Self-Describing Schemes for Interoperable MPEG-7 Multimedia Content Descriptions

Seungyup Paek, Ana B. Benitez, and Shih-Fu Chang[1]

Image & Advanced TV Lab, Department of Electrical Engineering
Columbia University, 1312 S.W. Mudd, Mail code 4712 Box F-4
New York, NY 10027, USA

ABSTRACT

In this paper, we present the self-describing schemes for interoperable image/video content descriptions, which are being developed as part of our proposal to the MPEG-7 standard. MPEG-7 aims to standardize content descriptions for multimedia data. The objective of this standard is to facilitate content-focused applications like multimedia searching, filtering, browsing, and summarization. To ensure maximum interoperability and flexibility, our descriptions are defined using the eXtensible Markup Language (XML), developed by the World Wide Web Consortium. We demonstrate the feasibility and efficiency of our self-describing schemes in our MPEG-7 testbed. First, we show how our scheme can accommodate image and video descriptions that are generated by a wide variety of systems. Then, we present two systems being developed that are enabled and enhanced by the proposed approach for multimedia content descriptions. The first system is an intelligent search engine with an associated expressive query interface. The second system is a new version of MetaSEEk, a metasearch system for mediation among multiple search engines for audio-visual information.

Keywords: MPEG-7, self-describing scheme, interoperability, audio-visual content description, visual information system, metasearch, XML.

1. INTRODUCTION

It is increasingly easier to access digital multimedia information. Correspondingly, it has become increasingly important to develop systems that process, filter, search and organize this information, so that useful knowledge can be derived from the exploding mass of information that is becoming accessible. To enable exciting new systems for processing, searching, filtering and organizing multimedia information, it has become clear that an interoperable method of describing multimedia content is necessary. This is the objective of the emerging MPEG-7 standardization effort.

In this paper, we first give a brief overview of the objectives of the MPEG-7 standard. MPEG-7 aims at the standardization of content descriptions of multimedia data. The objectives of this standard are to facilitate content-focused applications like multimedia searching, filtering, browsing, and summarization.

Then, we present self-describing schemes for interoperable image/video content descriptions, which are being developed as part of our proposal to MPEG-7. To ensure maximum interoperability and flexibility, our descriptions use the eXtensible Markup Language (XML), developed by the World Wide Web Consortium. Under the proposed self-describing schemes, an image is represented as a set of relevant objects that are organized in one or more object hierarchies. Similarly, a video is viewed as a set of relevant events that can be combined hierarchically in one ore more event hierarchies. Both, objects and events, are described by some feature descriptors that can link to external extraction and similarity code.

Finally, we demonstrate the feasibility and efficiency of our self-describing schemes in our MPEG-7 testbed. In our testbed, we will show how our scheme can accommodate image and video descriptions that are generated by a wide variety of systems. In addition, we introduce two systems being developed, which are enabled and enhanced by our approach for multimedia content descriptions. The first system is an intelligent search engine with an associated expressive query interface. The second system is a metasearch system for mediation among multiple search engines for audio-visual information.

1. Email: {syp, ana, sfchang}@ee.columbia.edu; WWW: http://www.ee.columbia.edu/~{syp, ana, sfchang}/

2. MPEG-7 STANDARD AND SCENARIOS

2.1. MPEG-7 standard

The MPEG-7 standard [14] has the objective of specifying a standard set of descriptors to describe various types of multimedia information. MPEG-7 will also standardize ways to define other descriptors as well as Description Schemes (DSs) for the structure of descriptors and their relationships. This description (i.e. the combination of descriptors and description schemes) will be associated with the content itself to allow fast and efficient searching for material of a user's interest. MPEG-7 will also standardize a language to specify description schemes, i.e. a Description Definition Language (DDL), and the schemes for encoding the descriptions of multimedia content.

2.2. MPEG-7 scenarios

MPEG-7 will improve existing applications and enable completely new ones. We will review three of the most relevantly impacted application scenarios [3]: distributed processing, exchange, and personalized viewing of multimedia content.

- **Distributed processing**

MPEG-7 will provide the ability to interchange descriptions of audio-visual material independently of any platform, any vendor, and any application, which will enable the distributed processing of multimedia content. This standard for interoperable content descriptions will mean that data from a variety of sources can be plugged into a variety of distributed applications such as multimedia processors, editors, retrieval systems, filtering agents, etc. Some of these applications may be provided by third parties, generating a sub-industry of providers of multimedia tools that can work with the standard descriptions of the multimedia data.

