Interactive 3-D Video Representation and Coding Technologies

ALJOSCHA SMOLIĆ AND PETER KAUFF

Invited Paper

Interactivity in the sense of being able to explore and navigate audio–visual scenes by freely choosing viewpoint and viewing direction, is an important key feature of new and emerging audio–visual media. This paper gives an overview of suitable technology for such applications, with a focus on international standards, which are beneficial for consumers, service providers, and manufacturers. We first give a general classification and overview of interactive scene representation formats as commonly used in computer graphics literature. Then, we describe popular standard formats for interactive three–dimensional (3-D) scene representation and creation of virtual environments, the virtual reality modeling language (VRML), and the MPEG-4 BInary Format for Scenes (BIFS) with some examples. Recent extensions to MPEG-4 BIFS, the Animation Framework eXtension (AFX), providing advanced computer graphics tools, are explained and illustrated. New technologies mainly targeted at reconstruction, modeling, and representation of dynamic real world scenes are further studied. The user shall be able to navigate photorealistic scenes within certain restrictions, which can be roughly defined as 3-D video. Omnidirectional video is an extension of the planar two–dimensional (2-D) image plane to a spherical or cylindrical image plane. Any 2-D view in any direction can be rendered from this overall recording to give the user the impression of looking around. In interactive stereo two views, one for each eye, are synthesized to provide the user with an adequate depth cue of the observed scene. Head motion parallax viewing can be supported in a certain operating range if sufficient depth or disparity data are delivered with the video data. In free viewpoint video, a dynamic scene is captured by a number of cameras. The input data are transformed into a special data representation that enables interactive navigation through the dynamic scene environment.

Keywords*—*Interactive media, MPEG standards, three–dimensional (3-D) video, video coding.

I. Introduction

Interactivity is an important key feature of new and emerging audio–visual media where the user has the opportunity to be active in some way instead of just being a passive consumer. One kind of interactivity is the ability to freely choose viewpoint and viewing direction within an audio–visual scene. The first media that provided such functionality were based on textured three–dimensional (3-D) mesh models, as they are well known from computer graphics. For representation and exchange of such data, the ISO/IEC has standardized a special language called virtual reality modeling language (VRML) that is widely used in the web. However, although VRML still holds merit, it is going to get dated due to the rapid progress in multimedia, in general, and virtual reality, in particular. VRML is invariably graphics-oriented and scene realism, therefore, is limited. Most of the scenes are either purely computer generated or contain static two–dimensional (2-D) views of real word objects represented by still pictures or moving textures.

Hence, as a complement to VRML, new standardization efforts have been launched to advance realism and functionality of 3-D representations in interactive audio–visual media. Most of them, such as the Web3D consortium, have collaborated closely with the moving picture experts group (MPEG) group of ISO/IEC. The reason for this liaison is that the most recent standard MPEG-4 is not only focused on audio–visual coding; indeed, it is much more multimedia-oriented than MPEG-1/2 and offers plenty of new functionalities like interactivity and 3-D scene representation.

Even in its basic version, MPEG-4 already supports interactivity with hybrid scenes containing both computer graphics and natural video objects. Furthermore, to be compatible with VRML, it includes all conventional tools and scene description formats for interactive 3-D graphics as a subset. In addition, it allows an easy integration of multiple audio and video streams compliantly coded with existing ISO or ITU standards. And the MPEG-4 scene description language called BInary Format for Scenes (BIFS) specifies a compressed binary format, which is suited for online transmission and streaming of 3-D scenes. The more recent

Manuscript received December 19, 2003; revised June 9, 2004.

The authors are with the Fraunhofer Institute for Telecommunications, Image Processing Department, Heinrich-Hertz-Institute (HHI) 10587 Berlin, Germany (e-mail: smolic@hhi.de; kauff@hhi.de).

