VIMOS: A Video Mosaic for Spatio-Temporal Representation of Visual Information

K. Selcuk Candan, Forouzan Golshani, Sethuraman Panchanathan, Youngchoon Park

Department of Computer Science and Engineering
Arizona State University
Tempe, AZ 85287-5406

ABSTRACT

The capability of extracting critical information from live video signal and presenting it to end-users in an easy-to-grasp form is essential in many application domains, including GIS, intelligence, surveillance, and manufacturing. We outline on the development of methods and algorithms that are necessary for real-time analysis of video data, both in compressed or uncompressed domains, generation of visual icons that embody the characteristics of the objects of interest (say enemy crafts, main actor, star player, etc.) and presentation of the panoramic spatio-temporal view of the entire scene in the form of video mosaics.

1. INTRODUCTION

Significant developments have been reported in the area of image cataloging and classification. However moving pictures deliver information in such an incredible rate that their cataloging via the usual techniques far exceeds the technological capabilities of text-based classification mechanisms.

Anyone who has ever tried to find a specific scene of interest in the family video library would know the difficulty in the task of video retrieval. Wouldn't it be nice to have a sequence of scenes each representing a segment of video along with a pointer to the actual frame number in which the scene occurs? Or wouldn't it be nice to be able to formulate a query specifying the characteristics of the desired scene and expecting the video delivery system to find the appropriate segments that contain the events or objects of interest? Such capabilities are considered luxury in the case family video sets, in many "industrial" application, they are absolutely essential. Examples include: industrial process monitoring, video on demand, security and surveillance systems, and traffic violation monitoring. In these applications, the shear volume of video data makes manual examination of huge volumes of video impossible, and the capability of extracting critical information from live video signal and presenting it to end-users in an easy-to-grasp form is vital.

Work on video compression is blossoming into a very sophisticated field. The successes of early MPEG algorithms have fostered research into yet another set of standards, MPEG7, that would embody all of the previous developments. The new standards is expected to embody significant studies in the area of content extraction of video.

Motion analysis is an important step in the process. This problem can be solved by using low level motion analysis based on optical flow analysis (which is costly and cannot be done on the fly), or by using spatio-temporal surface flow analysis (which is more coarse and requires less computational resources). In our earlier work [3], we designed a set of algorithms for high level motion extraction from video. This is different from the work presented in [2] which used a computational framework for intermediate level and high level motion analysis based on spatiotemporal surface flow and spatio-temporal flow curves.

A general architecture for the analysis of moving objects was presented by Kubota [9]. The process of motion analysis was divided into three stages: moving object candidate detection, object tracking, and final motion analysis. The experiments were conducted using human motion. Also related are the studies at the MIT Media Labs. Examples are salient video stills [10], which involves determining the optical flow between successive frames, applying affine transformations calculated from the flow warping transforms, like rotation, translation, etc., and applying a weighted median filter to the high resolution image data resulting in the final image; and a method for synthesizing panoramic overviews from a sequence of frames. Swanberg proposed a method for identifying desired objects, shots, and episodes from the video prior to insertion in the database [11].

Cut detection may be performed in a number of different ways [12]. One common method uses a projection detection filter which is based on finding the largest difference in consecutive frame histogram differences over a period of time. A model driven

0-7803-4876-1/98/$10.00 © 1998 IEEE

approach to digital video segmentation is presented in by Hampapur [13]. It deals with extracting features that correspond to cuts, spatial edits, and chromatic edits. Also relevant to our topic is the work on camera operation detection. The technique presented by Idris and Panchanathan uses the patterns created by the motion vectors of the motion compensation algorithm to recognize camera operations such as panning and zooming [6].

This paper, which is a brief and informal overview of the video processing system VIMOS, is organized as follows. The next section outlines the problem and identifies the essential steps in video content processing. Section 3 presents an overview of the VIMOS system. Section 4 contains more details of the how VIMOS works. The paper closes with some concluding remarks.

