetic coding conditioning(s)
Restart interval termination		
X'FFD0' through X'FFD7'	RST _m *	Restart with modulo 8 count "m"
Other markers		
X'FFD8' X'FFD9' X'FFDA' X'FFDB' X'FFDC' X'FFDD' X'FFDE' X'FFDF' X'FFE0' through X'FFE7' X'FFF0' through X'FFF7' X'FFFE'	SOI* EOI* SOS DQT DNL DRI DHP EXP APP _n JPG _n COM	Start of image End of image Start of scan Define quantization table(s) Define number of lines Define restart interval Define hierarchical progression Expand reference component(s) Reserved for application segments Reserved for JPEG extensions Comment
Reserved markers		
X'FF01' X'FF02' through X'FFBF'	TEM* RES	For temporary private use in arithmetic coding Reserved

B.1.1.4 Marker segments

A marker segment consists of a marker followed by a sequence of related parameters. The first parameter in a marker segment is the two-byte length parameter. This length parameter encodes the number of bytes in the marker segment, including the length parameter and excluding the two-byte marker. The marker segments identified by the SOF and SOS marker codes are referred to as headers: the frame header and the scan header respectively.

B.1.1.5 Entropy-coded data segments

An entropy-coded data segment contains the output of an entropy-coding procedure. It consists of an integer number of bytes, whether the entropy-coding procedure used is Huffman or arithmetic.

NOTES

1 Making entropy-coded segments an integer number of bytes is performed as follows: for Huffman coding, 1-bits are used, if necessary, to pad the end of the compressed data to complete the final byte of a segment. For arithmetic coding, byte alignment is performed in the procedure which terminates the entropy-coded segment (see D.1.8).

2 In order to ensure that a marker does not occur within an entropy-coded segment, any X'FF' byte generated by either a Huffman or arithmetic encoder, or an X'FF' byte that was generated by the padding of 1-bits described in NOTE 1 above, is followed by a "stuffed" zero byte (see D.1.6 and F.1.2.3).

B.1.2 Syntax

In B.2 and B.3 the interchange format syntax is specified. For the purposes of this Specification, the syntax specification consists of:

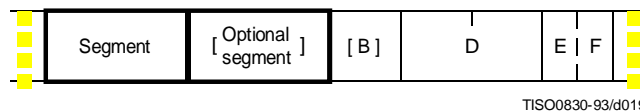
- the required ordering of markers, parameters, and entropy-coded segments;
- identification of optional or conditional constituent parts;
- the name, symbol, and definition of each marker and parameter;
- the allowed values of each parameter;
- any restrictions on the above which are specific to the various coding processes.

The ordering of constituent parts and the identification of which are optional or conditional is specified by the syntax figures in B.2 and B.3. Names, symbols, definitions, allowed values, conditions, and restrictions are specified immediately below each syntax figure.

B.1.3 Conventions for syntax figures

The syntax figures in B.2 and B.3 are a part of the interchange format specification. The following conventions, illustrated in Figure B.1, apply to these figures:

- **parameter/marker indicator:** A thin-lined box encloses either a marker or a single parameter;
- **segment indicator:** A thick-lined box encloses either a marker segment, an entropy-coded data segment, or combinations of these;
- **parameter length indicator:** The width of a thin-lined box is proportional to the parameter length (4, 8, or 16 bits, shown as E, B, and D respectively in Figure B.1) of the marker or parameter it encloses; the width of thick-lined boxes is not meaningful;
- **optional/conditional indicator:** Square brackets indicate that a marker or marker segment is only optionally or conditionally present in the compressed image data;
- **ordering:** In the interchange format a parameter or marker shown in a figure precedes all of those shown to its right, and follows all of those shown to its left;
- **entropy-coded data indicator:** Angled brackets indicate that the entity enclosed has been entropy encoded.



TISO0830-93/d019

Figure B.1 – Syntax notation conventions

B.1.4 Conventions for symbols, code lengths, and values

Following each syntax figure in B.2 and B.3, the symbol, name, and definition for each marker and parameter shown in the figure are specified. For each parameter, the length and allowed values are also specified in tabular form.

The following conventions apply to symbols for markers and parameters:

- all marker symbols have three upper-case letters, and some also have a subscript. Examples: SOI, SOF_n;
- all parameter symbols have one upper-case letter; some also have one lower-case letter and some have subscripts. Examples: Y, Nf, H_i, Tq_i.

B.2 General sequential and progressive syntax

This clause specifies the interchange format syntax which applies to all coding processes for sequential DCT-based, progressive DCT-based, and lossless modes of operation.

B.2.1 High-level syntax

Figure B.2 specifies the order of the high-level constituent parts of the interchange format for all non-hierarchical encoding processes specified in this Specification.

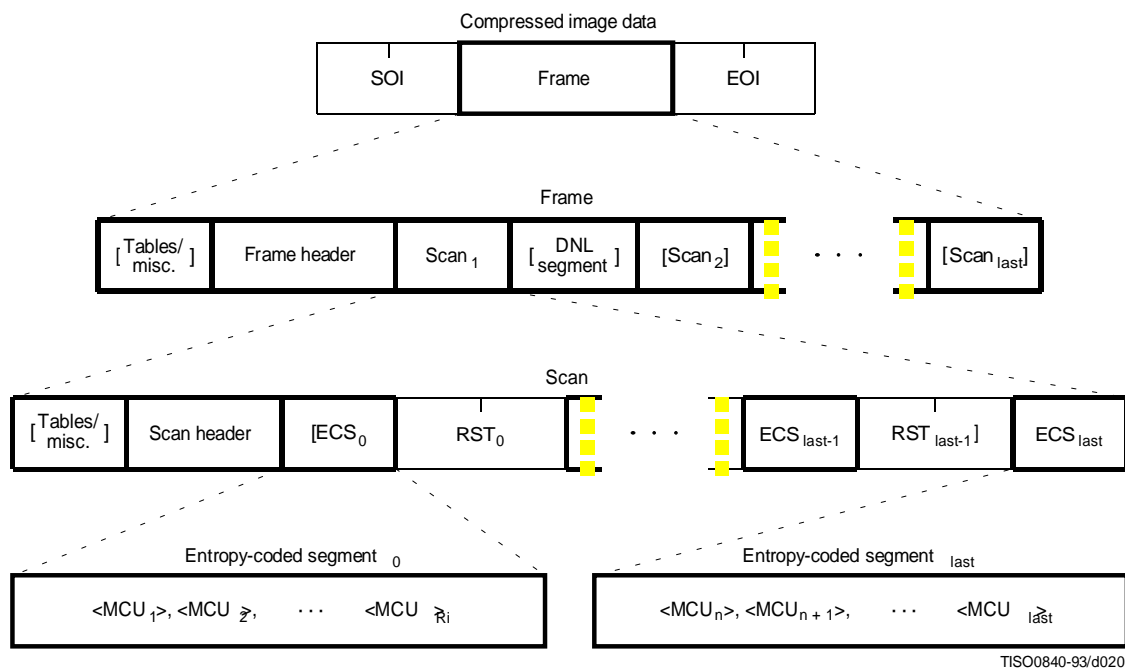


Figure B.2 – Syntax for sequential DCT-based, progressive DCT-based, and lossless modes of operation

The three markers shown in Figure B.2 are defined as follows:

SOI: Start of image marker – Marks the start of a compressed image represented in the interchange format or abbreviated format.

EOI: End of image marker – Marks the end of a compressed image represented in the interchange format or abbreviated format.

RST_m: Restart marker – A conditional marker which is placed between entropy-coded segments only if restart is enabled. There are 8 unique restart markers (m = 0 - 7) which repeat in sequence from 0 to 7, starting with zero for each scan, to provide a modulo 8 restart interval count.

The top level of Figure B.2 specifies that the non-hierarchical interchange format shall begin with an SOI marker, shall contain one frame, and shall end with an EOI marker.

The second level of Figure B.2 specifies that a frame shall begin with a frame header and shall contain one or more scans. A frame header may be preceded by one or more table-specification or miscellaneous marker segments as specified in B.2.4. If a DNL segment (see B.2.5) is present, it shall immediately follow the first scan.

For sequential DCT-based and lossless processes each scan shall contain from one to four image components. If two to four components are contained within a scan, they shall be interleaved within the scan. For progressive DCT-based processes each image component is only partially contained within any one scan. Only the first scan(s) for the components (which contain only DC coefficient data) may be interleaved.

The third level of Figure B.2 specifies that a scan shall begin with a scan header and shall contain one or more entropy-coded data segments. Each scan header may be preceded by one or more table-specification or miscellaneous marker segments. If restart is not enabled, there shall be only one entropy-coded segment (the one labeled "last"), and no restart markers shall be present. If restart is enabled, the number of entropy-coded segments is defined by the size of the image and the defined restart interval. In this case, a restart marker shall follow each entropy-coded segment except the last one.

The fourth level of Figure B.2 specifies that each entropy-coded segment is comprised of a sequence of entropy-coded MCUs. If restart is enabled and the restart interval is defined to be R_i , each entropy-coded segment except the last one shall contain R_i MCUs. The last one shall contain whatever number of MCUs completes the scan.

Figure B.2 specifies the locations where table-specification segments **may** be present. However, this Specification hereby specifies that the interchange format **shall** contain all table-specification data necessary for decoding the compressed image. Consequently, the required table-specification data **shall** be present at one or more of the allowed locations.

B.2.2 Frame header syntax

Figure B.3 specifies the frame header which shall be present at the start of a frame. This header specifies the source image characteristics (see A.1), the components in the frame, and the sampling factors for each component, and specifies the destinations from which the quantized tables to be used with each component are retrieved.

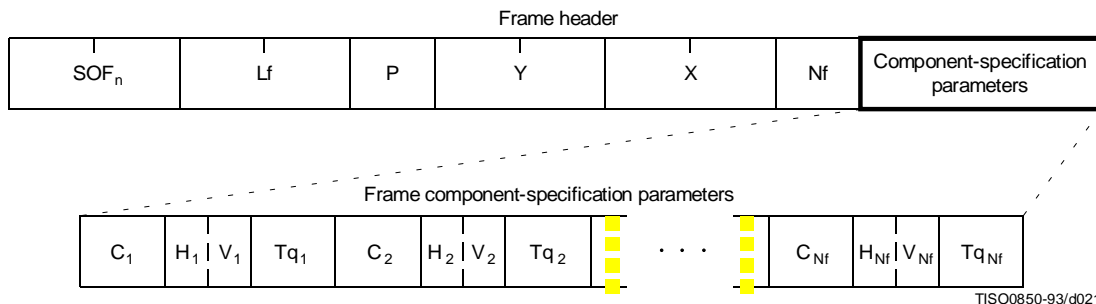


Figure B.3 – Frame header syntax

The markers and parameters shown in Figure B.3 are defined below. The size and allowed values of each parameter are given in Table B.2. In Table B.2 (and similar tables which follow), value choices are separated by commas (e.g. 8, 12) and inclusive bounds are separated by dashes (e.g. 0 - 3).

SOF_n: Start of frame marker – Marks the beginning of the frame parameters. The subscript n identifies whether the encoding process is baseline sequential, extended sequential, progressive, or lossless, as well as which entropy encoding procedure is used.

SOF₀: Baseline DCT

SOF₁: Extended sequential DCT, Huffman coding

SOF₂: Progressive DCT, Huffman coding

SOF₃: Lossless (sequential), Huffman coding

SOF₉: Extended sequential DCT, arithmetic coding

SOF₁₀: Progressive DCT, arithmetic coding

SOF₁₁: Lossless (sequential), arithmetic coding

Lf: Frame header length – Specifies the length of the frame header shown in Figure B.3 (see B.1.1.4).

P: Sample precision – Specifies the precision in bits for the samples of the components in the frame.

Y: Number of lines – Specifies the maximum number of lines in the source image. This shall be equal to the number of lines in the component with the maximum number of vertical samples (see A.1.1). Value 0 indicates that the number of lines shall be defined by the DNL marker and parameters at the end of the first scan (see B.2.5).

X: Number of samples per line – Specifies the maximum number of samples per line in the source image. This shall be equal to the number of samples per line in the component with the maximum number of horizontal samples (see A.1.1).

Nf: Number of image components in frame – Specifies the number of source image components in the frame. The value of Nf shall be equal to the number of sets of frame component specification parameters (C_i , H_i , V_i , and T_{qi}) present in the frame header.

C_i : Component identifier – Assigns a unique label to the i th component in the sequence of frame component specification parameters. These values shall be used in the scan headers to identify the components in the scan. The value of C_i shall be different from the values of C_1 through C_{i-1} .

H_i : Horizontal sampling factor – Specifies the relationship between the component horizontal dimension and maximum image dimension X (see A.1.1); also specifies the number of horizontal data units of component C_i in each MCU, when more than one component is encoded in a scan.

V_i : Vertical sampling factor – Specifies the relationship between the component vertical dimension and maximum image dimension Y (see A.1.1); also specifies the number of vertical data units of component C_i in each MCU, when more than one component is encoded in a scan.

T_{qi} : Quantization table destination selector – Specifies one of four possible quantization table destinations from which the quantization table to use for dequantization of DCT coefficients of component C_i is retrieved. If the decoding process uses the dequantization procedure, this table shall have been installed in this destination by the time the decoder is ready to decode the scan(s) containing component C_i . The destination shall not be re-specified, or its contents changed, until all scans containing C_i have been completed.

Table B.2 – Frame header parameter sizes and values

Parameter	Size (bits)	Values			
		Sequential DCT		Progressive DCT	Lossless
		Baseline	Extended		
Lf	16	8 + 3 × Nf			
P	8	8	8, 12	8, 12	2-16
Y	16	0-65 535			
X	16	1-65 535			
Nf	8	1-255	1-255	1-4	1-255
C_i	8	0-255			
H_i	4	1-4			
V_i	4	1-4			
T_{qi}	8	0-3	0-3	0-3	0

B.2.3 Scan header syntax

Figure B.4 specifies the scan header which shall be present at the start of a scan. This header specifies which component(s) are contained in the scan, specifies the destinations from which the entropy tables to be used with each component are retrieved, and (for the progressive DCT) which part of the DCT quantized coefficient data is contained in the scan. For lossless processes the scan parameters specify the predictor and the point transform.

NOTE – If there is only one image component present in a scan, that component is, by definition, non-interleaved. If there is more than one image component present in a scan, the components present are, by definition, interleaved.

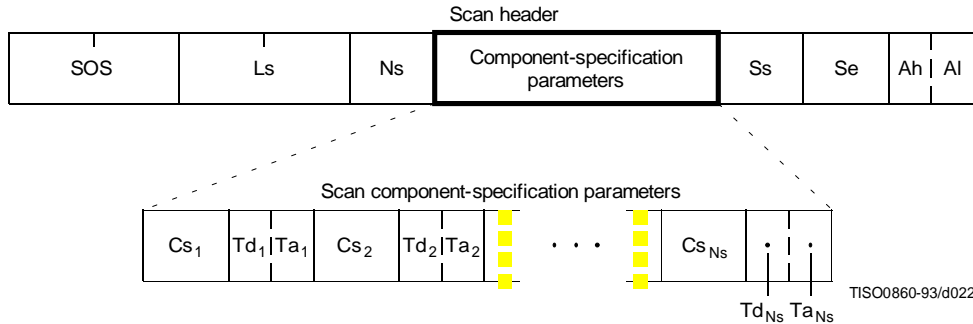


Figure B.4 – Scan header syntax

The marker and parameters shown in Figure B.4 are defined below. The size and allowed values of each parameter are given in Table B.3.

SOS: Start of scan marker – Marks the beginning of the scan parameters.

Ls: Scan header length – Specifies the length of the scan header shown in Figure B.4 (see B.1.1.4).

Ns: Number of image components in scan – Specifies the number of source image components in the scan. The value of Ns shall be equal to the number of sets of scan component specification parameters (Cs_j, Td_j, and Ta_j) present in the scan header.

Cs_j: Scan component selector – Selects which of the N_f image components specified in the frame parameters shall be the jth component in the scan. Each Cs_j shall match one of the C_i values specified in the frame header, and the ordering in the scan header shall follow the ordering in the frame header. If Ns > 1, the order of interleaved components in the MCU is Cs₁ first, Cs₂ second, etc. If Ns > 1, the following restriction shall be placed on the image components contained in the scan:

$$\sum_{j=1}^{N_s} H_j \times V_j \leq 10,$$

where H_j and V_j are the horizontal and vertical sampling factors for scan component j. These sampling factors are specified in the frame header for component i, where i is the frame component specification index for which frame component identifier C_i matches scan component selector Cs_j.

As an example, consider an image having 3 components with maximum dimensions of 512 lines and 512 samples per line, and with the following sampling factors:

Component 0	$H_0 = 4,$	$V_0 = 1$
Component 1	$H_1 = 1,$	$V_1 = 2$
Component 2	$H_2 = 2$	$V_2 = 2$

Then the summation of H_j × V_j is (4 × 1) + (1 × 2) + (2 × 2) = 10.

The value of Cs_j shall be different from the values of Cs₁ to Cs_{j-1}.

T_{dj}: DC entropy coding table destination selector – Specifies one of four possible DC entropy coding table destinations from which the entropy table needed for decoding of the DC coefficients of component C_{s_j} is retrieved. The DC entropy table shall have been installed in this destination (see B.2.4.2 and B.2.4.3) by the time the decoder is ready to decode the current scan. This parameter specifies the entropy coding table destination for the lossless processes.

T_{aj}: AC entropy coding table destination selector – Specifies one of four possible AC entropy coding table destinations from which the entropy table needed for decoding of the AC coefficients of component C_{s_j} is retrieved. The AC entropy table selected shall have been installed in this destination (see B.2.4.2 and B.2.4.3) by the time the decoder is ready to decode the current scan. This parameter is zero for the lossless processes.

S_s: Start of spectral or predictor selection – In the DCT modes of operation, this parameter specifies the first DCT coefficient in each block in zig-zag order which shall be coded in the scan. This parameter shall be set to zero for the sequential DCT processes. In the lossless mode of operations this parameter is used to select the predictor.

S_e: End of spectral selection – Specifies the last DCT coefficient in each block in zig-zag order which shall be coded in the scan. This parameter shall be set to 63 for the sequential DCT processes. In the lossless mode of operations this parameter has no meaning. It shall be set to zero.

A_h: Successive approximation bit position high – This parameter specifies the point transform used in the preceding scan (i.e. successive approximation bit position low in the preceding scan) for the band of coefficients specified by S_s and S_e. This parameter shall be set to zero for the first scan of each band of coefficients. In the lossless mode of operations this parameter has no meaning. It shall be set to zero.

A_l: Successive approximation bit position low or point transform – In the DCT modes of operation this parameter specifies the point transform, i.e. bit position low, used before coding the band of coefficients specified by S_s and S_e. This parameter shall be set to zero for the sequential DCT processes. In the lossless mode of operations, this parameter specifies the point transform, Pt.

The entropy coding table destination selectors, T_{dj} and T_{aj}, specify either Huffman tables (in frames using Huffman coding) or arithmetic coding tables (in frames using arithmetic coding). In the latter case the entropy coding table destination selector specifies both an arithmetic coding conditioning table destination and an associated statistics area.

Table B.3 – Scan header parameter size and values

Parameter	Size (bits)	Values			
		Sequential DCT		Progressive DCT	Lossless
		Baseline	Extended		
L _s	16	6 + 2 × N _s			
N _s	8	1-4			
C _{s_j}	8	0-255 ^{a)}			
T _{dj}	4	0-1	0-3	0-3	0-3
T _{aj}	4	0-1	0-3	0-3	0
S _s	8	0	0	0-63	1-7 ^{b)}
S _e	8	63	63	S _s -63 ^{c)}	0
A _h	4	0	0	0-13	0
A _l	4	0	0	0-13	0-15

a) C_{s_j} shall be a member of the set of C_i specified in the frame header.
 b) 0 for lossless differential frames in the hierarchical mode (see B.3).
 c) 0 if S_s equals zero.

B.2.4 Table-specification and miscellaneous marker segment syntax

Figure B.5 specifies that, at the places indicated in Figure B.2, any of the table-specification segments or miscellaneous marker segments specified in B.2.4.1 through B.2.4.6 may be present in any order and with no limit on the number of segments.

If any table specification for a particular destination occurs in the compressed image data, it shall replace any previous table specified for this destination, and shall be used whenever this destination is specified in the remaining scans in the frame or subsequent images represented in the abbreviated format for compressed image data. If a table specification for a given destination occurs more than once in the compressed image data, each specification shall replace the previous specification. The quantization table specification shall not be altered between progressive DCT scans of a given component.

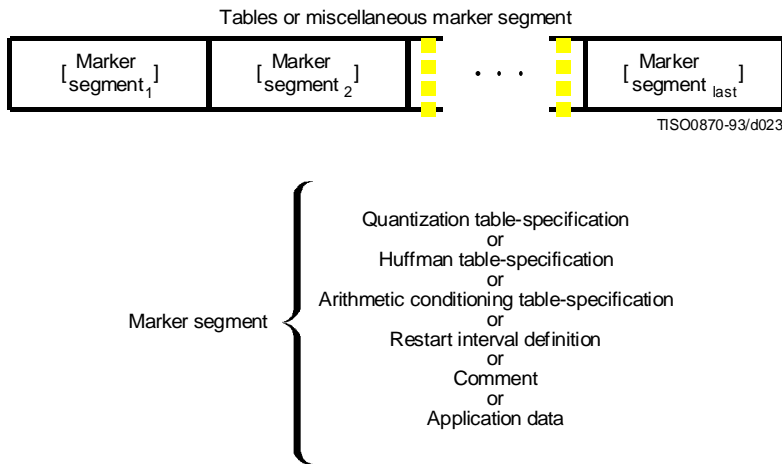


Figure B.5 – Tables/miscellaneous marker segment syntax

B.2.4.1 Quantization table-specification syntax

Figure B.6 specifies the marker segment which defines one or more quantization tables.

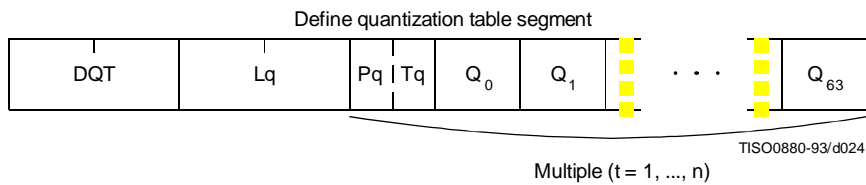


Figure B.6 – Quantization table syntax

The marker and parameters shown in Figure B.6 are defined below. The size and allowed values of each parameter are given in Table B.4.

DQT: Define quantization table marker – Marks the beginning of quantization table-specification parameters.

Lq: Quantization table definition length – Specifies the length of all quantization table parameters shown in Figure B.6 (see B.1.1.4).

P_q: Quantization table element precision – Specifies the precision of the Q_k values. Value 0 indicates 8-bit Q_k values; value 1 indicates 16-bit Q_k values. P_q shall be zero for 8 bit sample precision P (see B.2.2).

T_q: Quantization table destination identifier – Specifies one of four possible destinations at the decoder into which the quantization table shall be installed.

Q_k: Quantization table element – Specifies the kth element out of 64 elements, where k is the index in the zig-zag ordering of the DCT coefficients. The quantization elements shall be specified in zig-zag scan order.

Table B.4 – Quantization table-specification parameter sizes and values

Parameter	Size (bits)	Values			
		Sequential DCT		Progressive DCT	Lossless
		Baseline	Extended		
L _q	16	$2 + \sum_{t=1}^n (65 + 64 \times Pq(t))$			Undefined
P _q	4	0	0, 1	0, 1	Undefined
T _q	4	0-3			Undefined
Q _k	8, 16	1-255, 1-65 535			Undefined

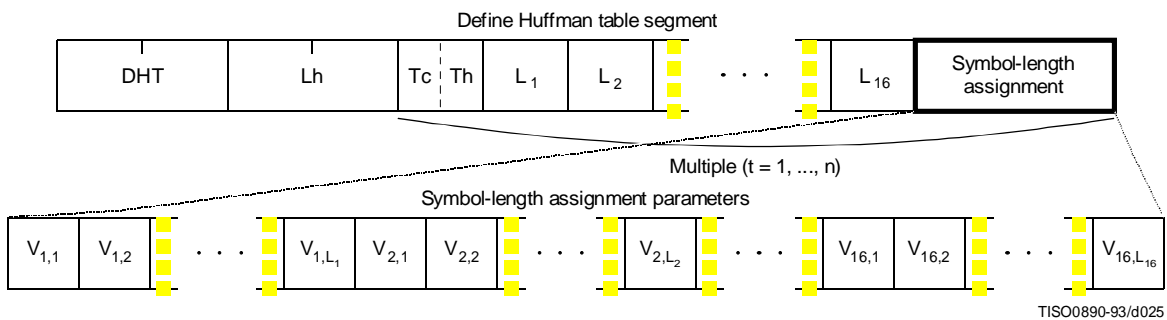
The value n in Table B.4 is the number of quantization tables specified in the DQT marker segment.

Once a quantization table has been defined for a particular destination, it replaces the previous tables stored in that destination and shall be used, when referenced, in the remaining scans of the current image and in subsequent images represented in the abbreviated format for compressed image data. If a table has never been defined for a particular destination, then when this destination is specified in a frame header, the results are unpredictable.

An 8-bit DCT-based process shall not use a 16-bit precision quantization table.

B.2.4.2 Huffman table-specification syntax

Figure B.7 specifies the marker segment which defines one or more Huffman table specifications.



TISO0890-93/d025

Figure B.7 – Huffman table syntax

The marker and parameters shown in Figure B.7 are defined below. The size and allowed values of each parameter are given in Table B.5.

DHT: Define Huffman table marker – Marks the beginning of Huffman table definition parameters.

Lh: Huffman table definition length – Specifies the length of all Huffman table parameters shown in Figure B.7 (see B.1.1.4).

Tc: Table class – 0 = DC table or lossless table, 1 = AC table.

Th: Huffman table destination identifier – Specifies one of four possible destinations at the decoder into which the Huffman table shall be installed.

L_i: Number of Huffman codes of length i – Specifies the number of Huffman codes for each of the 16 possible lengths allowed by this Specification. L_i's are the elements of the list BITS.

V_{i,j}: Value associated with each Huffman code – Specifies, for each i, the value associated with each Huffman code of length i. The meaning of each value is determined by the Huffman coding model. The V_{i,j}'s are the elements of the list HUFFVAL.

Table B.5 – Huffman table specification parameter sizes and values

Parameter	Size (bits)	Values			
		Sequential DCT		Progressive DCT	Lossless
		Baseline	Extended		
Lh	16	$2 + \sum_{t=1}^n (17 + m_t)$			
Tc	4	0, 1			0
Th	4	0, 1	0-3		
L _i	8	0-255			
V _{i,j}	8	0-255			

The value n in Table B.5 is the number of Huffman tables specified in the DHT marker segment. The value m_t is the number of parameters which follow the 16 L_i(t) parameters for Huffman table t, and is given by:

$$m_t = \sum_{i=1}^{16} L_i$$

In general, m_t is different for each table.

Once a Huffman table has been defined for a particular destination, it replaces the previous tables stored in that destination and shall be used when referenced, in the remaining scans of the current image and in subsequent images represented in the abbreviated format for compressed image data. If a table has never been defined for a particular destination, then when this destination is specified in a scan header, the results are unpredictable.

B.2.4.3 Arithmetic conditioning table-specification syntax

Figure B.8 specifies the marker segment which defines one or more arithmetic coding conditioning table specifications. These replace the default arithmetic coding conditioning tables established by the SOI marker for arithmetic coding processes. (See F.1.4.4.1.4 and F.1.4.4.2.1.)

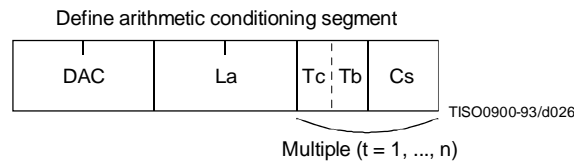


Figure B.8 – Arithmetic conditioning table-specification syntax

The marker and parameters shown in Figure B.8 are defined below. The size and allowed values of each parameter are given in Table B.6.

DAC: Define arithmetic coding conditioning marker – Marks the beginning of the definition of arithmetic coding conditioning parameters.

La: Arithmetic coding conditioning definition length – Specifies the length of all arithmetic coding conditioning parameters shown in Figure B.8 (see B.1.1.4).

Tc: Table class – 0 = DC table or lossless table, 1 = AC table.

Tb: Arithmetic coding conditioning table destination identifier – Specifies one of four possible destinations at the decoder into which the arithmetic coding conditioning table shall be installed.

Cs: Conditioning table value – Value in either the AC or the DC (and lossless) conditioning table. A single value of Cs shall follow each value of Tb. For AC conditioning tables Tc shall be one and Cs shall contain a value of Kx in the range $1 \leq Kx \leq 63$. For DC (and lossless) conditioning tables Tc shall be zero and Cs shall contain two 4-bit parameters, U and L. U and L shall be in the range $0 \leq L \leq U \leq 15$ and the value of Cs shall be $L + 16 \times U$.

The value n in Table B.6 is the number of arithmetic coding conditioning tables specified in the DAC marker segment. The parameters L and U are the lower and upper conditioning bounds used in the arithmetic coding procedures defined for DC coefficient coding and lossless coding. The separate value range 1-63 listed for DCT coding is the Kx conditioning used in AC coefficient coding.

Table B.6 – Arithmetic coding conditioning table-specification parameter sizes and values

Parameter	Size (bits)	Values			
		Sequential DCT		Progressive DCT	Lossless
		Baseline	Extended		
La	16	Undefined	$2 + 2 \times n$		
Tc	4	Undefined	0, 1	0	
Tb	4	Undefined	0-3		
Cs	8	Undefined	0-255 (Tc = 0), 1-63 (Tc = 1)	0-255	

B.2.4.4 Restart interval definition syntax

Figure B.9 specifies the marker segment which defines the restart interval.

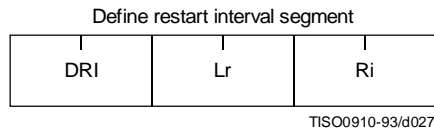


Figure B.9 – Restart interval definition syntax

The marker and parameters shown in Figure B.9 are defined below. The size and allowed values of each parameter are given in Table B.7.

DRI: Define restart interval marker – Marks the beginning of the parameters which define the restart interval.

Lr: Define restart interval segment length – Specifies the length of the parameters in the DRI segment shown in Figure B.9 (see B.1.1.4).

Ri: Restart interval – Specifies the number of MCU in the restart interval.

In Table B.7 the value *n* is the number of rows of MCU in the restart interval. The value *MCUR* is the number of MCU required to make up one line of samples of each component in the scan. The SOI marker disables the restart intervals. A DRI marker segment with *Ri* nonzero shall be present to enable restart interval processing for the following scans. A DRI marker segment with *Ri* equal to zero shall disable restart intervals for the following scans.

Table B.7 – Define restart interval segment parameter sizes and values

Parameter	Size (bits)	Values			
		Sequential DCT		Progressive DCT	Lossless
		Baseline	Extended		
Lr	16	4			
Ri	16	0-65 535			<i>n</i> × <i>MCUR</i>

B.2.4.5 Comment syntax

Figure B.10 specifies the marker segment structure for a comment segment.

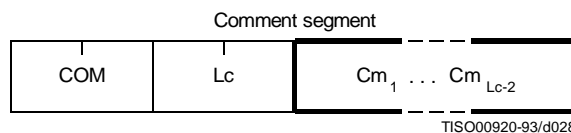


Figure B.10 – Comment segment syntax

The marker and parameters shown in Figure B.10 are defined below. The size and allowed values of each parameter are given in Table B.8.

COM: Comment marker – Marks the beginning of a comment.

Lc: Comment segment length – Specifies the length of the comment segment shown in Figure B.10 (see B.1.1.4).

Cm_i: Comment byte – The interpretation is left to the application.

Table B.8 – Comment segment parameter sizes and values

Parameter	Size (bits)	Values			
		Sequential DCT		Progressive DCT	Lossless
		Baseline	Extended		
Lc	16	2-65 535			
Cm _i	8	0-255			

B.2.4.6 Application data syntax

Figure B.11 specifies the marker segment structure for an application data segment.

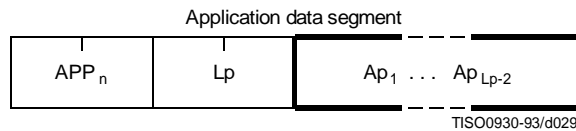


Figure B.11 – Application data syntax

The marker and parameters shown in Figure B.11 are defined below. The size and allowed values of each parameter are given in Table B.9.

APP_n: Application data marker – Marks the beginning of an application data segment.

Lp: Application data segment length – Specifies the length of the application data segment shown in Figure B.11 (see B.1.1.4).

Ap_i: Application data byte – The interpretation is left to the application.

The APP_n (Application) segments are reserved for application use. Since these segments may be defined differently for different applications, they should be removed when the data are exchanged between application environments.

Table B.9 – Application data segment parameter sizes and values

Parameter	Size (bits)	Values			
		Sequential DCT		Progressive DCT	Lossless
		Baseline	Extended		
Lp	16	2-65 535			
Ap _i	8	0-255			

B.2.5 Define number of lines syntax

Figure B.12 specifies the marker segment for defining the number of lines. The DNL (Define Number of Lines) segment provides a mechanism for defining or redefining the number of lines in the frame (the Y parameter in the frame header) at the end of the first scan. The value specified shall be consistent with the number of MCU-rows encoded in the first scan. This segment, if used, shall only occur at the end of the first scan, and only after coding of an integer number of MCU-rows. This marker segment is mandatory if the number of lines (Y) specified in the frame header has the value zero.

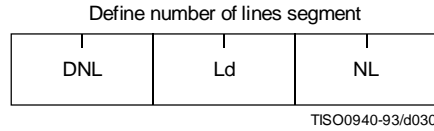


Figure B.12 – Define number of lines syntax

The marker and parameters shown in Figure B.12 are defined below. The size and allowed values of each parameter are given in Table B.10.

DNL: Define number of lines marker – Marks the beginning of the define number of lines segment.

Ld: Define number of lines segment length – Specifies the length of the define number of lines segment shown in Figure B.12 (see B.1.1.4).

NL: Number of lines – Specifies the number of lines in the frame (see definition of Y in B.2.2).

Table B.10 – Define number of lines segment parameter sizes and values

Parameter	Size (bits)	Values			
		Sequential DCT		Progressive DCT	Lossless
		Baseline	Extended		
Ld	16	4			
NL	16	1-65 535 ^{a)}			
^{a)} The value specified shall be consistent with the number of lines coded at the point where the DNL segment terminates the compressed data segment.					

B.3 Hierarchical syntax

B.3.1 High level hierarchical mode syntax

Figure B.13 specifies the order of the high level constituent parts of the interchange format for hierarchical encoding processes.



Figure B.13 – Syntax for the hierarchical mode of operation

Hierarchical mode syntax requires a DHP marker segment that appears before the non-differential frame or frames. The hierarchical mode compressed image data may include EXP marker segments and differential frames which shall follow the initial non-differential frame. The frame structure in hierarchical mode is identical to the frame structure in non-hierarchical mode.

The non-differential frames in the hierarchical sequence shall use one of the coding processes specified for SOF_n markers: SOF₀, SOF₁, SOF₂, SOF₃, SOF₉, SOF₁₀ and SOF₁₁. The differential frames shall use one of the processes specified for SOF₅, SOF₆, SOF₇, SOF₁₃, SOF₁₄ and SOF₁₅. The allowed combinations of SOF markers within one hierarchical sequence are specified in Annex J.

The sample precision (P) shall be constant for all frames and have the identical value as that coded in the DHP marker segment. The number of samples per line (X) for all frames shall not exceed the value coded in the DHP marker segment. If the number of lines (Y) is non-zero in the DHP marker segment, then the number of lines for all frames shall not exceed the value in the DHP marker segment.

B.3.2 DHP segment syntax

The DHP segment defines the image components, size, and sampling factors for the completed hierarchical sequence of frames. The DHP segment shall precede the first frame; a single DHP segment shall occur in the compressed image data.

The DHP segment structure is identical to the frame header syntax, except that the DHP marker is used instead of the SOF_n marker. The figures and description of B.2.2 then apply, except that the quantization table destination selector parameter shall be set to zero in the DHP segment.

B.3.3 EXP segment syntax

Figure B.14 specifies the marker segment structure for the EXP segment. The EXP segment shall be present if (and only if) expansion of the reference components is required either horizontally or vertically. The EXP segment parameters apply only to the next frame (which shall be a differential frame) in the image. If required, the EXP segment shall be one of the table-specification segments or miscellaneous marker segments preceding the frame header; the EXP segment shall not be one of the table-specification segments or miscellaneous marker segments preceding a scan header or a DHP marker segment.

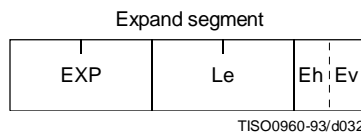


Figure B.14 – Syntax of the expand segment

The marker and parameters shown in Figure B.14 are defined below. The size and allowed values of each parameter are given in Table B.11.

EXP: Expand reference components marker – Marks the beginning of the expand reference components segment.

Le: Expand reference components segment length – Specifies the length of the expand reference components segment (see B.1.1.4).

Eh: Expand horizontally – If one, the reference components shall be expanded horizontally by a factor of two. If horizontal expansion is not required, the value shall be zero.

Ev: Expand vertically – If one, the reference components shall be expanded vertically by a factor of two. If vertical expansion is not required, the value shall be zero.

Both Eh and Ev shall be one if expansion is required both horizontally and vertically.

Table B.11 – Expand segment parameter sizes and values

Parameter	Size (bits)	Values			
		Sequential DCT		Progressive DCT	Lossless
		Baseline	Extended		
Le	16	3			
Eh	4	0, 1			
Ev	4	0, 1			

B.4 Abbreviated format for compressed image data

Figure B.2 shows the high-level constituent parts of the interchange format. This format includes all table specifications required for decoding. If an application environment provides methods for table specification other than by means of the compressed image data, some or all of the table specifications may be omitted. Compressed image data which is missing any table specification data required for decoding has the abbreviated format.

B.5 Abbreviated format for table-specification data

Figure B.2 shows the high-level constituent parts of the interchange format. If no frames are present in the compressed image data, the only purpose of the compressed image data is to convey table specifications or miscellaneous marker segments defined in B.2.4.1, B.2.4.2, B.2.4.5, and B.2.4.6. In this case the compressed image data has the abbreviated format for table specification data (see Figure B.15).

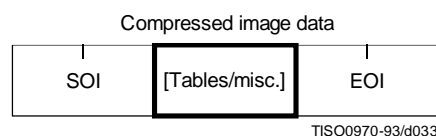
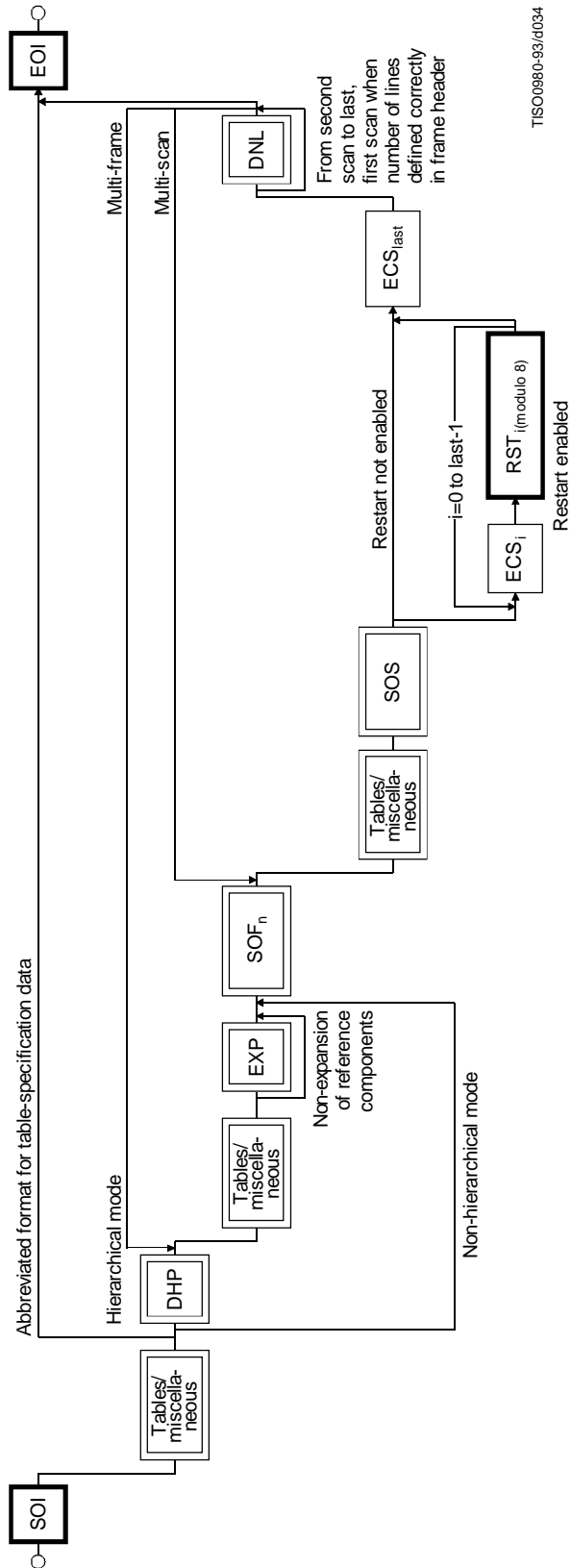


Figure B.15 – Abbreviated format for table-specification data syntax

B.6 Summary

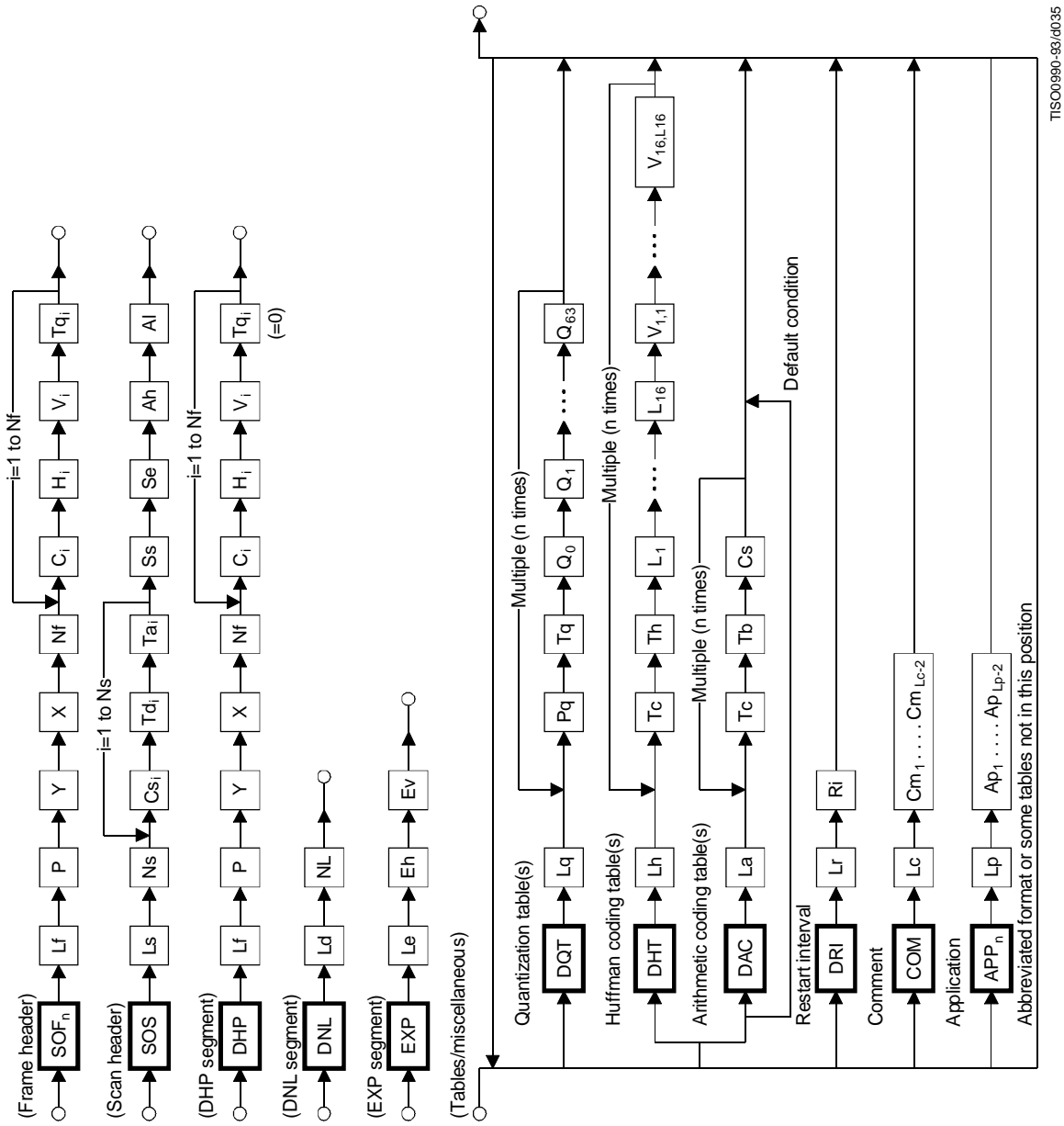
The order of the constituent parts of interchange format and all marker segment structures is summarized in Figures B.16 and B.17. Note that in Figure B.16 double-lined boxes enclose marker segments. In Figures B.16 and B.17 thick-lined boxes enclose only markers.

The EXP segment can be mixed with the other tables/miscellaneous marker segments preceding the frame header but not with the tables/miscellaneous marker segments preceding the DHP segment or the scan header.



TISO0980-93/d034

Figure B.16 – Flow of compressed data syntax



TISCO990-93/4035

Figure B.17 – Flow of marker segment

Annex C

Huffman table specification

(This annex forms an integral part of this Recommendation | International Standard)

A Huffman coding procedure may be used for entropy coding in any of the coding processes. Coding models for Huffman encoding are defined in Annexes F, G, and H. In this Annex, the Huffman table specification is defined.

Huffman tables are specified in terms of a 16-byte list (BITS) giving the number of codes for each code length from 1 to 16. This is followed by a list of the 8-bit symbol values (HUFFVAL), each of which is assigned a Huffman code. The symbol values are placed in the list in order of increasing code length. Code lengths greater than 16 bits are not allowed. In addition, the codes shall be generated such that the all-1-bits code word of any length is reserved as a prefix for longer code words.

NOTE – The order of the symbol values within HUFFVAL is determined only by code length. Within a given code length the ordering of the symbol values is arbitrary.

This annex specifies the procedure by which the Huffman tables (of Huffman code words and their corresponding 8-bit symbol values) are derived from the two lists (BITS and HUFFVAL) in the interchange format. However, the way in which these lists are generated is not specified. The lists should be generated in a manner which is consistent with the rules for Huffman coding, and it shall observe the constraints discussed in the previous paragraph. Annex K contains an example of a procedure for generating lists of Huffman code lengths and values which are in accord with these rules.

NOTE – There is **no requirement** in this Specification that any encoder or decoder shall implement the procedures in precisely the manner specified by the flow charts in this annex. It is necessary only that an encoder or decoder implement the **function** specified in this annex. The sole criterion for an encoder or decoder to be considered in compliance with this Specification is that it satisfy the requirements given in clause 6 (for encoders) or clause 7 (for decoders), as determined by the compliance tests specified in Part 2.

C.1 Marker segments for Huffman table specification

The DHT marker identifies the start of Huffman table definitions within the compressed image data. B.2.4.2 specifies the syntax for Huffman table specification.

C.2 Conversion of Huffman table specifications to tables of codes and code lengths

Conversion of Huffman table specifications to tables of codes and code lengths uses three procedures. The first procedure (Figure C.1) generates a table of Huffman code sizes. The second procedure (Figure C.2) generates the Huffman codes from the table built in Figure C.1. The third procedure (Figure C.3) generates the Huffman codes in symbol value order.

Given a list BITS (1 to 16) containing the number of codes of each size, and a list HUFFVAL containing the symbol values to be associated with those codes as described above, two tables are generated. The HUFFSIZE table contains a list of code lengths; the HUFFCODE table contains the Huffman codes corresponding to those lengths.

Note that the variable LASTK is set to the index of the last entry in the table.

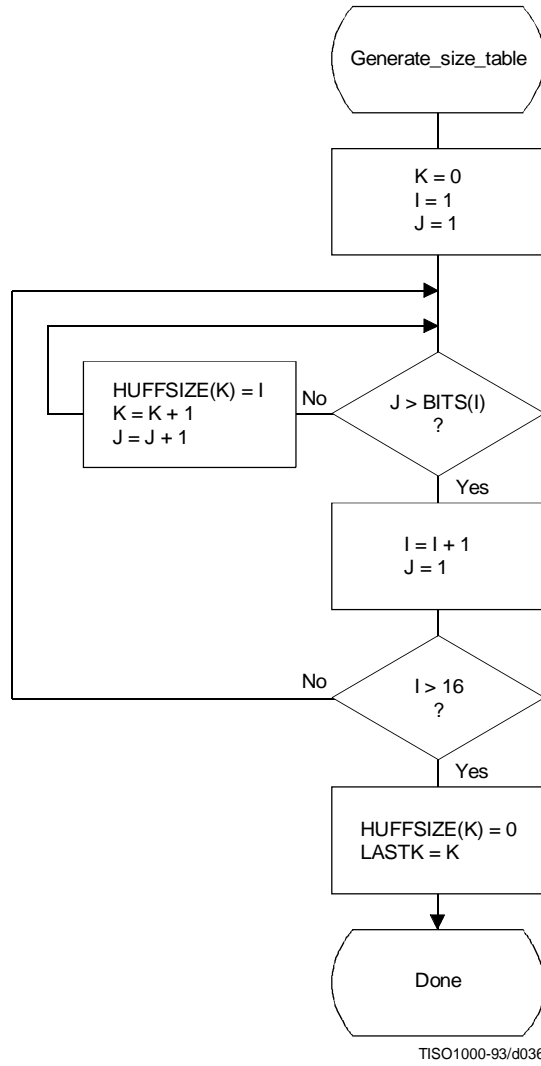


Figure C.1 – Generation of table of Huffman code sizes

A Huffman code table, HUFFCODE, containing a code for each size in HUFFSIZE is generated by the procedure in Figure C.2. The notation "SLL CODE 1" in Figure C.2 indicates a shift-left-logical of CODE by one bit position.

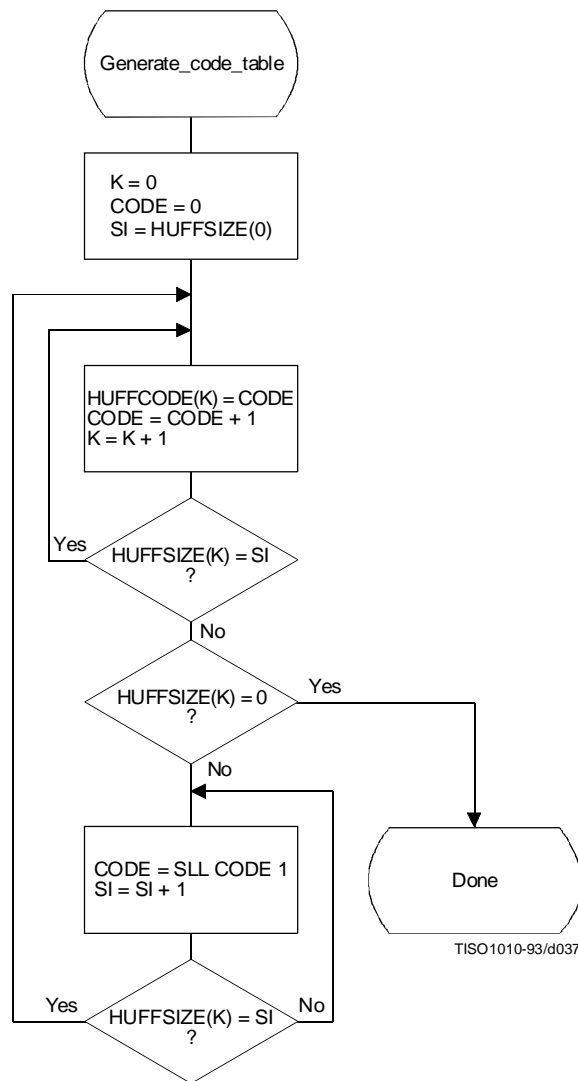


Figure C.2 – Generation of table of Huffman codes

Two tables, HUFFCODE and HUFFSIZE, have now been generated. The entries in the tables are ordered according to increasing Huffman code numeric value and length.

The encoding procedure code tables, EHUFCA and EHUFSA, are created by reordering the codes specified by HUFFCODE and HUFFSIZE according to the symbol values assigned to each code in HUFFVAL.

Figure C.3 illustrates this ordering procedure.

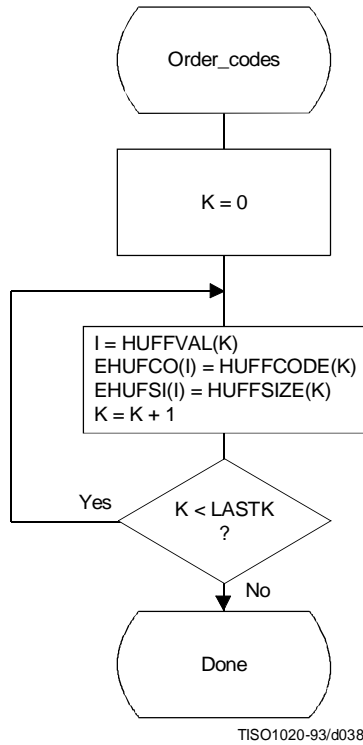


Figure C.3 – Ordering procedure for encoding procedure code tables

C.3 Bit ordering within bytes

The root of a Huffman code is placed toward the MSB (most-significant-bit) of the byte, and successive bits are placed in the direction MSB to LSB (least-significant-bit) of the byte. Remaining bits, if any, go into the next byte following the same rules.

Integers associated with Huffman codes are appended with the MSB adjacent to the LSB of the preceding Huffman code.

Annex D
Arithmetic coding

(This annex forms an integral part of this Recommendation | International Standard)

An adaptive binary arithmetic coding procedure may be used for entropy coding in any of the coding processes except the baseline sequential process. Coding models for adaptive binary arithmetic coding are defined in Annexes F, G, and H. In this annex the arithmetic encoding and decoding procedures used in those models are defined.

In K.4 a simple test example is given which should be helpful in determining if a given implementation is correct.

NOTE – There is **no requirement** in this Specification that any encoder or decoder shall implement the procedures in precisely the manner specified by the flow charts in this annex. It is necessary only that an encoder or decoder implement the **function** specified in this annex. The sole criterion for an encoder or decoder to be considered in compliance with this Specification is that it satisfy the requirements given in clause 6 (for encoders) or clause 7 (for decoders), as determined by the compliance tests specified in Part 2.

D.1 Arithmetic encoding procedures

Four arithmetic encoding procedures are required in a system with arithmetic coding (see Table D.1).

Table D.1 – Procedures for binary arithmetic encoding

Procedure	Purpose
Code_0(S)	Code a “0” binary decision with context-index S
Code_1(S)	Code a “1” binary decision with context-index S
Initenc	Initialize the encoder
Flush	Terminate entropy-coded segment

The “Code_0(S)” and “Code_1(S)” procedures code the 0-decision and 1-decision respectively; S is a context-index which identifies a particular conditional probability estimate used in coding the binary decision. The “Initenc” procedure initializes the arithmetic coding entropy encoder. The “Flush” procedure terminates the entropy-coded segment in preparation for the marker which follows.

D.1.1 Binary arithmetic encoding principles

The arithmetic coder encodes a series of binary symbols, zeros and ones, each symbol representing one possible result of a binary decision.

Each “binary decision” provides a choice between two alternatives. The binary decision might be between positive and negative signs, a magnitude being zero or nonzero, or a particular bit in a sequence of binary digits being zero or one.

The output bit stream (entropy-coded data segment) represents a binary fraction which increases in precision as bytes are appended by the encoding process.

D.1.1.1 Recursive interval subdivision

Recursive probability interval subdivision is the basis for the binary arithmetic encoding procedures. With each binary decision the current probability interval is subdivided into two sub-intervals, and the bit stream is modified (if necessary) so that it points to the base (the lower bound) of the probability sub-interval assigned to the symbol which occurred.

In the partitioning of the current probability interval into two sub-intervals, the sub-interval for the less probable symbol (LPS) and the sub-interval for the more probable symbol (MPS) are ordered such that usually the MPS sub-interval is closer to zero. Therefore, when the LPS is coded, the MPS sub-interval size is added to the bit stream. This coding convention requires that symbols be recognized as either MPS or LPS rather than 0 or 1. Consequently, the size of the LPS sub-interval and the sense of the MPS for each decision must be known in order to encode that decision.

The subdivision of the current probability interval would ideally require a multiplication of the interval by the probability estimate for the LPS. Because this subdivision is done approximately, it is possible for the LPS sub-interval to be larger than the MPS sub-interval. When that happens a “conditional exchange” interchanges the assignment of the sub-intervals such that the MPS is given the larger sub-interval.

Since the encoding procedure involves addition of binary fractions rather than concatenation of integer code words, the more probable binary decisions can sometimes be coded at a cost of much less than one bit per decision.

D.1.1.2 Conditioning of probability estimates

An adaptive binary arithmetic coder requires a statistical model – a model for selecting conditional probability estimates to be used in the coding of each binary decision. When a given binary decision probability estimate is dependent on a particular feature or features (the context) already coded, it is “conditioned” on that feature. The conditioning of probability estimates on previously coded decisions must be identical in encoder and decoder, and therefore can use only information known to both.

Each conditional probability estimate required by the statistical model is kept in a separate storage location or “bin” identified by a unique context-index *S*. The arithmetic coder is adaptive, which means that the probability estimates at each context-index are developed and maintained by the arithmetic coding system on the basis of prior coding decisions for that context-index.

D.1.2 Encoding conventions and approximations

The encoding procedures use fixed precision integer arithmetic and an integer representation of fractional values in which X'8000' can be regarded as the decimal value 0.75. The probability interval, *A*, is kept in the integer range X'8000' ≤ *A* < X'10000' by doubling it whenever its integer value falls below X'8000'. This is equivalent to keeping *A* in the decimal range 0.75 ≤ *A* < 1.5. This doubling procedure is called renormalization.

The code register, *C*, contains the trailing bits of the bit stream. *C* is also doubled each time *A* is doubled. Periodically – to keep *C* from overflowing – a byte of data is removed from the high order bits of the *C*-register and placed in the entropy-coded segment.

Carry-over into the entropy-coded segment is limited by delaying X'FF' output bytes until the carry-over is resolved. Zero bytes are stuffed after each X'FF' byte in the entropy-coded segment in order to avoid the accidental generation of markers in the entropy-coded segment.

Keeping *A* in the range 0.75 ≤ *A* < 1.5 allows a simple arithmetic approximation to be used in the probability interval subdivision. Normally, if the current estimate of the LPS probability for context-index *S* is *Qe(S)*, precise calculation of the sub-intervals would require:

$$\begin{array}{ll} Qe(S) \times A & \text{Probability sub-interval for the LPS;} \\ A - (Qe(S) \times A) & \text{Probability sub-interval for the MPS.} \end{array}$$

Because the decimal value of *A* is of order unity, these can be approximated by

$$\begin{array}{ll} Qe(S) & \text{Probability sub-interval for the LPS;} \\ A - Qe(S) & \text{Probability sub-interval for the MPS.} \end{array}$$

Whenever the LPS is coded, the value of *A - Qe(S)* is added to the code register and the probability interval is reduced to *Qe(S)*. Whenever the MPS is coded, the code register is left unchanged and the interval is reduced to *A - Qe(S)*. The precision range required for *A* is then restored, if necessary, by renormalization of both *A* and *C*.

With the procedure described above, the approximations in the probability interval subdivision process can sometimes make the LPS sub-interval larger than the MPS sub-interval. If, for example, the value of *Qe(S)* is 0.5 and *A* is at the minimum allowed value of 0.75, the approximate scaling gives one-third of the probability interval to the MPS and two-thirds to the LPS. To avoid this size inversion, conditional exchange is used. The probability interval is subdivided using the simple approximation, but the MPS and LPS sub-interval assignments are exchanged whenever the LPS sub-interval is larger than the MPS sub-interval. This MPS/LPS conditional exchange can only occur when a renormalization will be needed.

Each binary decision uses a context. A context is the set of prior coding decisions which determine the context-index, *S*, identifying the probability estimate used in coding the decision.

Whenever a renormalization occurs, a probability estimation procedure is invoked which determines a new probability estimate for the context currently being coded. No explicit symbol counts are needed for the estimation. The relative probabilities of renormalization after coding of LPS and MPS provide, by means of a table-based probability estimation state machine, a direct estimate of the probabilities.

D.1.3 Encoder code register conventions

The flow charts in this annex assume the register structures for the encoder as shown in Table D.2.

Table D.2 – Encoder register connections

	MSB		LSB	
C-register	0000cbbb,	bbbbssss,	xxxxxxxx,	xxxxxxxx
A-register	00000000,	00000000,	aaaaaaaa,	aaaaaaaa

The “a” bits are the fractional bits in the A-register (the current probability interval value) and the “x” bits are the fractional bits in the code register. The “s” bits are optional spacer bits which provide useful constraints on carry-over, and the “b” bits indicate the bit positions from which the completed bytes of data are removed from the C-register. The “c” bit is a carry bit. Except at the time of initialization, bit 15 of the A-register is always set and bit 16 is always clear (the LSB is bit 0).

These register conventions illustrate one possible implementation. However, any register conventions which allow resolution of carry-over in the encoder and which produce the same entropy-coded segment may be used. The handling of carry-over and the byte stuffing following X'FF' will be described in a later part of this annex.

D.1.4 Code_1(S) and Code_0(S) procedures

When a given binary decision is coded, one of two possibilities occurs – either a 1-decision or a 0-decision is coded. Code_1(S) and Code_0(S) are shown in Figures D.1 and D.2. The Code_1(S) and Code_0(S) procedures use probability estimates with a context-index S. The context-index S is determined by the statistical model and is, in general, a function of the previous coding decisions; each value of S identifies a particular conditional probability estimate which is used in encoding the binary decision.

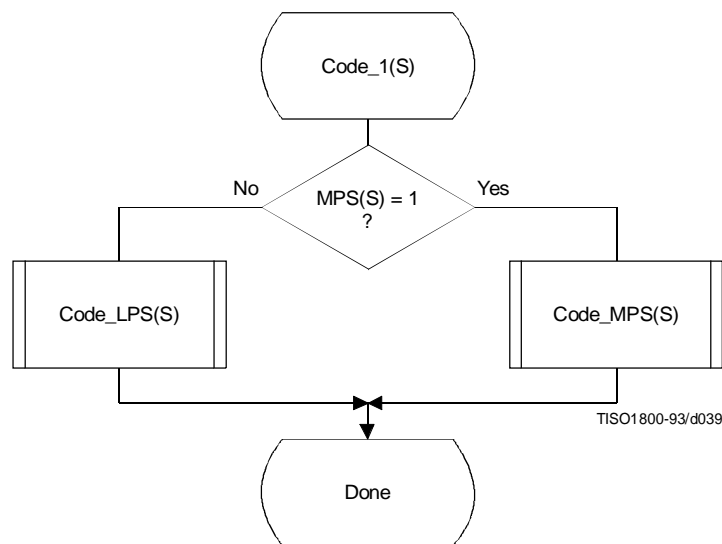


Figure D.1 – Code_1(S) procedure

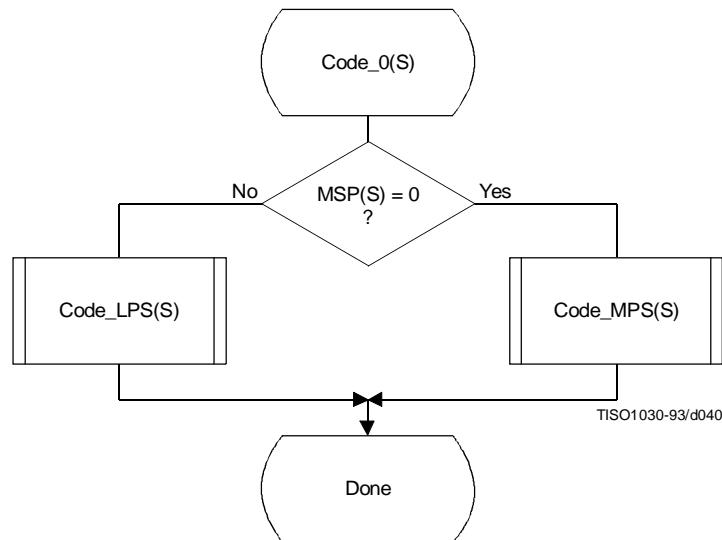


Figure D.2 – Code_0(S) procedure

The context-index S selects a storage location which contains $\text{Index}(S)$, an index to the tables which make up the probability estimation state machine. When coding a binary decision, the symbol being coded is either the more probable symbol or the less probable symbol. Therefore, additional information is stored at each context-index identifying the sense of the more probable symbol, $\text{MPS}(S)$.

For simplicity, the flow charts in this subclause assume that the context storage for each context-index S has an additional storage field for $Q_e(S)$ containing the value of $Q_e(\text{Index}(S))$. If only the value of $\text{Index}(S)$ and $\text{MPS}(S)$ are stored, all references to $Q_e(S)$ should be replaced by $Q_e(\text{Index}(S))$.

The $\text{Code_LPS}(S)$ procedure normally consists of the addition of the MPS sub-interval $A - Q_e(S)$ to the bit stream and a scaling of the interval to the sub-interval, $Q_e(S)$. It is always followed by the procedures for obtaining a new LPS probability estimate ($\text{Estimate_}Q_e(S)_\text{after_LPS}$) and renormalization (Renorm_e) (see Figure D.3).

However, in the event that the LPS sub-interval is larger than the MPS sub-interval, the conditional MPS/LPS exchange occurs and the MPS sub-interval is coded.

The $\text{Code_MPS}(S)$ procedure normally reduces the size of the probability interval to the MPS sub-interval. However, if the LPS sub-interval is larger than the MPS sub-interval, the conditional exchange occurs and the LPS sub-interval is coded instead. Note that conditional exchange cannot occur unless the procedures for obtaining a new LPS probability estimate ($\text{Estimate_}Q_e(S)_\text{after_MPS}$) and renormalization (Renorm_e) are required after the coding of the symbol (see Figure D.4).

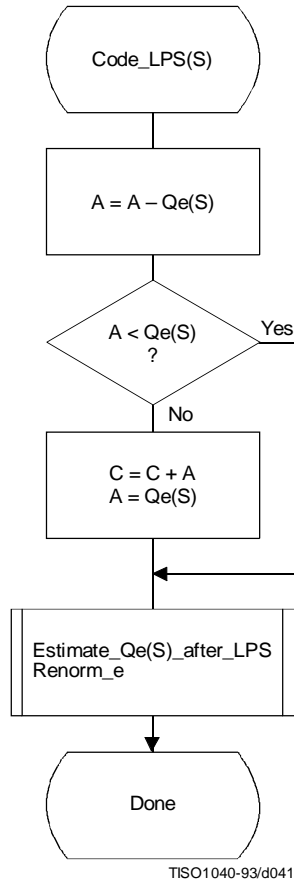
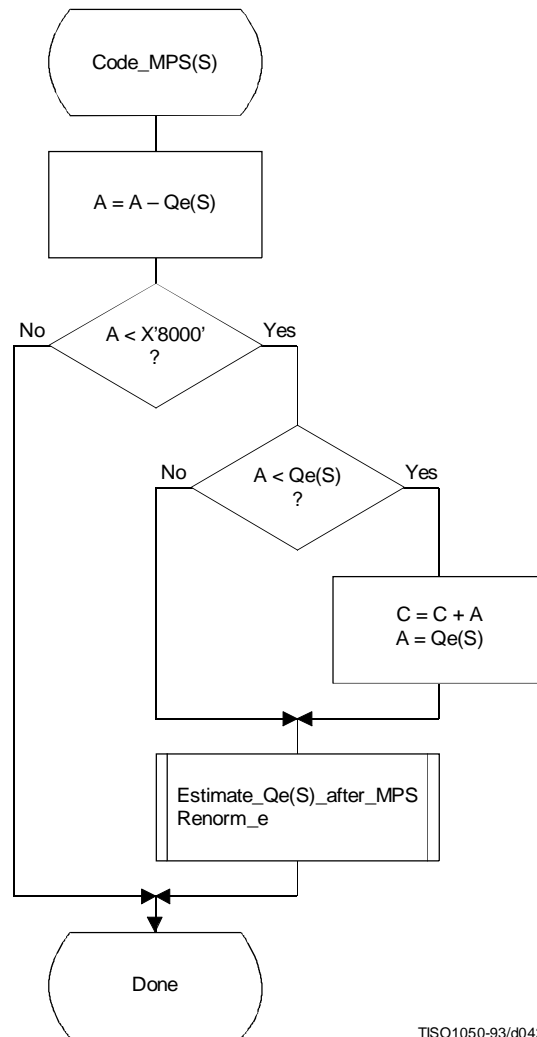


Figure D.3 – Code_LPS(S) procedure with conditional MPS/LPS exchange



TISO1050-93/d042

Figure D.4 – Code_MPS(S) procedure with conditional MPS/LPS exchange

D.1.5 Probability estimation in the encoder

D.1.5.1 Probability estimation state machine

The probability estimation state machine consists of a number of sequences of probability estimates. These sequences are interlinked in a manner which provides probability estimates based on approximate symbol counts derived from the arithmetic coder renormalization. Some of these sequences are used during the initial “learning” stages of probability estimation; the rest are used for “steady state” estimation.

