Video Indexing Based on Mosaic Representations

MICHAL IRANI, MEMBER, IEEE, AND P. ANANDAN, MEMBER, IEEE

Video is a rich source of information. It provides visual information about scenes. This information is implicitly buried inside the raw video data, however, and is provided with the cost of very high temporal redundancy. While the standard sequential form of video storage is adequate for viewing in a "movie mode," it fails to support rapid access to information of interest that is required in many of the emerging applications of video. This paper presents an approach for efficient access, use, and manipulation of video data. The video data are first transformed from their sequential and redundant frame-based representation, in which the information about the scene is distributed over many frames, to an explicit and compact scene-based representation, to which each frame can be directly related.

This compact reorganization of the video data supports nonlinear browsing and efficient indexing to provide rapid access directly to information of interest. This paper describes a new set of methods for indexing into the video sequence based on the scene-based representation. These indexing methods are based on geometric and dynamic information contained in the video. These methods complement the more traditional "content-based indexing" methods, which utilize image-appearance information (namely, color and texture properties) but are considerably simpler to achieve and are highly computationally efficient.

Keywords*—*Compact video representations, mosaics, video annotation, video browsing, video compression, video data bases, video indexing, video manipulation.

I. Introduction

The emergence of video as data and a source of information on the computer opens the potential for new ways of accessing, viewing, and manipulating the contents of video. These include direct nonlinear access to video frames and sequences of interest, new modes of viewing that give the viewer control over how the video is viewed, annotation and manipulation of objects and scenes in the video, and merging of text and graphics with the video data.

While the standard manner of representing video as a sequence of frames is adequate for viewing it in a movie mode, it does not support the type of interaction with video information described above. Currently, the only way to access the information of interest is by sequentially scanning the video. The only way to manipulate, annotate, or edit the video is by processing the video frame by frame. This process is both slow and tedious.

This paper presents a new approach for efficient access, storage, and manipulation of video data. Our approach is based on the fact that a video sequence contains many views of the same *scene* taken over time, from either a moving or a stationary camera. Hence, the information that is common to all the frames is the scene itself. This information is distributed over many frames, however, at the cost of very high temporal redundancy, and is found only implicitly in the video data. We transform the video data from a sequential *frame-based* representation, in which this common scene information is *distributed* over many frames, into a single common *scene-based* representation to which each frame can be *directly* related. This representation then allows *direct* and *immediate* access to the *scene* information, such as static locations and dynamically moving objects. It also eliminates the redundancy between the different views of the scene contained in the frames and results in a highly efficient and compact representation of the video information. Hence, the scene-based representation forms the basis for direct and efficient access to and manipulation of the video information and supports efficient storage and transmission of the video data.

The scene representation is composed of three components.

1) *Extended spatial information:* this captures the appearance of the entire scene imaged in the video clip and is represented in the form of a few (often just one) panoramic mosaic images constructed by composing the information from the different views of the scene in the individual frames into a single image.

2) *Extended temporal information:* this captures the motion of independently moving objects in the scene (e.g., in the form of their trajectories).

3) *Geometric information:* this captures the three-dimensional (3-D) scene structure, as well as the geometric transformations that are induced by the motion of the camera, and maps the frames to the common mosaic image.

Manuscript received July 20, 1997; revised November 30, 1997. The Guest Editor coordinating the review of this paper and approving it for publication was A. M. Tekalp.

M. Irani was with the Sarnoff Corporation, Princeton, NJ 08540 USA. She is now with the Department of Applied Math and Computer Science, The Weizmann Institute of Science, Rehovot 76100 Israel.

P. Anandan was with the Sarnoff Corporation, Princeton, NJ 08540 USA. He is now with Microsoft Corporation, Redmond, WA 98052 USA.

Publisher Item Identifier S 0018-9219(98)03562-2.

DOCKET ALARM
Find authenticated court documents without watermarks at docketalarm.com.

Taken together, these three components provide a *compact* description of the video data.