2. PROBLEM OVERVIEW

Real-time processing of video data, including content analysis and representation, is essential in many applications. Example applications range from commercial environments, say, subject tracking and surveillance, manufacturing, and space explorations, to modern military combat environments which require complex information processing to enable a vast assortment of input, continuous assessment of the circumstances, coordination of resources, and response to threats. An important source of information in all these environments is the input obtained from stationery or moving cameras. Unless the visual data recorded/transmitted by cameras are processed immediately, either the visual content becomes obsolete or the accumulated data becomes so large that the task of processing would be impossible. An example of the latter case is the vast amount of video data sent down by NASA's Magellan spacecraft. Such capabilities require the fulfillment of the following tasks:

- capturing the raw video and delivering it to a data and information processing unit,
- conversion of the raw data into a format suitable for storage and indexing,
- extraction of the relevant (or critical) information from the raw and/or processed video data before its storage,
- annotation of the video data for off-line processing and future access,
- generation of a "world model" from the extracted information,
- analysis of a world model for target recognition and tracking and early signaling of suspicious events,
- visualization of the world model and the analysis results for human evaluation and intervention, and
- storage of the world model and the analysis results for off-line processing and future accesses.

The processing of the above tasks in a timely manner requires on-line processing of the video data, real-time retrieval of relevant information from information stores, and dynamic and easy-to-grasp visualization of results. Also, since the end-users may or may-not know what exactly they should be looking for at a given instant, both "pull" and "push" based information access methods are needed.

3. OVERVIEW OF VIMOS

Our video content retrieval tool, called VIMOS, is designed for capturing, indexing, storing and visually presenting the information contained in video streams. Fig. 1 illustrates the overall architecture. VIMOS can handle visual data of different formats, in both compressed and uncompressed domains. Visual information has two dimensions; namely, spatial and temporal. The spatial content of video is represented by the usual combination of structural information (regions, objects and portions of objects), features (color, texture, object sketch and shape), and spatial relationships [4][5]. The temporal information present in a video, on the other hand, can be represented by the motion (trajectory) of the objects, camera operations, viewing perspective, and the temporal relationships [4][6]. In VIMOS, the video index captures both spatial and temporal contents.