The vision of the near future is one in which a user can access various content providers' web sites to download content and associated indexing data, obtained by some low-level or high-level processing. The user can then proceed to access several tool providers' web sites to download tools (e.g. Java applets) to manipulate the heterogeneous data descriptions in particular ways, according to the user's personal interests. An example of such a multimedia tool will be a video editor. A MPEG-7 compliant video editor will be able to manipulate and process video content from a variety of sources if the description associated with each video is MPEG-7 compliant. Each video may come with varying degrees of description detail such as camera motion, scene cuts, annotations, and object segmentations.

- **Content exchange**

A second scenario that will greatly benefit from an interoperable content-description standard is the exchange of multimedia content among heterogeneous audio-visual databases. MPEG-7 will provide the means to express, exchange, translate, and reuse existing descriptions of audio-visual material.

Currently, TV broadcasters, radio broadcasters, and other content providers manage and store an enormous amount of audio-visual material. This material is currently described manually using textual information and proprietary databases. Describing audio-visual material is an expensive and time-consuming task, so it is desirable to minimize the re-indexing of data that has been processed before.

Consider a media company that purchases videos from a TV broadcaster. The TV broadcaster has already described and indexed the content in their proprietary description scheme. Without an interoperable content description, the purchasing company will have to invest manpower to translate manually the description of the broadcaster into their proprietary scheme. Interchange of multimedia content descriptions would be possible if all the content providers embraced the same scheme and system. As this is unlikely to happen, MPEG-7 proposes to adopt a single industry-wide interoperable interchange format that is system and vendor independent.

- **Customized views**

Finally, multimedia playera and viewera compliant with the multimedia description standard will provide the users with innovative capabilities such as multiple views of the data configured by the user. The user could change the display's configuration without requiring the data to be downloaded again in a different format from the content broadcaster.

The ability to capture and transmit semantic and structural annotations of the audio-visual data, made possible by MPEG-7, greatly expands the range of possibilities for client-side manipulation of the data for displaying purposes. For example, a browsing system can allow users to quickly browse through videos if they receive information about their corresponding semantic structure. For example, when modeling a tennis match video, the viewer can choose to view only the third game of the second set, all the overhead smashes made by one player, etc.

These examples only hint at the possible uses that creative multimedia-application designers will find for richly structured data delivered in a standardized way based on MPEG-7.

3. SELF-DESCRIBING SCHEMES

In this section, we present description schemes for interoperable image/video content descriptions. The proposed description schemes are self-describing in the sense that they combine the data and the structure of the data in the same format. The advantages of such a type of descriptions are flexibility, easy validation, and efficient exchange.

3.1. eXtensible Markup Language (XML)

SGML (Standard Generalized Markup Language, ISO 8879) is a standard language for defining and using document formats. SGML allows documents to be self-describing, i.e. they describe their own grammar by specifying the tag set used in the document and the structural relationships that those tags represent. SGML makes it possible to define your own formats for your own documents, to handle large and complex documents, and to manage large information repositories. However, full SGML contains many optional features that are not needed for Web applications and has proven to be too complex to current vendors of Web browsers.

The World Wide Web Consortium (W3C) has created an SGML Working Group to build a set of specifications to make it easy and straightforward to use the beneficial features of SGML on the Web [21]. The goal of the W3C SGML activity is to enable the delivery of self-describing data structures of arbitrary depth and complexity to applications that require such structures. The first phase of this effort is the specification of a simplified subset of SGML specially designed for Web applications. This subset, called XML (Extensible Markup Language), retains the key SGML advantages in a language that is designed to be vastly easier to learn, use, and implement than full SGML.

Before describing the image and video DSs, we present some of the core features of XML. Let's start with a simple XML element:

```
<image> Hello MPEG-7 world! </image>
```

<image> is the start tag and </image> is the end tag; Hello MPEG-7 world! is the content of the element. What does the image tag mean? In short, it means anything you want it to mean. XML predefines no tags at all. Rather than relying on a few hundred predefined tags, XML lets you create the tags you need to describe your data. Users define what is allowed in each document by providing rules, collectively known as the Document Type Definition (DTD). The DTD states the element types with their characteristics, the notations, and the entities allowed in the document. Apart from the DTD, XML documents must follow some basic well-form rules. This is the minimum criterion for XML parsers and processors.

Text in XML documents consists of characters. A document's text is divided into character data and markup. In a first approximation, markup describes a document's logical structure while character data is the basic content of the document. Generally, anything inside a pair of <> angle brackets is markup and anything that is not inside these brackets is character data. Start tags and empty tags may optionally contain attributes. An attribute is a name-value pair separated by an equal sign. Work is in progress to include binary data in XML tags. Currently, XML allows defining binary entities pointing to binary data (e.g. images). They require an associated notation describing the type of resource (e.g. GIF and JPG).