We construct the common scene-based representation by measuring and interpreting the image motion within the video clip. Regions of the video frames corresponding to the static and dynamic portions of the scene are determined. The geometric transformations and the 3-D scene structure are recovered as a part of this process. This process is done automatically, without any information about the camera calibration or the scene.

Once the common scene-based representation is constructed, it forms the basis for direct and efficient browsing, indexing, and manipulation of the video data. *Browsing* is done by skimming a collection of images that "summarize" the video data. We refer to these images as *visual summaries*. These summaries visually describe the video information in a compact and succinct fashion and can serve as a *visual table of contents* for the video.

Since the mosaics capture the information that is common to all the frames, they provide the means *directly* to index into and manipulate the individual frames. Both the static and dynamic portions of the video sequence can be accessed this way. These indexing methods are based on *geometric* and *dynamic* information contained in the video. These complement the more traditional approach to "content-based indexing," which utilizes image *appearance* information (namely, color and texture properties) [7], [9], [10], [26], but are considerably simpler to achieve and are computationally highly efficient. The existing appearance-based methods themselves can also be used more efficiently within the scene-based representation when applied directly to the mosaic image (i.e., to the appearance component of our representation), rather than to the individual video frames one by one.

The rest of this paper is organized as follows. Section II presents the common and compact scene-based representation, to which each frame is *directly* related. Section III explains how to use the scene-based representation to browse, index, and manipulate video data efficiently and rapidly. Section IV reviews the techniques used for constructing the scene-based representation from raw video sequences. Section V concludes this paper.

II. From Frames to Scenes

Video is a rich data source. It provides information about scenes. This information is buried inside the raw video data, however, and is provided at the cost of very high temporal redundancy (e.g., every scene point is displayed repeatedly in numerous consecutive frames). In this section, we first review the fundamental components of information in a video stream (Section II-A). Then we make use of these information components to *transform* the video from an implicit and redundant frame-based representation to an explicit and nonredundant scene-based representation that

A. The Three Fundamental Information Components of Video

Video extends the imaging capabilities of a still camera in three ways. First, although the field of view of each single image frame may be small, the camera can be panned or otherwise moved around in order to cover an extended spatial area. However, the *extended spatial information* acquired by the video is not available in a coherent form. It is distributed among a sequence of frames and is hard to use.

The second, and perhaps the most common, use of video is to record the evolution of events over time. Again, however, this *extended temporal information* is not explicitly represented but distributed over a sequence of video frames. While it is natural for a human to view it as a movie, this representation is not particularly suitable for analytic purposes.

Third, a video camera can be moved in order to acquire views from a continuously varying set of vantage points. This induces image motion, which depends on the 3-D geometric layout of the scene and the motion of the camera. However, this *geometric information* is also only implicitly present and is not directly accessible from the standard sequential video representation.

Thus, the total information contained in the video data consists of the three *scene* components mentioned above. However, this information is distributed among the frames and is implicitly encoded in terms of image motion. Therefore, a natural way to reorganize the video data is in terms of these three scene components. Moreover, such a reorganization removes the tremendous redundancy that is present in the source video data. This scene-based organization is highly efficient since it directly and *uniquely* maps onto the information in the scene. Therefore, it facilitates efficient interaction and manipulation and supports very efficient storage and transmission.

B. The Scene-Based Representation

To bring out the *common* scene information contained in the video, and make it more directly accessible, we first transform the video from its *implicit* and *redundant* frame-based representation to an *explicit* and *compact* scene-based representation. In this section, we introduce the scene-based representation. In Section IV, we elaborate on the details of the representation and explain how it is constructed from the video data.

The video stream is first *temporally* segmented into *scene segments*, which are subsequences of the input video sequence. A beginning or an end of a scene segment is automatically detected wherever a scene cut or scene change occurs in the video. The scene cuts are characterized typically by *drastic* changes in the frame content, which are directly reflected in the distribution of color and the gray levels in the image, or in the image motion (e.g., see [9] and [37]). These changes are relatively simple to detect.