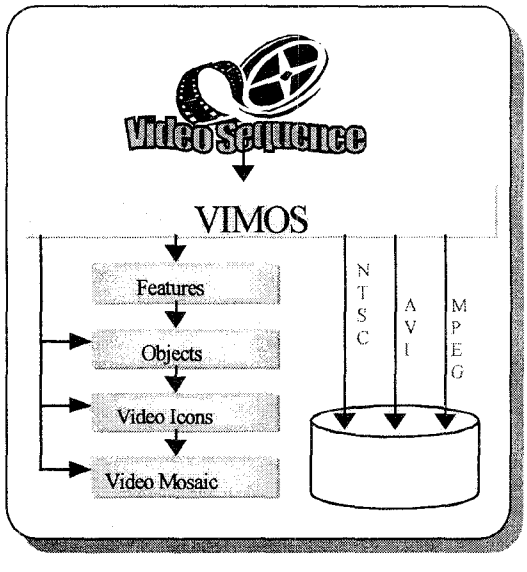

Fig. 1: Overall system functionality

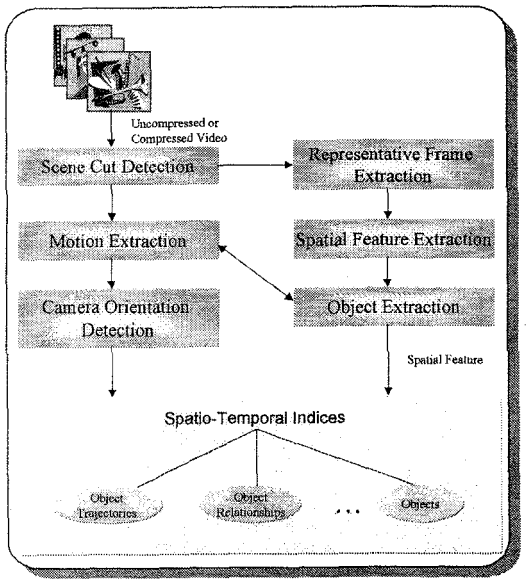

Fig. 2: Visual content indexing process

The steps involved in visual content indexing are presented in Fig. 2. The first step is to segment the video into elemental scenes called shots. The purpose of segmentation is to partition the incoming video stream into a set of meaningful and manageable segments, which then serve as basic units for indexing. We employ a variety of algorithms (based on the nature and the contents of the video) which have been developed for our previous image and video processing application. These algorithms are particularly efficient in detecting both abrupt and gradual scene changes. [7]

Once a video is segmented into shots, one or more representative frames (also called key frames) that best represent the contents of the shot are selected. These selections must reflect in the most effective manner the significant information pertaining to the specific application. The selection process may require human intervention for certain domains, but could automated for simpler cases. The two methods of presenting visual information that are the outcome of this project are the concepts of visual icons (VICON) and video mosaic (VIMOS).

To facilitate the creation of the appropriate VICONs for quick browsing of the incoming visual information as well as navigation and search through stored visual data, we need to extract the spatio-temporal features corresponding to each shot. The spatial features are extracted by processing the representative frame of each shot, while the temporal features are obtained by computing the motion trajectories and camera operations/viewing angle within a shot.

We first classify the motion information within a shot into two classes, namely, local, and global. The global motion vectors that characterize the temporal activity over the entire frame are used to derive the camera operations and viewing angle, while the local motion vectors that are specific to regions are clustered to identify the specific regions as well as object segmentation. For example, the zooming operation on the target and the angle of fly of an aircraft can be captured using the global motion vectors. The sub-regions or specific targets, such as the tanks on the ground, are identified using their corresponding local motion trajectories. We have devised an algorithm that extracts motion information from a video sequence. [4] Our algorithm provides a low-cost extension to the motion compensation component of MPEG compression algorithm.

Spatial feature processing includes decompression, enhancement, filtering, normalization, segmentation, and object identification. The system extracts primitive features such as color, texture, shape, sketch, etc. that are quantitative in nature either automatically or, in strategic cases, semi-automatically. The feature set chosen to represent the information is domain-specific and must be optimized for each application. We note that the target/object and region segmentation procedures would employ both the spatial and temporal information to ensure robust target detection.

In summary, the spatial and temporal processing of the video data generates the following outputs: target objects, their trajectories, camera operations/viewing angle, spatial and temporal relationships, annotations, and features. These outputs are subsequently assembled together to create VICONs and VIMOS. Since VICONs are end-users' windows to the real-world information, they both provide a complete description of the visual content of the video data and they highlight the most relevant information for quick recognition. Hence, generating VICONs requires the creation of a world-model to represent the visual content and the filtering of the extracted visual information to find the most critical subset. This process includes the use of a knowledge-base for the identification of the semantics of the objects (e.g. friend vs. enemy) and their behaviors (e.g. attacking) and the merging of these semantics with the visual content to create a prioritized world model. This model, then, can be used in the creation of the VICONs and VIMOS as well as in the storage of the extracted information for future accesses. The priorities play an essential role both in the highlighting of the critical information in VICONs and in the creation of storage hierarchies to speed up access to relevant information.

4. VIMOS: A CLOSER LOOK

As described in the earlier sections, the input to the system is a compressed domain digital video stream. For this paper, we use a short video of a basketball game, as presented in Figure 3. Initially, the information content of this stream is not yet extracted. As the first step of the information extraction process, we perform shot detection and identification on the input stream. This algorithm identifies the shot boundaries and returns the individual shots. This process is based on the technique presented in [14] which combines color histogram comparisons and model based object detection along with optical flow analysis. We define a shot boundary as the time point at which there has been a significant change in the visual content of the stream. Hence, this algorithm pinpoints the significant points in time, such as when a new object enters into the view, an existing object leaves the view space, or there is a signature of an artificial editing of the video stream. Each shot, then, corresponds to an atomic sequence of events that are closely related with each other.

