InsightVideo: Toward Hierarchical Video Content Organization for Efficient Browsing, Summarization and Retrieval

Xingquan Zhu, Ahmed K. Elmagarmid, *Senior Member, IEEE*, Xiangyang Xue, Lide Wu, *Senior Member, IEEE*, and Ann Christine Catlin

***Abstract*—Hierarchical video browsing and feature-based video retrieval are two standard methods for accessing video content. Very little research, however, has addressed the benefits of integrating these two methods for more effective and efficient video content access. In this paper, we introduce InsightVideo, a video analysis and retrieval system, which joins video content hierarchy, hierarchical browsing and retrieval for efficient video access. We propose several video processing techniques to organize the content hierarchy of the video. We first apply a camera motion classification and key-frame extraction strategy that operates in the compressed domain to extract video features. Then, shot grouping, scene detection and pairwise scene clustering strategies are applied to construct the video content hierarchy. We introduce a video similarity evaluation scheme at different levels (key-frame, shot, group, scene, and video.) By integrating the video content hierarchy and the video similarity evaluation scheme, hierarchical video browsing and retrieval are seamlessly integrated for efficient content access. We construct a progressive video retrieval scheme to refine user queries through the interactions of browsing and retrieval. Experimental results and comparisons of camera motion classification, key-frame extraction, scene detection, and video retrieval are presented to validate the effectiveness and efficiency of the proposed algorithms and the performance of the system.**

***Index Terms*—Video browsing, video content organization, video retrieval, video similarity assessment.**

I. Introduction

RECENT advances in high-performance networking and improvements in computer hardware have led to the emergence and proliferation of video and image-based applications. Database management techniques for traditional textual and numeric data cannot handle video data; therefore, new models for storage and retrieval must be developed. In general, a video database management system should address two different problems: 1) the presentation of video content for browsing and 2) the retrieval of video content based on user queries.

Some methods have been developed for presenting video content by hierarchical video shot clustering [1], [2], organizing storyboards [3], or joining spatial-temporal content analysis and progressive retrieval for video browsing [4]. These methods allow a viewer to rapidly browse through a video sequence, navigate from one segment to another, and then either get a quick overview of video content or zoom to different levels of detail to locate segments of interest. These systems are efficient in video browsing and content presentation; however, they fail either in detecting semantically related units for browsing [1], [2], [4] or in integrating efficient video retrieval with video browsing [3].

In comparison with video content presentation, more extensive research has been done in the area of video retrieval. Several research prototype systems have been developed which provide automatic indexing, query and retrieval based on visual features, such as color and texture [5]–[12]; others execute queries on textual annotation [13]. Elmagarmid, *et al.* [14] has published a comprehensive overview of this topic.

The first video parsing, indexing and retrieval framework was presented by Zhang, *et al.* [2]. It uses the annotations and visual features of key-frames for video browsing and retrieval. QBIC [12] supports shape queries for semi-manually extracted objects. The Virage [8] system supports feature layout queries, and users can assign different weights to different features. The Photobook system [9] enables users to plug in their own content analysis procedures. Cypress [11] allows users to define concepts using visual features like color. VisualSEEk [10] allows localized feature queries and histogram refinements for feedback using a web-based tool. Systems such as CVEPS [15] and JACOB [16] support automatic video segmentation and video indexing based on key-frames or objects. The web-based retrieval system, WebSEEK [17], builds several indexes for images and video based on visual features and nonvisual features. The Informedia digital video library project [18] has done extensive research in exploring video knowledge by integrating visual features, closed caption, speech recognition etc. A more advanced content-based system, VideoQ [7], supports video query by single or multiple objects, using many visual features such as color, texture, shape, and motion.

Manuscript received August 13, 2001; revised February 17, 2004. This work was supported by the National Science Foundation (NSF) under Grants 9972883-EIA, 9974255-IIS, 9983248-EIA, and 0209120-IIS, by a grant from the state of Indiana 21 Century Fund, and by the U.S. Army Research Laboratory and the U.S. Army Research Office under Grant DAAD19-02-1-0178. The work of X. Xue and L. Wu were supported by the NSF of China under Contracts 60003017, 60373020 and 69935010. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Sankar Basu.

X. Zhu is with the Department of Computer Science, University of Vermont, Burlington, VT 05405 USA (e-mail: xqzhu@cs.uvm.edu).

A. K. Elmagarmid and A. C. Catlin are with the Department of Computer Science, Purdue University, West Lafayette, IN 47907 USA (e-mail: ake@cs.purdue.edu; acc@cs.purdue.edu).

X. Xue and L. Wu are with the Department of Computer Science, Fudan University, Shanghai 200433 China (e-mail: xyxue@fudan.edu.cn; ldwu@fudan.edu.cn).

Digital Object Identifier 10.1109/TMM.2005.850977

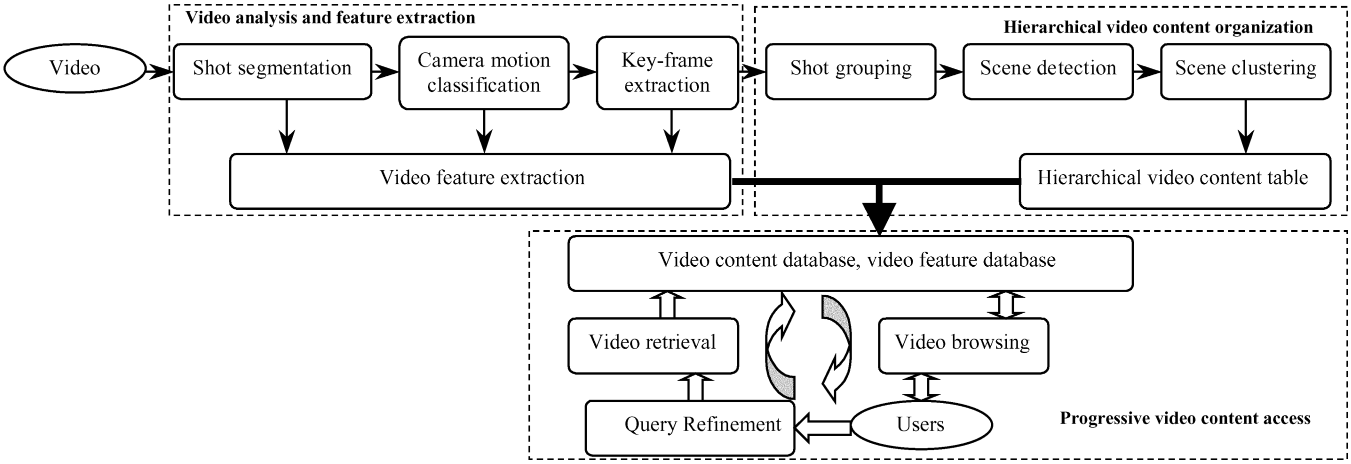


Fig. 1. System flowchart for InsightVideo.

However, the video retrieval approaches introduced above usually just add the functionalities for shot segmentation and key-frames extraction to existing image retrieval systems. After shot detection and key-frame extraction, they merely apply similarity measurements based on low-level features of the video frames or shots. This is not satisfactory because video is a temporal media, so the sequencing of individual frames creates new semantics that may not be present in any of the individually retrieved shots.

A naïve user is interested in querying at the semantic level, rather than having to use features to describe his (her) concept. In most cases, it is difficult to express concepts using feature matching, and even a good match in terms of feature metrics may yield poor query results for the user. For example, in multiple domain recall, a query for 60% green and 40% blue may return an image of a grass and sky, a green board on a blue wall or a blue car parked in front of a park, as well as many others. Helping users to find query examples and refine their queries is also an import feature for video retrieval systems. However, instead of integrating the efficient video browsing and retrieval together, the systems described above emphasize either browsing or retrieval. A progressive strategy should be developed to join video browsing and retrieval schemes together to improve the effectiveness and efficiency of both.

Motivated by the above observations, we propose a novel video content organization and accessing model for video browsing and retrieval. A progressive video retrieval scheme is formed by executing the browsing and retrieval iteratively. The distinct features of our system are the following: 1) several novel video processing techniques are introduced which improve existing algorithms in important areas; 2) a video content hierarchy is constructed which allows hierarchical video browsing and summarization to be executed directly and efficiently; 3) by addressing video similarity at different levels and granularity, our retrieved results mostly consist of visually and semantically related units; and 4) the seamless integration of video browsing and retrieval allows users to efficiently shrink and refine their queries.