Each entry in the probability estimation state machine is assigned an index, and each index has associated with it a Q_e value and two Next_Index values. The Next_Index_MPS gives the index to the new probability estimate after an MPS renormalization; the Next_Index_LPS gives the index to the new probability estimate after an LPS renormalization. Note that both the index to the estimation state machine and the sense of the MPS are kept for each context-index S. The sense of the MPS is changed whenever the entry in the Switch_MPS is one.

The probability estimation state machine is given in Table D.3. Initialization of the arithmetic coder is always with an MPS sense of zero and a Q_e index of zero in Table D.3.

The Q_e values listed in Table D.3 are expressed as hexadecimal integers. To approximately convert the 15-bit integer representation of Q_e to a decimal probability, divide the Q_e values by $(4/3) \times (X'8000)$.

Table D.3 – Qe values and probability estimation state machine

Index	Qe _Value	Next_Index		Switch _MPS	Index	Qe _Value	Next_Index		Switch _MPS
		_LPS	_MPS				_LPS	_MPS	
0	X'5A1D'	1	1	1	57	X'01A4'	55	58	0
1	X'2586'	14	2	0	58	X'0160'	56	59	0
2	X'1114'	16	3	0	59	X'0125'	57	60	0
3	X'080B'	18	4	0	60	X'00F6'	58	61	0
4	X'03D8'	20	5	0	61	X'00CB'	59	62	0
5	X'01DA'	23	6	0	62	X'00AB'	61	63	0
6	X'00E5'	25	7	0	63	X'008F'	61	32	0
7	X'006F'	28	8	0	64	X'5B12'	65	65	1
8	X'0036'	30	9	0	65	X'4D04'	80	66	0
9	X'001A'	33	10	0	66	X'412C'	81	67	0
10	X'000D'	35	11	0	67	X'37D8'	82	68	0
11	X'0006'	9	12	0	68	X'2FE8'	83	69	0
12	X'0003'	10	13	0	69	X'293C'	84	70	0
13	X'0001'	12	13	0	70	X'2379'	86	71	0
14	X'5A7F'	15	15	1	71	X'1EDF'	87	72	0
15	X'3F25'	36	16	0	72	X'1AA9'	87	73	0
16	X'2CF2'	38	17	0	73	X'174E'	72	74	0
17	X'207C'	39	18	0	74	X'1424'	72	75	0
18	X'17B9'	40	19	0	75	X'119C'	74	76	0
19	X'1182'	42	20	0	76	X'0F6B'	74	77	0
20	X'0CEF'	43	21	0	77	X'0D51'	75	78	0
21	X'09A1'	45	22	0	78	X'0BB6'	77	79	0
22	X'072F'	46	23	0	79	X'0A40'	77	48	0
23	X'055C'	48	24	0	80	X'5832'	80	81	1
24	X'0406'	49	25	0	81	X'4D1C'	88	82	0
25	X'0303'	51	26	0	82	X'438E'	89	83	0
26	X'0240'	52	27	0	83	X'3BDD'	90	84	0
27	X'01B1'	54	28	0	84	X'34EE'	91	85	0
28	X'0144'	56	29	0	85	X'2EAE'	92	86	0
29	X'00F5'	57	30	0	86	X'299A'	93	87	0
30	X'00B7'	59	31	0	87	X'2516'	86	71	0
31	X'008A'	60	32	0	88	X'5570'	88	89	1
32	X'0068'	62	33	0	89	X'4CA9'	95	90	0
33	X'004E'	63	34	0	90	X'44D9'	96	91	0
34	X'003B'	32	35	0	91	X'3E22'	97	92	0
35	X'002C'	33	9	0	92	X'3824'	99	93	0
36	X'5AE1'	37	37	1	93	X'32B4'	99	94	0
37	X'484C'	64	38	0	94	X'2E17'	93	86	0
38	X'3A0D'	65	39	0	95	X'56A8'	95	96	1
39	X'2EF1'	67	40	0	96	X'4F46'	101	97	0
40	X'261F'	68	41	0	97	X'47E5'	102	98	0
41	X'1F33'	69	42	0	98	X'41CF'	103	99	0
42	X'19A8'	70	43	0	99	X'3C3D'	104	100	0
43	X'1518'	72	44	0	100	X'375E'	99	93	0
44	X'1177'	73	45	0	101	X'5231'	105	102	0
45	X'0E74'	74	46	0	102	X'4C0F'	106	103	0
46	X'0BFB'	75	47	0	103	X'4639'	107	104	0
47	X'09F8'	77	48	0	104	X'415E'	103	99	0
48	X'0861'	78	49	0	105	X'5627'	105	106	1
49	X'0706'	79	50	0	106	X'50E7'	108	107	0
50	X'05CD'	48	51	0	107	X'4B85'	109	103	0
51	X'04DE'	50	52	0	108	X'5597'	110	109	0
52	X'040F'	50	53	0	109	X'504F'	111	107	0
53	X'0363'	51	54	0	110	X'5A10'	110	111	1
54	X'02D4'	52	55	0	111	X'5522'	112	109	0
55	X'025C'	53	56	0	112	X'59EB'	112	111	1
56	X'01F8'	54	57	0					

D.1.5.2 Renormalization driven estimation

The change in state in Table D.3 occurs only when the arithmetic coder interval register is renormalized. This must always be done after coding an LPS, and whenever the probability interval register is less than X'8000' (0.75 in decimal notation) after coding an MPS.

When the LPS renormalization is required, Next_Index_LPS gives the new index for the LPS probability estimate. When the MPS renormalization is required, Next_Index_MPS gives the new index for the LPS probability estimate. If Switch_MPS is 1 for the old index, the MPS symbol sense must be inverted after an LPS.

D.1.5.3 Estimation following renormalization after MPS

The procedure for estimating the probability on the MPS renormalization path is given in Figure D.5. Index(S) is part of the information stored for context-index S. The new value of Index(S) is obtained from Table D.3 from the column labeled Next_Index_MPS, as that is the next index after an MPS renormalization. This next index is stored as the new value of Index(S) in the context storage at context-index S, and the value of Qe at this new Index(S) becomes the new Qe(S). MPS(S) does not change.

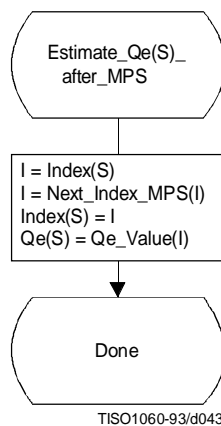


Figure D.5 – Probability estimation on MPS renormalization path

D.1.5.4 Estimation following renormalization after LPS

The procedure for estimating the probability on the LPS renormalization path is shown in Figure D.6. The procedure is similar to that of Figure D.5 except that when Switch_MPS(I) is 1, the sense of MPS(S) must be inverted.

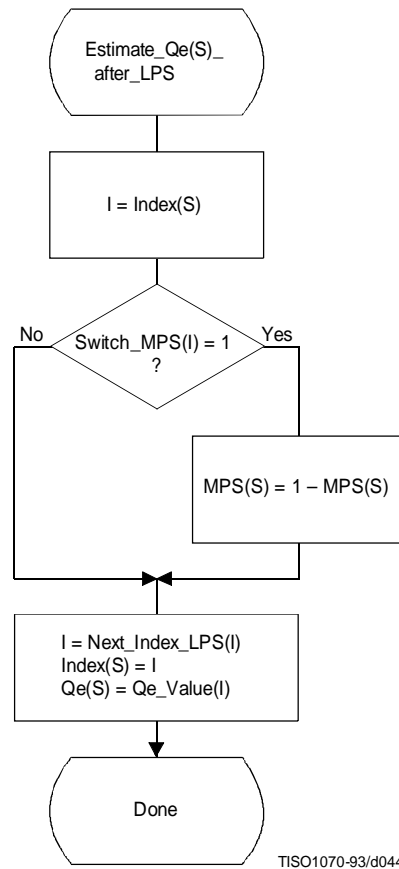


Figure D.6 – Probability estimation on LPS renormalization path

D.1.6 Renormalization in the encoder

The Renorm_e procedure for the encoder renormalization is shown in Figure D.7. Both the probability interval register A and the code register C are shifted, one bit at a time. The number of shifts is counted in the counter CT; when CT is zero, a byte of compressed data is removed from C by the procedure Byte_out and CT is reset to 8. Renormalization continues until A is no longer less than X'8000'.

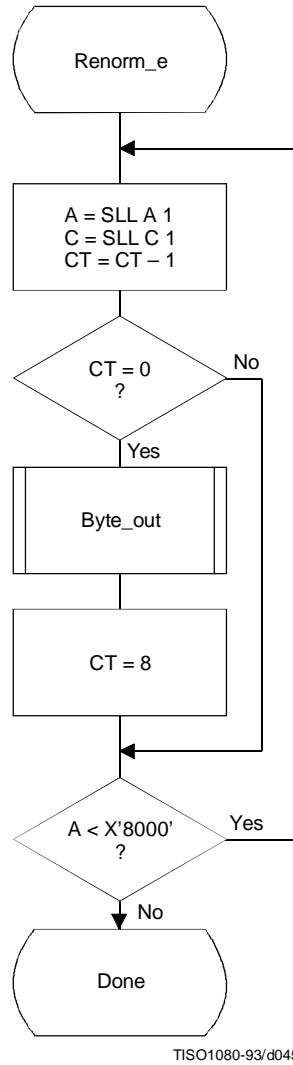


Figure D.7 – Encoder renormalization procedure

The Byte_out procedure used in Renorm_e is shown in Figure D.8. This procedure uses byte-stuffing procedures which prevent accidental generation of markers by the arithmetic encoding procedures. It also includes an example of a procedure for resolving carry-over. For simplicity of exposition, the buffer holding the entropy-coded segment is assumed to be large enough to contain the entire segment.

In Figure D.8 BP is the entropy-coded segment pointer and B is the compressed data byte pointed to by BP. T in Byte_out is a temporary variable which is used to hold the output byte and carry bit. ST is the stack counter which is used to count X'FF' output bytes until any carry-over through the X'FF' sequence has been resolved. The value of ST rarely exceeds 3. However, since the upper limit for the value of ST is bounded only by the total entropy-coded segment size, a precision of 32 bits is recommended for ST.

Since large values of ST represent a latent output of compressed data, the following procedure may be needed in high speed synchronous encoding systems for handling the burst of output data which occurs when the carry is resolved.

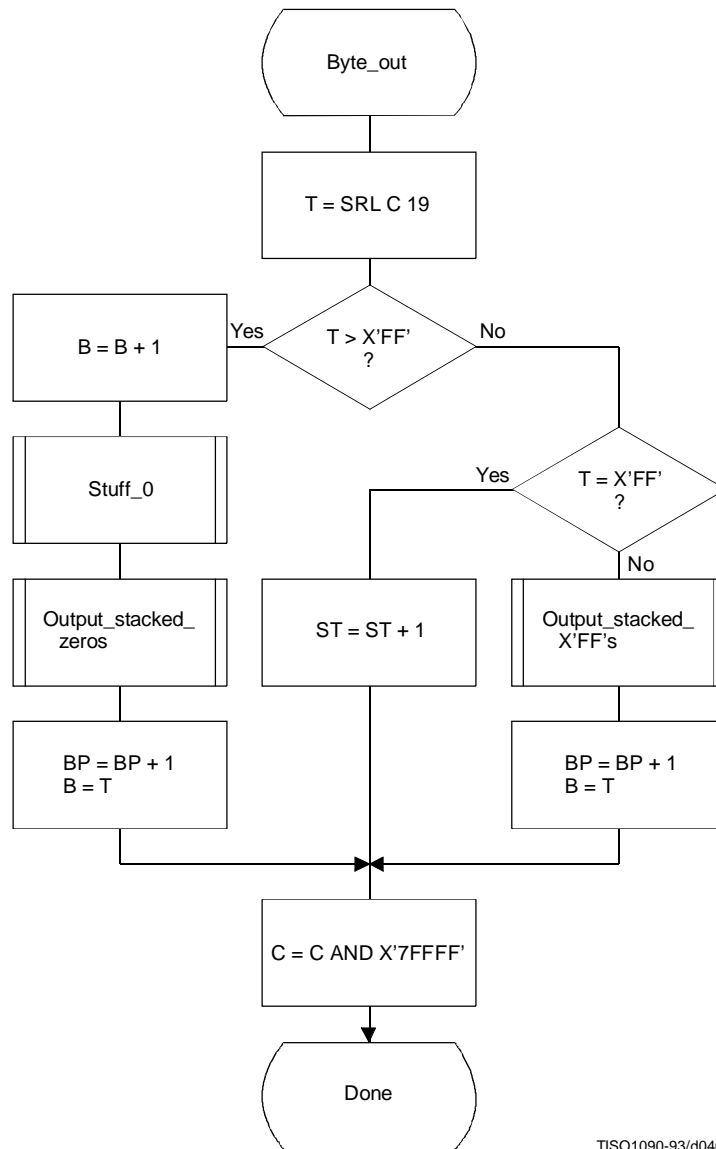


Figure D.8 – Byte_out procedure for encoder

When the stack count reaches an upper bound determined by output channel capacity, the stack is emptied and the stacked X'FF' bytes (and stuffed zero bytes) are added to the compressed data before the carry-over is resolved. If a carry-over then occurs, the carry is added to the final stuffed zero, thereby converting the final X'FF00' sequence to the X'FF01' temporary private marker. The entropy-coded segment must then be post-processed to resolve the carry-over and remove the temporary marker code. For any reasonable bound on ST this post processing is very unlikely.

Referring to Figure D.8, the shift of the code register by 19 bits aligns the output bits with the low order bits of T. The first test then determines if a carry-over has occurred. If so, the carry must be added to the previous output byte before advancing the segment pointer BP. The Stuff_0 procedure stuffs a zero byte whenever the addition of the carry to the data already in the entropy-coded segments creates a X'FF' byte. Any stacked output bytes – converted to zeros by the carry-over – are then placed in the entropy-coded segment. Note that when the output byte is later transferred from T to the entropy-coded segment (to byte B), the carry bit is ignored if it is set.

If a carry has not occurred, the output byte is tested to see if it is X'FF'. If so, the stack count ST is incremented, as the output must be delayed until the carry-over is resolved. If not, the carry-over has been resolved, and any stacked X'FF' bytes must then be placed in the entropy-coded segment. Note that a zero byte is stuffed following each X'FF'.

The procedures used by Byte_out are defined in Figures D.9 through D.11.

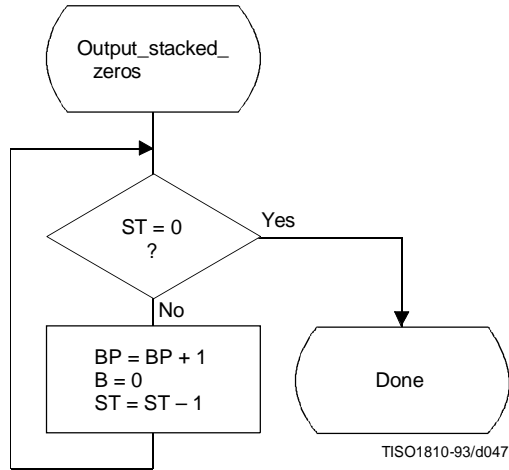


Figure D.9 – Output_stacked_zeros procedure for encoder

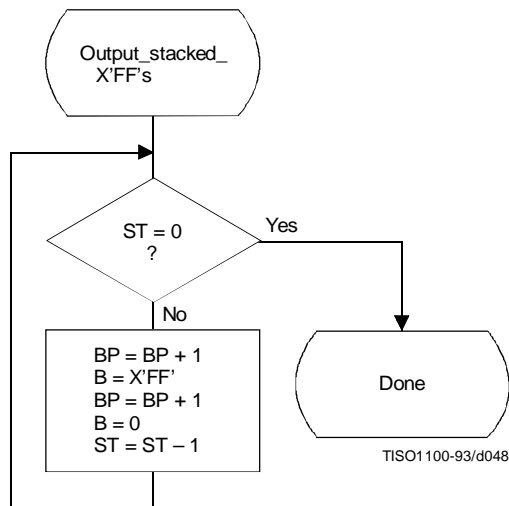


Figure D.10 – Output_stacked_X'FF's procedure for encoder

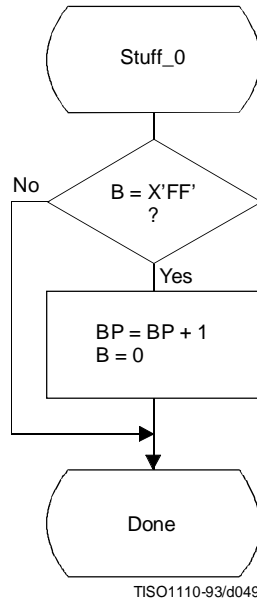


Figure D.11 – Stuff_0 procedure for encoder

D.1.7 Initialization of the encoder

The Initenc procedure is used to start the arithmetic coder. The basic steps are shown in Figure D.12.

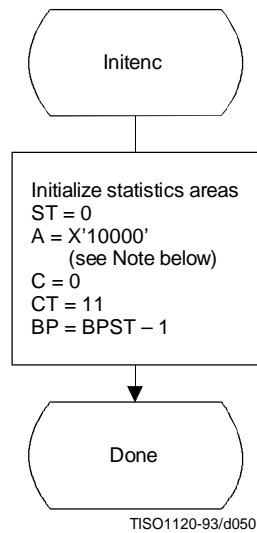


Figure D.12 – Initialization of the encoder

The probability estimation tables are defined by Table D.3. The statistics areas are initialized to an MPS sense of 0 and a Qe index of zero as defined by Table D.3. The stack count (ST) is cleared, the code register (C) is cleared, and the interval register is set to X'10000'. The counter (CT) is set to 11, reflecting the fact that when A is initialized to X'10000' three spacer bits plus eight output bits in C must be filled before the first byte is removed. Note that BP is initialized to point to the byte before the start of the entropy-coded segment (which is at BPST). Note also that the statistics areas are initialized for all values of context-index S to MPS(S) = 0 and Index(S) = 0.

NOTE – Although the probability interval is initialized to X'10000' in both Initenc and Initdec, the precision of the probability interval register can still be limited to 16 bits. When the precision of the interval register is 16 bits, it is initialized to zero.

D.1.8 Termination of encoding

The Flush procedure is used to terminate the arithmetic encoding procedures and prepare the entropy-coded segment for the addition of the X'FF' prefix of the marker which follows the arithmetically coded data. Figure D.13 shows this flush procedure. The first step in the procedure is to set as many low order bits of the code register to zero as possible without pointing outside of the final interval. Then, the output byte is aligned by shifting it left by CT bits; Byte_out then removes it from C. C is then shifted left by 8 bits to align the second output byte and Byte_out is used a second time. The remaining low order bits in C are guaranteed to be zero, and these trailing zero bits shall not be written to the entropy-coded segment.

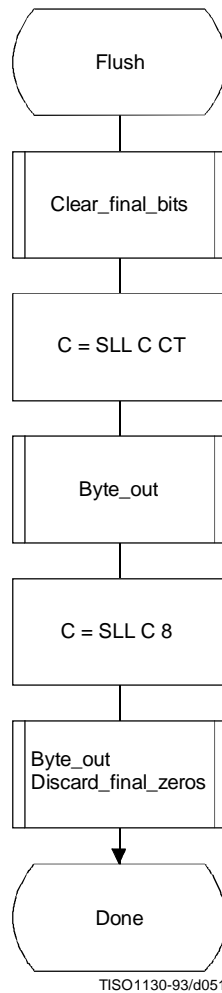


Figure D.13 – Flush procedure

Any trailing zero bytes already written to the entropy-coded segment and not preceded by a X'FF' may, optionally, be discarded. This is done in the Discard_final_zeros procedure. Stuffed zero bytes shall not be discarded.

Entropy coded segments are always followed by a marker. For this reason, the final zero bits needed to complete decoding shall not be included in the entropy coded segment. Instead, when the decoder encounters a marker, zero bits shall be supplied to the decoding procedure until decoding is complete. This convention guarantees that when a DNL marker is used, the decoder will intercept it in time to correctly terminate the decoding procedure.

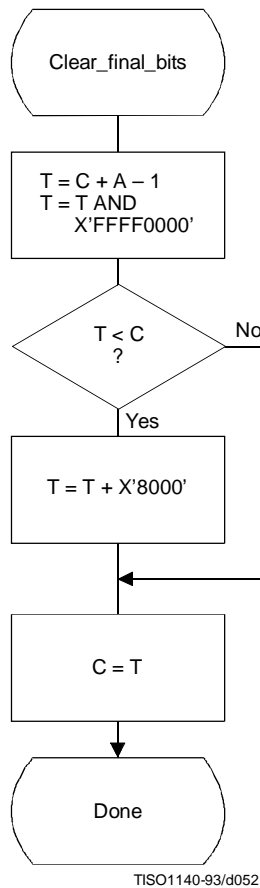


Figure D.14 – Clear_final_bits procedure in Flush

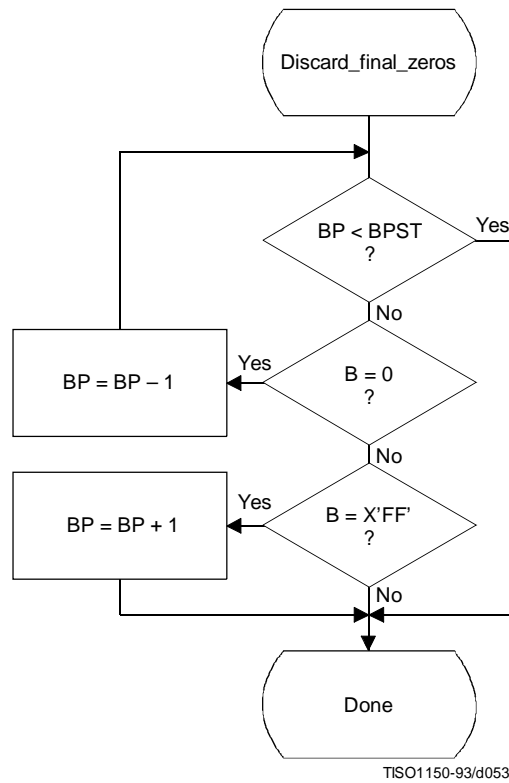


Figure D.15 – Discard_final_zeros procedure in Flush

D.2 Arithmetic decoding procedures

Two arithmetic decoding procedures are used for arithmetic decoding (see Table D.4).

The “Decode(S)” procedure decodes the binary decision for a given context-index S and returns a value of either 0 or 1. It is the inverse of the “Code_0(S)” and “Code_1(S)” procedures described in D.1. “Initdec” initializes the arithmetic coding entropy decoder.

Table D.4 – Procedures for binary arithmetic decoding

Procedure	Purpose
Decode(S)	Decode a binary decision with context-index S
Initdec	Initialize the decoder

D.2.1 Binary arithmetic decoding principles

The probability interval subdivision and sub-interval ordering defined for the arithmetic encoding procedures also apply to the arithmetic decoding procedures.

Since the bit stream always points within the current probability interval, the decoding process is a matter of determining, for each decision, which sub-interval is pointed to by the bit stream. This is done recursively, using the same probability interval sub-division process as in the encoder. Each time a decision is decoded, the decoder subtracts from the bit stream any interval the encoder added to the bit stream. Therefore, the code register in the decoder is a pointer into the current probability interval relative to the base of the interval.

If the size of the sub-interval allocated to the LPS is larger than the sub-interval allocated to the MPS, the encoder invokes the conditional exchange procedure. When the interval sizes are inverted in the decoder, the sense of the symbol decoded must be inverted.

D.2.2 Decoding conventions and approximations

The approximations and integer arithmetic defined for the probability interval subdivision in the encoder must also be used in the decoder. However, where the encoder would have added to the code register, the decoder subtracts from the code register.

D.2.3 Decoder code register conventions

The flow charts given in this section assume the register structures for the decoder as shown in Table D.5:

Table D.5 – Decoder register conventions

	MSB	LSB
Cx register	xxxxxxx,	xxxxxxx
C-low	bbbbbbb,	0000000
A-register	aaaaaaaa,	aaaaaaaa

Cx and C-low can be regarded as one 32-bit C-register, in that renormalization of C shifts a bit of new data from bit 15 of C-low to bit 0 of Cx. However, the decoding comparisons use Cx alone. New data are inserted into the “b” bits of C-low one byte at a time.

NOTE – The comparisons shown in the various procedures use arithmetic comparisons, and therefore assume precisions greater than 16 bits for the variables. Unsigned (logical) comparisons should be used in 16-bit precision implementations.

D.2.4 The decode procedure

The decoder decodes one binary decision at a time. After decoding the decision, the decoder subtracts any amount from the code register that the encoder added. The amount left in the code register is the offset from the base of the current probability interval to the sub-interval allocated to the binary decisions not yet decoded. In the first test in the decode procedure shown in Figure D.16 the code register is compared to the size of the MPS sub-interval. Unless a conditional exchange is needed, this test determines whether the MPS or LPS for context-index S is decoded. Note that the LPS for context-index S is given by $1 - \text{MPS}(S)$.

When a renormalization is needed, the MPS/LPS conditional exchange may also be needed. For the LPS path, the conditional exchange procedure is shown in Figure D.17. Note that the probability estimation in the decoder is identical to the probability estimation in the encoder (Figures D.5 and D.6).

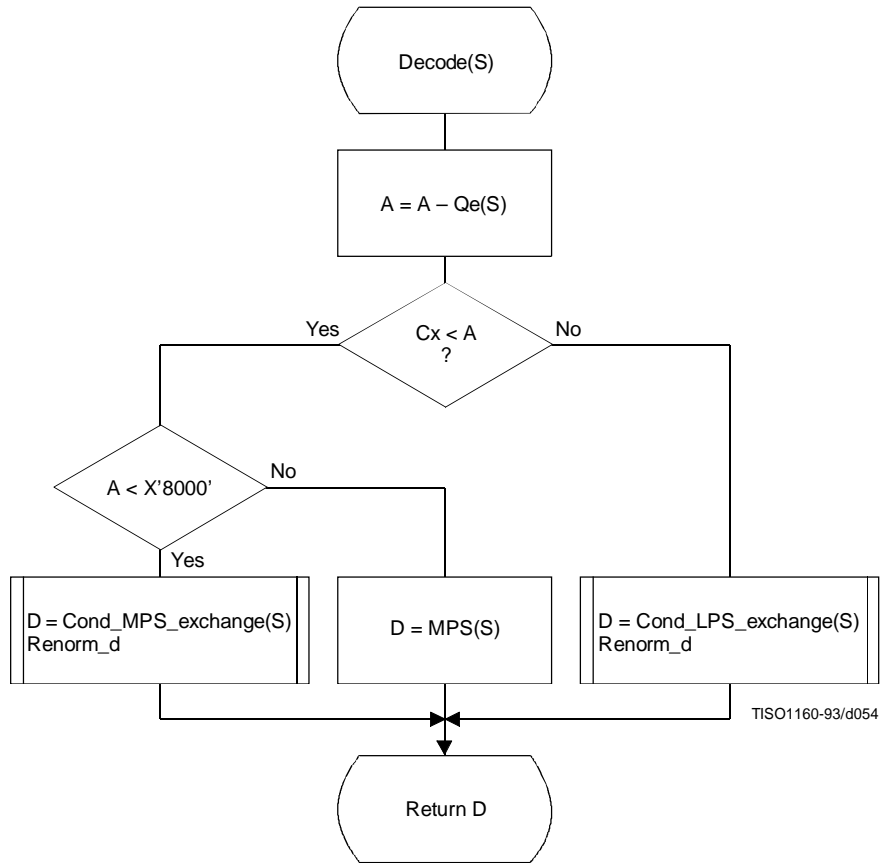


Figure D.16 – Decode(S) procedure

For the MPS path of the decoder the conditional exchange procedure is given in Figure D.18.

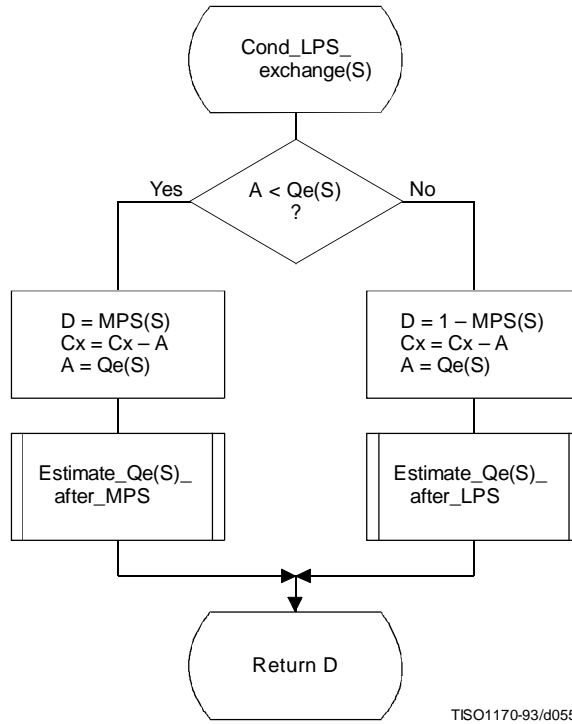


Figure D.17 – Decoder LPS path conditional exchange procedure

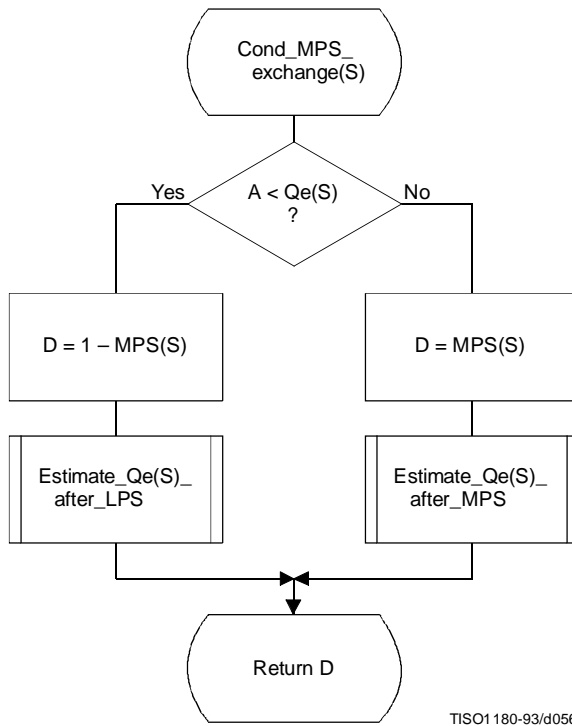


Figure D.18 – Decoder MPS path conditional exchange procedure

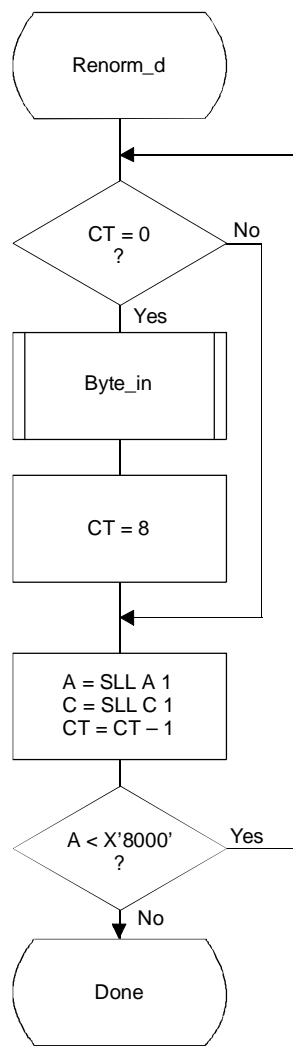
D.2.5 Probability estimation in the decoder

The procedures defined for obtaining a new LPS probability estimate in the encoder are also used in the decoder.

D.2.6 Renormalization in the decoder

The Renorm_d procedure for the decoder renormalization is shown in Figure D.19. CT is a counter which keeps track of the number of compressed bits in the C-low section of the C-register. When CT is zero, a new byte is inserted into C-low by the procedure Byte_in and CT is reset to 8.

Both the probability interval register A and the code register C are shifted, one bit at a time, until A is no longer less than X'8000'.



TISO1190-93/d057

Figure D.19 – Decoder renormalization procedure

The Byte_in procedure used in Renorm_d is shown in Figure D.20. This procedure fetches one byte of data, compensating for the stuffed zero byte which follows any X'FF' byte. It also detects the marker which must follow the entropy-coded segment. The C-register in this procedure is the concatenation of the Cx and C-low registers. For simplicity of exposition, the buffer holding the entropy-coded segment is assumed to be large enough to contain the entire segment.

B is the byte pointed to by the entropy-coded segment pointer BP. BP is first incremented. If the new value of B is not a X'FF', it is inserted into the high order 8 bits of C-low.

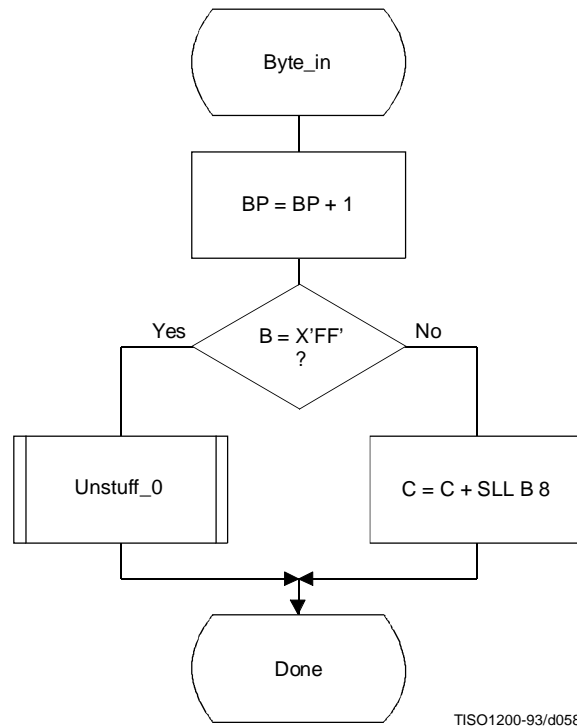


Figure D.20 – Byte_in procedure for decoder

The Unstuff_0 procedure is shown in Figure D.21. If the new value of B is X'FF', BP is incremented to point to the next byte and this next B is tested to see if it is zero. If so, B contains a stuffed byte which must be skipped. The zero B is ignored, and the X'FF' B value which preceded it is inserted in the C-register.

If the value of B after a X'FF' byte is not zero, then a marker has been detected. The marker is interpreted as required and the entropy-coded segment pointer is adjusted ("Adjust BP" in Figure D.21) so that 0-bytes will be fed to the decoder until decoding is complete. One way of accomplishing this is to point BP to the byte preceding the marker which follows the entropy-coded segment.

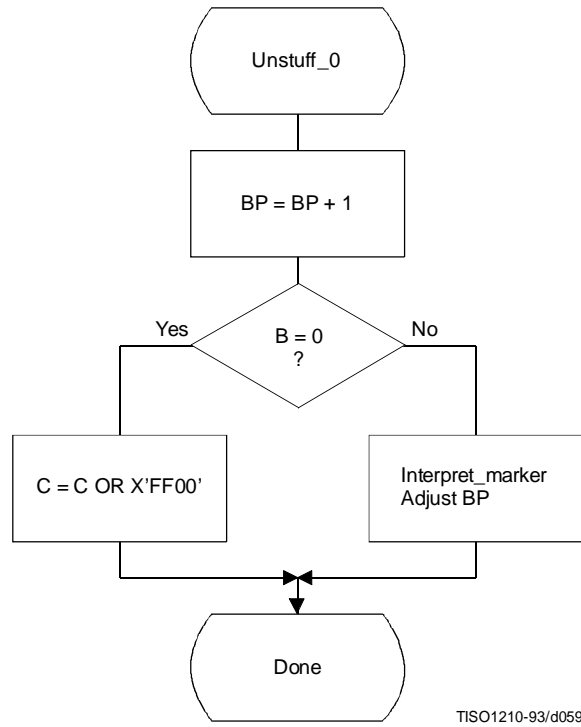


Figure D.21 – Unstuff_0 procedure for decoder

D.2.7 Initialization of the decoder

The Initdec procedure is used to start the arithmetic decoder. The basic steps are shown in Figure D.22.

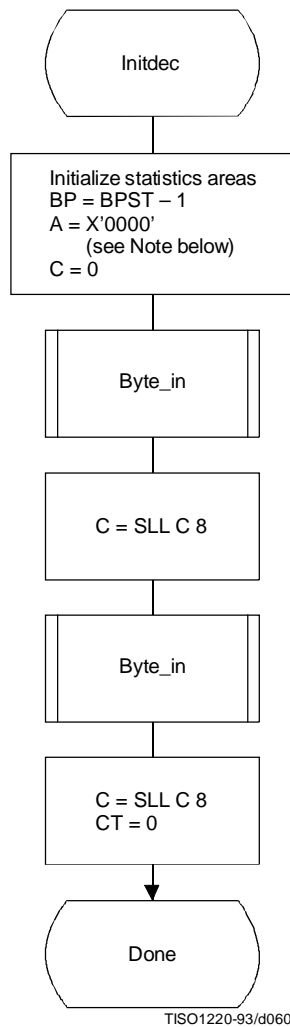


Figure D.22 – Initialization of the decoder

The estimation tables are defined by Table D.3. The statistics areas are initialized to an MPS sense of 0 and a Qe index of zero as defined by Table D.3. BP, the pointer to the entropy-coded segment, is then initialized to point to the byte before the start of the entropy-coded segment at BPST, and the interval register is set to the same starting value as in the encoder. The first byte of compressed data is fetched and shifted into Cx. The second byte is then fetched and shifted into Cx. The count is set to zero, so that a new byte of data will be fetched by Renorm_d.

NOTE – Although the probability interval is initialized to X'10000' in both Initenc and Initdec, the precision of the probability interval register can still be limited to 16 bits. When the precision of the interval register is 16 bits, it is initialized to zero.

D.3 Bit ordering within bytes

The arithmetically encoded entropy-coded segment is an integer of variable length. Therefore, the ordering of bytes and the bit ordering within bytes is the same as for parameters (see B.1.1.1).

Annex E

Encoder and decoder control procedures

(This annex forms an integral part of this Recommendation | International Standard)

This annex describes the encoder and decoder control procedures for the sequential, progressive, and lossless modes of operation.

The encoding and decoding control procedures for the hierarchical processes are specified in Annex J.

NOTES

1 There is **no requirement** in this Specification that any encoder or decoder shall implement the procedures in precisely the manner specified by the flow charts in this annex. It is necessary only that an encoder or decoder implement the **function** specified in this annex. The sole criterion for an encoder or decoder to be considered in compliance with this Specification is that it satisfy the requirements given in clause 6 (for encoders) or clause 7 (for decoders), as determined by the compliance tests specified in Part 2.

2 Implementation-specific setup steps are not indicated in this annex and may be necessary.

E.1 Encoder control procedures

E.1.1 Control procedure for encoding an image

The encoder control procedure for encoding an image is shown in Figure E.1.

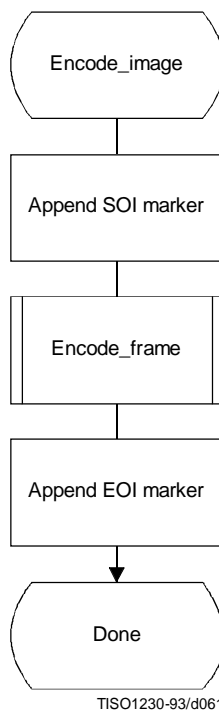


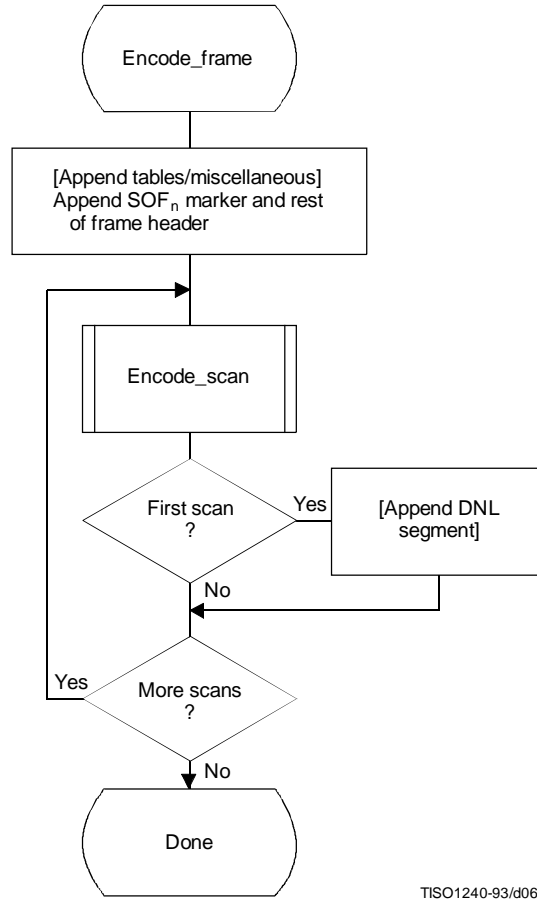
Figure E.1 – Control procedure for encoding an image

E.1.2 Control procedure for encoding a frame

In all cases where markers are appended to the compressed data, optional X'FF' fill bytes may precede the marker.

The control procedure for encoding a frame is oriented around the scans in the frame. The frame header is first appended, and then the scans are coded. Table specifications and other marker segments may precede the SOF_n marker, as indicated by [tables/miscellaneous] in Figure E.2.

Figure E.2 shows the encoding process frame control procedure.



TISO1240-93/d062

Figure E.2 – Control procedure for encoding a frame

E.1.3 Control procedure for encoding a scan

A scan consists of a single pass through the data of each component in the scan. Table specifications and other marker segments may precede the SOS marker. If more than one component is coded in the scan, the data are interleaved. If restart is enabled, the data are segmented into restart intervals. If restart is enabled, a RST_m marker is placed in the coded data between restart intervals. If restart is disabled, the control procedure is the same, except that the entire scan contains a single restart interval. The compressed image data generated by a scan is always followed by a marker, either the EOI marker or the marker of the next marker segment.

Figure E.3 shows the encoding process scan control procedure. The loop is terminated when the encoding process has coded the number of restart intervals which make up the scan. “m” is the restart interval modulo counter needed for the RST_m marker. The modulo arithmetic for this counter is shown after the “Append RST_m marker” procedure.

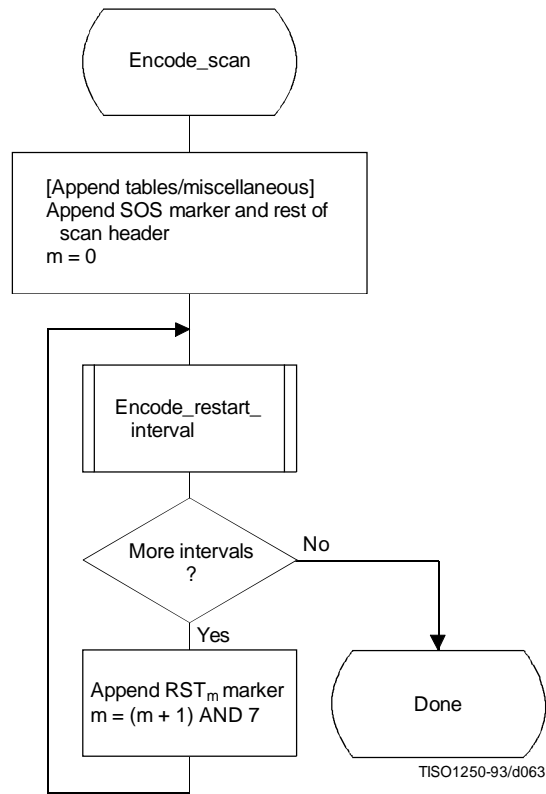


Figure E.3 – Control procedure for encoding a scan

E.1.4 Control procedure for encoding a restart interval

Figure E.4 shows the encoding process control procedure for a restart interval. The loop is terminated either when the encoding process has coded the number of minimum coded units (MCU) in the restart interval or when it has completed the image scan.

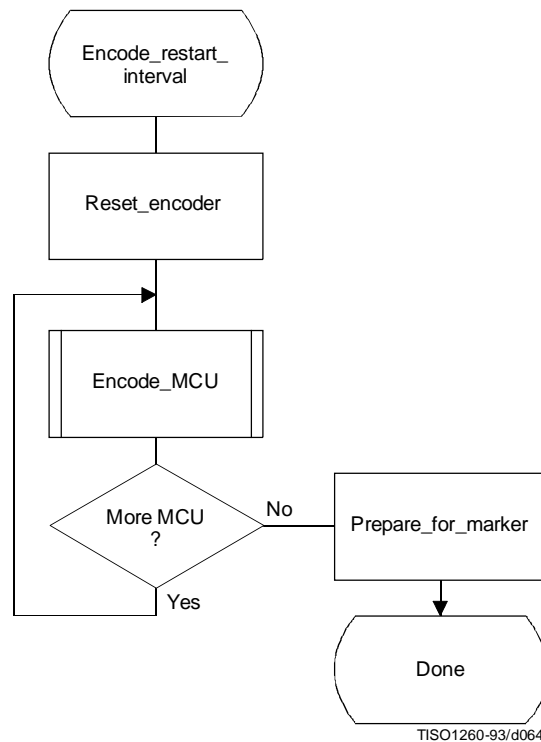


Figure E.4 – Control procedure for encoding a restart interval

The “Reset_encoder” procedure consists at least of the following:

- a) if arithmetic coding is used, initialize the arithmetic encoder using the “Initenc” procedure described in D.1.7;
- b) for DCT-based processes, set the DC prediction (PRED) to zero for all components in the scan (see F.1.1.5.1);
- c) for lossless processes, reset the prediction to a default value for all components in the scan (see H.1.1);
- d) do all other implementation-dependent setups that may be necessary.

The procedure “Prepare_for_marker” terminates the entropy-coded segment by:

- a) padding a Huffman entropy-coded segment with 1-bits to complete the final byte (and if needed stuffing a zero byte) (see F.1.2.3); or
- b) invoking the procedure “Flush” (see D.1.8) to terminate an arithmetic entropy-coded segment.

NOTE – The number of minimum coded units (MCU) in the final restart interval must be adjusted to match the number of MCU in the scan. The number of MCU is calculated from the frame and scan parameters. (See Annex B.)

E.1.5 Control procedure for encoding a minimum coded unit (MCU)

The minimum coded unit is defined in A.2. Within a given MCU the data units are coded in the order in which they occur in the MCU. The control procedure for encoding a MCU is shown in Figure E.5.

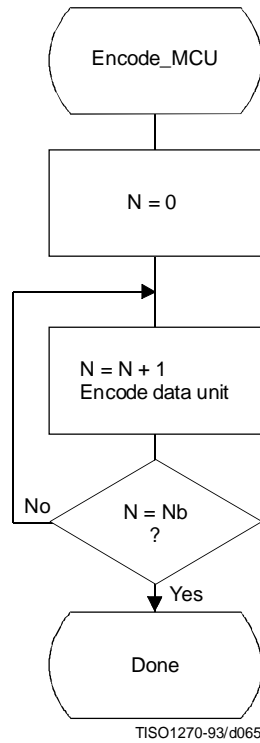


Figure E.5 – Control procedure for encoding a minimum coded unit (MCU)

In Figure E.5, N_b refers to the number of data units in the MCU. The order in which data units occur in the MCU is defined in A.2. The data unit is an 8×8 block for DCT-based processes, and a single sample for lossless processes.

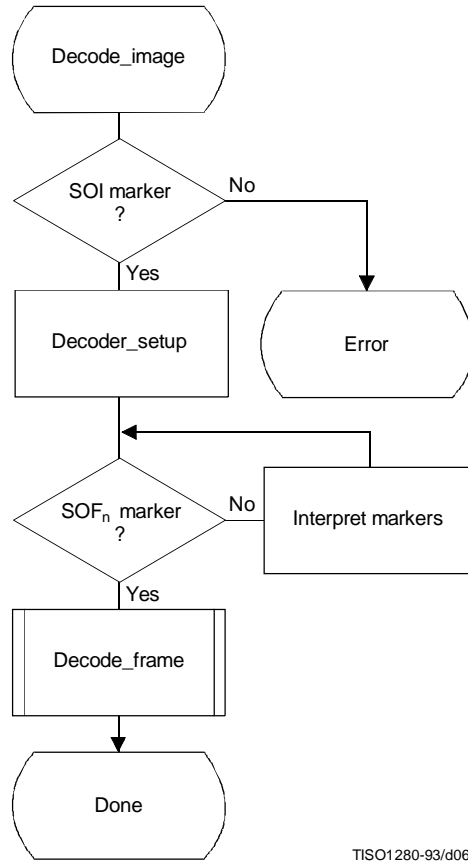
The procedures for encoding a data unit are specified in Annexes F, G, and H.

E.2 Decoder control procedures

E.2.1 Control procedure for decoding compressed image data

Figure E.6 shows the decoding process control for compressed image data.

Decoding control centers around identification of various markers. The first marker must be the SOI (Start Of Image) marker. The “Decoder_setup” procedure resets the restart interval ($R_i = 0$) and, if the decoder has arithmetic decoding capabilities, sets the conditioning tables for the arithmetic coding to their default values. (See F.1.4.4.1.4 and F.1.4.4.2.1.) The next marker is normally a SOF_n (Start Of Frame) marker; if this is not found, one of the marker segments listed in Table E.1 has been received.



TISO1280-93/d066

Figure E.6 – Control procedure for decoding compressed image data

Table E.1 – Markers recognized by “Interpret markers”

Marker	Purpose
DHT	Define Huffman Tables
DAC	Define Arithmetic Conditioning
DQT	Define Quantization Tables
DRI	Define Restart Interval
APP _n	Application defined marker
COM	Comment

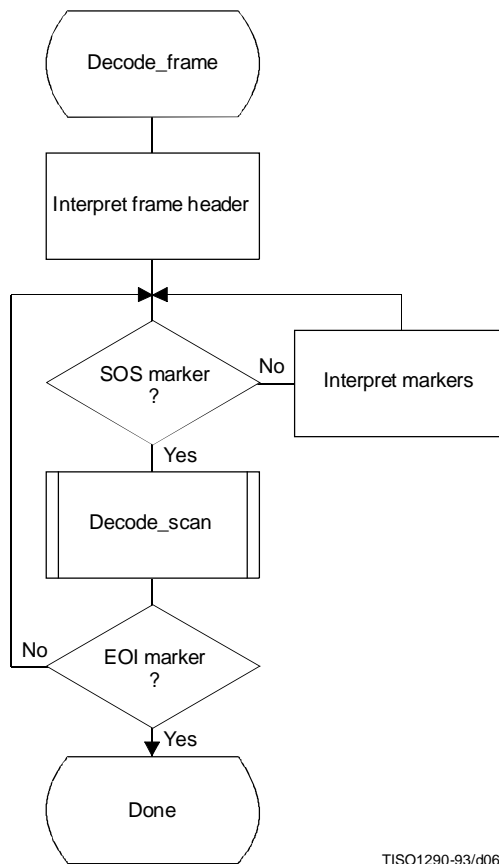
Note that optional X'FF' fill bytes which may precede any marker shall be discarded before determining which marker is present.

The additional logic to interpret these various markers is contained in the box labeled “Interpret markers”. DHT markers shall be interpreted by processes using Huffman coding. DAC markers shall be interpreted by processes using arithmetic coding. DQT markers shall be interpreted by DCT-based decoders. DRI markers shall be interpreted by all decoders. APPn and COM markers shall be interpreted only to the extent that they do not interfere with the decoding.

By definition, the procedures in “Interpret markers” leave the system at the next marker. Note that if the expected SOI marker is missing at the start of the compressed image data, an error condition has occurred. The techniques for detecting and managing error conditions can be as elaborate or as simple as desired.

E.2.2 Control procedure for decoding a frame

Figure E.7 shows the control procedure for the decoding of a frame.



TISO1290-93/d067

Figure E.7 – Control procedure for decoding a frame

The loop is terminated if the EOI marker is found at the end of the scan.

The markers recognized by “Interpret markers” are listed in Table E.1. Subclause E.2.1 describes the extent to which the various markers shall be interpreted.

E.2.3 Control procedure for decoding a scan

Figure E.8 shows the decoding of a scan.

The loop is terminated when the expected number of restart intervals has been decoded.

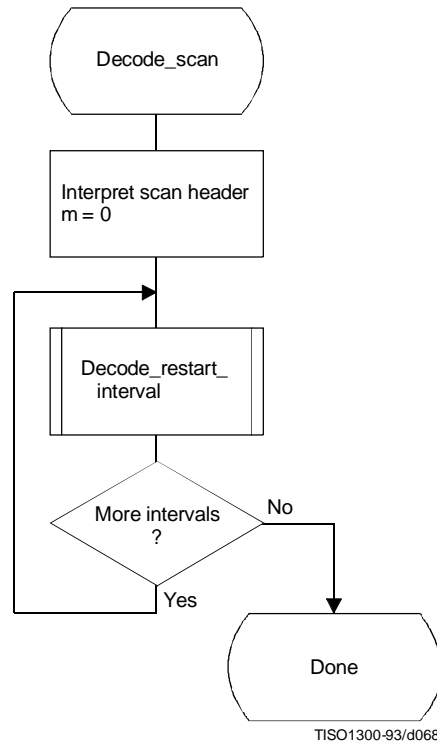


Figure E.8 – Control procedure for decoding a scan

E.2.4 Control procedure for decoding a restart interval

The procedure for decoding a restart interval is shown in Figure E.9. The “Reset_decoder” procedure consists at least of the following:

- a) if arithmetic coding is used, initialize the arithmetic decoder using the “Initdec” procedure described in D.2.7;
- b) for DCT-based processes, set the DC prediction (PRED) to zero for all components in the scan (see F.2.1.3.1);
- c) for lossless process, reset the prediction to a default value for all components in the scan (see H.2.1);
- d) do all other implementation-dependent setups that may be necessary.

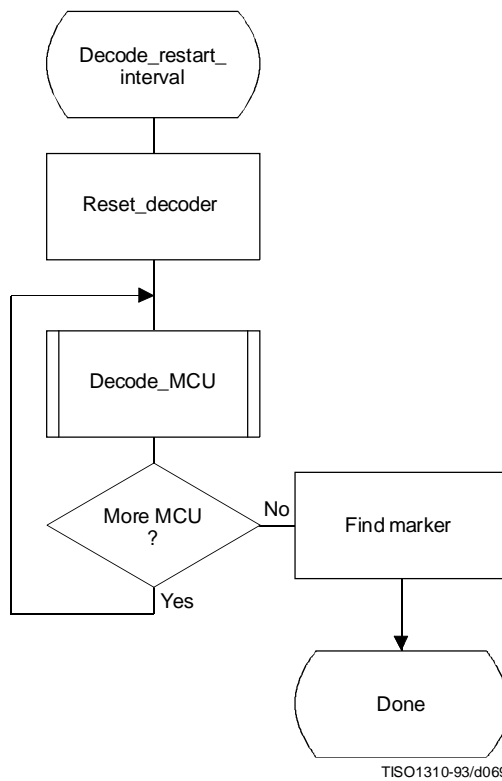


Figure E.9 – Control procedure for decoding a restart interval

At the end of the restart interval, the next marker is located. If a problem is detected in locating this marker, error handling procedures may be invoked. While such procedures are optional, the decoder shall be able to correctly recognize restart markers in the compressed data and reset the decoder when they are encountered. The decoder shall also be able to recognize the DNL marker, set the number of lines defined in the DNL segment, and end the “Decode_restart_interval” procedure.

NOTE – The final restart interval may be smaller than the size specified by the DRI marker segment, as it includes only the number of MCUs remaining in the scan.

E.2.5 Control procedure for decoding a minimum coded unit (MCU)

The procedure for decoding a minimum coded unit (MCU) is shown in Figure E.10.

In Figure E.10 Nb is the number of data units in a MCU.

The procedures for decoding a data unit are specified in Annexes F, G, and H.

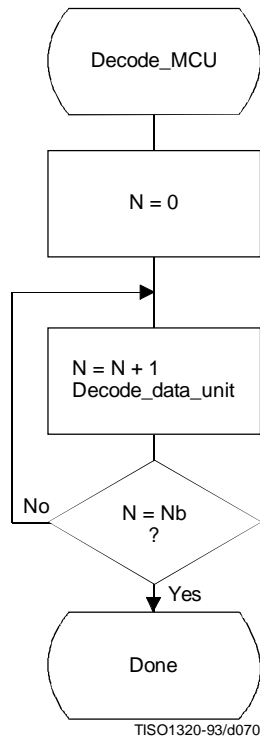


Figure E.10 – Control procedure for decoding a minimum coded unit (MCU)

Annex F

Sequential DCT-based mode of operation

(This annex forms an integral part of this Recommendation | International Standard)

This annex provides a **functional specification** of the following coding processes for the sequential DCT-based mode of operation:

- 1) baseline sequential;
- 2) extended sequential, Huffman coding, 8-bit sample precision;
- 3) extended sequential, arithmetic coding, 8-bit sample precision;
- 4) extended sequential, Huffman coding, 12-bit sample precision;
- 5) extended sequential, arithmetic coding, 12-bit sample precision.

For each of these, the encoding process is specified in F.1, and the decoding process is specified in F.2. The functional specification is presented by means of specific flow charts for the various procedures which comprise these coding processes.

NOTE – There is **no requirement** in this Specification that any encoder or decoder which embodies one of the above-named processes shall implement the procedures in precisely the manner specified by the flow charts in this annex. It is necessary only that an encoder or decoder implement the **function** specified in this annex. The sole criterion for an encoder or decoder to be considered in compliance with this Specification is that it satisfy the requirements given in clause 6 (for encoders) or clause 7 (for decoders), as determined by the compliance tests specified in Part 2.

F.1 Sequential DCT-based encoding processes

F.1.1 Sequential DCT-based control procedures and coding models

F.1.1.1 Control procedures for sequential DCT-based encoders

The control procedures for encoding an image and its constituent parts – the frame, scan, restart interval and MCU – are given in Figures E.1 to E.5. The procedure for encoding a MCU (see Figure E.5) repetitively calls the procedure for encoding a data unit. For DCT-based encoders the data unit is an 8×8 block of samples.

F.1.1.2 Procedure for encoding an 8×8 block data unit

For the sequential DCT-based processes encoding an 8×8 block data unit consists of the following procedures:

- a) level shift, calculate forward 8×8 DCT and quantize the resulting coefficients using table destination specified in frame header;
- b) encode DC coefficient for 8×8 block using DC table destination specified in scan header;
- c) encode AC coefficients for 8×8 block using AC table destination specified in scan header.

F.1.1.3 Level shift and forward DCT (FDCT)

The mathematical definition of the FDCT is given in A.3.3.

Prior to computing the FDCT the input data are level shifted to a signed two's complement representation as described in A.3.1. For 8-bit input precision the level shift is achieved by subtracting 128. For 12-bit input precision the level shift is achieved by subtracting 2048.

F.1.1.4 Quantization of the FDCT

The uniform quantization procedure described in Annex A is used to quantize the DCT coefficients. One of four quantization tables may be used by the encoder. No default quantization tables are specified in this Specification. However, some typical quantization tables are given in Annex K.

The quantized DCT coefficient values are signed, two's complement integers with 11-bit precision for 8-bit input precision and 15-bit precision for 12-bit input precision.

F.1.1.5 Encoding models for the sequential DCT procedures

The two dimensional array of quantized DCT coefficients is rearranged in a zig-zag sequence order defined in A.3.6. The zig-zag order coefficients are denoted ZZ (0) through ZZ(63) with:

$$ZZ(0) = Sq_{00}, ZZ(1) = Sq_{01}, ZZ(2) = Sq_{10}, \dots, ZZ(63) = Sq_{77}$$

Sq_{vu} are defined in Figure A.6.

Two coding procedures are used, one for the DC coefficient ZZ(0) and the other for the AC coefficients ZZ(1)..ZZ(63). The coefficients are encoded in the order in which they occur in zig-zag sequence order, starting with the DC coefficient. The coefficients are represented as two's complement integers.

F.1.1.5.1 Encoding model for DC coefficients

The DC coefficients are coded differentially, using a one-dimensional predictor, PRED, which is the quantized DC value from the most recently coded 8 × 8 block from the same component. The difference, DIFF, is obtained from

$$DIFF = ZZ(0) - PRED$$

At the beginning of the scan and at the beginning of each restart interval, the prediction for the DC coefficient prediction is initialized to 0. (Recall that the input data have been level shifted to two's complement representation.)

F.1.1.5.2 Encoding model for AC coefficients

Since many coefficients are zero, runs of zeros are identified and coded efficiently. In addition, if the remaining coefficients in the zig-zag sequence order are all zero, this is coded explicitly as an end-of-block (EOB).

F.1.2 Baseline Huffman encoding procedures

The baseline encoding procedure is for 8-bit sample precision. The encoder may employ up to two DC and two AC Huffman tables within one scan.

F.1.2.1 Huffman encoding of DC coefficients

F.1.2.1.1 Structure of DC code table

The DC code table consists of a set of Huffman codes (maximum length 16 bits) and appended additional bits (in most cases) which can code any possible value of DIFF, the difference between the current DC coefficient and the prediction. The Huffman codes for the difference categories are generated in such a way that no code consists entirely of 1-bits ('X'FF' prefix marker code avoided).

The two's complement difference magnitudes are grouped into 12 categories, SSSS, and a Huffman code is created for each of the 12 difference magnitude categories (see Table F.1).

For each category, except SSSS = 0, an additional bits field is appended to the code word to uniquely identify which difference in that category actually occurred. The number of extra bits is given by SSSS; the extra bits are appended to the LSB of the preceding Huffman code, most significant bit first. When DIFF is positive, the SSSS low order bits of DIFF are appended. When DIFF is negative, the SSSS low order bits of (DIFF - 1) are appended. Note that the most significant bit of the appended bit sequence is 0 for negative differences and 1 for positive differences.

F.1.2.1.2 Defining Huffman tables for the DC coefficients

The syntax for specifying the Huffman tables is given in Annex B. The procedure for creating a code table from this information is described in Annex C. No more than two Huffman tables may be defined for coding of DC coefficients. Two examples of Huffman tables for coding of DC coefficients are provided in Annex K.

Table F.1 – Difference magnitude categories for DC coding

SSSS	DIFF values
0	0
1	-1,1
2	-3,-2,2,3
3	-7..-4,4..7
4	-15..-8,8..15
5	-31..-16,16..31
6	-63..-32,32..63
7	-127..-64,64..127
8	-255..-128,128..255
9	-511..-256,256..511
10	-1 023..-512,512..1 023
11	-2 047..-1 024,1 024..2 047

F.1.2.1.3 Huffman encoding procedures for DC coefficients

The encoding procedure is defined in terms of a set of extended tables, XHUFÇO and XHUFŞI, which contain the complete set of Huffman codes and sizes for all possible difference values. For full 12-bit precision the tables are relatively large. For the baseline system, however, the precision of the differences may be small enough to make this description practical.

XHUFÇO and XHUFŞI are generated from the encoder tables EHUFÇO and EHUFŞI (see Annex C) by appending to the Huffman codes for each difference category the additional bits that completely define the difference. By definition, XHUFÇO and XHUFŞI have entries for each possible difference value. XHUFÇO contains the concatenated bit pattern of the Huffman code and the additional bits field; XHUFŞI contains the total length in bits of this concatenated bit pattern. Both are indexed by DIFF, the difference between the DC coefficient and the prediction.

The Huffman encoding procedure for the DC difference, DIFF, is:

$$\text{SIZE} = \text{XHUFŞI}(\text{DIFF})$$

$$\text{CODE} = \text{XHUFÇO}(\text{DIFF})$$

$$\text{code SIZE bits of CODE}$$

where DC is the quantized DC coefficient value and PRED is the predicted quantized DC value. The Huffman code (CODE) (including any additional bits) is obtained from XHUFÇO and SIZE (length of the code including additional bits) is obtained from XHUFŞI, using DIFF as the index to the two tables.

F.1.2.2 Huffman encoding of AC coefficients

F.1.2.2.1 Structure of AC code table

Each non-zero AC coefficient in ZZ is described by a composite 8-bit value, RS, of the form

$$\text{RS} = \text{binary 'RRRRSSSS'}$$

The 4 least significant bits, 'SSSS', define a category for the amplitude of the next non-zero coefficient in ZZ, and the 4 most significant bits, 'RRRR', give the position of the coefficient in ZZ relative to the previous non-zero coefficient (i.e. the run-length of zero coefficients between non-zero coefficients). Since the run length of zero coefficients may exceed 15, the value 'RRRRSSSS' = X'F0' is defined to represent a run length of 15 zero coefficients followed by a coefficient of zero amplitude. (This can be interpreted as a run length of 16 zero coefficients.) In addition, a special value 'RRRRSSSS' = '00000000' is used to code the end-of-block (EOB), when all remaining coefficients in the block are zero.

The general structure of the code table is illustrated in Figure F.1. The entries marked "N/A" are undefined for the baseline procedure.

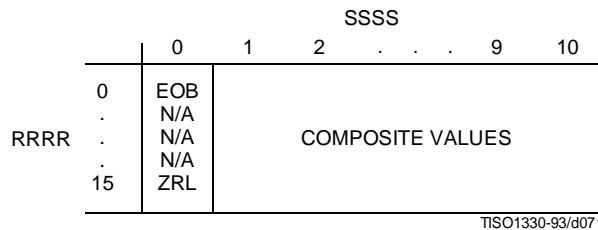


Figure F.1 – Two-dimensional value array for Huffman coding

The magnitude ranges assigned to each value of SSSS are defined in Table F.2.

Table F.2 – Categories assigned to coefficient values

SSSS	AC coefficients
1	-1,1
2	-3,-2,2,3
3	-7..-4,4..7
4	-15..-8,8..15
5	-31..-16,16..31
6	-63..-32,32..63
7	-127..-64,64..127
8	-255..-128,128..255
9	-511..-256,256..511
10	-1 023..-512,512..1 023

The composite value, RRRRSSSS, is Huffman coded and each Huffman code is followed by additional bits which specify the sign and exact amplitude of the coefficient.

The AC code table consists of one Huffman code (maximum length 16 bits, not including additional bits) for each possible composite value. The Huffman codes for the 8-bit composite values are generated in such a way that no code consists entirely of 1-bits.

The format for the additional bits is the same as in the coding of the DC coefficients. The value of SSSS gives the number of additional bits required to specify the sign and precise amplitude of the coefficient. The additional bits are either the low-order SSSS bits of ZZ(K) when ZZ(K) is positive or the low-order SSSS bits of ZZ(K) – 1 when ZZ(K) is negative. ZZ(K) is the Kth coefficient in the zig-zag sequence of coefficients being coded.

F.1.2.2.2 Defining Huffman tables for the AC coefficients

The syntax for specifying the Huffman tables is given in Annex B. The procedure for creating a code table from this information is described in Annex C.

In the baseline system no more than two Huffman tables may be defined for coding of AC coefficients. Two examples of Huffman tables for coding of AC coefficients are provided in Annex K.

F.1.2.2.3 Huffman encoding procedures for AC coefficients

As defined in Annex C, the Huffman code table is assumed to be available as a pair of tables, EHUF0 (containing the code bits) and EHUF1 (containing the length of each code in bits), both indexed by the composite value defined above.

The procedure for encoding the AC coefficients in a block is shown in Figures F.2 and F.3. In Figure F.2, K is the index to the zig-zag scan position and R is the run length of zero coefficients.

The procedure “Append EHUF1(X'F0') bits of EHUF0(X'F0')” codes a run of 16 zero coefficients (ZRL code of Figure F.1). The procedure “Code EHUF1(0) bits of EHUF0(0)” codes the end-of-block (EOB code). If the last coefficient (K = 63) is not zero, the EOB code is bypassed.

CSIZE is a procedure which maps an AC coefficient to the SSSS value as defined in Table F.2.

F.1.2.3 Byte stuffing

In order to provide code space for marker codes which can be located in the compressed image data without decoding, byte stuffing is used.

Whenever, in the course of normal encoding, the byte value X'FF' is created in the code string, a X'00' byte is stuffed into the code string.

If a X'00' byte is detected after a X'FF' byte, the decoder must discard it. If the byte is not zero, a marker has been detected, and shall be interpreted to the extent needed to complete the decoding of the scan.