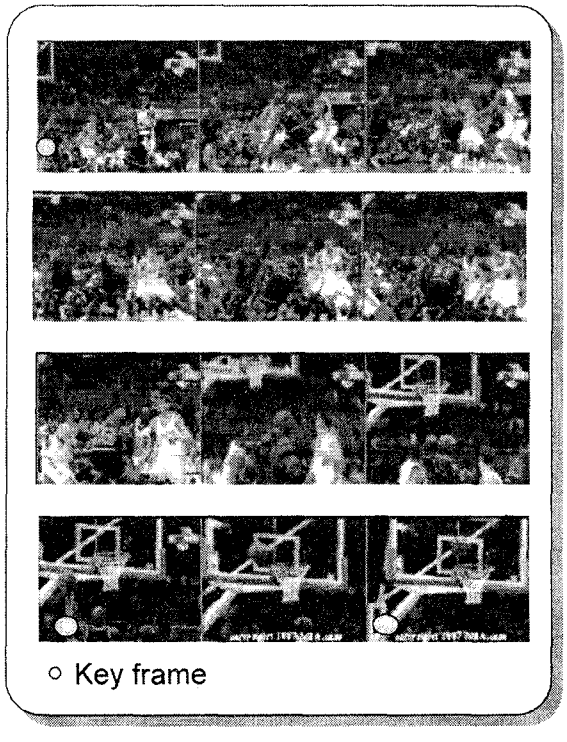

Fig. 3: A sample digital video stream.

Given a video sequence (a shot), we first create a representative environment for the shot. We use the world knowledge, stored in the knowledge base, to match the background of the image to the world-model and locate the position and orientation of the camera. If we do not have pre-acquired knowledge about the environment at which the shot is taking place, we select (or synthesize) a key-frame from the frames in the video sequence. The selected key-frames are shown in Figure 4. Each key-frame is the overall representation of the environment.

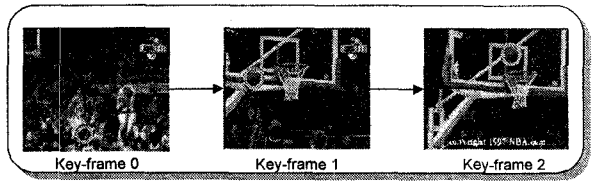

Fig. 4: Selected key-frames

The next step in the information extraction process is the identification of the objects within the shots. We use image segmentation techniques and a variety of other image processing filters to identify the objects within individual frames [5]. We also derive the associations between objects, including spatial relationships, in the frames of the given shot. Next, we identify the camera motions (such as pan, tilt, zoom) using the motion vectors between consecutive frames. We, then, use the camera motion information and the motion vectors associated with the objects in a given shot to find the trajectories of the objects with respect to a reference point in the world-model.

The above information allows us to create <object, trajectory, location> triplets which describe the behavior of the object within the given shot. Note that at the specification level, we do not make any assumptions on the internal representations and implementation of the three components of the <object, trajectory, location> triplets. The first component, Object, may be represented by its collection of its attributes, both visual and alpha-numeric, organized in an object oriented fashion. Trajectory may have a number of different representations, such as chain codes, directional vectors, and polynomial splines. These representations would be 2D projections of the spatial movement. The Location component is the simplest and refers to a region on the grid. These triplets constitute the lowest level building blocks of VICONs. Figure 5 illustrates a typical <object, trajectory, location> triplet for the object ball appearing in the sample video sequence.

When there is a need to track the movements of an object through a series of key frames, the position and the partial trajectory of the object in each frame will have to be reflected in the subsequent key frame(s). Such transformations require spatial translations followed by interpolation. Fig. 6. illustrates this process. See [15] for more details.