Each scene segment is subsequently parsed into the three fundamental components of video (see Section II-A).

DOCKET ALARM
Find authenticated court documents without watermarks at docketalarm.com.

objects, and the geometric information. These components are organized as described below.

Corresponding to the three fundamental components, the scene-based representation is divided into three parts.

1) Panoramic Mosaic Image: This captures an extended spatial view of the entire scene visible in the video clip in a single (or sometimes a few) “snapshot” image(s) (e.g., see Fig. 1). This image captures the appearance of the *static* portions of the scene.

The mosaic image is constructed by first aligning all the frames with respect to the common coordinate system (which becomes also the mosaic coordinate system) and then integrating all these frames to form a single image. Different methods of integration can be employed (e.g., temporal average, temporal median, superresolution, etc.). These are described in more detail in [12].

The mosaic representation removes the redundancy contained in the overlap between successive frames and represents each spatial point only once. Mosaics have been previously used as an effective way of creating panoramic views of a scene from video sequences [3], [16], [20], [23], [31], [32]. Until now, however, they have not been used as an information component within a scene-based representation, which provides direct and efficient access to video data.

Section IV describes a hierarchy of mosaic representations. The hierarchy corresponds to increasing complexity levels in the camera motion and in the 3-D scene structure.

2) Geometric Transformations: These relate the different video frames to the mosaic coordinate system. The geometric transformations contain the information necessary to map the location of each scene point back and forth between the panoramic mosaic image(s) and the individual frames. Corresponding to the hierarchy of the panoramic mosaic representations, there exists a hierarchy of representations of the geometric transformations. These range from global parametric two-dimensional (2-D) transformations to more complex 3-D transformations and are described in Section IV.

3) Dynamic Information: This is the information about *moving objects*, which are not captured by the static panoramic mosaic image. Moving-object information is *completely* captured by representing the extended time trajectories of those objects as well as their appearance. Such a complete representation is needed, e.g., for video compression (since the video frames need to be reconstructed from the scene representation). To access, browse, index, and annotate the video (as presented in Section III), however, the trajectory information alone is sufficient. The trajectory of the center of mass of each detected moving object (i.e., a single image point per moving object per frame) is maintained. These trajectories are represented in the coordinate system of the mosaic image, which is common to all the frames. In the common coordinate system, time continuity, continuous tracking, and the temporal behavior of the moving object can be analyzed more effectively (see Figs. 3 and

Thus, the three components of our scene-based representation form a *compact* representation of the video clip. The compactness results from the fact that every scene point is presented *only once* in the mosaic image, while in the original video clip, it is observed in multiple frames. This compactness of the scene-based representation facilitates very high *compression* (and we have developed such algorithms for very-low-bit-rate compression [13]). In this paper, we focus on the power of this representation for video indexing and manipulation. Section III describes how this representation can be used for efficiently accessing and manipulating the video data. Section IV describes the methods for constructing the scene-based representation.

III. From Scenes to Visual Summaries and Indexing

Once a video sequence is transformed from the frame-based representation to the scene-based representation, it forms the basis for the user’s interaction with the video. The user can initially preview the video by browsing through *visual summaries* of the various video clips. These visual summaries can serve as a *visual table of contents* of the video data. When a scene of interest is detected by the user, he can either request to view only that portion of the video or further index into individual video frames. The detected frames of interest can then be either *viewed* or *manipulated* by the user.

A. Visual Summaries—A Visual Table of Contents

There are two types of visual summaries of video clips through which a user can browse. These are captured by two types of mosaic images, which are constructed from the video clip of a scene.

1) The Static Background Mosaic: The video frames of a single video segment (clip) are aligned and integrated into a single mosaic image. This image provides an extended (panoramic) spatial view of the entire static background scene viewed in the clip in a single “snapshot” image and represents the scene better than any single frame. This image does not include any moving objects. The user can visually browse through the collection of such mosaic images to select a scene (clip) of interest.

