INTERACTIVE CONTENT-BASED VIDEO INDEXING AND BROWSING

Michal Irani[1]
Dept. of Applied Math and CS
The Weizmann Institute of Science
76100 Rehovot, Israel

H.S. Sawhney R. Kumar P. Anandan
David Sarnoff Research Center
Princeton, NJ 08543-5300, USA

Abstract - In this paper we present a framework for efficient representation, access, and manipulation of video data. Our approach is based on decomposing video information into its spatial (appearance), temporal (dynamics), geometric components. This derived information is organized into data representations that support *non-linear* browsing and efficient indexing to provide rapid access directly to the information of interest.

INTRODUCTION

Videos provide a flexible, and cost-effective method for collecting and recording information about scenes, objects, and events. However, due to a lack of tools for efficient access to information of interest, they are primarily acquired and stored largely for the purposes of viewing only. In this paper, a new approach is presented for efficient and rapid browsing, indexing, storage, and manipulation of video data. The input video sequence is automatically parsed into its three fundamental information components: (i) the *spatial* (appearance) information, (ii) the *temporal* (dynamics) information, and (iii) the *geometric* (3D structure and camera-induced motion) information. These provide a *complete* and *efficient* representation of the video data, and enable to *non-linear* browsing of the video, and rapid access to the video information, just like browsing a book. A hierarchy of appearance, temporal, and geometric based *visual indices* is constructed, corresponding to each of the three fundamental information components A *visual table of contents*, consisting of *visual summaries* of the video clips is also constructed. These visual summaries allow a user to "preview" the video data, and provide the interface through which the user can access the video segments/frames using the abovementioned visual *indices*. Furthermore, these compact video representations allow *efficient* video editing, annotation, and other types of video manipulations (e.g., fly-throughs). They are also well suited for efficient and scalable storage and transmission.

[1] This work was done while the author was at the David Sarnoff Research Center.

0-7803-3780-8/97/$10.00 ©1997 IEEE

PARSING THE VIDEO: FROM FRAMES TO SCENES

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. Second, and perhaps the most common use of video is to record the evolution of events over time. 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 three-dimensional geometric layout of the scene and the motion of the camera. These three *scene* components form the total information contained in the video data. However, this information is distributed among a sequence of frames, and is implicitly encoded in terms of image motion. As a result, it is hard to directly access and use the scene information contained in the video sequences.

A natural way to reorganize the video data is to decompose it into its three information components: the extended spatial, temporal, and geometric information about the scene. To do this, the video stream is first *temporally* segmented into *scene* segments. A beginning or an end of a scene segment is automatically detected wherever a scene-cut or scene-change occurs in the video (e.g., see [1]). Each scene segment is then decomposed into its fundamental scene components.

The first step is the determination of the alignment parameters that compensate for the camera-induced motion [2, 5]. These can vary from simple 2D to more complex 3D geometric transformations, depending on the complexity of the scene and the camera motion. The 3D structure of the scene can be explicitly represented when the input data supports the recovery of such information [8, 3]. The geometric coordinate transformations, along with the reconstructed 3D scene structure, form the *geometric component* of that video segment.

Once these geometric relationships are established, the image frames are aligned and the camera-induced motion is removed. The separation of the video data into dynamic and static scene parts can therefore be performed. The static portions of the images (i.e., those which are aligned after removal of the camera-induced motion) are integrated into a single panoramic image of the scene [3, 7]. These representative views of the static background scene form the *extended spatial (appearance) information component* of the scene.

Moving objects are detected as the remaining misaligned parts in the image after removal of the camera-induced motion. Each moving object is segmented and tracked over time, and are represented by its appearance and its trajectory over time. This forms the *extended temporal information*.

The spatial background mosaic, the temporal foreground information, and the geometric coordinate transformations constitute a complete and efficient representation of the video data [3]. The video frames can be fully recovered from these three components.

VISUAL SUMMARIES

There are two types of visual summaries of video clips that a user can browse through. These also serve as a "table-of-contents" of the video data.