Byte alignment of markers is achieved by padding incomplete bytes with 1-bits. If padding with 1-bits creates a X'FF' value, a zero byte is stuffed before adding the marker.

F.1.3 Extended sequential DCT-based Huffman encoding process for 8-bit sample precision

This process is identical to the Baseline encoding process described in F.1.2, with the exception that the number of sets of Huffman table destinations which may be used within the same scan is increased to four. Four DC and four AC Huffman table destinations is the maximum allowed by this Specification.

F.1.4 Extended sequential DCT-based arithmetic encoding process for 8-bit sample precision

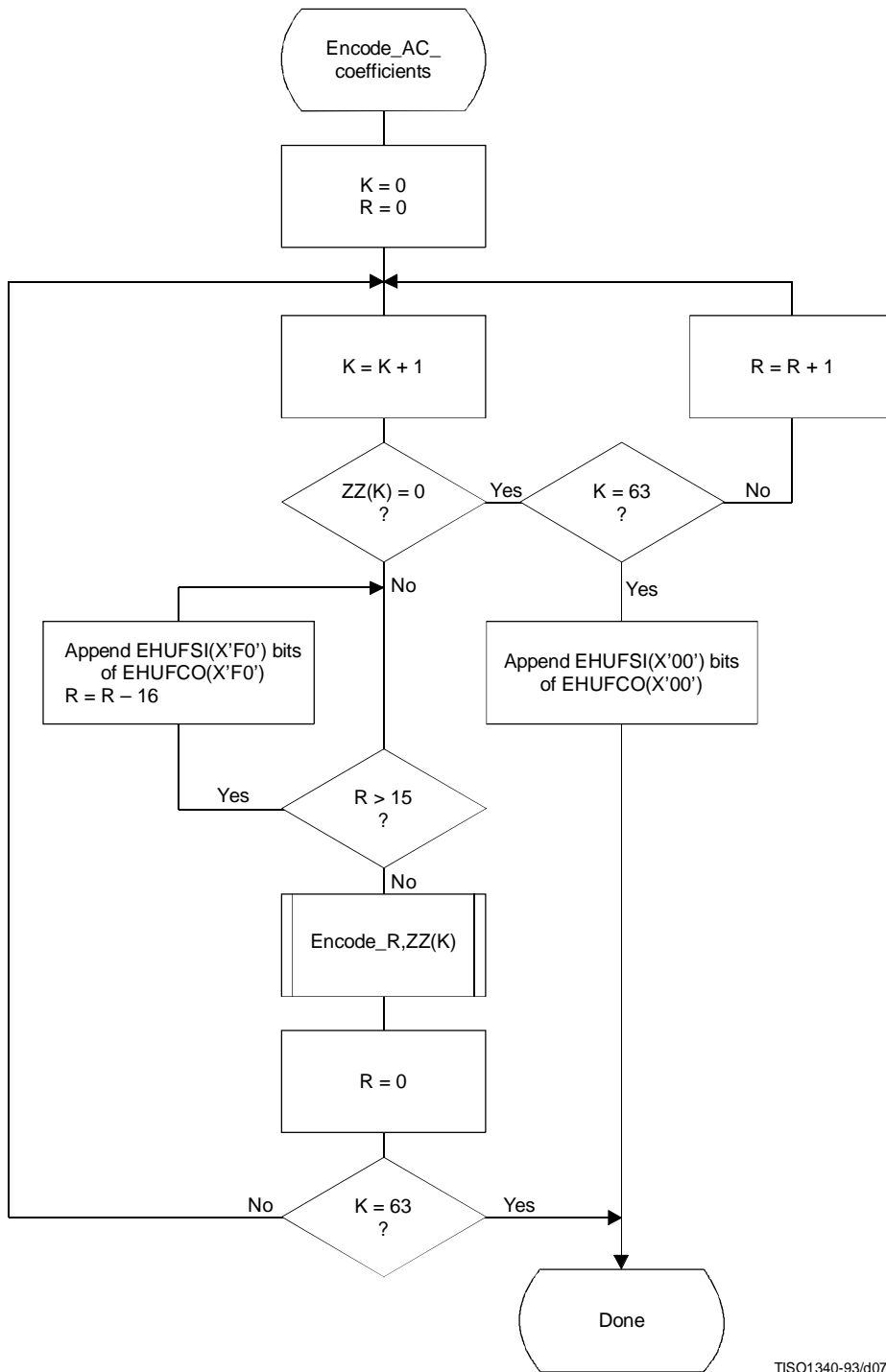
This subclause describes the use of arithmetic coding procedures in the sequential DCT-based encoding process.

NOTE – The arithmetic coding procedures in this Specification are defined for the maximum precision to encourage interchangeability.

The arithmetic coding extensions have the same DCT model as the Baseline DCT encoder. Therefore, Annex F.1.1 also applies to arithmetic coding. As with the Huffman coding technique, the binary arithmetic coding technique is lossless. It is possible to transcode between the two systems without either FDCT or IDCT computations, and without modification of the reconstructed image.

The basic principles of adaptive binary arithmetic coding are described in Annex D. Up to four DC and four AC conditioning table destinations and associated statistics areas may be used within one scan.

The arithmetic encoding procedures for encoding binary decisions, initializing the statistics area, initializing the encoder, terminating the code string, and adding restart markers are listed in Table D.1 of Annex D.



TISO1340-93/d072

Figure F.2 – Procedure for sequential encoding of AC coefficients with Huffman coding

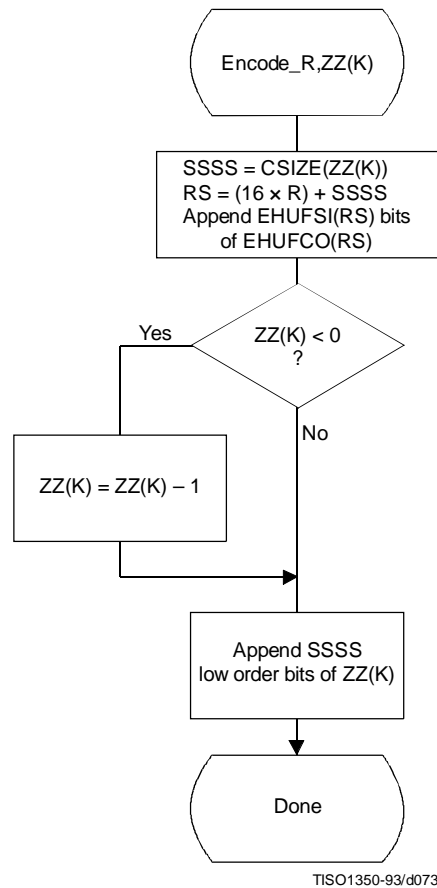


Figure F.3 – Sequential encoding of a non-zero AC coefficient

Some of the procedures in Table D.1 are used in the higher level control structure for scans and restart intervals described in Annex E. At the beginning of scans and restart intervals, the probability estimates used in the arithmetic coder are reset to the standard initial value as part of the Initenc procedure which restarts the arithmetic coder. At the end of scans and restart intervals, the Flush procedure is invoked to empty the code register before the next marker is appended.

F.1.4.1 Arithmetic encoding of DC coefficients

The basic structure of the decision sequence for encoding a DC difference value, DIFF, is shown in Figure F.4.

The context-index S_0 and other context-indices used in the DC coding procedures are defined in Table F.4 (see F.1.4.4.1.3). A 0-decision is coded if the difference value is zero and a 1-decision is coded if the difference is not zero. If the difference is not zero, the sign and magnitude are coded using the procedure Encode_V(S_0), which is described in F.1.4.3.1.

F.1.4.2 Arithmetic encoding of AC coefficients

The AC coefficients are coded in the order in which they occur in the zig-zag sequence $ZZ(1, \dots, 63)$. An end-of-block (EOB) binary decision is coded before coding the first AC coefficient in ZZ , and after each non-zero coefficient. If the EOB occurs, all remaining coefficients in ZZ are zero. Figure F.5 illustrates the decision sequence. The equivalent procedure for the Huffman coder is found in Figure F.2.

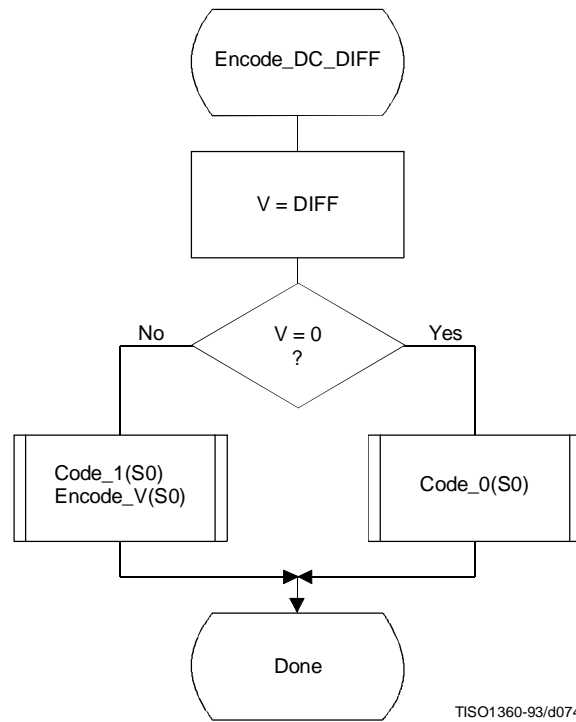


Figure F.4 – Coding model for arithmetic coding of DC difference

The context-indices SE and S0 used in the AC coding procedures are defined in Table F.5 (see F.1.4.4.2). In Figure F.5, K is the index to the zig-zag sequence position. For the sequential scan, Kmin is 1 and Se is 63. The V = 0 decision is part of a loop which codes runs of zero coefficients. Whenever the coefficient is non-zero, “Encode_V(S0)” codes the sign and magnitude of the coefficient. Each time a non-zero coefficient is coded, it is followed by an EOB decision. If the EOB occurs, a 1-decision is coded to indicate that the coding of the block is complete. If the coefficient for K = Se is not zero, the EOB decision is skipped.

F.1.4.3 Encoding the binary decision sequence for non-zero DC differences and AC coefficients

Both the DC difference and the AC coefficients are represented as signed two’s complement integer values. The decomposition of these signed integer values into a binary decision tree is done in the same way for both the DC and AC coding models.

Although the binary decision trees for this section of the DC and AC coding models are the same, the statistical models for assigning statistics bins to the binary decisions in the tree are quite different.

F.1.4.3.1 Structure of the encoding decision sequence

The encoding sequence can be separated into three procedures, a procedure which encodes the sign, a second procedure which identifies the magnitude category, and a third procedure which identifies precisely which magnitude occurred within the category identified in the second procedure.

At the point where the binary decision sequence in Encode_V(S0) starts, the coefficient or difference has already been determined to be non-zero. That determination was made in the procedures in Figures F.4 and F.5.

Denoting either DC differences (DIFF) or AC coefficients as V, the non-zero signed integer value of V is encoded by the sequence shown in Figure F.6. This sequence first codes the sign of V. It then (after converting V to a magnitude and decrementing it by 1 to give Sz) codes the magnitude category of Sz (code_log2_Sz), and then codes the low order magnitude bits (code_Sz_bits) to identify the exact magnitude value.

There are two significant differences between this sequence and the similar set of operations described in F.1.2 for Huffman coding. First, the sign is encoded before the magnitude category is identified, and second, the magnitude is decremented by 1 before the magnitude category is identified.

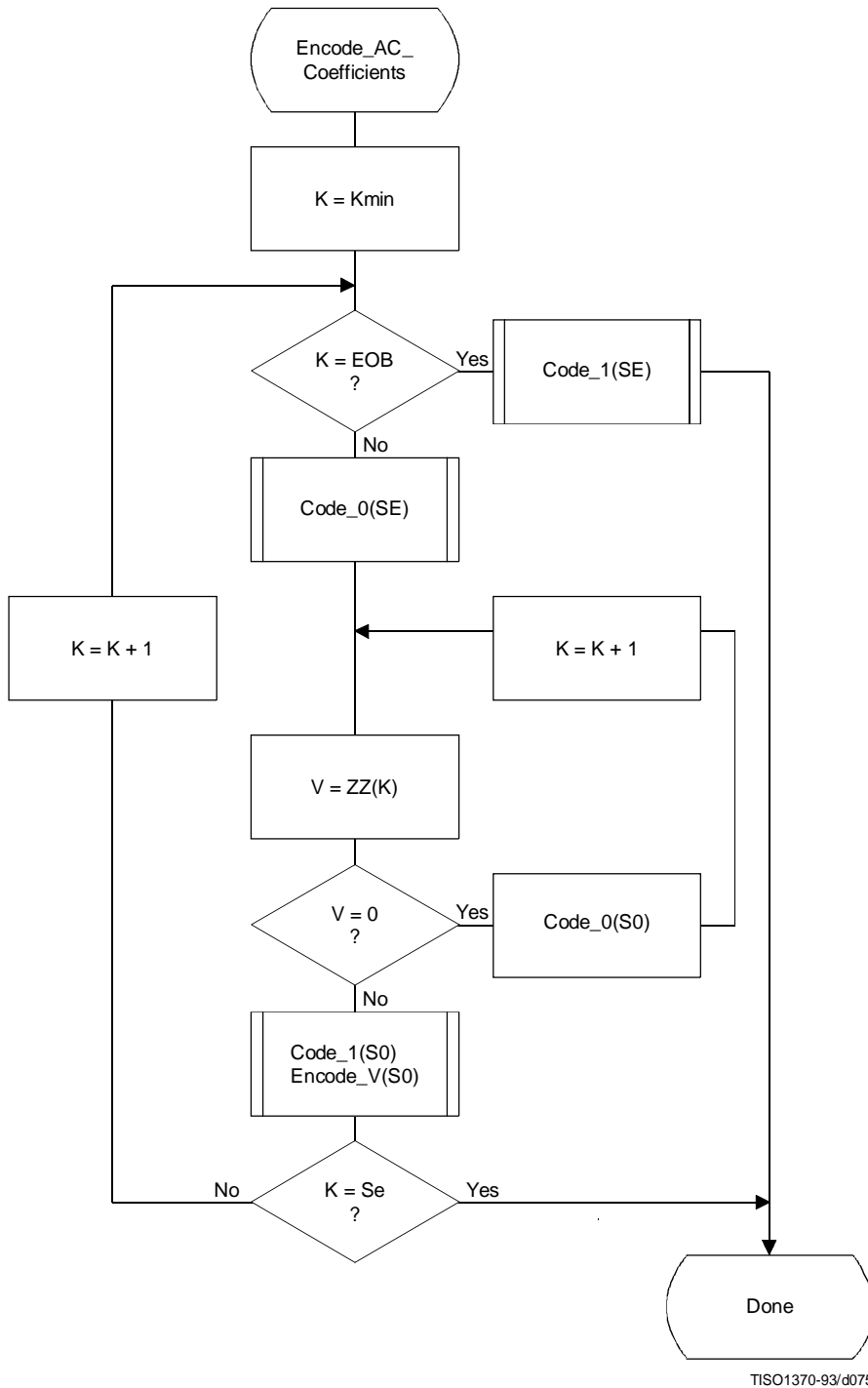


Figure F.5 – AC coding model for arithmetic coding

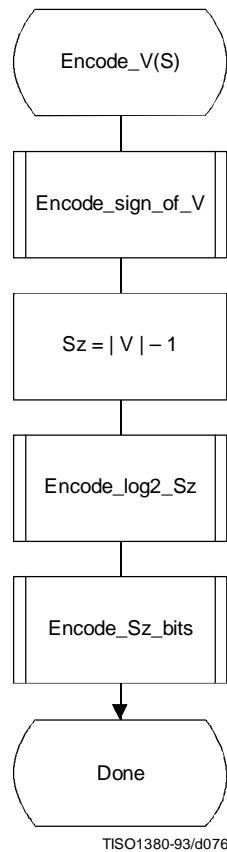


Figure F.6 – Sequence of procedures in encoding non-zero values of V

F.1.4.3.1.1 Encoding the sign

The sign is encoded by coding a 0-decision when the sign is positive and a 1-decision when the sign is negative (see Figure F.7).

The context-indices SS, SN and SP are defined for DC coding in Table F.4 and for AC coding in Table F.5. After the sign is coded, the context-index S is set to either SN or SP, establishing an initial value for Encode_log2_Sz.

F.1.4.3.1.2 Encoding the magnitude category

The magnitude category is determined by a sequence of binary decisions which compares Sz against an exponentially increasing bound (which is a power of 2) in order to determine the position of the leading 1-bit. This establishes the magnitude category in much the same way that the Huffman encoder generates a code for the value associated with the difference category. The flow chart for this procedure is shown in Figure F.8.

The starting value of the context-index S is determined in Encode_sign_of_V, and the context-index values X1 and X2 are defined for DC coding in Table F.4 and for AC coding in Table F.5. In Figure F.8, M is the exclusive upper bound for the magnitude and the abbreviations “SLL” and “SRL” refer to the shift-left-logical and shift-right-logical operations – in this case by one bit position. The SRL operation at the completion of the procedure aligns M with the most significant bit of Sz (see Table F.3).

The highest precision allowed for the DCT is 15 bits. Therefore, the highest precision required for the coding decision tree is 16 bits for the DC coefficient difference and 15 bits for the AC coefficients, including the sign bit.

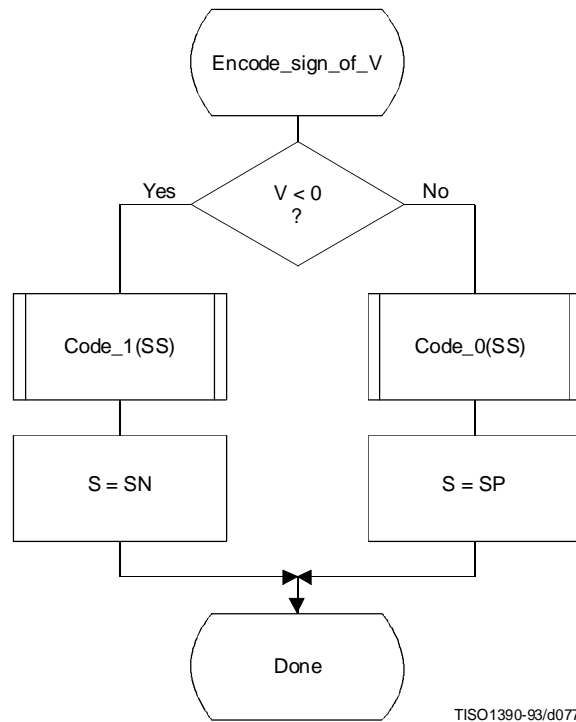


Figure F.7 – Encoding the sign of V

Table F.3 – Categories for each maximum bound

Exclusive upper bound (M)	Sz range	Number of low order magnitude bits
1	0	0
2	1	0
4	2,3	1
8	4,...,7	2
16	8,...,15	3
32	16,...,31	4
64	32,...,63	5
128	64,...,127	6
256	128,...,255	7
512	256,...,511	8
1 024	512,...,1 023	9
2 048	1 024,...,2 047	10
4 096	2 048,...,4 095	11
8 192	4 096,...,8 191	12
16 384	8 192,...,16 383	13
32 768	16 384,...,32 767	14

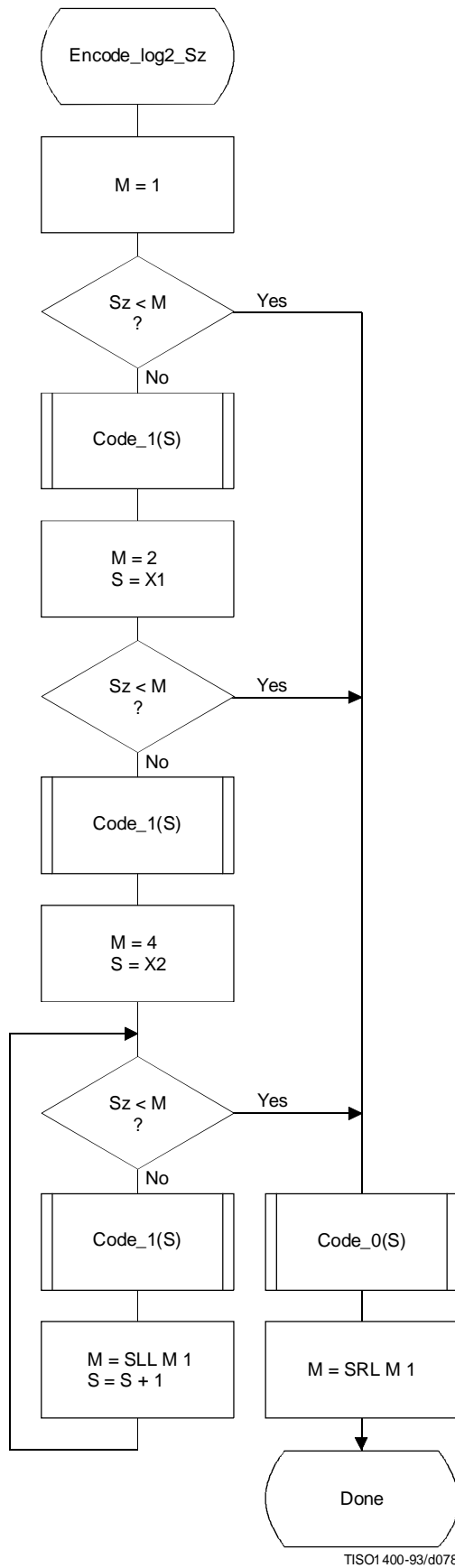


Figure F.8 – Decision sequence to establish the magnitude category

F.1.4.3.1.3 Encoding the exact value of the magnitude

After the magnitude category is encoded, the low order magnitude bits are encoded. These bits are encoded in order of decreasing bit significance. The procedure is shown in Figure F.9. The abbreviation “SRL” indicates the shift-right-logical operation, and M is the exclusive bound established in Figure F.8. Note that M has only one bit set – shifting M right converts it into a bit mask for the logical “AND” operation.

The starting value of the context-index S is determined in Encode_log2_Sz. The increment of S by 14 at the beginning of this procedure sets the context-index to the value required in Tables F.4 and F.5.

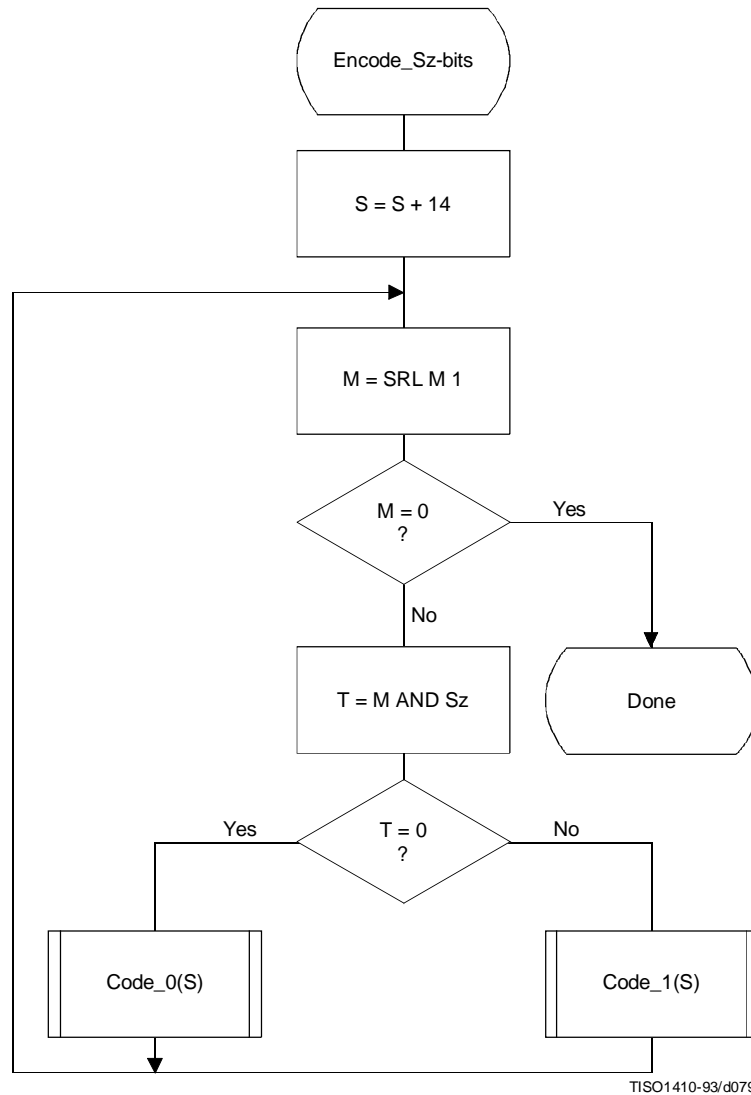


Figure F.9 – Decision sequence to code the magnitude bit pattern

F.1.4.4 Statistical models

An adaptive binary arithmetic coder requires a statistical model. The statistical model defines the contexts which are used to select the conditional probability estimates used in the encoding and decoding procedures.

Each decision in the binary decision trees is associated with one or more contexts. These contexts identify the sense of the MPS and the index in Table D.3 of the conditional probability estimate Q_e which is used to encode and decode the binary decision.

The arithmetic coder is adaptive, which means that the probability estimates for each context are developed and maintained by the arithmetic coding system on the basis of prior coding decisions for that context.

F.1.4.4.1 Statistical model for coding DC prediction differences

The statistical model for coding the DC difference conditions some of the probability estimates for the binary decisions on previous DC coding decisions.

F.1.4.4.1.1 Statistical conditioning on sign

In coding the DC coefficients, four separate statistics bins (probability estimates) are used in coding the zero/not-zero ($V = 0$) decision, the sign decision and the first magnitude category decision. Two of these bins are used to code the $V = 0$ decision and the sign decision. The other two bins are used in coding the first magnitude decision, $S_z < 1$; one of these bins is used when the sign is positive, and the other is used when the sign is negative. Thus, the first magnitude decision probability estimate is conditioned on the sign of V .

F.1.4.4.1.2 Statistical conditioning on DC difference in previous block

The probability estimates for these first three decisions are also conditioned on D_a , the difference value coded for the previous DCT block of the same component. The differences are classified into five groups: zero, small positive, small negative, large positive and large negative. The relationship between the default classification and the quantization scale is shown in Figure F.10.

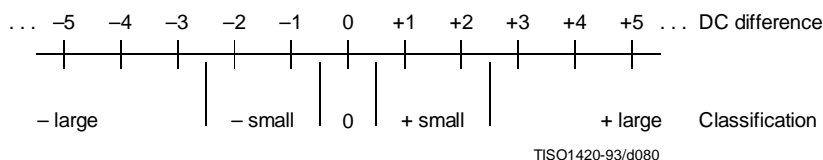


Figure F.10 – Conditioning classification of difference values

The bounds for the “small” difference category determine the classification. Defining L and U as integers in the range 0 to 15 inclusive, the lower bound (exclusive) for difference magnitudes classified as “small” is zero for $L = 0$, and is 2^{L-1} for $L > 0$.

The upper bound (inclusive) for difference magnitudes classified as “small” is 2^U .

L shall be less than or equal to U .

These bounds for the conditioning category provide a segmentation which is identical to that listed in Table F.3.

F.1.4.4.1.3 Assignment of statistical bins to the DC binary decision tree

As shown in Table F.4, each statistics area for DC coding consists of a set of 49 statistics bins. In the following explanation, it is assumed that the bins are contiguous. The first 20 bins consist of five sets of four bins selected by a context-index S_0 . The value of S_0 is given by $DC_Context(D_a)$, which provides a value of 0, 4, 8, 12 or 16, depending on the difference classification of D_a (see F.1.4.4.1.2). The remaining 29 bins, $X_1, \dots, X_{15}, M_2, \dots, M_{15}$, are used to code magnitude category decisions and magnitude bits.

Table F.4 – Statistical model for DC coefficient coding

Context-index	Value	Coding decision
S0	DC_Context(Da)	V = 0
SS	S0 + 1	Sign of V
SP	S0 + 2	Sz < 1 if V > 0
SN	S0 + 3	Sz < 1 if V < 0
X1	20	Sz < 2
X2	X1 + 1	Sz < 4
X3	X1 + 2	Sz < 8
.	.	.
.	.	.
X15	X1 + 14	Sz < 2 ¹⁵
M2	X2 + 14	Magnitude bits if Sz < 4
M3	X3 + 14	Magnitude bits if Sz < 8
.	.	.
.	.	.
M15	X15 + 14	Magnitude bits if Sz < 2 ¹⁵

F.1.4.4.1.4 Default conditioning for DC statistical model

The bounds, L and U, for determining the conditioning category have the default values L = 0 and U = 1. Other bounds may be set using the DAC (Define Arithmetic coding Conditioning) marker segment, as described in Annex B.

F.1.4.4.1.5 Initial conditions for DC statistical model

At the start of a scan and at the beginning of each restart interval, the difference for the previous DC value is defined to be zero in determining the conditioning state.

F.1.4.4.2 Statistical model for coding the AC coefficients

As shown in Table F.5, each statistics area for AC coding consists of a contiguous set of 245 statistics bins. Three bins are used for each value of the zig-zag index K, and two sets of 28 additional bins X2,...,X15,M2,...,M15 are used for coding the magnitude category and magnitude bits.

The value of SE (and also S0, SP and SN) is determined by the zig-zag index K. Since K is in the range 1 to 63, the lowest value for SE is 0 and the largest value for SP is 188. SS is not assigned a value in AC coefficient coding, as the signs of the coefficients are coded with a fixed probability value of approximately 0.5 (Qe = X'5A1D', MPS = 0).

The value of X2 is given by AC_Context(K). This gives X2 = 189 when K ≤ Kx and X2 = 217 when K > Kx, where Kx is defined using the DAC marker segment (see B.2.4.3).

Note that a X1 statistics bin is not used in this sequence. Instead, the 63 × 1 array of statistics bins for the magnitude category is used for two decisions. Once the magnitude bound has been determined – at statistics bin Xn, for example – a single statistics bin, Mn, is used to code the magnitude bit sequence for that bound.

F.1.4.4.2.1 Default conditioning for AC coefficient coding

The default value of Kx is 5. This may be modified using the DAC marker segment, as described in Annex B.

F.1.4.4.2.2 Initial conditions for AC statistical model

At the start of a scan and at each restart, all statistics bins are re-initialized to the standard default value described in Annex D.

Table F.5 – Statistical model for AC coefficient coding

Context-index	Value	Coding decision
SE	$3 \times (K - 1)$	K = EOB
S0	SE + 1	V = 0
SS	Fixed estimate	Sign of V
SN,SP	S0 + 1	Sz < 1
X1	S0 + 1	Sz < 2
X2	AC_Context(K)	Sz < 4
X3	X2 + 1	Sz < 8
.	.	.
.	.	.
X15	X2 + 13	Sz < 2 ¹⁵
M2	X2 + 14	Magnitude bits if Sz < 4
M3	X3 + 14	Magnitude bits if Sz < 8
.	.	.
.	.	.
M15	X15 + 14	Magnitude bits if Sz < 2 ¹⁵

F.1.5 Extended sequential DCT-based Huffman encoding process for 12-bit sample precision

This process is identical to the sequential DCT process for 8-bit precision extended to four Huffman table destinations as documented in F.1.3, with the following changes.

F.1.5.1 Structure of DC code table for 12-bit sample precision

The two's complement difference magnitudes are grouped into 16 categories, SSSS, and a Huffman code is created for each of the 16 difference magnitude categories.

The Huffman table for DC coding (see Table F.1) is extended as shown in Table F.6.

Table F.6 – Difference magnitude categories for DC coding

SSSS	Difference values
12	-4 095..-2 048,2 048..4 095
13	-8 191..-4 096,4 096..8 191
14	-16 383..-8 192,8 192..16 383
15	-32 767..-16 384,16 384..32 767

F.1.5.2 Structure of AC code table for 12-bit sample precision

The general structure of the code table is extended as illustrated in Figure F.11. The Huffman table for AC coding is extended as shown in Table F.7.

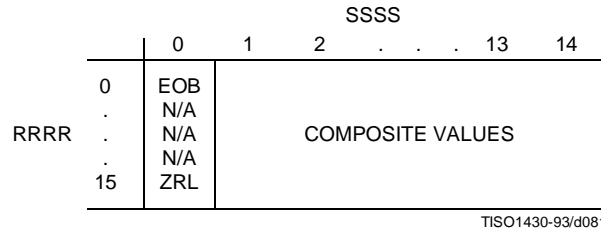


Figure F.11 – Two-dimensional value array for Huffman coding

Table F.7 – Values assigned to coefficient amplitude ranges

SSSS	AC coefficients
11	-2 047..-1 024,1 024..2 047
12	-4 095..-2 048,2 048..4 095
13	-8 191..-4 096,4 096..8 191
14	-16 383..-8 192,8 192..16 383

F.1.6 Extended sequential DCT-based arithmetic encoding process for 12-bit sample precision

The process is identical to the sequential DCT process for 8-bit precision except for changes in the precision of the FDCT computation.

The structure of the encoding procedure is identical to that specified in F.1.4 which was already defined for a 12-bit sample precision.

F.2 Sequential DCT-based decoding processes

F.2.1 Sequential DCT-based control procedures and coding models

F.2.1.1 Control procedures for sequential DCT-based decoders

The control procedures for decoding compressed image data and its constituent parts – the frame, scan, restart interval and MCU – are given in Figures E.6 to E.10. The procedure for decoding a MCU (Figure E.10) repetitively calls the procedure for decoding a data unit. For DCT-based decoders the data unit is an 8 × 8 block of samples.

F.2.1.2 Procedure for decoding an 8 × 8 block data unit

In the sequential DCT-based decoding process, decoding an 8 × 8 block data unit consists of the following procedures:

- a) decode DC coefficient for 8 × 8 block using the DC table destination specified in the scan header;
- b) decode AC coefficients for 8 × 8 block using the AC table destination specified in the scan header;
- c) dequantize using table destination specified in the frame header and calculate the inverse 8 × 8 DCT.

F.2.1.3 Decoding models for the sequential DCT procedures

Two decoding procedures are used, one for the DC coefficient ZZ(0) and the other for the AC coefficients ZZ(1)...ZZ(63). The coefficients are decoded in the order in which they occur in the zig-zag sequence order, starting with the DC coefficient. The coefficients are represented as two's complement integers.

F.2.1.3.1 Decoding model for DC coefficients

The decoded difference, DIFF, is added to PRED, the DC value from the most recently decoded 8×8 block from the same component. Thus $ZZ(0) = \text{PRED} + \text{DIFF}$.

At the beginning of the scan and at the beginning of each restart interval, the prediction for the DC coefficient is initialized to zero.

F.2.1.3.2 Decoding model for AC coefficients

The AC coefficients are decoded in the order in which they occur in ZZ. When the EOB is decoded, all remaining coefficients in ZZ are initialized to zero.

F.2.1.4 Dequantization of the quantized DCT coefficients

The dequantization of the quantized DCT coefficients as described in Annex A, is accomplished by multiplying each quantized coefficient value by the quantization table value for that coefficient. The decoder shall be able to use up to four quantization table destinations.

F.2.1.5 Inverse DCT (IDCT)

The mathematical definition of the IDCT is given in A.3.3.

After computation of the IDCT, the signed output samples are level-shifted, as described in Annex A, converting the output to an unsigned representation. For 8-bit precision the level shift is performed by adding 128. For 12-bit precision the level shift is performed by adding 2 048. If necessary, the output samples shall be clamped to stay within the range appropriate for the precision (0 to 255 for 8-bit precision and 0 to 4 095 for 12-bit precision).

F.2.2 Baseline Huffman Decoding procedures

The baseline decoding procedure is for 8-bit sample precision. The decoder shall be capable of using up to two DC and two AC Huffman tables within one scan.

F.2.2.1 Huffman decoding of DC coefficients

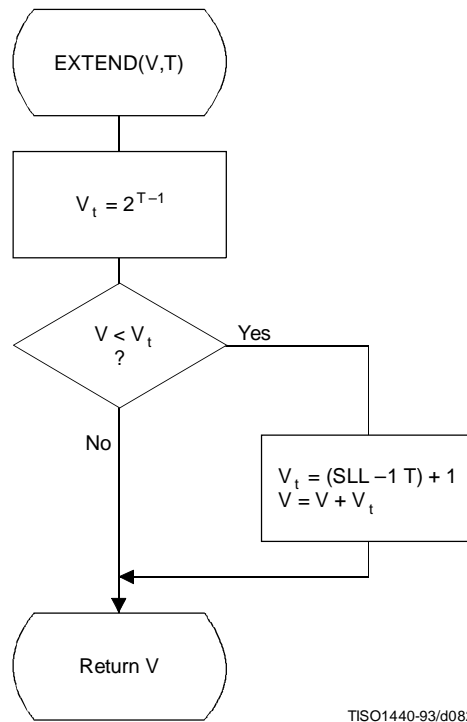
The decoding procedure for the DC difference, DIFF, is:

$$T = \text{DECODE}$$

$$\text{DIFF} = \text{RECEIVE}(T)$$

$$\text{DIFF} = \text{EXTEND}(\text{DIFF}, T)$$

where DECODE is a procedure which returns the 8-bit value associated with the next Huffman code in the compressed image data (see F.2.2.3) and RECEIVE(T) is a procedure which places the next T bits of the serial bit string into the low order bits of DIFF, MSB first. If T is zero, DIFF is set to zero. EXTEND is a procedure which converts the partially decoded DIFF value of precision T to the full precision difference. EXTEND is shown in Figure F.12.

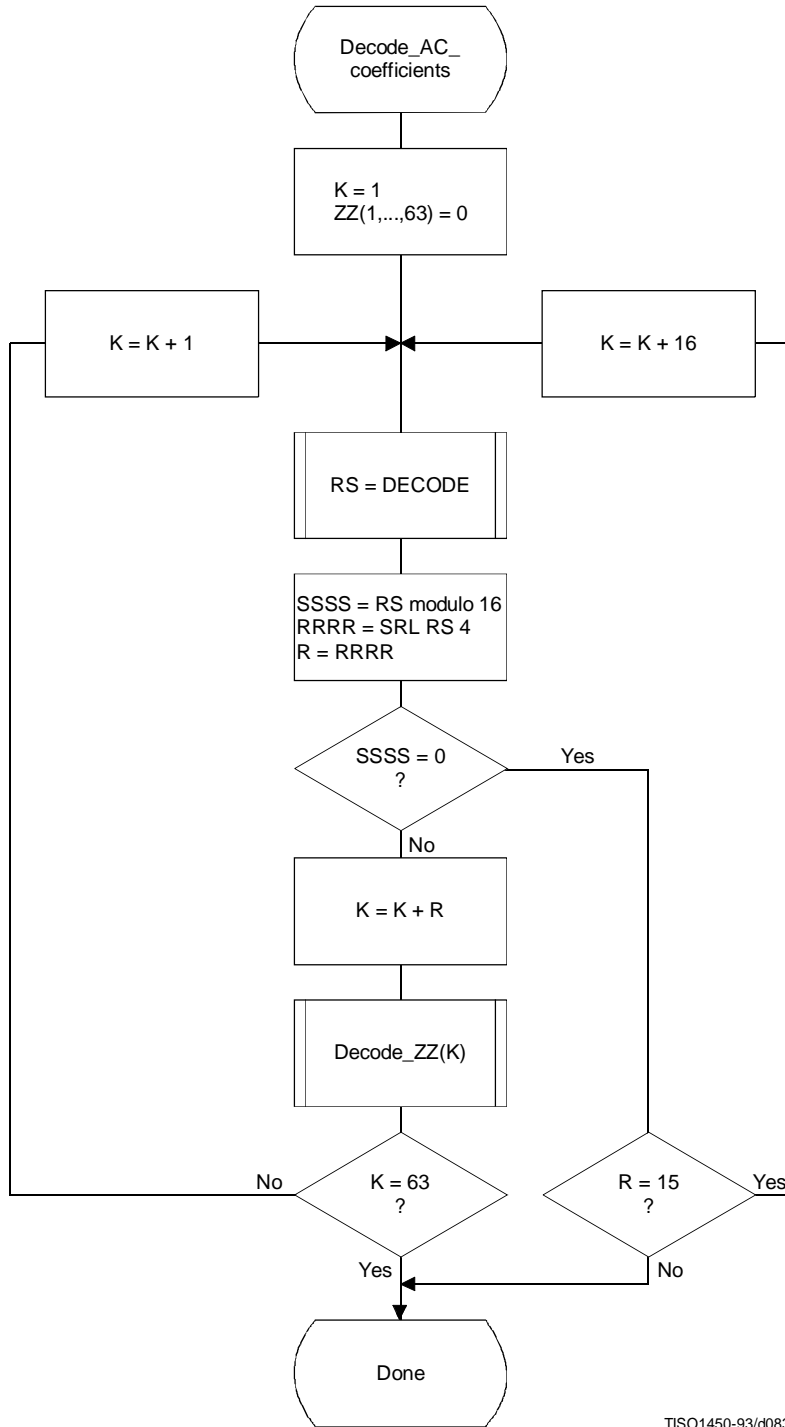


TISO1440-93/d082

Figure F.12 – Extending the sign bit of a decoded value in V

F.2.2.2 Decoding procedure for AC coefficients

The decoding procedure for AC coefficients is shown in Figures F.13 and F.14.



TISO1450-93/d083

Figure F.13 – Huffman decoding procedure for AC coefficients

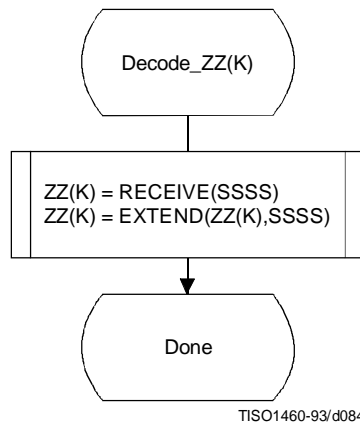


Figure F.14 – Decoding a non-zero AC coefficient

The decoding of the amplitude and sign of the non-zero coefficient is done in the procedure “Decode_ZZ(K)”, shown in Figure F.14.

DECODE is a procedure which returns the value, RS, associated with the next Huffman code in the code stream (see F.2.2.3). The values SSSS and R are derived from RS. The value of SSSS is the four low order bits of the composite value and R contains the value of RRRR (the four high order bits of the composite value). The interpretation of these values is described in F.1.2.2. EXTEND is shown in Figure F.12.

F.2.2.3 The DECODE procedure

The DECODE procedure decodes an 8-bit value which, for the DC coefficient, determines the difference magnitude category. For the AC coefficient this 8-bit value determines the zero run length and non-zero coefficient category.

Three tables, HUFFVAL, HUFFCODE, and HUFFSIZE, have been defined in Annex C. This particular implementation of DECODE makes use of the ordering of the Huffman codes in HUFFCODE according to both value and code size. Many other implementations of DECODE are possible.

NOTE – The values in HUFFVAL are assigned to each code in HUFFCODE and HUFFSIZE in sequence. There are no ordering requirements for the values in HUFFVAL which have assigned codes of the same length.

The implementation of DECODE described in this subclause uses three tables, MINCODE, MAXCODE and VALPTR, to decode a pointer to the HUFFVAL table. MINCODE, MAXCODE and VALPTR each have 16 entries, one for each possible code size. MINCODE(I) contains the smallest code value for a given length I, MAXCODE(I) contains the largest code value for a given length I, and VALPTR(I) contains the index to the start of the list of values in HUFFVAL which are decoded by code words of length I. The values in MINCODE and MAXCODE are signed 16-bit integers; therefore, a value of -1 sets all of the bits.

The procedure for generating these tables is shown in Figure F.15. The procedure for DECODE is shown in Figure F.16. Note that the 8-bit “VALUE” is returned to the procedure which invokes DECODE.

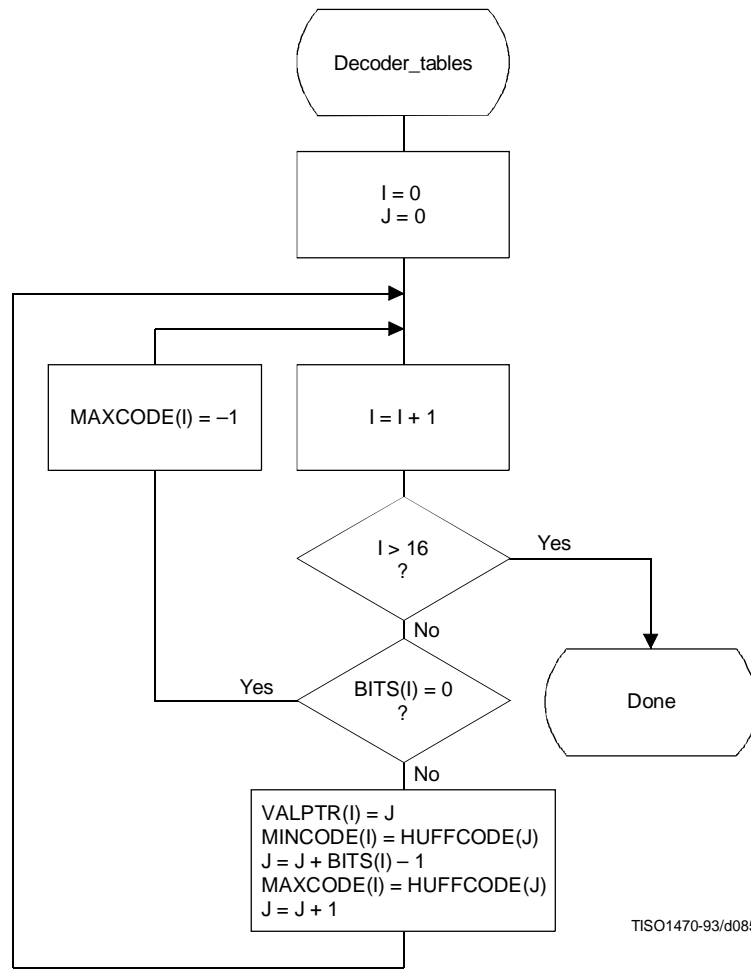


Figure F.15 – Decoder table generation

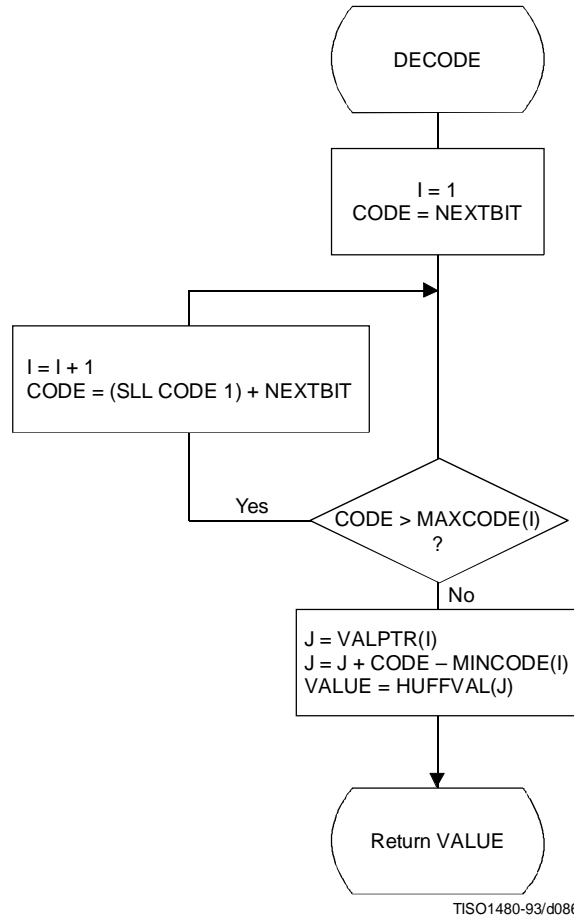


Figure F.16 – Procedure for DECODE

F.2.2.4 The RECEIVE procedure

RECEIVE(SSSS) is a procedure which places the next SSSS bits of the entropy-coded segment into the low order bits of DIFF, MSB first. It calls NEXTBIT and it returns the value of DIFF to the calling procedure (see Figure F.17).

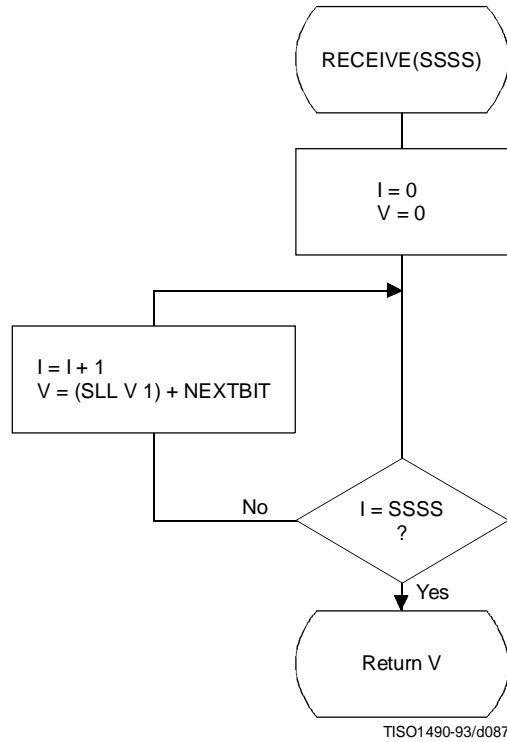


Figure F.17 – Procedure for RECEIVE(SSSS)

F.2.2.5 The NEXTBIT procedure

NEXTBIT reads the next bit of compressed data and passes it to higher level routines. It also intercepts and removes stuff bytes and detects markers. NEXTBIT reads the bits of a byte starting with the MSB (see Figure F.18).

Before starting the decoding of a scan, and after processing a RST marker, CNT is cleared. The compressed data are read one byte at a time, using the procedure NEXTBYTE. Each time a byte, B, is read, CNT is set to 8.

The only valid marker which may occur within the Huffman coded data is the RST_m marker. Other than the EOI or markers which may occur at or before the start of a scan, the only marker which can occur at the end of the scan is the DNL (define-number-of-lines).

Normally, the decoder will terminate the decoding at the end of the final restart interval before the terminating marker is intercepted. If the DNL marker is encountered, the current line count is set to the value specified by that marker. Since the DNL marker can only be used at the end of the first scan, the scan decode procedure must be terminated when it is encountered.

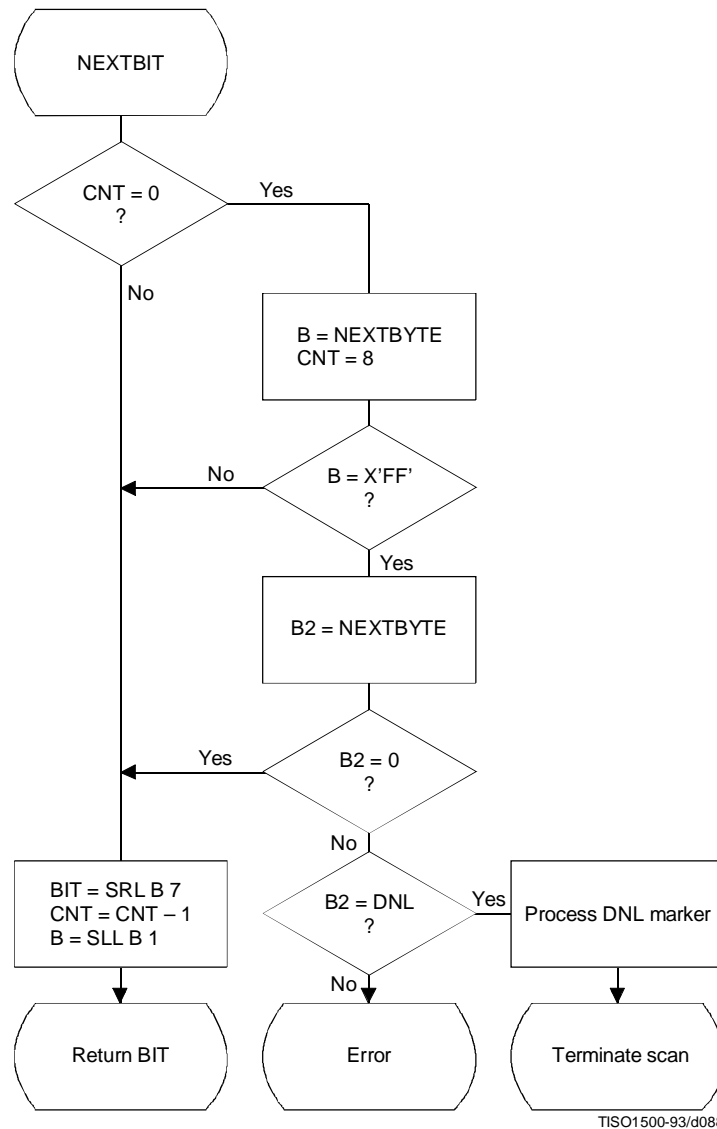


Figure F.18 – Procedure for fetching the next bit of compressed data

F.2.3 Sequential DCT decoding process with 8-bit precision extended to four sets of Huffman tables

This process is identical to the Baseline decoding process described in F.2.2, with the exception that the decoder shall be capable of using up to four DC and four AC Huffman tables within one scan. Four DC and four AC Huffman tables is the maximum allowed by this Specification.

F.2.4 Sequential DCT decoding process with arithmetic coding

This subclause describes the sequential DCT decoding process with arithmetic decoding.

The arithmetic decoding procedures for decoding binary decisions, initializing the statistical model, initializing the decoder, and resynchronizing the decoder are listed in Table D.4 of Annex D.

Some of the procedures in Table D.4 are used in the higher level control structure for scans and restart intervals described in F.2. At the beginning of scans and restart intervals, the probability estimates used in the arithmetic decoder are reset to the standard initial value as part of the Initdec procedure which restarts the arithmetic coder.

The statistical models defined in F.1.4.4 also apply to this decoding process.

The decoder shall be capable of using up to four DC and four AC conditioning tables and associated statistics areas within one scan.

F.2.4.1 Arithmetic decoding of DC coefficients

The basic structure of the decision sequence for decoding a DC difference value, DIFF, is shown in Figure F.19. The equivalent structure for the encoder is found in Figure F.4.

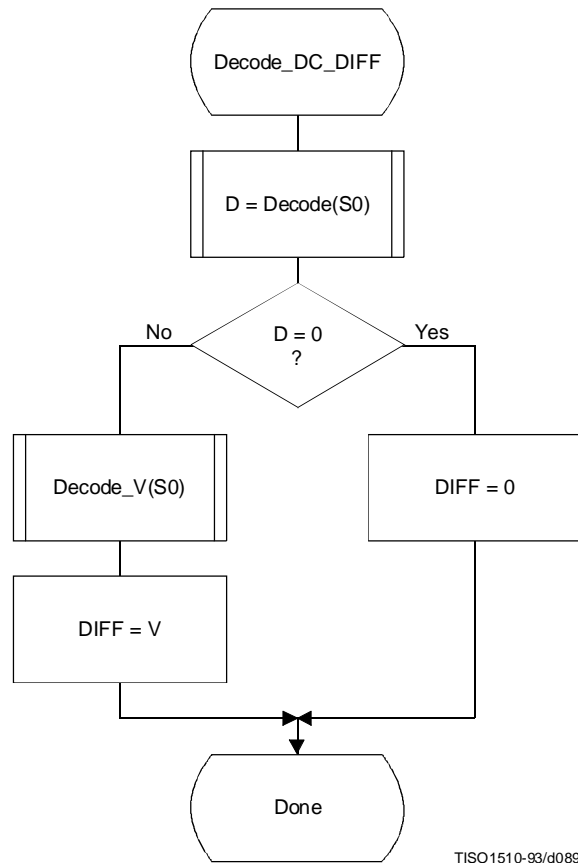


Figure F.19 – Arithmetic decoding of DC difference

The context-indices used in the DC decoding procedures are defined in Table F.4 (see F.1.4.4.1.3).

The “Decode” procedure returns the value “D” of the binary decision. If the value is not zero, the sign and magnitude of the non-zero DIFF must be decoded by the procedure “Decode_V(S0)”.

F.2.4.2 Arithmetic Decoding of AC coefficients

The AC coefficients are decoded in the order that they occur in ZZ(1,...,63). The encoder procedure for the coding process is found in Figure F.5. Figure F.20 illustrates the decoding sequence.

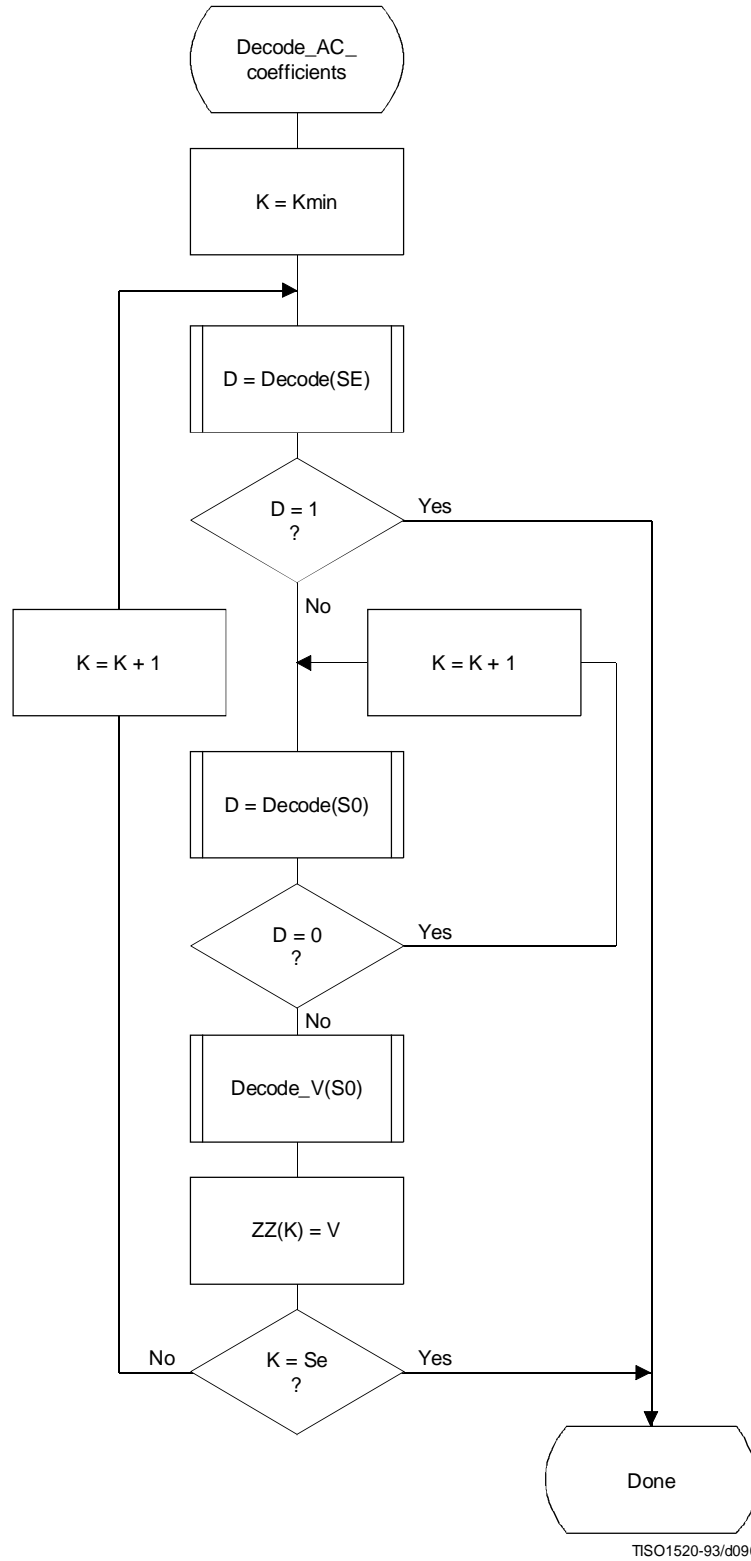


Figure F.20 – Procedure for decoding the AC coefficients

The context-indices used in the AC decoding procedures are defined in Table F.5 (see F.1.4.4.2).

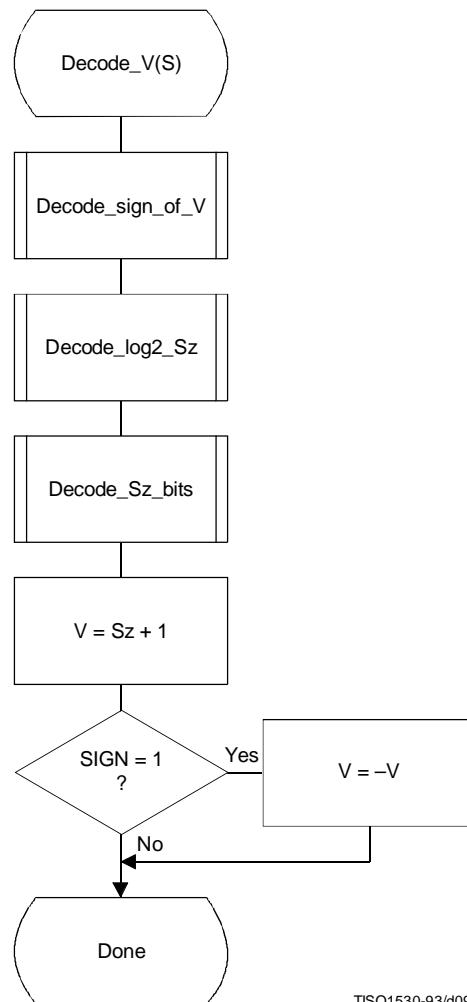
In Figure F.20, K is the index to the zig-zag sequence position. For the sequential scan, $K_{min} = 1$ and $S_e = 63$. The decision at the top of the loop is the EOB decision. If the EOB occurs ($D = 1$), the remaining coefficients in the block are set to zero. The inner loop just below the EOB decoding decodes runs of zero coefficients. Whenever the coefficient is non-zero, "Decode_V" decodes the sign and magnitude of the coefficient. After each non-zero coefficient is decoded, the EOB decision is again decoded unless $K = S_e$.

F.2.4.3 Decoding the binary decision sequence for non-zero DC differences and AC coefficients

Both the DC difference and the AC coefficients are represented as signed two's complement 16-bit integer values. The decoding decision tree for these signed integer values is the same for both the DC and AC coding models. Note, however, that the statistical models are not the same.

F.2.4.3.1 Arithmetic decoding of non-zero values

Denoting either DC differences or AC coefficients as V, the non-zero signed integer value of V is decoded by the sequence shown in Figure F.21. This sequence first decodes the sign of V. It then decodes the magnitude category of V (Decode_log2_Sz), and then decodes the low order magnitude bits (Decode_Sz_bits). Note that the value decoded for Sz must be incremented by 1 to get the actual coefficient magnitude.



TISO1530-93/d091

Figure F.21 – Sequence of procedures in decoding non-zero values of V

F.2.4.3.1.1 Decoding the sign

The sign is decoded by the procedure shown in Figure F.22.

The context-indices are defined for DC decoding in Table F.4 and AC decoding in Table F.5.

If SIGN = 0, the sign of the coefficient is positive; if SIGN = 1, the sign of the coefficient is negative.

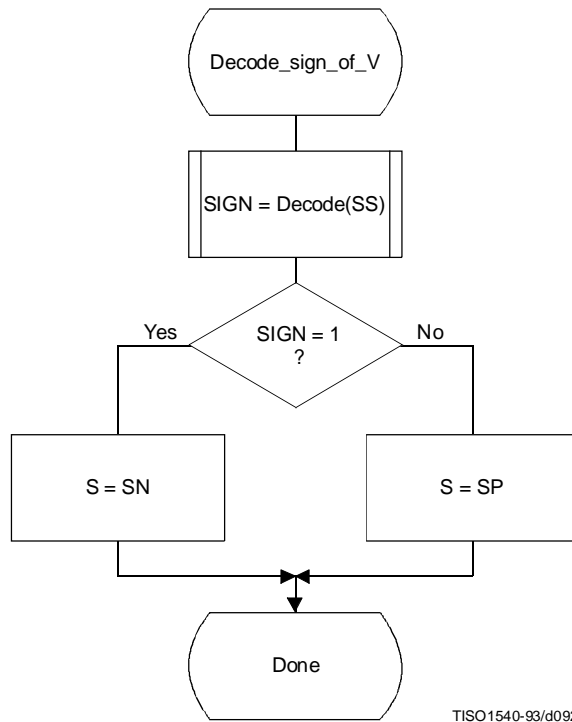


Figure F.22 – Decoding the sign of V

F.2.4.3.1.2 Decoding the magnitude category

The context-index S is set in Decode_sign_of_V and the context-index values X1 and X2 are defined for DC coding in Table F.4 and for AC coding in Table F.5.

In Figure F.23, M is set to the upper bound for the magnitude and shifted left until the decoded decision is zero. It is then shifted right by 1 to become the leading bit of the magnitude of Sz.

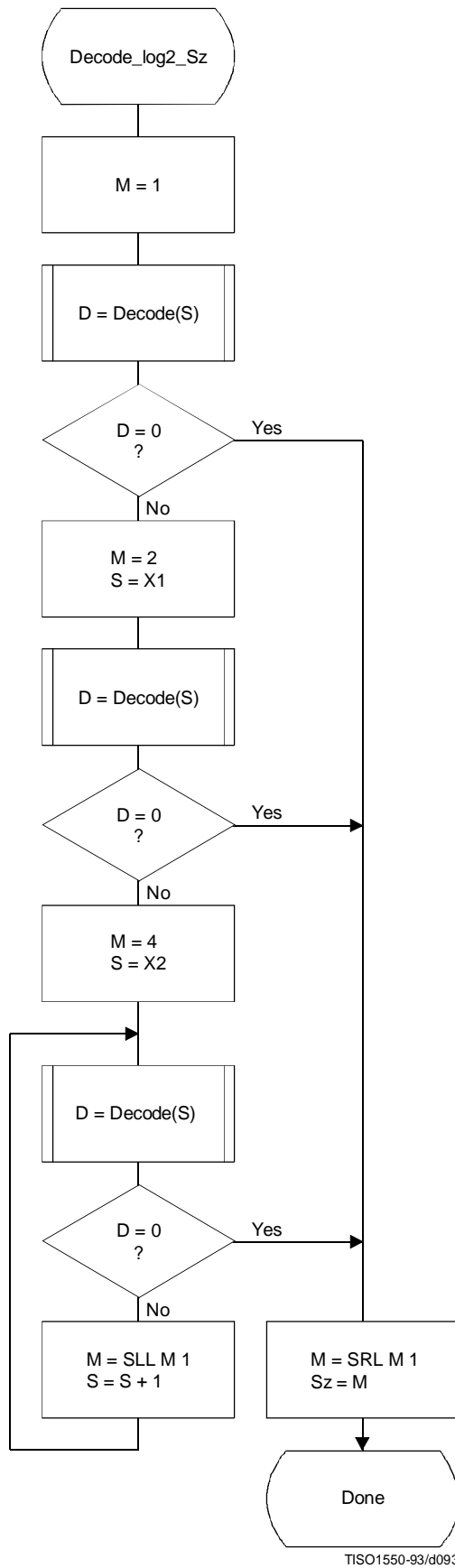


Figure F.23 – Decoding procedure to establish the magnitude category

F.2.4.3.1.3 Decoding the exact value of the magnitude

After the magnitude category is decoded, the low order magnitude bits are decoded. These bits are decoded in order of decreasing bit significance. The procedure is shown in Figure F.24.

The context-index S is set in Decode_log2_Sz.

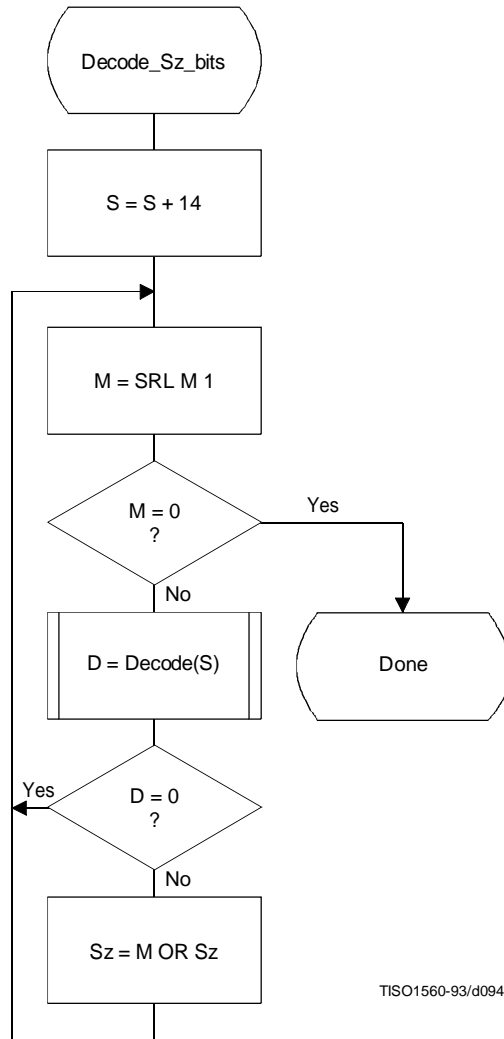


Figure F.24 – Decision sequence to decode the magnitude bit pattern

F.2.4.4 Decoder restart

The RST_m markers which are added to the compressed data between each restart interval have a two byte value which cannot be generated by the coding procedures. These two byte sequences can be located without decoding, and can therefore be used to resynchronize the decoder. RST_m markers can therefore be used for error recovery.