II. System Overview

The process flow for the InsightVideo system is illustrated in Fig. 1. The system consists of three parts: 1) video analysis and feature extraction; 2) hierarchical video content organization; and 3) progressive video content access. To extract video features, a shot segmentation algorithm is applied to each input video. Then, for each segmented shot, the camera motion classification strategy is utilized to qualitatively classify camera motions. Based on identified motion information, key-frame extraction is executed to select the key-frame(s) for each shot. The detected camera motions and low-level features are utilized for video similarity evaluation. After the video features have been extracted, the video content table is constructed by shot grouping, scene detection, and scene clustering strategies to generate a three-layer video content hierarchy (group, scene, clustered scene).

Based on this video content hierarchy and extracted video features, we propose a progressive video content access scheme in which we first address the video similarity evaluation scheme at different levels and then integrate the hierarchical video browsing and retrieval for video content access and progressive retrieval. Using hierarchical video browsing, a user is provided with an overview of video content from which a query example can be selected. Then, video retrieval is invoked to produce a list of similar units, and the user can browse the content hierarchy of retrieved results to refine the query. With interactions between the retrieval and browsing, a user's query can be quickly refined to retrieve the unit of interest.

The remainder of this paper is organized as follows. Section III presents several video analysis and feature extraction techniques, including camera motion classification and key-frame extraction schemes. Then, based on extracted video features, Section IV introduces techniques for hierarchical video content organization. In Section V, the video similarity assessment scheme is applied at different levels of the video content hierarchy. Section VI presents techniques that join hierarchical video browsing and retrieval for efficient video content access. The conclusion and remarks are given in Section VII.

III. Video Analysis and Feature Extraction

Most schemes for video feature extraction begin by segmenting contiguous frames into separate shots, and then selecting key-frames to represent shot content. With this scheme, a video database is treated somewhat like an image database, because the motion information in the video (or shot) is missed. In our system, the motion information in the video is detected and extracted as a shot feature to help in identifying video content. We first apply shot segmentation to the video, and then execute the camera motion classification scheme. Based on extracted motion information, a key-frame extraction scheme is proposed and the camera motion in the shot will also be utilized as the features to evaluate similarity between shots.

A great deal of research has been done in shot boundary detection, and many approaches achieve satisfactory performance [1], [19]. In previous work, we have developed a shot segmentation approach with an adaptive threshold selection for break and gradual shot detection [20]. In the next sections, we will introduce the camera motion classification and key-frame extraction schemes.

A. *Camera Motion Classification*

Motion characterization plays an important role in content-based video indexing. It is an essential step in creating compact video representation automatically. For example, a mosaic image can represent a panning sequence [21]; the frames before and after a zoom can represent the zoom sequence. As the research work in [53] has demonstrated, in addition to various visual features, the motion information in video shots can also be explored for content-based video retrieval. Thus, an effective characterization of camera motion greatly facilitates the video representation, indexing, and retrieval tasks. And the proposed multimedia content description standard MPEG-7 [59] has also adopted various descriptors (DS) to qualitatively (different types of motions) and quantitatively (the amount of motions) describe the camera motion in each shot [57], [58].

To extract the camera motion, Ngo *et al.* [22] proposed a method using temporal slice analysis for motion characterization, however, to distinguish different types of motion patterns in the slice is a challenging task for videos with cluttered background or containing moving objects. Srinivasan *et al.* [24] introduced a qualitative camera motion extraction method that separates the optical flow into two parts, parallel and rotation, for motion characterization. Xiong *et al.* [23] presented a method that analyzed spatial optical flow distribution. However, these last two methods can only be used when the *focus of expansion (FOE)* or *focus of contraction (FOC)* [25] is at the center of the image, and this is not always the case in generic videos.

To analyze camera motion in the compressed domain, Tan *et al.* [26], Kobla *et al.* [27] and Dorai *et al.* [28] presented three methods based on motion vectors in MPEG streams. In [26], a 6-parameters transformation model is utilized to classify camera motions into panning, tilting and zooming. The methods in [27], [28] map motion vectors in the current frame into eight directions. Motion classification was developed based on the values in these eight directions. However, these strategies are sensitive

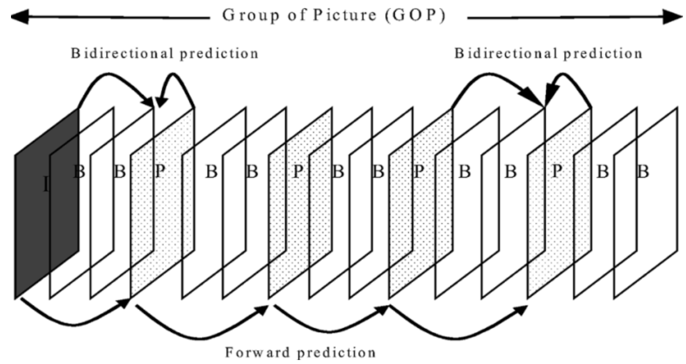


Fig. 2. GOP structure.

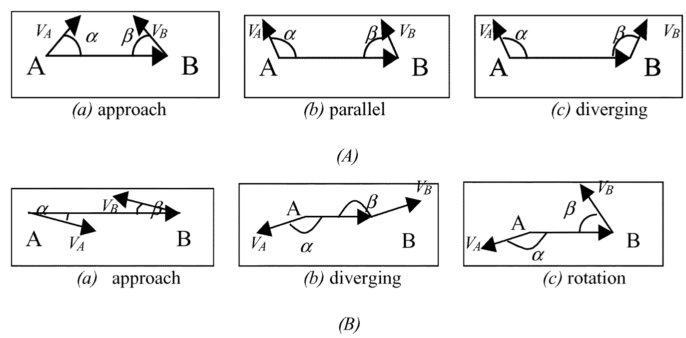


Fig. 3. Mutual relationship between motion vectors. (A) The mutual relationship between motion vectors on the same side of vector $\vec{V}_{AB}$. (B) The mutual relationship between motion vectors on different sides of vector $\vec{V}_{AB}$.

to noise in motion vectors and fail to detect the camera rolling. Furthermore, extracted optical flow or motion vectors may contain considerable noise or errors, which significantly reduces the efficiency of their strategies.

We have found that the statistical information for the mutual relationship between any two motion vectors is relatively robust to noise (see Fig. 4) For a given type of camera motion contained in the current frame, the statistical mutual relationship in the frame will show a distinct distribution tendency. Based on this observation, we propose a qualitative camera motion classification method. In addition to detecting most common camera motions (pan, tilt, zoom, still), our method can also detect camera rolling, and various detected camera motions will directly comply with the motion descriptors in MPEG-7 standard [59].

1) Problem Formulation: Our objective is to efficiently process videos stored in MPEG format for camera motion classification. As shown in Fig. 2, the syntax of MPEG-1 video defines four types of coded pictures: intracoded pictures (I-frames), predicted pictures (P-frames), bidirectionally predicted pictures (B-frames), and DC encoded frames (which are now rarely used). These pictures are organized into sequences of groups of pictures (*GOP*). Each video frame is divided into a sequence of nonoverlapping macroblocks (*MB*), such that each *MB* is then either intracoded or intercoded. An I-frame is completely intracoded, and the *MB* in P-frame may be separated into two types: intracoded (containing forward prediction motion vectors) and intercoded (containing no motion vectors). In this paper, we use only the motion vectors from P-frames, that is, we are sampling the camera motion. For example, if