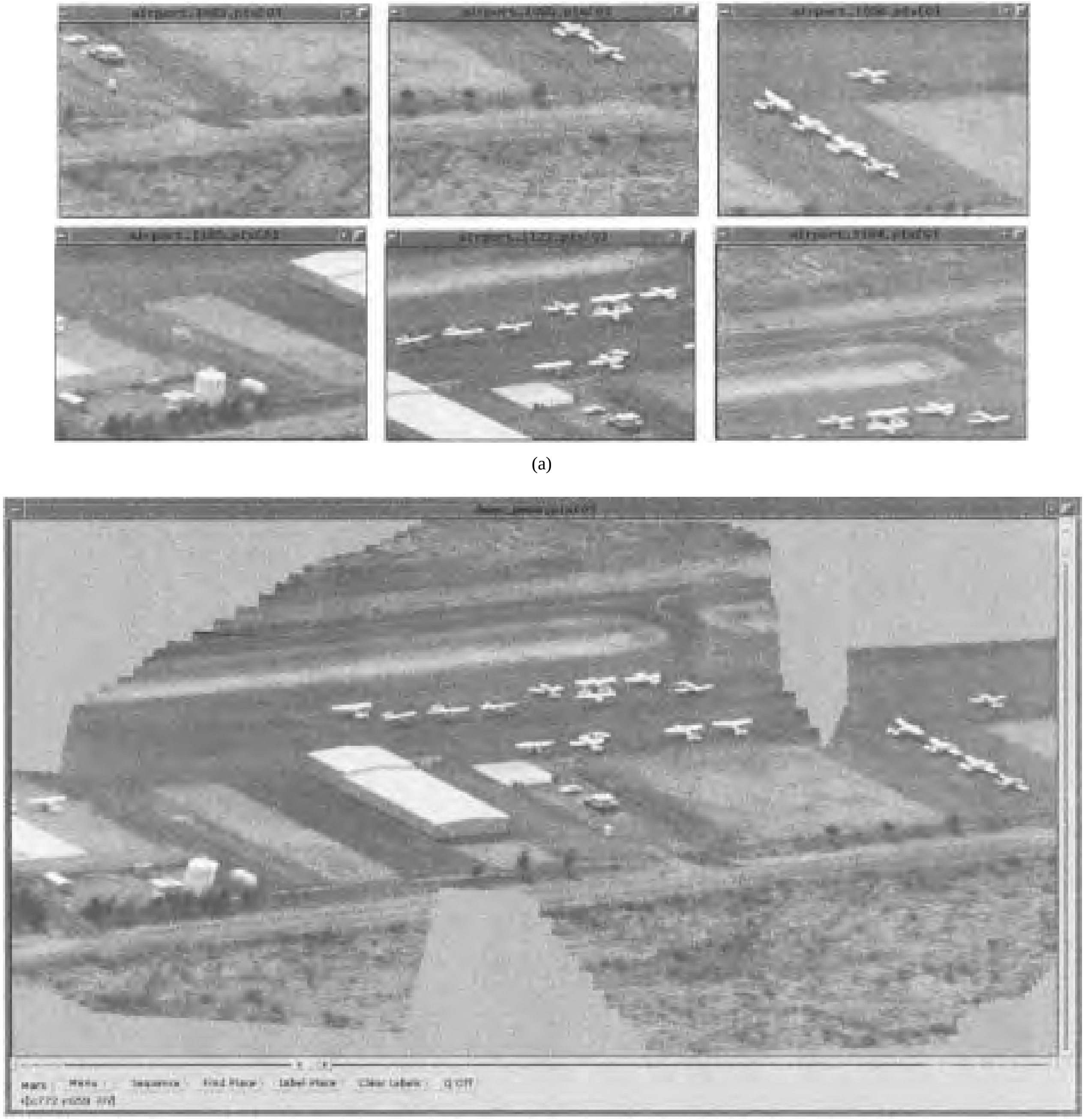
Figs. 1 and 2(b) display some examples of static background mosaic images.