Before error recovery procedures can be invoked, the error condition must first be detected. Errors during decoding can show up in two places:

- a) The decoder fails to find the expected marker at the point where it is expecting resynchronization.
- b) Physically impossible data are decoded. For example, decoding a magnitude beyond the range of values allowed by the model is quite likely when the compressed data are corrupted by errors. For arithmetic decoders this error condition is extremely important to detect, as otherwise the decoder may reach a condition where it uses the compressed data very slowly.

NOTE – Some errors will not cause the decoder to lose synchronization. In addition, recovery is not possible for all errors; for example, errors in the headers are likely to be catastrophic. The two error conditions listed above, however, almost always cause the decoder to lose synchronization in a way which permits recovery.

In regaining synchronization, the decoder can make use of the modulo 8 coding restart interval number in the low order bits of the RST_m marker. By comparing the expected restart interval number to the value in the next RST_m marker in the compressed image data, the decoder can usually recover synchronization. It then fills in missing lines in the output data by replication or some other suitable procedure, and continues decoding. Of course, the reconstructed image will usually be highly corrupted for at least a part of the restart interval where the error occurred.

F.2.5 Sequential DCT decoding process with Huffman coding and 12-bit precision

This process is identical to the sequential DCT process defined for 8-bit sample precision and extended to four Huffman tables, as documented in F.2.3, but with the following changes.

F.2.5.1 Structure of DC Huffman decode table

The general structure of the DC Huffman decode table is extended as described in F.1.5.1.

F.2.5.2 Structure of AC Huffman decode table

The general structure of the AC Huffman decode table is extended as described in F.1.5.2.

F.2.6 Sequential DCT decoding process with arithmetic coding and 12-bit precision

The process is identical to the sequential DCT process for 8-bit precision except for changes in the precision of the IDCT computation.

The structure of the decoding procedure in F.2.4 is already defined for a 12-bit input precision.

Annex G

Progressive DCT-based mode of operation

(This annex forms an integral part of this Recommendation | International Standard)

This annex provides a **functional specification** of the following coding processes for the progressive DCT-based mode of operation:

- 1) spectral selection only, Huffman coding, 8-bit sample precision;
- 2) spectral selection only, arithmetic coding, 8-bit sample precision;
- 3) full progression, Huffman coding, 8-bit sample precision;
- 4) full progression, arithmetic coding, 8-bit sample precision;
- 5) spectral selection only, Huffman coding, 12-bit sample precision;
- 6) spectral selection only, arithmetic coding, 12-bit sample precision;
- 7) full progression, Huffman coding, 12-bit sample precision;
- 8) full progression, arithmetic coding, 12-bit sample precision.

For each of these, the encoding process is specified in G.1, and the decoding process is specified in G.2. The functional specification is presented by means of specific flow charts for the various procedures which comprise these coding processes.

NOTE – There is **no requirement** in this Specification that any encoder or decoder which embodies one of the above-named processes shall implement the procedures in precisely the manner specified by the flow charts in this annex. It is necessary only that an encoder or decoder implement the **function** specified in this annex. The sole criterion for an encoder or decoder to be considered in compliance with this Specification is that it satisfy the requirements given in clause 6 (for encoders) or clause 7 (for decoders), as determined by the compliance tests specified in Part 2.

The number of Huffman or arithmetic conditioning tables which may be used within the same scan is four.

Two complementary progressive procedures are defined, spectral selection and successive approximation.

In spectral selection the DCT coefficients of each block are segmented into frequency bands. The bands are coded in separate scans.

In successive approximation the DCT coefficients are divided by a power of two before coding. In the decoder the coefficients are multiplied by that same power of two before computing the IDCT. In the succeeding scans the precision of the coefficients is increased by one bit in each scan until full precision is reached.

An encoder or decoder implementing a full progression uses spectral selection within successive approximation. An allowed subset is spectral selection alone.

Figure G.1 illustrates the spectral selection and successive approximation progressive processes.

G.1 Progressive DCT-based encoding processes

G.1.1 Control procedures and coding models for progressive DCT-based procedures

G.1.1.1 Control procedures for progressive DCT-based encoders

The control procedures for encoding an image and its constituent parts – the frame, scan, restart interval and MCU – are given in Figures E.1 through E.5.

The control structure for encoding a frame is the same as for the sequential procedures. However, it is convenient to calculate the FDCT for the entire set of components in a frame before starting the scans. A buffer which is large enough to store all of the DCT coefficients may be used for this progressive mode of operation.

The number of scans is determined by the progression defined; the number of scans may be much larger than the number of components in the frame.

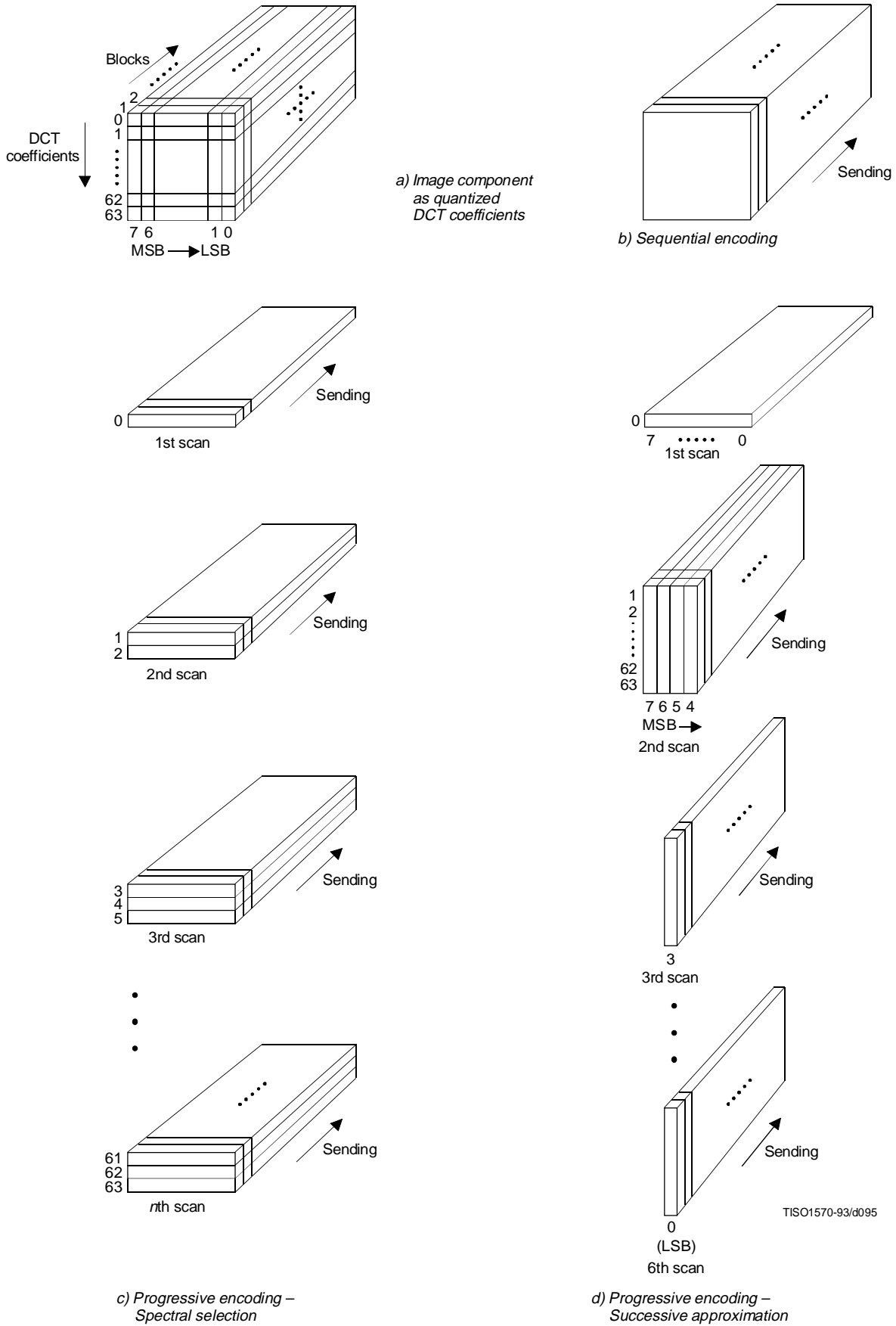


Figure G.1 – Spectral selection and successive approximation progressive processes

The procedure for encoding a MCU (see Figure E.5) repetitively invokes the procedure for coding a data unit. For DCT-based encoders the data unit is an 8×8 block of samples.

Only a portion of each 8×8 block is coded in each scan, the portion being determined by the scan header parameters S_s , S_e , A_h , and A_l (see B.2.3). The procedures used to code portions of each 8×8 block are described in this annex. Note, however, that where these procedures are identical to those used in the sequential DCT-based mode of operation, the sequential procedures are simply referenced.

G.1.1.1.1 Spectral selection control

In spectral selection the zig-zag sequence of DCT coefficients is segmented into bands. A band is defined in the scan header by specifying the starting and ending indices in the zig-zag sequence. One band is coded in a given scan of the progression. DC coefficients are always coded separately from AC coefficients, and only scans which code DC coefficients may have interleaved blocks from more than one component. All other scans shall have only one component. With the exception of the first DC scans for the components, the sequence of bands defined in the scans need not follow the zig-zag ordering. For each component, a first DC scan shall precede any AC scans.

G.1.1.1.2 Successive approximation control

If successive approximation is used, the DCT coefficients are reduced in precision by the point transform (see A.4) defined in the scan header (see B.2.3). The successive approximation bit position parameter A_l specifies the actual point transform, and the high four bits (A_h) – if there are preceding scans for the band – contain the value of the point transform used in those preceding scans. If there are no preceding scans for the band, A_h is zero.

Each scan which follows the first scan for a given band progressively improves the precision of the coefficients by one bit, until full precision is reached.

G.1.1.2 Coding models for progressive DCT-based encoders

If successive approximation is used, the DCT coefficients are reduced in precision by the point transform (see A.4) defined in the scan header (see B.2.3). These models also apply to the progressive DCT-based encoders, but with the following changes.

G.1.1.2.1 Progressive encoding model for DC coefficients

If A_l is not zero, the point transform for DC coefficients shall be used to reduce the precision of the DC coefficients. If A_l is zero, the coefficient values (as modified by the point transform) shall be coded, using the procedure described in Annex F. If A_h is not zero, the least significant bit of the point transformed DC coefficients shall be coded, using the procedures described in this annex.

G.1.1.2.2 Progressive encoding model for AC coefficients

If A_l is not zero, the point transform for AC coefficients shall be used to reduce the precision of the AC coefficients. If A_l is zero, the coefficient values (as modified by the point transform) shall be coded using modifications of the procedures described in Annex F. These modifications are described in this annex. If A_h is not zero, the precision of the coefficients shall be improved using the procedures described in this annex.

G.1.2 Progressive encoding procedures with Huffman coding

G.1.2.1 Progressive encoding of DC coefficients with Huffman coding

The first scan for a given component shall encode the DC coefficient values using the procedures described in F.1.2.1. If the successive approximation bit position parameter A_l is not zero, the coefficient values shall be reduced in precision by the point transform described in Annex A before coding.

In subsequent scans using successive approximation the least significant bits are appended to the compressed bit stream without compression or modification (see G.1.2.3), except for byte stuffing.

G.1.2.2 Progressive encoding of AC coefficients with Huffman coding

In spectral selection and in the first scan of successive approximation for a component, the AC coefficient coding model is similar to that used by the sequential procedures. However, the Huffman code tables are extended to include coding of runs of End-Of-Bands (EOBs). See Table G.1.

Table G.1 – EOBn code run length extensions

EOBn code	Run length
EOB0	1
EOB1	2,3
EOB2	4..7
EOB3	8..15
EOB4	16..31
EOB5	32..63
EOB6	64..127
EOB7	128..255
EOB8	256..511
EOB9	512..1 023
EOB10	1 024..2 047
EOB11	2 048..4 095
EOB12	4 096..8 191
EOB13	8 192..16 383
EOB14	16 384..32 767

The end-of-band run structure allows efficient coding of blocks which have only zero coefficients. An EOB run of length 5 means that the current block and the next four blocks have an end-of-band with no intervening non-zero coefficients. The EOB run length is limited only by the restart interval.

The extension of the code table is illustrated in Figure G.2.

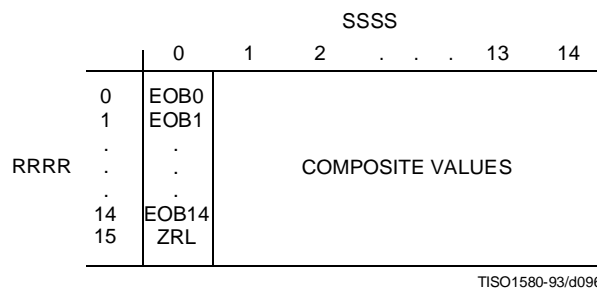


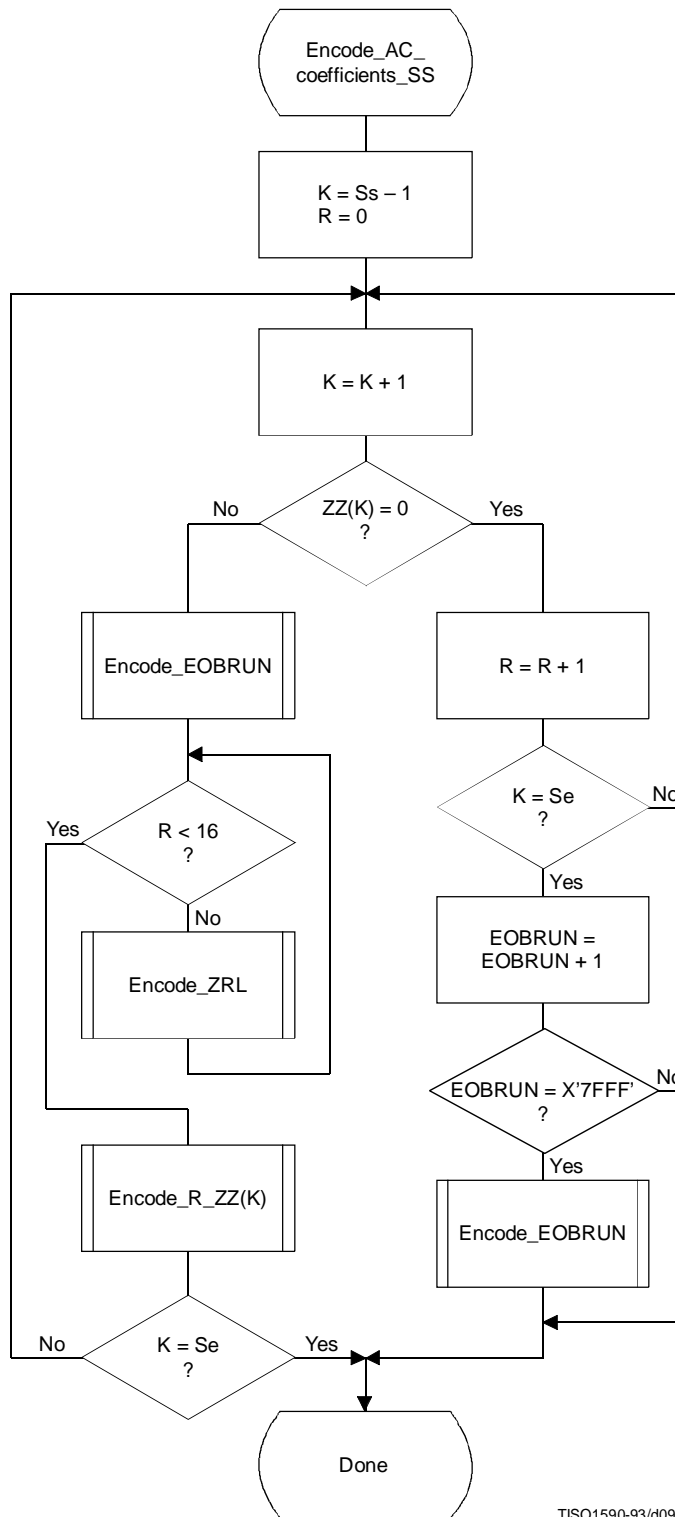
Figure G.2 – Two-dimensional value array for Huffman coding

The EOBn code sequence is defined as follows. Each EOBn code is followed by an extension field similar to the extension field for the coefficient amplitudes (but with positive numbers only). The number of bits appended to the EOBn code is the minimum number required to specify the run length.

If an EOB run is greater than 32 767, it is coded as a sequence of EOB runs of length 32 767 followed by a final EOB run sufficient to complete the run.

At the beginning of each restart interval the EOB run count, EOBRUN, is set to zero. At the end of each restart interval any remaining EOB run is coded.

The Huffman encoding procedure for AC coefficients in spectral selection and in the first scan of successive approximation is illustrated in Figures G.3, G.4, G.5, and G.6.



TISO1590-93/d097

Figure G.3 – Procedure for progressive encoding of AC coefficients with Huffman coding

In Figure G.3, S_s is the start of spectral selection, S_e is the end of spectral selection, K is the index into the list of coefficients stored in the zig-zag sequence ZZ , R is the run length of zero coefficients, and $EOBRUN$ is the run length of EOBs. $EOBRUN$ is set to zero at the start of each restart interval.

If the scan header parameter A_l (successive approximation bit position low) is not zero, the DCT coefficient values $ZZ(K)$ in Figure G.3 and figures which follow in this annex, including those in the arithmetic coding section, shall be replaced by the point transformed values $ZZ'(K)$, where $ZZ'(K)$ is defined by:

$$ZZ'(K) = \frac{ZZ(K) \times X}{2^{A_l}}$$

$EOBSIZE$ is a procedure which returns the size of the EOB extension field given the EOB run length as input. $CSIZE$ is a procedure which maps an AC coefficient to the $SSSS$ value defined in the subclauses on sequential encoding (see F.1.1 and F.1.3).

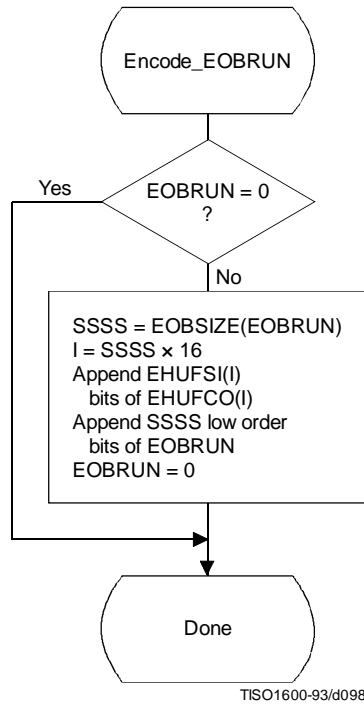


Figure G.4 – Progressive encoding of a non-zero AC coefficient

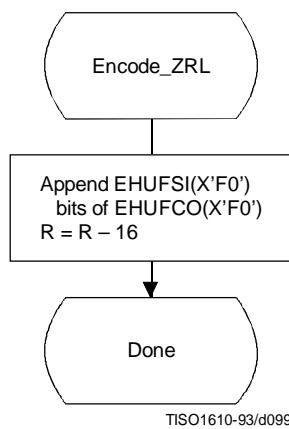


Figure G.5 – Encoding of the run of zero coefficients

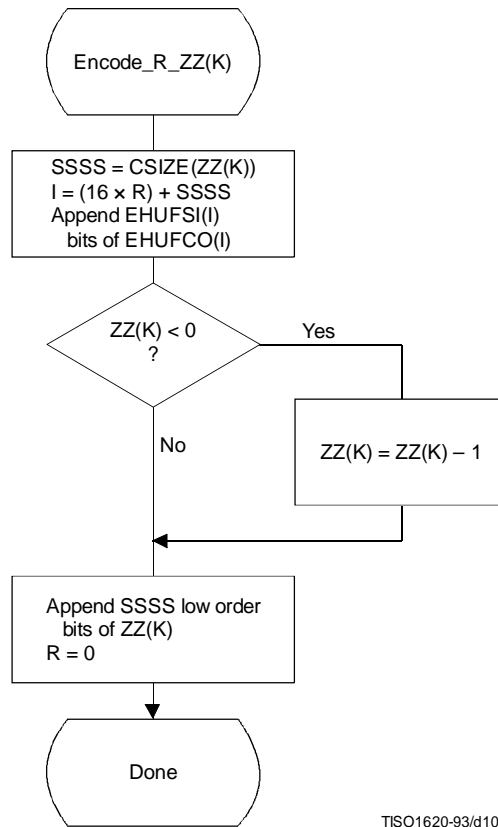


Figure G.6 – Encoding of the zero run and non-zero coefficient

G.1.2.3 Coding model for subsequent scans of successive approximation

The Huffman coding structure of the subsequent scans of successive approximation for a given component is similar to the coding structure of the first scan of that component.

The structure of the AC code table is identical to the structure described in G.1.2.2. Each non-zero point transformed coefficient that has a zero history (i.e. that has a value ± 1 , and therefore has not been coded in a previous scan) is defined by a composite 8-bit run length-magnitude value of the form:

RRRRSSSS

The four most significant bits, RRRR, give the number of zero coefficients that are between the current coefficient and the previously coded coefficient (or the start of band). Coefficients with non-zero history (a non-zero value coded in a previous scan) are skipped over when counting the zero coefficients. The four least significant bits, SSSS, provide the magnitude category of the non-zero coefficient; for a given component the value of SSSS can only be one.

The run length-magnitude composite value is Huffman coded and each Huffman code is followed by additional bits:

- a) One bit codes the sign of the newly non-zero coefficient. A 0-bit codes a negative sign; a 1-bit codes a positive sign.
- b) For each coefficient with a non-zero history, one bit is used to code the correction. A 0-bit means no correction and a 1-bit means that one shall be added to the (scaled) decoded magnitude of the coefficient.

Non-zero coefficients with zero history are coded with a composite code of the form:

$$\text{HUFFCO(RRRRSSSS)} + \text{additional bit (rule a)} + \text{correction bits (rule b)}$$

In addition whenever zero runs are coded with ZRL or EOB_n codes, correction bits for those coefficients with non-zero history contained within the zero run are appended according to rule b above.

For the Huffman coding version of Encode_AC_Coefficients_SA the EOB is defined to be the position of the last point transformed coefficient of magnitude 1 in the band. If there are no coefficients of magnitude 1, the EOB is defined to be zero.

NOTE – The definition of EOB is different for Huffman and arithmetic coding procedures.

In Figures G.7 and G.8 BE is the count of buffered correction bits at the start of coding of the block. BE is initialized to zero at the start of each restart interval. At the end of each restart interval any remaining buffered bits are appended to the bit stream following the last EOB_n Huffman code and associated appended bits.

In Figures G.7 and G.9, BR is the count of buffered correction bits which are appended to the bit stream according to rule b. BR is set to zero at the beginning of each Encode_AC_Coefficients_SA. At the end of each restart interval any remaining buffered bits are appended to the bit stream following the last Huffman code and associated appended bits.

G.1.3 Progressive encoding procedures with arithmetic coding

G.1.3.1 Progressive encoding of DC coefficients with arithmetic coding

The first scan for a given component shall encode the DC coefficient values using the procedures described in F.1.4.1. If the successive approximation bit position parameter is not zero, the coefficient values shall be reduced in precision by the point transform described in Annex A before coding.

In subsequent scans using successive approximation the least significant bits shall be coded as binary decisions using a fixed probability estimate of 0.5 ($Q_e = X'5A1D'$, MPS = 0).

G.1.3.2 Progressive encoding of AC coefficients with arithmetic coding

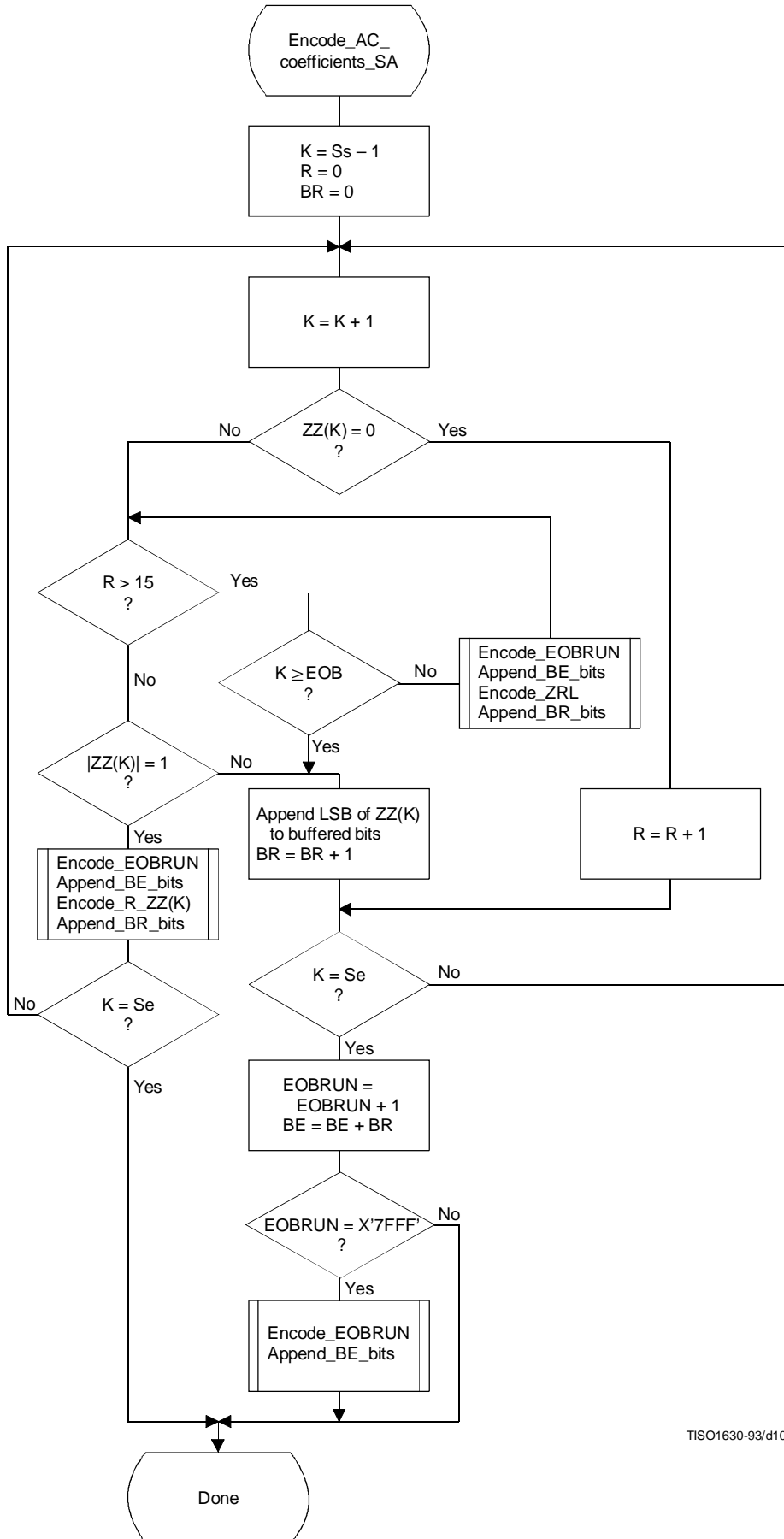
Except for the point transform scaling of the DCT coefficients and the grouping of the coefficients into bands, the first scan(s) of successive approximation is identical to the sequential encoding procedure described in F.1.4. If K_{min} is equated to S_s, the index of the first AC coefficient index in the band, the flow chart shown in Figure F.5 applies. The EOB decision in that figure refers to the “end-of-band” rather than the “end-of-block”. For the arithmetic coding version of Encode_AC_Coefficients_SA (and all other AC coefficient coding procedures) the EOB is defined to be the position following the last non-zero coefficient in the band.

NOTE - The definition of EOB is different for Huffman and arithmetic coding procedures.

The statistical model described in F.1.4 also holds. For this model the default value of K_x is 5. Other values of K_x may be specified using the DAC marker code (Annex B). The following calculation for K_x has proven to give good results for 8-bit precision samples:

$$K_x = K_{min} + SRL (8 + S_e - K_{min}) / 4$$

This expression reduces to the default of K_x = 5 when the band is from index 1 to index 63.



TISO1630-93/d101

Figure G.7 – Successive approximation coding of AC coefficients using Huffman coding

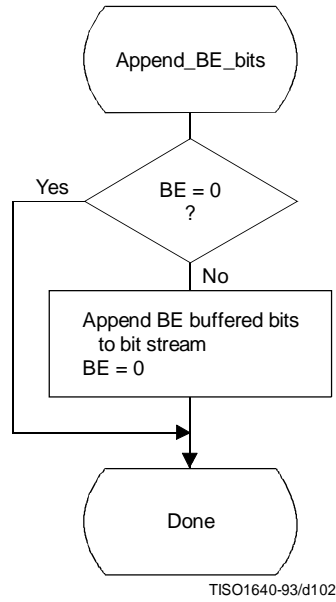


Figure G.8 – Transferring BE buffered bits from buffer to bit stream

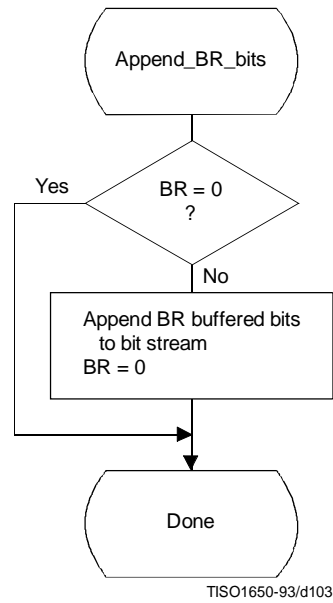


Figure G.9 – Transferring BR buffered bits from buffer to bit stream

G.1.3.3 Coding model for subsequent scans of successive approximation

The procedure "Encode_AC_Coefficient_SA" shown in Figure G.10 increases the precision of the AC coefficient values in the band by one bit.

As in the first scan of successive approximation for a component, an EOB decision is coded at the start of the band and after each non-zero coefficient.

However, since the end-of-band index of the previous successive approximation scan for a given component, EOB_x , is known from the data coded in the prior scan of that component, this decision is bypassed whenever the current index, K , is less than EOB_x . As in the first scan(s), the EOB decision is also bypassed whenever the last coefficient in the band is not zero. The decision $ZZ(K) = 0$ decodes runs of zero coefficients. If the decoder is at this step of the procedure, at least one non-zero coefficient remains in the band of the block being coded. If $ZZ(K)$ is not zero, the procedure in Figure G.11 is followed to code the value.

The context-indices in Figures G.10 and G.11 are defined in Table G.2 (see G.1.3.3.1). The signs of coefficients with magnitude of one are coded with a fixed probability value of approximately 0.5 ($Q_e = X'5A1D'$, $MPS = 0$).

G.1.3.3.1 Statistical model for subsequent successive approximation scans

As shown in Table G.2, each statistics area for subsequent successive approximation scans of AC coefficients consists of a contiguous set of 189 statistics bins. The signs of coefficients with magnitude of one are coded with a fixed probability value of approximately 0.5 ($Q_e = X'5A1D'$, $MPS = 0$).

G.2 Progressive decoding of the DCT

The description of the computation of the IDCT and the dequantization procedure contained in A.3.3 and A.3.4 apply to the progressive operation.

Progressive decoding processes must be able to decompress compressed image data which requires up to four sets of Huffman or arithmetic coder conditioning tables within a scan.

In order to avoid repetition, detailed flow diagrams of progressive decoder operation are not included. Decoder operation is defined by reversing the function of each step described in the encoder flow charts, and performing the steps in reverse order.

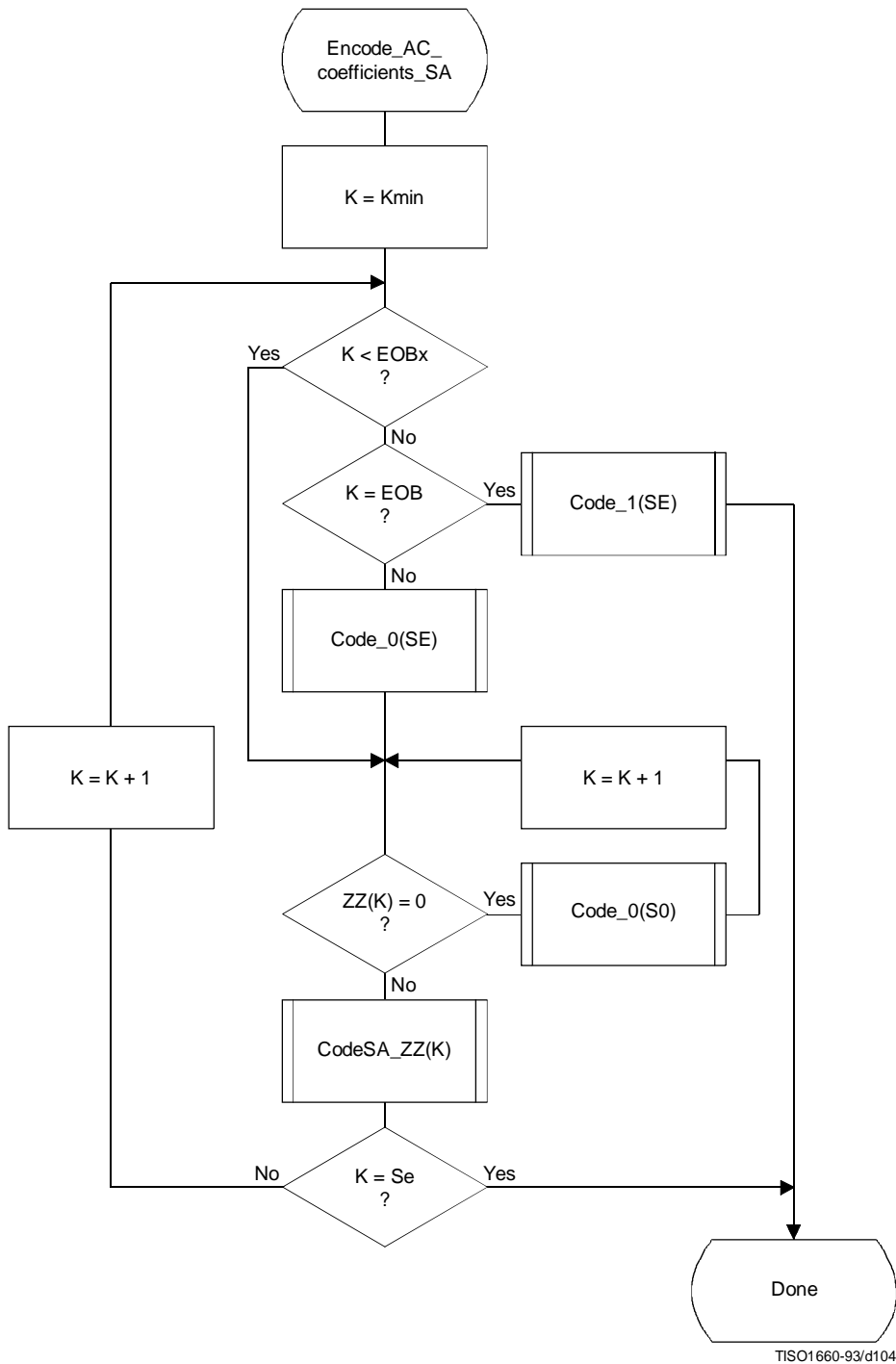


Figure G.10 – Subsequent successive approximation scans for coding of AC coefficients using arithmetic coding

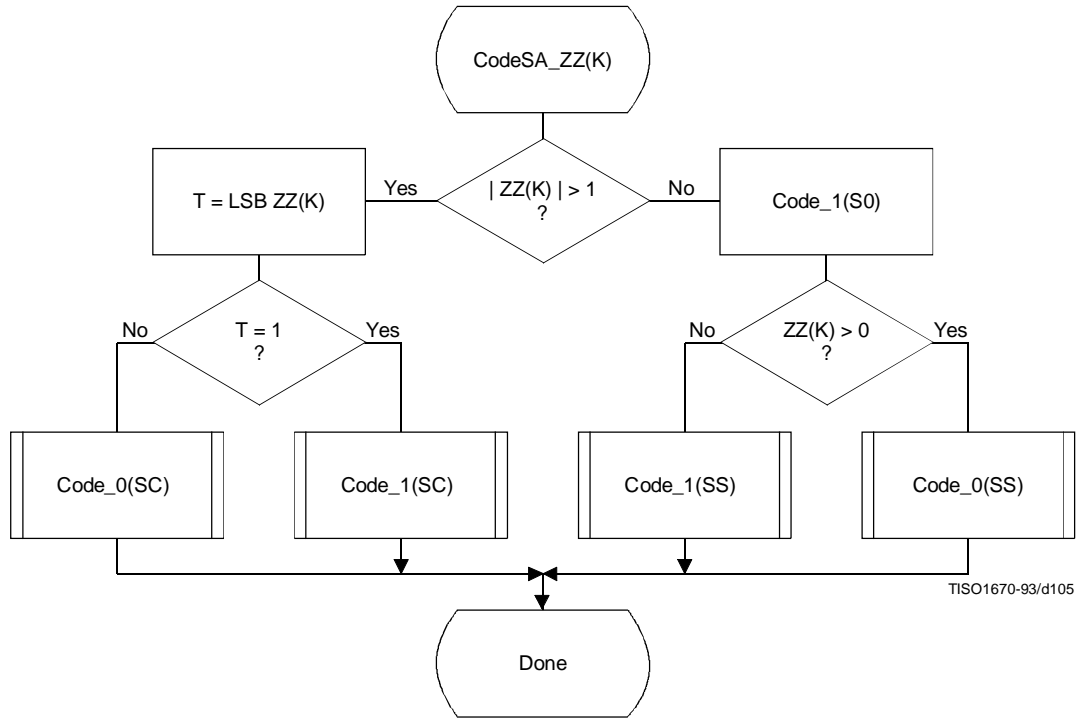


Figure G.11 – Coding non-zero coefficients for subsequent successive approximation scans

Table G.2 – Statistical model for subsequent scans of successive approximation coding of AC coefficient

Context-index	AC coding	Coding decision
SE	$3 \times (K-1)$	$K = \text{EOB}$
S0	$SE + 1$	$V = 0$
SS	Fixed estimate	Sign
SC	$S0 + 1$	$\text{LSB } ZZ(K) = 1$

Annex H

Lossless mode of operation

(This annex forms an integral part of this Recommendation | International Standard)

This annex provides a **functional specification** of the following coding processes for the lossless mode of operation:

- 1) lossless processes with Huffman coding;
- 2) lossless processes with arithmetic coding.

For each of these, the encoding process is specified in H.1, and the decoding process is specified in H.2. The functional specification is presented by means of specific procedures which comprise these coding processes.

NOTE – There is **no requirement** in this Specification that any encoder or decoder which embodies one of the above-named processes shall implement the procedures in precisely the manner specified in this annex. It is necessary only that an encoder or decoder implement the **function** specified in this annex. The sole criterion for an encoder or decoder to be considered in compliance with this Specification is that it satisfy the requirements given in clause 6 (for encoders) or clause 7 (for decoders), as determined by the compliance tests specified in Part 2.

The processes which provide for sequential lossless encoding and decoding are not based on the DCT. The processes used are spatial processes based on the coding model developed for the DC coefficients of the DCT. However, the model is extended by incorporating a set of selectable one- and two-dimensional predictors, and for interleaved data the ordering of samples for the one-dimensional predictor can be different from that used in the DCT-based processes.

Either Huffman coding or arithmetic coding entropy coding may be employed for these lossless encoding and decoding processes. The Huffman code table structure is extended to allow up to 16-bit precision for the input data. The arithmetic coder statistical model is extended to a two-dimensional form.

H.1 Lossless encoder processes

H.1.1 Lossless encoder control procedures

Subclause E.1 contains the encoder control procedures. In applying these procedures to the lossless encoder, the data unit is one sample.

Input data precision may be from 2 to 16 bits/sample. If the input data path has different precision from the input data, the data shall be aligned with the least significant bits of the input data path. Input data is represented as unsigned integers and is not level shifted prior to coding.

When the encoder is reset in the restart interval control procedure (see E.1.4), the prediction is reset to a default value. If arithmetic coding is used, the statistics are also reset.

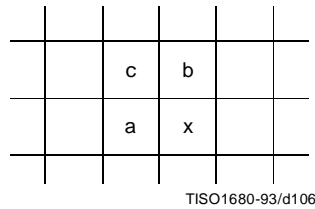
For the lossless processes the restart interval shall be an integer multiple of the number of MCU in an MCU-row.

H.1.2 Coding model for lossless encoding

The coding model developed for encoding the DC coefficients of the DCT is extended to allow a selection from a set of seven one-dimensional and two-dimensional predictors. The predictor is selected in the scan header (see Annex B). The same predictor is used for all components of the scan. Each component in the scan is modeled independently, using predictions derived from neighbouring samples of that component.

H.1.2.1 Prediction

Figure H.1 shows the relationship between the positions (a, b, c) of the reconstructed neighboring samples used for prediction and the position of x, the sample being coded.



TISO1680-93/d106

Figure H.1 – Relationship between sample and prediction samples

Define P_x to be the prediction and R_a , R_b , and R_c to be the reconstructed samples immediately to the left, immediately above, and diagonally to the left of the current sample. The allowed predictors, one of which is selected in the scan header, are listed in Table H.1.

Table H.1 – Predictors for lossless coding

Selection-value	Prediction
0	No prediction (See Annex J)
1	$P_x = R_a$
2	$P_x = R_b$
3	$P_x = R_c$
4	$P_x = R_a + R_b - R_c$
5	$P_x = R_a + ((R_b - R_c)/2)^{a)}$
6	$P_x = R_b + ((R_a - R_c)/2)^{a)}$
7	$P_x = (R_a + R_b)/2$
a) Shift right arithmetic operation	

Selection-value 0 shall only be used for differential coding in the hierarchical mode of operation. Selections 1, 2 and 3 are one-dimensional predictors and selections 4, 5, 6, and 7 are two-dimensional predictors.

The one-dimensional horizontal predictor (prediction sample R_a) is used for the first line of samples at the start of the scan and at the beginning of each restart interval. The selected predictor is used for all other lines. The sample from the line above (prediction sample R_b) is used at the start of each line, except for the first line. At the beginning of the first line and at the beginning of each restart interval the prediction value of $2^P - 1$ is used, where P is the input precision.

If the point transformation parameter (see A.4) is non-zero, the prediction value at the beginning of the first lines and the beginning of each restart interval is $2^P - P_t - 1$, where P_t is the value of the point transformation parameter.

Each prediction is calculated with full integer arithmetic precision, and without clamping of either underflow or overflow beyond the input precision bounds. For example, if R_a and R_b are both 16-bit integers, the sum is a 17-bit integer. After dividing the sum by 2 (predictor 7), the prediction is a 16-bit integer.

For simplicity of implementation, the divide by 2 in the prediction selections 5 and 6 of Table H.1 is done by an arithmetic-right-shift of the integer values.

The difference between the prediction value and the input is calculated modulo 2^{16} . In the decoder the difference is decoded and added, modulo 2^{16} , to the prediction.

H.1.2.2 Huffman coding of the modulo difference

The Huffman coding procedures defined in Annex F for coding the DC coefficients are used to code the modulo 2^{16} differences. The table for DC coding contained in Tables F.1 and F.6 is extended by one additional entry. No extra bits are appended after SSSS = 16 is encoded. See Table H.2.

Table H.2 – Difference categories for lossless Huffman coding

SSSS	Difference values
0	0
1	-1,1
2	-3,-2,2,3
3	-7..-4,4..7
4	-15..-8,8..15
5	-31..-16,16..31
6	-63..-32,32..63
7	-127..-64,64..127
8	-255..-128,128..255
9	-511..-256,256..511
10	-1 023..-512,512..1 023
11	-2 047..-1 024,1 024..2 047
12	-4 095..-2 048,2 048..4 095
13	-8 191..-4 096,4 096..8 191
14	-16 383..-8 192,8 192..16 383
15	-32 767..-16 384,16 384..32 767
16	32 768

H.1.2.3 Arithmetic coding of the modulo difference

The statistical model defined for the DC coefficient arithmetic coding model (see F.1.4.4.1) is generalized to a two-dimensional form in which differences coded for the sample to the left and for the line above are used for conditioning.

H.1.2.3.1 Two-dimensional statistical model

The binary decisions are conditioned on the differences coded for the neighbouring samples immediately above and immediately to the left from the same component. As in the coding of the DC coefficients, the differences are classified into 5 categories: zero(0), small positive (+S), small negative (-S), large positive (+L), and large negative (-L). The two independent difference categories combine to give 25 different conditioning states. Figure H.2 shows the two-dimensional array of conditioning indices. For each of the 25 conditioning states probability estimates for four binary decisions are kept.

At the beginning of the scan and each restart interval the conditioning derived from the line above is set to zero for the first line of each component. At the start of each line, the difference to the left is set to zero for the purposes of calculating the conditioning.

		Difference above (position b)				
		0	+S	-S	+L	-L
Difference to left (position a)	0	0	4	8	12	16
	+S	20	24	28	32	36
	-S	40	44	48	52	56
	+L	60	64	68	72	76
	-L	80	84	88	92	96

TISO1690-93/d107

Figure H.2 – 5 × 5 Conditioning array for two-dimensional statistical model

H.1.2.3.2 Assignment of statistical bins to the DC binary decision tree

Each statistics area for lossless coding consists of a contiguous set of 158 statistics bins. The first 100 bins consist of 25 sets of four bins selected by a context-index S0. The value of S0 is given by L_Context(Da,Db), which provides a value of 0, 4, ..., 92 or 96, depending on the difference classifications of Da and Db (see H.1.2.3.1). The value for S0 provided by L_Context(Da,Db) is from the array in Figure H.2.

The remaining 58 bins consist of two sets of 29 bins, X1, ..., X15, M2, ..., M15, which are used to code magnitude category decisions and magnitude bits. The value of X1 is given by X1_Context(Db), which provides a value of 100 when Db is in the zero, small positive or small negative categories and a value of 129 when Db is in the large positive or large negative categories.

The assignment of statistical bins to the binary decision tree used for coding the difference is given in Table H.3.

Table H.3 – Statistical model for lossless coding

Context-index	Value	Coding decision
S0	L_Context(Da,Db)	V = 0
SS	S0 + 1	Sign
SP	S0 + 2	Sz < 1 if V > 0
SN	S0 + 3	Sz < 1 if V < 0
X1	X1_Context(Db)	Sz < 2
X2	X1 + 1	Sz < 4
X3	X1 + 2	Sz < 8
.	.	.
.	.	.
X15	X1 + 14	Sz < 2 ¹⁵
M2	X2 + 14	Magnitude bits if Sz < 4
M3	X3 + 14	Magnitude bits if Sz < 8
.	.	.
.	.	.
M15	X15 + 14	Magnitude bits if Sz < 2 ¹⁵

H.1.2.3.3 Default conditioning bounds

The bounds, L and U, for determining the conditioning category have the default values $L = 0$ and $U = 1$. Other bounds may be set using the DAC (Define-Arithmetic-Conditioning) marker segment, as described in Annex B.

H.1.2.3.4 Initial conditions for statistical model

At the start of a scan and at each restart, all statistics bins are re-initialized to the standard default value described in Annex D.

H.2 Lossless decoder processes

Lossless decoders may employ either Huffman decoding or arithmetic decoding. They shall be capable of using up to four tables in a scan. Lossless decoders shall be able to decode encoded image source data with any input precision from 2 to 16 bits per sample.

H.2.1 Lossless decoder control procedures

Subclause E.2 contains the decoder control procedures. In applying these procedures to the lossless decoder the data unit is one sample.

When the decoder is reset in the restart interval control procedure (see E.2.4) the prediction is reset to the same value used in the encoder (see H.1.2.1). If arithmetic coding is used, the statistics are also reset.

Restrictions on the restart interval are specified in H.1.1.

H.2.2 Coding model for lossless decoding

The predictor calculations defined in H.1.2 also apply to the lossless decoder processes.

The lossless decoders, decode the differences and add them, modulo 2^{16} , to the predictions to create the output. The lossless decoders shall be able to interpret the point transform parameter, and if non-zero, multiply the output of the lossless decoder by 2^{Pt} .

In order to avoid repetition, detailed flow charts of the lossless decoding procedures are omitted.

Annex J

Hierarchical mode of operation

(This annex forms an integral part of this Recommendation | International Standard)

This annex provides a **functional specification** of the coding processes for the hierarchical mode of operation.

In the hierarchical mode of operation each component is encoded or decoded in a non-differential frame. Such frames may be followed by a sequence of differential frames. A non-differential frame shall be encoded or decoded using the procedures defined in Annexes F, G and H. Differential frame procedures are defined in this annex.

The coding process for a hierarchical encoding containing DCT-based processes is defined as the highest numbered process listed in Table J.1 which is used to code any non-differential DCT-based or differential DCT-based frame in the compressed image data format. The coding process for a hierarchical encoding containing only lossless processes is defined to be the process used for the non-differential frames.

Table J.1 – Coding processes for hierarchical mode

Process	Non-differential frame specification	
1	Extended sequential DCT, Huffman, 8-bit	Annex F, process 2
2	Extended sequential DCT, arithmetic, 8-bit	Annex F, process 3
3	Extended sequential DCT, Huffman, 12-bit	Annex F, process 4
4	Extended sequential DCT, arithmetic, 12-bit	Annex F, process 5
5	Spectral selection only, Huffman, 8-bit	Annex G, process 1
6	Spectral selection only, arithmetic, 8-bit	Annex G, process 2
7	Full progression, Huffman, 8-bit	Annex G, process 3
8	Full progression, arithmetic, 8-bit	Annex G, process 4
9	Spectral selection only, Huffman, 12-bit	Annex G, process 5
10	Spectral selection only, arithmetic, 12-bit	Annex G, process 6
11	Full progression, Huffman, 12-bit	Annex G, process 7
12	Full progression, arithmetic, 12-bit	Annex G, process 8
13	Lossless, Huffman, 2 through 16 bits	Annex H, process 1
14	Lossless, arithmetic, 2 through 16 bits	Annex H, process 2

Hierarchical mode syntax requires a DHP marker segment that appears before the non-differential frame or frames. It may include EXP marker segments and differential frames which shall follow the initial non-differential frame. The frame structure in hierarchical mode is identical to the frame structure in non-hierarchical mode.

Either all non-differential frames within an image shall be coded with DCT-based processes, or all non-differential frames shall be coded with lossless processes. All frames within an image must use the same entropy coding procedure, either Huffman or arithmetic, with the exception that non-differential frames coded with the baseline process may occur in the same image with frames coded with arithmetic coding processes.

If the non-differential frames use DCT-based processes, all differential frames except the final frame for a component shall use DCT-based processes. The final differential frame for each component may use a differential lossless process.

If the non-differential frames use lossless processes, all differential frames shall use differential lossless processes.

For each of the processes listed in Table J.1, the encoding processes are specified in J.1, and decoding processes are specified in J.2.

NOTE – There is **no requirement** in this Specification that any encoder or decoder which embodies one of the above-named processes shall implement the procedures in precisely the manner specified by the flow charts in this annex. It is necessary only that an encoder or decoder implement the **function** specified in this annex. The sole criterion for an encoder or decoder to be considered in compliance with this Specification is that it satisfy the requirements given in clause 6 (for encoders) or clause 7 (for decoders), as determined by the compliance tests specified in Part 2.

In the hierarchical mode of operation each component is encoded or decoded in a non-differential frame followed by a sequence of differential frames. A non-differential frame shall use the procedures defined in Annexes F, G, and H. Differential frame procedures are defined in this annex.

J.1 Hierarchical encoding

J.1.1 Hierarchical control procedure for encoding an image

The control structure for encoding of an image using the hierarchical mode is given in Figure J.1.

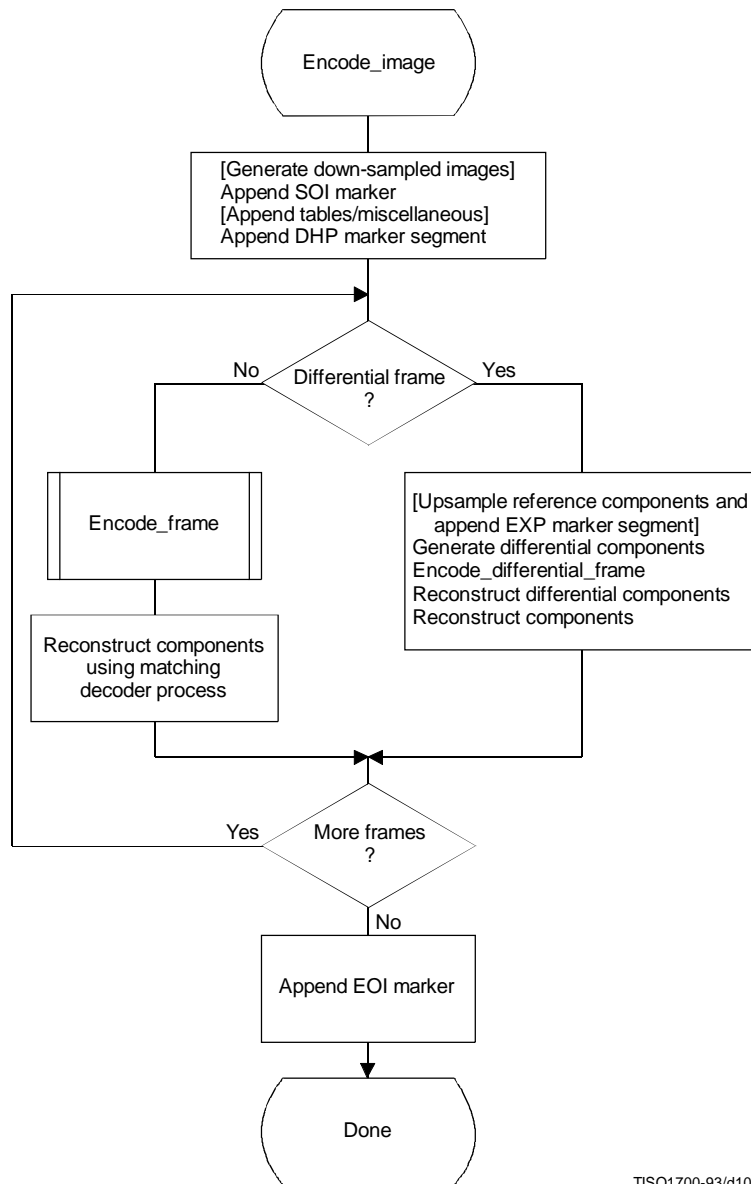


Figure J.1 – Hierarchical control procedure for encoding an image

In Figure J.1 procedures in brackets shall be performed whenever the particular hierarchical encoding sequence being followed requires them.

In the hierarchical mode the define-hierarchical-progression (DHP) marker segment shall be placed in the compressed image data before the first start-of-frame. The DHP segment is used to signal the size of the image components of the completed image. The syntax of the DHP segment is specified in Annex B.

The first frame for each component or group of components in a hierarchical process shall be encoded by a non-differential frame. Differential frames shall then be used to encode the two's complement differences between source input components (possibly downsampled) and the reference components (possibly upsampled). The reference components are reconstructed components created by previous frames in the hierarchical process. For either differential or non-differential frames, reconstructions of the components shall be generated if needed as reference components for a subsequent frame in the hierarchical process.

Resolution changes may occur between hierarchical frames in a hierarchical process. These changes occur if downsampling filters are used to reduce the spatial resolution of some or all of the components of the source image. When the resolution of a reference component does not match the resolution of the component input to a differential frame, an upsampling filter shall be used to increase the spatial resolution of the reference component. The EXP marker segment shall be added to the compressed image data before the start-of-frame whenever upsampling of a reference component is required. No more than one EXP marker segment shall precede a given frame.

Any of the marker segments allowed before a start-of-frame for the encoding process selected may be used before either non-differential or differential frames.

For 16-bit input precision (lossless encoder), the differential components which are input to a differential frame are calculated modulo 2^{16} . The reconstructed components calculated from the reconstructed differential components are also calculated modulo 2^{16} .

If a hierarchical encoding process uses a DCT encoding process for the first frame, all frames in the hierarchical process except for the final frame for each component shall use the DCT encoding processes defined in either Annex F or Annex G, or the modified DCT encoding processes defined in this annex. The final frame may use a modified lossless process defined in this annex.

If a hierarchical encoding process uses a lossless encoding process for the first frame, all frames in the hierarchical process shall use a lossless encoding process defined in Annex H, or a modified lossless process defined in this annex.

J.1.1.1 Downsampling filter

The downsampled components are generated using a downsampling filter that is not specified in this Specification. This filter should, however, be consistent with the upsampling filter. An example of a downsampling filter is provided in K.5.

J.1.1.2 Upsampling filter

The upsampling filter increases the spatial resolution by a factor of two horizontally, vertically, or both. Bi-linear interpolation is used for the upsampling filter, as illustrated in Figure J.2.

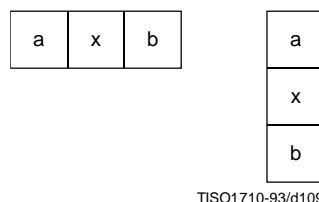


Figure J.2 – Diagram of sample positions for upsampling rules

The rule for calculating the interpolated value is:

$$P_x = (Ra + Rb) / 2$$

where Ra and Rb are sample values from adjacent positions a and b of the lower resolution image and Px is the interpolated value. The division indicates truncation, not rounding. The left-most column of the upsampled image matches the left-most column of the lower resolution image. The top line of the upsampled image matches the top line of the lower resolution image. The right column and the bottom line of the lower resolution image are replicated to provide the values required for the right column edge and bottom line interpolations. The upsampling process always doubles the line length or the number of lines.

If both horizontal and vertical expansions are signalled, they are done in sequence – first the horizontal expansion and then the vertical.

J.1.2 Control procedure for encoding a differential frame

The control procedures in Annex E for frames, scans, restart intervals, and MCU also apply to the encoding of differential frames, and the scans, restart intervals, and MCU from which the differential frame is constructed. The differential frames differ from the frames of Annexes F, G, and H only at the coding model level.

J.1.3 Encoder coding models for differential frames

The coding models defined in Annexes F, G, and H are modified to allow them to be used for coding of two’s complement differences.

J.1.3.1 Modifications to encoder DCT encoding models for differential frames

Two modifications are made to the DCT coding models to allow them to be used in differential frames. First, the FDCT of the differential input is calculated without the level shift. Second, the DC coefficient of the DCT is coded directly – without prediction.

J.1.3.2 Modifications to lossless encoding models for differential frames

One modification is made to the lossless coding models. The difference is coded directly – without prediction. The prediction selection parameter in the scan header shall be set to zero. The point transform which may be applied to the differential inputs is defined in Annex A.

J.1.4 Modifications to the entropy encoders for differential frames

The coding of two’s complement differences requires one extra bit of precision for the Huffman coding of AC coefficients. The extension to Tables F.1 and F.7 is given in Table J.2.

Table J.2 – Modifications to table of AC coefficient amplitude ranges

SSSS	AC coefficients
15	–32 767..–16 384, 16 384..32 767

The arithmetic coding models are already defined for the precision needed in differential frames.

J.2 Hierarchical decoding

J.2.1 Hierarchical control procedure for decoding an image

The control structure for decoding an image using the hierarchical mode is given in Figure J.3.

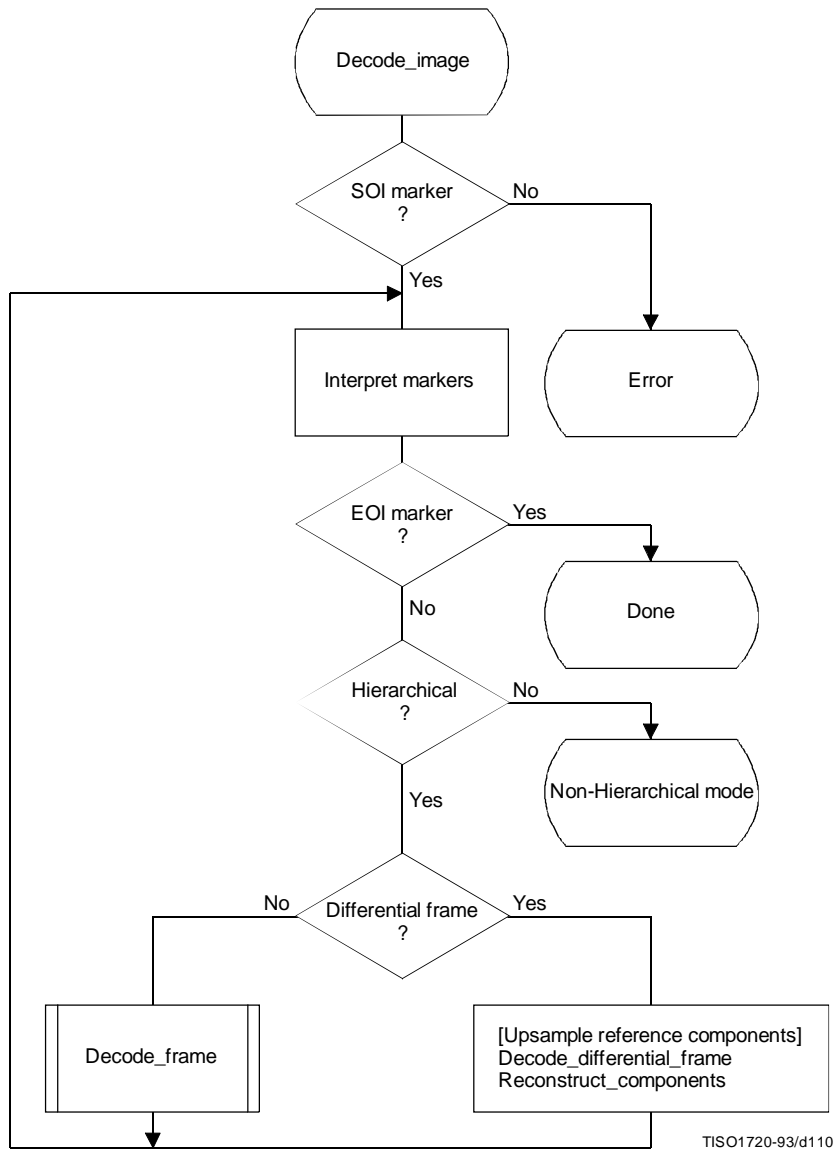


Figure J.3 – Hierarchical control procedure for decoding an image

The Interpret markers procedure shall decode the markers which may precede the SOF marker, continuing this decoding until either a SOF or EOI marker is found. If the DHP marker is encountered before the first frame, a flag is set which selects the hierarchical decoder at the “hierarchical?” decision point. In addition to the DHP marker (which shall precede any SOF) and the EXP marker (which shall precede any differential SOF requiring resolution changes in the reference components), any other markers which may precede a SOF shall be interpreted to the extent required for decoding of the compressed image data.

If a differential SOF marker is found, the differential frame path is followed. If the EXP was encountered in the Interpret markers procedure, the reference components for the frame shall be upsampled as required by the parameters in the EXP segment. The upsampling procedure described in J.1.1.2 shall be followed.

The Decode_differential_frame procedure generates a set of differential components. These differential components shall be added, modulo 2^{16} , to the upsampled reference components in the Reconstruct_components procedure. This creates a new set of reference components which shall be used when required in subsequent frames of the hierarchical process.

J.2.2 Control procedure for decoding a differential frame

The control procedures in Annex E for frames, scans, restart intervals, and MCU also apply to the decoding of differential frames and the scans, restart intervals, and MCU from which the differential frame is constructed. The differential frame differs from the frames of Annexes F, G, and H only at the decoder coding model level.

J.2.3 Decoder coding models for differential frames

The decoding models described in Annexes F, G, and H are modified to allow them to be used for decoding of two's complement differential components.

J.2.3.1 Modifications to the differential frame decoder DCT coding model

Two modifications are made to the decoder DCT coding models to allow them to code differential frames. First, the IDCT of the differential output is calculated without the level shift. Second, the DC coefficient of the DCT is decoded directly – without prediction.

J.2.3.2 Modifications to the differential frame decoder lossless coding model

One modification is made to the lossless decoder coding model. The difference is decoded directly – without prediction. If the point transformation parameter in the scan header is not zero, the point transform, defined in Annex A, shall be applied to the differential output.

J.2.4 Modifications to the entropy decoders for differential frames

The decoding of two's complement differences requires one extra bit of precision in the Huffman code table. This is described in J.1.4. The arithmetic coding models are already defined for the precision needed in differential frames.

Annex K

Examples and guidelines

(This annex does not form an integral part of this Recommendation | International Standard)

This annex provides examples of various tables, procedures, and other guidelines.

K.1 Quantization tables for luminance and chrominance components

Two examples of quantization tables are given in Tables K.1 and K.2. These are based on psychovisual thresholding and are derived empirically using luminance and chrominance and 2:1 horizontal subsampling. These tables are provided as examples only and are not necessarily suitable for any particular application. These quantization values have been used with good results on 8-bit per sample luminance and chrominance images of the format illustrated in Figure 13. Note that these quantization values are appropriate for the DCT normalization defined in A.3.3.

If these quantization values are divided by 2, the resulting reconstructed image is usually nearly indistinguishable from the source image.

Table K.1 – Luminance quantization table

16	11	10	16	24	40	51	61
12	12	14	19	26	58	60	55
14	13	16	24	40	57	69	56
14	17	22	29	51	87	80	62
18	22	37	56	68	109	103	77
24	35	55	64	81	104	113	92
49	64	78	87	103	121	120	101
72	92	95	98	112	100	103	99

Table K.2 – Chrominance quantization table

17	18	24	47	99	99	99	99
18	21	26	66	99	99	99	99
24	26	56	99	99	99	99	99
47	66	99	99	99	99	99	99
99	99	99	99	99	99	99	99
99	99	99	99	99	99	99	99
99	99	99	99	99	99	99	99
99	99	99	99	99	99	99	99

K.2 A procedure for generating the lists which specify a Huffman code table

A Huffman table is generated from a collection of statistics in two steps. The first step is the generation of the list of lengths and values which are in accord with the rules for generating the Huffman code tables. The second step is the generation of the Huffman code table from the list of lengths and values.

The first step, the topic of this section, is needed only for custom Huffman table generation and is done only in the encoder. In this step the statistics are used to create a table associating each value to be coded with the size (in bits) of the corresponding Huffman code. This table is sorted by code size.

A procedure for creating a Huffman table for a set of up to 256 symbols is shown in Figure K.1. Three vectors are defined for this procedure:

FREQ(V)	Frequency of occurrence of symbol V
CODESIZE(V)	Code size of symbol V
OTHERS(V)	Index to next symbol in chain of all symbols in current branch of code tree

where V goes from 0 to 256.

Before starting the procedure, the values of FREQ are collected for $V = 0$ to 255 and the FREQ value for $V = 256$ is set to 1 to reserve one code point. FREQ values for unused symbols are defined to be zero. In addition, the entries in CODESIZE are all set to 0, and the indices in OTHERS are set to -1, the value which terminates a chain of indices. Reserving one code point guarantees that no code word can ever be all "1" bits.

The search for the entry with the least value of FREQ(V) selects the largest value of V with the least value of FREQ(V) greater than zero.

The procedure "Find V1 for least value of FREQ(V1) > 0" always selects the value with the largest value of V1 when more than one V1 with the same frequency occurs. The reserved code point is then guaranteed to be in the longest code word category.