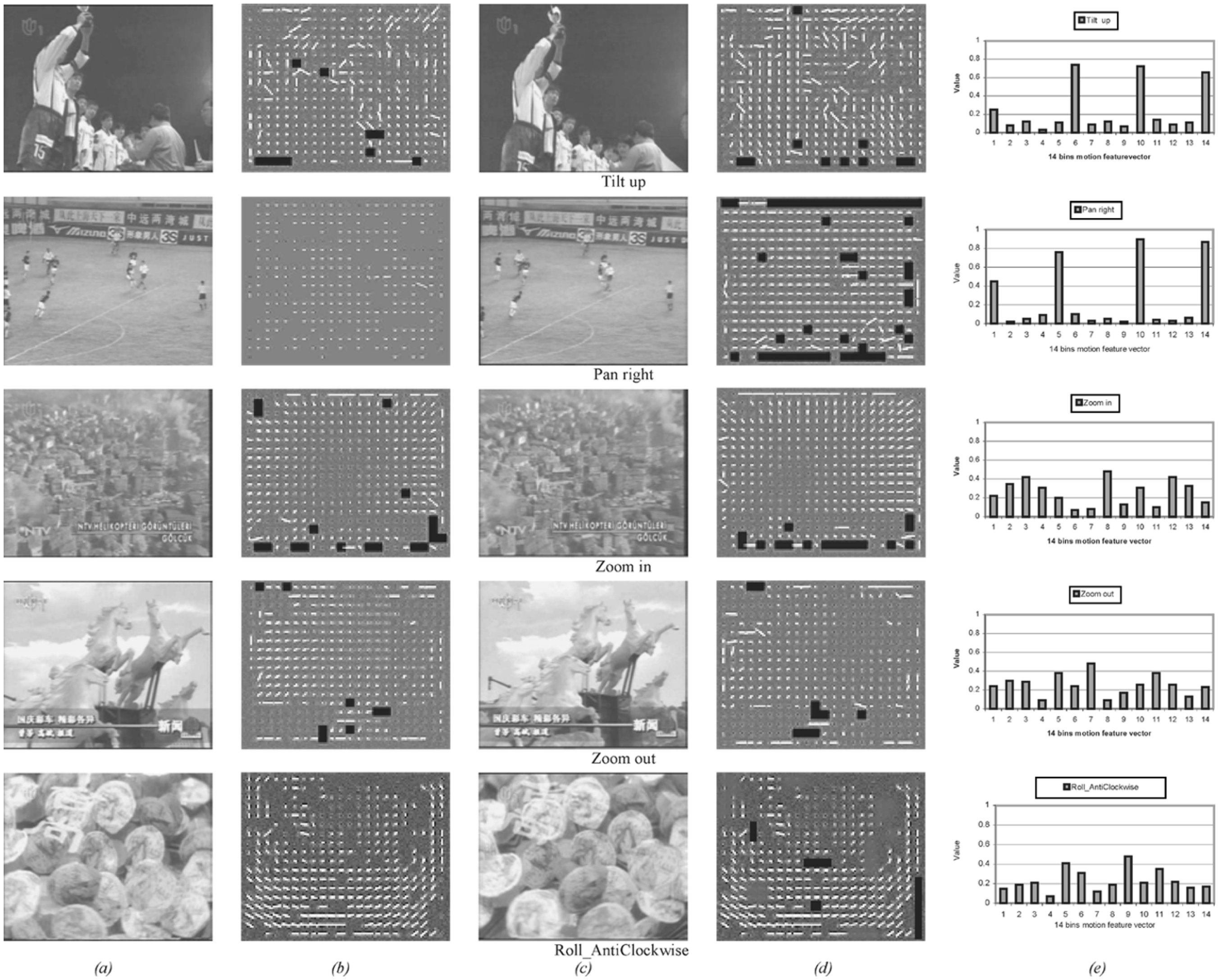


Fig. 4. Relationship between camera motion and motion vectors. The columns *(a), (b), (c), (d),* and *(e)* indicate the current P-frame (P_i), motion vectors in P_i, the succeeding P-frame (P_{i+1}), motion vectors in P_{i+1}, and the 14-bin motion feature vector distribution for *(d)* respectively. The black blocks in motion vectors indicate the "intracoded macroblock"; hence, no motion vector is available for those blocks.

the MPEG video is coded at a rate of 30 frames per second using the *GOP* in Fig. 2, there are 8 P-frames per second in the video. We will require the underlying camera motion rates (per frame) to have a bandwidth of less than 4 Hz. For most videos, this is a reasonable assumption. Using motion vectors from both P and B-frames has the potential to yield better accuracy, but at the cost of increased computation. In our classification scheme, we assume that there is no large object motion or the motion caused by large objects can be ignored. Thus, only the dominant camera motion is detected.

2) The Mutual Relation Between Motion Vectors: Given two points A, B in current frame P_i with positions $p_A = (x_A, y_A)$, $p_B = (x_B, y_B)$ and motion vectors $V_A = (u_A, v_A)$ and $V_B = (u_B, v_B)$, we denote the vector from point A to B as $\vec{V}_{AB}$, and the line crossing point A and B as $y = ((y_A - y_B)/(x_A - x_B))x + (x_A y_B - y_A x_B/(x_A - x_B))$. As shown in Fig. 3, there are four types of mutual relationships between V_A and V_B: approach, parallel, diverging, and rotation.

To classify the mutual relationship between V_A and V_B, we first measure whether they are on the same side [Fig. 3(a)] or different sides [Fig. 3(b)] of vector $\vec{V}_{AB}$. Based on the geometrical relationship among the four points (x_A, y_A), $(x_A + u_A, y_A + v_A)$, (x_B, y_B) and $(x_B + u_B, y_B + v_B)$, it is obvious that if V_A and V_B are on the same side of vector $\vec{V}_{AB}$, both points $(x_A + u_A, y_A + v_A)$ and $(x_B + u_B, y_B + v_B)$ should be above or below the line which crosses point A and B at the same time. Hence, we multiply y_1 and y_2 (from (1)). If the product is nonnegative, we will claim V_A and V_B are on the same side of vector $\vec{V}_{AB}$; otherwise, V_A and V_B are on different sides of vector $\vec{V}_{AB}$

$$\begin{cases} y_1 = y_A + v_A - \frac{y_A - y_B}{x_A - x_B} \cdot (x_A + u_A) - \frac{x_A y_B - y_A x_B}{x_A - x_B} \\ y_2 = y_B + v_B - \frac{y_A - y_B}{x_A - x_B} \cdot (x_B + u_B) - \frac{x_A y_B - y_A x_B}{x_A - x_B} \end{cases}. \quad (1)$$

As shown in Fig. 3, if we assume that α denotes the angle between $\vec{V}_{AB}$ and V_A, and β denotes the angle between V_B

and $\vec{V}_{AB}$, if V_A and V_B are on the same side of $\vec{V}_{AB}$, then their mutual relationship is classified as follows.

- If $\alpha+\beta < 180° - T_{PARA}$, the mutual relationship between V_A and V_B is approach.
- If $\alpha+\beta > 180° + T_{PARA}$, the mutual relationship between V_A and V_B is diverging.
- Otherwise, the mutual relationship between V_A and V_B is parallel.

If V_A and V_B are on different sides of $\vec{V}_{AB}$, then their mutual relationship is classified as follows.

- If $\alpha+\beta < T_{CLOSE}$, the mutual relationship between V_A and V_B is approach.
- If $\alpha+\beta > T_{FAR}$, the mutual relationship between V_A and V_B is diverging.
- Otherwise, the mutual relationship between V_A and V_B is rotation.

In our system, we set T_{PARA}, T_{CLOSE} and T_{FAR} to 15°, 60°, and 250° respectively.

3) The Relationship Between Camera Motion and Motion Vectors: Fig. 4 shows the relationship between the camera motion and motion vectors contained in the frame:

- If the camera pans or tilts, most motion vectors' mutual relationships in the frame are parallel.
- If the motion of the current frame is zooming, most motion vectors' mutual relationships in current frame either approach to (zoom out) *FOC* or diverge from (zoom in) *FOE*.
- If the camera rolls, most vertical vectors' (defined in Section III-A4d) mutual relationship in the frame either approach to (Roll_Clockwise) *FOC* or diverge from (Roll_AntiClockwise) *FOE*.

Based on these observations, a motion feature vector is constructed to characterize the motion vectors.