Digital Object Identifier 10.1109/JPROC.2004.839608

0018-9219/$20.00 © 2005 IEEE

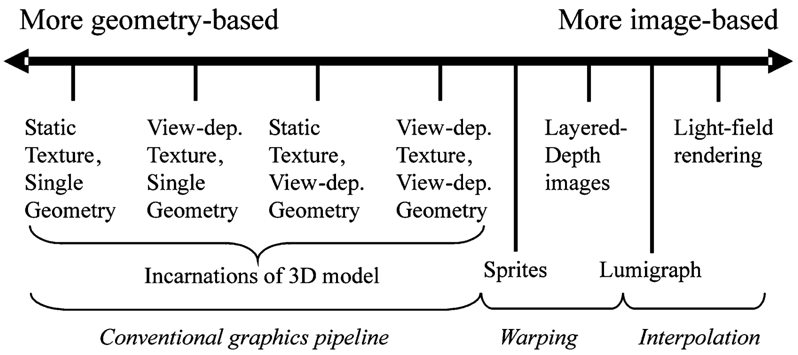
More geometry-based
More image-based
Static Texture, Single Geometry
View-dep. Texture, Single Geometry
Static Texture, View-dep. Geometry
View-dep. Texture, View-dep. Geometry
Layered-Depth images
Light-field rendering
Incarnations of 3D model
Sprites
Lumigraph
Conventional graphics pipeline
Warping
Interpolation

Fig. 1. Categorization of scene representations [27].

versions of the MPEG-4 standard support further extensions toward advanced 3-D representation. One example is the Animation Framework eXtension (AFX) that provides new 3-D formats for natural-looking scene objects. In this context, AFX also offers new tools such as surface light fields and depth image-based rendering for efficient 3-D modeling of scene objects captured by real imagery.

Further approaches for modeling and rendering real world scenes are investigated in still ongoing 3-D integration activities of a new MPEG group, 3-D audio–visual (3DAV). This includes video-based rendering as well as advanced 3-D reconstruction methods. Application scenarios investigated in this context are omnidirectional video, interactive stereo video, and free viewpoint video.

Omnidirectional video is an extension of the planar 2-D image plane to a spherical or cylindrical image plane. Other kinds of planes (e.g., hyperbolic) are also possible. Video is captured at a certain viewpoint (which may move over time) in multiple directions. Any 2-D view in any direction can be rendered from this overall recording. Such an omnidirectional video can be displayed with a suitable player. Key functions of interaction are zoom and rotation of the viewing direction to give the user the impression of looking around.

In interactive stereo two views, one for each eye, are synthesized to provide the user with an adequate depth cue of the observed scene. Head motion parallax viewing can be supported if sufficient depth or disparity data are delivered with the video data. In that case, it is possible to generate stereoscopic virtual views that correspond to various head positions around a given zero position. This is an important feature of envisaged immersive media, such as 3-D-TV, that seem to be applicable in the near future.

The most general case is free viewpoint video. A dynamic scene is captured by a number of cameras. In addition to the video signals, other information such as camera calibration and derived scene geometry is also acquired or estimated. These input data are transformed into a special data representation that enables interactive navigation through the dynamic scene environment. It is also possible to generate 3-D video objects that are composed from multiple view video information. Free viewpoint video representation formats can either be purely image-based or rely on a certain kind of 3-D reconstruction.

This paper gives an overview of available and emerging technology for modeling, coding, and rendering of dynamic real world scenes for interactive applications, at the convergence point of computer graphics, computer vision, and classical media. The main focus is on such formats that are or will probably be available in open standards such as ISO/IEC MPEG. A general classification of interactive 3-D scene representation approaches is given in the next section. Section III describes the scene description language MPEG-4 BIFS in more detail. A few examples are given and the link to the common computer graphics format VRML is explained. Following these fundamentals, the recently established AFX extension of MPEG-4 is presented in Section IV. Then, Section V gives an overview of the new technology that is under investigation in the working group 3DAV of MPEG. Finally, Section VI summarizes the paper and gives an outlook to future developments in these fields.

II. Classification of Interactive 3-D Scene Representation Formats

In computer graphics literature, methods for scene representation are often classified as a continuum in between two extremes as illustrated in Fig. 1 [27]. The one extreme is represented by classical 3-D computer graphics. This approach can also be called geometry-based modeling. In most cases, scene geometry is described on the basis of 3-D meshes. Real world objects are reproduced using geometric 3-D surfaces with an associated texture mapped onto them. More sophisticated attributes can be assigned as well. For instance, appearance properties (opacity, reflectance, specular lights, etc.) can enhance the realism of the models significantly.