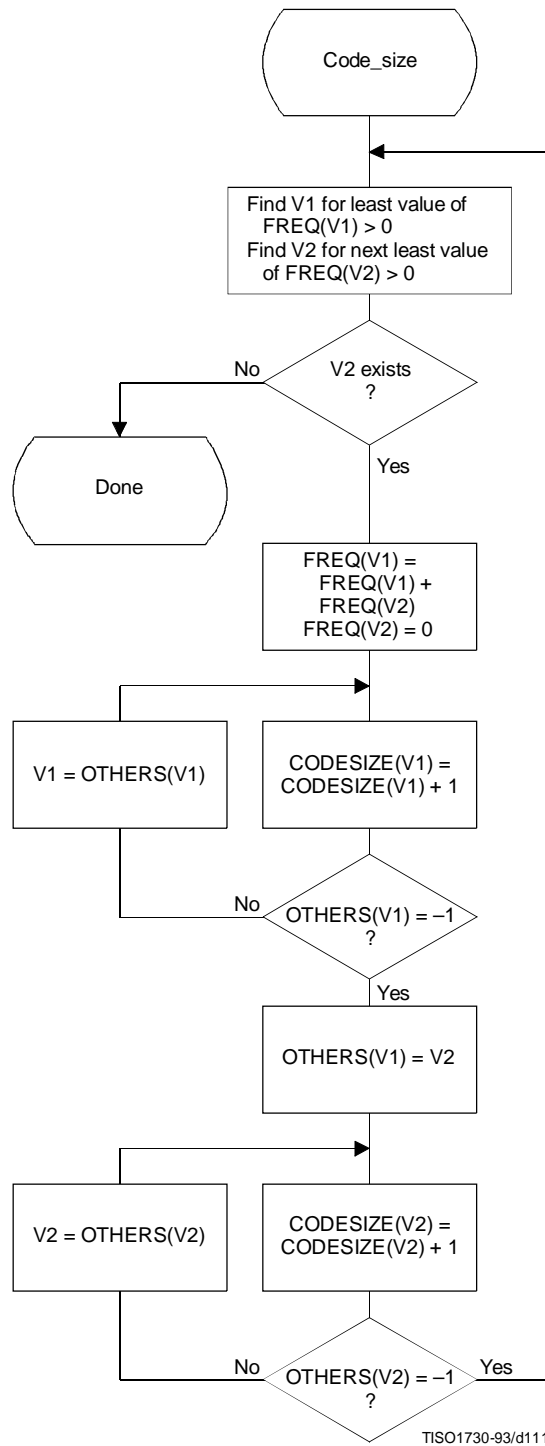
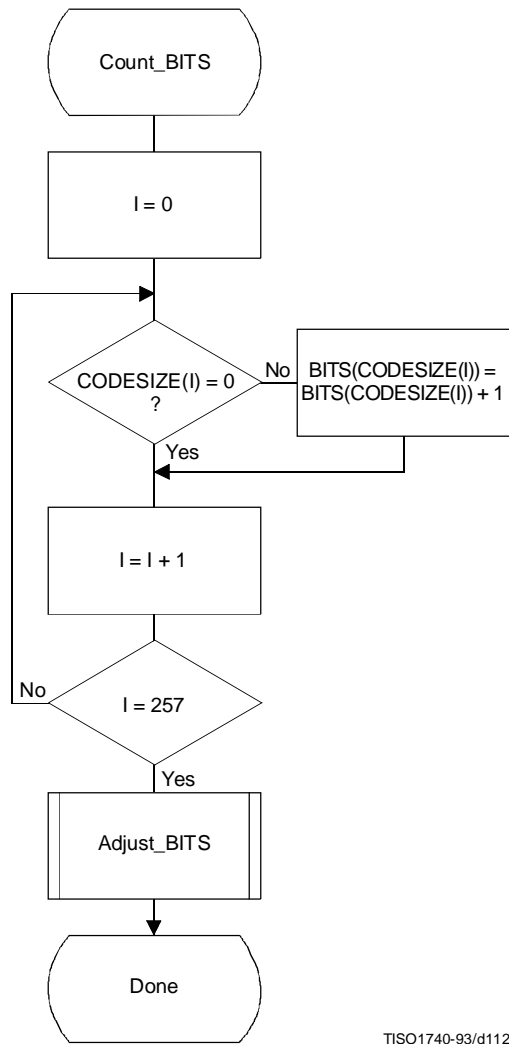


Figure K.1 – Procedure to find Huffman code sizes

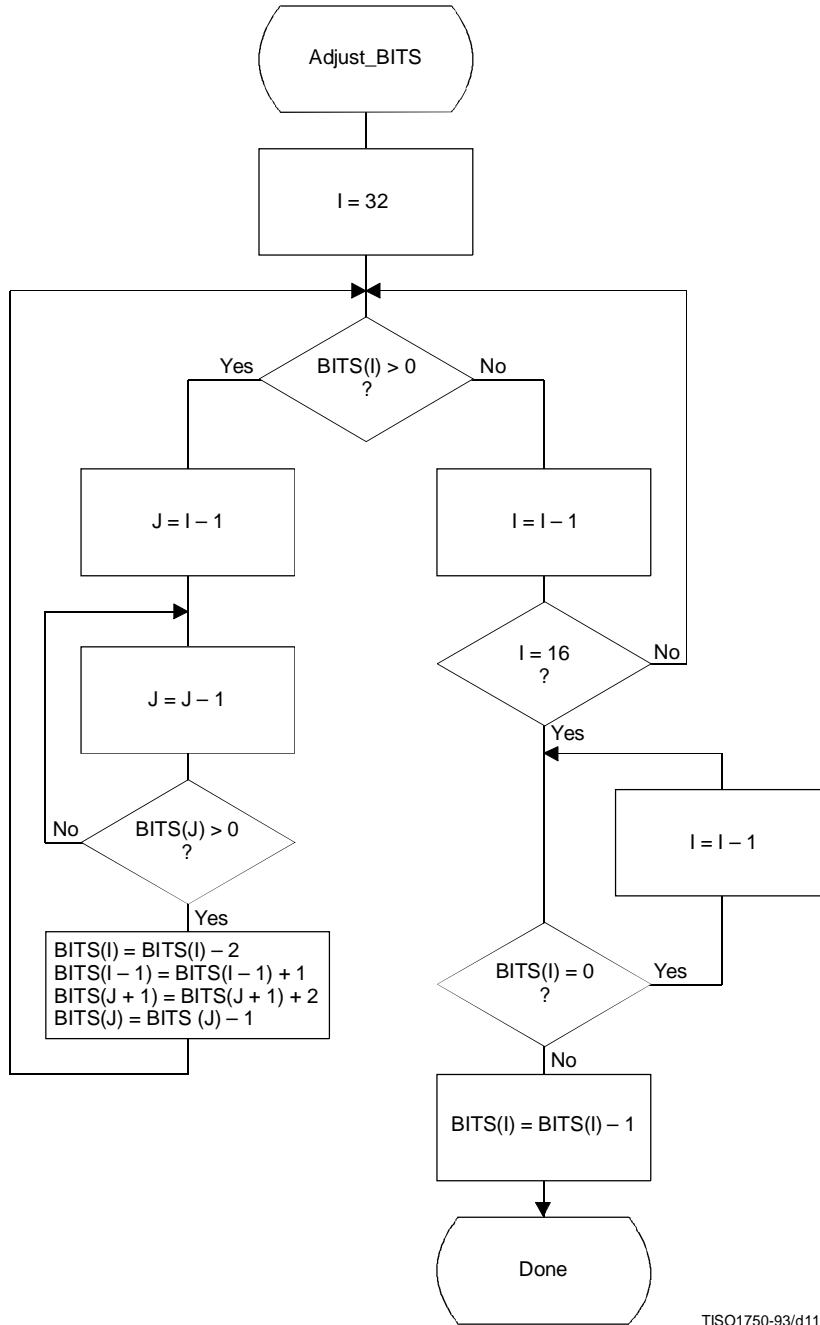
Once the code lengths for each symbol have been obtained, the number of codes of each length is obtained using the procedure in Figure K.2. The count for each size is contained in the list, BITS. The counts in BITS are zero at the start of the procedure. The procedure assumes that the probabilities are large enough that code lengths greater than 32 bits never occur. Note that until the final Adjust_BITS procedure is complete, BITS may have more than the 16 entries required in the table specification (see Annex C).



TISO1740-93/d112

Figure K.2 – Procedure to find the number of codes of each size

Figure K.3 gives the procedure for adjusting the BITS list so that no code is longer than 16 bits. Since symbols are paired for the longest Huffman code, the symbols are removed from this length category two at a time. The prefix for the pair (which is one bit shorter) is allocated to one of the pair; then (skipping the BITS entry for that prefix length) a code word from the next shortest non-zero BITS entry is converted into a prefix for two code words one bit longer. After the BITS list is reduced to a maximum code length of 16 bits, the last step removes the reserved code point from the code length count.

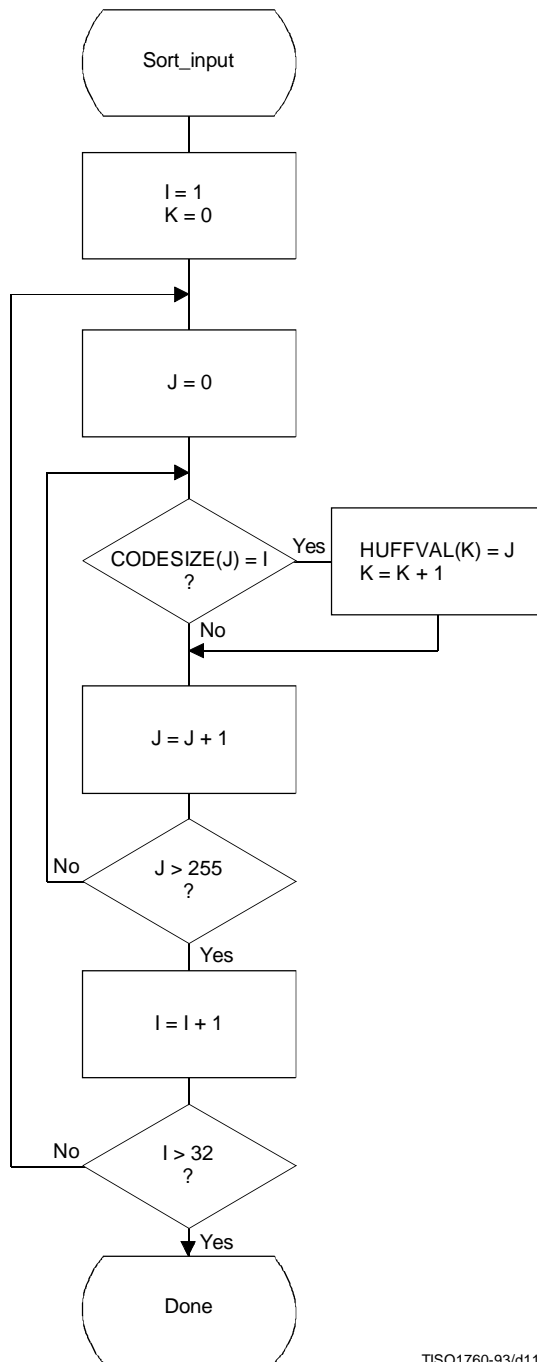


TISO1750-93/d113

Figure K.3 – Procedure for limiting code lengths to 16 bits

The input values are sorted according to code size as shown in Figure K.4. HUFFVAL is the list containing the input values associated with each code word, in order of increasing code length.

At this point, the list of code lengths (BITS) and the list of values (HUFFVAL) can be used to generate the code tables. These procedures are described in Annex C.



TISO1760-93/d114

Figure K.4 – Sorting of input values according to code size

K.3 Typical Huffman tables for 8-bit precision luminance and chrominance

Huffman table-specification syntax is specified in B.2.4.2.

K.3.1 Typical Huffman tables for the DC coefficient differences

Tables K.3 and K.4 give Huffman tables for the DC coefficient differences which have been developed from the average statistics of a large set of video images with 8-bit precision. Table K.3 is appropriate for luminance components and Table K.4 is appropriate for chrominance components. Although there are no default tables, these tables may prove to be useful for many applications.

Table K.3 – Table for luminance DC coefficient differences

Category	Code length	Code word
0	2	00
1	3	010
2	3	011
3	3	100
4	3	101
5	3	110
6	4	1110
7	5	11110
8	6	111110
9	7	1111110
10	8	11111110
11	9	111111110

Table K.4 – Table for chrominance DC coefficient differences

Category	Code length	Code word
0	2	00
1	2	01
2	2	10
3	3	110
4	4	1110
5	5	11110
6	6	111110
7	7	1111110
8	8	11111110
9	9	111111110
10	10	1111111110
11	11	11111111110

K.3.2 Typical Huffman tables for the AC coefficients

Tables K.5 and K.6 give Huffman tables for the AC coefficients which have been developed from the average statistics of a large set of images with 8-bit precision. Table K.5 is appropriate for luminance components and Table K.6 is appropriate for chrominance components. Although there are no default tables, these tables may prove to be useful for many applications.

Table K.5 – Table for luminance AC coefficients (sheet 1 of 4)

Run/Size	Code length	Code word
0/0 (EOB)	4	1010
0/1	2	00
0/2	2	01
0/3	3	100
0/4	4	1011
0/5	5	11010
0/6	7	1111000
0/7	8	11111000
0/8	10	1111110110
0/9	16	1111111110000010
0/A	16	1111111110000011
1/1	4	1100
1/2	5	11011
1/3	7	1111001
1/4	9	111110110
1/5	11	11111110110
1/6	16	1111111110000100
1/7	16	1111111110000101
1/8	16	1111111110000110
1/9	16	1111111110000111
1/A	16	1111111110001000
2/1	5	11100
2/2	8	11111001
2/3	10	1111110111
2/4	12	111111110100
2/5	16	1111111110001001
2/6	16	1111111110001010
2/7	16	1111111110001011
2/8	16	1111111110001100
2/9	16	1111111110001101
2/A	16	1111111110001110
3/1	6	111010
3/2	9	111110111
3/3	12	111111110101
3/4	16	1111111110001111
3/5	16	1111111110010000
3/6	16	1111111110010001
3/7	16	1111111110010010
3/8	16	1111111110010011
3/9	16	1111111110010100
3/A	16	1111111110010101

Table K.5 (sheet 2 of 4)

Run/Size	Code length	Code word
4/1	6	111011
4/2	10	1111111000
4/3	16	111111110010110
4/4	16	111111110010111
4/5	16	111111110011000
4/6	16	111111110011001
4/7	16	111111110011010
4/8	16	111111110011011
4/9	16	111111110011100
4/A	16	111111110011101
5/1	7	1111010
5/2	11	11111110111
5/3	16	111111110011110
5/4	16	111111110011111
5/5	16	111111110100000
5/6	16	111111110100001
5/7	16	111111110100010
5/8	16	111111110100011
5/9	16	111111110100100
5/A	16	111111110100101
6/1	7	1111011
6/2	12	111111110110
6/3	16	111111110100110
6/4	16	111111110100111
6/5	16	111111110101000
6/6	16	111111110101001
6/7	16	111111110101010
6/8	16	111111110101011
6/9	16	111111110101100
6/A	16	111111110101101
7/1	8	11111010
7/2	12	111111110111
7/3	16	111111110101110
7/4	16	111111110101111
7/5	16	111111110110000
7/6	16	111111110110001
7/7	16	111111110110010
7/8	16	111111110110011
7/9	16	111111110110100
7/A	16	111111110110101
8/1	9	111111000
8/2	15	11111111000000

Table K.5 (sheet 3 of 4)

Run/Size	Code length	Code word
8/3	16	111111110110110
8/4	16	111111110110111
8/5	16	111111110111000
8/6	16	111111110111001
8/7	16	111111110111010
8/8	16	111111110111011
8/9	16	111111110111100
8/A	16	111111110111101
9/1	9	111111001
9/2	16	111111110111110
9/3	16	111111110111111
9/4	16	1111111111000000
9/5	16	1111111111000001
9/6	16	1111111111000010
9/7	16	1111111111000011
9/8	16	1111111111000100
9/9	16	1111111111000101
9/A	16	1111111111000110
A/1	9	111111010
A/2	16	1111111111000111
A/3	16	1111111111001000
A/4	16	1111111111001001
A/5	16	1111111111001010
A/6	16	1111111111001011
A/7	16	1111111111001100
A/8	16	1111111111001101
A/9	16	1111111111001110
A/A	16	1111111111001111
B/1	10	1111111001
B/2	16	111111111010000
B/3	16	111111111010001
B/4	16	111111111010010
B/5	16	111111111010011
B/6	16	111111111010100
B/7	16	111111111010101
B/8	16	111111111010110
B/9	16	111111111010111
B/A	16	111111111011000
C/1	10	1111111010
C/2	16	111111111011001
C/3	16	111111111011010
C/4	16	111111111011011

Table K.5 (sheet 4 of 4)

Run/Size	Code length	Code word
C/5	16	111111111011100
C/6	16	111111111011101
C/7	16	111111111011110
C/8	16	111111111011111
C/9	16	111111111100000
C/A	16	111111111100001
D/1	11	1111111000
D/2	16	111111111100010
D/3	16	111111111100011
D/4	16	111111111100100
D/5	16	111111111100101
D/6	16	111111111100110
D/7	16	111111111100111
D/8	16	111111111101000
D/9	16	111111111101001
D/A	16	111111111101010
E/1	16	111111111101011
E/2	16	111111111101100
E/3	16	111111111101101
E/4	16	111111111101110
E/5	16	111111111101111
E/6	16	111111111110000
E/7	16	111111111110001
E/8	16	111111111110010
E/9	16	111111111110011
E/A	16	111111111110100
F/0 (ZRL)	11	1111111001
F/1	16	111111111110101
F/2	16	111111111110110
F/3	16	111111111110111
F/4	16	111111111111000
F/5	16	111111111111001
F/6	16	111111111111010
F/7	16	111111111111011
F/8	16	111111111111100
F/9	16	111111111111101
F/A	16	111111111111110

Table K.6 – Table for chrominance AC coefficients (sheet 1 of 4)

Run/Size	Code length	Code word
0/0 (EOB)	2	00
0/1	2	01
0/2	3	100
0/3	4	1010
0/4	5	11000
0/5	5	11001
0/6	6	111000
0/7	7	1111000
0/8	9	111110100
0/9	10	1111110110
0/A	12	111111110100
1/1	4	1011
1/2	6	111001
1/3	8	11110110
1/4	9	111110101
1/5	11	11111110110
1/6	12	111111110101
1/7	16	1111111110001000
1/8	16	1111111110001001
1/9	16	1111111110001010
1/A	16	1111111110001011
2/1	5	11010
2/2	8	11110111
2/3	10	1111110111
2/4	12	111111110110
2/5	15	111111111000010
2/6	16	1111111110001100
2/7	16	1111111110001101
2/8	16	1111111110001110
2/9	16	1111111110001111
2/A	16	1111111110010000
3/1	5	11011
3/2	8	11111000
3/3	10	1111111000
3/4	12	111111110111
3/5	16	1111111110010001
3/6	16	1111111110010010
3/7	16	1111111110010011
3/8	16	1111111110010100
3/9	16	1111111110010101
3/A	16	1111111110010110
4/1	6	111010

Table K.6 (sheet 2 of 4)

Run/Size	Code length	Code word
4/2	9	111110110
4/3	16	111111110010111
4/4	16	111111110011000
4/5	16	111111110011001
4/6	16	111111110011010
4/7	16	111111110011011
4/8	16	111111110011100
4/9	16	111111110011101
4/A	16	111111110011110
5/1	6	111011
5/2	10	1111111001
5/3	16	111111110011111
5/4	16	111111110100000
5/5	16	111111110100001
5/6	16	111111110100010
5/7	16	111111110100011
5/8	16	111111110100100
5/9	16	111111110100101
5/A	16	111111110100110
6/1	7	1111001
6/2	11	11111110111
6/3	16	111111110100111
6/4	16	111111110101000
6/5	16	111111110101001
6/6	16	111111110101010
6/7	16	111111110101011
6/8	16	111111110101100
6/9	16	111111110101101
6/A	16	111111110101110
7/1	7	1111010
7/2	11	1111111000
7/3	16	111111110101111
7/4	16	111111110110000
7/5	16	111111110110001
7/6	16	111111110110010
7/7	16	111111110110011
7/8	16	111111110110100
7/9	16	111111110110101
7/A	16	111111110110110
8/1	8	11111001
8/2	16	111111110110111
8/3	16	111111110111000

Table K.6 (sheet 3 of 4)

Run/Size	Code length	Code word
8/4	16	111111110111001
8/5	16	111111110111010
8/6	16	111111110111011
8/7	16	111111110111100
8/8	16	111111110111101
8/9	16	111111110111110
8/A	16	111111110111111
9/1	9	111110111
9/2	16	1111111111000000
9/3	16	1111111111000001
9/4	16	1111111111000010
9/5	16	1111111111000011
9/6	16	1111111111000100
9/7	16	1111111111000101
9/8	16	1111111111000110
9/9	16	1111111111000111
9/A	16	1111111111001000
A/1	9	111111000
A/2	16	1111111111001001
A/3	16	1111111111001010
A/4	16	1111111111001011
A/5	16	1111111111001100
A/6	16	1111111111001101
A/7	16	1111111111001110
A/8	16	1111111111001111
A/9	16	1111111111010000
A/A	16	1111111111010001
B/1	9	111111001
B/2	16	1111111111010010
B/3	16	1111111111010011
B/4	16	1111111111010100
B/5	16	1111111111010101
B/6	16	1111111111010110
B/7	16	1111111111010111
B/8	16	1111111111011000
B/9	16	1111111111011001
B/A	16	1111111111011010
C/1	9	111111010
C/2	16	1111111111011011
C/3	16	1111111111011100
C/4	16	1111111111011101
C/5	16	1111111111011110

Table K.6 (sheet 4 of 4)

Run/Size	Code length	Code word
C/6	16	1111111111011111
C/7	16	1111111111100000
C/8	16	1111111111100001
C/9	16	1111111111100010
C/A	16	1111111111100011
D/1	11	11111111001
D/2	16	1111111111100100
D/3	16	1111111111100101
D/4	16	1111111111100110
D/5	16	1111111111100111
D/6	16	1111111111101000
D/7	16	1111111111101001
D/8	16	1111111111101010
D/9	16	1111111111101011
D/A	16	1111111111101100
E/1	14	1111111100000
E/2	16	1111111111101101
E/3	16	1111111111101110
E/4	16	1111111111101111
E/5	16	1111111111110000
E/6	16	1111111111110001
E/7	16	1111111111110010
E/8	16	1111111111110011
E/9	16	1111111111110100
E/A	16	1111111111110101
F/0 (ZRL)	10	1111111010
F/1	15	11111111000011
F/2	16	1111111111110110
F/3	16	1111111111110111
F/4	16	1111111111111000
F/5	16	1111111111111001
F/6	16	1111111111111010
F/7	16	1111111111111011
F/8	16	1111111111111100
F/9	16	1111111111111101
F/A	16	1111111111111110

K.3.3 Huffman table-specification examples

K.3.3.1 Specification of typical tables for DC difference coding

A set of typical tables for DC component coding is given in K.3.1. The specification of these tables is as follows:

For Table K.3 (for luminance DC coefficients), the 16 bytes which specify the list of code lengths for the table are

X'00 01 05 01 01 01 01 01 01 00 00 00 00 00 00'

The set of values following this list is

X'00 01 02 03 04 05 06 07 08 09 0A 0B'

For Table K.4 (for chrominance DC coefficients), the 16 bytes which specify the list of code lengths for the table are

X'00 03 01 01 01 01 01 01 01 01 01 00 00 00 00'

The set of values following this list is

X'00 01 02 03 04 05 06 07 08 09 0A 0B'

K.3.3.2 Specification of typical tables for AC coefficient coding

A set of typical tables for AC component coding is given in K.3.2. The specification of these tables is as follows:

For Table K.5 (for luminance AC coefficients), the 16 bytes which specify the list of code lengths for the table are

X'00 02 01 03 03 02 04 03 05 05 04 04 00 00 01 7D'

The set of values which follows this list is

X'01	02	03	00	04	11	05	12	21	31	41	06	13	51	61	07
22	71	14	32	81	91	A1	08	23	42	B1	C1	15	52	D1	F0
24	33	62	72	82	09	0A	16	17	18	19	1A	25	26	27	28
29	2A	34	35	36	37	38	39	3A	43	44	45	46	47	48	49
4A	53	54	55	56	57	58	59	5A	63	64	65	66	67	68	69
6A	73	74	75	76	77	78	79	7A	83	84	85	86	87	88	89
8A	92	93	94	95	96	97	98	99	9A	A2	A3	A4	A5	A6	A7
A8	A9	AA	B2	B3	B4	B5	B6	B7	B8	B9	BA	C2	C3	C4	C5
C6	C7	C8	C9	CA	D2	D3	D4	D5	D6	D7	D8	D9	DA	E1	E2
E3	E4	E5	E6	E7	E8	E9	EA	F1	F2	F3	F4	F5	F6	F7	F8
F9	FA'														

For Table K.6 (for chrominance AC coefficients), the 16 bytes which specify the list of code lengths for the table are

X'00 02 01 02 04 04 03 04 07 05 04 04 00 01 02 77'

The set of values which follows this list is:

X'00	01	02	03	11	04	05	21	31	06	12	41	51	07	61	71
13	22	32	81	08	14	42	91	A1	B1	C1	09	23	33	52	F0
15	62	72	D1	0A	16	24	34	E1	25	F1	17	18	19	1A	26
27	28	29	2A	35	36	37	38	39	3A	43	44	45	46	47	48
49	4A	53	54	55	56	57	58	59	5A	63	64	65	66	67	68
69	6A	73	74	75	76	77	78	79	7A	82	83	84	85	86	87
88	89	8A	92	93	94	95	96	97	98	99	9A	A2	A3	A4	A5
A6	A7	A8	A9	AA	B2	B3	B4	B5	B6	B7	B8	B9	BA	C2	C3
C4	C5	C6	C7	C8	C9	CA	D2	D3	D4	D5	D6	D7	D8	D9	DA
E2	E3	E4	E5	E6	E7	E8	E9	EA	F2	F3	F4	F5	F6	F7	F8
F9	FA														

K.4 Additional information on arithmetic coding

K.4.1 Test sequence for a small data set for the arithmetic coder

The following 256-bit test sequence (in hexadecimal form) is structured to test many of the encoder and decoder paths:

X'00020051 000000C0 0352872A AAAAAAAAA 82C02000 FCD79EF6 74EAABF7 697EE74C'

Tables K.7 and K.8 provide a symbol-by-symbol list of the arithmetic encoder and decoder operation. In these tables the event count, EC, is listed first, followed by the value of Qe used in encoding and decoding that event. The decision D to be encoded (and decoded) is listed next. The column labeled MPS contains the sense of the MPS, and if it is followed by a CE (in the "CX" column), the conditional MPS/LPS exchange occurs when encoding and decoding the decision (see Figures D.3, D.4 and D.17). The contents of the A and C registers are the values before the event is encoded and decoded. ST is the number of X'FF' bytes stacked in the encoder waiting for a resolution of the carry-over. Note that the A register is always greater than X'7FFF'. (The starting value has an implied value of X'10000'.)

In the encoder test, the code bytes (B) are listed if they were completed during the coding of the preceding event. If additional bytes follow, they were also completed during the coding of the preceding event. If a byte is listed in the Bx column, the preceding byte in column B was modified by a carry-over.

In the decoder the code bytes are listed if they were placed in the code register just prior to the event EC.

For this file the coded bit count is 240, including the overhead to flush the final data from the C register. When the marker X'FFD9' is appended, a total of 256 bits are output. The actual compressed data sequence for the encoder is (in hexadecimal form)

X'655B5144 F7969D51 7855BFFF 00FC5184 C7CEF939 00287D46 708ECBC0 F6FFD900'

Table K.7 – Encoder test sequence (sheet 1 of 7)

EC	D	MPS	CX	Qe (hexadecimal)	A (hexadecimal)	C (hexadecimal)	CT	ST	Bx	B
1	0	0	CE	5A1D	0000	00000000	11	0		
2	0	0		5A1D	A5E3	00000000	11	0		
3	0	0		2586	B43A	0000978C	10	0		
4	0	0		2586	8EB4	0000978C	10	0		
5	0	0		1114	D25C	00012F18	9	0		
6	0	0		1114	C148	00012F18	9	0		
7	0	0		1114	B034	00012F18	9	0		
8	0	0		1114	9F20	00012F18	9	0		
9	0	0		1114	8E0C	00012F18	9	0		
10	0	0		080B	F9F0	00025E30	8	0		
11	0	0		080B	F1E5	00025E30	8	0		
12	0	0		080B	E9DA	00025E30	8	0		
13	0	0		080B	E1CF	00025E30	8	0		
14	0	0		080B	D9C4	00025E30	8	0		
15	1	0		080B	D1B9	00025E30	8	0		
16	0	0		17B9	80B0	00327DE0	4	0		
17	0	0		1182	D1EE	0064FBC0	3	0		
18	0	0		1182	C06C	0064FBC0	3	0		
19	0	0		1182	AEEA	0064FBC0	3	0		
20	0	0		1182	9D68	0064FBC0	3	0		
21	0	0		1182	8BE6	0064FBC0	3	0		
22	0	0		0CEF	F4C8	00C9F780	2	0		
23	0	0		0CEF	E7D9	00C9F780	2	0		
24	0	0		0CEF	DAEA	00C9F780	2	0		
25	0	0		0CEF	CDFB	00C9F780	2	0		
26	1	0		0CEF	C10C	00C9F780	2	0		
27	0	0		1518	CEF0	000AB9D0	6	0		65
28	1	0		1518	B9D8	000AB9D0	6	0		
29	0	0		1AA9	A8C0	005AF480	3	0		
30	0	0		1AA9	8E17	005AF480	3	0		
31	0	0		174E	E6DC	00B5E900	2	0		
32	1	0		174E	CF8E	00B5E900	2	0		
33	0	0		1AA9	BA70	00050A00	7	0		5B
34	0	0		1AA9	9FC7	00050A00	7	0		
35	0	0		1AA9	851E	00050A00	7	0		
36	0	0		174E	D4EA	000A1400	6	0		

Table K.7 – Encoder test sequence (sheet 2 of 7)

EC	D	MPS	CX	Qe (hexadecimal)	A (hexadecimal)	C (hexadecimal)	CT	ST	Bx	B
37	0	0		174E	BD9C	000A1400	6	0		
38	0	0		174E	A64E	000A1400	6	0		
39	0	0		174E	8F00	000A1400	6	0		
40	0	0		1424	EF64	00142800	5	0		
41	0	0		1424	DB40	00142800	5	0		
42	0	0		1424	C71C	00142800	5	0		
43	0	0		1424	B2F8	00142800	5	0		
44	0	0		1424	9ED4	00142800	5	0		
45	0	0		1424	8AB0	00142800	5	0		
46	0	0		119C	ED18	00285000	4	0		
47	0	0		119C	DB7C	00285000	4	0		
48	0	0		119C	C9E0	00285000	4	0		
49	0	0		119C	B844	00285000	4	0		
50	0	0		119C	A6A8	00285000	4	0		
51	0	0		119C	950C	00285000	4	0		
52	0	0		119C	8370	00285000	4	0		
53	0	0		0F6B	E3A8	0050A000	3	0		
54	0	0		0F6B	D43D	0050A000	3	0		
55	0	0		0F6B	C4D2	0050A000	3	0		
56	0	0		0F6B	B567	0050A000	3	0		
57	1	0		0F6B	A5FC	0050A000	3	0		
58	1	0		1424	F6B0	00036910	7	0		51
59	0	0		1AA9	A120	00225CE0	4	0		
60	0	0		1AA9	8677	00225CE0	4	0		
61	0	0		174E	D79C	0044B9C0	3	0		
62	0	0		174E	C04E	0044B9C0	3	0		
63	0	0		174E	A900	0044B9C0	3	0		
64	0	0		174E	91B2	0044B9C0	3	0		
65	0	0		1424	F4C8	00897380	2	0		
66	0	0		1424	E0A4	00897380	2	0		
67	0	0		1424	CC80	00897380	2	0		
68	0	0		1424	B85C	00897380	2	0		
69	0	0		1424	A438	00897380	2	0		
70	0	0		1424	9014	00897380	2	0		
71	1	0		119C	F7E0	0112E700	1	0		
72	1	0		1424	8CE0	001E6A20	6	0		44
73	0	0		1AA9	A120	00F716E0	3	0		

Table K.7 – Encoder test sequence (sheet 3 of 7)

EC	D	MPS	CX	Qe (hexadecimal)	A (hexadecimal)	C (hexadecimal)	CT	ST	Bx	B
74	1	0		1AA9	8677	00F716E0	3	0		
75	0	0		2516	D548	00041570	8	0		F7
76	1	0		2516	B032	00041570	8	0		
77	0	0		299A	9458	00128230	6	0		
78	0	0		2516	D57C	00250460	5	0		
79	1	0		2516	B066	00250460	5	0		
80	0	0		299A	9458	00963EC0	3	0		
81	1	0		2516	D57C	012C7D80	2	0		
82	0	0		299A	9458	0004B798	8	0		96
83	0	0		2516	D57C	00096F30	7	0		
84	0	0		2516	B066	00096F30	7	0		
85	0	0		2516	8B50	00096F30	7	0		
86	1	0		1EDF	CC74	0012DE60	6	0		
87	1	0		2516	F6F8	009C5FA8	3	0		
88	1	0		299A	9458	0274C628	1	0		
89	0	0		32B4	A668	0004C398	7	0		9D
90	0	0		2E17	E768	00098730	6	0		
91	1	0		2E17	B951	00098730	6	0		
92	0	0		32B4	B85C	002849A8	4	0		
93	1	0		32B4	85A8	002849A8	4	0		
94	0	0		3C3D	CAD0	00A27270	2	0		
95	1	0		3C3D	8E93	00A27270	2	0		
96	0	0		415E	F0F4	00031318	8	0		51
97	1	0		415E	AF96	00031318	8	0		
98	0	0	CE	4639	82BC	000702A0	7	0		
99	1	0		415E	8C72	000E7E46	6	0		
100	0	0	CE	4639	82BC	001D92B4	5	0		
101	1	0		415E	8C72	003B9E6E	4	0		
102	0	0	CE	4639	82BC	0077D304	3	0		
103	1	0		415E	8C72	00F01F0E	2	0		
104	0	0	CE	4639	82BC	01E0D444	1	0		
105	1	0		415E	8C72	0002218E	8	0		78
106	0	0	CE	4639	82BC	0004D944	7	0		
107	1	0		415E	8C72	000A2B8E	6	0		
108	0	0	CE	4639	82BC	0014ED44	5	0		
109	1	0		415E	8C72	002A538E	4	0		
110	0	0	CE	4639	82BC	00553D44	3	0		

Table K.7 – Encoder test sequence (sheet 4 of 7)

EC	D	MPS	CX	Qe (hexadecimal)	A (hexadecimal)	C (hexadecimal)	CT	ST	Bx	B
111	1	0		415E	8C72	00AAF38E	2	0		
112	0	0	CE	4639	82BC	01567D44	1	0		
113	1	0		415E	8C72	0005738E	8	0		55
114	0	0	CE	4639	82BC	000B7D44	7	0		
115	1	0		415E	8C72	0017738E	6	0		
116	0	0	CE	4639	82BC	002F7D44	5	0		
117	1	0		415E	8C72	005F738E	4	0		
118	0	0	CE	4639	82BC	00BF7D44	3	0		
119	1	0		415E	8C72	017F738E	2	0		
120	0	0	CE	4639	82BC	02FF7D44	1	0		
121	1	0		415E	8C72	0007738E	8	0		BF
122	0	0	CE	4639	82BC	000F7D44	7	0		
123	1	0		415E	8C72	001F738E	6	0		
124	0	0	CE	4639	82BC	003F7D44	5	0		
125	1	0		415E	8C72	007F738E	4	0		
126	0	0	CE	4639	82BC	00FF7D44	3	0		
127	1	0		415E	8C72	01FF738E	2	0		
128	0	0	CE	4639	82BC	03FF7D44	1	0		
129	1	0		415E	8C72	0007738E	8	1		
130	0	0	CE	4639	82BC	000F7D44	7	1		
131	0	0		415E	8C72	001F738E	6	1		
132	0	0		3C3D	9628	003EE71C	5	1		
133	0	0		375E	B3D6	007DCE38	4	1		
134	0	0		32B4	F8F0	00FB9C70	3	1		
135	1	0		32B4	C63C	00FB9C70	3	1		
136	0	0		3C3D	CAD0	03F0BFE0	1	1		
137	1	0		3C3D	8E93	03F0BFE0	1	1		
138	1	0		415E	F0F4	000448D8	7	0		FF00FC
139	0	0	CE	4639	82BC	0009F0DC	6	0		
140	0	0		415E	8C72	00145ABE	5	0		
141	0	0		3C3D	9628	0028B57C	4	0		
142	0	0		375E	B3D6	00516AF8	3	0		
143	0	0		32B4	F8F0	00A2D5F0	2	0		
144	0	0		32B4	C63C	00A2D5F0	2	0		
145	0	0		32B4	9388	00A2D5F0	2	0		
146	0	0		2E17	C1A8	0145ABE0	1	0		

Table K.7 – Encoder test sequence (sheet 5 of 7)

EC	D	MPS	CX	Qe (hexadecimal)	A (hexadecimal)	C (hexadecimal)	CT	ST	Bx	B
147	1	0		2E17	9391	0145ABE0	1	0		
148	0	0		32B4	B85C	00084568	7	0		51
149	0	0		32B4	85A8	00084568	7	0		
150	0	0		2E17	A5E8	00108AD0	6	0		
151	0	0		299A	EFA2	002115A0	5	0		
152	0	0		299A	C608	002115A0	5	0		
153	0	0		299A	9C6E	002115A0	5	0		
154	0	0		2516	E5A8	00422B40	4	0		
155	0	0		2516	C092	00422B40	4	0		
156	0	0		2516	9B7C	00422B40	4	0		
157	0	0		1EDF	ECCC	00845680	3	0		
158	0	0		1EDF	CDDE	00845680	3	0		
159	0	0		1EDF	AF0E	00845680	3	0		
160	0	0		1EDF	902F	00845680	3	0		
161	1	0		1AA9	E2A0	0108AD00	2	0		
162	1	0		2516	D548	000BA7B8	7	0		84
163	1	0		299A	9458	00315FA8	5	0		
164	1	0		32B4	A668	00C72998	3	0		
165	1	0		3C3D	CAD0	031E7530	1	0		
166	1	0		415E	F0F4	000C0F0C	7	0		C7
167	0	0	CE	4639	82BC	00197D44	6	0		
168	0	0		415E	8C72	0033738E	5	0		
169	1	0		3C3D	9628	0066E71C	4	0		
170	1	0		415E	F0F4	019D041C	2	0		
171	0	0	CE	4639	82BC	033B6764	1	0		
172	1	0		415E	8C72	000747CE	8	0		CE
173	0	0	CE	4639	82BC	000F25C4	7	0		
174	1	0		415E	8C72	001EC48E	6	0		
175	1	0	CE	4639	82BC	003E1F44	5	0		
176	1	0		4B85	F20C	00F87D10	3	0		
177	1	0	CE	504F	970A	01F2472E	2	0		
178	0	0	CE	5522	8D76	03E48E5C	1	0		
179	0	0		504F	AA44	00018D60	8	0		F9
180	1	0		4B85	B3EA	00031AC0	7	0		
181	1	0	CE	504F	970A	0007064A	6	0		
182	1	0	CE	5522	8D76	000E0C94	5	0		
183	1	0		59EB	E150	00383250	3	0		

Table K.7 – Encoder test sequence (sheet 6 of 7)

EC	D	MPS	CX	Qe (hexadecimal)	A (hexadecimal)	C (hexadecimal)	CT	ST	Bx	B
184	0	1		59EB	B3D6	0071736A	2	0		
185	1	0		59EB	B3D6	00E39AAA	1	0		
186	1	1		59EB	B3D6	0007E92A	8	0		38
187	1	1		5522	B3D6	000FD254	7	0		
188	1	1		504F	BD68	001FA4A8	6	0		
189	0	1		4B85	DA32	003F4950	5	0		
190	1	1	CE	504F	970A	007FAFFA	4	0		
191	1	1		4B85	A09E	00FFED6A	3	0		
192	0	1		4639	AA32	01FFDAD4	2	0		
193	0	1	CE	4B85	8C72	04007D9A	1	0		
194	1	1	CE	504F	81DA	0000FB34	8	0	39	00
195	1	1		4B85	A09E	0002597E	7	0		
196	1	1		4639	AA32	0004B2FC	6	0		
197	0	1		415E	C7F2	000965F8	5	0		
198	1	1	CE	4639	82BC	0013D918	4	0		
199	0	1		415E	8C72	00282B36	3	0		
200	0	1	CE	4639	82BC	0050EC94	2	0		
201	1	1		4B85	F20C	0003B250	8	0		28
202	1	1		4B85	A687	0003B250	8	0		
203	1	1		4639	B604	000764A0	7	0		
204	0	1		415E	DF96	000EC940	6	0		
205	1	1	CE	4639	82BC	001ECEFO	5	0		
206	0	1		415E	8C72	003E16E6	4	0		
207	1	1	CE	4639	82BC	007CC3F4	3	0		
208	0	1		415E	8C72	00FA00EE	2	0		
209	1	1	CE	4639	82BC	01F49804	1	0		
210	0	1		415E	8C72	0001A90E	8	0		7D
211	1	1	CE	4639	82BC	0003E844	7	0		
212	0	1		415E	8C72	0008498E	6	0		
213	1	1	CE	4639	82BC	00112944	5	0		
214	0	1		415E	8C72	0022CB8E	4	0		
215	1	1	CE	4639	82BC	00462D44	3	0		
216	1	1		415E	8C72	008CD38E	2	0		
217	1	1		3C3D	9628	0119A71C	1	0		
218	1	1		375E	B3D6	00034E38	8	0		46
219	1	1		32B4	F8F0	00069C70	7	0		
220	1	1		32B4	C63C	00069C70	7	0		

Table K.7 – Encoder test sequence (sheet 7 of 7)

EC	D	MPS	CX	Qe (hexadecimal)	A (hexadecimal)	C (hexadecimal)	CT	ST	Bx	B
221	0	1		32B4	9388	00069C70	7	0		
222	1	1		3C3D	CAD0	001BF510	5	0		
223	1	1		3C3D	8E93	001BF510	5	0		
224	1	1		375E	A4AC	0037EA20	4	0		
225	0	1		32B4	DA9C	006FD440	3	0		
226	1	1		3C3D	CAD0	01C1F0A0	1	0		
227	1	1		3C3D	8E93	01C1F0A0	1	0		
228	0	1		375E	A4AC	0003E140	8	0		70
229	1	1		3C3D	DD78	00113A38	6	0		
230	0	1		3C3D	A13B	00113A38	6	0		
231	0	1		415E	F0F4	00467CD8	4	0		
232	1	1	CE	4639	82BC	008E58DC	3	0		
233	0	1		415E	8C72	011D2ABE	2	0		
234	1	1	CE	4639	82BC	023AEBA4	1	0		
235	1	1		415E	8C72	0006504E	8	0		8E
236	1	1		3C3D	9628	000CA09C	7	0		
237	1	1		375E	B3D6	00194138	6	0		
238	1	1		32B4	F8F0	00328270	5	0		
239	1	1		32B4	C63C	00328270	5	0		
240	0	1		32B4	9388	00328270	5	0		
241	1	1		3C3D	CAD0	00CB8D10	3	0		
242	1	1		3C3D	8E93	00CB8D10	3	0		
243	1	1		375E	A4AC	01971A20	2	0		
244	0	1		32B4	DA9C	032E3440	1	0		
245	0	1		3C3D	CAD0	000B70A0	7	0		CB
246	1	1		415E	F0F4	002FFCCC	5	0		
247	1	1		415E	AF96	002FFCCC	5	0		
248	1	1		3C3D	DC70	005FF998	4	0		
249	0	1		3C3D	A033	005FF998	4	0		
250	1	1		415E	F0F4	01817638	2	0		
251	0	1		415E	AF96	01817638	2	0		
252	0	1	CE	4639	82BC	0303C8E0	1	0		
253	1	1		4B85	F20C	000F2380	7	0		C0
254	1	1		4B85	A687	000F2380	7	0		
255	0	1		4639	B604	001E4700	6	0		
256	0	1	CE	4B85	8C72	003D6D96	5	0		
Flush:					81DA	007ADB2C	4	0		F6 FFD9

Table K.8 – Decoder test sequence (sheet 1 of 7)

EC	D	MPS	CX	Qe (hexadecimal)	A (hexadecimal)	C (hexadecimal)	CT	B
1	0	0	CE	5A1D	0000	655B0000	0	65 5B
2	0	0		5A1D	A5E3	655B0000	0	
3	0	0		2586	B43A	332AA200	7	51
4	0	0		2586	8EB4	332AA200	7	
5	0	0		1114	D25C	66554400	6	
6	0	0		1114	C148	66554400	6	
7	0	0		1114	B034	66554400	6	
8	0	0		1114	9F20	66554400	6	
9	0	0		1114	8E0C	66554400	6	
10	0	0		080B	F9F0	CCAA8800	5	
11	0	0		080B	F1E5	CCAA8800	5	
12	0	0		080B	E9DA	CCAA8800	5	
13	0	0		080B	E1CF	CCAA8800	5	
14	0	0		080B	D9C4	CCAA8800	5	
15	1	0		080B	D1B9	CCAA8800	5	
16	0	0		17B9	80B0	2FC88000	1	
17	0	0		1182	D1EE	5F910000	0	
18	0	0		1182	C06C	5F910000	0	
19	0	0		1182	AEEA	5F910000	0	
20	0	0		1182	9D68	5F910000	0	
21	0	0		1182	8BE6	5F910000	0	
22	0	0		0CEF	F4C8	BF228800	7	44
23	0	0		0CEF	E7D9	BF228800	7	
24	0	0		0CEF	DAEA	BF228800	7	
25	0	0		0CEF	CDFB	BF228800	7	
26	1	0		0CEF	C10C	BF228800	7	
27	0	0		1518	CEF0	B0588000	3	
28	1	0		1518	B9D8	B0588000	3	
29	0	0		1AA9	A8C0	5CC40000	0	
30	0	0		1AA9	8E17	5CC40000	0	
31	0	0		174E	E6DC	B989EE00	7	F7
32	1	0		174E	CF8E	B989EE00	7	
33	0	0		1AA9	BA70	0A4F7000	4	
34	0	0		1AA9	9FC7	0A4F7000	4	
35	0	0		1AA9	851E	0A4F7000	4	
36	0	0		174E	D4EA	149EE000	3	
37	0	0		174E	BD9C	149EE000	3	
38	0	0		174E	A64E	149EE000	3	
39	0	0		174E	8F00	149EE000	3	
40	0	0		1424	EF64	293DC000	2	

Table K.8 – Decoder test sequence (sheet 2 of 7)

EC	D	MPS	CX	Qe (hexadecimal)	A (hexadecimal)	C (hexadecimal)	CT	B
41	0	0		1424	DB40	293DC000	2	
42	0	0		1424	C71C	293DC000	2	
43	0	0		1424	B2F8	293DC000	2	
44	0	0		1424	9ED4	293DC000	2	
45	0	0		1424	8AB0	293DC000	2	
46	0	0		119C	ED18	527B8000	1	
47	0	0		119C	DB7C	527B8000	1	
48	0	0		119C	C9E0	527B8000	1	
49	0	0		119C	B844	527B8000	1	
50	0	0		119C	A6A8	527B8000	1	
51	0	0		119C	950C	527B8000	1	
52	0	0		119C	8370	527B8000	1	
53	0	0		0F6B	E3A8	A4F70000	0	
54	0	0		0F6B	D43D	A4F70000	0	
55	0	0		0F6B	C4D2	A4F70000	0	
56	0	0		0F6B	B567	A4F70000	0	
57	1	0		0F6B	A5FC	A4F70000	0	
58	1	0		1424	F6B0	E6696000	4	96
59	0	0		1AA9	A120	1EEB0000	1	
60	0	0		1AA9	8677	1EEB0000	1	
61	0	0		174E	D79C	3DD60000	0	
62	0	0		174E	C04E	3DD60000	0	
63	0	0		174E	A900	3DD60000	0	
64	0	0		174E	91B2	3DD60000	0	
65	0	0		1424	F4C8	7BAD3A00	7	9D
66	0	0		1424	E0A4	7BAD3A00	7	
67	0	0		1424	CC80	7BAD3A00	7	
68	0	0		1424	B85C	7BAD3A00	7	
69	0	0		1424	A438	7BAD3A00	7	
70	0	0		1424	9014	7BAD3A00	7	
71	1	0		119C	F7E0	F75A7400	6	
72	1	0		1424	8CE0	88B3A000	3	
73	0	0		1AA9	A120	7FBD0000	0	
74	1	0		1AA9	8677	7FBD0000	0	
75	0	0		2516	D548	9F7A8800	5	51
76	1	0		2516	B032	9F7A8800	5	
77	0	0		299A	9458	517A2000	3	
78	0	0		2516	D57C	A2F44000	2	
79	1	0		2516	B066	A2F44000	2	
80	0	0		299A	9458	5E910000	0	

Table K.8 – Decoder test sequence (sheet 3 of 7)

EC	D	MPS	CX	Qe (hexadecimal)	A (hexadecimal)	C (hexadecimal)	CT	B
81	1	0		2516	D57C	BD22F000	7	78
82	0	0		299A	9458	32F3C000	5	
83	0	0		2516	D57C	65E78000	4	
84	0	0		2516	B066	65E78000	4	
85	0	0		2516	8B50	65E78000	4	
86	1	0		1EDF	CC74	CBCF0000	3	
87	1	0		2516	F6F8	F1D00000	0	
88	1	0		299A	9458	7FB95400	6	55
89	0	0		32B4	A668	53ED5000	4	
90	0	0		2E17	E768	A7DAA000	3	
91	1	0		2E17	B951	A7DAA000	3	
92	0	0		32B4	B85C	72828000	1	
93	1	0		32B4	85A8	72828000	1	
94	0	0		3C3D	CAD0	7E3B7E00	7	BF
95	1	0		3C3D	8E93	7E3B7E00	7	
96	0	0		415E	F0F4	AF95F800	5	
97	1	0		415E	AF96	AF95F800	5	
98	0	0	CE	4639	82BC	82BBF000	4	
99	1	0		415E	8C72	8C71E000	3	
100	0	0	CE	4639	82BC	82BBC000	2	
101	1	0		415E	8C72	8C718000	1	
102	0	0	CE	4639	82BC	82BB0000	0	
103	1	0		415E	8C72	8C71FE00	7	FF 00
104	0	0	CE	4639	82BC	82BBFC00	6	
105	1	0		415E	8C72	8C71F800	5	
106	0	0	CE	4639	82BC	82BBF000	4	
107	1	0		415E	8C72	8C71E000	3	
108	0	0	CE	4639	82BC	82BBC000	2	
109	1	0		415E	8C72	8C718000	1	
110	0	0	CE	4639	82BC	82BB0000	0	
111	1	0		415E	8C72	8C71F800	7	FC
112	0	0	CE	4639	82BC	82BBF000	6	
113	1	0		415E	8C72	8C71E000	5	
114	0	0	CE	4639	82BC	82BBC000	4	
115	1	0		415E	8C72	8C718000	3	
116	0	0	CE	4639	82BC	82BB0000	2	
117	1	0		415E	8C72	8C700000	1	
118	0	0	CE	4639	82BC	82B80000	0	
119	1	0		415E	8C72	8C6AA200	7	51
120	0	0	CE	4639	82BC	82AD4400	6	

Table K.8 – Decoder test sequence (sheet 4 of 7)

EC	D	MPS	CX	Qe (hexadecimal)	A (hexadecimal)	C (hexadecimal)	CT	B
121	1	0		415E	8C72	8C548800	5	
122	0	0	CE	4639	82BC	82811000	4	
123	1	0		415E	8C72	8BFC2000	3	
124	0	0	CE	4639	82BC	81D04000	2	
125	1	0		415E	8C72	8A9A8000	1	
126	0	0	CE	4639	82BC	7F0D0000	0	
127	1	0		415E	8C72	85150800	7	84
128	0	0	CE	4639	82BC	74021000	6	
129	1	0		415E	8C72	6EFE2000	5	
130	0	0	CE	4639	82BC	47D44000	4	
131	0	0		415E	8C72	16A28000	3	
132	0	0		3C3D	9628	2D450000	2	
133	0	0		375E	B3D6	5A8A0000	1	
134	0	0		32B4	F8F0	B5140000	0	
135	1	0		32B4	C63C	B5140000	0	
136	0	0		3C3D	CAD0	86331C00	6	C7
137	1	0		3C3D	8E93	86331C00	6	
138	1	0		415E	F0F4	CF747000	4	
139	0	0	CE	4639	82BC	3FBCE000	3	
140	0	0		415E	8C72	0673C000	2	
141	0	0		3C3D	9628	0CE78000	1	
142	0	0		375E	B3D6	19CF0000	0	
143	0	0		32B4	F8F0	339F9C00	7	CE
144	0	0		32B4	C63C	339F9C00	7	
145	0	0		32B4	9388	339F9C00	7	
146	0	0		2E17	C1A8	673F3800	6	
147	1	0		2E17	9391	673F3800	6	
148	0	0		32B4	B85C	0714E000	4	
149	0	0		32B4	85A8	0714E000	4	
150	0	0		2E17	A5E8	0E29C000	3	
151	0	0		299A	EFA2	1C538000	2	
152	0	0		299A	C608	1C538000	2	
153	0	0		299A	9C6E	1C538000	2	
154	0	0		2516	E5A8	38A70000	1	
155	0	0		2516	C092	38A70000	1	
156	0	0		2516	9B7C	38A70000	1	
157	0	0		1EDF	ECCC	714E0000	0	
158	0	0		1EDF	CDED	714E0000	0	
159	0	0		1EDF	AF0E	714E0000	0	
160	0	0		1EDF	902F	714E0000	0	

Table K.8 – Decoder test sequence (sheet 5 of 7)

EC	D	MPS	CX	Qe (hexadecimal)	A (hexadecimal)	C (hexadecimal)	CT	B
161	1	0		1AA9	E2A0	E29DF200	7	F9
162	1	0		2516	D548	D5379000	4	
163	1	0		299A	9458	94164000	2	
164	1	0		32B4	A668	A5610000	0	
165	1	0		3C3D	CAD0	C6B4E400	6	39
166	1	0		415E	F0F4	E0879000	4	
167	0	0	CE	4639	82BC	61E32000	3	
168	0	0		415E	8C72	4AC04000	2	
169	1	0		3C3D	9628	95808000	1	
170	1	0		415E	F0F4	EE560000	7	00
171	0	0	CE	4639	82BC	7D800000	6	
172	1	0		415E	8C72	81FA0000	5	
173	0	0	CE	4639	82BC	6DCC0000	4	
174	1	0		415E	8C72	62920000	3	
175	1	0	CE	4639	82BC	2EFC0000	2	
176	1	0		4B85	F20C	BBF00000	0	
177	1	0	CE	504F	970A	2AD25000	7	28
178	0	0	CE	5522	8D76	55A4A000	6	
179	0	0		504F	AA44	3AA14000	5	
180	1	0		4B85	B3EA	75428000	4	
181	1	0	CE	504F	970A	19BB0000	3	
182	1	0	CE	5522	8D76	33760000	2	
183	1	0		59EB	E150	CDD80000	0	
184	0	1		59EB	B3D6	8CE6FA00	7	7D
185	1	0		59EB	B3D6	65F7F400	6	
186	1	1		59EB	B3D6	1819E800	5	
187	1	1		5522	B3D6	3033D000	4	
188	1	1		504F	BD68	6067A000	3	
189	0	1		4B85	DA32	C0CF4000	2	
190	1	1	CE	504F	970A	64448000	1	
191	1	1		4B85	A09E	3B130000	0	
192	0	1		4639	AA32	76268C00	7	46
193	0	1	CE	4B85	8C72	245B1800	6	
194	1	1	CE	504F	81DA	48B63000	5	
195	1	1		4B85	A09E	2E566000	4	
196	1	1		4639	AA32	5CACCC00	3	
197	0	1		415E	C7F2	B9598000	2	
198	1	1	CE	4639	82BC	658B0000	1	
199	0	1		415E	8C72	52100000	0	
200	0	1	CE	4639	82BC	0DF8E000	7	70

Table K.8 – Decoder test sequence (sheet 6 of 7)

EC	D	MPS	CX	Qe (hexadecimal)	A (hexadecimal)	C (hexadecimal)	CT	B
201	1	1		4B85	F20C	37E38000	5	
202	1	1		4B85	A687	37E38000	5	
203	1	1		4639	B604	6FC70000	4	
204	0	1		415E	DF96	DF8E0000	3	
205	1	1	CE	4639	82BC	82AC0000	2	
206	0	1		415E	8C72	8C520000	1	
207	1	1	CE	4639	82BC	827C0000	0	
208	0	1		415E	8C72	8BF31C00	7	8E
209	1	1	CE	4639	82BC	81BE3800	6	
210	0	1		415E	8C72	8A767000	5	
211	1	1	CE	4639	82BC	7EC4E000	4	
212	0	1		415E	8C72	8483C000	3	
213	1	1	CE	4639	82BC	72DF8000	2	
214	0	1		415E	8C72	6CB90000	1	
215	1	1	CE	4639	82BC	434A0000	0	
216	1	1		415E	8C72	0D8F9600	7	CB
217	1	1		3C3D	9628	1B1F2C00	6	
218	1	1		375E	B3D6	363E5800	5	
219	1	1		32B4	F8F0	6C7CB000	4	
220	1	1		32B4	C63C	6C7CB000	4	
221	0	1		32B4	9388	6C7CB000	4	
222	1	1		3C3D	CAD0	2EA2C000	2	
223	1	1		3C3D	8E93	2EA2C000	2	
224	1	1		375E	A4AC	5D458000	1	
225	0	1		32B4	DA9C	BA8B0000	0	
226	1	1		3C3D	CAD0	4A8F0000	6	C0
227	1	1		3C3D	8E93	4A8F0000	6	
228	0	1		375E	A4AC	951E0000	5	
229	1	1		3C3D	DD78	9F400000	3	
230	0	1		3C3D	A13B	9F400000	3	
231	0	1		415E	F0F4	E9080000	1	
232	1	1	CE	4639	82BC	72E40000	0	
233	0	1		415E	8C72	6CC3EC00	7	F6
234	1	1	CE	4639	82BC	435FD800	6	
235	1	1		415E	8C72	0DB9B000	5	
236	1	1		3C3D	9628	1B736000	4	
237	1	1		375E	B3D6	36E6C000	3	
238	1	1		32B4	F8F0	6DCD8000	2	
239	1	1		32B4	C63C	6DCD8000	2	
240	0	1		32B4	9388	6DCD8000	2	

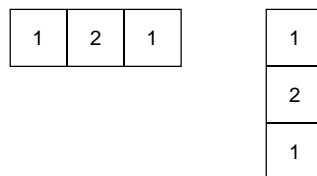
Table K.8 – Decoder test sequence (sheet 7 of 7)

EC	D	MPS	CX	Qe (hexadecimal)	A (hexadecimal)	C (hexadecimal)	CT	B
241	1	1		3C3D	CAD0	33E60000	0	
242	1	1		3C3D	8E93	33E60000	0	
Marker detected: zero byte fed to decoder								
243	1	1		375E	A4AC	67CC0000	7	
244	0	1		32B4	DA9C	CF980000	6	
245	0	1		3C3D	CAD0	9EC00000	4	
246	1	1		415E	F0F4	40B40000	2	
247	1	1		415E	AF96	40B40000	2	
248	1	1		3C3D	DC70	81680000	1	
249	0	1		3C3D	A033	81680000	1	
Marker detected: zero byte fed to decoder								
250	1	1		415E	F0F4	75C80000	7	
251	0	1		415E	AF96	75C80000	7	
252	0	1	CE	4639	82BC	0F200000	6	
253	1	1		4B85	F20C	3C800000	4	
254	1	1		4B85	A687	3C800000	4	
255	0	1		4639	B604	79000000	3	
256	0	1	CE	4B85	8C72	126A0000	2	

K.5 Low-pass downsampling filters for hierarchical coding

In this section simple examples are given of downsampling filters which are compatible with the upsampling filter defined in J.1.1.2.

Figure K.5 shows the weighting of neighbouring samples for simple one-dimensional horizontal and vertical low-pass filters. The output of the filter must be normalized by the sum of the neighbourhood weights.



TISO1770-93/d115

Figure K.5 – Low-pass filter example

The centre sample in Figure K.5 should be aligned with the left column or top line of the high resolution image when calculating the left column or top line of the low resolution image. Sample values which are situated outside of the image boundary are replicated from the sample values at the boundary to provide missing edge values.

If the image being downsampled has an odd width or length, the odd dimension is increased by 1 by sample replication on the right edge or bottom line before downsampling.

K.6 Domain of applicability of DCT and spatial coding techniques

The DCT coder is intended for lossy coding in a range from quite visible loss to distortion well below the threshold for visibility. However in general, DCT-based processes cannot be used for true lossless coding.

The lossless coder is intended for completely lossless coding. The lossless coding process is significantly less effective than the DCT-based processes for distortions near and above the threshold of visibility.

The point transform of the input to the lossless coder permits a very restricted form of lossy coding with the “lossless” coder. (The coder is still lossless after the input point transform.) Since the DCT is intended for lossy coding, there may be some confusion about when this alternative lossy technique should be used.

Lossless coding with a point transformed input is intended for applications which cannot be addressed by DCT coding techniques. Among these are

- true lossless coding to a specified precision;
- lossy coding with precisely defined error bounds;
- hierarchical progression to a truly lossless final stage.

If lossless coding with a point transformed input is used in applications which can be met effectively by DCT coding, the results will be significantly less satisfactory. For example, distortion in the form of visible contours usually appears when precision of the luminance component is reduced to about six bits. For normal image data, this occurs at bit rates well above those for which the DCT gives outputs which are visually indistinguishable from the source.

K.7 Domain of applicability of the progressive coding modes of operation

Two very different progressive coding modes of operation have been defined, progressive coding of the DCT coefficients and hierarchical progression. Progressive coding of the DCT coefficients has two complementary procedures, spectral selection and successive approximation. Because of this diversity of choices, there may be some confusion as to which method of progression to use for a given application.

K.7.1 Progressive coding of the DCT

In progressive coding of the DCT coefficients two complementary procedures are defined for decomposing the 8×8 DCT coefficient array, spectral selection and successive approximation. Spectral selection partitions zig-zag array of DCT coefficients into “bands”, one band being coded in each scan. Successive approximation codes the coefficients with reduced precision in the first scan; in each subsequent scan the precision is increased by one bit.

A single forward DCT is calculated for these procedures. When all coefficients are coded to full precision, the DCT is the same as in the sequential mode. Therefore, like the sequential DCT coding, progressive coding of DCT coefficients is intended for applications which need very good compression for a given level of visual distortion.

The simplest progressive coding technique is spectral selection; indeed, because of this simplicity, some applications may choose – despite the limited progression that can be achieved – to use only spectral selection. Note, however, that the absence of high frequency bands typically leads – for a given bit rate – to a significantly lower image quality in the intermediate stages than can be achieved with the more general progressions. The net coding efficiency at the completion of the final stage is typically comparable to or slightly less than that achieved with the sequential DCT.

A much more flexible progressive system is attained at some increase in complexity when successive approximation is added to the spectral selection progression. For a given bit rate, this system typically provides significantly better image quality than spectral selection alone. The net coding efficiency at the completion of the final stage is typically comparable to or slightly better than that achieved with the sequential DCT.

K.7.2 Hierarchical progression

Hierarchical progression permits a sequence of outputs of increasing spatial resolution, and also allows refinement of image quality at a given spatial resolution. Both DCT and spatial versions of the hierarchical progression are allowed, and progressive coding of DCT coefficients may be used in a frame of the DCT hierarchical progression.

The DCT hierarchical progression is intended for applications which need very good compression for a given level of visual distortion; the spatial hierarchical progression is intended for applications which need a simple progression with a truly lossless final stage. Figure K.6 illustrates examples of these two basic hierarchical progressions.

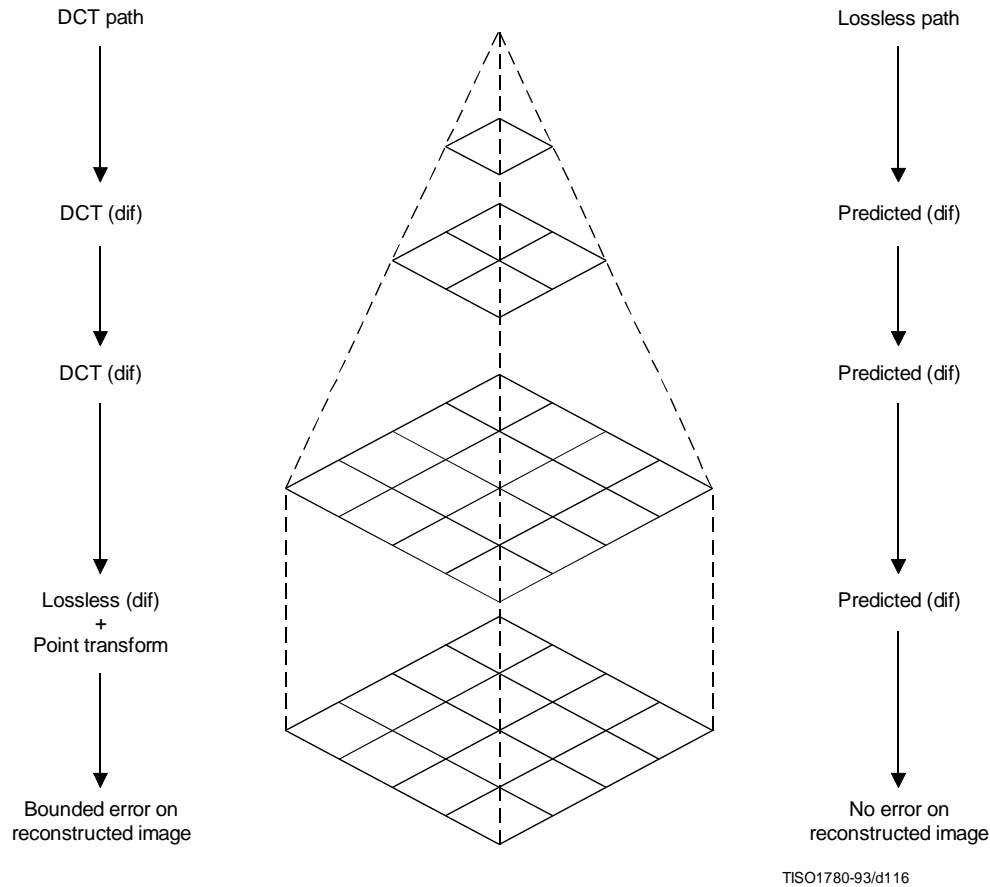


Figure K.6 – Sketch of the basic operations of the hierarchical mode

K.7.2.1 DCT Hierarchical progression

If a DCT hierarchical progression uses reduced spatial resolution, the early stages of the progression can have better image quality for a given bit rate than the early stages of non-hierarchical progressive coding of the DCT coefficients. However, at the point where the distortion between source and output becomes indistinguishable, the coding efficiency achieved with a DCT hierarchical progression is typically significantly lower than the coding efficiency achieved with a non-hierarchical progressive coding of the DCT coefficients.

While the hierarchical DCT progression is intended for lossy progressive coding, a final spatial differential coding stage can be used. When this final stage is used, the output can be almost lossless, limited only by the difference between the encoder and decoder IDCT implementations. Since IDCT implementations can differ significantly, truly lossless coding after a DCT hierarchical progression cannot be guaranteed. An important alternative, therefore, is to use the input point transform of the final lossless differential coding stage to reduce the precision of the differential input. This allows a bounding of the difference between source and output at a significantly lower cost in coded bits than coding of the full precision spatial difference would require.

K.7.2.2 Spatial hierarchical progression

If lossless progression is required, a very simple hierarchical progression may be used in which the spatial lossless coder with point transformed input is used as a first stage. This first stage is followed by one or more spatial differential coding stages. The first stage should be nearly lossless, such that the low order bits which are truncated by the point transform are essentially random – otherwise the compression efficiency will be degraded relative to non-progressive lossless coding.

K.8 Suppression of block-to-block discontinuities in decoded images

A simple technique is available for suppressing the block-to-block discontinuities which can occur in images compressed by DCT techniques.

The first few (five in this example) low frequency DCT coefficients are predicted from the nine DC values of the block and the eight nearest-neighbour blocks, and the predicted values are used to suppress blocking artifacts in smooth areas of the image.

The prediction equations for the first five AC coefficients in the zig-zag sequence are obtained as follows:

K.8.1 AC prediction

The sample field in a 3 by 3 array of blocks (each block containing an 8 × 8 array of samples) is modeled by a two-dimensional second degree polynomial of the form:

$$P(x,y) = A1(x^2y^2) + A2(x^2y) + A3(xy^2) + A4(x^2) + A5(xy) + A6(y^2) + A7(x) + A8(y) + A9$$

The nine coefficients A1 through A9 are uniquely determined by imposing the constraint that the mean of P(x,y) over each of the nine blocks must yield the correct DC-values.

Applying the DCT to the quadratic field predicting the samples in the central block gives a prediction of the low frequency AC coefficients depicted in Figure K.7.

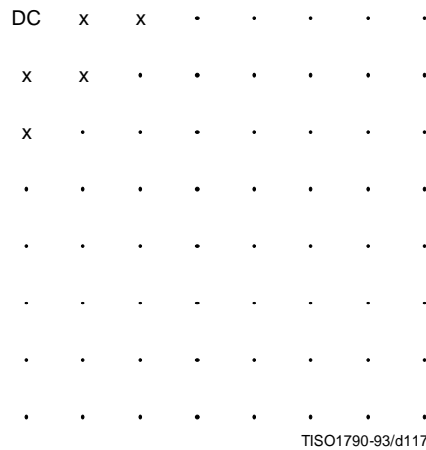


Figure K.7 – DCT array positions of predicted AC coefficients

The prediction equations derived in this manner are as follows:

For the two dimensional array of DC values shown

DC ₁	DC ₂	DC ₃
DC ₄	DC ₅	DC ₆
DC ₇	DC ₈	DC ₉

The unquantized prediction equations are

- AC₀₁ = 1,13885 (DC₄ – DC₆)
- AC₁₀ = 1,13885 (DC₂ – DC₈)
- AC₂₀ = 0,27881 (DC₂ + DC₈ – 2 × DC₅)
- AC₁₁ = 0,16213 ((DC₁ – DC₃) – (DC₇ – DC₉))
- AC₀₂ = 0,27881 (DC₄ + DC₆ – 2 × DC₅)

The scaling of the predicted AC coefficients is consistent with the DCT normalization defined in A.3.3.