<O, T, L> where:
O is defined as
(Global_ID: 123456, Shot_ID: (5678-5722
World_name: {"ball", "basketball", ...},
Real_color: <200, 120, 35>,
picture: 2gif, shape: "round, ...)
T is the chain code represented as
(4, 4, 4, 3, 2, 2, 1, 1, 7)
L is the centroid of the region occupied by this object in the initial key frame, say
(.42, .12)

Fig. 5: An example of the <object, trajectory, location> triplet for the object ball

To summarize, each video sequence is described by its visual content (appearance of objects) and their associated semantics, including motion and trajectory information.

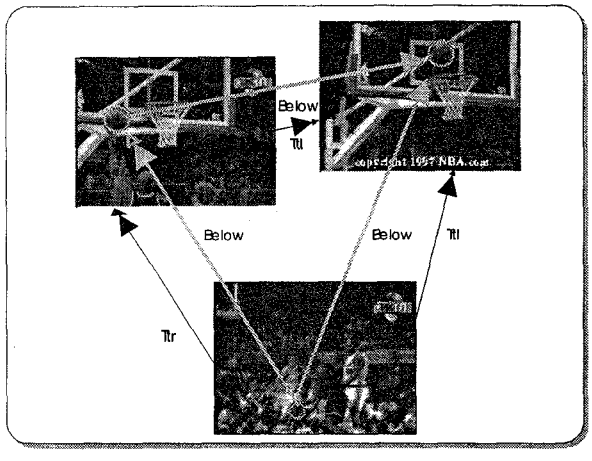

Fig. 6: Object tracking through key frames

For any particular object of interest in a given shot, the system identifies the corresponding object in the representative frame that was selected or created (as described earlier). It then plots the trajectory of the object, and superimposes it onto the representative frame at the appropriate location. This artificially created image is a VICON. A sample VICON is presented in Figure 7.

When we are interested in the overall content of the shot, the triplets corresponding to all (or the significant) objects in the sequence are drawn on the representative frame. Such an artificially created image is also a VICON. Another way to create a VICON for the shot is to use the knowledge-base to create a footprint of the environment (for instance a bird's eye view of the scene) and then to place the shot-object triplets on this footprint. For those VICONs that contain more than one object, we benefit from the user preferences and object semantics to highlight those triplets that are critical or high profile.

In some cases, users are not interested in the contents of a single shot, but a series of shots which are related to each other. These shots may correspond to a consecutive sequence of shots that a complex action (with many objects coming in and leaving the frame boundaries) took place. They, also, may correspond to a non-consecutive set of frames which, even though they are not continuous in the temporal order of the video shots, constitutes a single connected activity. In such cases, the VICONs representing the content of the shots in the given set must be put together to create a representative mosaic for the set.

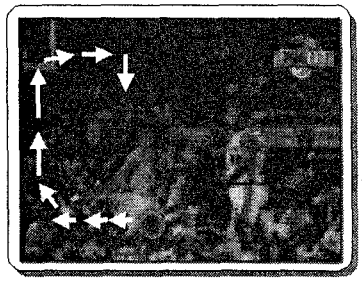

Fig. 7 : An example VICON

This process (i.e., scene composition) requires identification of the shot-objects that correspond to the same entities in different shots and merging of the backgrounds of different shots to create a unique visual world representation (Figure 8).