4) Motion Feature Vector Construction: In this subsection, we introduce four histograms to characterize motion vectors in each frame. A 14-bin feature vector is then formed by packing these four histograms sequentially from bin 1 to bin 14.

a) Motion vector energy histogram (H_{me})*:* For any P-frame P_i and its motion vectors, we assume there are N *MB* contained in P_i. We denote AP_i as the aggregation of all available motion vectors (intercoded *MB*) in P_i, and the number of motion vectors in AP_i is denoted by N_{mv}. Given point A $(P_A = (x_A, y_A))$ in P_i and its motion vector $V_A = (u_A, v_A)$, then (2) defines the energy of V_A

$$||V_A||^2 = u_A^2 + v_A^2. \quad (2)$$

Assuming SP_i denotes the aggregation of motion vectors in AP_i with energy smaller than a given threshold T_{SMALL}, the number of vectors in SP_i is denoted by N_{small}. We calculate the mean μ and variance δ of the motion vectors in SP_i. If we assume LP_i denotes the aggregation of motion vectors in AP_i whose distance to μ is larger than T_{LOC}, and the number of vectors in LP_i is denoted by N_{loc}. The motion vector energy histogram (H_{me}) is constructed using (3).

$$H_{\mathrm{me}}[0] = \frac{(N - N_{\mathrm{mv}} + N_{\mathrm{loc}})}{N}; \quad H_{\mathrm{me}}[1] = \frac{N_{\mathrm{small}}}{N}. \quad (3)$$

In our system, we set $T_{\mathrm{LOC}} = 1.5\delta$ and $T_{\mathrm{SMALL}} = 2$ respectively. In the next section, the motion vectors in aggregation VP_i, $VP_i = AP_i \cap (\overline{SP_i \cup LP_i})$, is referred to as the *valid motion vectors* in P_i, i.e., the *valid motion vectors* are those with relatively high energy and low variance.

b) Motion vector orientation histogram (H_{mo})*:* Clearly, the orientations of the *valid motion vectors* VP_i in P_i will help us determine the direction of the camera motion. For each motion vector $V_A = (u_A, v_A)$ in VP_i, we denote $D(V_A)$ as its orientation, and then divide all valid motion vectors' orientations into four categories: (−45°, 45°), (45°, 135°), (135°, 225°), (225°, 315°). The motion vector orientation histogram is constructed using (4)

$$H_{\mathrm{mo}}(k) = \frac{\sum_{V_A; V_A \in VP_i, -45° + 90° \cdot k < D(V_A) \le 45° + 90° \cdot k} 1}{N_{\mathrm{mv}} - N_{\mathrm{small}}}; \quad k = 0, 1, 2, 3. \quad (4)$$

c) Motion vector mutual relationship histogram (H_{mr})*:* Given two motion vectors in VP_i, their mutual relationship is classified with the strategy given in Section III-A2. The histogram of the mutual relationships in P_i are then calculated and put into different bins of histogram H_{mr}, with $H_{\mathrm{mr}}[0]$, $H_{\mathrm{mr}}[1]$, $H_{\mathrm{mr}}[2]$, and $H_{\mathrm{mr}}[3]$ corresponding to approach, diverging, rotation and parallel, respectively.

d) Motion vector vertical mutual relationship histogram (H_{mvr})*:* As described in Section III-A3, if the camera rolls, the mutual relationships of most motion vectors' vertical lines will approach to *FOC* or diverge from *FOE*. Hence, given any motion vector $V_A = (u_A, v_A)$ in VP_i, its vertical vector is defined by $V'_A = (-v_A, u_A)$, and we can use the strategy in Section III-A2 to calculate the mutual relationship for any two vertical vectors V'_A and V'_B in VP_i. The histogram constructed in this way is denoted as the vertical mutual relationship histogram (H_{mvr}) with $H_{\mathrm{mvr}}[0]$, $H_{\mathrm{mvr}}[1]$, $H_{\mathrm{mvr}}[2]$ and $H_{\mathrm{mvr}}[3]$ representing approach, diverging, rotation and parallel, respectively.

5) Camera Motion Classification: The experimental results in Fig. 4(e) show that for any type of camera motion, the 14-bin motion feature vector will have a distinct distribution mode. For example, when the camera pans, $H_{\mathrm{mr}}[3]$ will contain the largest value in H_{mr}, and the bin with the largest value in H_{mo} will indicate the direction of the panning. For zooming operations, either $H_{\mathrm{mr}}[0]$ or $H_{\mathrm{mr}}[1]$ will have the largest value in H_{mr}. If the camera rolls, $H_{\mathrm{mr}}[2]$ will have the largest value in H_{mr}, and either $H_{\mathrm{mvr}}[0]$ or $H_{\mathrm{mvr}}[1]$ will have the largest value in H_{mvr}. Hence, based on the 14-bin vector, a qualitative camera motion classification strategy is presented.

Input: 14-bin motion feature vector of current P-frame P_i.
Output: The motion category (pan left, pan right, tilt up, tilt down, zoom in, zoom out, roll_clockwise, roll_anticlockwise, still, unknown) P_i belongs to, denoted as "$P_i \leftarrow ?$".
Procedure:

1) If $H_{\mathrm{me}}[0]$ is larger than threshold T_{UNK}, $P_i \leftarrow$ "unknown", otherwise go to step 2.
2) If $H_{\mathrm{me}}[1]$ is larger than threshold T_{STILL}, $P_i \leftarrow$ "still", if not go to step 3.

3) If $H_{\mathrm{me}}[0] + H_{\mathrm{me}}[1]$ is larger than threshold T_{UNION}, $P_i \leftarrow$ "unknown", otherwise, go to step 4.
4) Find the largest and second largest values among H_{mr} and denote them as $H_{\mathrm{mr}}^{\max}$ and $H_{\mathrm{mr}}^{\mathrm{sec}}$ respectively. If the ratio between $H_{\mathrm{mr}}^{\mathrm{sec}}$ and $H_{\mathrm{mr}}^{\max}$ is larger than threshold T_{REL}, $P_i \leftarrow$ "unknown", otherwise, the steps below are used for classification.
- If $H_{\mathrm{mr}}^{\max} = H_{\mathrm{mr}}[0]$, then $P_i \leftarrow$ "zoom out"; else if $H_{\mathrm{mr}}^{\max} = H_{\mathrm{mr}}[1]$, $P_i \leftarrow$ "zoom in".
- If $H_{\mathrm{mr}}^{\max} = H_{\mathrm{mr}}[2]$, go to step 6; else if $H_{\mathrm{mr}}^{\max} = H_{\mathrm{mr}}[3]$, go to step 5.

5) Find the maximal value among H_{mo} and denote it as $H_{\mathrm{mo}}^{\max}$:
- If $H_{\mathrm{mo}}^{\max} = H_{\mathrm{mo}}[0]$, $P_i \leftarrow$ "panning left"; else if $H_{\mathrm{mo}}^{\max}$ equals $H_{\mathrm{mo}}[1]$, $P_i \leftarrow$ "tilting down".
- If $H_{\mathrm{mo}}^{\max} = H_{\mathrm{mo}}[2]$, $P_i \leftarrow$ "panning right"; else if $H_{\mathrm{mo}}^{\max}$ equals $H_{\mathrm{mo}}[3]$, $P_i \leftarrow$ "tilting up".

6) Find the maximal value among H_{mvr} and denote it as $H_{\mathrm{mvr}}^{\max}$:
- If $H_{\mathrm{mvr}}^{\max} = H_{\mathrm{mvr}}[0]$, $P_i \leftarrow$ "roll_anticlockwise"; else If $H_{\mathrm{mvr}}^{\max} = H_{\mathrm{mvr}}[1]$, $P_i \leftarrow$ "roll_clockwise"; Otherwise, $P_i \leftarrow$ "unknown".

The thresholds $T_{UNK}, T_{STILL}, T_{\mathrm{REL}}$ and T_{UNION} may be determined by empirical studies; in our system we set them to 0.55, 0.5, 0.8, and 0.8, respectively.

There is little doubt that some conditions might result in an incorrect classification of the camera motion. Since the camera motion should be consistent over a certain length of time, the temporal filter operation is used to eliminate those errors, that is, any camera motion lasting less than 3 P-frames is absorbed by preceding or succeeding camera motions. The filtered camera motion information is then stored as the motion feature of the shot.

B. Key-Frame Extraction

Key-frame(s) summarize the content of a video shot. Other research has addressed the problem of key-frame extraction [30]–[35], and a recent survey can be found in [29] and [55]. A first attempt in key-frame extraction was to choose the frame appearing at the beginning of each shot as the key-frame [35]. However, if the shot is dynamic, this strategy will not provide good results. In order to address this problem, clustering [31] and low-level features [30] are utilized for key-frame extraction by clustering all frames into M clusters or calculating accumulated frame differences. Due to the fact that the motions in video shots imply the content evolution and changes, a motion activity-based key-frame extraction method has been proposed in [54], [55], where the MPEG-7 motion intensity in each shot is used to guide the key-frame selection. Given a user specified number, the system selects the corresponding number of key-frames by using the cumulative motion intensity, where more key-frames are extracted from the high motion frame regions. However, determining the number of key-frames that optimally addresses the video content change is a difficulty. On the other hand, if there is a large camera or object motion in the shot, the selected key-frames may be blurred, and thus not suitable for the key-frame.