Everyone is familiar with this type of computer graphics from games, Internet, TV, movies, etc. The achievable performance might be extremely good if the scenes are purely computer generated. The available technology for both production and rendering has been highly optimized over the last few years, especially in the case of common 3-D mesh representations. In addition, state-of-the-art PC graphics cards are able to render highly complex scenes with an impressive quality in terms of refresh rate, levels of detail, spatial resolution, reproduction of motion, and accuracy of textures.

A drawback to this approach is the high cost for content creation. Aiming at photorealism, 3-D scene and object modeling is complex and time consuming, and it becomes even more complex if a dynamically changing environment simulating real life is being created. Furthermore, an automatic 3-D object and scene reconstruction implies an estimation of camera geometry, depth structures, and 3-D shapes. Inherently, all these processes tend to produce occasional errors. Therefore high-quality production, e.g., for movies, has to be done user assisted, supervised by a skilled operator.

The other extreme is given by scene representations that do not use any 3-D geometry at all. It is usually called image-based modeling. In this case, virtual intermediate views are generated from available real views by interpolation. The main advantages are a high quality of virtual view synthesis and an avoidance of 3-D scene reconstruction. However, these benefits have to be paid by dense sampling of the real world with plenty of original view images. In general, the synthesis quality increases with the number of available views. Hence, a large amount of cameras has to be set up to achieve high-performance rendering, and plenty of image data needs to be processed therefore. To the contrary, if the number of used cameras is too low, interpolation and occlusion artifacts will appear in the synthesized images, possibly affecting the quality.

The image-based methods can be derived from the theory of the plenoptic function. The expression descends from the Latin root plenus, meaning complete or full and optic, pertaining to vision [37]. This 7-dimensional function has initially been postulated by Adelson and Bergen [1]:

$$p = P(V_x,\, V_x,\, V_x,\, \theta,\, \phi,\, \lambda,\, t). \tag{1}$$

It describes the intensity of every light ray at every position in space (V_x, V_y, V_z, 3-D), in every direction (θ, ϕ, 2-D), for every wavelength (λ, 1-D), at any time (t, 1-D). It represents everything that can be seen from all positions in space, into any direction, anytime. As such, it might be called the universal formula of vision, but it has only theoretical relevance. In practice, it is simplified by omitting dimensions, for example the wavelength, time, or some spatial dimension. Moreover, it is not possible to apply the plenoptic function in its continuous form, i.e., it has to be sampled. Views are taken at a number of discrete positions, probably into some discrete directions. Against this background it is possible to formulate a complete plenoptic sampling theory in analogy to the common sampling theorem of signal theory [4].

Examples of image-based representations are ray-space [12]–[16] or light-field rendering [32] and panoramic configurations including concentric and cylindrical mosaics [5], [37], [51], [55]. All these methods do not make any use of geometry, but they either have to cope with an enormous complexity in terms of data acquisition or they execute simplifications restricting the level of interactivity.

In between the two extremes there exists a continuum of methods that make more or less use of both approaches and combine the advantages in a particular manner. For instance, a Lumigraph [3], [18] uses a similar representation as a light field but adds a rough 3-D model. This provides information on the depth structure of the scene and therefore allows for reducing the number of views.

Other representations do not use explicit 3-D models but depth or disparity maps. Such maps assign a depth value to each pixel of an image. Together with the original 2-D image the depth map builds a 3-D-like representation, often called 2.5-D. This can be extended to layered depth images [50] where multiple color and depth values are stored in consecutively ordered depth layers (see Section IV).

Closer to the geometry-based end of the spectrum, we can find methods that use view-dependent geometry and/or view dependent texture [7], [43]. Surface light fields combine the idea of light fields with an explicit 3-D model [6], [57]. Furthermore, volumetric representations such as voxels (from volume elements) can be used instead of a complete 3-D mesh model to describe 3-D geometry [8], [29], [41], [42], [49].