- **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 background scene viewed in the clip (see Figure 1 of a baseball video), and represents the scene better than any single frame. The user can visually browse through the collection of such mosaic images to select a scene (clip) of interest.
- **Synopsis mosaic:** While the background mosaic is useful for capturing the background scene, it is desirable to also obtain a synopsis of the event that occurs within the video clip. This is achieved by overlaying the trajectories of the moving objects on top of the background mosaic (Figure 1). This single "snapshot" image provides a visual summary of the entire foreground event that occurred in the video clip. Each moving object is represented separately and uniquely (see Figure 3 of an airborne UAV video).

A HIERARCHY OF VISUAL INDICES

Given our organization of the video in terms of its three fundamental information components, there are two natural modes of interaction for the user. The user can browse the table-of-contents to identify a few scenes of interest. This process is made more efficient by creating a set of appearance-based indices which can be used to select relevant scenes/clips of interest based on visual similarity. This is especially useful to speed-up the access from large video databases. The other mode of interaction is for the user to directly access and/or manipulate the video frames associated with an identified scene of interest. This is made possible via the alignment parameters and moving objects information which was estimated in the formation of the background mosaics. To facilitate both modes of search, we propose a hierarchy of indices, corresponding to each of the three fundamental information components.

- **Indexing based on appearance (spatial) attributes:** Visual appearance based indices enable the user to look for *similar* scenes and objects across video clips. We employ two kinds of visual indices: (i) global scene based, and (ii) object/patch/pattern based. Global indices capture color, texture and spatial distributions of color and texture of mosaics. Object and pattern based indices use multi-dimensional feature vectors that are invariant/quasi-invariant with respect to view variations, and that capture the saliency of patterns in patches at multiple scales. Each of these appearance indices is organized into multi-dimensional tree-like search structures so that sub-linear indexing can be done. Views of the database generated through these indices can be further browsed and manipulated using the spatial and object indexing methods specific to each video clip. One example of non-linear ordering of a 260-clip baseball game using similarity of texture distributions is shown in Figure 2. The first frame is used as a query frame. Texture distributions computed on the query frame and the rest of the clips are used to order the whole database with respect to the query. The top 16 matches ordered by a similarity measure are shown in the figure.

• **Indices based on geometric attributes:** Once a few scenes of interest (in the form of visual summaries) have been selected, the user proceeds to access the video frames themselves. The user selects a scene point by by clicking on a point on the mosaic image, or automatically through a pattern matching search on the mosaic. All frames containing the selected scene point are immediately determined through the geometric coordinate transformations that relate the video frames to the mosaic image. The same mechanism is also used to efficiently inherit annotations from the mosaic image onto scene locations in the video frames (see Figure 3 of airborne UAV video). The annotation is specified by the user *once* on the mosaic image, rather than tediously specifying it for each and every frame.

• **Indices based on temporal attributes:** Similarly, all frames containing a selected moving object can be immediately determined and accessed. The user can either interactively select an object of interest (e.g., via clicking on an object marked on the synopsis mosaic), or automatically through appearance-based search. Since the geometric relationship as well as the trajectories of the moving objects are precomputed, all frames containing that object are immediately accessed and viewed. Similarly, the moving objects in the video frames are efficiently annotated or manipulated via the same mechanism used for indexing, without the need for the user to repeatedly perform the operation on a frame-by-frame basis. (see Figure 3 of airborne UAV video).

References

[1] H.-J. Zhang, A. Kankanhalli, and S.W. Smoliar. Automatic partitioning of full-motion video. *Multimedia Systems*, 1(1):10–28, 1993.

[2] J.R. Bergen, P. Anandan, K.J. Hanna, and R. Hingorani. Hierarchical model-based motion estimation. In *ECCV-92*, pp. 237–252, Santa Margarita Ligure, May 1992.

[3] M. Irani, P. Anandan, J. Bergen, R. Kumar and S. Hsu. Efficient Representations of Video Sequences and their Applications *Signal Processing: Image Communication*, 8:327–351, 1996.