3.2. Image description scheme

In this section, we present the proposed description scheme for images. To clarify the explanation, we will use the example shown in Figure 1. Using this example, we will walk through the image DS expressed in XML. Along the way, we will explain the use of various XML elements that are defined for the proposed image DS. The complete set of rules of the tags in the image and video description schemes is defined in our document type definitions [15]. Another advantage of using XML as the DLL is that it provides the capability to import external description schemes' DTDs to incorporate them in one description by using namespaces. We will see an example later in this section.

The basic description element of our image description scheme is the object element (<object>). An object element represents a region of the image for which some features are available. There are two different types of objects: physical and logical objects. Physical objects usually correspond to continuous regions of the image with some descriptors in common (semantics, features, etc.) - in other words, real objects in the image. Logical objects are groupings of objects based on some high-level semantic relationships (e.g. faces). The object element comprises the concepts of group of objects, objects, and regions in the visual literature. The set of all objects identified in an image is included within the object set element (<object_set>).

For the image example of Figure 1.a, we have chosen to describe the objects listed below. Each object element has a unique identifier within an image description. The identifier is expressed as an attribute of the object element (id). Another attribute of the object element (type) distinguishes between physical and logical objects. We have left the content of each object element empty to show clearly the overall structure of the image description. Later in the section, we will describe the features that can be included within the object element.

```
<object_set>
    <object id="0" type="PHYSICAL" > </object> <!-- Family portrait -->
    <object id="1" type="PHYSICAL" > </object> <!-- Father -->
    <object id="2" type="PHYSICAL" > </object> <!-- Mother -->
    <object id="3" type="LOGICAL" > </object> <!-- Faces -->
    <object id="4" type="PHYSICAL" > </object> <!-- Father's face -->
    <object id="5" type="PHYSICAL" > </object> <!-- Mother's face -->
</object_set>
```

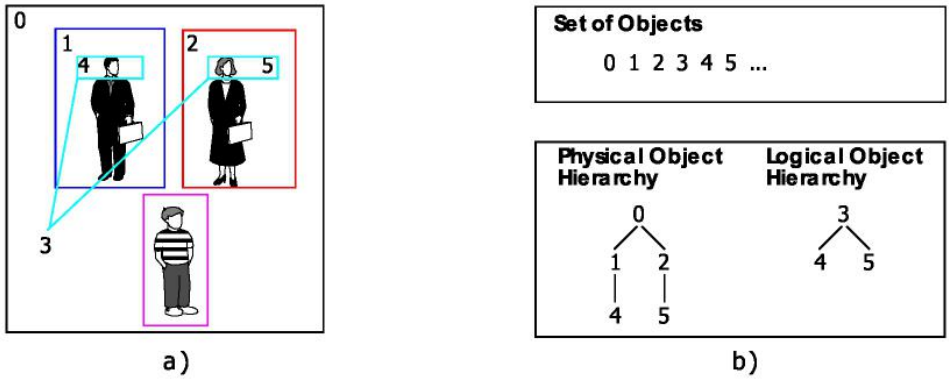

Figure 1: a) Image example. b) High-level description of the image by proposed image description scheme.

The image description scheme is comprised of object elements that are combined hierarchically in one or more object hierarchy elements (<object_hierarchy>). The hierarchy is a way to organize the object elements in the object set element. Each object hierarchy consists of a tree of object node elements (<object_node>). Each object node points to an object. The objects in an image can be organized by their location in the image or by their semantic relationships. These two ways to group objects generate two types of hierarchies: physical and logical hierarchies. A physical hierarchy describes the physical location of the objects in the image. On the other hand, a logical hierarchy organizes the objects based on a higher level understanding of their semantics, similar to semantic clustering.

Continuing with the image example in Figure 1.a, two possible hierarchies are shown in Figure 1.b. These hierarchies are expressed in XML below. The type of hierarchy is included in the object hierarchy element as an attribute (type). The object node element has associated a unique identifier in the form of an attribute (id). The object node element references an object element by using the latter's unique identifier. The reference to the object element is included as an attribute (object_ref). An object element can include links back to nodes in the object hierarchy as an attribute too (object_node_ref).

```
<object_hierarchy type="PHYSICAL"> <!-- Physical hierarchy -->
    <object_node id="10" object_ref="0"> <!-- Portrait -->
        <object_node id="11" object_ref="1"> <!-- Father -->
            <object_node id="12" object_ref="4"/> <!-- Father's face -->
        </object_node>
        <object_node id="13" object_ref="2"> <!-- Mother -->
            <object_node id="14" object_ref="5"/> <!-- Mother's face -->
        </object_node>
    </object_node>
</object_hierarchy>
<object_hierarchy type="LOGICAL"> <!-- Logical hierarchy: faces in the image -->
    <object_node id="15" object_ref="3"> <!-- Faces -->
```

```
        <object_node id="16" object_ref="4"/> <!– Father's face –>
        <object_node id="17" object_ref="5"/> <!– Mother's face –>
    </object_node>
</object_hierarchy>
```

An object set element and one or more object hierarchy elements form the image element (<image>). The image element symbolizes the image or picture being described.