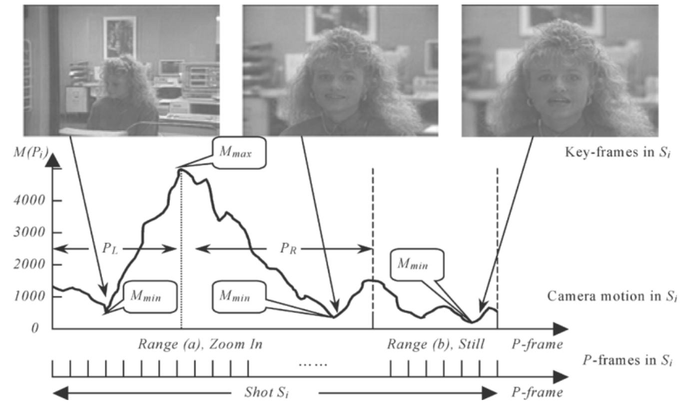


Fig. 5. Camera motion based key-frame selection.

The authors in [34] and [32] avoid these problems by proposing threshold-free methods for extracting key-frames. In [34], the temporal behavior of a suitable feature vector is followed along a sequence of frames, and a key-frame is extracted at each point on the curve where the magnitude of its second derivative reaches the local maximum. A similar approach is presented in [32], where the local minima of motion is utilized for key-frame extraction. However, two problems remain: 1) locating the best range to find the local minimum is also determined by a critical threshold and 2) since the small motion of the video can cause large variations in the optical flow, these methods may focus on many activity details but not the shot content overview.

To extract key-frames using these strategies, the video must be fully decoded. In the next section, we introduce a threshold-free method that extracts key-frames in the compressed domain. Our method is based on the method from literature [32], however, there are several distinguishing elements: 1) our method is executed in the compressed domain (only a very limited number of frames need to be decoded); 2) instead of using optical flow, we use motion vectors from the MPEG video; and 3) instead of using the threshold, we use camera motions in the shot to determine the local maximum or minimum.

1) The Algorithm: Our key-frame extraction algorithm is executed using the following steps.

1) Given any shot S_i, use the camera motion classification and temporal motion filter to detect and classify the camera motions, as shown in Fig. 5.
2) Find the representative frame(s) for each type of motion (see Fig. 5), and the collection of all representative frames is taken as the key-frames for S_i.

From the start frame to the end frame in shot S_i, for any given P-frames (P_i), denote AP_i the aggregation of all available motion vectors in P_i, then (5) is used to calculate the motion magnitude of P_i

$$M(P_i) = \sum_{V_k, V_k=(u_k,v_k), V_k \in AP_i, AP_i \subset P_i} \left(u_k^2 + v_k^2\right) \quad (5)$$

where V_k denotes the motion vectors in P_i. Given P_i, its $M(P_i)$ is influenced by two factors.

- The motion information contained in P_i. The smaller the amount of motion, the smaller $M(P_i)$ is.

- The number of intracoded *MB* in P_i. The more intracoded *MB*, the smaller $M(P_i)$ is.

We then determine the motion magnitude for each P-frame in S_i. These values will help us select the representative frame for each type of camera motion in S_i. As shown in Fig. 5, from the start frame of S_i to the end frame, we sequentially select one type of camera motion and execute the following steps.

1) If the camera motion is still (as shown in Fig. 5 *range (b)*), find the smallest value of $M(P_i)$ among all frames in that range, and denote it as $M_{\min}$. The corresponding P-frame is selected as the representative frame for all frames in this range.
2) For all other types of camera motion, find the largest value of $M(P_i)$ among all frames in that range [as shown in Fig. 5 *range (a)*]. Denote this value as $M_{\max}$. We then use $M_{\max}$ to separate the frames in this range into two separate and consecutive parts, P_L and P_R, as shown in Fig. 5.
3) For any part of P_L and P_R, find the smallest value of $M(P_i)$ among all frames in that part, and use the corresponding P-frame as the representative frame. Denote the selected representative frame for part P_L and P_R as R_{P_L} and R_{P_R} respectively.
4) Some small camera motions may cause very little frame difference, and two representative frames from this range might be redundant. Hence, we calculate the visual feature based frame difference between R_{P_L} and R_{P_R} (using the (29) introduced in Section V-A). If this value is smaller than threshold $T_{\rm merg}$ ($T_{\rm merg} = 0.35$ in our system), only R_{P_R} is used as the representative frame in the range. Otherwise both R_{P_L} and R_{P_R} are selected as the representative frames.
5) Iteratively execute steps 1–4 until all camera motions in S_i have been processed successfully, the selected representative frames are considered the key-frames for S_i.

Since content changes in shots are usually caused by camera motions, we first use the camera motion classification strategy to separate the frames in S_i into different ranges, with each range containing one distinct camera motion. The collection of representative frames for all ranges forms the set of key-frames for the shot. Furthermore, the representative frames are selected with the local minimal $M(P_i)$, so they will have higher definition and more “fresh” content.

By adopting motion activity in camera motion selection, our method is also similar to the scheme in [54], [55]. However, there are two key distinctions: 1) with the method in [54], [55], it is the authors but not the system that determine the number of key-frames to be extracted from each shot. Given a video that contains hundreds of shots, it would be very hard (or even unreasonable) for users to specify the number of key-frames for each shot. Consequently, the naïve users may simply specify a constant key-frame number for all shots. In that case, the proposed scheme may introduce redundancy in low motion shots and miss the content change in high motion shots and 2) the method in [54] and [55] does not consider the local motion minimum but uses only the accumulative motion activity, as a result, the select key-frames may be blurred and not clear enough for the content presentation purpose.

TABLE I
CAMERA MOTION CLASSIFICATION RESULT

Camera Motion	Frame Numbers	P-Frame Numbers	Precision (A)	Precision (B)
Pan	7780	2022	0.84	0.82
Tilt	2004	501	0.85	0.81
Zoom	2948	761	0.73	0.65
Rotation	890	233	0.65	*N/A*
Still	4589	1684	0.87	0.84
Average	*20011*	*5201*	*0.804*	*0.756*

C. Experimental Results

1) Camera Motion Classification Results: Table I shows the results produced by our camera motion detection algorithm. We evaluated the efficiency of our algorithm (denoted by A) through an experimental comparison with the transformation model based method [26] [1] (denoted by B). Several standard MPEG-I streams (about 11 711 frames) were downloaded from http://www.open-video.org and used as our test bed. One edited MPEG-I file (about 16 075 frames) containing a large number of zooming and roll motions was also used as a test dataset. For better evaluation, the precision defined in (6) is used, where n_c, n_f denote the correctly and falsely detected camera motion in the P-frames

$$Precision = \frac{n_c}{(n_c + n_f)}. \tag{6}$$

Among all 27 786 frames in the video, the sequential frame regions with a distinct camera motion (pan, tilt, zoom, roll, still) are selected as our ground truth. These frames (about 20 011 frames) occupy about 72% of the entire video, and contain about 5201 P frames. Our experiment is executed with these 5201 P-frames. From Table I, we find that, on average, our method has a precision of approximately 80.4%, about 5% higher than the transformation model based method [26]. In detecting pure panning and tilting, both methods have about the same precision. However, while some abnormal motion vectors caused by objects motion or other reasons contained or *FOE/FOC* is not at the center of the image, the efficiency of this method is rather reduced, since those motion vectors cannot be characterized by the proposed transformation model. However, our method is a statistical strategy, the abnormal or distorted motion vectors would not have much influence on unfolding the dominant camera motion in the frames, thus resulting in a relatively higher precision. Furthermore, while method B is not able to detect roll motion, our method produces a precision of 68% for roll detection.

On a *PC* with *PIII 900-MHz CUP*, the average time to process one P-frame is three times faster than real time and four times faster than method B.

2) Key-Frame Extraction Results: Since there is no comprehensive user study that validates the applicability of key-frames

[1]**Remark**: We compare our method with the method in literature [32], since it also works in compressed domain and utilizes only the motion vector of P-frame for classification.