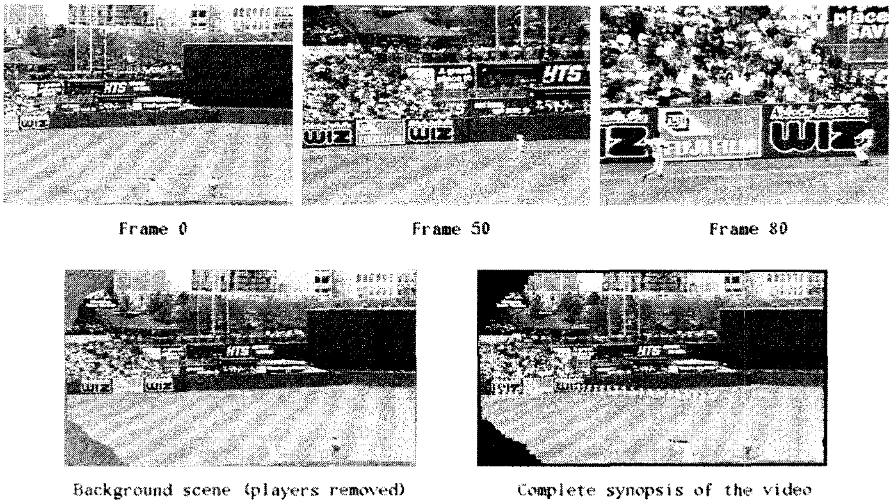
[4] M. Irani, S. Hsu, and P. Anandan. Video Compression Using Mosaic Representations *Signal Processing: Image Communication*, 7:529–552, 1995.

[5] M. Irani, B. Rousso, and S. Peleg. Recovery of Ego-Motion Using Image Stabilization In *Proc. IEEE CVPR-94*, June 1994.

[6] S. Mann and R.W. Picard. Virtual bellows: Constructing high quality stills from video. In *Proc. IEEE ICIP-94.*, November 1994.

[7] H. S. Sawhney and S. Ayer. Compact Representations of Motion Video using Dominant and Multiple Motion Estimation. *IEEE Trans. PAMI*, vol. 18, 1996.

[8] Rakesh Kumar, P. Anandan, M. Irani, J. R. Bergen, and K. J. Hanna. Representation of scenes from collections of images. In *Workshop on Representations of Visual Scenes*, 1995.



Frame 0

Frame 50

Frame 80

Background scene (players removed)

Complete synopsis of the video

Figure 1: Synopsis mosaic image of the baseball game sequence.

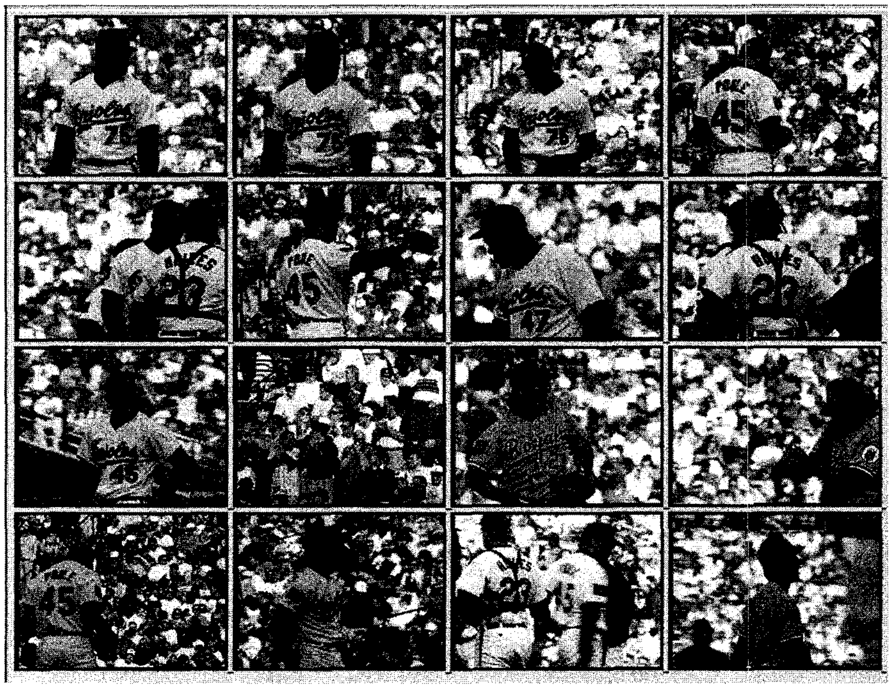


Figure 2: Representative frames from a 260 clip baseball video ordered by similarity of texture content. Top 16 frames are shown ordered based on texture similarity with respect to the first frame that is used as query.