In our image description scheme, the object element contains the feature elements; they include location, color, texture, shape, size, motion, time, and annotation elements, among others. Time and motion descriptors with have sense when the object belongs to a video sequence. The location element contains pointers to the locations of the image. Note that annotations can be textual, visual or audio. These features can be extracted or assigned automatically or manually. For those features extracted automatically, the feature descriptors can include links to external extraction and similarity matching code. An example is included below. This example also shows how external DSs can be imported and combined with ours.

```
<object id="4" type="PHYSICAL" object_node_ref="12 16"> <!– Father's face –>
    <color> </color>
    <texture>
        <tamura>
            <tamura_value coarseness="0.01" contrast="0.39" orientation="0.7"/>
            <code type="EXTRACTION" language="JAVA" version="1.2"> <!– Link to extraction code –>
                <location> <location_site href="ftp://extraction.tamura.java"/> </location>
            </code>
        </tamura>
    </texture>
    <shape> </shape>
    <position> </position>
    <!– Import and use of external annotation DS's DTD –>
    <text_annotation xmlns:extAnDS="http://www.other.ds/annotations.dtd">
        <extAnDS:Class>Face</extAnDS:Class>
     </text_annotation>
</object>
```

In summary, both, the object hierarchy and object set elements, are part of the image element (<image>). The objects in the object set are combined hierarchically in one or more object hierarchy elements. For efficient transversal of the image description, links are provided to traverse from objects in the object set to corresponding object nodes in the object hierarchy and viceversa. The objects include various feature descriptors that can link to external extraction and similarity matching code.

3.3. Video description scheme

In this section, we present the proposed description scheme (DS) for videos. To clarify the explanation, we will use the example shown in Figure 2. Using this example, we will walk through the video DS expressed in XML. Along the way, we will explain the use of various XML elements that are defined for the proposed MPEG-7 video DS. The structure of the image description scheme and the video description scheme are very similar.

The basic description element of our video description scheme is the event element (<event>). An event represents one or more shots of the video for which some features are available. We distinguish three different types of events: a shot, a continuous group of shots, and a discontinuous group of shots. Discontinuous group of shots will usually be associated together based on common features (e.g. background color) or high-level semantic relationships (e.g. actor on screen). The event element comprises the concepts of story, scene, and shot in the visual literature. The set of all events identified in a video is included within the event set element (<event_set>).

For the video example of Figure 2.b, we have chosen to describe the events listed below. Each event element has a unique identifier within a video description. The identifier is expressed as an attribute of the event element (id). Another attribute of the event element (type) distinguishes between the three different types of events. We have left each event element empty to show clearly the overall structure of the video description. Later in the section, we will describe the features that can be included within the event element.

```
<event_set>
    <event id="0" type="SHOT" > </event> <!– The tiger –>
    <event id="1" type="SHOT" > </event> <!– Stalking the prey –>
```

DOCKET ALARM

Explore Litigation Insights

Docket Alarm provides insights to develop a more informed litigation strategy and the peace of mind of knowing you're on top of things.

Real-Time Litigation Alerts

Keep your litigation team up-to-date with **real-time alerts** and advanced team management tools built for the enterprise, all while greatly reducing PACER spend.

Our comprehensive service means we can handle Federal, State, and Administrative courts across the country.

Advanced Docket Research

With over 230 million records, Docket Alarm's cloud-native docket research platform finds what other services can't. Coverage includes Federal, State, plus PTAB, TTAB, ITC and NLRB decisions, all in one place.

Identify arguments that have been successful in the past with full text, pinpoint searching. Link to case law cited within any court document via Fastcase.

Analytics At Your Fingertips

Learn what happened the last time a particular judge, opposing counsel or company faced cases similar to yours.

Advanced out-of-the-box PTAB and TTAB analytics are always at your fingertips.

API

Docket Alarm offers a powerful API (application programming interface) to developers that want to integrate case filings into their apps.

LAW FIRMS
Build custom dashboards for your attorneys and clients with live data direct from the court.

Automate many repetitive legal tasks like conflict checks, document management, and marketing.

FINANCIAL INSTITUTIONS
Litigation and bankruptcy checks for companies and debtors.

E-DISCOVERY AND LEGAL VENDORS
Sync your system to PACER to automate legal marketing.