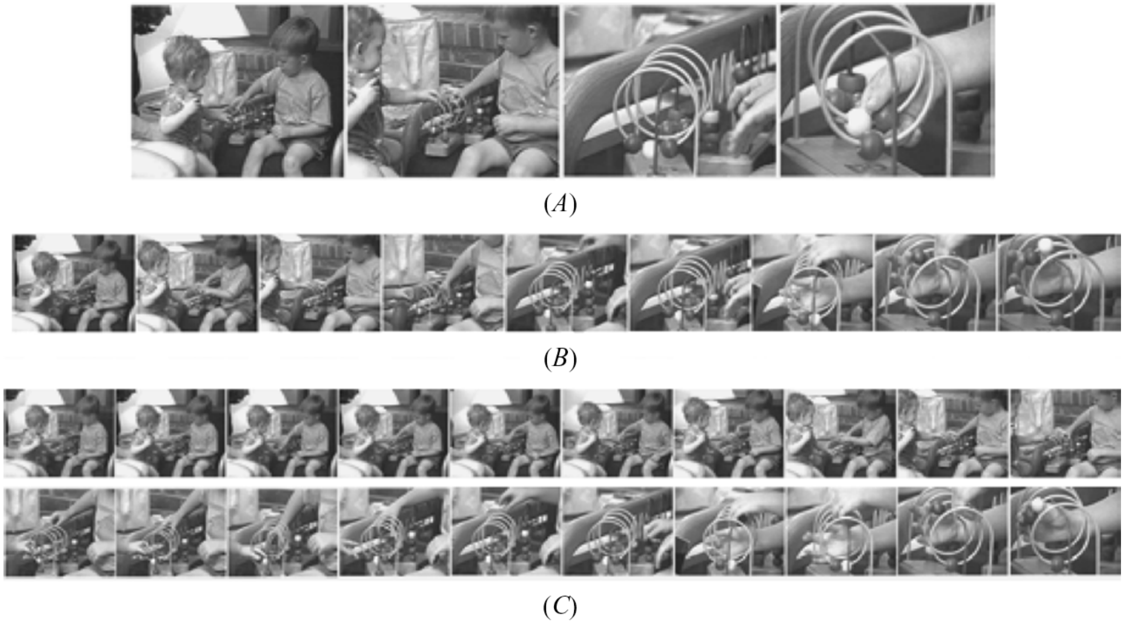


Fig. 6. Key-frame extraction results. (a) Indicates the results of our method. (b) Indicates results with the method in [32]. (c) Represents the sampling of the shot with 15 frames stepsize, from top left to bottom right;

extracted with different methods, a quantitative comparison between our method and other strategies is not available. We thus present some pictorial experimental results. Fig. 6 illustrates a comparison of our strategy with the literature [32]. (Instead of using optical flow, we use motion vector to calculate $M(t)$). Fig. 6(c) denotes the sampled frames in shot S_i with a stepsize of 15 frames. The shot starts with a still camera motion focusing on the two children playing with the rings. Then, the camera zooms in to emphasize the rings. Finally, some close up frames of the ring are shown. With our camera motion classification strategy, this shot was separated into three motion ranges: still, zoom in, irregular (due to the motion of the hands) sequentially. Hence, four key-frames [as shown in Fig. 6(a)] are extracted with our method: the first key-frame is produced by still, the second and third are produced by zoom in, and the last one is produced by irregular motion (since the two representative frames in the irregular motion range are similar, only one is used.) Using the strategy in [32], nine key-frames are extracted, as shown in Fig. 6(b), where most of the details of the shot have also been addressed, even the movement of the hand. Although our strategy did lose some detail information, the content information is very well maintained. We believe that key-frames should be used to get the overview of the video shot content (not the details), and hence, we believe our method maintains a relatively good balance between overall shot content and details.

IV. Hierarchical Video Content Organization

Generally, videos can be represented using a hierarchy of five levels (video, scene, group, shot, and key-frame),[2] increasing in granularity from top to bottom. Much research has addressed the problem of constructing semantically richer video entities by visual feature based shot grouping [35]–[38] or joint semantic rules and knowledge information for scene detection [39]–[41], [50]. However, these strategies only solve the problem of semantic units detection and visualization. Since similar scenes may appear repeatedly in a video, redundant scene information should be reduced by clustering beyond the scene level. In this way, a concise video content table can be created for hierarchical browsing or summarization. Instead of using the semantic unit for video content table construction, other strategies utilize the video shot (or key-frames) based clustering strategy [1], [2], [42] to construct video content hierarchy. However, the constructed hierarchy just addresses some low-level feature based frame differences.

To address this problem, we generate a three level hierarchy from clustered scenes to groups. By integrating video key-frames and shots, a five level video content hierarchy (clustered scene, scene, group, shot, key-frame) is successfully constructed.

As shown in Fig. 1, we construct the video content hierarchy in three steps: 1) group detection; 2) scene detection; and 3) scene clustering. The video shots are first grouped into semantically richer groups. Then, similar neighboring groups are merged into scenes. Beyond the scene level, a pairwise cluster scheme is utilized to eliminate repeated scenes in the video, thus reducing the redundant information. Using the content structure constructed by this strategy, the hierarchical video browsing and summarization is accessed directly. In addition, we have also addressed the problem of the representative unit selection for groups, scenes, and clustered scene units for visualizing the generated video content information.

Generally, the quality of most proposed methods is heavily based on the selection of thresholds [36]–[38], however, the content and low-level features among different videos vary greatly. Thus, we use the entropic thresholding technique to select the

[2]***Remark***: In this paper, the video group and scene are defined as in [37]: 1) a *video scene* is a collection of semantically related and temporally adjacent shots, depicting and conveying a high-level concept or story and 2) a *video group* is an intermediate entity between the physical shots and semantic scenes; examples of groups are temporally or spatially related shots.

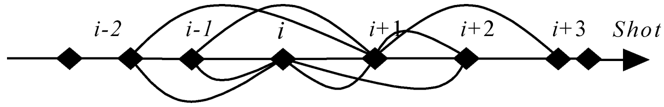


Fig. 7. Shot grouping strategy.

optimal threshold for video group and scene detection; it has been shown to be highly efficient for the two-class data classification problem.

A. Video Group Detection

The shots in one group usually share similar background or have a high correlation in time series. Therefore, to segment spatially or temporally related video shots into groups, a given shot is compared with the shots that precede and succeed it (no more than two shots) to determine the correlation between them, as shown in Fig. 7. Assume $StSim(S_i, S_j)$ denotes the similarity between shot S_i and S_j, which was given in (35). Our group detection procedure is stated as below.

Input: Video shots. **Output**: Video groups
Procedure:

1) Given any shot S_i, if CR_i is larger than TH_2:
   a) If $R(i)$ is larger than TH_1, claim a new group starts at shot S_i.
   b) Otherwise, go to step 1 to process other shots.
2) Otherwise:
   a) If both CR_i and CL_i are smaller than TH_2, claim a new group starts at shot S_i.
   b) Otherwise, go to step 1 to process other shots.
3) Iteratively execute step 1 and 2 until all shots are parsed successfully.

The definitions of CR_i, CL_i, $R(i)$ are given in (7)–(11)

$$CL_i = Max\{StSim(S_i, S_{i-1}), StSim(S_i, S_{i-2})\} \quad (7)$$
$$CR_i = Max\{StSim(S_i, S_{i+1}), StSim(S_i, S_{i+2})\} \quad (8)$$
$$CL_{i+1} = Max\{StSim(S_{i+1}, S_{i-1}), StSim(S_{i+1}, S_{i-2})\} \quad (9)$$
$$CR_{i+1} = Max\{StSim(S_{i+1}, S_{i+2}), StSim(S_{i+1}, S_{i+3})\} \quad (10)$$
$$R(i) = \frac{(CR_i + CR_{i+1})}{(CL_i + CL_{i+1})}. \quad (11)$$

Since closed captions and speech information are not available in our strategy, the visual features such as color and texture play an essential role in determining the shots in one group. Hence, to calculate the similarity between S_i and S_j with (35), we set W_H, W_M, W_F and W_L equal to 0.5, 0.0, 0.5, and 0.0, respectively; that is, we use only the visual features for similarity evaluation. Meanwhile, to evaluate the similarity between key-frames K_i and K_j with (29), we set W_c, W_T equal to 0.7 and 0.3 respectively.

Using the shot grouping strategy above, two kinds of shots are absorbed into a given group (as shown in Fig. 8): 1) shots related in temporal series, where similar shots are shown back and forth. Shots in this group are *temporally related* and 2) shots similar in visual perception, where all shots in the group are similar in visual features. Shots in this group are *spatially related*.

1) Group Classification and Representative Shot Selection: Given any group G_i, we assign it to one of two categories: temporally vs spatially related group. Assuming there are T shots (S_i, $i = 1, \ldots, T$) contained in G_i, the group classification strategy is described below.