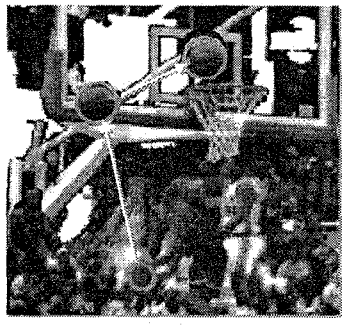

Fig. 8: An artificially synthesized VIMOS

If there is a pre-determined world model of the scene, this task will be more tractable. If, on the other hand, there is no available world-model, then this task will represent a significant challenge. In this case, instead of trying to create a single background that describes all the shots, we use the directions of the trajectories of the significant objects to place the representative frames on to a canvas and we link the trajectories of such objects to create a single

representation for the given set of shots. We call this visual representation of a set of shots a video mosaic, or a VIMOS presented in Figure 8. Readers interested in this topic are referred to the extended version of this paper that contains detailed presentations on the architecture of VIMOS and the important algorithms. [15]

5. CONCLUSIONS

We propose a framework along with a set of algorithms and technologies for immediate processing of information contained in video signal and its global representation in the form of video mosaics. VIMOS is a complete toolkit for the automated spatio-temporal analysis of the video data, real-time visualization of the results through VICONs and VIMOS, and storage and querying of visual information. VICONs and VIMOS provide a personalized (or role-oriented) highlighting which enables end-users to quickly access information most relevant to them. The pyramid-based VIMOS structure also provides different information granularities for different end-users.

6. REFERENCES

[1] S. Adali, K.S. Candan, S.-S. Chen, K. Erol, and V.S. Subrahmanian, "Advanced Video Information System", ACM-Springer Multimedia Systems Journal, vol. 4, pp.172-186, August 96.

[2] M. Allmen "Image Sequence description using spatiotemporal flow: Toward motion-based recognition," Ph.D thesis, University of Wisconsin-Madison, 1991.

[3] J. Chung-Mong Lee, Qing LI and Wei Xiong, "VIMS: A Video Information Management System", Multimedia Tools and Applications, vol.9, pp.1-25, 1992.

[4] N. Dimitrova, F. Golshani, "Motion Recovery for Video Content Classification", ACM Trans. On Office Information Systems, Vol. 13, No. 4, 1995, pp. 408-439.

[5] F. Golshani, Y.C. Park, "Content-based Image Indexing and Retrieval in ImageRoadMap", Proc. SPIE's International Symposium on Voice, Video and Data Communications – Multimedia Storage and Archiving Systems II, Dallas, TX, November 1997.

[6] F. Idris and S. Panchanathan, "Review of Image and Video Indexing Techniques", Journal of Visual Communication and Image Representation-Special Issue on Indexing, Storage and Retrieval of Images and Video, June 1997.

[7] F. Idris and S. Panchanathan, "Storage and Retrieval of Compressed Images", IEEE Transactions on Consumer Electronics , vol. 41, no. 3, pp. 937 - 941, August 1995.

[8] N. Dimitrova, F. Golshani, "Rx for semantic video retrieval" Proceedings of the ACM Multimedia Conference, San Francisco, October 1994, ACM Press, pp 219-226

[9] H. Kobuta, et al, "Vision processor system for moving object analysis" Machine Vision and Applications, Vol. 7, 1993, pp 37-43.

[10] L. Teodosio and W. Bender, "Sailent video stills: Content and context preserved," Proceedings of ACM Multimedia '93, Anaheim, CA, Aug '93.

[11] D. Swanberg, C. F. Shu, R. Jain, "Knowledge guided parsing in video databases", Image and Video Processing Conference, SPIE, Vol. 1908, Feb. 1993, pp 13-24.

[12] K. Ostoji, Y. Tonomura, "Projection etection filter for video cut detection", Proc. ACM Multimedia'93, pp 251-257.

[13] A. Hampapur, T. Weymouth, R. Jain, " Digital video segmentation" Proc. ACM Multimedia'94, pp 354-367.

[14] J. Y. A. Wang and Edward H.Anderson, "Spatio-temporal Segmentation of Video Data", M.I.T Media Lab. Vision and Modeling Group, Tech. Report No.262, Feb., 1994.

[15] K. S. Candan, F. Golshani, S. Panchanathan, Y.C. Park, "Video content representation by VIMOS", Tech. Report, Computer Science Department, Arizona State University, 1998.