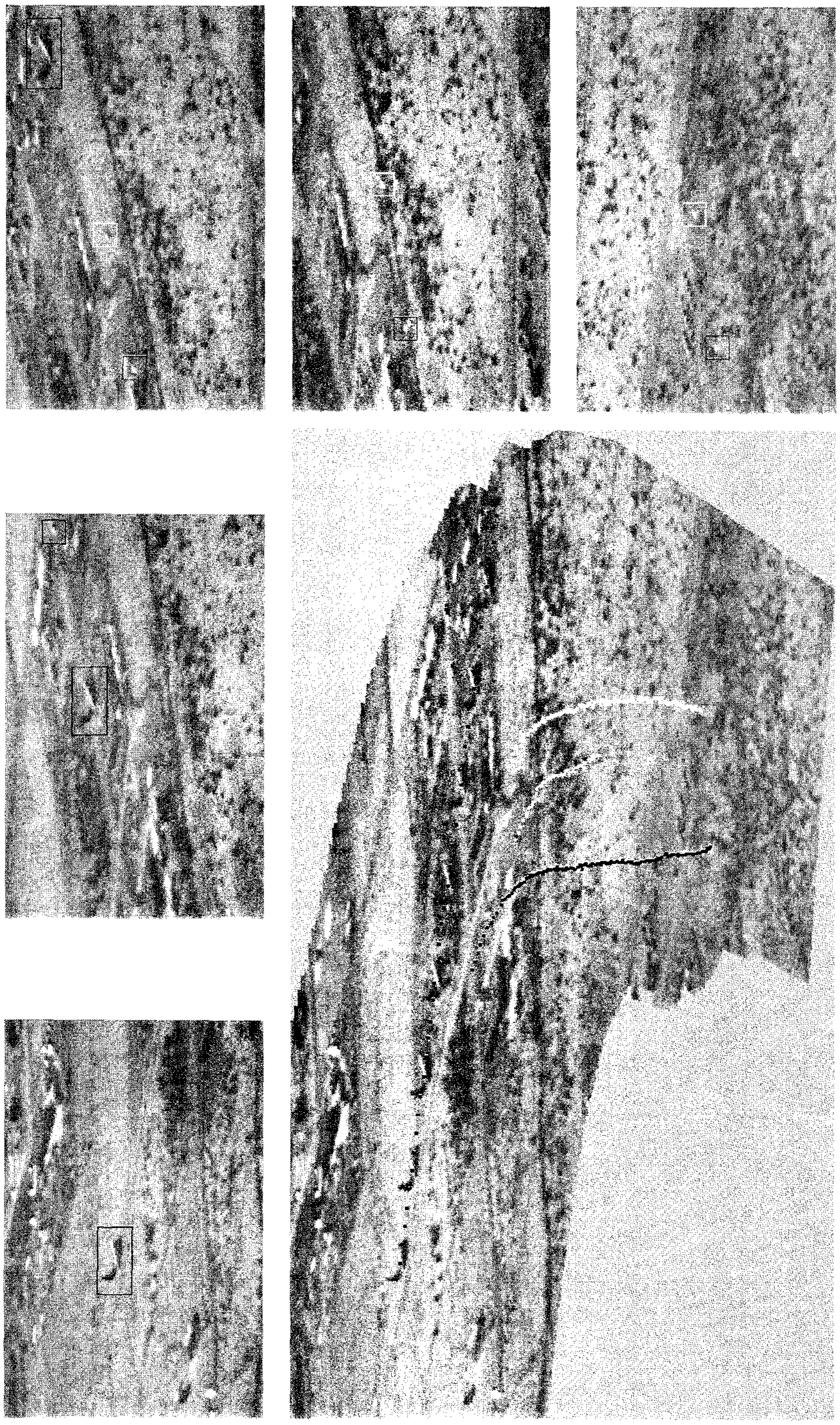


Figure 3: : *Big frame on bottom right*: Synopsis mosaic with multiple object trajectories highlighted, and *small frames*: corresponding objects outlined in a few frames. *Red*: airplane, *Green, Cyan, Yellow*: Parachuters.