Input: Video group G_i, and shots S_i $(i = 1, \ldots, T)$ in G_i. **Output**: Clusters (C_{Nc}, $N_c = 1, \ldots, U$) of shots in G_i.
Procedure:

1) Initially, set variant $N_c = 1$, cluster C_{Nc} has no members.
2) Select the shot (S_k) in G_i with the smallest shot number as the seed for cluster C_{Nc}, and subtract S_k from G_i. If there are no more shots contained in G_i, go to step 5.
3) Calculate the similarity between S_k and other shot S_j in G_i, If $StSim(S_k, S_j)$ is larger than threshold T_h, absorb shot S_j in cluster C_{Nc}. Subtract S_j from G_i.
4) Iteratively execute step 3, until there are no more shots that can be absorbed in current cluster C_{Nc}. Increase N_c by 1 and go to step 2.
5) If N_c is larger than 1, we claim G_i is a *temporally related* group, otherwise, it is a *spatially related* group.

After the video group has been classified, the representative shot(s) of each group are selected to represent and visualize the content information in G_i. We denote this procedure as *SelectRepShot()*. The key-frames of all representative shot(s) are selected as representative frames for the video group.

[SelectRepShot]

The representative shot of group G_i is defined as the shot that represents the most content in G_i. As semantic content is not available in our system, visual features information is used to select representative shots. We have merged all shots in G_i into N_c clusters, and these clusters help us select the representative shots. Given group G_i with N_c clusters (C_i), we denote by $ST(C_i)$ the number of shots contained in cluster C_i. The representative shot of G_i is selected as follows.

1) Given N_c clusters C_i $(i = 1, \ldots, N_c)$ in G_i, use steps 2, 3, and 4 to extract one representative shot for each cluster C_i. In all, N_c representative shots will be selected for each G_i.
2) Given any cluster C_i with more than two shots, the representative shot of $C_i(R_S(C_i))$ is obtained from (12)

$$R_s(C_i) = \arg\max_{S_j \atop 1 \le j \le ST(C_i);\ S_j \subset C_i} \left\{ \frac{1}{ST(C_i)} \sum_{k=1}^{ST(C_i)} StSim(S_j, S_k); S_k \subset C_i \right\}. \quad (12)$$

3) If there are two shots contained in C_i, the shot that has more key-frames usually has more content information, and hence is selected as the representative shot for C_i. If all shots in C_i have the same key-frame numbers, the shot with larger time duration is selected as the representative shot.

If there is only one shot contained in cluster C_i, it is selected as the representative shot for C_i.

Fig. 8. Examples of detected video groups.

B. Group Merging for Scene Detection

Since our shot grouping strategy places more emphasis on the details of the scene, one scene may be grouped into several groups, as shown in Fig. 8. However, groups in the same scene usually have higher correlation with each other when compared with other groups in different scenes. Hence, a group merging method is introduced to merge adjacent groups with higher correlation into one scene.

Input: Video groups ($G_i, i = 1, \ldots, M$) **Output**: Video scenes ($SE_j, j = 1, \ldots, N$).
Procedure:

1) Given groups G_i, $i = 1, \ldots, M$, calculate similarities between all neighboring groups ($SG_i, i = 1, \ldots, M-1$) using (13), where $GpSim(G_i, G_j)$ denotes the similarity between group G_i and G_j (defined in (37))

$$SG_i = GpSim(G_i, G_{i+1}) \quad i = 1, \ldots, M-1 \tag{13}$$

2) Use the automatic threshold detection strategy in Section IV-D to find the best group merging threshold (T_G) for $SG_i, i = 1, \ldots, M-1$, with $T_G = ATD(SG_i)$.
3) Adjacent groups with similarity larger than T_G are merged into a new group. If there are more than two sequentially adjacent groups with larger similarity than T_G, all are merged into a new group.
4) The reserved and newly generated groups are formed as a video scene. Scenes containing only two shots are eliminated, since they usually convey less semantic information than scenes with more shots. The *SelectRepGroup()* strategy is used to select the representative group for each scene.

[SelectRepGroup]

For any scene SE_i, the representative group is the group in SE_i that contains the most content information for SE_i. As noted previously, we use the low-level features associated with each group in our strategy.

1) For any scene SE_i that contains three or more groups $G_j(j = 1, \ldots, N_i)$, the representative group of $SE_i(R_p(SE_i))$ is given by (14), as shown at the bottom of the page. That is, $R_p(SE_i)$ is the group in SE_i which has the largest average similarity with all other groups.
2) If there are only two groups in SE_i, we use the average motion information and the time duration of the group as the measurement. Usually, a group containing more motion will have more key-frames. Hence, we calculate the ratio between the sum of key-frame numbers and shot numbers in each group, and choose the one with the highest ratio as the representative group. If both groups

$$R_p(SE_i) = \arg\max_{\substack{G_j \\ 1 \le j \le N_i}} \left\{ \frac{1}{N_i} \sum_{k=1}^{N_i} GpSim(G_j, G_k); G_k \subset SE_i, G_j \subset SE_i \right\}. \tag{14}$$

have the same ratio, the group with longer time duration is selected as the representative group.

3) If there is only one group in SE_i, this group is selected as the representative group of SE_i.

In the sections below, the selected representative group $R_p(SE_i)$ is also taken as the centroid of SE_i.

C. Video Scene Clustering

Using the results of group merging, video scene information can be constructed. However, in most situations, many similar scenes would appear several times in the video. Clustering those similar scenes into one unit can eliminate redundancy and produce a more concise video content summary. Since the general K-meaning cluster algorithm needs to seed the initial cluster center, and the initial guess of cluster centroids and the order in which feature vectors are classified can affect the clustering result, we introduce a seedless *pairwise cluster scheme (PCS)* for video scene clustering.

Input: Video scenes ($SE_j,\ j = 1,\ldots,M$) and all member groups ($G_i, i = 1,\ldots,NG$).

Output: Clustered scene structure ($SE_k, k = 1,\ldots,N$).

Procedure:

1) Given video groups $G_i, i = 1,\ldots,NG$, we first calculate the similarities between any group G_i and G_j ($i = 1,\ldots,NG-1; j = 1,\ldots,NG-1$). The similarity matrix (SM_{ij}) for all groups is computed using (15)

$$SM_{ij}(G_i,G_j) = GpSim(G_i,G_j), \quad i=1,\ldots,NG-1;\ j=1,\ldots,NG-1 \quad (15)$$

where $GpSim(G_i,G_j)$ denotes the similarity between G_i and G_j given by (37). For any scene SE_j, it consists of either one or several groups. Hence, the similarity matrix of all scenes (SM'_{ij}) can be derived from the group similarity matrix (SM_{ij}) with (16)

$$SM'_{ij}(SE_i,SE_j) = GpSim\left(R_p(SE_i),R_p(SE_j)\right); \quad i=1,\ldots,M;\ j=1,\ldots,M. \quad (16)$$

2) Find the largest value in matrix SM'_{ij}, and merge the corresponding scenes into a new scene, and use *SelectRepGroup()* to find the representative group (scene centroid) for a newly generated scene.

3) After we have obtained the desired number of clusters, go to the end; if not, go to step 4.

4) Based on the group similarity matrix SM_{ij} and the updated centroid of the newly generated scene, update the scene similarity matrix SM'_{ij} with (16) directly, then go to step 2.

In order to determine the end of the scene clustering at step 3, the number of clusters N needs to be explicitly specified. Our experimental results have shown that for a great deal of interesting videos, if we have M video scenes, then using a clustering algorithm to reduce the number of scenes by 40% produces a relatively good result with respect to eliminating the redundancy and reserving important video scenes. However, a fixed threshold often loses the adaptive ability of the algorithm. Hence, to find an optimal number of clusters, we have employed the cluster validity analysis [49]. The intuitive approach is to find cluster numbers that minimize intra-cluster distance while maximizing the inter-cluster distance. Assume that N indicates the number of clusters. Then the optimal cluster would result in measurement $\rho(N)$ with the smallest value, where $\rho(N)$ is defined in (17)

$$\rho(N) = \frac{1}{N}\sum_{i=1}^{N} \max_{1\le j\le N}\left\{\frac{\varsigma_i+\varsigma_j}{\xi_{ij}}\right\} \quad (17)$$

$$\varsigma_i = \frac{1}{N_j}\sum_{i=1}^{N_j}\left(1-GpSim\left(C_i^j,u_j\right)\right); \qquad \xi_{ij} = 1-GpSim(u_i,u_j) \quad (18)$$

and N_j is the number of scenes in cluster j, u_j is the centroid of the cluster j. ς_i is the intra-cluster distance of the cluster i, ξ_{ij} is the inter-cluster distance of clusters i and j, and $C_{\min}$, $C_{\max}$ are the range of the cluster number we seek for optimal values. We set these two numbers $C_{\min} = [M\cdot 0.5]$ and $C_{\max} = [M\cdot 0.7]$, where the operator $[x]$ indicates the maximal integer which is not larger than x. That is, we seek an optimal cluster number by clustering 30% to 50% of the original scenes (M). Hence, the optimal number of cluster $\hat{N}$ is selected as