K.8.2 Quantized AC prediction

The prediction equations can be mapped to a form which uses quantized values of the DC coefficients and which computes quantized AC coefficients using integer arithmetic. The quantized DC coefficients need to be scaled, however, such that the predicted coefficients have fractional bit precision.

First, the prediction equation coefficients are scaled by 32 and rounded to the nearest integer. Thus,

$$1,13885 \times 32 = 36$$

$$0,27881 \times 32 = 9$$

$$0,16213 \times 32 = 5$$

The multiplicative factors are then scaled by the ratio of the DC and AC quantization factors and rounded appropriately. The normalization defined for the DCT introduces another factor of 8 in the unquantized DC values. Therefore, in terms of the quantized DC values, the predicted quantized AC coefficients are given by the equations below. Note that if (for example) the DC values are scaled by a factor of 4, the AC predictions will have 2 fractional bits of precision relative to the quantized DCT coefficients.

$$QAC_{01} = ((R_d \times Q_{01}) + (36 \times Q_{00} \times (QDC_4 - QDC_6)))/(256 \times Q_{01})$$

$$QAC_{10} = ((R_d \times Q_{10}) + (36 \times Q_{00} \times (QDC_2 - QDC_8)))/(256 \times Q_{10})$$

$$QAC_{20} = ((R_d \times Q_{20}) + (9 \times Q_{00} \times (QDC_2 + QDC_8 - 2 \times QDC_5)))/(256 \times Q_{20})$$

$$QAC_{11} = ((R_d \times Q_{11}) + (5 \times Q_{00} \times ((QDC_1 - QDC_3) - (QDC_7 - QDC_9))))/(256 \times Q_{11})$$

$$QAC_{02} = ((R_d \times Q_{02}) + (9 \times Q_{00} \times (QDC_4 + QDC_6 - 2 \times QDC_5)))/(256 \times Q_{02})$$

where QDC_x and QAC_{xy} are the quantized and scaled DC and AC coefficient values. The constant R_d is added to get a correct rounding in the division. R_d is 128 for positive numerators, and -128 for negative numerators.

Predicted values should not override coded values. Therefore, predicted values for coefficients which are already non-zero should be set to zero. Predictions should be clamped if they exceed a value which would be quantized to a non-zero value for the current precision in the successive approximation.

K.9 Modification of dequantization to improve displayed image quality

For a progression where the first stage successive approximation bit, A_1 , is set to 3, uniform quantization of the DCT gives the following quantization and dequantization levels for a sequence of successive approximation scans, as shown in Figure K.8:

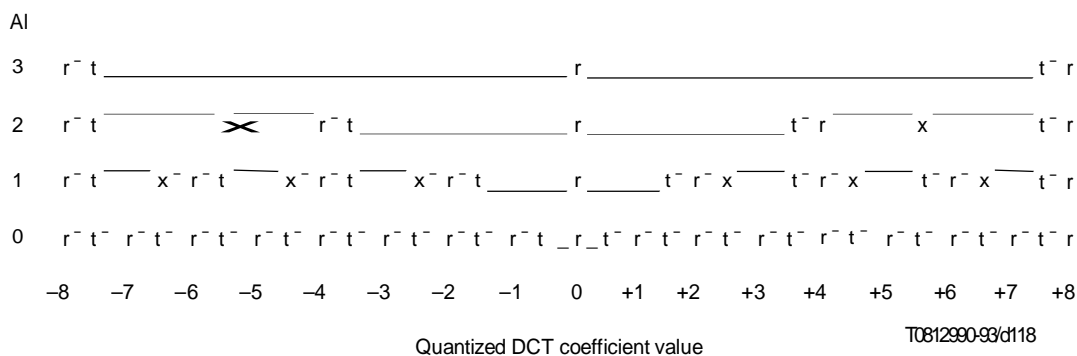


Figure K.8 - Illustration of two reconstruction strategies

The column to the left labelled “ A_1 ” gives the bit position specified in the scan header. The quantized DCT coefficient magnitudes are therefore divided by 2^{A_1} during that scan.

Referring to the final scan ($A_l = 0$), the points marked with “t” are the threshold values, while the points marked with “r” are the reconstruction values. The unquantized output is obtained by multiplying the horizontal scale in Figure K.8 by the quantization value.

The quantization interval for a coefficient value of zero is indicated by the depressed interval of the line. As the bit position A_l is increased, a “fat zero” quantization interval develops around the zero DCT coefficient value. In the limit where the scaling factor is very large, the zero interval is twice as large as the rest of the quantization intervals.

Two different reconstruction strategies are shown. The points marked “r” are the reconstruction obtained using the normal rounding rules for the DCT for the complete full precision output. This rule seems to give better image quality when high bandwidth displays are used. The points marked “x” are an alternative reconstruction which tends to give better images on lower bandwidth displays. “x” and “r” are the same for slice 0. The system designer must determine which strategy is best for the display system being used.

K.10 Example of point transform

The difference between the arithmetic-shift-right by P_t and divide by 2^{P_t} can be seen from the following:

After the level shift the DC has values from +127 to -128. Consider values near zero (after the level shift), and the case where $P_t = 1$:

Before level shift	Before point transform	After divide by 2	After shift-right-arithmetic 1
131	+3	+1	+1
130	+2	+1	+1
129	+1	0	0
128	0	0	0
127	-1	0	-1
126	-2	-1	-1
125	-3	-1	-2
124	-4	-2	-2
123	-5	-2	-3

The key difference is in the truncation of precision. The divide truncates the magnitude; the arithmetic shift truncates the LSB. With a divide by 2 we would get non-uniform quantization of the DC values; therefore we use the shift-right-arithmetic operation.

For positive values, the divide by 2 and the shift-right-arithmetic by 1 operations are the same. Therefore, the shift-right-arithmetic by 1 operation effectively is a divide by 2 when the point transform is done before the level shift.

Annex L

Patents

(This annex does not form an integral part of this Recommendation | International Standard)

L.1 Introductory remarks

The user's attention is called to the possibility that – for some of the coding processes specified in Annexes F, G, H, and J – compliance with this Specification may require use of an invention covered by patent rights.

By publication of this Specification, no position is taken with respect to the validity of this claim or of any patent rights in connection therewith. However, for each patent listed in this annex, the patent holder has filed with the Information Technology Task Force (ITTF) and the Telecommunication Standardization Bureau (TSB) a statement of willingness to grant a license under these rights on reasonable and non-discriminatory terms and conditions to applicants desiring to obtain such a license.

The criteria for including patents in this annex are:

- a) the patent has been identified by someone who is familiar with the technical fields relevant to this Specification, and who believes use of the invention covered by the patent is *required* for implementation of one or more of the coding processes specified in Annexes F, G, H, or J;
- b) the patent-holder has written a letter to the ITTF and TSB, stating willingness to grant a license to an unlimited number of applicants throughout the world under reasonable terms and conditions that are demonstrably free of any unfair discrimination.

This list of patents shall be updated, if necessary, upon publication of any revisions to the Recommendation | International Standard.

L.2 List of patents

The following patents may be required for implementation of any one of the processes specified in Annexes F, G, H, and J which uses arithmetic coding:

US 4,633,490, December 30, 1986, IBM, MITCHELL (J.L.) and GOERTZEL (G.): *Symmetrical Adaptive Data Compression/Decompression System*.

US 4,652,856, February 4, 1986, IBM, MOHIUDDIN (K.M.) and RISSANEN (J.J.): *A Multiplication-free Multi-Alphabet Arithmetic Code*.

US 4,369,463, January 18, 1983, IBM, ANASTASSIOU (D.) and MITCHELL (J.L.): *Grey Scale Image Compression with Code Words a Function of Image History*.

US 4,749,983, June 7, 1988, IBM, LANGDON (G.): *Compression of Multilevel Signals*.

US 4,935,882, June 19, 1990, IBM, PENNEBAKER (W.B.) and MITCHELL (J.L.): *Probability Adaptation for Arithmetic Coders*.

US 4,905,297, February 27, 1990, IBM, LANGDON (G.G.), Jr., MITCHELL (J.L.), PENNEBAKER (W.B.), and RISSANEN (J.J.): *Arithmetic Coding Encoder and Decoder System*.

US 4,973,961, November 27, 1990, AT&T, CHAMZAS (C.), DUTTWEILER (D.L.): *Method and Apparatus for Carry-over Control in Arithmetic Entropy Coding*.

US 5,025,258, June 18, 1991, AT&T, DUTTWEILER (D.L.): *Adaptive Probability Estimator for Entropy Encoding/Decoding*.

US 5,099,440, March 24, 1992, IBM, PENNEBAKER (W.B.) and MITCHELL (J.L.): *Probability Adaptation for Arithmetic Coders*.

Japanese Patent Application 2-46275, February 26, 1990, MEL ONO (F.), KIMURA (T.), YOSHIDA (M.), and KINO (S.): *Coding System*.

The following patent may be required for implementation of any one of the hierarchical processes specified in Annex H when used with a lossless final frame:

US 4,665,436, May 12, 1987, EI OSBORNE (J.A.) and SEIFFERT (C.): *Narrow Bandwidth Signal Transmission*.

No other patents required for implementation of any of the other processes specified in Annexes F, G, H, or J had been identified at the time of publication of this Specification.

L.3 Contact addresses for patent information

Director, Telecommunication Standardization Bureau (formerly CCITT)
International Telecommunication Union
Place des Nations
CH-1211 Genève 20, Switzerland
Tel. +41 (22) 730 5111
Fax: +41 (22) 730 5853

Information Technology Task Force
International Organization for Standardization
1, rue de Varembe
CH-1211 Genève 20, Switzerland
Tel: +41 (22) 734 0150
Fax: +41 (22) 733 3843

Program Manager, Licensing
Intellectual Property and Licensing Services
IBM Corporation
208 Harbor Drive
P.O. Box 10501
Stamford, Connecticut 08904-2501, USA
Tel: +1 (203) 973 7935
Fax: +1 (203) 973 7981 or +1 (203) 973 7982

Mitsubishi Electric Corp.
Intellectual Property License Department
1-2-3 Morunouchi, Chiyoda-ku
Tokyo 100, Japan
Tel: +81 (3) 3218 3465
Fax: +81 (3) 3215 3842

AT&T Intellectual Property Division Manager
Room 3A21
10 Independence Blvd.
Warren, NJ 07059, USA
Tel: +1 (908) 580 5392
Fax: +1 (908) 580 6355

Senior General Manager
Corporate Intellectual Property and Legal Headquarters
Canon Inc.
30-2 Shimomaruko 3-chome
Ohta-ku Tokyo 146 Japan
Tel: +81 (3) 3758 2111
Fax: +81 (3) 3756 0947

Chief Executive Officer
Electronic Imagery, Inc.
1100 Park Central Boulevard South
Suite 3400
Pompano Beach, FL 33064, USA
Tel: +1 (305) 968 7100
Fax: +1 (305) 968 7319

Annex M

Bibliography

(This annex does not form an integral part of this Recommendation | International Standard)

M.1 General references

LEGER (A.), OMACHI (T.), and WALLACE (G.K.): JPEG Still Picture Compression Algorithm, *Optical Engineering*, Vol. 30, No. 7, pp. 947-954, 1991.

RABBANI (M.) and JONES (P.): Digital Image Compression Techniques, *Tutorial Texts in Optical Engineering*, Vol. TT7, SPIE Press, 1991.

HUDSON (G.), YASUDA (H.) and SEBESTYEN (I.): The International Standardization of a Still Picture Compression Technique, *Proc. of IEEE Global Telecommunications Conference*, pp. 1016-1021, 1988.

LEGER (A.), MITCHELL (J.) and YAMAZAKI (Y.): Still Picture Compression Algorithm Evaluated for International Standardization, *Proc. of the IEEE Global Telecommunications Conference*, pp. 1028-1032, 1988.

WALLACE (G.), VIVIAN (R.) and POULSEN (H.): Subjective Testing Results for Still Picture Compression Algorithms for International Standardization, *Proc. of the IEEE Global Telecommunications Conference*, pp. 1022-1027, 1988.

MITCHELL (J.L.) and PENNEBAKER (W.B.): Evolving JPEG Colour Data Compression Standard, *Standards for Electronic Imaging Systems*, M. Nier, M.E. Courtot, Editors, SPIE, Vol. CR37, pp. 68-97, 1991.

WALLACE (G.K.): The JPEG Still Picture Compression Standard, *Communications of the ACM*, Vol. 34, No. 4, pp. 31-44, 1991.

NETRAVALI (A.N.) and HASKELL (B.G.): *Digital Pictures: Representation and Compression*, Plenum Press, New York 1988.

PENNEBAKER (W.B.) and MITCHELL (J.L.): *JPEG: Still Image Data Compression Standard*, Van Nostrand Reinhold, New York 1993.

M.2 DCT references

CHEN (W.), SMITH (C.H.) and FRALICK (S.C.): A Fast Computational Algorithm for the Discrete Cosine Transform, *IEEE Trans. on Communications*, Vol. COM-25, pp. 1004-1009, 1977.

AHMED (N.), NATARAJAN (T.) and RAO (K.R.): Discrete Cosine Transform, *IEEE Trans. on Computers*, Vol. C-23, pp. 90-93, 1974.

NARASINHA (N.J.) and PETERSON (A.M.): On the Computation of the Discrete Cosine Transform, *IEEE Trans. on Communications*, Vol. COM-26, No. 6, pp. 966-968, 1978.

DUHAMEL (P.) and GUILLEMOT (C.): Polynomial Transform Computation of the 2-D DCT, *Proc. IEEE ICASSP-90*, pp. 1515-1518, Albuquerque, New Mexico 1990.

FEIG (E.): A Fast Scaled DCT Algorithm, in *Image Processing Algorithms and Techniques*, Proc. SPIE, Vol. 1244, K.S. Pennington and R. J. Moorhead II, Editors, pp. 2-13, Santa Clara, California, 1990.

HOU (H.S.): A Fast Recursive Algorithm for Computing the Discrete Cosine Transform, *IEEE Trans. Acoust. Speech and Signal Processing*, Vol. ASSP-35, No. 10, pp. 1455-1461.

LEE (B.G.): A New Algorithm to Compute the Discrete Cosine Transform, *IEEE Trans. on Acoust., Speech and Signal Processing*, Vol. ASSP-32, No. 6, pp. 1243-1245, 1984.

LINZER (E.N.) and FEIG (E.): New DCT and Scaled DCT Algorithms for Fused Multiply/Add Architectures, *Proc. IEEE ICASSP-91*, pp. 2201-2204, Toronto, Canada, 1991.

VETTERLI (M.) and NUSSBAUMER (H.J.): Simple FFT and DCT Algorithms with Reduced Number of Operations, *Signal Processing*, 1984.

VETTERLI (M.): Fast 2-D Discrete Cosine Transform, *Proc. IEEE ICASSP-85*, pp. 1538-1541, Tampa, Florida, 1985.

ARAI (Y.), AGUI (T.), and NAKAJIMA (M.): A Fast DCT-SQ Scheme for Images, *Trans. of IEICE*, Vol. E.71, No. 11, pp. 1095-1097, 1988.

SUEHIRO (N.) and HATORI (M.): Fast Algorithms for the DFT and other Sinusoidal Transforms, *IEEE Trans. on Acoust., Speech and Signal Processing*, Vol ASSP-34, No. 3, pp. 642-644, 1986.

M.3 Quantization and human visual model references

CHEN (W.H.) and PRATT (W.K.): Scene adaptive coder, *IEEE Trans. on Communications*, Vol. COM-32, pp. 225-232, 1984.

GRANRATH (D.J.): The role of human visual models in image processing, *Proceedings of the IEEE*, Vol. 67, pp. 552-561, 1981.

LOHSCHELLER (H.): Vision adapted progressive image transmission, *Proceedings of EUSIPCO*, Vol. 83, pp. 191-194, 1983.

LOHSCHELLER (H.) and FRANKE (U.): Colour picture coding – Algorithm optimization and technical realization, *Frequenze*, Vol. 41, pp. 291-299, 1987.

LOHSCHELLER (H.): A subjectively adapted image communication system, *IEEE Trans. on Communications*, Vol. COM-32, pp. 1316-1322, 1984.

PETERSON (H.A.) *et al*: Quantization of colour image components in the DCT domain, *SPIE/IS&T 1991 Symposium on Electronic Imaging Science and Technology*, 1991.

M.4 Arithmetic coding references

LANGDON (G.): An Introduction to Arithmetic Coding, *IBM J. Res. Develop.*, Vol. 28, pp. 135-149, 1984.

PENNEBAKER (W.B.), MITCHELL (J.L.), LANGDON (G.) Jr., and ARPS (R.B.): An Overview of the Basic Principles of the Q-Coder Binary Arithmetic Coder, *IBM J. Res. Develop.*, Vol. 32, No. 6, pp. 717-726, 1988.

MITCHELL (J.L.) and PENNEBAKER (W.B.): Optimal Hardware and Software Arithmetic Coding Procedures for the Q-Coder Binary Arithmetic Coder, *IBM J. Res. Develop.*, Vol. 32, No. 6, pp. 727-736, 1988.

PENNEBAKER (W.B.) and MITCHELL (J.L.): Probability Estimation for the Q-Coder, *IBM J. Res. Develop.*, Vol. 32, No. 6, pp. 737-752, 1988.

MITCHELL (J.L.) and PENNEBAKER (W.B.): Software Implementations of the Q-Coder, *IBM J. Res. Develop.*, Vol. 32, No. 6, pp. 753-774, 1988.

ARPS (R.B.), TRUONG (T.K.), LU (D.J.), PASCO (R.C.) and FRIEDMAN (T.D.): A Multi-Purpose VLSI Chip for Adaptive Data Compression of Bilevel Images, *IBM J. Res. Develop.*, Vol. 32, No. 6, pp. 775-795, 1988.

ONO (F.), YOSHIDA (M.), KIMURA (T.) and KINO (S.): Subtraction-type Arithmetic Coding with MPS/LPS Conditional Exchange, *Annual Spring Conference of IECEED*, Japan, D-288, 1990.

DUTTWEILER (D.) and CHAMZAS (C.): Probability Estimation in Arithmetic and Adaptive-Huffman Entropy Coders, submitted to *IEEE Trans. on Image Processing*.

JONES (C.B.): An Efficient Coding System for Long Source Sequences, *IEEE Trans. Inf. Theory*, Vol. IT-27, pp. 280-291, 1981.

LANGDON (G.): Method for Carry-over Control in a Fifo Arithmetic Code String, *IBM Technical Disclosure Bulletin*, Vol. 23, No.1, pp. 310-312, 1980.

M.5 Huffman coding references

HUFFMAN (D.A.): A Method for the Construction of Minimum Redundancy codes, *Proc. IRE*, Vol. 40, pp. 1098-1101, 1952.

Image Compression and the Discrete Cosine Transform

Ken Cabeen and Peter Gent

Math 45

College of the Redwoods

Abstract. The mathematical equations of the DCT and its uses with image compression are explained.

Introduction

As our use of and reliance on computers continues to grow, so too does our need for efficient ways of storing large amounts of data. For example, someone with a web page or online catalog – that uses dozens or perhaps hundreds of images – will more than likely need to use some form of image compression to store those images. This is because the amount of space required to hold unadulterated images can be prohibitively large in terms of cost. Fortunately, there are several methods of image compression available today. These fall into two general categories: lossless and lossy image compression. The JPEG process is a widely used form of lossy image compression that centers around the Discrete Cosine Transform. The DCT works by separating images into parts of differing frequencies. During a step called quantization, where part of compression actually occurs, the less important frequencies are discarded, hence the use of the term "lossy." Then, only the most important frequencies that remain are used to retrieve the image in the decompression process. As a result, reconstructed images contain some distortion; but as we shall soon see, these levels of distortion can be adjusted during the compression stage. The JPEG method is used for both color and black-and-white images, but the focus of this article will be on compression of the latter.

The Process

The following is a general overview of the JPEG process. Later, we will take the reader through a detailed tour of JPEG's method so that a more comprehensive understanding of the process may be acquired.

1. The image is broken into 8x8 blocks of pixels.
2. Working from left to right, top to bottom, the DCT is applied to each block.
3. Each block is compressed through quantization.
4. The array of compressed blocks that constitute the image is stored in a drastically reduced amount of space.
5. When desired, the image is reconstructed through decompression, a process that uses the Inverse Discrete Cosine Transform (IDCT).

The DCT Equation

The DCT equation (Eq. 1) computes the i, j^{th} entry of the DCT of an image.

$$D(i, j) = \frac{1}{\sqrt{2N}} C(i)C(j) \sum_{x=0}^{N-1} \sum_{y=0}^{N-1} p(x, y) \cos \left[\frac{(2x+1)i\pi}{2N} \right] \cos \left[\frac{(2y+1)j\pi}{2N} \right] \quad 1$$

$$C(u) = \begin{cases} \frac{1}{\sqrt{2}} & \text{if } u = 0 \\ 1 & \text{if } u > 0 \end{cases} \quad 2$$

$p(x,y)$ is the x,y^{th} element of the image represented by the matrix p . N is the size of the block that the DCT is done on. The equation calculates one entry (i,j^{th}) of the transformed image from the pixel values of the original image matrix. For the standard 8x8 block that JPEG compression uses, N equals 8 and x and y range from 0 to 7. Therefore $D(i,j)$ would be as in Equation (3).

$$D(i,j) = \frac{1}{4}C(i)C(j) \sum_{x=0}^7 \sum_{y=0}^7 p(x,y) \cos\left[\frac{(2x+1)i\pi}{16}\right] \cos\left[\frac{(2y+1)j\pi}{16}\right] \quad 3$$

Because the DCT uses cosine functions, the resulting matrix depends on the horizontal, diagonal, and vertical frequencies. Therefore an image block with a lot of change in frequency has a very random looking resulting matrix, while an image matrix of just one color, has a resulting matrix of a large value for the first element and zeroes for the other elements.

The DCT Matrix

To get the matrix form of Equation (1), we will use the following equation

$$T_{i,j} = \left\{ \begin{array}{ll} \frac{1}{\sqrt{N}} & \text{if } i = 0 \\ \sqrt{\frac{2}{N}} \cos\left[\frac{(2j+1)i\pi}{2N}\right] & \text{if } i > 0 \end{array} \right\} \quad 4$$

For an 8x8 block it results in this matrix:

$$T = \begin{bmatrix} .3536 & .3536 & .3536 & .3536 & .3536 & .3536 & .3536 & .3536 \\ .4904 & .4157 & .2778 & .0975 & -.0975 & -.2778 & -.4157 & -.4904 \\ .4619 & .1913 & -.1913 & -.4619 & -.4619 & -.1913 & .1913 & .4619 \\ .4157 & -.0975 & -.4904 & -.2778 & .2778 & .4904 & .0975 & -.4157 \\ .3536 & -.3536 & -.3536 & .3536 & .3536 & -.3536 & -.3536 & .3536 \\ .2778 & -.4904 & .0975 & .4157 & -.4157 & -.0975 & .4904 & -.2778 \\ .1913 & -.4619 & .4619 & -.1913 & -.1913 & .4619 & -.4619 & .1913 \\ .0975 & -.2778 & .4157 & -.4904 & .4904 & -.4157 & .2778 & -.0975 \end{bmatrix}$$

The first row ($i = 1$) of the matrix has all the entries equal to $1/\sqrt{8}$ as expected from Equation (4).

The columns of T form an orthonormal set, so T is an orthogonal matrix. When doing the inverse DCT the orthogonality of T is important, as the inverse of T is T' which is easy to calculate.

Doing the DCT on an 8x8 Block

Before we begin, it should be noted that the pixel values of a black-and-white image range from 0 to 255 in steps of 1, where pure black is represented by 0, and pure white by 255. Thus it can be seen how a photo, illustration, etc. can be accurately represented by these 256 shades of gray.

Since an image comprises hundreds or even thousands of 8x8 blocks of pixels, the following description of what happens to one 8x8 block is a microcosm of the JPEG process;

what is done to one block of image pixels is done to all of them, in the order earlier specified.

Now, let's start with a block of image-pixel values. This particular block was chosen from the very upper- left-hand corner of an image.

$$Original = \begin{bmatrix} 154 & 123 & 123 & 123 & 123 & 123 & 123 & 136 \\ 192 & 180 & 136 & 154 & 154 & 154 & 136 & 110 \\ 254 & 198 & 154 & 154 & 180 & 154 & 123 & 123 \\ 239 & 180 & 136 & 180 & 180 & 166 & 123 & 123 \\ 180 & 154 & 136 & 167 & 166 & 149 & 136 & 136 \\ 128 & 136 & 123 & 136 & 154 & 180 & 198 & 154 \\ 123 & 105 & 110 & 149 & 136 & 136 & 180 & 166 \\ 110 & 136 & 123 & 123 & 123 & 136 & 154 & 136 \end{bmatrix}$$

Because the DCT is designed to work on pixel values ranging from -128 to 127, the original block is "leveled off" by subtracting 128 from each entry. This results in the following matrix.

$$M = \begin{bmatrix} 26 & -5 & -5 & -5 & -5 & -5 & -5 & 8 \\ 64 & 52 & 8 & 26 & 26 & 26 & 8 & -18 \\ 126 & 70 & 26 & 26 & 52 & 26 & -5 & -5 \\ 111 & 52 & 8 & 52 & 52 & 38 & -5 & -5 \\ 52 & 26 & 8 & 39 & 38 & 21 & 8 & 8 \\ 0 & 8 & -5 & 8 & 26 & 52 & 70 & 26 \\ -5 & -23 & -18 & 21 & 8 & 8 & 52 & 38 \\ -18 & 8 & -5 & -5 & -5 & 8 & 26 & 8 \end{bmatrix}$$

We are now ready to perform the Discrete Cosine Transform, which is accomplished by matrix multiplication.

$$D = TMT' \tag{5}$$

In Equation (5) matrix *M* is first multiplied on the left by the DCT matrix *T* from the previous section; this transforms the rows. The columns are then transformed by multiplying on the right by the transpose of the DCT matrix. This yields the following matrix.

$$D = \begin{bmatrix} 162.3 & 40.6 & 20.0 & 72.3 & 30.3 & 12.5 & -19.7 & -11.5 \\ 30.5 & 108.4 & 10.5 & 32.3 & 27.7 & -15.5 & 18.4 & -2.0 \\ -94.1 & -60.1 & 12.3 & -43.4 & -31.3 & 6.1 & -3.3 & 7.1 \\ -38.6 & -83.4 & -5.4 & -22.2 & -13.5 & 15.5 & -1.3 & 3.5 \\ -31.3 & 17.9 & -5.5 & -12.4 & 14.3 & -6.0 & 11.5 & -6.0 \\ -0.9 & -11.8 & 12.8 & 0.2 & 28.1 & 12.6 & 8.4 & 2.9 \\ 4.6 & -2.4 & 12.2 & 6.6 & -18.7 & -12.8 & 7.7 & 12.0 \\ -10.0 & 11.2 & 7.8 & -16.3 & 21.5 & 0.0 & 5.9 & 10.7 \end{bmatrix}$$

This block matrix now consists of 64 DCT coefficients, c_{ij} , where i and j range from 0 to 7. The top-left coefficient, c_{00} , correlates to the low frequencies of the original image block. As we move away from c_{00} in all directions, the DCT coefficients correlate to higher and higher frequencies of the image block, where c_{77} corresponds to the highest frequency. It is important to note that the human eye is most sensitive to low frequencies, and results from the quantization step will reflect this fact.

Quantization

Our 8x8 block of DCT coefficients is now ready for compression by quantization. A remarkable and highly useful feature of the JPEG process is that in this step, varying levels of image compression and quality are obtainable through selection of specific quantization matrices. This enables the user to decide on quality levels ranging from 1 to 100, where 1 gives the poorest image quality and highest compression, while 100 gives the best quality and lowest compression. As a result, the quality/compression ratio can be tailored to suit different needs.

Subjective experiments involving the human visual system have resulted in the JPEG standard quantization matrix. With a quality level of 50, this matrix renders both high compression and excellent decompressed image quality.

$$Q_{50} = \begin{bmatrix} 16 & 11 & 10 & 16 & 24 & 40 & 51 & 61 \\ 12 & 12 & 14 & 19 & 26 & 58 & 60 & 55 \\ 14 & 13 & 16 & 24 & 40 & 57 & 69 & 56 \\ 14 & 17 & 22 & 29 & 51 & 87 & 80 & 62 \\ 18 & 22 & 37 & 56 & 68 & 109 & 103 & 77 \\ 24 & 35 & 55 & 64 & 81 & 104 & 113 & 92 \\ 49 & 64 & 78 & 87 & 103 & 121 & 120 & 101 \\ 72 & 92 & 95 & 98 & 112 & 100 & 103 & 99 \end{bmatrix}$$

If, however, another level of quality and compression is desired, scalar multiples of the JPEG standard quantization matrix may be used. For a quality level greater than 50 (less compression, higher image quality), the standard quantization matrix is multiplied by $(100\text{-quality level})/50$. For a quality level less than 50 (more compression, lower image quality), the standard quantization matrix is multiplied by $50/\text{quality level}$. The scaled

quantization matrix is then rounded and clipped to have positive integer values ranging from 1 to 255. For example, the following quantization matrices yield quality levels of 10 and 90.

$$Q_{10} = \begin{bmatrix} 80 & 60 & 50 & 80 & 120 & 200 & 255 & 255 \\ 55 & 60 & 70 & 95 & 130 & 255 & 255 & 255 \\ 70 & 65 & 80 & 120 & 200 & 255 & 255 & 255 \\ 70 & 85 & 110 & 145 & 255 & 255 & 255 & 255 \\ 90 & 110 & 185 & 255 & 255 & 255 & 255 & 255 \\ 120 & 175 & 255 & 255 & 255 & 255 & 255 & 255 \\ 245 & 255 & 255 & 255 & 255 & 255 & 255 & 255 \\ 255 & 255 & 255 & 255 & 255 & 255 & 255 & 255 \end{bmatrix}$$

$$Q_{90} = \begin{bmatrix} 3 & 2 & 2 & 3 & 5 & 8 & 10 & 12 \\ 2 & 2 & 3 & 4 & 5 & 12 & 12 & 11 \\ 3 & 3 & 3 & 5 & 8 & 11 & 14 & 11 \\ 3 & 3 & 4 & 6 & 10 & 17 & 16 & 12 \\ 4 & 4 & 7 & 11 & 14 & 22 & 21 & 15 \\ 5 & 7 & 11 & 13 & 16 & 12 & 23 & 18 \\ 10 & 13 & 16 & 17 & 21 & 24 & 24 & 21 \\ 14 & 18 & 19 & 20 & 22 & 20 & 20 & 20 \end{bmatrix}$$

Quantization is achieved by dividing each element in the transformed image matrix D by the corresponding element in the quantization matrix, and then rounding to the nearest integer value. For the following step, quantization matrix Q_{50} is used.

$$C_{i,j} = \text{round}\left(\frac{D_{i,j}}{Q_{i,j}}\right) \tag{6}$$

$$C = \begin{bmatrix} 10 & 4 & 2 & 5 & 1 & 0 & 0 & 0 \\ 3 & 9 & 1 & 2 & 1 & 0 & 0 & 0 \\ -7 & -5 & 1 & -2 & -1 & 0 & 0 & 0 \\ -3 & -5 & 0 & -1 & 0 & 0 & 0 & 0 \\ -2 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

Recall that the coefficients situated near the upper-left corner correspond to the lower frequencies – to which the human eye is most sensitive – of the image block. In addition, the zeros represent the less important, higher frequencies that have been discarded, giving rise to

the lossy part of compression. As mentioned earlier, only the remaining nonzero coefficients will be used to reconstruct the image. It is also interesting to note the effect of different quantization matrices; use of Q_{10} would give C significantly more zeros, while Q_{90} would result in very few zeros.

Coding

The quantized matrix C is now ready for the final step of compression. Before storage, all coefficients of C are converted by an encoder to a stream of binary data (01101011...). In-depth coverage of the coding process is beyond the scope of this article. However, we can point out one key aspect that the reader is sure to appreciate. After quantization, it is quite common for most of the coefficients to equal zero. JPEG takes advantage of this by encoding quantized coefficients in the zig-zag sequence shown in Figure 1. The advantage lies in the consolidation of relatively large runs of zeros, which compress very well. The sequence in Figure 1 (4x4) continues for the entire 8x8 block.

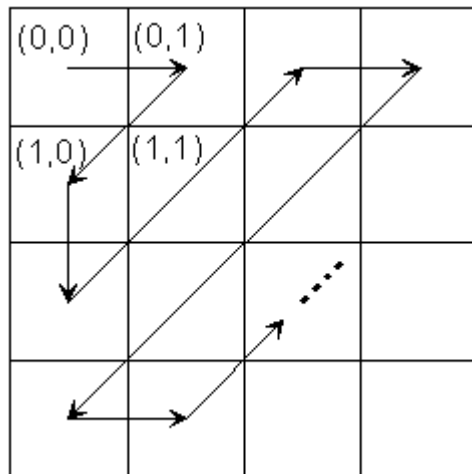


Figure 1

Decompression

Reconstruction of our image begins by decoding the bit stream representing the quantized matrix C . Each element of C is then multiplied by the corresponding element of the quantization matrix originally used.

$$R_{i,j} = Q_{i,j} \times C_{i,j} \quad 7$$

$$R = \begin{bmatrix} 160 & 44 & 20 & 80 & 24 & 0 & 0 & 0 \\ 36 & 108 & 14 & 38 & 26 & 0 & 0 & 0 \\ -98 & -65 & 16 & -48 & -40 & 0 & 0 & 0 \\ -42 & -85 & 0 & -29 & 0 & 0 & 0 & 0 \\ -36 & 22 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

The IDCT is next applied to matrix R , which is rounded to the nearest integer. Finally, 128 is added to each element of that result, giving us the decompressed JPEG version N of our original 8x8 image block M .

$$N = \text{round}(T' R T) + 128$$

8

Comparison of Matrices

Let us now see how the JPEG version of our original pixel block compares.

$$\begin{array}{l} \textit{Original} = \\ \textit{Decompressed} = \end{array} \begin{bmatrix} 154 & 123 & 123 & 123 & 123 & 123 & 123 & 136 \\ 192 & 180 & 136 & 154 & 154 & 154 & 136 & 110 \\ 254 & 198 & 154 & 154 & 180 & 154 & 123 & 123 \\ 239 & 180 & 136 & 180 & 180 & 166 & 123 & 123 \\ 180 & 154 & 136 & 167 & 166 & 149 & 136 & 136 \\ 128 & 136 & 123 & 136 & 154 & 180 & 198 & 154 \\ 123 & 105 & 110 & 149 & 136 & 136 & 180 & 166 \\ 110 & 136 & 123 & 123 & 123 & 136 & 154 & 136 \\ 149 & 134 & 119 & 116 & 121 & 126 & 127 & 128 \\ 204 & 168 & 140 & 144 & 155 & 150 & 135 & 125 \\ 253 & 195 & 155 & 166 & 183 & 165 & 131 & 111 \\ 245 & 185 & 148 & 166 & 184 & 160 & 124 & 107 \\ 188 & 149 & 132 & 155 & 172 & 159 & 141 & 136 \\ 132 & 123 & 125 & 143 & 160 & 166 & 168 & 171 \\ 109 & 119 & 126 & 128 & 139 & 158 & 168 & 166 \\ 111 & 127 & 127 & 114 & 118 & 141 & 147 & 135 \end{bmatrix}$$

This is a remarkable result, considering that nearly 70% of the DCT coefficients were discarded prior to image block decompression/reconstruction. Given that similar results will occur with the rest of the blocks that constitute the entire image, it should be no surprise that

the JPEG image will be scarcely distinguishable from the original. Remember, there are 256 possible shades of gray in a black-and-white picture, and a difference of, say, 10, is barely noticeable to the human eye.

Pepper Example

We can do the DCT and quantization process on the peppers image.



Figure 2 – Peppers

Each eight by eight block is hit with the DCT, resulting in the image shown in Figure 3.

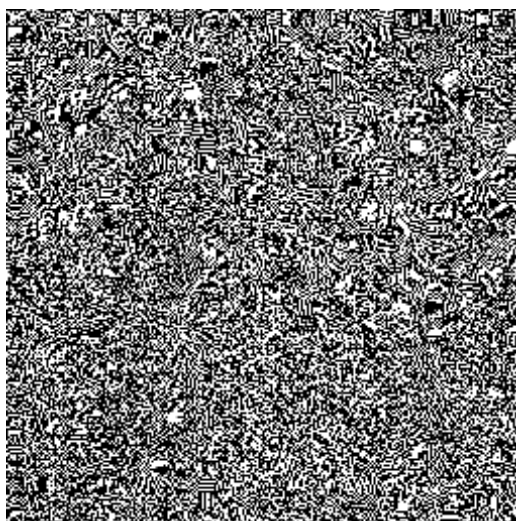


Figure 3 – DCT of Peppers

Each element in each block of the image is then quantized using a quantization matrix of quality level 50. At this point many of the elements become zeroed out, and the image takes up much less space to store.

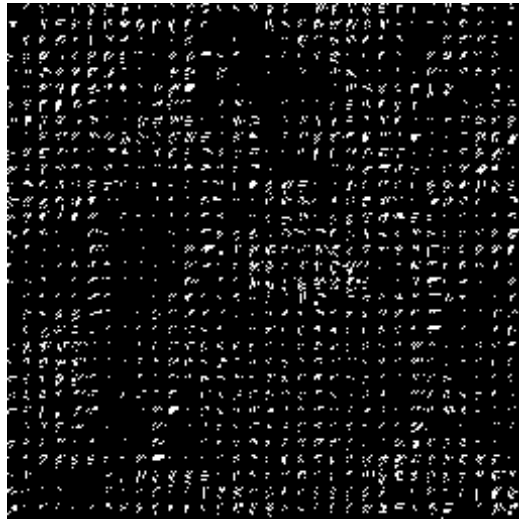


Figure 4 – Quantized DCT of Peppers

The image can now be decompressed using the inverse discrete cosine transform. At quality level 50 there is almost no visible loss in this image, but there is high compression. At lower quality levels, the quality goes down by a lot, but the compression does not increase very much.



Figure 5 – Original Peppers



Figure 6 – Quality 50 – 84% Zeros



Figure 7 – Quality 20 – 91% Zeros



Figure 8 – Quality 10 – 94% Zeros

More Examples

We can see what the compression does to other images. High contrast images, or images with a lot of high frequencies do not compress as well as smooth, low frequency images.



Figure 9 – Original



Figure 10 – Quality 15 – 90% Zeros

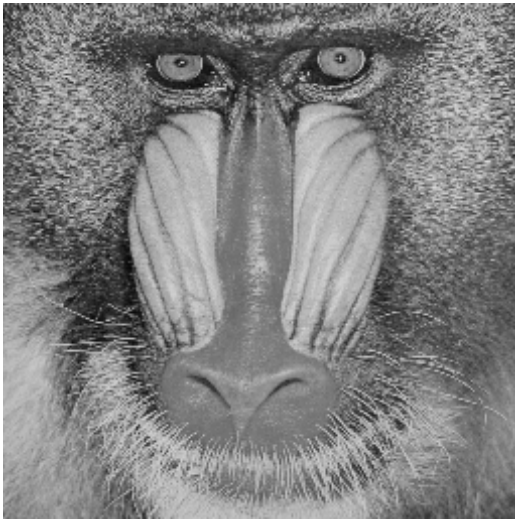


Figure 11 – Original

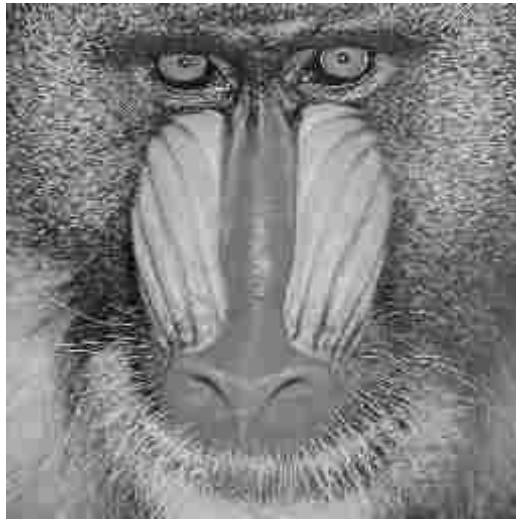


Figure 12 – Quality 15 – 88% Zeros

Bibliography

- Kesavan, Hareesh. *Choosing a DCT Quantization Matrix for JPEG Encoding*. Web page.
<http://www-ise.Stanford.EDU/class/ee392c/demos/kesavan/>
- McGowan, John. *The Discrete Cosine Transform*. Web page.
<http://www.rahul.net/jfm/dct.html>
- Wallace, Gregory K. *The JPEG Still Picture Compression Standard*. Paper submitted in December 1991 for publication in *IEEE Transactions on Consumer Electronics*.
- Wolfgang, Ray. *JPEG Tutorial*. Web page.
<http://www.imaging.org/tutorial/jpegtut1.html>
- Our special thanks to David Arnold, math instructor extraordinaire.

Image Coding Using Wavelet Transform

Marc Antonini, Michel Barlaud, *Member, IEEE*, Pierre Mathieu, and Ingrid Daubechies, *Member, IEEE*

Abstract—Image compression is now essential for applications such as transmission and storage in data bases. This paper proposes a new scheme for image compression taking into account psychovisual features both in the space and frequency domains; this new method involves two steps. First, we use a *wavelet transform* in order to obtain a set of biorthogonal subclasses of images; the original image is decomposed at different scales using a pyramidal algorithm architecture. The decomposition is along the vertical and horizontal directions and maintains constant the number of pixels required to describe the image. Second, according to Shannon's rate distortion theory, the wavelet coefficients are *vector quantized* using a multi-resolution codebook. Furthermore, to encode the wavelet coefficients, we propose a noise shaping bit allocation procedure which assumes that details at high resolution are less visible to the human eye. Finally, in order to allow the receiver to recognize a picture as quickly as possible at minimum cost, we present a progressive transmission scheme. It is shown that the wavelet transform is particularly well adapted to progressive transmission.

Keywords—Wavelet, biorthogonal wavelet, multiscale pyramidal algorithm, vector quantization, noise shaping, progressive transmission.

I. INTRODUCTION

IN many different fields, *digitized images* are replacing conventional analog images as photograph or x-rays. The volume of data required to describe such images greatly slow transmission and makes storage prohibitively costly. The information contained in the images must, therefore, be compressed by extracting only the visible elements, which are then encoded. The quantity of data involved is thus reduced substantially.

A fundamental goal of data compression is to reduce the bit rate for transmission or storage while maintaining an acceptable fidelity or image quality. Compression can be achieved by transforming the data, projecting it on a basis of functions, and then encoding this transform. Because of the nature of the image signal and the mechanisms of human vision, the transform used must accept nonstationarity and be well localized in both the space and frequency domains. To avoid redundancy, which hinders compression, the transform must be at least biorthogonal and lastly, in order to save CPU time, the corresponding algorithm must be fast. The two-dimensional wavelet transform defined by Meyer and Lemarié [31], [24], [25],

together with its implementation as described by Mallat [27], satisfies each of these conditions.

The compression method we have developed associates a wavelet transform and a vector quantization coding scheme. The wavelet coefficients are coded considering a noise shaping bit allocation procedure. This technique exploits the psychovisual as well as statistical redundancies in the image data, enabling bit rate reduction.

Section II describes the wavelet transforms used in this paper. After a quick review of wavelets in general, we explain in more detail the properties and construction of regular biorthogonal wavelet bases. We then extend this one-dimensional construction to a two-dimensional scheme with separable filters. The new coding scheme is next presented in Section III. We focus particularly in this section on the statistical properties of wavelet coefficients, on the asymptotic coding gain that can be achieved using vector quantization in the subimages, and on the optimal allocation across the subimages. Experimental results are given in Section IV for images taken within and outside of the training set.

II. WAVELETS

A. A Short Review of Wavelet Analysis

Wavelets are functions generated from one single function ψ by dilations and translations

$$\psi^{a,b}(t) = |a|^{-1/2} \psi\left(\frac{t-b}{a}\right).$$

(For this introduction we assume t is a one-dimensional variable). The *mother wavelet* ψ has to satisfy $\int dx \psi(x) = 0$, which implies at least some oscillations. (Technically speaking, the condition on ψ should be $\int d\omega |\Psi(\omega)|^2 |\omega|^{-1} < \infty$, where Ψ is the Fourier transform of ψ ; if $\psi(t)$ decays faster than $|t|^{-1}$ for $t \rightarrow \infty$, then this condition is equivalent to the one above). The definition of wavelets as dilates of one function means that high frequency wavelets correspond to $a < 1$ or narrow width, while low frequency wavelets have $a > 1$ or wider width.

The basic idea of the wavelet transform is to represent any arbitrary function f as a superposition of wavelets. Any such superposition decomposes f into different scale levels, where each level is then further decomposed with a resolution adapted to the level. One way to achieve such a decomposition writes f as an integral over a and b of $\psi^{a,b}$ with appropriate weighting coefficients [22]. In practice, one prefers to write f as a discrete superposition (sum rather than integral). Therefore, one introduces a discrete

Manuscript received February 7, 1990; revised March 26, 1991.

M. Antonini, M. Barlaud, and P. Mathieu are with LASSY 13S CNRS, Université de Nice-Sophia Antipolis, 06560 Valbonne, France.

I. Daubechies is with AT&T Bell Laboratories, Murray Hill, NJ 07974. IEEE Log Number 9106073.

tization, $a = a_0^m$, $b = nb_0a_0^m$, with $m, n \in \mathbb{Z}$, and $a_0 > 1$, $b_0 > 0$ fixed. The wavelet decomposition is then

$$f = \sum c_{m,n}(f) \psi_{m,n} \quad (1)$$

with $\psi_{m,n}(t) = \psi^{a_0^m, nb_0a_0^m}(t) = a_0^{-m/2} \psi(a_0^{-m}t - nb_0)$. Decompositions of this type were studied in [14], [15]. For $a_0 = 2$, $b_0 = 1$ there exist very special choices of ψ such that the $\psi_{m,n}$ constitute an orthonormal basis, so that

$$c_{m,n}(f) = \langle \psi_{m,n}, f \rangle = \int dx \psi_{m,n}(x) f(x)$$

in this case. Different bases of this nature were constructed by Stromberg [36], Meyer [31], Lemarié [24], Battle [7], and Daubechies [16]. All these examples correspond to a multiresolution analysis, a mathematical tool invented by Mallat [27], which is particularly well adapted to the use of wavelet bases in image analysis, and which gives rise to a fast computation algorithm.

In a multiresolution analysis, one really has *two* functions: the mother wavelet ψ and a *scaling function* ϕ . One also introduces dilated and translated versions of the scaling function, $\phi_{m,n}(x) = 2^{-m/2} \phi(2^{-m}x - n)$. For fixed m , the $\phi_{m,n}$ are orthonormal. We denote by V_m the space spanned by the $\phi_{m,n}$; these spaces V_m describe successive approximation spaces, $\dots V_2 \subset V_1 \subset V_0 \subset V_{-1} \subset V_{-2} \dots$, each with resolution 2^m . For each m , the $\psi_{m,n}$ span a space W_m which is exactly the orthogonal complement in V_{m-1} of V_m ; the coefficients $\langle \psi_{m,n}, f \rangle$, therefore, describe the information lost when going from an approximation of f with resolution 2^{m-1} to the coarser approximation with resolution 2^m . All this is translated into the following algorithm for the computation of the $c_{m,n}(f) = \langle \psi_{m,n}, f \rangle$ (for more details, see [27]):

$$\begin{aligned} c_{m,n}(f) &= \sum_k g_{2n-k} a_{m-1,k}(f) \\ a_{m,n}(f) &= \sum_k h_{2n-k} a_{m-1,k}(f) \end{aligned} \quad (2)$$

where $g_l = (-1)^l h_{-l+1}$ and $h_n = 2^{1/2} \int dx \phi(x-n) \phi(2x)$. In fact the $a_{m,n}(f)$ are coefficients characterizing the projection of f onto V_m . If the function f is given in sampled form, then one can take these samples for the highest order resolution approximation coefficients $a_{0,n}$, and (2) describes a subband coding algorithm on these sampled values, with low-pass filter h and high-pass filter g . Because of their association with orthonormal wavelet bases, these filters give exact reconstruction, i.e.:

$$a_{m-1,l}(f) = \sum_n [h_{2n-l} a_{m,n}(f) + g_{2n-l} c_{m,n}(f)]. \quad (3)$$

Most of the orthonormal wavelet bases have infinitely supported ψ , corresponding to filters h and g with infinitely many taps. The construction in [16] gives ψ with finite support, and therefore, corresponds to FIR filters. It follows that the orthonormal bases in [16] correspond to a subband coding scheme with exact reconstruction property, using the same FIR filters for reconstruction as

for decomposition. Such filters are well known since the work of Smith and Barnwell [35] and of Vetterli [37]. The extra ingredient in the orthonormal wavelet decomposition is that it writes the signal to be decomposed as a superposition of reasonably *smooth* elementary building blocks. The filters must satisfy the additional condition:

$$\prod_{k=1}^{\infty} H(2^{-k}\xi)$$

decay faster than $C(1 + |\xi|)^{-\epsilon-0.5}$ as $|\xi| \rightarrow \infty$, for some $\epsilon > 0$, where

$$H(\xi) = 2^{-1/2} \sum_n h_n e^{-jn\xi}.$$

This extra regularity requirement is usually not satisfied by the exact reconstruction filters in the ASSP literature.

B. Applications of Wavelet Bases to Image Analysis

1) *Biorthogonal Wavelet Bases*: Since images are mostly smooth (except for occasional edges) it seems appropriate that an exact reconstruction subband coding scheme for image analysis should correspond to an orthonormal basis with a reasonably smooth mother wavelet. In order to have fast computation, the filters should be short (short filters lead to less smoothness, however, so they cannot be too short). On the other hand it is desirable that the FIR filters used be linear phase, since such filters can be easily cascaded in pyramidal filter structures without the need for phase compensation. Unfortunately, there are no nontrivial orthonormal linear phase FIR filters with the exact reconstruction property [35], regardless of any regularity considerations. The only symmetric exact reconstruction filters are those corresponding to the Haar basis, i.e., $h_0 = h_1 = 2^{1/2}$ and $g_0 = -g_1 = 2^{1/2}$, with all other $h_n, g_n = 0$.

One can preserve linear phase (corresponding to symmetry for the wavelet) by relaxing the orthonormality requirement, and using biorthogonal bases. It is then still possible to construct examples where the mother wavelets have arbitrarily high regularity.

In such a scheme, we still decompose as in (2), but reconstruction becomes

$$a_{m-1,l}(f) = \sum_n [\tilde{h}_{2n-l} a_{m,n}(f) + \tilde{g}_{2n-l} c_{m,n}(f)] \quad (4)$$

where the filters \tilde{h}, \tilde{g} may be different from h, g . In order to have exact reconstruction, we impose:

$$\begin{aligned} \tilde{g}_n &= (-1)^n h_{-n+1} \\ g_n &= (-1)^n \tilde{h}_{-n+1} \end{aligned} \quad \sum_n h_n \tilde{h}_{n+2k} = \delta_{k,0}. \quad (5)$$

So far, we have not performed anything differently from the usual exact reconstruction subband coding schemes with synthesis filters different from the decomposition filters. If the filters satisfy the additional condition that:

$$\prod_{k=1}^{\infty} \tilde{H}(2^{-k}\xi) \quad \text{and} \quad \prod_{k=1}^{\infty} H(2^{-k}\xi) \quad (6a)$$

decay faster than $C(1 + |\xi|)^{-\epsilon - 0.5}$ as $|\xi| \rightarrow \infty$, for some $\epsilon > 0$, where

$$\tilde{H}(\xi) = 2^{-1/2} \sum_n \tilde{h}_n e^{-jn\xi} \quad H(\xi) = 2^{-1/2} \sum_n h_n e^{-jn\xi} \quad (6b)$$

then we can give the following interpretation to (2) and (4). Define functions ϕ and $\tilde{\phi}$ by

$$\phi(x) = \sum_n h_n \phi(2x - n) \quad \tilde{\phi}(x) = \sum_n \tilde{h}_n \tilde{\phi}(2x - n).$$

Their Fourier transforms are exactly the infinite products (6a), and they are, therefore, well-defined square integrable functions, compactly supported if the filters h and \tilde{h} are FIR. Define also

$$\psi(x) = \sum_n g_n \phi(2x - n) \quad \tilde{\psi}(x) = \sum_n \tilde{g}_n \tilde{\phi}(2x - n).$$

Then, the $a_{m,n}(f)$ and $c_{m,n}(f)$ in (2) can be rewritten as:

$$a_{m,n}(f) = \langle \phi_{m,n}, f \rangle = 2^{-m/2} \int dx \phi_{m,n}(x) f(x)$$

$$c_{m,n}(f) = \langle \psi_{m,n}, f \rangle = 2^{-m/2} \int dx \psi_{m,n}(x) f(x)$$

and reconstruction is simply:

$$f = \sum_{m,n} \langle \psi_{m,n}, f \rangle \tilde{\psi}_{m,n}. \quad (7)$$

The filter bank structure with the associating wavelets and scaling functions is depicted on the following subband coding scheme (Fig. 1).

If the infinite products in (6a) decay even faster than imposed above, then ϕ and $\tilde{\phi}$ and consequently ψ and $\tilde{\psi}$ will be reasonably smooth. Note that (7) is very similar to the orthonormal decomposition described in Section II-A; the only difference is that the expansion of f with respect to the basis $\tilde{\psi}_{m,n}$ uses coefficients computed via the dual basis $\psi_{m,n}$ with $\tilde{\psi}$ different from ψ . This interpretation is not possible for all exact reconstruction subband coding schemes; in particular, convergence of the infinite products (6a) is only possible if

$$\sum_n h_n = 2^{1/2} \quad \text{and} \quad \sum_n \tilde{h}_n = 2^{1/2}.$$

Moreover, (7) can only hold if

$$\sum_n (-1)^n h_n = 0 \quad \text{and} \quad \sum_n (-1)^n \tilde{h}_n = 0.$$

Most exact reconstruction subband coding schemes do not satisfy these conditions.

Biorthogonal bases of wavelets have recently been constructed, with regularity simultaneously but independently, by Cohen, Daubechies and Feauveau [12] and by Herley and Vetterli [38]. Reference [12] contains a detailed mathematical study, with proofs that, under the conditions stated above, the wavelets do indeed constitute numerically stable bases (Riesz bases) and a discussion of necessary and sufficient conditions for regularity. In [18]

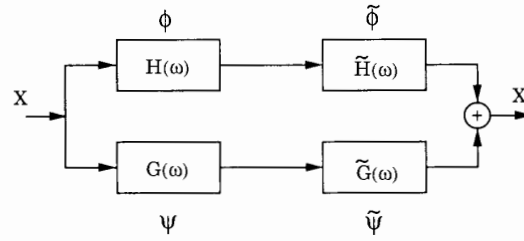


Fig. 1. Filter bank structure and the associating wavelets.

Feauveau explores the construction from the point of view of multiresolution spaces rather than from the filters. Basically one has two hierarchies of spaces in the biorthogonal case, each corresponding to one pair of filters.

It is shown in [12] that arbitrarily high regularity can be achieved by both ψ and $\tilde{\psi}$, provided one chooses sufficiently long filters. In particular, if the functions ψ and $\tilde{\psi}$ are, respectively, $(k - 1)$ and $(\tilde{k} - 1)$ times continuously differentiable, then the trigonometric polynomials $H(\xi)$ and $\tilde{H}(\xi)$ have to be divisible by $(1 + e^{-j\xi})^k$ and $(1 + e^{-j\xi})^{\tilde{k}}$, respectively, so that the length of the corresponding filters h, \tilde{h} has to exceed k, \tilde{k} .

By (5), divisibility of $\tilde{H}(\xi)$ by $(1 + e^{-j\xi})^{\tilde{k}}$ means that $\tilde{\psi}$ will have \tilde{k} consecutive moments zero:

$$\int dx x^l \tilde{\psi}(x) = 0, \quad \text{for } l = 0, 1, \dots, \tilde{k} - 1.$$

For more details concerning this discussion, see [12].

It is well known (and it can easily be checked by using Taylor expansions) that if ψ has \tilde{k} moments zero, then the coefficients $\langle \psi_{m,n}, f \rangle$ will represent functions f , which are \tilde{k} times differentiable, with a high compression potential (many coefficients will be negligibly small).

Many examples of biorthogonal wavelet bases with reasonably regular ψ and $\tilde{\psi}$ can be constructed; for our applications, the regularity of the elementary building blocks $\tilde{\psi}_{m,n}$, which is linked to the number of zero moments of $\tilde{\psi}$, is more important than the regularity of the $\psi_{m,n}$ or the number of zero moments of ψ . Within the limits imposed by the support widths, we will, therefore, try to choose \tilde{k} as large as possible.

In terms of trigonometric polynomials $H(\xi)$ and $\tilde{H}(\xi)$, the exact reconstruction requirement condition on h and \tilde{h} given in (5) reduces to (for symmetric filters)

$$H(\xi)\tilde{H}(\xi) + H(\xi + \pi)\tilde{H}(\xi + \pi) = 1. \quad (8)$$

Together with divisibility of H and \tilde{H} , respectively, by $(1 + e^{-j\xi})^k$ and $(1 + e^{-j\xi})^{\tilde{k}}$, this leads to (see [12])

$$H(\xi)\tilde{H}(\xi) = \cos(\xi/2)^{2l} \left[\sum_{p=0}^{l-1} \binom{l-1+p}{p} \cdot \sin(\xi/2)^{2p} + \sin(\xi/2)^{2l} R(\xi) \right] \quad (9)$$

where $R(\xi)$ is an odd polynomial in $\cos(\xi)$, and where $2l = k + \tilde{k}$ (symmetry of h and \tilde{h} forces $k + \tilde{k}$ to be even).

TABLE I
FILTER COEFFICIENTS FOR THE SPLINE FILTERS WITH $l = 3, k = 4, \bar{k} = 2$

n	0	± 1	± 2	± 3	± 4
$2^{-1/2}h_n$	45/64	19/64	-1/8	-3/64	3/128
$2^{-1/2}\tilde{h}_n$	1/2	1/4	0	0	0

Many examples are possible. We have studied in particular the following three examples, which belong to three different families.

2) *Spline Filters*: One can choose, e.g., $R \equiv 0$, with $\tilde{H}(\xi) = \cos(\xi/2)^{\bar{k}} e^{-j\kappa\xi/2}$ where $\kappa = 0$ if \bar{k} is even, $\kappa = 1$ if \bar{k} is odd. This corresponds to the filters called ‘‘spline filters’’ in [12] (because the corresponding function $\tilde{\phi}$ is a *B-spline* function) or ‘‘binomial filters’’ in [38] (because the \tilde{h} are simply binomial coefficients). It then follows that:

$$H(\xi) = \cos(\xi/2)^{2l-\bar{k}} e^{j\kappa\xi/2} \cdot \left[\sum_{p=0}^{l-1} \binom{l-1+p}{p} \sin(\xi/2)^{2p} \right]. \quad (10)$$

We have looked at one example from this family; it corresponds to $l = 3, \bar{k} = 2$. The coefficients h_n and \tilde{h}_n are listed in Table I; the corresponding scaling functions and wavelets are plotted in Fig. 2.

It is clear that the two filters in the first example have very uneven length. This is typical for all the examples in this family of ‘‘spline filters.’’

3) *A Spline Variant with Less Dissimilar Lengths*: This family still uses $R \equiv 0$ in (9), but factorizes the right-hand side of (9), breaking up the polynomial of degree $l - 1$ in $\sin(\xi/2)$ into a product of two polynomials in $\sin(\xi/2)$ with real coefficients, one to be allocated to H , the other to \tilde{H} , so as to make the lengths of h and \tilde{h} as close as possible.

The example presented here is the ‘‘smallest’’ one in this family (shortest h and \tilde{h}); it corresponds to $l = 4$ and $k = 4$. The filter coefficients are listed in Table II; the corresponding scaling functions and wavelets are plotted in Fig. 3.

Note that, unlike examples 1 and 3 where the $2^{-1/2}h_n, 2^{-1/2}\tilde{h}_n$ are rational, the entries in Table II are truncated decimal expansions of irrational numbers. The functions ϕ in examples 1 and 2 look very similar (compare Figs. 2(a) and 3(a)); a more detailed analysis shows that the one in example 2 is more regular, however. Both correspond to 4 vanishing moments for $\tilde{\psi}$.

4) *Filters Close to Orthonormal Filters*: Finally, there exist many examples for which $R \neq 0$. In particular there exists a special choice of R for which the two filters are very close to each other, and both very close to an orthonormal wavelet filter.

Surprisingly, for the first example of this series, one of the two filters is a Laplacian pyramid filter proposed in [9]. It corresponds to $l = 2, k = 2$ and $R(\xi) = 48 \cos(\xi)/175$. The filter coefficients are listed in Table III; the corresponding scaling functions and

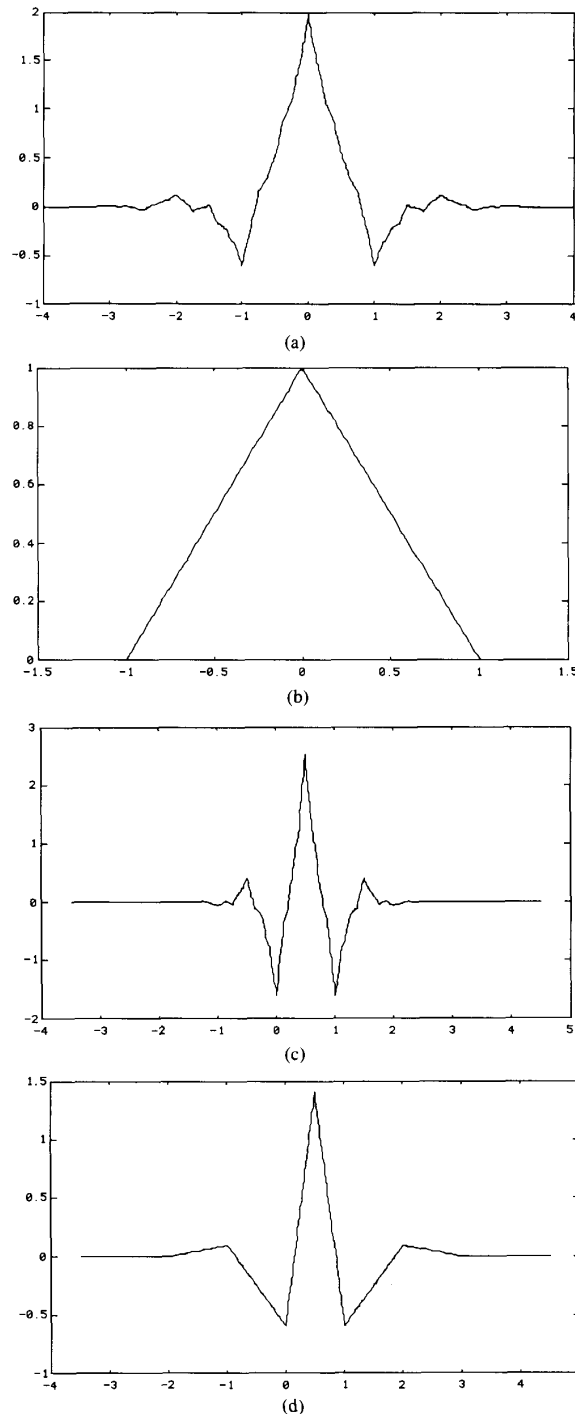


Fig. 2. Scaling functions $\phi, \tilde{\phi}$ and wavelets $\psi, \tilde{\psi}$ for example 1 (spline filters with $l = 3, k = 4, \bar{k} = 2$). (a) Scaling function ϕ . (b) Scaling function $\tilde{\phi}$. (c) Wavelet ψ . (d) Wavelet $\tilde{\psi}$.

wavelets are plotted in Fig. 4. It is clear that the scaling functions ϕ and $\tilde{\phi}$ are very similar, corresponding to very similar ψ and $\tilde{\psi}$. Note that in this case, the filter coefficients are again rational.

TABLE II
FILTER COEFFICIENTS FOR THE SPLINE VARIANT WITH LESS DISSIMILAR LENGTHS, WITH $l = 4 = k, \bar{k} = 4$

n	0	± 1	± 2	± 3	± 4
$2^{-1/2}h_n$	0.602 949	0.266 864	-0.078 223	-0.016 864	0.026 749
$2^{-1/2}\bar{h}_n$	0.557 543	0.295 636	-0.028 772	-0.045 636	0

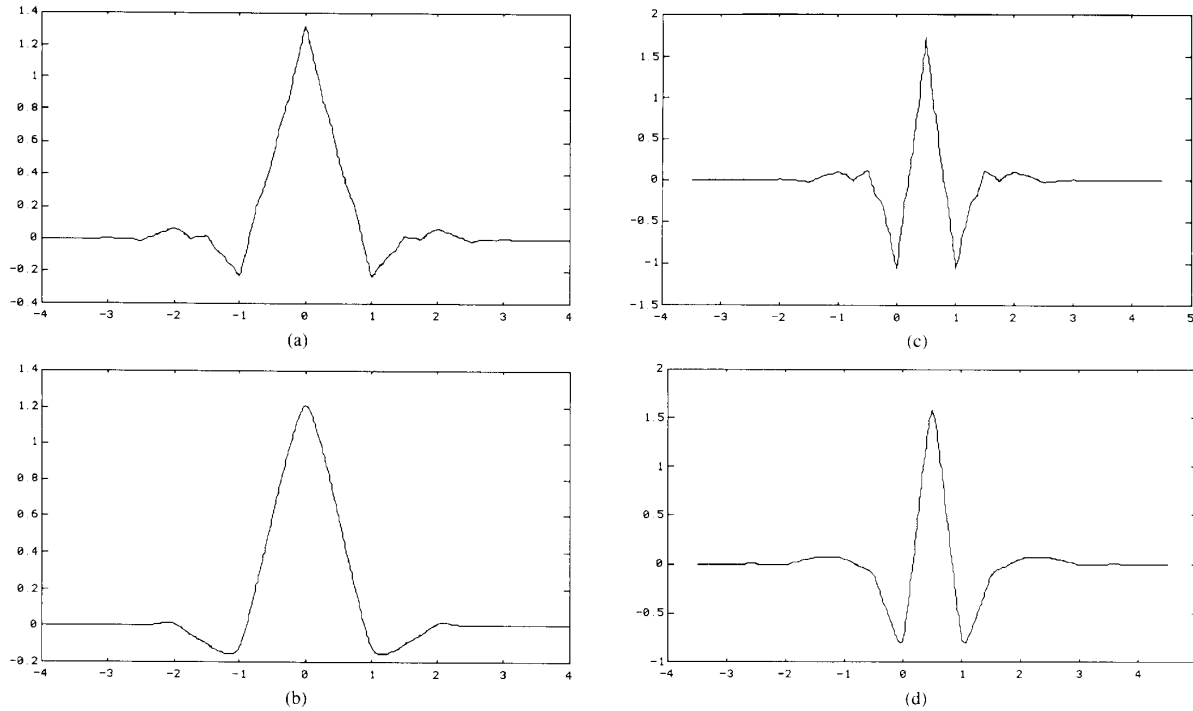


Fig. 3. Scaling functions $\phi, \tilde{\phi}$ and wavelets $\psi, \tilde{\psi}$ for example 2 (spline variant with less dissimilar lengths: $l = 4 = k, \bar{k} = 4$). (a) Scaling function ϕ . (b) Scaling function $\tilde{\phi}$. (c) Wavelet ψ . (d) Wavelet $\tilde{\psi}$.

TABLE III
FILTER COEFFICIENTS FOR EXAMPLE 3. THE ENTRIES ARE RATIONAL, AND THE TWO FILTERS ARE VERY CLOSE. THE h -FILTER COINCIDES WITH A LAPLACIAN PYRAMID FILTER PROPOSED IN [9]. IN THIS CASE $l = 2 = k, \bar{k} = 2$

n	0	± 1	± 2	± 3	± 4
$2^{-1/2}h_n$	0.6	0.25	-0.05	0	0
$2^{-1/2}\bar{h}_n$	17/28	73/280	-3/56	-3/280	0

The two biorthogonal filters in this example are both close to an orthonormal wavelet filter of length 6 constructed in [17], where it was called a ‘coiflet.’ Being an orthonormal wavelet filter, the coiflet is nonsymmetric. The filters in this example are shorter than in examples 1 and 2, but k is also smaller. The next example in this family corresponds to $k = 4$ (and $l = 4$); the filters h and \bar{h} then have length 9 and 15; they are both close to a coiflet of length 12.