2) The Synopsis Mosaic: While the static mosaic image effectively captures the background scene, it contains no representation of the dynamic events in the scene. To provide a summary of the events, we create a new type of mosaic called the *synopsis* mosaic. This is constructed by overlaying the trajectories of the moving objects on top of the background mosaic. This single “snapshot” image provides a visual summary of the entire dynamic foreground *event* that occurred in the video clip.

Fig. 3 graphically illustrates the trajectory associated with a moving object in a synopsis mosaic.

Fig. 2(c) provides a summary of the entire event in the

DOCKET ALARM
Find authenticated court documents without watermarks at docketalarm.com.



(a)

(b)

Fig. 1. Static background mosaic of an airport video clip. (a) A few representative frames from the minute-long video clip. The video shows an airport being imaged from the air with a moving camera. The scene itself is static (i.e., no moving objects). (b) The static background mosaic image, which provides an extended view of the entire scene imaged by the camera in the one-minute video clip.

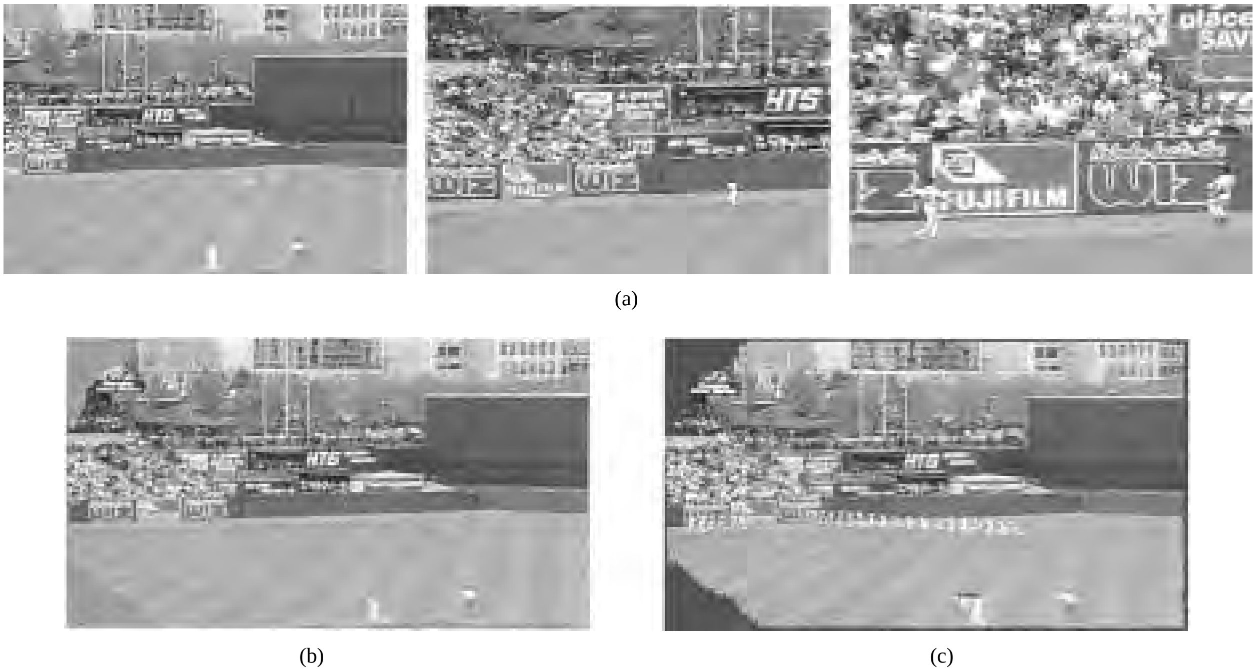
To allow for comprehensive display of multiple trajectories (corresponding to multiple moving objects), the trajectory of each moving object is uniquely color coded.