The complete processing chain of such systems can be divided into the parts of acquisition/capturing, processing, scene representation, coding, transmission/streaming/storage, interactive rendering, and 3-D displays. The design has to take into account all parts, since there are strong interrelations between all of them. For instance, an interactive display that requires random access to 3-D data will affect the performance of a coding scheme that is based on data prediction. A complete system for efficient representation and interactive streaming of high-resolution panoramic views has been presented in [20]. Other coding and transmission aspects of such data have also been studied, for example in [16], [21], [28], [31], [33], [39], [40], [46], [52], [60], and [61]. The European IST project ATTEST has studied a complete processing chain for interactive 3-D broadcast including 3-D-TV acquisition, data representation, joint coding of video and depth maps, auto-stereoscopic 3-D displays, and parallax viewing based on head tracking [10].

III. MPEG-4 Binary Format for Scenes (BIFS)

Classical 3-D computer graphics representations using textured 3-D mesh models are widely spread over various applications. A lot of software tools and specialized hardware are available and standard APIs such as OpenGL, DirectX, or Java3D provide developers with easy access to state-of-the-art functionalities of graphics cards. Due to historical implementation reasons, many applications (e.g., games) use proprietary formats for data representation. However, for exchange of 3-D graphics data between different systems it is necessary to define standardized formats ensuring interoperability. For that purpose, ISO/IEC has specified VRML. It was mainly developed for transmission of 3-D graphics over the Internet but can as well be used for other applications. VRML data can be visualized with an appropriate player and the user can navigate through the virtual environments. Apart from the formats and attributes for object description (geometric primitives, 3-D meshes,

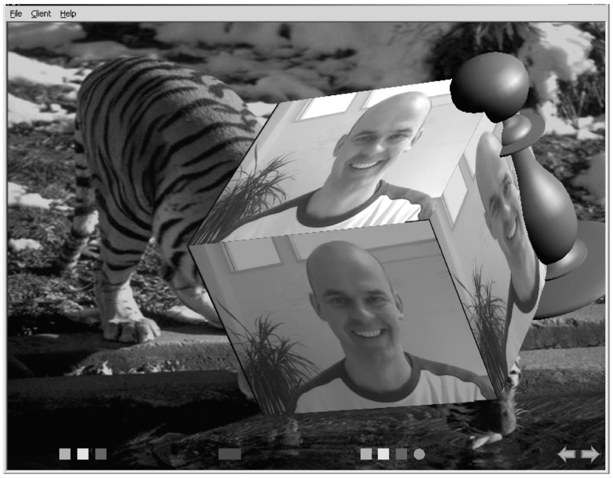

Fig. 2. Example of MPEG-4 BIFS scene.

textures, appearance, etc.), VRML also contains other elements necessary to define interactive 3-D worlds (e.g., light sources, collision, sensors, interpolators, viewpoint).

Later on, ISO/IEC issued a standard for multimedia data representation known as MPEG-4. It builds on BIFS, which is an extension of VRML. In addition to the functionality of VRML, BIFS provides, for instance, a better integration of natural audio and video, advanced 3-D audio features, a timing model, an update mechanism to modify the scene in time, a script to animate the scene temporally, new graphics elements (e.g., face and body animation), and an efficient binary encoding for the scene description. The last point is particularly important for streaming and broadcast applications, since scene description files in text format might be insufficiently large and not well suited for real-time streaming.

As such, MPEG-4 is much more than just another video and audio codec, although advanced audio–visual coding is again an important part of MPEG-4. In fact, it is a real multimedia standard, combining all types of media within a standardized format.

An example image of a rendered MPEG-4 BIFS scene is shown in Fig. 2. The background is a still image coded in JPEG format. The foreground scene contains 3-D graphics elements (box, pawn in the game) that the user can interact with (move, rotate, change of object properties like shape, size, texture, opacity, etc.). In addition, the interaction elements at the bottom allow a change of the image background and a browsing from scene to scene. Such scene changes can be downloaded on demand and the scene graph is updated online while viewing the scene. Furthermore, a live video stream showing a person is decoded and mapped onto the surface of the 3-D box. This simple example illustrates some basic features of MPEG-4 BIFS and its potential for content creators. The possibility of online scene updates and the live streaming of audio–visual data and their seamless integration into virtual worlds represent a clear progress over VRML.