5) *Extension to the Two-Dimensional Case:* There ex-

ist various extensions of the one-dimensional wavelet transform to higher dimensions. We follow Mallat [27] and use a two-dimensional wavelet transform in which horizontal and vertical orientations are considered preferential.

In two-dimensional wavelet analysis one introduces, like in the one-dimensional case, a scaling function $\phi(x, y)$ such that:

$$\phi(x, y) = \phi(x)\phi(y) \tag{11}$$

where $\phi(x)$ is a one-dimensional scaling function.

Let $\psi(x)$ be the one-dimensional wavelet associated with the scaling function $\phi(x)$. Then, the three two-dimensional wavelets are defined as:

$$\begin{aligned} \psi^H(x, y) &= \phi(x)\psi(y) \\ \psi^V(x, y) &= \psi(x)\phi(y) \\ \psi^D(x, y) &= \psi(x)\psi(y). \end{aligned} \tag{12}$$

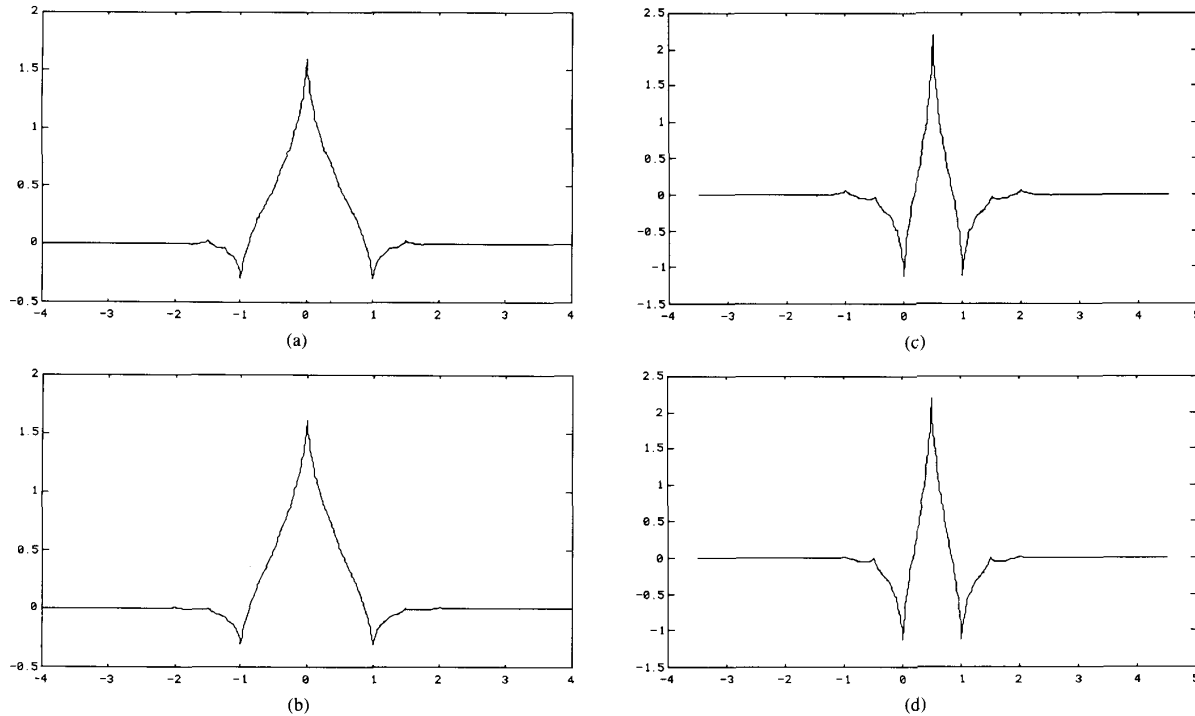


Fig. 4. Scaling functions ϕ , $\tilde{\phi}$ and wavelets ψ , $\tilde{\psi}$ for example 3 (biorthogonal filters close to an orthonormal wavelet filter, $l = 2 = k$, $\tilde{l} = 2$). (a) Scaling function ϕ . (b) Scaling function $\tilde{\phi}$. (c) Wavelet ψ . (d) Wavelet $\tilde{\psi}$.

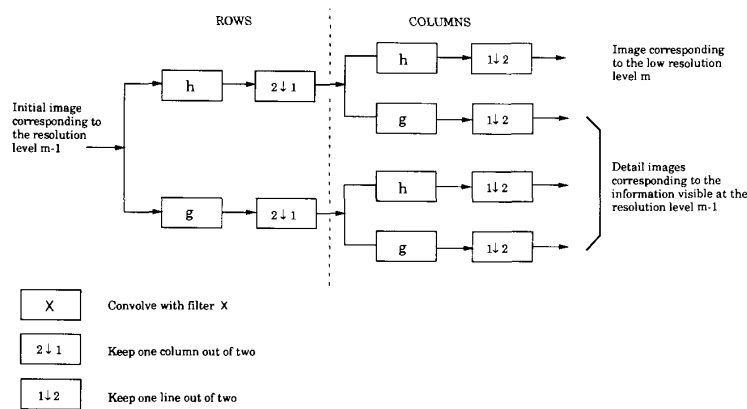


Fig. 5. One stage in a multiscale image decomposition.

Fig. 5 represents one stage in a *multiscale pyramidal* decomposition of an image: wavelet coefficients of the image are computed, as in the one-dimensional case (Sections II-A and II-B.1), using a subband coding algorithm. The filters h and g are one-dimensional filters. This decomposition provides subimages corresponding to different resolution levels and orientations (see Fig. 6). The reconstruction scheme of the image is presented Fig. 7.

To compare the three different filters presented in this paper, we have decomposed the image Lena (Fig. 16) with each of these filters. The results are presented in Fig. 8.

In Fig. 8(a) we can see the normalized detail subimages at different resolution levels $m = 1$, $m = 2$, and $m = 3$ (wavelet coefficients) and in Fig. 8(b) the low resolution level subimages.

III. IMAGE CODING APPLICATION

A. Statistical Properties of Wavelet Coefficients

The performance of a coder used for a given resolution and direction can be determined by the statistics of the corresponding subimage, i.e., its probability density function (PDF).

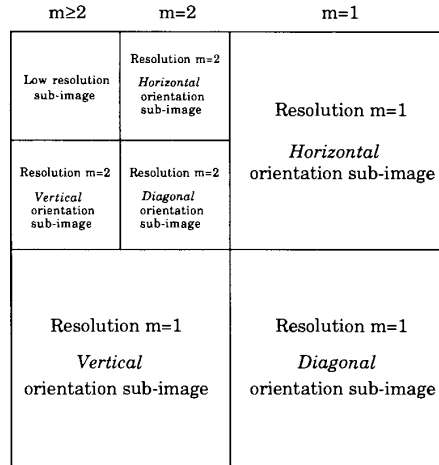


Fig. 6. Image decomposition.

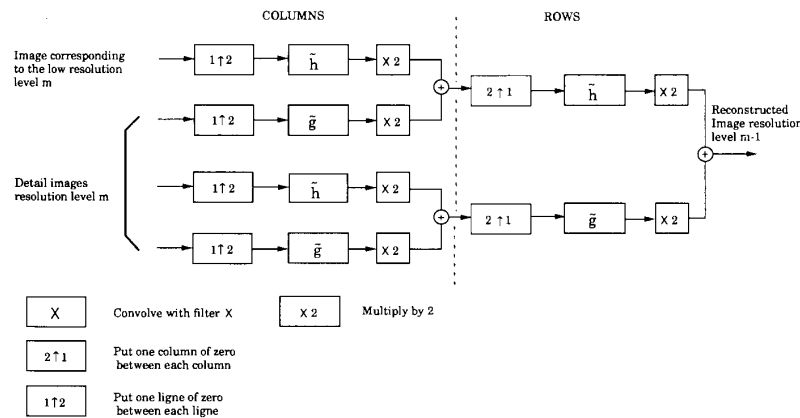


Fig. 7. One stage in a multiscale image reconstruction.

A typical PDF and different approximations are given in Fig. 9, where we plot the true PDF for resolution level $m = 1$ and direction $d =$ vertical together with three model functions: a Gaussian, a Laplacian, and an intermediate function, the so-called generalized Gaussian [2].

This generalized Gaussian law is given explicitly by

$$p_{m,d}(x) = a_{m,d} \exp(-|b_{m,d}x|^{r_{m,d}})$$

with

$$a_{m,d} = \frac{b_{m,d} r_{m,d}}{2\Gamma\left(\frac{1}{r_{m,d}}\right)} \quad \text{and} \quad b_{m,d} = \frac{1}{\sigma_{m,d}} \frac{\Gamma\left(\frac{3}{r_{m,d}}\right)^{1/2}}{\Gamma\left(\frac{1}{r_{m,d}}\right)^{1/2}} \quad (13)$$

where $\sigma_{m,d}$ is the standard deviation of the subimage (m, d) , and $\Gamma(\cdot)$ is the usual Gamma function.

The general formula (13) contains the other two examples as particular cases:

- $r_{m,d} = 2$ leads to the well-known Gaussian PDF;
- $r_{m,d} = 1$ leads to a Laplacian PDF.

The variance of this approximation model is set equal to the variance of the corresponding subimage. Thus the parameter $r_{m,d}$ is computed in order to match the real PDF using the well-known chi-squared test. In this case the optimum parameter was 0.7. Other experiments for other resolutions (except the lowest resolution) lead to very similar results.

We can see in Fig. 9 that the real PDF (scale $m = 1$ and vertical orientation) is closely approximated by a generalized Gaussian law with parameter $r_{1,v} = 0.7$.

B. Encoding of Wavelet Coefficients Using Vector Quantization

Different techniques involving vector or scalar quantization can be used to encode wavelet coefficients.

According to Shannon's rate distortion theory, better results are always obtained when vectors rather than sca-

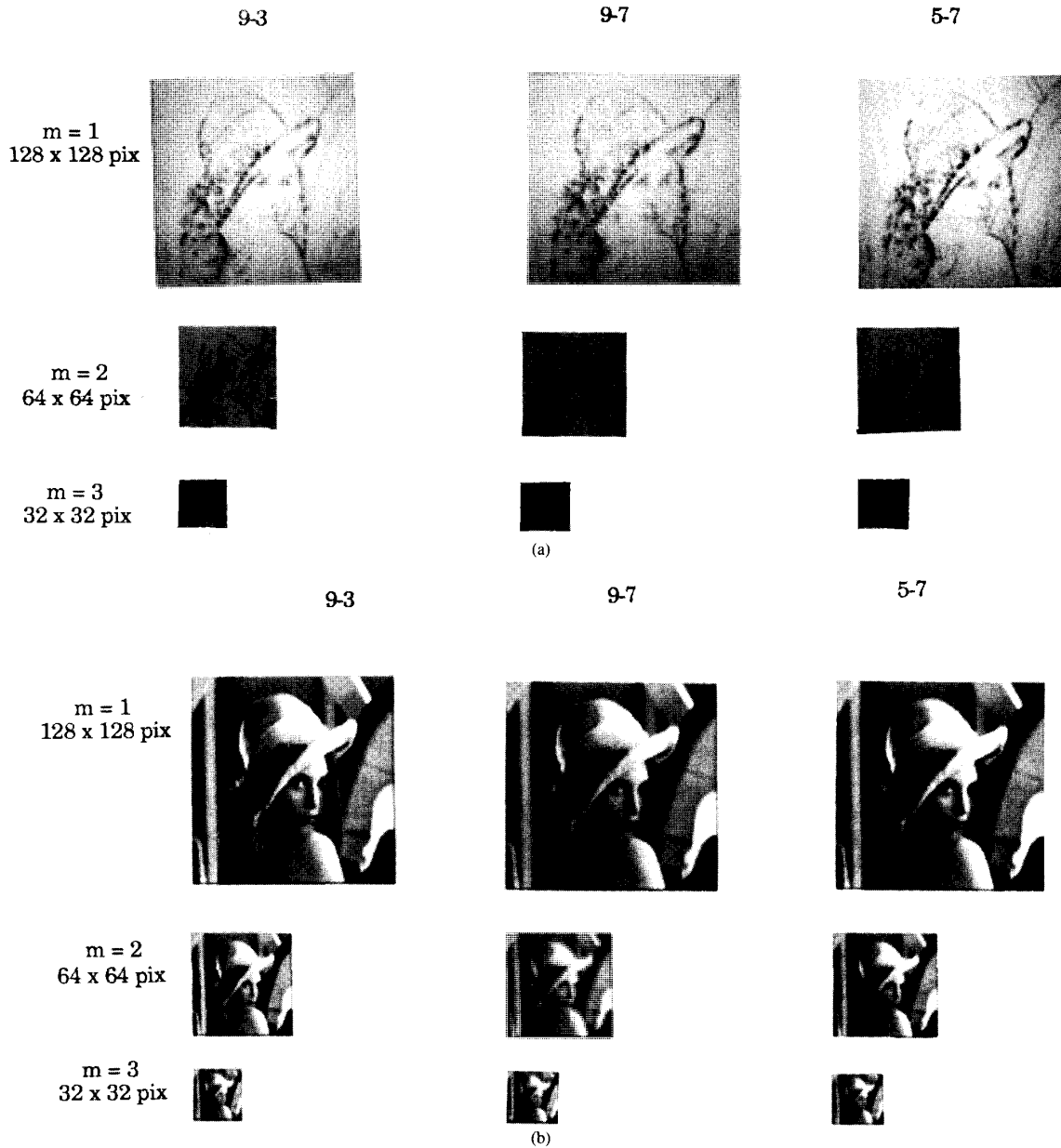


Fig. 8. Comparison among the different subimages. (a) Comparison among the normalized detail subimages. (b) Comparison among the low resolution level subimages.

lars are encoded. Therefore, the present application uses vector quantization.

1. Principle of Vector Quantization: Developed recently by Gersho and Gray (1980) [20], [21], vector quantization has proven to be a powerful tool for digital image compression [4], [29], [30], [32], [39]. The principle involves encoding a sequence of samples (vector) rather than encoding each sample individually. Encoding is performed by approximating the sequence to be coded by a vector belonging to a catalogue of shapes, usually known as a codebook.

The codebook is created and optimized using the well-known Linde-Buzo-Gray (LBG) [26] classification al-

gorithm with a mean squared error (MSE) criterion. This algorithm is designed to perform a classification based on a training set comprised of vectors belonging to different images; it converges iteratively toward a locally optimal codebook.

Each of the vectors in the codebook is indexed. At the encoding stage, the index of the vector in the codebook most closely describing (in terms of MSE criterion) the sample set to be encoded is selected to represent this set. Of course, in order to reconstruct the sample set, the decoder must have the same codebook as the coder.

The encoding/decoding scheme depicted in Fig. 10 was proposed in [29] and [30] for orthonormal wavelets.

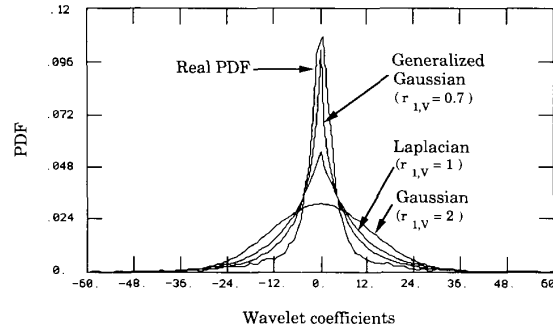


Fig. 9. Real PDF of subimage at scale $m = 1$ for vertical orientation, and its different approximations.

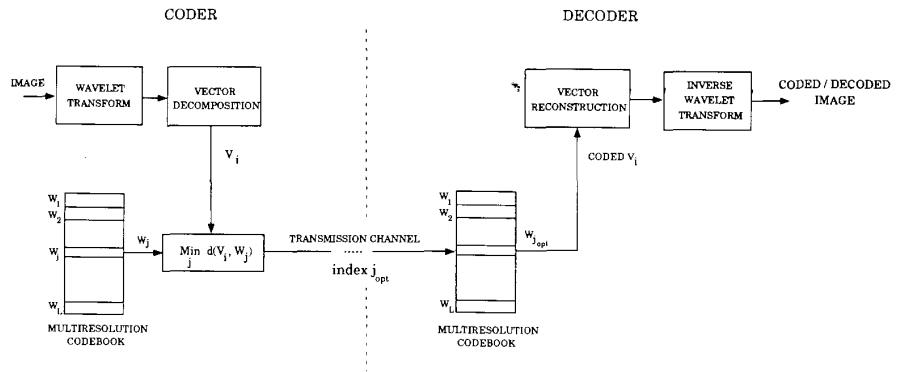


Fig. 10. Encoding/decoding scheme.

2) *Comparative Performances of Vector Quantization (VQ) and Scalar Quantization (SQ)*: According to [3], [13], [19], [43], [30] the asymptotic lower bound distortion gain obtained when VQ, rather than SQ, is applied to a subimage is expressed as:

$$G_{m,d}^{VQ} \geq \frac{2^{-c}}{(c + 1)A(k_{m,d}, c)} \times \frac{\left[\int [p_{m,d}(x)]^{1/(c+1)} dx \right]^{(c+1)}}{\left[\int [p_{m,d}(x)]^{k_{m,d}/(c+k_{m,d})} dx \right]^{(c+k_{m,d})}} \quad (14)$$

for a subimage corresponding to resolution m and direction d . $p_{m,d}(x)$ is the PDF of wavelet coefficients of the subimage with resolution m and direction d .

Here, the MSE criterion is used as a distortion measure ($c = 2$). The values of $A(k_{m,d}, 2)$ used are the upper bounds of the MSE computed and tabulated by Conway and Sloane for vector size $k_{m,d}$ [13]. This formula gives an indication of the minimum theoretical gain that can be obtained.

However, this approximation is valid only for small quantization errors, i.e., for a high bit rate $R_{m,d}$. Thus the gain $G_{m,d}^{VQ}$ only gives here an asymptotic indication.

In Fig. 11, the curves of $G_{m,d}^{VQ}$ are plotted as a function of the vector dimension $k_{m,d}$ for the Laplacian, Gaussian,

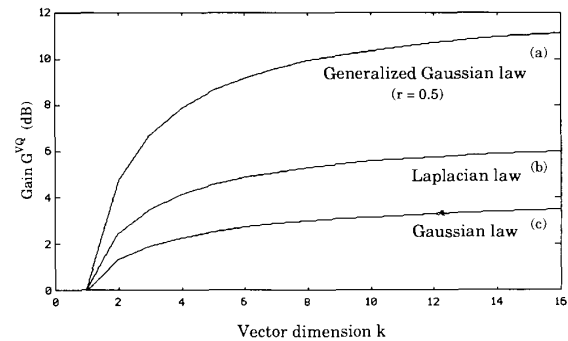


Fig. 11. Asymptotic lower bound distortion gain $G_{m,d}^{VQ} = \text{function}(k_{m,d})$.

and generalized Gaussian approximation laws, and for a subimage at scale $m = 1$ and vertical orientation. Experimental results are closely matched by the theoretical results for a generalized Gaussian law with $r_{m,d} = 0.7$ except for the lower subband. Therefore, all computations based on this approximation law show that, in each subband, VQ outperforms SQ (see Fig. 11).

In summary VQ performs better for coding wavelet coefficients.

3) *Generation of a Multiresolution Codebook*: The preceding paragraph explained why VQ outperforms other methods. Nonetheless, major problems are encountered in the VQ of images.

- It is impossible to create a universal codebook (efficient for each image to be encoded).

- The LBG algorithm smooths high frequencies (loss of resolution).

- There is a trade-off between low distortion and high compression rate (computational cost).

- It is not easy to take into account the properties of the human visual system [28], [33].

The use of the wavelet transform (i.e., multiresolution) is one way of overcoming these different problems.

The wavelet decomposition of an image enables the generation of a codebook containing two-dimensional vectors for *each resolution level and preferential direction* (horizontal, vertical, and diagonal). Each of these subcodebooks (see Fig. 12) is generated using the LBG algorithm.

- The training set is comprised of vectors belonging to different images corresponding to the resolution and orientation under consideration.

- The initial codebook is generated by splitting the centroid (center of gravity) of this training set [21].

A multiresolution codebook can thus be obtained by assembling all of these resulting subcodebooks. Each subcodebook has a low distortion level and contains few words, which clearly facilitates the search for the best coding vector; the coding computational load is reduced, because only the appropriate subcodebook (resolution direction) of the multiresolution codebook is checked for each input vector. In addition, the quality of the coded image is better. The multiresolution codebook is depicted in Fig. 12.

Global codebook design has drawbacks in that it results in edge smoothing while the proposed method preserves edges. In fact, each subcodebook contains the shape of the wavelet coefficients which are most highly representative in terms of the MSE criterion.

Since the spatial and frequency aspects of the image are taken into account in the wavelet decomposition, the classification and search during the encoding of a subimage vector can be achieved using a simple criterion such as least mean squares. This frees us from using distortion measurements such as weighted least mean squares or other measurements involving perceptual factors. These algorithms are indeed costly in computation time.

C. Optimal Bit Allocation

Multiresolution exploits the eye's masking effects, and therefore, enables us to refine and select the type of coding according to the resolution level and the contour orientation. Although a flat noise shape minimizes the MSE criterion, it is generally not optimal for a subjective quality of image. To apply *noise shaping* across the VQ subimages, we define a total weighted MSE distortion $D_T^*(R_T)$ ((17)) for a total bit rate R_T ((18)).

Let us define $D_{m,d}(R_{m,d})$ the average distortion in the coding of the subimage (m, d) for $R_{m,d}$ bits per pixel:

$$D_{m,d}(R_{m,d}) = E(|x - q(x)|^c) = d(x, q(x)) \quad c \geq 1 \quad (15)$$

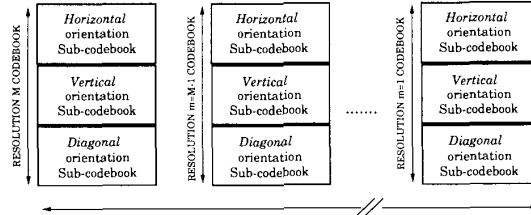


Fig. 12. Multiresolution codebook.

for all coefficients x belonging to the subimage, $q(x)$ being the quantization of x .

Total distortion of the image for a total rate of R_T bits per pixel is then given by:

$$D_T(R_T) = \frac{1}{2^{2M}} D_M^{SQ}(R_M^{SQ}) + \sum_{m=1}^M \frac{1}{2^{2m}} \sum_{d=1}^3 D_{m,d}(R_{m,d}) \quad (16)$$

where $D_M^{SQ}(R_M^{SQ})$ corresponds to the distortion in the subimage of lowest resolution M (texture subimage).

The problem of finding an optimal bit assignment (in bits per pixel) for each subimage vector quantizer is then formulated as:

$$\text{Min}_{R_{m,d}} \left[D_T^*(R_T) = \frac{1}{2^{2M}} D_M^{SQ}(R_M^{SQ}) + \sum_{m=1}^M \frac{1}{2^{2m}} \sum_{d=1}^3 D_{m,d}(R_{m,d}) \times B_{m,d} \right] \quad (17)$$

$$\text{subject to: } R_T = \frac{1}{2^{2M}} R_M^{SQ} + \sum_{m=1}^M \frac{1}{2^{2m}} \sum_{d=1}^3 R_{m,d} \quad (18)$$

where R_M^{SQ} corresponds to the bit allocation, in bits per pixel, of lowest resolution M subimage.

Assignment of the weights is based on the fact that the human eye is not equally sensitive to signals at all spatial frequencies. On the basis of contrast sensitivity data collected by Campbell and Robson [10], and to obtain a controlled degree of noise shaping across the subimages, we consider a function $B_{m,d}$ such that:

$$B_{m,d} = \gamma^m \log(\sigma_{m,d}^{2\beta_{m,d}}) \quad (19)$$

where $\sigma_{m,d}$ is the standard deviation corresponding to subimage (m, d) and the values of γ and $\beta_{m,d}$ are chosen experimentally in order to match human vision.

$D_T^*(R_T)$ is the total weighted encoding distortion function, and M is the lowest resolution considered.

The expression of $D_{m,d}(R_{m,d})$ is given by [19]

$$D_{m,d}(R_{m,d}) = 2^{-cR_{m,d}} \times \alpha_{m,d}(p, c), \quad c \geq 1$$

with

$$\alpha_{m,d}(p, c) = A(k_{m,d}, c) \times \left[\int [p_{m,d}(x)]^{k_{m,d}/(c+k_{m,d})} dx \right]^{(c+k_{m,d})} \quad (20)$$

This minimization problem can be solved by using Lagrangian multipliers. Using this technique, we must solve the following equation:

$$\frac{\partial}{\partial R_{m,d}} \left[D_T^*(R_T) - \lambda \left(R_T - \frac{1}{2^{2M}} R_M^{SQ} - \sum_{m=1}^M \frac{1}{2^{2m}} \sum_{d=1}^3 R_{m,d} \right) \right] = 0 \quad (21)$$

where λ is a Lagrangian multiplier.

Using (17) and (20), this equation becomes:

$$\begin{aligned} \frac{\partial}{\partial R_{m,d}} \left[\frac{1}{2^{2m}} D_M^{SQ}(R_M^{SQ}) \right. \\ \left. + \sum_{m=1}^M \frac{1}{2^{2m}} \sum_{d=1}^3 (2^{-cR_{m,d}} \alpha_{m,d}(p, c) B_{m,d}) \right. \\ \left. - \lambda \left(R_T - \frac{1}{2^{2M}} R_M^{SQ} - \sum_{m=1}^M \frac{1}{2^{2m}} \sum_{d=1}^3 R_{m,d} \right) \right] = 0. \end{aligned} \quad (22)$$

Taking the partial derivative with respect to $R_{m,d}$ yields an expression for $R_{m,d}$ in terms of λ :

$$R_{m,d} = \frac{1}{c} \log_2 \left[\frac{(c \ln 2) \alpha_{m,d}(p, c) B_{m,d}}{\lambda} \right]. \quad (23)$$

By substituting (23) into the constraint (18) of the minimization problem we obtain an expression of the Lagrangian multiplier λ

$$\lambda = c \ln 2 \left[2^{-c(R_T - (1/4^M) R_M^{SQ})} \prod_{m=1}^M \prod_{d=1}^3 [\alpha_{m,d}(p, c) B_{m,d}]^{1/4^m} \right]^{4^{M/4^M - 1}}. \quad (24)$$

Finally, substituting λ into (23) results in an expression of the optimal bit assignment $R_{m,d_{opt}}$ (in bits per pixel (bpp)) to the vector quantizer of subimage (m, d) :

$$R_{m,d_{opt}} = \frac{4^M R_T - R_M^{SQ}}{4^M - 1} + \frac{1}{c} \log_2 \left[\frac{\alpha_{m,d}(p, c) B_{m,d}}{\left[\prod_{m'=1}^M \prod_{d'=1}^3 [\alpha_{m',d'}(p, c) B_{m',d'}]^{1/4^{m'}} \right]^{4^{M/4^M - 1}}} \right]. \quad (25)$$

This expression requires the knowledge of the subimage's PDF's.

The optimal distortion of the quantizer, $D_{T_{opt}}^*(R_T)$, is then computed by combining (25) and (17). We find:

$$\begin{aligned} D_{T_{opt}}^*(R_T) = \frac{1}{2^{2M}} D_M^{SQ}(R_M^{SQ}) + \frac{4^M - 1}{4^M} 2^{-c(4^M R_T - R_M^{SQ})/4^M - 1} \\ \cdot \left[\prod_{m=1}^M \sum_{d=1}^3 [\alpha_{m,d}(p, c) B_{m,d}]^{1/4^m} \right]^{4^{M/4^M - 1}}. \end{aligned} \quad (26)$$

Finally, bit allocation which is a function of the image will be transmitted as side information requiring only a few bits.

IV. EXPERIMENTAL RESULTS

The images used are sampled 256 by 256 black and white images. The intensity of each pixel is coded on 256 grey levels (8 bpp).

The numerical evaluation of the coder's performance is achieved by computing the peak signal-to-noise ratio (PSNR) between the original image and the coded image.

For each coded image, we can use a variable length code. We also give the corresponding \mathcal{R}_T if an optimal entropy coding was performed, defined as follows.

To the L codewords $w_j; j = 1, 2, \dots, L$ of the vector quantizer corresponds to L regions (clusters) of \mathbb{R}^k , $\mathcal{P}_j; j = 1, 2, \dots, L$. The j th region is defined by

$$\mathcal{P}_j = \{x \in \mathbb{R}^k / Q(x) = w_j\}$$

and represents the subset of vectors of \mathbb{R}^k which are well matched by the codeword w_j of the codebook.

Thus for each resolution and direction, we can introduce the average information of the codebook, called the entropy measure:

$$\mathcal{R}_{m,d} = -\frac{1}{k_{m,d}} \times \sum_{j=1}^L p(w_j) \log_2 p(w_j) \text{ bpp}$$

where $p(w_j)$ is the probability of selecting the source vector w_j , belonging to the codebook at scale m and corresponding to the orientation d , during the coding of the image (m, d) .

Then, as in (18), \mathcal{R}_T is the sum of the estimated entropy in each subimage as follows:

$$\mathcal{R}_T = \frac{1}{2^{2M}} \mathcal{R}_M^{SQ} + \sum_{m=1}^M \frac{1}{2^{2m}} \sum_{d=1}^3 \mathcal{R}_{m,d} \text{ bpp}.$$

The vector quantizer used is a *full search* quantizer, i.e., during the coding, all of the vectors in the subcodebook corresponding to the resolution and direction to be encoded are searched. The selection criterion used is the MSE criterion.

A. Comparison Between the Different Wavelets

In the following, we present results obtained with the Lena image (image within the training set) for a real bit rate of 1 bpp and using the three different filters proposed in Section II-B. (Fig. 13 corresponds to filters 9-3 presented in example 1, Fig. 14 corresponds to filters 9-7 presented in example 2, and Fig. 15 corresponds to filters 5-7 presented in example 3.) Here, the Lena image is taken as part of the training set in order to minimize the effects of quantization noise: this enables the influence of the filters to be taken into account.

For a given set of filters, separate codebooks are trained for each resolution-orientation subimage, and bit alloca-



Fig. 13. Filters no. 1-3, PSNR = 31.82 dB, $R_T = 0.80$ bpp.



Fig. 15. Filters no. 3, 5-7, PSNR = 31.46 dB, $R_T = 0.80$ bpp.



Fig. 14. Filters no. 2, 9-7, PSNR = 32.10 dB, $R_T = 0.78$ bpp.



Fig. 16. Original 256 by 256 Lena, 8 bpp.

tion is carried out according to (25). For the Lena image, the bit assignment is represented in Fig. 17. Resolution 1 (diagonal orientation) is discarded. Resolution 1 (horizontal and vertical orientations) and resolution 2 (diagonal orientation) are coded using 256-vector codebooks (codeword size 4 by 4) resulting in a 0.5-b/pixel rate, while resolution 2 (horizontal and vertical orientations) is coded at a 2-b/pixel rate using 256-vector codebooks

(codeword size 2 by 2). Finally, the lowest resolution is coded at 8 b/pixel.

B. Results as a Function of Regularity and Vanishing Moments

In Section II-B, we mentioned our belief that both the regularity of the reconstruction wavelet $\tilde{\psi}$ and the number

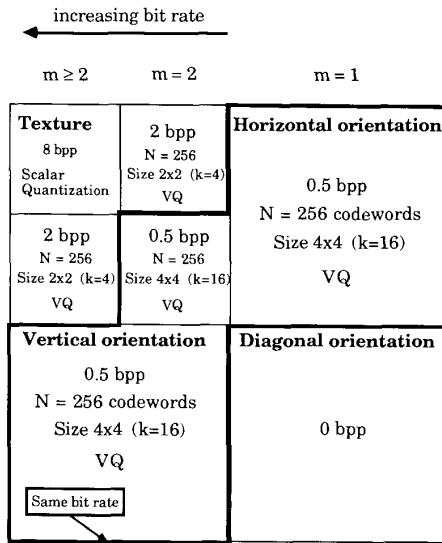


Fig. 17. Subimages bit rate allocation: example of a bit allocation for a total bit rate of 1 bpp and for the 256 by 256 Lena image.

of vanishing moments of the analyzing wavelet ψ are important in applications. To illustrate this we carried out the following experiments. For a given pair, h, \tilde{h} , we analyzed the same image twice: once as described above, and a second time after exchanging the roles of the filters h and \tilde{h} .

The filter pairs in example 2 both have the same number of vanishing moments, $k = \tilde{k} = 4$. However, $\tilde{\psi}$ is considerably more regular than ψ (see Fig. 3). With this filter pair, our experiment on the Lena image led to a PSNR of 32.10 dB in the first case, and to a PSNR of 31.51 dB if the roles of h and \tilde{h} are inverted. The case where the reconstruction wavelet has the highest regularity, therefore, performs best.

In example 1 the functions ψ and $\tilde{\psi}$ have comparable regularity: both are continuous and neither has a continuous derivative. In fact $\tilde{\psi}$ is a bit more regular than ψ : $\tilde{\psi}$ is differentiable almost everywhere, and is Hölder continuous with exponent 1, while ψ is Hölder continuous with the exponent only at 0.83. On the other hand, ψ has 2 vanishing moments, while $\tilde{\psi}$ has 4 ($k = 4, \tilde{k} = 2$). The same experiment, again with the Lena image, now leads to a PSNR of 31.82 dB if h, \tilde{h} are taken as in Table I, and to a PSNR of 31.13 dB when the roles of h and \tilde{h} are reversed. The situation where $\tilde{\psi}$ is most regular but ψ has fewer vanishing moments, therefore, performs better (gain of 0.69 dB) than the case where ψ has more vanishing moments but $\tilde{\psi}$ is less regular. This seems to suggest that the regularity of $\tilde{\psi}$ has a larger effect than the number of vanishing moments of ψ . However, in this example the difference in overall regularity, as measured by the differences between Hölder exponents, is much smaller here

than in example 2 (0.17 as compared to 0.63 in example 2), and it seems hard to explain how this smaller difference in Hölder exponent could account for a comparable gain in PSNR. In fact, the Hölder exponent is not a very good measure for the regularity of $\tilde{\psi}$ in this case: it is completely determined by the discontinuity of the derivative of $\tilde{\psi}$ in only a few points, and it is insensitive to the fact that $\tilde{\psi}$ is infinitely differentiable in all other points. If this is taken into account, then $\tilde{\psi}$ looks much more regular than ψ (the Hölder exponent of which is determined by its behavior near a dense set of points), which might explain the gain in PSNR.

We conclude from all this that: 1) for the same number of vanishing moments for ψ , the scheme with most regular $\tilde{\psi}$ is likely to perform best; and 2) increasing the regularity of $\tilde{\psi}$, even at the expense of the number of vanishing moments for ψ , may lead to better results.

Based on theoretical arguments (Taylor expansions) and results from numerical analysis [8], we also expect: 3) for comparable regularity of $\tilde{\psi}$, the scheme with largest vanishing moments for ψ is likely to perform best.

C. Comparison with Other Coders

If the PSNR is chosen as a criterion of comparison, these results are close to those obtained by Woods and O’Neil [42] and Westerink *et al.* [40]. However, in their subband coding algorithm, they use 32-taps Johnston filters, while only 9 or 7 taps are necessary for our method. According to Westerink’s results in [41], the PSNR decreases by about 2 dB when using 8-taps Johnston filters. However, some others new QMF designs can also lead to good results with about 9 taps for image coding [1].

In this section, we present both numerical and qualitative comparison between our coding scheme and other previously published results. Since the most popular image in the recent literature has been the 512 by 512 Lena image, the comparison is made using this image taken outside the training set.

Among the different methods published, we consider the three following well-known methods: Ho and Gersho obtained a 30.93-dB PSNR at 0.36 bpp, result using “variable-rate multi stage VQ” [23]. Riskin and Gray improved on the full search VQ (PSNR = 29.29 dB, 0.32 bpp) using pruned tree structured VQ (PSNR = 30.92 dB, 0.32 bpp) [34]. High PSNR values were obtained by Woods and Cohen using entropy coded and predictive VQ (PSNR = 32.5 dB, 0.45 bpp) [11].

Our aim is not to optimize the PSNR but rather a weighted function of the MSE in order to match human vision. We give two examples at low bit rate using wavelet VQ.

Our initial result at 0.37 bpp presented Fig. 18 with a 30.85-dB PSNR is very close to those of Ho and Gersho [23] and Riskin *et al.* [34]. The perceptual quality of our coded images is better than indicated by the PSNR value



Fig. 18. 512 by 512 Lena image. Filters no. 2 9-7, PSNR = 30.85 dB, $\rho_T = 0.37$ bpp.



Fig. 19. 512 by 512 Lena image. Filters no. 2 9-7, PSNR = 29.11 dB, $\rho_T = 0.21$ bpp.

mainly due to the regularity of the wavelet and the bit allocation. These images do not suffer from the blocking effects obtained when using VQ in the spatial domain. No ringing effects can be observed.

The second result at 0.21 bpp presented in Fig. 19 with a 29.11-dB PSNR shows that a very low bit rate can be achieved with our method, without severe degradation.

Our method using a new class of filters derived from

wavelet theory using full search VQ can be improved by any of the three above-mentioned methods.

In fact the LBG clustering algorithm is a very simple algorithm but not optimal for variable length code. The PSNR of the method could be improved by about 3 dB, for example, using ECVQ [34] but CPU time becomes prohibitively expensive.

D. Progressive Transmission Scheme

The main objective of progressive transmission is to allow the receiver to recognize a picture as quickly as possible at minimum cost, by sending a low resolution level picture first. Then, it can be decided to receive further picture details or to abort the transmission. Further details of the picture are obtained by sequentially receiving the encoded wavelet coefficients at different resolution levels and directions.

Following the example of [40], we will display each picture level during the progressive transmission with a size that matches the resolution of that particular level.

To test the efficiency of the vector quantizer, the image to be coded is taken outside the training set.

Fig. 20 represents 5 stages in the progressive transmission of a 256 by 256 image using filters 9-7 given in example 2. According to the bit allocation procedure (Section III-C) with a generalized Gaussian PDF approximation law, only the wavelet coefficients corresponding to the $m = 1$ and $m = 2$ high resolution levels are vector quantized, while the low level subimages ($m \geq 2$) are scalar quantized.

V. CONCLUSION

This paper describes a new image coding scheme combining the wavelet transform and VQ.

A new family of filters has been derived from the wavelet theory. We have shown the importance of regularity and vanishing moments for image coding. Furthermore, these filters require few taps, unlike standard QMF methods.

The wavelet transform used here attempts to exploit the masking effect of the human eye, yielding encouraging results. Indeed, the proposed method enables high compression bit rates while maintaining good visual quality through the use of bit allocation in the subimages. The blocking effects seen when spatial VQ is performed are avoided.

This method is well adapted to progressive transmission as well as very low bit rate compression. Furthermore, using a simple full-search VQ provides good results, comparable to the best results published currently.

Further research should include some new derivation such as entropy constraint and predictive VQ. We would improve this coding scheme, if we accept a heavier computational load.

<i>Resolution</i>	<i>m=4</i>	<i>m=3</i>	<i>m=2</i>	<i>m=1</i>
Size	16 × 16 pix	32 × 32 pix	64 × 64 pix	128 × 128 pix
R_T	0.031 bpp	0.125 bpp	0.5 bpp	0.781 bpp
\mathcal{R}_T	0.0264 bpp	0.0919 bpp	0.3354 bpp	0.5039 bpp



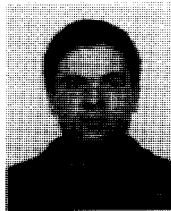
Resolution m=0
256 × 256 pix
 $R_T = 1 \text{ bpp}$ $\mathcal{R}_T = 0.6297 \text{ bpp}$
 PSNR = 31.28 dB

Fig. 20. Progressive transmission-filters no. 2 9-7.

REFERENCES

- [1] E. H. Adelson and E. Simoncelli, "Non-separable extensions of quadrature mirror filters to multiple dimensions," *Proc. IEEE*, vol. 78, Apr. 1990.
- [2] M. Abramowitz, I. A. Stegun, *Handbook of Mathematical Functions*. New York: Dover, 1965.
- [3] V. R. Algazi, "Useful approximation to optimum quantization," *IEEE Trans. Commun.*, vol. COM-14, pp. 297-301, June 1966.
- [4] M. Antonini, M. Barlaud, P. Mathieu, and I. Daubechies, "Image

- coding using vector quantization in the wavelet transform domain," in *Proc. IEEE ICASSP*, April 1990, pp. 2297-2300.
- [5] M. Barlaud, L. Blanc-Féraud, P. Mathieu, J. Menez, and M. Antonini, "2D linear predictive image coding with vector quantization," in *Proc. EUSIPCO*, Grenoble, France, Sept. 5-8, 1988, pp. 1637-1640.
- [6] M. Barlaud, P. Mathieu, and M. Antonini, "Wavelet transform image coding using vector quantization," presented at 6th Workshop on MDSP, Monterey, CA, Sept. 1989.
- [7] G. Battle, "A block spin construction of wavelets. Part I Lemarié functions," *Comm. Math. Phys.*, vol. 110, pp. 601-615, 1987.
- [8] G. Beylkin, R. Coifman, and V. Rokhlin, "Fast wavelet transforms and numerical analysis. I," to be published.
- [9] P. Burt and E. Adelson, "The Laplacian pyramid as a compact image code," *IEEE Trans. Commun.*, vol. 31, pp. 482-540, 1983.
- [10] F. W. Campbell and J. G. Robson, "Application of Fourier analysis to the visibility of gratings," *J. Phys.*, vol. 197, pp. 551-566, 1968.
- [11] R. A. Cohen and J. W. Woods, "Sliding block entropy coding of images," in *Proc. IEEE ICASSP*, Glasgow, Scotland, May 23-26, 1989, pp. 1731-1733.
- [12] A. Cohen, I. Daubechies, and J. C. Feauveau, "Biorthogonal bases of compactly supported wavelets," AT&T Bell Lab., Tech. Rep., TM 11217-900529-07, 1990.
- [13] J. H. Conway and N. J. A. Sloane, "A lower bound on the average error of vector quantizers," *IEEE Trans. Inform. Theory*, vol. IT-31, pp. 106-109, Jan. 1985.
- [14] I. Daubechies, A. Grossman, and Y. Meyer, "Painless nonorthogonal expansions," *J. Math. Phys.*, vol. 27, pp. 1271-1283, 1986.
- [15] I. Daubechies, "The wavelet transform, time-frequency localization and signal analysis," to be published.
- [16] —, "Orthonormal bases of compactly supported wavelets," *Comm. Pure Appl. Math.*, vol. 41, pp. 909-996, 1988.
- [17] —, "Orthonormal bases of compactly supported wavelets. II. Variations on a theme," AT&T Bell Lab., Tech. Rep. TM 11217-891116-17, 1990.
- [18] J. C. Feauveau, "Analyse multirésolution par ondelettes non orthogonales et bancs de filtres numériques," Ph.D. dissertation, Univ. Paris Sud, France, Jan. 1990.
- [19] A. Gersho, "Asymptotically optimal block quantization," *IEEE Trans. Inform. Theory*, vol. IT-25, July 1979.
- [20] —, "On the structure of vector quantizers," *IEEE Trans. Inform. Theory*, vol. IT-28, Mar. 1982.
- [21] R. M. Gray, "Vector quantization," *IEEE ASSP Mag.*, pp. 4-29, Apr. 1984.
- [22] A. Grossman and J. Morlet, "Decomposition of hardy functions into square integrable wavelets of constant shape," *SIAM J. Math. Anal.*, vol. 15, pp. 723-736, 1984.
- [23] Y. Ho and A. Gersho, "Variable-rate multi-stage vector quantization for image coding," in *Proc. IEEE ICASSP*, New York, Apr. 1988.
- [24] P. G. Lemarié, "Une nouvelle base d'ondelettes de $L^2(\mathbb{R})$," *J. Math. Pures et Appl.*, vol. 67, pp. 227-238, 1988.
- [25] P. G. Lemarié and Y. Meyer, "Ondelettes et bases hilbertiennes," *Rev. Mat. Iberoamericana*, vol. 2, pp. 1-18, 1986.
- [26] Y. Linde, A. Buzo, and R. M. Gray, "An algorithm for vector quantizer design," *IEEE Trans. Commun.*, vol. COM-28, pp. 84-95, Jan. 1980.
- [27] S. Mallat, "A theory for multiresolution signal decomposition: The wavelet representation," *IEEE Trans. Pattern Anal. Mach. Intel.*, vol. 11, July 89.
- [28] D. Marr, *Vision*. New York: Freeman, 1982.
- [29] P. Mathieu, M. Barlaud, and M. Antonini, "Compression d'Images par transformée en ondelette," *12ième colloque GRETSI, Juan les Pins*, June 12-16, 1989.
- [30] P. Mathieu, M. Barlaud, and M. Antonini, "Compression d'Image par transformée en ondelette et quantification vectorielle," *Traitement du Signal*, vol. 7, no. 2, 1990.
- [31] Y. Meyer, "Principe d'incertitude, bases hilbertiennes et algèbres d'opérateurs," *Seminaire Bourbaki*, no. 662, 1985-1986.
- [32] N. M. Nasrabadi and R. A. King, "Image coding using vector quantization: A review," *IEEE Trans. Commun.*, vol. 36, Aug. 1988.
- [33] W. K. Pratt, *Digital Image Processing*. New York: Wiley, 1978.
- [34] E. Riskin, E. M. Daly, and R. M. Gray, "Pruned tree-structured vector quantization in image coding," in *Proc. IEEE ICASSP*, Glasgow, Scotland, May 1989, pp. 1735-1738.
- [35] M. J. Smith and D. P. Barnwell, "Exact reconstruction for tree-structured subband coders," *IEEE Trans. Acoust., Speech, Signal Proc.*, vol. ASSP-34, pp. 434-441, 1986.
- [36] J. O. Stromberg, "A modified haar system and higher order spline systems," in *Conf. in Harmonic Analysis in Honor of Antoni Zygmund*, Vol. II, pp. 475-493.
- [37] M. Vetterli, "Splitting a signal into subsampled channels allowing perfect reconstruction," in *Proc. IASTED Conf. Appl. Signal Processing Digital Filtering*, Paris, France, June 1985.
- [38] M. Vetterli and C. Herley, "Wavelets and filter banks: Relationships and new results," in *Proc. IEEE ICASSP*, Albuquerque, Apr. 1990.
- [39] P. H. Westerink, D. E. Boekee, J. Biemond, and J. W. Woods, "Subband coding of image using vector quantization," *IEEE Trans. Commun.*, vol. 36, pp. 713-719, 1988.
- [40] P. H. Westerink, J. Biemond, and D. E. Boekee, "Progressive transmission of images using subband coding," in *Proc. IEEE ICASSP*, 1989, pp. 1811-1814.
- [41] P. H. Westerink, "Subband coding of images," Ph.D. dissertation Delft Univ., 1989.
- [42] J. W. Woods and S. D. O'Neil, "Subband coding of images," *IEEE Trans. Acoust., Speech, Signal Proc.*, vol. ASSP-34, Oct. 1986.
- [43] P. Zador, "Asymptotic quantization error of continuous signals and their quantization dimension," *IEEE Trans. Inform. Theory*, vol. IT-28, pp. 139-149, 1982.



Marc Antonini was born in France on August 29, 1965. He received the DEA degree in signal processing in 1988 from the University of Nice-Sophia Antipolis, France, and the Ph.D. degree from the Laboratory of Signaux et Systèmes, URA 13S, CNRS and the University of Nice-Sophia Antipolis in 1991.

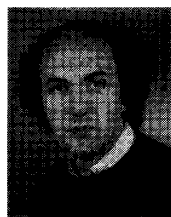
His research interests include multidimensional image processing, wavelet analysis, and image coding.



Michel Barlaud (M'88) was born in France on November 24, 1945. He received the "Doctorat d'Etat" degree from University of Paris XII.

He is currently a Professor and a member of the Laboratory of Signaux et Systèmes, URA 13S both from CNRS and University of Nice-Sophia Antipolis. After some work on non-stationary signal processing, his research interests move towards multidimensional image processing, wavelet analysis, image coding, inverse problems, image restoration, and edge detection.

Dr. Barlaud is member of the IEEE-ASSP MDSP committee.



Pierre Mathieu was born in Alger on May 10, 1956. He received the Ingenieur ENSEEIHT and Ph.D. degrees from INP Toulouse.

He is currently Maître de Conférences in the Laboratory of Signaux et Systèmes, URA 13S both from CNRS and University of Nice-Sophia Antipolis. His research interests include multidimensional image processing, wavelet analysis, image coding, and image restoration.



Ingrid Daubechies (M'89) received the B.S. and Ph.D. degrees from the Vrije Universiteit Brussel, Belgium in 1975 and 1980, both in physics.

She is currently a Member of Technical Staff in the Mathematics Center of AT&T Bell Laboratories, Murray Hill, NJ. Her current research interests include mathematical problems in connection with signal analysis, in particular applications of time-frequency representations.



US005321520A

United States Patent [19]

[11] Patent Number: 5,321,520

Inga et al.

[45] Date of Patent: Jun. 14, 1994

- [54] **AUTOMATED HIGH DEFINITION/RESOLUTION IMAGE STORAGE, RETRIEVAL AND TRANSMISSION SYSTEM**
- [75] Inventors: **Jorge J. Inga; Thomas V. Saliga**, both of Tampa, Fla.
- [73] Assignee: **Automated Medical Access Corporation**, Tampa, Fla.
- [21] Appl. No.: **915,298**
- [22] Filed: **Jul. 20, 1992**
- [51] Int. Cl.⁵ **H04N 1/21; H04N 1/41**
- [52] U.S. Cl. **358/403; 358/426; 358/428**
- [58] Field of Search **358/403, 426, 261.1, 358/444, 428; 395/114**

5,170,266 12/1992 Marsh et al. 358/426

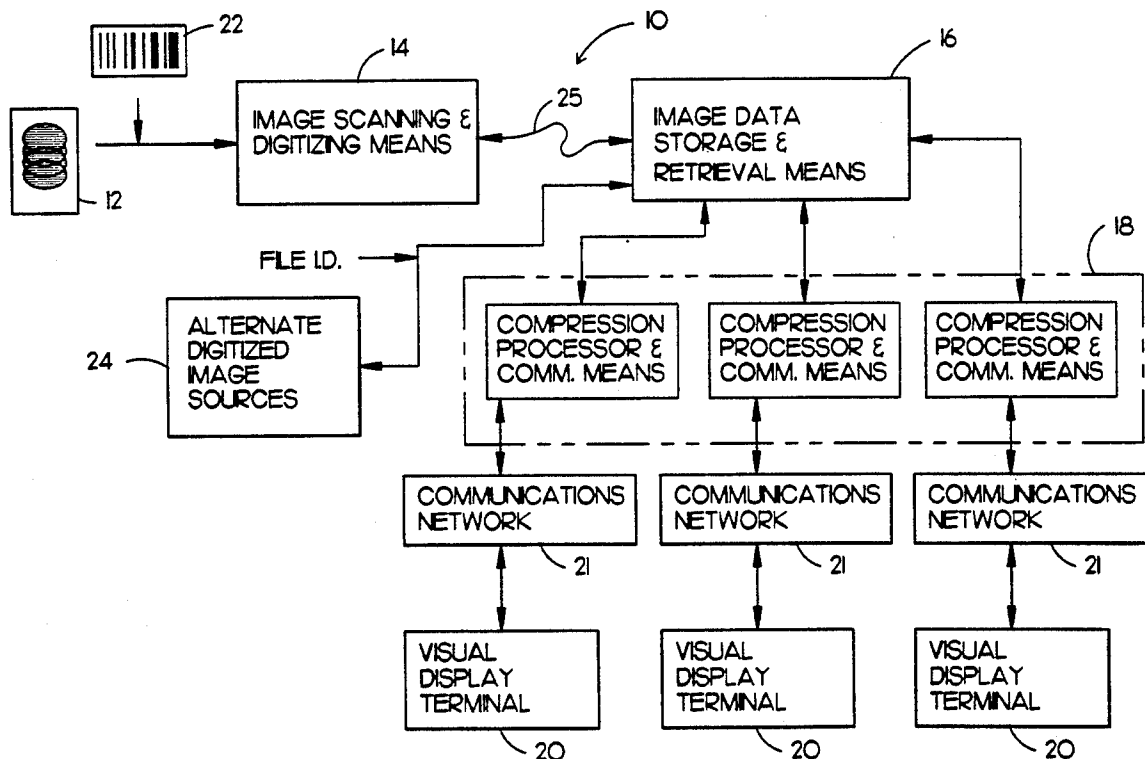
Primary Examiner—Edward L. Coles, Sr.
Assistant Examiner—Scott A. Rogers
Attorney, Agent, or Firm—David Kiewit

[57] **ABSTRACT**

An automated high definition/resolution image storage, retrieval and transmission system for use with medical X-ray film or other documents to provide simultaneous automated access to a common data base by a plurality of remote subscribers upon request, the automated high definition/resolution image storage, retrieval and transmission system comprising an image scanning and digitizing subsystem to scan and digitize visual image information from an image film or the like; an image data storage and retrieval subsystem to receive and store the digitized information and to selectively provide the digitized information upon request from a remote site, a telecommunication subsystem to selectively transmit the requested digitized information from the image data storage and retrieval subsystem to the requesting remote visual display terminal for conversion to a visual image at the remote site to visually display the requested information from the image data storage and retrieval subsystem.

- [56] **References Cited**
- U.S. PATENT DOCUMENTS**
- | | | | |
|-----------|---------|-------------------------|-----------|
| 4,719,514 | 1/1988 | Kurahayashi et al. | 358/404 |
| 4,768,099 | 8/1988 | Mukai | 358/426 |
| 4,933,025 | 2/1991 | Vesel et al. | 370/94.1 |
| 4,958,283 | 9/1990 | Tawara et al. | 364/413.3 |
| 5,068,745 | 11/1991 | Shimura | 358/426 |
| 5,111,044 | 5/1992 | Agano | 250/327.2 |
| 5,111,306 | 5/1992 | Kanno et al. | 358/426 |