Figs. 4 and 5 provide visual summaries of airborne video clips each with multiple moving objects. Fig. 4 shows a flying airplane and a moving car on the road. Fig. 5 shows a flying airplane, three parachuters that were dropped from the plane, and a moving car.

The natural mode of operation for the user is first to

few scenes of interest. Once the user has identified a scene (i.e., mosaic) of interest, he proceeds to directly access and/or manipulate individual video frames associated with only a *portion* of the scene that is of interest to him. The scene-based representation supports this type of indexing. Two new types of indexing methods are presented: 1) indexing based on *location* (geometric) information, and 2) indexing based on *dynamic* information. These are made possible directly via the geometric coordinate transforma-

DOCKET ALARM Find authenticated court documents without watermarks at docketalarm.com.



(a)

(b)

(c)

Fig. 2. Visual summaries of a baseball video clip. (a) A few representative frames from the video clip. The video shows two outfielders running, while the camera is panning to the left and zooming on the two baseball players. (b) The static background mosaic image, which provides an extended view of the entire scene captured by the camera in the video clip. The "missing" regions at the top left and bottom left were never imaged by the camera because at that point, it was zoomed on the two players (e.g., frame 80). (c) The synopsis mosaic, which provides a visual summary of the entire event. It shows the trajectories of the two outfielders in the context of the mosaic image.

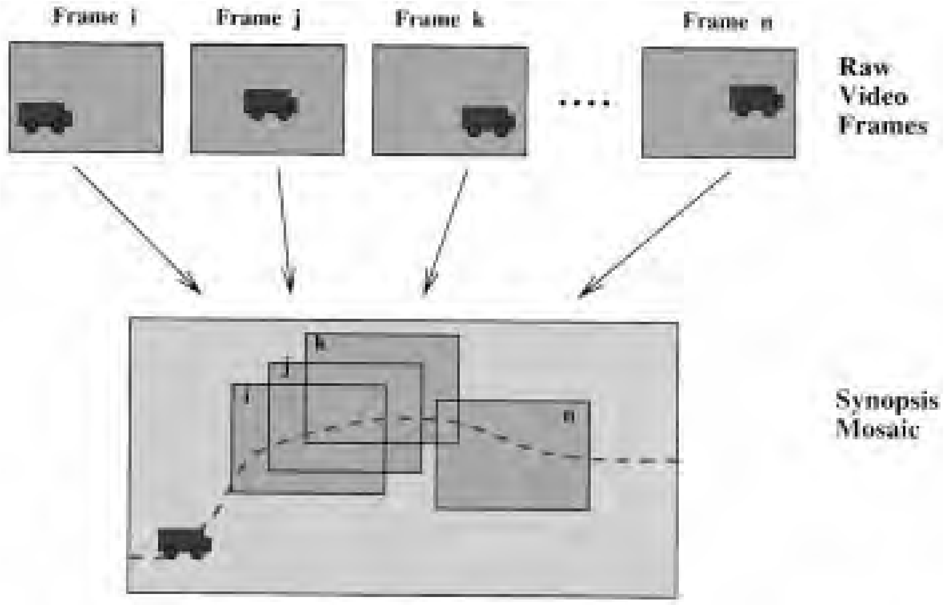


Fig. 3. Synopsis of a moving object. The trajectory of the moving object is depicted in the synopsis mosaic. This shows the motion of the moving object after cancellation of the background (camera-induced) motion. With each point on the trajectory is associated a frame number (i.e., the "time" when the moving object was at that location).

through the moving objects information that was estimated in the formation of the mosaic-based scene representation (Section II-B). The access and manipulation of selected video frames is done directly from the mosaic-based visual summaries. These location and dynamic indexing methods based indexing," which utilizes image-appearance information (e.g., color and texture) [7], [9], [10], [26]. However, our methods are considerably simpler to achieve and are highly computationally efficient.

The remainder of this section describes these modes of

DOCKET ALARM
Find authenticated court documents without watermarks at docketalarm.com.

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**