A further example, which efficiently takes advantage of BIFS, is interactive streaming and rendering of high-resolution panoramic views [20]. Panoramic views are widely used on the Internet to provide complete views of real environments. Navigation is restricted to rotation and zoom. A panorama is a purely image-based representation as explained in Section II.

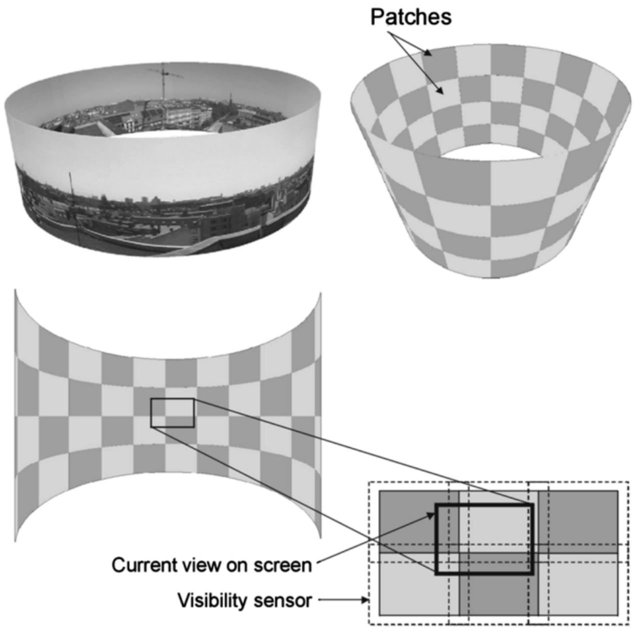

Fig. 3. BIFS representation for high-resolution panoramic views.

Basically, only a small portion of the panorama is displayed at a certain time with an interactive player (if the video is not projected as a whole as done, e.g., in a dome application). This can, for instance, be a 60° field of view. A rough estimate for a minimum resolution to get a good quality for each rendered view is 600 × 600 pixels. This means that a full spherical panorama would need a total resolution of at least 3600 × 1800 pixels. Transmitting such a large image (or even higher possible resolutions as reported in [20]) over the Internet prior to display causes unacceptable long download times. Furthermore, the usage of one large picture for the whole panorama is a heavy burden for the player as it has to keep the whole image in the memory for rendering purposes.

In this context, BIFS can be used for an efficient representation of interactive panoramas. Since the user only sees a portion of the panorama at a time, streaming and rendering can be limited to this particular image area. For this purpose, the panorama is divided into patches that can be decoded separately as illustrated in Fig. 3. In practice, these patches are JPEG coded images arranged around a 3-D cylinder using the BIFS syntax. Each of the patches is assigned to a visibility sensor that is slightly bigger than the patch. While the user navigates over the panorama, the actual visible patches are streamed, loaded, and rendered. This process is started as soon as the corresponding visibility sensor gets into the active window shown at the screen. A look ahead and prefetching strategy guarantees that all image data are available on time. This is simply achieved by oversizing the visibility sensor. Vice versa, a patch is unloaded if the corresponding sensor gets out of the actual view. In a streaming application, copies of the transmitted patches can be stored locally to avoid renewed transmission when revisiting same areas of the panorama. This procedure allows smooth rendering of even very high-resolution panoramas.

In a complete interactive multimedia application, the interactive video must be accompanied by corresponding interactive audio. If the user rotates the viewpoint, the associated sound should follow the interaction, by changing its direction of origin. If the user approaches a sound source, the associated sound should become louder. These functionalities are provided by MPEG-4 AudioBIFS [48], which provides the means for setting up 3-D audio scenes. Sound sources with various attributes and properties can be placed anywhere in a virtual 3-D space. With a suitable 3-D audio player the user can navigate arbitrarily within such a scene and corresponding audio is rendered for every position and orientation.