12 Claims, 9 Drawing Sheets



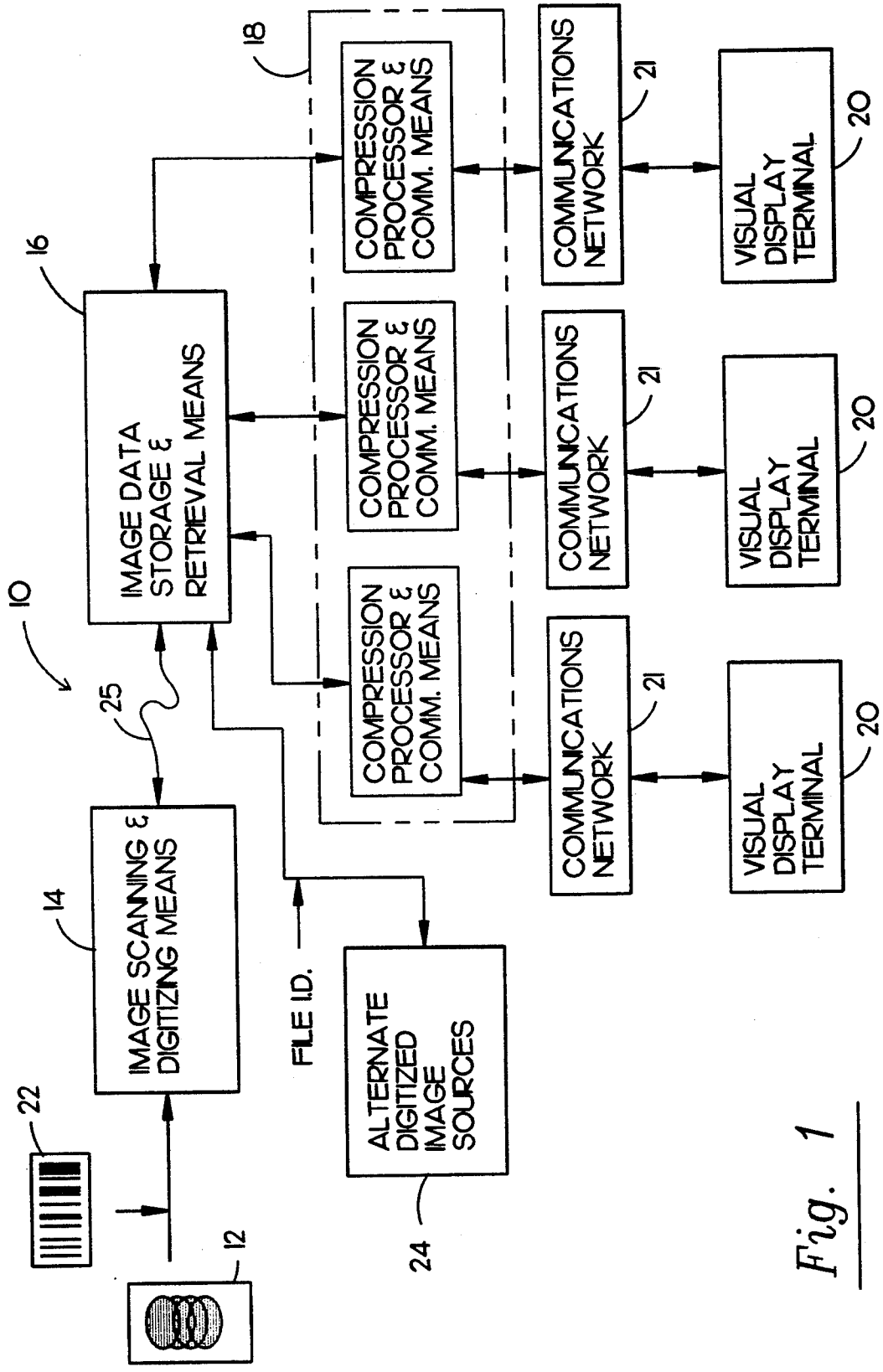


Fig. 1

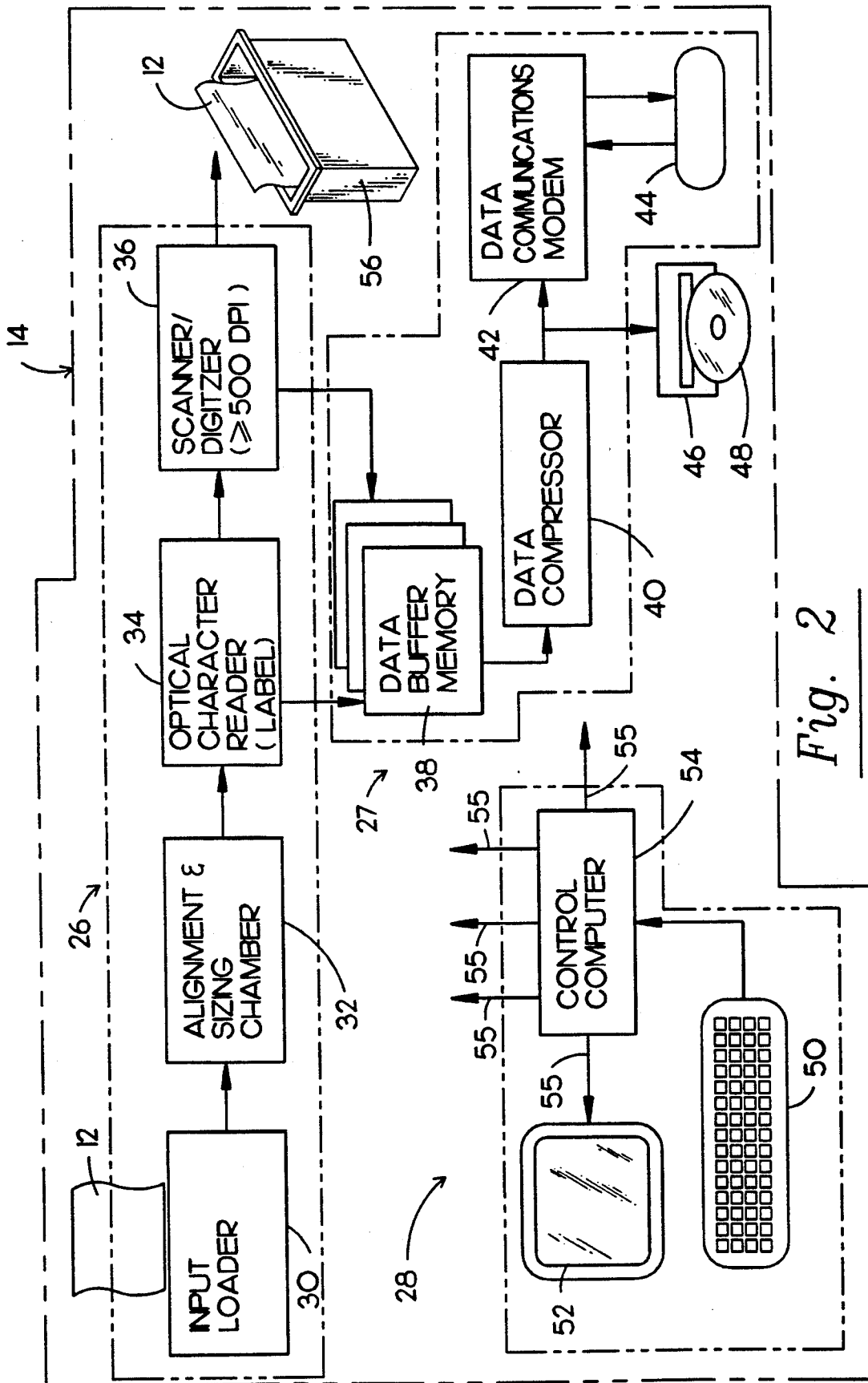


Fig. 2

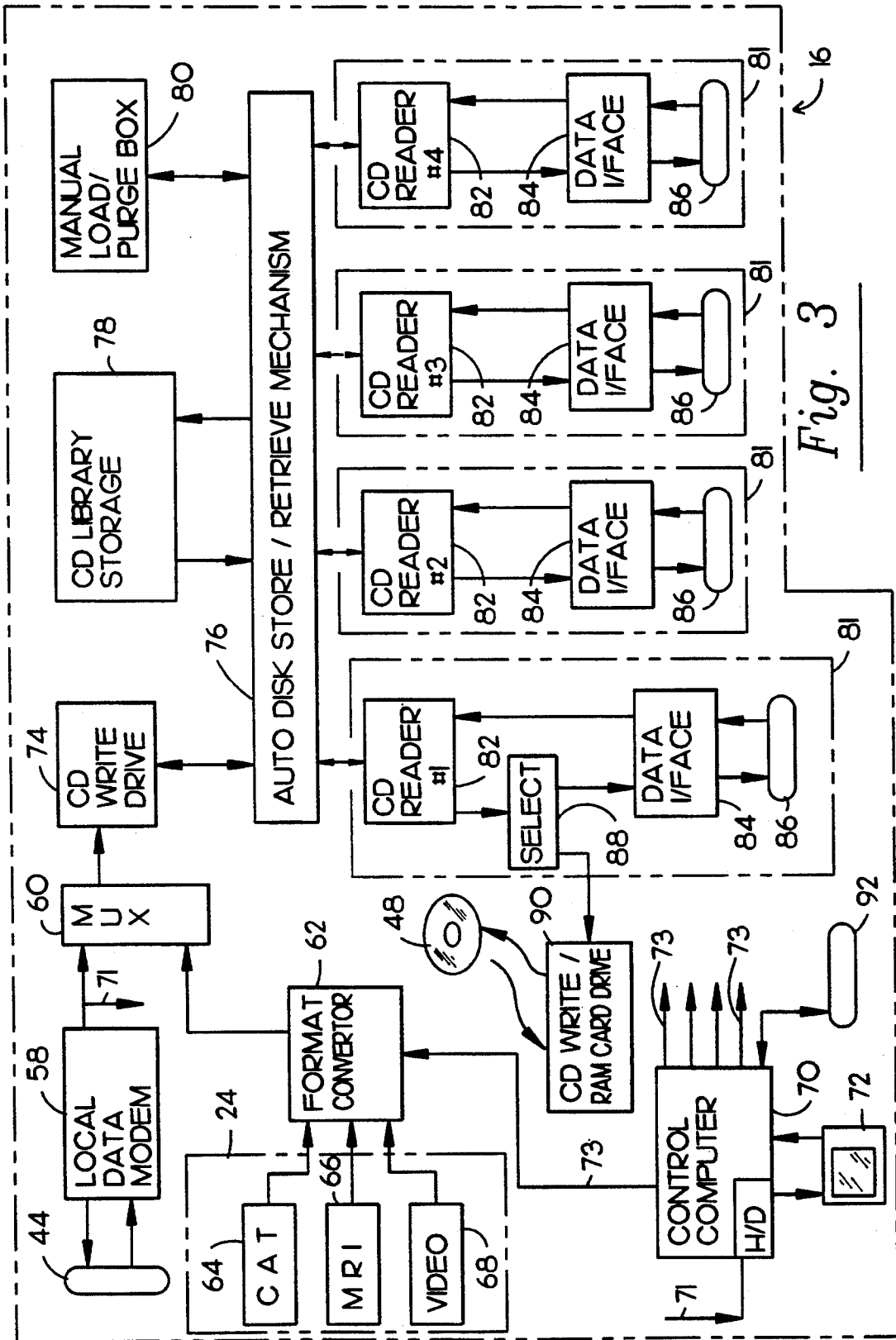


Fig. 3

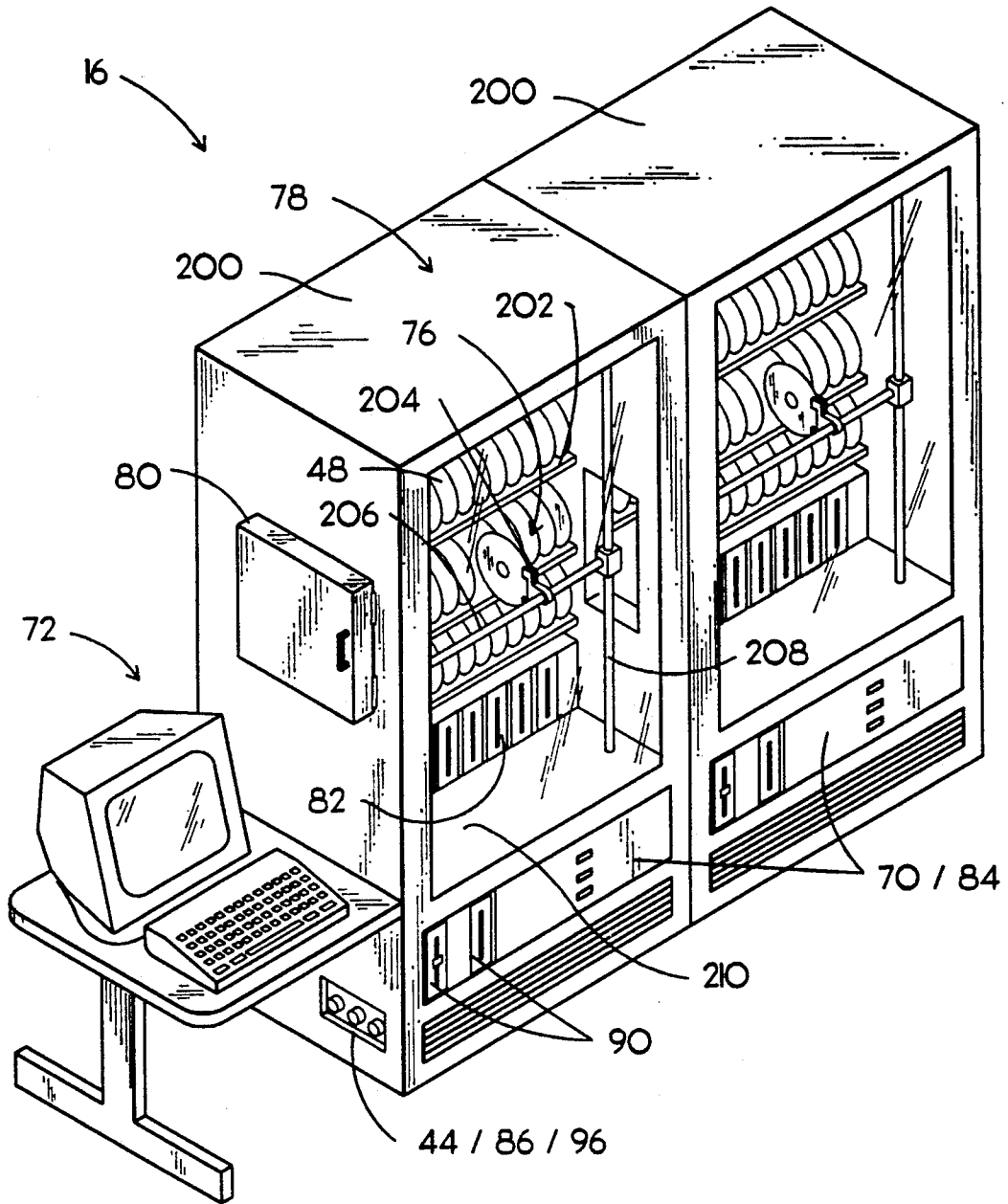


Fig. 4

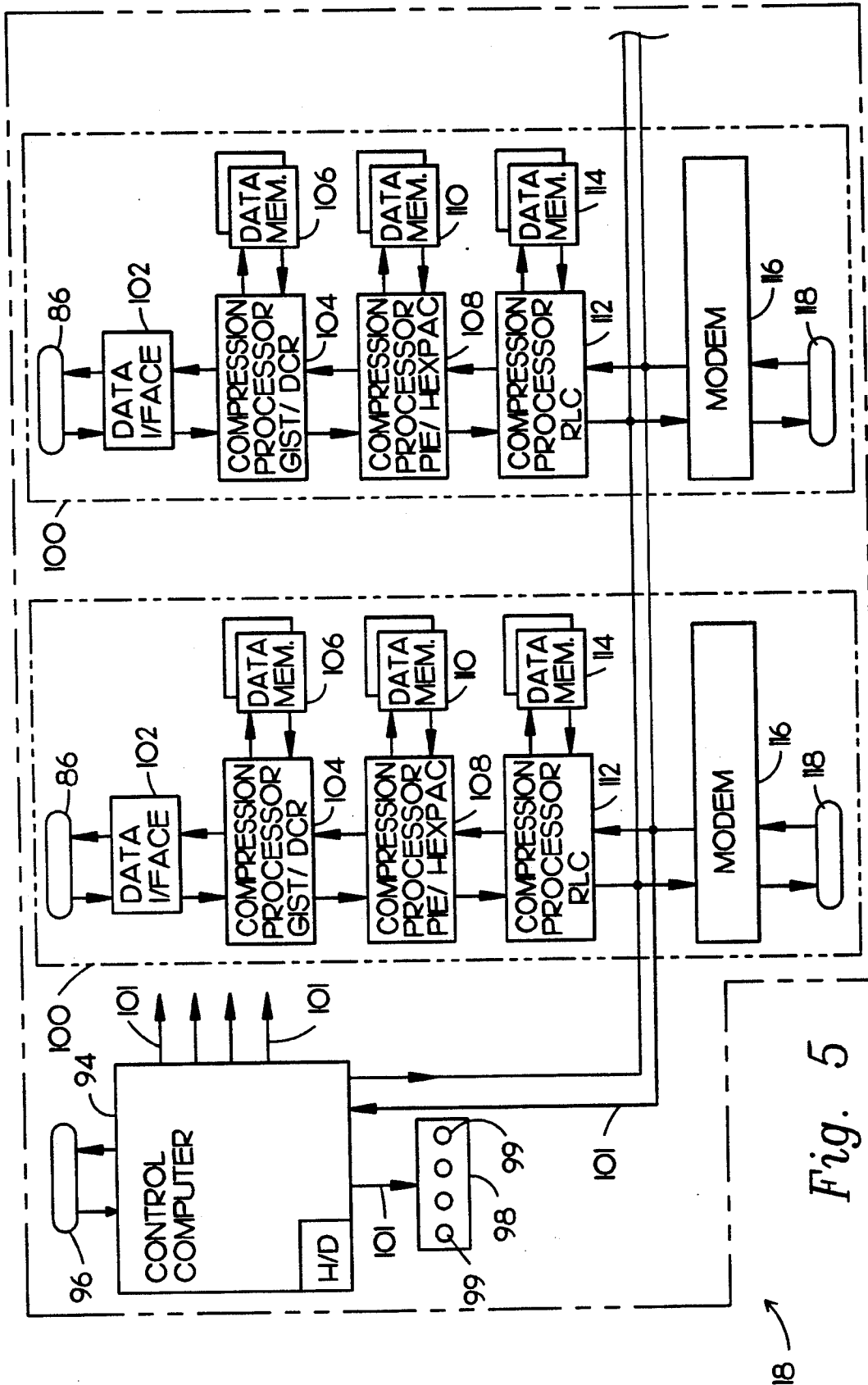


Fig. 5

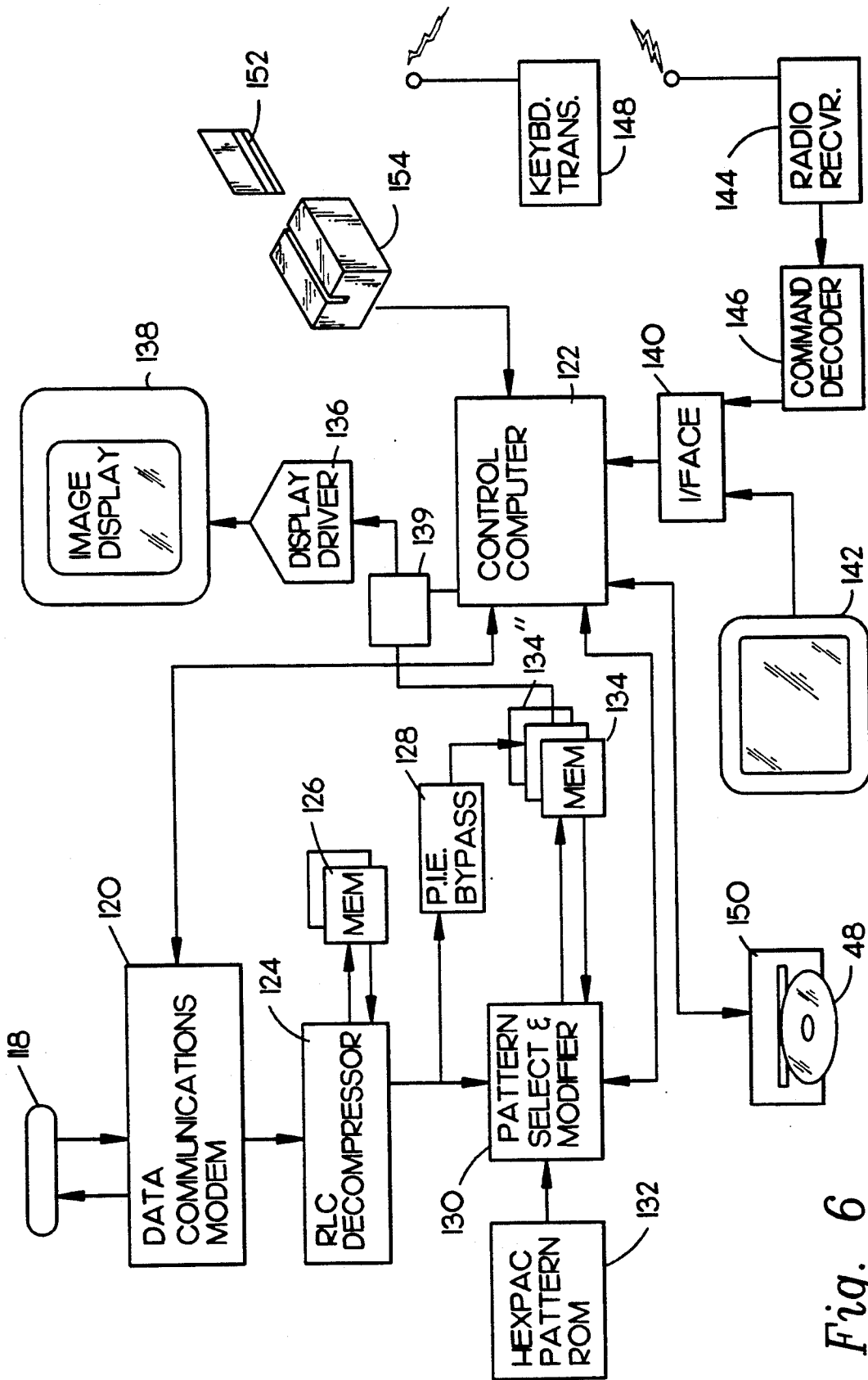


Fig. 6

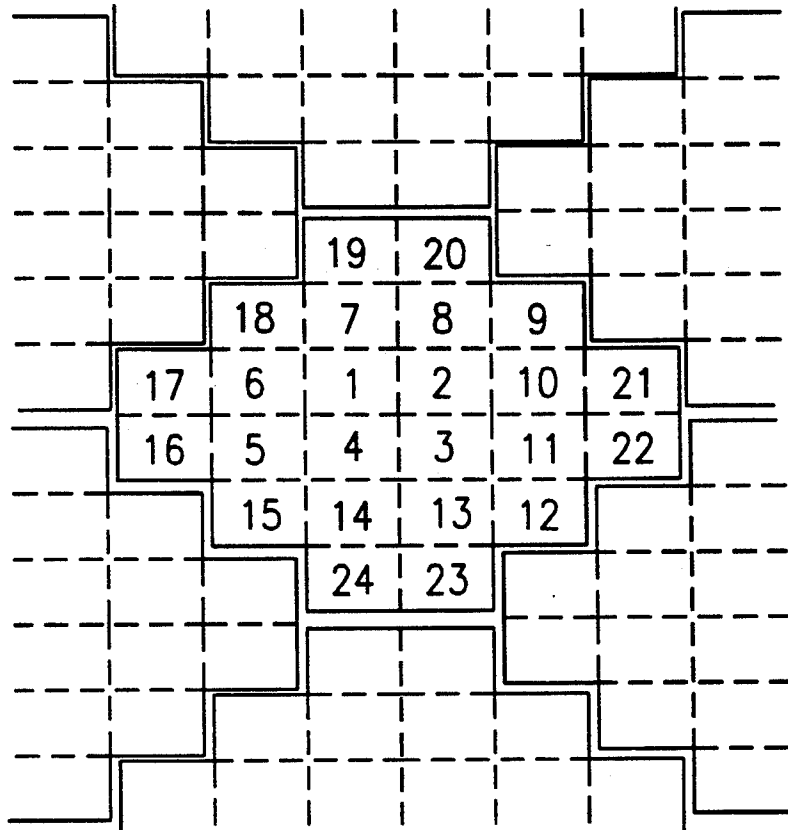


Fig. 7

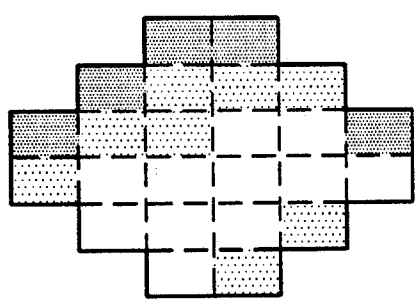


Fig. 8

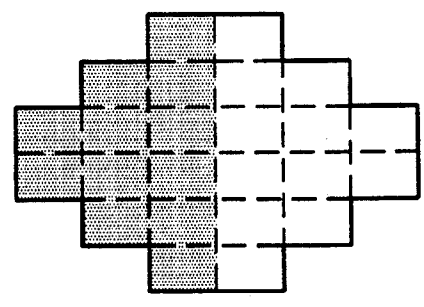


Fig. 9

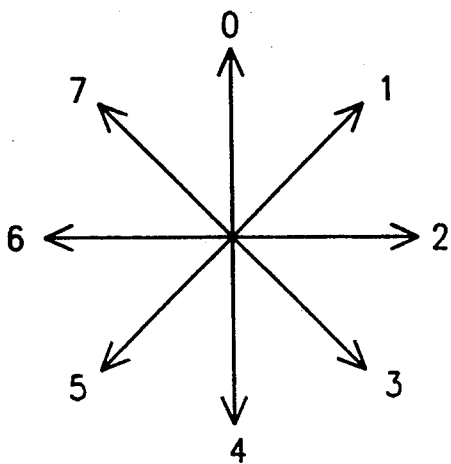


Fig. 10

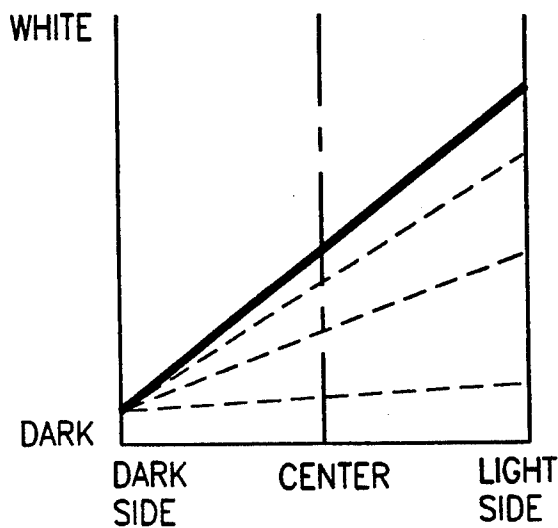


Fig. 11



Fig. 12

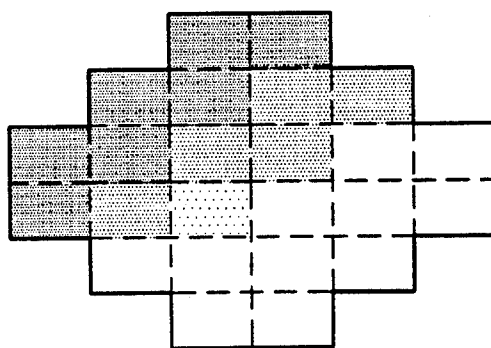


Fig. 13

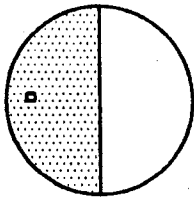


Fig. 14-A

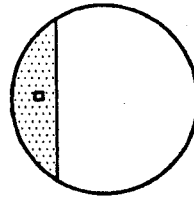


Fig. 14-B

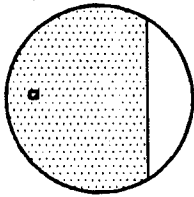


Fig. 14-C

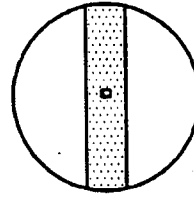


Fig. 14-D

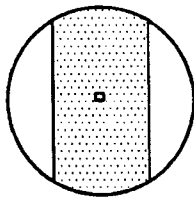


Fig. 14-E

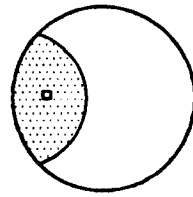


Fig. 14-F

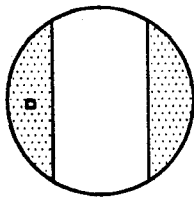


Fig. 14-G

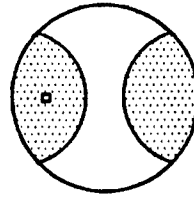


Fig. 14-H

AUTOMATED HIGH DEFINITION/RESOLUTION IMAGE STORAGE, RETRIEVAL AND TRANSMISSION SYSTEM

BACKGROUND OF THE INVENTION

1. Field of the Invention

An automated high definition/resolution image storage, retrieval and transmission system for use with medical X-ray films and the like.

2. Description of the Prior Art

Many industries have their own unique inventory control means for the particular needs of the respective trade. The food industry, for example, has largely standardized on bar-code technology to track food items. The medical profession, however, in many areas has not kept technological pace notwithstanding the obvious need therefor. Storage and retrieval systems for medical image data such as X-ray films, CAT scans, angiograms, tomograms and MRI are commonly antiquated and often employing methods of the 1920's. For example, when image films are used in the physician must display these photo films on a light box.

Moreover, due to the diffuse responsibilities of multiple attending physicians and treatment sites involving patients with particularly complex conditions, image data for such patients is often misplaced, lost, or at best, difficult to find when needed. Hospitals maintain large "file rooms" to store the patient image data. The film image data is typically stored in a large brown envelope approximately 14 by 17 inches and open at one end. These become bulky to handle and store especially in a complex situation in which several of these folders are needed. Weight alone can build up to 15 pounds. It has proven time consuming to obtain image data from file rooms either due to administrative backlogs, lack of specialized filing personnel and misfiling.

Typically, the physician examines the patient in his office after the radiographical studies have been made in a hospital or diagnostic facility. These films and the information contained therein are often unavailable at the time of the examination. Thus, there is a need for remote access to these image data for rapid patient assessment and therapy recommendation.

U.S. Pat. No. 4,603,254 teaches a stimuable phosphor sheet carrying a radiation image stored therein scanned with stimulating rays. The light emitted from the stimuable phosphor sheet in proportion to the radiation energy stored therein is detected and converted into an electric signal converted to a digital signal. Digital data is created to reproduce the radiation image for use in diagnosis and storage.

U.S. Pat. No. 4,764,870 describes a system for transferring medical diagnostic information from the diagnostic site to remote stations. An internal analog video signal from imaging diagnostic equipment such as a CAT scanner or MRI equipment, is converted to an analog video signal of different, preferably standard, format that is stored and transmitted in the reformatted image information to the remote terminal. The received signal is stored, decoded and applied in appropriate analog video form to an associated CRT display for reproduction of the diagnostic images.

U.S. Pat. No. 5,005,126 shows a system for transferring medical diagnostic information from the diagnostic site to remote stations similar to that found in U.S. Pat. No. 4,764,870.

U.S. Pat. No. 5,019,975 teaches a method for constructing a data base in a medical image filing system comprising the steps or recording information indicating the time at which each medical image is recorded and a rank of importance for each medical image as image retrieval signal data for image signals corresponding to each medical image; recording the number of times the image signals corresponding to each medical image have been retrieved as image retrieval signal data and incrementing the number each time the image signals are retrieved; and when the data base is full of image retrieval signal data, deleting the image retrieval signal data corresponding to the image signals of the medical image in which at least (1) the time at which the medical image was recorded earlier than a predetermined time and (2) the rank of importance of the medical image is lower than a predetermined value.

U.S. Pat. No. 4,611,247 describes a radiation image reproducing apparatus to read a radiation image from a first recording medium as a visible image. Input devices of the apparatus enter data which are associated with a method of exposing an object to a radiation and object's exposed part. In response to the input data, a processing condition determining unit determines conditions optimum for a gradation processing and a spatial frequency processing. A processor system is provided for reading the radiation image stored in the first recording medium and processing the radiation image on the basis of conditions which the processing condition determining unit determines in response to the input data associated with the radiation image.

U.S. Pat. No. 4,750,137 discloses a method and a computer program for performing the method for optimizing signals being exchanged between a host unit and an addressable-buffer peripheral device. The program optimizes an outgoing signal from the host unit by (1) creating an updated-state map representing the state of the peripheral device buffer expected to exist after processing by the peripheral device of the outgoing signal, (2) performing an exclusive-or (XOR) operation using the updated-state map and a present-state map representing the existing state of the buffer, and (3) constructing and transmitting a substitute outgoing signal which represents only changes to the buffer, and in which all premodified field flags are turned off. Position-dependent characters, such as attribute bytes, are translated into nondata characters prior to incorporation into a map, and are retranslated into their original form for use in the substitute signal.

U.S. Pat. No. 4,858,129 teaches an X-ray CT apparatus in which a plurality of dynamic tomographic images obtained by repeatedly photographing a region of interest of a subject under examination are stored in an image memory for subsequent display on a display device. A processing device extracts data of pixels along a certain the tomographic images and stores the pixel data in the image memory, in the order of photographing time of the tomographic images, thus forming a time sequence image formed of picked-up pixels. The processing device reduces a tomographic image and the time sequence image and rearranges the reduced images in one frame area of the image memory for simultaneous display thereof on the display device.

U.S. Pat. No. 5,021,770 discloses an image display system having a plurality of CRT display screens. The system is of the type in which a number of images of specific portions of a patient having a specific identification code are selected from among a multitude of X-ray

image taken by a plurality of shooting methods, and when the regions of interest are specific, a plurality of appropriate images are further selected using the previously stored aptitude values for the regions and shooting methods and displayed on the plurality of CRT display screens. In order that the segments to be inspected can be pointed to on the screen on which the image of the patient is displayed, a memory is provided which is adapted to previously store codes corresponding to the specific image of the patient and to specify the respective regions of the image in such a manner that they correspond to the pixel positions of the image.

U.S. Pat. No. 4,879,665 teaches a medical picture filing system composed of a picture data memory device, a picture data input-output device for inputting/outputting the picture data, a retrieving device for storing the picture data into the memory device and extracting it therefrom on the basis of retrieving data, a retrieving data input device for inputting the retrieving data into the retrieving device, a retrieving data storing device for storing the retrieving data, the retrieving data being classified by block of information obtained in one-time examination. When medical pictures are filed, retrieving data collected for each examination is utilized for reducing the amount of retrieving data, while when reproduced, retrieval is carried out for each one-time examination thereby shortening the time required for retrieval.

In light of recent advances in computer data basing, digitization and compression of image data, image enhancement algorithms and cost effective computer technology, the means for improved storage and retrieval of vital patient image data is now possible.

Such system should include the following major features:

- 1) means to more compactly store and more efficiently retrieve image data and automatically identify the data by patient name, image type, date and the like;
- 2) means for physicians to quickly and remotely access particular patient image data at the medical facility even if achieved at several different locations;
- 3) means to prevent loss of vital image data due to ordinary human handling and misplacement errors;
- 4) means to quickly and affordably access image data from the physician's office;
- 5) means to enhance the medical images by both contrast enhancement and zooming for improved diagnostics and/or surgical guidance; and
- 6) means to quickly and conveniently access image data and display on a large screen in the operating room with any desired enhancement or expansion.

As described more fully hereinafter, the present invention provides means to accomplish these goals. The system uses both general purpose system elements well known to those practiced in electronic arts and specific elements having significant novelty.

SUMMARY OF THE INVENTION

The present invention relates to an automated high definition/resolution image storage, retrieval and transmission system capable for storing, transmitting and displaying medical diagnostic quality images for use with medical X-ray films or the like.

The system comprises means to process the image data from patient imaging to physician usage. The major or significant processing stages are described hereinafter. Specifically, the major steps in the image data flow include:

PATIENT RADIOGRAPHY: The patient's body is imaged and a film is exposed as in an X ray room, MRI or CAT scan lab.

FILM PREPARATION: The film(s) is developed to create a visible image with OCR readable patient identification information superimposed thereon.

FILM INTERPRETATION: Commonly, a radiologist drafts an opinion letter for the film(s). This document preferably includes an optical character reader, or OCR, readable patient identification label or standard marking area.

IMAGE SCANNING AND DIGITIZING SUBSYSTEM: A scanner subsystem digitizes each patient image film and/or document on a high resolution scanner. This digitized data is transmitted by a local high speed data link to a separate or remote master storage unit. Patient identification information is read from a standard format on each file by OCR techniques and efficiently stored with the digitized image data. Enhanced scanner resolution and gray scale requirements are provided. Further, to reduce data rate requirements, data compaction or compression is accomplished within the scanner subsystem.

To back-up possible data link down time or scanner down time, the scanner subsystem may include a CD-ROM data storage drive so that image data may continue to be digitized. The CD-ROM disk may then be manually delivered to the file room unit for subsequent use.

In an optional embodiment, the digitized data of one or two images may be written to a compact semiconductor memory card "RAM Cards". This form of data storage may be used to send selected images for special purposes such as when the image data is needed in another city for second opinion purposes.

At this point in the image data flow, there is a split in which the original film data is stored as a "master" in a file room and the image disk is made available for active "on-line" use in an image storage and retrieval subsystem.

FILM FILING: The patient image films may be placed in the industry standard 14 by 17 inch brown paper folders and placed on conventional filing shelves. However, it is preferred that older films be tagged and stored off-site to reduce the current excessive bulk of films in many hospital file rooms. The system would now make this practical since the original films would seldom need to be accessed.

In the preferred embodiment of the system, the patient may have his entire image data collected and written to one or more of the storage CD-ROM disks for archiving at the hospital.

IMAGE STORAGE AND RETRIEVAL SUBSYSTEM: This subsystem is a remotely controllable, automatically accessible image data subsystem to store and automatically retrieve, on-demand, the compressed digital information contained on the CD-ROM disks.

The image storage and retrieval subsystem has a high-speed data link connection to the scanning and digitizing subsystem and has a write drive recording mechanism which is dedicated to receiving the data from the scanning and digitizing subsystem. This CD-ROM write drive can operate without interrupting remote access operations.

Remote access may be made to the image storage and retrieval subsystem by a variety of telecommunication links. Access will be granted only if a valid user code has been presented. By means of several read-only CD

disk drives and electronic buffering, virtually simultaneous access can be granted to several or more users.

As explained more fully hereinafter, the medical image disk will contain relatively huge quantities of data making it impractical to send over conventional data communication links without very efficient data compression technology. While there are a variety of data compression techniques available, none are well tailored to this application. Thus, novel compression means are in the a remote telecommunication access subsystem.

TELECOMMUNICATION SUBSYSTEM: Occasionally circumstances may warrant manually making an extra copy of the patient's image files to be physically delivered to an authorized requester. However, for the system to fulfill broad service to the health care industry it must be able to efficiently telecommunicate image files to remote locations both cost effectively and within a reasonable time interval.

A novel medical facsimile technology is in the preferred embodiment which works interactively with a remote requester to send only what is needed at acceptable resolutions, and the presented image is progressively updated as the communications connection is maintained until the resolution limits of the user display are reached, after which time, other images are either sent or further enhanced.

The specific technical means for accomplishing this uses the following novel technologies: a) guided image selection and transmission (GIST); b) progressive image enhancement (PIE); c) display compatible resolution (DCR); d) hexagonal pattern classification compression (HexPac) and e) run length coding (RLC), RLC is well known to those skilled in the arts.

It appears practical to send immediately useful patient data in less than one minute over a phone line (9600 baud) whereas it take many hours by conventional coding and transmission means. When wide-band telecommunications as satellite, fiber optic, micro-wave links becomes more generally available at affordable prices, then the more complex data compression techniques described here will be less important, but until that time, these types of techniques are believed essential to overall system success.

This combination of technologies to efficiently compress the image data and transmit remotely comprises the telecommunication access subsystem. In practice, these technologies may be implemented for the most part with available computer modules although several special signal processor boards are needed.

REMOTE DISPLAY TERMINAL: The quality of the image available to the user is limited or determined by the receiving presentation terminal or monitor. Two specific presentation terminal types are envisioned in the preferred embodiment of the system, a modified personal computer terminal for use in a physician's office, hospital nurses' station and the like, and a large screen presentation terminal with remote controlled interaction primarily for operating room use.

Both terminals have means to show the available patient directory of images, and means to select an image, enhancement and zoom on selected areas. Image enhancement has heretofore been impractical for film based images and thus much subtle but important pathological information has been largely lost. This is especially true of X-ray data The ability to enhance subtle contrasted tissues areas is considered to be an important feature and benefit of the system.

An optional high-resolution printer (300 dpi or better) permits the physician to print out selected images. This will be especially valuable when the physician expands and enhances selected critical image areas since a cost effective printer would otherwise not have adequate gray scale or pixel resolution to give diagnostically useful output.

Each terminal consists of a standard high performance personal computer with one or more data source interfaces such as RAM card, CD-ROM disk or data modem, a decompression graphics interface circuit and graphics display. The large screen presentation terminal has a large screen display for easy viewing for a surgeon who may be ten or more feet distant. The large screen presentation terminal also has an optional remote control so that an attending technician or nurse can scroll images, enhance and zoom, at the surgeon's request.

A keynote of each terminal design is a very simple user interface based upon a limited selection menu and obviously pointed-to graphical icons.

The invention accordingly comprises the features of construction, combination of elements, and arrangement of parts which will be exemplified in the construction hereinafter set forth, and the scope of the invention will be indicated in the claims.

BRIEF DESCRIPTION OF THE DRAWINGS

For a fuller understanding of the nature and object of the invention, reference should be had to the following detailed description taken in connection with the accompanying drawings in which:

FIG. 1 is a functional block diagram of the entire system of the present invention.

FIG. 2 is a functional block diagram of the image scanning and digitizing means.

FIG. 3 is a functional block diagram of the image data storage and retrieval means.

FIG. 4 is a perspective view of the image data storage and retrieval means.

FIG. 5 is a functional block diagram of the telecommunication means.

FIG. 6 is a functional block diagram of the remote display terminal means.

FIG. 7 depicts the hexagonal pattern of the hexagonal compression method.

FIG. 8 depicts an actual hexagonal pattern from an X-ray film. FIG. 9 depicts the selected predetermined hexagonal pattern most closely corresponding to the actual hexagonal pattern shown in FIG. 8.

FIG. 10 graphically represents the predetermined rotational orientations for the predetermined hexagonal patterns.

FIG. 11 graphically depicts a selected gray level slope of the selected predetermined hexagonal pattern of FIG. 9.

FIG. 12 depicts a single pixel from the predetermined hexagonal pattern.

FIG. 13 depicts a hexagonal pattern reconstructed by a remote display terminal means corresponding to the actual hexagonal pattern shown in FIG. 8.

FIGS. 14-A through 14-H depict the predetermined set of orthogonal gray level patterns.

Similar reference characters refer to similar parts throughout the several views of the drawings.

DETAILED DESCRIPTION OF THE PREFERRED EMBODIMENT

As shown in FIG. 1, the present invention relates to an automated high definition/resolution image storage, retrieval and transmission system generally indicated as 10 for use with medical X-ray film 12 or other documents to provide simultaneous automated access to a common data base by a plurality of remote subscribers upon request from the remote subscribers.

The automated high definition/resolution image storage, retrieval and transmission system 10 comprises an image scanning and digitizing means 14 to transform the visual image from the medical X-ray film 12 or other documents into digital data, an image data storage and retrieval means 16 to store and selectively transfer digital data upon request, a telecommunication means 18 to selectively receive digital data from the image data storage and retrieval means 16 for transmission to one of a plurality of remote visual display terminals each indicated as 20 upon request from the respective remote visual display terminal 20 through a corresponding communications network 21 such as a telephone line, satellite link, cable network or local area network such as Ethernet or an ISDN service for conversion to a visual image for display at the remote requesting site.

To improve automation and tracking, a machine readable indicia or label 22 containing key patient information may be used in association with the medical X-ray film 12. As shown, the machine readable indicia or label 22 is affixed to the medical X-ray film 12 prior to scanning by the image scanning and digitizing means 14 to provide file access and identification. Furthermore, digital data from alternate digitized image sources collectively indicated as 24 and file identification may be fed to the image data storage and retrieval means 16 for storage and retrieval.

FIG. 2 is a functional block diagram of the image scanning and digitizing means 14 capable of converting the visual image from the medical X-ray film 12 to digitized image data for transmission to the image data storage and retrieval means 16 over a bi-directional high speed data link 25. Specifically, the image scanning and digitizing means 14 comprises a film loading and scanning section and a data compression and transmission section generally indicated as 26 and 27 respectively and a display and control section generally indicated as 28. The film loading and scanning section 26 comprises a film input loader 30, alignment and sizing chamber 32, optical character reader 34 and film scanner/digitizer 36 capable of at least 500 dots per inch resolution 36; while, the data compression and transmission section 27 comprises a data buffer memory 38, low-loss data compression means 40, local data modem 42 and transmission connector 44 to operatively couple the image scanning and digitizing means 14 to the image data storage and retrieval means 16. The low-loss data compressor 40 is also operatively coupled to a compact disk data storage drive 46 capable of writing or storing compressed digitized patient image data on a compact disk 48. The display and control section 28 comprises a keyboard/control console 50, display terminal 52 and control computer 54 which is operatively coupled to the other components of the image scanning and digitizing means 14 through a plurality of conductors each indicated as 55. A film collector tray 56 may be disposed adjacent the film scanner/digitizer 36 to receive the medical X-ray film 12 therefrom following processing.

To reduce the approximately 238 Megapixels required to digitize a 14 inch by 17 inch, medical X-ray film 12 with 700 dots or pixels per inch with a two byte level to a manageable size without significant information loss, a linear gray level prediction, modified run-length code generating logic circuitry is embodied within the low-loss data compression means 40 to dynamically compress the digitized data before storage or recording. The image data is compressed with acceptable diagnostic resolution loss. The low-loss data compression means 40 measures the "local" slope of the pixel gray level and continues to compare that estimated gray level for up to an entire scan line until a pixel region is reached which differs from the linear estimate by more than a predetermined amount. The data actually sent for that region consists of the slope of the line, actual level at the origin of the slope line and the number of pixels comprising that region. The circuitry will discard linear gray level slope differences of the original film which can be reliably determined to be noise or image "artifacts". A sudden pixel (if at 1000 dots per inch) dramatic change in gray level could be rejected as dust or film noise for example. The compressed data is a trade-off between complexity, speed and minimum data loss to reduce the total data quantity stored by a factor of approximately three. Thus, about 80 Megapixels of data may still have to be stored per 14 inch by 17 inch film image.

In the preferred embodiment, the bi-directional high speed communications link 25 transmits the low-loss compressed digitized data from the developing lab room to the hospital file room where the image data storage and retrieval means 16 will transfer and store the patient and image data in a new patient file on a compact disk 48.

Two way communications between the image scanning and digitizing means 14 and the image data storage and retrieval means 16 minimizes data loss by insuring that a compact disk 48 be available to receive and store data. Moreover, the compact disk data storage drive 46 with re-writable ROM technology can record data even if communications with the image data storage and retrieval means 16 is disrupted. Thus the image scanning and digitizing means 14 can automatically start writing data to the compact disk data storage drive 46 as soon as an image data storage and retrieval means 16 fault is sense. The display and control section 28 informs the operator of the system status.

In operation, the film lab technician may stack one or more medical X-ray films 12 onto the input loader 30 as shown in FIG. 2. A "read" button is depressed on the keyboard/control console 50 and each film 12 is thereafter fed in automatically, digitized and transmitted to the image data storage and retrieval means 16 located in the file room. As the reading of each film 12 is completed, the film 12 is deposited into the film collector tray 56. System status, number-of-films read logging and so forth are shown on the display terminal 52.

Initially, the image scanning and digitizing means 14 positions the film 12 in the alignment and sizing chamber 32 on a precision carrying platen for subsequent optical scanning. This platen contains optical sensors to sense the exact film size so only the useful image area is digitized. Once the film 12 is secured onto the movable platen, the film 12 is passed through the optical character reader 34 and then to the film scanner/digitizer 36.

The patient data and image identification is first recorded onto the remote CD-ROM file directory in the image data storage and retrieval means 16 from the

OCR "pass" and then the compressed scanned image data is sequentially written to a compact disk 48 by a CD write drive for storage with the CD library storage of the image data storage and retrieval means 16 as described more fully hereinafter as the film 12 slowly passes through the film scanner/digitizer 36.

Specifically, the film scanner/digitizer 36 converts the image to a digital representation of preferably at least a 700 dot per inch resolution. This digital data is temporarily stored in the data buffer memory 38 where the patient data from the optical character reader 34 and corresponding digitized image data from the file scanner/digitizer 36 are properly formatted for subsequent compression and transmission to the image data storage and retrieval means 16. The stored data is then accessed by and compressed by the data compression means 40 as previously described and transmitted through the local data modem 42 and transmission connector 44 to the image data storage and retrieval means 16 or a compact disk data storage drive 46. The display and control section 28 permits the X-ray lab staff to monitor system status, report quantity of documents and films processed and allow for scheduling local recording of image data on compact disks 48.

FIGS. 3 and 4 show the image data storage and retrieval means 16 to receive and store the low-loss compressed digitized patient information and image data from the image scanning and digitizing means 14 and to selectively transmit the stored low-loss compressed digitized patient information and image data to one or more of the remote visual display terminal(s) 20 through corresponding telecommunication means 18 and corresponding communications network(s) 21 upon request from one or more of the remote display terminal(s) 20.

The image data and retrieval means 16 is essentially a central data storage library for medical subscribers to remotely access and visually display patient data and information.

As described hereinafter, the image data storage and retrieval means 16 is robotically automated to minimize hospital staff requirements. At any given time, it is estimated that a typical hospital may have several hundred active patients with requirements for physician access to corresponding image files. An active patient may require one to three compact disks 48. Thus, the image data storage and retrieval means 16 should have sufficient means to store and retrieve at least 500 compact disks 48.

Further, to minimize personnel requirements, the image data storage and retrieval means 16 has a semi-automatic log-in mechanism for updating the compact disk inventory and an automatic mechanism for retrieving and reading the compact disks 48 remotely via communication link interfaces similar to juke box playback mechanisms. Except for the occasional loading of new empty compact disks 48 and removal of inactive compact disks 48, the operation of the image data storage and retrieval means 16 is fully automatic permitting authorized access at any time.

As described more fully hereinafter, several playback drives with electronic buffering are incorporated so that essentially simultaneous access can be provided to several remote requesting users. An optional duplicating CD write drive and RAM-Card drive permits additional copies to be made locally upon demand for either back-up or other use. The image data storage and retrieval means 16 has an operator's console/desk ar-

angement for file maintenance and duplicating control by the hospital file room clerk. Control software is a simple menu selection design so that relatively unskilled personnel can maintain the central data storage library or image data bank.

As shown in the functional block diagram of FIG. 3, the image data storage and retrieval means 16 comprises a local data modem 58 operatively coupled between the image scanning and digitizing means 14 through the transmission connectors 44 and bi-directional high speed communication link 25, and a selector or multiplexer 60. A format convertor 62 is operatively coupled between the alternate digitized image source(s) 24 such as CAT 64, MRI 66 and/or video 68 and control computer 70 which is, in turn, coupled to a control console 72 including a visual display and input means such as a keyboard. The local data mode 58 is also coupled to the hard disk (H/D) of the control computer 70 through a conductor 71. The other components of the image data storage and retrieval means 16 are coupled to the control computer 70 through a plurality of conductors each indicated as 73. A CD write drive 74 is operatively coupled between the multiplexer or selector 60 and an auto disk storage/retrieve mechanism 76 which is, in turn, operatively coupled to a CD library storage 78, a manual load/purge box 80 and a plurality of data retrieval and transmission channels each indicated as 81. Each data retrieval and transmission channel 81 comprises a CD reader drive 82 operatively coupled through a corresponding data interface 84 to a corresponding transmission connector 86. In addition, one of the CD reader drives 82 is operatively coupled through a selector switch 88 to an optional CD write/RAM card drive 90 configured to manually receive a compact disk 48 or RAM card.

As shown in FIG. 4, the CD library storage 78 comprises at least one cabinet 200 to operatively house 800 compact disks 48 arranged on four shelves each indicated as 202 and the auto disk storage/retrieve mechanism 76 which comprises a CD coupler 204 to engage and grasp a selected compact disk 48 and move horizontally on a support member 206 that moves vertically on a pair of end support members each indicated as 208. An access door 210 permits movement of compact disks 48 to and from the cabinet 200. However, in normal operation, "old" patient data is removed by writing collected image data to a single compact disk 48 through the CD write/RAM card drive 90 thus freeing internally disposed compact disks 48 for new data. The CD write/RAM card drive 90 may also be used to collect a patient's image data on a single compact disk 48 for use in the operating room's display terminal. This obviates the need for a high speed internal hospital local area network.

The computer associated with the CD robotic arm and drive mechanism performs ordinary library maintenance functions such as retrieval of outdated files, access statistics, entry of access validation codes, and so forth. This computer subsystem also handles data communication interface functions.

Internal to the environmentally controlled cabinet 200 are a plurality of playback mechanisms (field expandable to six) which are automatically controlled by the accessing physicians via the coupled communications system. Yet another CD-ROM write drive can record new data from the image scanning and digitizing means 14 or perform library functions such as consoli-

11 dation of a patient's data from several compact disks 48 to a single patient-dedicated compact disk 48.

The internal computer maintains a file log of which compact disks 48 are empty and where each patient's image data is stored by disk number and track on a disk location. When the image scanning and digitizing means 14 requests to down-load data, the auto disk storage/retrieval mechanism 76, of the image data storage and retrieval means 16 retrieves the "current" compact disk 48 which is being written with data (if not already loaded), then loads the compact disk 48 into the CD write drive 74, and signals to the image scanning and digitizing means 14 to transmit. Image data is then recorded with a typical record time of 4 minutes for a full-size, high density image.

Once the robotic arm has delivered the compact disk 48 to the CD write drive 74, the robotic arm is free to access and place other compact disks 48 onto CD reader drive 82 as commanded by its communications interface. The robotic arm can find and place a disk 48 into the appropriate CD reader 82 in approximately 10 seconds. Thus, there is minimal waiting time for disk access unless all CD readers drives 82 are in use.

As shown in FIG. 3, data is received through the input transmission connector 44 to the CD write drive 74 through the selector switch 60. Alternately, other image data from other sources such as CAT scanners 64 or MRI medical equipment 66 may be fed through the format convertor 62 for storage on a compact disk 48. If the other image sources are written to CD write drive 74, file identification data must be supplied to the format convertor 62 from the control computer 70.

The image file data received from the image scanning and digitizing means 14 is directly written to free space on a compact disk 48 in the CD write drive 74. No other data compression or special formatting is required as the image scanning and digitizing means 14 has performed these functions. As new image data is received from the image scanning and digitizing means 14 or another image source 24, the image data is sequentially appended to the last file on the compact disk 48 currently being written to. Thus, no attempt is made to organize a single patient's image files onto a single compact disk 48. However, each file received is logged into the control computer 70 through the conductor 71. Therefore, the control computer 70 always knows what disk location in the CD library storage 78 contains any specified file. Once a compact disk 48 is filled with image data, the auto disk storage/retrieve mechanism 76 removes the compact disk 48 from the CD write drive 74 and stores the compact disk 48 in an empty location in the CD library storage 78.

The plurality of data retrieval and transmission storage 78. channels 81 service the data requests from subscribers. As previously indicated, a single data retrieval and transmission channel 81 includes the select switch 88 to direct image file data to the optional CD write/-RAM card drive 90. By this means, all image data for an individual patient may be collected on one or more selected compact disks 48 for archiving or other use. However, normally, the control computer 70 will automatically remove old image data by removing the compact disk 48 from the CD library storage 78 and placing the compact disk 48 in the manual load/purge box 80. The removal age and exceptions information are selected by the system operator from the control console 72.

The control console 72 is also used to enter and maintain subscriber access identification codes in an "authorization file". This updated user authorization file data is sent through a transmission connector 92 to the telecommunications means 18 internal computer memory accessed by the control computer 70 as needed to accept or reject subscriber data link access requests. The user authorization file normally residing in the telecommunications means 18 may be remotely updated by authorized persons.

The number of data retrieval and transmission channels 81 depends on intended subscriber demand. The image data storage and retrieval means 16 is modular and may be upgraded as demand increases. Each data interface 84 operates cooperatively with the telecommunications means 18 to send only as much information as the telecommunications means 18 can compress and transmit to a remote visual display terminal 20 of a requesting subscriber in a given time interval. Thus, the interface is an asynchronous block-buffered type.

Since the entire system 10 is designed to provide easy and quick access to a patient's medical images, it is vital that these images be transmitted to a variety of locations in a timely and cost effective manner and further data compression is imperative. The telephone network is still the most commonly available network but has a severe data rate limitation of about 1200 bytes per second (9600 baud). While other high speed telecommunication channels such as time-shared cable, satellite link may eventually become commonly available, for the immediately foreseeable future, the "phone" network must be used if system 10 is to be practical today.

As noted earlier, a typical medical image may be stored as 119 megabytes of data. At 1200 bytes per second, it could take 27 hours to completely transmit the already compressed medical image data. This is obviously unacceptable. To overcome this obstacle, the telecommunications means 18 as shown in FIG. 5 utilizes five distinct data handling technologies to achieve useful data image transmission in less than one minute:

(1) Guided Image Selection and Transmission or GIST depends upon interactive use by the physician to identify what portions of an image are needed for enhancement or better resolution. Thus the data actually transmitted to the subscriber's visual display terminal 20 is guided by the subscriber observing the image. In particular, once the user has an image displayed on his or her visual display terminal 20, the user may outline a specific region of interest such as a lesion or tumorous growth for more detailed study. The operator may select this region using a "mouse" or light pen or similar well-known computer display terminal peripheral device. Having selected this region, the visual display terminal 20 will display the more detailed pixel data be sent on this region. The telecommunications means 18 will continue to send further precision data until the natural resolution limits of the display are reached or all available data is sent and received. This process of expanding an image region is known as "zooming" in computer-aided design systems. The novel feature here is that the image is further refined in resolution when "zoomed". The means for doing this and knowing when to "stop" further pixel transmission is defined by the PIE and DCR technology described hereinafter,

(2) Progressive Image Enhancement or PIE utilizes the transmission time from the instant a first "crude" image is presented to the subscriber to the present time of observation to progressively enhance the quality of

the presented image. The longer the user observes a selected image, the "better" the image becomes in the sense of pixel resolution and quantity of gray levels. In the preferred embodiment, hexagonal pixel groups are first transmitted using the HexPac pattern compression technology described hereinafter. Once a full terminal screen display has been made composed of these hexagonal patterns, then the telecommunications means 18 transmits more precise pixel detail. First all pixels located on the periphery of each hexagonal group are updated with their exact gray level values and thereafter, all inner pixels are similarly updated. If the display terminal's resolution is less than the 1000 dots per inch of the source image data, then pixel groups are sent, such as a square of four pixels, which match the display resolution and "zoom" expansion selected. This display matching technique is further defined hereinafter as DCR,

(3) Display Compatible Resolution or DCR transmits information about the user's terminal 20 back to the telecommunications means 18. Only data with a resolution compatible with that terminal 20 will be sent. Any excess data-link connect time can be used to send other image data which is likely to be requested or has been pre-specified to be sent.

(4) An image pattern compression method comprising a Hexagonal Pattern Classification or HexPac exploits the two dimensional nature of images. The data received by telecommunications means 18 is first uncompressed and placed into a multi-scanline digital buffer. This image data is then divided up into hexagonal cells and matched against predefined patterns. Many fewer bits of data can be used to represent these predefined patterns thus substantially compressing the image data for phone-line transmission. The pixels of these hexagonal patterns may easily be "refined" by the PIE technology described earlier. If the DCR subsystem determines that the user terminal has a pixel area of, say, 1500 by 1000 dots, then the HexPac technology recreates a new super pixel which is the average gray level of all actual pixels within that super pixel area. This immediately reduces the quantity of pixels to be sent (to only 1500 by 1000 pixels). Without further data compression, this quantity of data would still require about 26 minutes of data transmission time at 9600 baud, the highest available phone network data rate.

(5) Run length coding or RLC permits data to be compressed by specifying how many pixels have the same gray level in a sequence or "run length" of scanning. The image data sent by a CD reader drive 82 to telecommunications means 18 is compressed with run-length coding but is nearly loss-less in the duplication of the original film data. To substantially reduce the quantity of data needed to send an acceptable medical image to a remote user terminal 20 over the data-rate limited phone-line modem, a "lossy" compression is used. Since the PIE and DCR techniques described earlier will eventually provide any degree of diagnostic image integrity desired, it is believed acceptable to initially transmit a "lossy" image provided it gives adequate resolution for the user to begin the analysis and guided image selection. Many fewer bits can describe this "run" of similar gray levels thus compressing the amount of data sent. This technique is well known and often used in facsimile transmission. A one dimensional RLC is incorporated in the preferred embodiment but since HexPac elements are being coded, it can be con-

sidered more accurately a quasi two dimensional RLC compression.

FIG. 5 is a functional block diagram of the telecommunications means 18 including a control computer 94 operatively coupled to the image data storage and retrieval means 16 through a transmission connector 96. The various components of the telecommunications means 18 including a status panel 98 with a plurality of system indicators each indicated as 99 and a plurality of data compression channels each generally indicated as 100 coupled to the control computer 94 by a plurality of conductors each indicated as 101.

Each data compression channel 100 comprises a transmission connector 86, a communications data interface 102, a first compression processor or means 104 including logic means to generate the GIST and DCR data compressions and corresponding first data memory 106, a second compression processor or means 108 including logic means to generate the PIE and HEXPAC data compressions and corresponding second data memory 110 and a third compression processor or means 12 including logic means to generate the RLC data compression and corresponding third data memory 114, a corresponding modem 116 and a transmission connector 118.

The control computer 94 coordinates or controls data flow to and from the plurality of data compression channels 100 through the transmission connectors 86 and 118 respectively. Validated subscriber image data requests are transmitted to the image data storage and retrieval means 16 which searches the image library file 78 for availability of the requested compact disk 48. If available, the image data storage and retrieval means 16 loads the appropriate disk 48 from CD library storage 78 into a CD reader drive 82 and informs telecommunications means from 18 through the transmission connector 96 to the control computer 94 that a specific data interface 84 has data available to be transmitted through the corresponding transmission connectors 86. Once a subscriber transaction has been turned over to a specific data retrieval and transmission channel 81, the data compression channel 100 receives the data therefrom unless commanded to stop by a feedback control line. The data interface 102 is used to inform the CD reader drive 82 as to what portion of the image is requested by the first compression means 104. Generally, the complete image is first requested. Thus the CD reader drive 82 is requested to read the image data from the start.

The data is temporarily stored in the first data memory 106. Here the pixel data is first expanded from the RLC code into uncompressed pixel data. This is only done on a relatively few number of scan lines—about one tenth of an inch height of original image data. This uncompressed data is then remapped by the first compression means 104 into "larger" pixels whose average intensity is the average of all combined pixels compatible with the display resolution receiving remote visual display terminal 20. This "super pixel" data is then fed to the second data memory 110. The super pixel data in memory 110 is then processed by the second compression means 108. Initially, the lowest resolution image will be transmitted to rapidly form a useful remote image on the requesting remote visual display terminal 20 through a communications network 21. This will be done by combining super pixels in the second data memory 110 into hexagonal patterns which approximate the group of super pixels. These HexPac data packets are then sent to the third data memory 114. There the Hex-

Pac data packets are further compressed by the third compression means 112. These packets of run-length coded HexPac data packets are then transmitted through the corresponding modem 116 and transmission connector 118 over the selected communications network 21. The modem 116 includes state of the art error control techniques such as block retransmission when a remote error has been detected. Thus, data transmission is essentially error free as needed for compressed data handling.

The control computer 94 includes circuitry means to monitor the activity of each data channel. The identification of each subscriber is logged along with the total connect time for billing purposes. Thus the control computer 94 generally coordinates the plurality of communication links and their connections to the particular data retrieval and transmission channel 81 within the image data storage and retrieval means 16 as well as granting access and performing connection accounting tasks. The status panel 98, connected to the control computer 94 is used to aide in system debug and indicate operation of the data compression channels 100. The status panel 98 would not normally be used by hospital personnel but by system service technicians.

The control computer 94 also has a permanent memory such as a hard disk to record subscriber usage data and internally sensed hardware problems. This data may be downloaded on any of the transmission connectors 118 when a correct authorization code has been received. Thus, the servicing company can acquire subscriber usage information remotely for billing purposes and system diagnostic purposes.

The preferred embodiment of the telecommunications means 18 uses modular communication channel hardware. Thus, the module may be customized to function with any type of communication channel such as satellite links, cable networks or a local area network such as Ethernet or ISDN services.

It is important to note that all communications is bidirectional so that if, say, a remote visual display terminal 20 should become temporarily "overloaded" with image data due to decompression processing delays or due to a detected data error, then, the remote visual display terminal 20 may request that data transmission be stopped or a block of data be repeated until it is received correctly.

FIG. 7 graphically shows a hexagonal group of the hexagonal compression method comprising a group of square image pixels partitioned into a hexagonal group. The pixels are numbered for convenience of reference from the inside to the outside in a clock-wise manner. Each hexagonal group or packet comprises 24 super pixels as earlier described but other numbers are possible. It is assumed that each pixel is gray level coded using 2 bytes of data. Thus, the hexagonal group requires (24×2) 48 bytes of data to fully represent the 24 super pixels comprising the image pattern at the user terminal 20.

FIG. 8 shows a typical pattern as may occur in a region of an X-ray film 12. The method predefines a group of likely patterns, one of which is represented as a "best" match as in FIG. 9 with the actual pattern in FIG. 8. As shown in FIGS. 14A through 14H, there are 8 possible predetermined representative gray level patterns represented by 3 bits. These patterns are specifically selected to be essentially uncorrelated with each other even rotated relative to each other. As shown in FIG. 10, these patterns may be rotated through 8 equal

angles (another 3 bits of data) to best match the actual pattern. Rotation angle "1" is shown in FIG. 10 as the best match for the given example. Thus far, six bits have been used to approximate the actual pattern of FIG. 8. As shown in FIGS. 14A through 14H, each fictitious pattern includes a dark and light regions and origin. Although FIG. 11 discloses a straight gray level slope corresponding to the pattern shown in FIG. 14A, the gray level slope will vary with the fictitious pattern. For example, the gray level slope of the fictitious pattern shown in FIG. 14D would closely approximate a V shape.

FIG. 11 shows how the gray level slope may be discretely selected to best match the slope of the actual pattern. Two bits are used to approximate this slope.

FIG. 12 shows that one particular pixel, such as the darkest pixel, has been selected to be fairly precisely gray level represented by means of 8 bits (256 gray levels).

The total bits required to approximate the actual pattern is 16 or two bytes. FIG. 13 shows how this fictitious or reconstructed pattern may be reproduced at the user terminal 20 when decoded.

In this example, only two bytes were required to represent "adequately" an original 48 bytes of image data. Thus, a 24 to 1 compression ratio has been achieved. Further, run-length encoding (RLC) may be used on these HexPac Groups to further reduce redundant spans of white and black. It is estimated that the combined compression ratio of HexPac and RLC on the super-pixel image is about 36 to 1 for this particular set of parameters. This combined compression technology reduces data transmission time (at 9600 baud) to approximately 43 seconds for an initial useful medical image.

For medical images, further enhancements through the PIE compression should favor the elimination of artificial lining between hexagonal patterns first. As the user continues to view the same image, then the PIE compression will progressively improve the gray level integrity by updating all number 24 pixels to 8 bits of gray level resolution and updating all number 23 pixels to 8 bits of gray level and so forth for all remaining pixels in descending order. This process takes about 10 minutes at 9600 baud to update all peripheral hexagonal pixels and about 20 minutes total for all pixels.

If the user continues to observe or request further image resolution, the telcommunication means 20 causes each pixel gray level to be updated by one additional bit in descending order again until the full 16 bits of gray level is received and stored at the terminal 20 for each super pixel. Each doubling of gray level resolution takes between 1 and 2.6 minutes at 9600 baud depending on the run length statistics of the gray levels.

FIG. 6 is a functional block diagram of a remote visual display terminal 20 to be operatively coupled to one of the data compression channels 100 of the telecommunication means 18 by a communications network 21 and a transmission connector 118. The visual display terminal 20 comprises a data communications modem 120 operatively coupled to a control computer 122 and RLC decompression means 124. The RLC decompression means 124 is, in turn, operatively coupled to a memory 126, a PIE bypass 128 and a pattern select and modifier 130 which is operatively coupled to a Hexpac pattern ROM 132 and the control computer 122. A memory 134 is operatively coupled between the PIE bypass 128 and pattern selector and modifier 130 and a display drive 136 which is operatively coupled to an

image display 138. In addition, an image enhancing processor means 139 including circuitry to generate edge contrast enhancement, gray level contrast enhancement by means of gray level region expansion or differential gray level tracking and gray level enhancement or other state of the art image enhancement method well known to those skilled in the art. The control computer 122 is operatively coupled to an interface 140 to a first control or selector means 142 and a second control or selector means including a radio receiver 144 and signal command decoder 146 for use with portable keyboard transmitter 148. In addition, an optional CD read/write drive 150 may be provided for use with a compact disk 48.

The modem 120 has built-in compatible error correction technology to communicate with corresponding transmitting data compression channel 100. After the user has selected the image data storage and retrieval means 16 and validated authority by swiping through an identification magnetic card 152 or otherwise through a magnetic card reader 154 entered an assigned security code, the operator may select a patient and one or more image files presented to him on the display screen 138. Selection is accomplished by a touch-screen overlay on the first control or selector means 142 or by the keyboard transmitter 148 of the second control or selector means.

Once one or more images have been selected by the user, the modem 120 writes image data to the temporary memory 126 which is actively accessed by the RLC decompression means 124. This decompressed data describes the HexPac patterns or packets as stripes of the image running, for example sequentially from left to right. These Hexpac pattern specifications, typically 2 data bytes or 16 bits are then routed to a pattern selection processor 130 which accesses the predefined patterns from Read-Only Memory device 132. Each pattern is then rotated and gray-level modified by processor 130 according to the HexPac 16 bit pattern specification received from the RLC decompression means 124. Each modified pattern is then written to a graphics display memory circuit 134. As the graphics display memory circuit 134, develops the pattern data, the display driver 136 and image display 138 show the image on the screen as it is received. In this manner, the entire "first pass" medical image is painted on the image display 138 screen.

If the user makes no further intervention, then once the image is fully displayed on this "first pass", then the progressive image enhancement technology requests pixel enhancement data. This enhancement data bypasses pattern selector and modifier 130 and is routed through the PIE bypass 128. In the PIE bypass 128, the enhanced pixel information is directed to the correct graphic memory locations in the graphics display means circuit 134'. Thus, the display driver 136 and image display 138 are continually resolution enhanced.

If the image is fully enhanced to the limits set by the DCR in the control computer 122, the image storage and retrieval means 16 is directed by the control computer 122 to begin sending new image data on the next selected image and begin storing this image data in a second graphics display memory circuit 134". This second data memory 134" can hold one more images and may be selected immediately by the user when he is finished inspecting an earlier image. The user may further direct by touch screen command 142 that these be

stored in the computer's hard disk or archived by the optional CD read/write drive 150.

The user may at any time select a portion of the displayed image for further expansion by enabling or selecting the Guided Image Selection & Transmission (GIST) circuitry in the control computer 122 or image enhancement through the image enhancing processor means 139. This may be accomplished either by touch screen control means 142 or the second remote control means 144/146/148. This remote keyboard and transmitter unit 144/146/148 duplicates the on-display simulated push-buttons of the touch screen control means 142. Coded command signals sent by 148 are received by radio receiver 144 and decoded by 146. These commands are then accepted by control computer 122 as though they were normal keyboard commands.

The user may terminate a session with the image data storage and retrieval means 16 at any time by selection of stop and escape command. While a printer is not shown in this description, it can be an optional addition to terminal 20.

In summary, the image data storage and retrieval means 16 selects the first image and writes that data to a temporary memory buffer in the telecommunication means 18. Information about the subscriber's terminal is uploaded to the telecommunications means 18 so that the Display Compatible Resolution (DCR) logic circuitry knows when to stop sending added data for the requested first image. A special interactive compression computer then compresses this first image data using the HexPac circuitry and sends that data over the data link modem to the subscriber terminal 20. Error detection and correction methods will generally be used in this communications link protocol.

Once a first "crude" image is sent to the subscriber visual display terminal 20, then the Progress Image Enhancement (PIE) circuitry begins to send additional data to further refine the resolution of each hexagonal pixel region. If no further guidance is given by the subscriber, the PIE will continue to refine the picture's resolution until its natural limit is sent for the terminal 20. Thereafter, the PIE will begin sending image data from the second specified film and loading it into yet another memory buffer. Thus, the data link connection is always transmitting useful data even though the subscriber may only be analyzing one image for some time.

However, should the user desire to zoom in on a particular region of an image, he or she may define that region desired on the terminal 20 by the Guided Image Selection & Transmission (GIST) to expand the visual display accordingly. The DCR will recognize the requirement for additional resolution and command the PIE to begin transmitting additional pixel information until such time as the DCR informs it that once again the natural display resolution limit has been reached.

The following image enhancement means present in the instant invention: edge contrast enhancement, gray level contrast enhancement by means of gray level region expansion or differential gray level tracking and gray level enhancement and may be accomplished by the image enhancing processor means 139 in the visual display terminals 20.

The human eye cannot reliably discern gray level differences less than approximately 2%. Yet, significant tissue density information causes X-ray gray level differences in this range and below. The enhancement technologies above will cause these tissue density differ-

ences to be magnified thus revealing hitherto unseen image data.

To ensure ease of use, the following features are incorporated: touch-screen selection of commands, magnetic card identification of the subscriber or user, icon based menus or selection buttons on the CRT display and split image display screen overlays

It will thus be seen that the objects set forth above, among those made apparent from the preceding description are efficiently attained and since certain changes may be made in the above construction without departing from the scope of the invention, it is intended that all matter contained in the above description or shown in the accompanying drawing shall be interpreted as illustrative and not in a limiting sense.

It is also to be understood that the following claims are intended to cover all of the generic and specific features of the invention herein described, and all statements of the scope of the invention which, as a matter of language, might be said to fall therebetween.

Now that the invention has been described,

What is claimed is:

1. A medical image storage, retrieval and transmission system providing simultaneous automated access to an image data base by a plurality of remote subscribers upon request over a communications network, said system comprising

image digitizing means forming a first digitized representation of said image,

first data compression means generating from said first digitized representation a low-loss second digitized representation of said image,

image data storage and retrieval means comprising means to receive and store said second digitized representation and to selectively provide said second digitized representation to data channel compression means compressing said second digitized representation to form a third digitized representation, and

telecommunication means including means to selectively transmit said requested third digitized representation to a requesting remote visual display terminal, said telecommunication means initially telecommunicating a first portion of said third digital representation, said remote terminal including means to convert said first portion of said third representation to a visual image having an initial resolution less than a resolution limit of said requesting terminal, said telecommunication means subsequently telecommunicating a second portion of said third digital representation, said second portion usable with said first portion to form an image having a resolution intermediate said initial resolution and said resolution limit.

2. A system of claim 1 wherein said first data compression means includes logic means to generate a run-length compressed digitized image data signal as said second representation.

3. A system of claim 1 wherein said first data compression means is operatively coupled to an external data storage drive to store said second representation.

4. A system of claim 1 wherein said image data storage and retrieval means comprises a data modem coupled to said image scanning and digitizing means, a write drive operatively coupled to said data modem to receive and to store said second digitized representation, and a plurality of data retrieval and transmission channels, each said channel comprising an image data reader means operatively coupled to said telecommunication means to selectively receive said second digitized representation of said image for transmission to a said requesting remote visual display terminal.

5. A system of claim 4, wherein a said image data reader means is operatively coupled to an external data write drive configured to receive a storage medium and to store compressed digital image data thereon.

6. A system of claim 4 wherein said telecommunication means comprises a control computer operatively coupled to said image data storage and retrieval means to selectively control data flow between said image data storage and retrieval means and said remote visual display terminal and a plurality of data compression channels coupled to said control computer, wherein each said data compression channel comprises a data memory including means to decompress said low-loss second representation of said image data received from said data retrieval and transmission channel and a compression means including logic means to compress and decompress second representation of said image data to form said third digitized representation of said image for transmission over said communication network to a said requesting remote visual display terminal.

7. A system of claim 1 wherein said first portion of said third representation of said image comprises a plurality of super pixels, and said second portion of said third representation comprises data representative of exact gray levels of a first subset of said super pixels and wherein a third portion of said third representation comprises similar data for a third subset of said super pixels.

8. A system of claim 1 wherein said remote terminal further includes means to select a region of a said image and said telecommunication means includes means to transmit a third portion of said third digitized representation, said third portion specific to said selected region, thereby providing an expanded visual display of said selected region, said expanded visual display containing more pixels than were included in said selected region.

9. A system of claim 1 wherein said remote visual display terminal further includes logic means to enhance an edge contrast of a displayed image.

10. A system of claim 1 wherein said remote visual display terminal further includes logic means to enhance gray level contrast by means of gray level region expansion.

11. A system of claim 1 wherein said remote visual display terminal further includes logic means for differential gray level tracking and gray level enhancement.

12. A system of claim 1 wherein individual patient information corresponding to said image is read by an optical character reader for compression and transmission with a said corresponding second digital representation to said image data storage and retrieval means.

* * * * *

UNITED STATES PATENT AND TRADEMARK OFFICE
CERTIFICATE OF CORRECTION

PATENT NO. : 5,321,520
DATED : June 14, 1994
INVENTOR(S) : Jorge J. Inga, et al

It is certified that error appears in the above-identified patent and that said Letters Patent is hereby corrected as shown below:

Column 20, lines 27 and 28, "and decompress" should read --said decompressed--.

Signed and Sealed this
Twenty-fifth Day of October, 1994

Attest:



BRUCE LEHMAN

Attesting Officer

Commissioner of Patents and Trademarks

(12) **EX PARTE REEXAMINATION CERTIFICATE** (8039th)
United States Patent
Inga et al.

(10) **Number:** US 5,321,520 C1
(45) **Certificate Issued:** Feb. 22, 2011

(54) **AUTOMATED HIGH DEFINITION/RESOLUTION IMAGE STORAGE, RETRIEVAL AND TRANSMISSION SYSTEM**

4,903,317 A 2/1990 Nishihara et al.
4,941,190 A 7/1990 Joyce

OTHER PUBLICATIONS

(75) Inventors: **Jorge J. Inga**, Tampa, FL (US); **Thomas V. Saliga**, Tampa, FL (US)

Sharaf E. Elnahas, Progressive Coding and Transmission of Digital Diagnostic Pictures, MI-5 IEEE Transactions on Med. Imaging 73 (Jun. 1986).*

(73) Assignee: **Automated Medical Access Corporation**, Tampa, FL (US)

* cited by examiner

Reexamination Request:
No. 90/011,260, Sep. 30, 2010

Primary Examiner—Henry N Tran

Reexamination Certificate for:
Patent No.: 5,321,520
Issued: Jun. 14, 1994
Appl. No.: 07/915,298
Filed: Jul. 20, 1992

(57) **ABSTRACT**

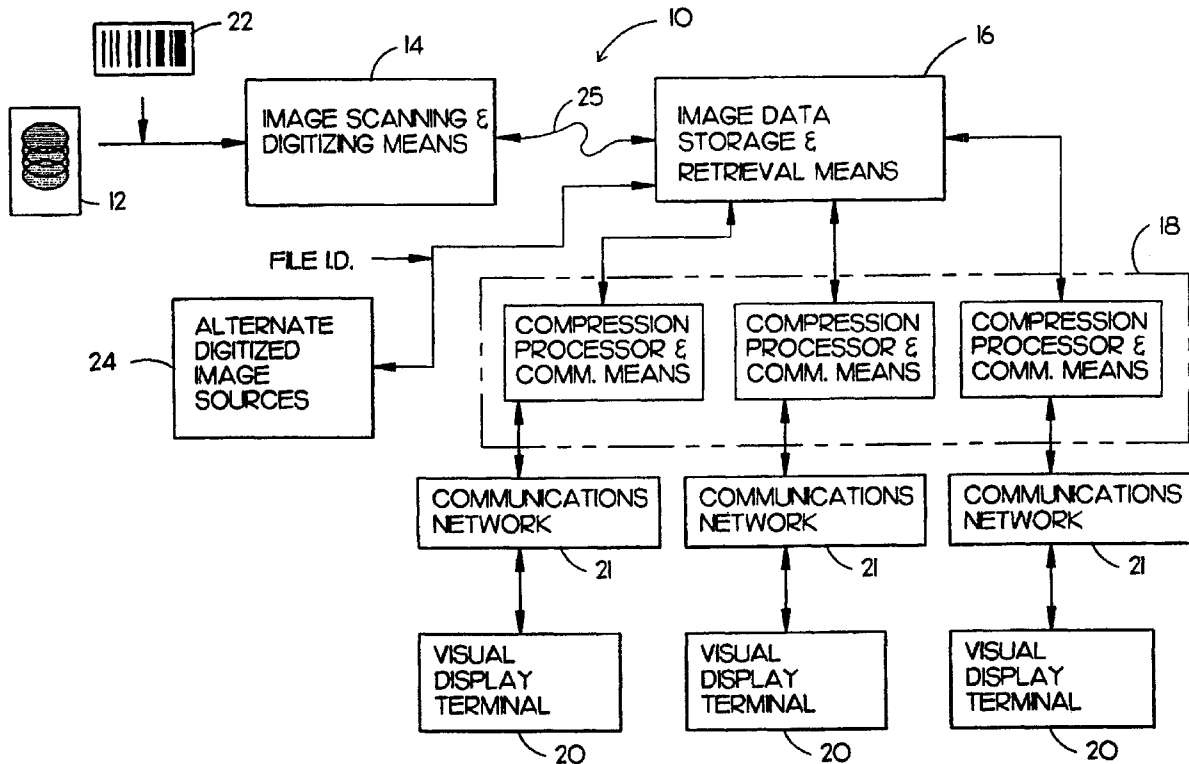
An automated high definition/resolution image storage, retrieval and transmission system for use with medical X-ray film or other documents to provide simultaneous automated access to a common data base by a plurality of remote subscribers upon request, the automated high definition/resolution image storage, retrieval and transmission system comprising an image scanning and digitizing subsystem to scan and digitize visual image information from an image film or the like; an image data storage and retrieval subsystem to receive and store the digitized information and to selectively provide the digitized information upon request from a remote site, a telecommunication subsystem to selectively transmit the requested digitized information from the image data storage and retrieval subsystem to the requesting remote visual display terminal for conversion to a visual image at the remote site to visually display the requested information from the image data storage and retrieval subsystem.

Certificate of Correction issued Oct. 25, 1994.

(51) **Int. Cl.**
H04N 1/21 (2006.01)
H04N 1/41 (2006.01)

(52) **U.S. Cl.** 358/403; 358/1.9
(58) **Field of Classification Search** 358/403
See application file for complete search history.

(56) **References Cited**
U.S. PATENT DOCUMENTS
4,520,671 A 6/1985 Hardin



US 5,321,520 C1

1
EX PARTE
REEXAMINATION CERTIFICATE
ISSUED UNDER 35 U.S.C. 307

NO AMENDMENTS HAVE BEEN MADE TO
THE PATENT

2
AS A RESULT OF REEXAMINATION, IT HAS BEEN
DETERMINED THAT:

5 The patentability of claims **1-12** is confirmed.

* * * * *