IV. MPEG-4 AFX

Computer graphics research has of course continued successfully since the initial version of MPEG-4 BIFS was finalized. Some of these developments had been integrated into an extension of MPEG-4 called AFX [25]. Two of the new tools are of specific interest in the scope of this paper: light-field mapping (LFM) and depth image-based representation (DIBR). The first one addresses the concept of surface light fields and the second one the concept of layered depth images.

As mentioned before, surface light fields combine a classical 3-D mesh representation with the concept of a light-field rendering. A light-field representation is used as texture that is mapped onto the 3-D mesh model. Such a 3-D model is typically built out of thousands of triangles that approximate the 3-D surface of an object.

Texture mapping means the assignment of a colored pixel map onto the 3-D surface. The simplest way is to assign a single still image as texture. However, this leads to poor rendering results. In reality, natural materials look different from changing view angles depending on reflectance properties, micro-structures, and lighting conditions. It is not possible to reproduce these properties by a single texture that looks the same from any direction. Therefore, conventional computer graphics employ sophisticated tools to model these material properties as best as is possible. The results might look fine for purely computer-generated objects. However, it is extremely difficult to set these parameters such that they precisely mirror the material properties of real world objects.

The promising solution to this problem is to incorporate ideas from image-based rendering. As explained in Section II, the idea of this method is to describe the real world by multiple view images instead of graphical 3-D models. As a consequence, view-dependent texture mapping assigns more than only one single texture to a triangle [7], [43]. Depending on the actual view direction, a realistic texture reproducing natural material appearance is calculated from the available original views. The same concept is exploited for a surface light field such as the LFM tool in AFX [6], [57]. The result is an extremely realistic rendering of static objects. However, data acquisition requires special equipment and a quite complex manual procedure of content creation. A reliable automatic generation of view-dependent textures for moving objects does not seem applicable for the near future. Therefore, applications to dynamic video objects are not realistic so far.

The AFX tool DIBR implements the concept of layered depth images [50]. In this case, a 3-D object or scene is represented by a number of views with associated depth maps as shown in Fig. 4 [2]. The depth maps define a depth value for every single pixel of the 2-D images. Together with appropriate scaling and information from camera calibration, it is possible to render virtual intermediate views as shown in the middle image in Fig. 4. The quality of the rendered views and the possible range of navigation depend on the number of original views and the setting of the cameras. A special case of this method is stereo-vision, where two views are generated accordingly to the geometry of the human eyes basis. In this case, the depth is often calculated using disparity estimation. Supposing that the capturing cameras are fully calibrated and their 3-D geometry is known, therefore, corresponding depth values can be recalculated one-to-one from the estimated disparity results. In the case of simple camera configurations (such as a conventional stereo rig or a multi baseline video system) this disparity estimation can even be used for fully automatic real-time depth reconstruction in 3-D video or 3-D-TV applications as explained in Section V-C.

V. MPEG Exploration on 3DAV

To this end, the considerations have mainly been concentrated on computer graphics with integrated 2-D video. Some of the concepts like panoramic views or DIBR can easily be extended toward 3-D video. In this context, the term 3-D video shall refer to interactive and navigable representations of real world dynamic scenes as captured by real imagery. A lot of research has been done in this field during the last few years resulting in different types of formats and technology for different types of application scenarios. To investigate the needs for standardization in this area, MPEG has established a working group called 3DAV [53].

Three main application scenarios have been extracted: omnidirectional video, interactive stereo video, and free-viewpoint video and related requirements for standardization have been derived [23]. Suitable technology for realization has been reviewed and evaluated experimentally [24]. The following section gives an overview of the results.

A. Common Issues

It has been identified that the definition of suitable quality measures is a common problem of all technology investigated in the context of 3DAV. For example, there are no original data to assess algorithms that compute intermediate views. Therefore, the definition of suitable quality measures is a critical task for all 3DAV technology. In many cases, only subjective criteria can be used.

Another common issue of technology that makes use of interpolated views, such as interactive stereo and free viewpoint video, is the need for accurate 3-D camera calibration information. A suitable data format, which is based on Tsai's fundamental work [56], has already been proposed in 3DAV.

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.

WHAT WILL YOU BUILD? | sales@docketalarm.com | 1-866-77-FASTCASE

fastcase®
Smarter legal research.