$$\hat{N} = \underset{C_{\min}\le N\le C_{\max}}{Min}(\rho(N)) \quad (19)$$

D. Automatic Threshold Detection (ATD)

As we have discussed in sections above, thresholds TH_1, TH_2, and T_G are the key elements for group and scene detection. An entropic threshold technique is applied in this section to select the optimal thresholds for these three factors. A fast entropy calculation method is also presented. To illustrate, assume the maximal difference of $R(i)$ in (11) is in the range $[0,M]$. In an input *MPEG* video, assume there are f_i shots whose $R(i)$ has the value i ($i \in [0,M]$). Given a threshold, say T, the probability for the *nongroup-boundary* shots $P_n(i)$ and *group-boundary* shots $P_e(i)$ can be defined as (20) and (21), respectively

$$P_n(i) = \frac{f_i}{\sum_{h=0}^{T} f_h}, \quad 0\le i\le T \quad (20)$$

$$P_e(i) = \frac{f_i}{\sum_{h=T+1}^{M} f_h}, \quad T+1\le i\le M \quad (21)$$

where $\sum_{h=0}^{T} f_h$ gives the total number of shots with ratio $R(i)$ in the range $0\le R(i)\le T$. The entropies for these two classes are then given by

$$H_n(T) = -\sum_{i=0}^{T} P_n(i)\log P_n(i);$$

$$H_e(T) = -\sum_{i=T+1}^{M} P_e(i)\log P_e(i). \quad (22)$$

The optimal threshold vector T_C for classification has to satisfy the following criterion function [52]:

$$H(T_c) = \max_{T=0,\ldots,M}\left\{H_n(T)+H_e(T)\right\}. \quad (23)$$

To find the global maximum of (23), the computational burden is bounded by $O(M^2)$. To reduce the search burden, a fast search algorithm is proposed which exploits the recursive iterations for the probability calculations for $P_{\rm n}(i)$, $P_{\rm e}(i)$ and the entropies $H_n(T)$, $H_e(T)$, where the computational burden is induced by calculating the re-normalized part repeatedly. We first define the total number of the pairs in the *nongroup-boundary* and *group-boundary* classes [the re-normalized parts used in (20) and (21)] when the threshold is set to T

$$P_0(T) = \sum_{h=0}^{T} f_h; \quad P_1(T) = \sum_{h=T+1}^{M} f_h \tag{24}$$

The corresponding total number of pairs at global threshold $T+1$ can be calculated as

$$\begin{aligned} P_0(T+1) &= \sum_{h=0}^{T+1} f_h = \sum_{h=0}^{T} f_h + f_{T+1} \\ &= P_0(T) + f_{T+1} \\ P_1(T+1) &= \sum_{h=T+2}^{M} f_h = \sum_{h=T+1}^{M} f_h - f_{T+1} \\ &= P_1(T) - f_{T+1}. \end{aligned} \tag{25}$$

The recursive iteration property of the two corresponding entropies can then be exploited as denoted by (26).

The recursive iteration is reduced by adding only the incremental part, and the search burden is reduced to $O(M)$. We denote the above automatic threshold selection strategy as ATD. The optimal threshold for Th_1 is determined with $Th_1 = ATD(R(i))$. The same strategy can be applied to find the optimal threshold for Th_2 and T_G, with $Th_2 = Min(ATD(CR_i), ATD(CL_i))$ and $T_G = ATD(SG_i)$.

Figs. 8 and 9 present the experimental results for video group and scene detection. By utilizing the automatic threshold detection, most groups and scenes are correctly detected

$$\begin{aligned} &H_n(T+1) \\ &= -\sum_{i=0}^{T+1} \frac{f_i}{P_0(T+1)} \log \frac{f_i}{P_0(T+1)} \\ &= -\frac{P_0(T)}{P_0(T+1)} \sum_{i=0}^{T+1} \frac{f_i}{P_0(T)} \log \left\{ \frac{f_i}{P_0(T)} \frac{P_0(T)}{P_0(T+1)} \right\} \\ &= \frac{P_0(T)}{P_0(T+1)} H_n(T) - \frac{f_{T+1}}{P_0(T+1)} \log \frac{f_{T+1}}{P_0(T+1)} \\ &\quad - \frac{P_0(T)}{P_0(T+1)} \log \frac{P_0(T)}{P_0(T+1)} \\ &H_e(T+1) \\ &= -\sum_{i=T+2}^{M} \frac{f_i}{P_1(T+1)} \log \frac{f_i}{P_1(T+1)} \\ &= -\frac{P_1(T)}{P_1(T+1)} \sum_{i=T+2}^{M} \frac{f_i}{P_1(T)} \log \left\{ \frac{f_i}{P_1(T)} \frac{P_1(T)}{P_1(T+1)} \right\} \\ &= \frac{P_1(T)}{P_1(T+1)} H_e(T) + \frac{f_{T+1}}{P_1(T+1)} \log \frac{f_{T+1}}{P_1(T+1)} \\ &\quad - \frac{P_1(T)}{P_1(T+1)} \log \frac{P_1(T)}{P_1(T+1)}. \end{aligned} \tag{26}$$

E. Scene Detection Experimental Results

Table II presents the experimental results and comparisons between our scene detection algorithm and other strategies [36], [37]. The scene detection is executed among two medical videos and four news programs. As *scene* is a semantic level concept, an absolute scene boundary cannot be concretely defined in common videos (especially in medical videos, where the story unit boundary is not distinct). However, we believe that in semantic unit detection, it is often worse to fail to segment distinct boundaries than to oversegment a scene. Hence, to judge the quality of the detected results, the following rule is applied: the scene is judged to be correctly detected if and only if all shots in the current scene belong to the same semantic unit (scene), otherwise the current scene is judged to be falsely detected. Thus, the scene detection precision (P) in (27) is utilized for performance evaluation

$$P = \frac{\#\ of\ correctly\ detected\ scenes}{\#\ of\ detected\ scenes}. \tag{27}$$

Clearly, without any scene detection (treating each shot as one scene), the scene detection precision would be 100%, and hence another *compression rate factor (CRF)* is defined in (28)

$$CRF = \frac{\#\ of\ detected\ scene}{\#\ of\ shot\ in\ the\ video}. \tag{28}$$

To distinguish our method with others, we denote our method as A, the other two methods in [36] and [37] as B and C, respectively. From the results in Table II, some observations can be made: 1) our scene detection algorithm achieves the best precision among all three methods, about 67% shots are assigned in the right semantic unit; 2) method C [36] achieves the highest compression rate, unfortunately the precision of this method is also the lowest. On the other hand, this strategy is a threshold based method, and hence there is no doubt that some scenes are over segmented or missed; and 3) as a tradeoff with the precision, the compression ratio of our method is the lowest (8.6%) (each scene consists of about 11 shots). However, as previously mentioned, during video browsing or retrieval, it is worse to fail to segment distinct boundaries than to oversegment a scene. From this point of view, our method is better than the other two methods.

V. Video Similarity Assessment

In measuring the similarity between videos, Dimitrova *et al.* [44] regarded the average distance of corresponding frames between two videos as the similarity measure, and took the temporal order of the frames into account. Lienhart *et al.* [45] considered the video similarity from different hierarchies, and defined the measure by different degrees of aggregation based on either a set or a sequence representation. Adjeroh *et al.* [43] formulated the problem of video sequence-to-sequence matching as a pattern matching problem and introduced new "string edit" operations required for the special characteristics of video sequences. Zhao *et al.* [46] presented a method to use feature lines [47] to evaluate the distances between the query image and video shot. To consider the influencing factors of the subjectivity of humans, Liu *et al.* [48] presents a video retrieval system to simulate the visual judgment of a human.

Fig. 9. Examples of detected video scenes.

TABLE II
VIDEO SCENE DETECTION RESULTS

Movie content	*Shots*	*Method A*			*Method B*			*Method C*		
		Scenes	*P*	*CRF*	*Scenes*	*P*	*CRF*	*Scenes*	*P*	*CRF*
Medical 1	265	29	0.69	0.23	23	0.63	0.13	21	0.52	0.098
Medical 2	221	26	0.54	0.32	21	0.57	0.17	17	0.50	0.081
News 1	189	25	0.71	0.31	22	0.76	0.12	16	0.64	0.074
News 2	178	19	0.65	0.26	13	0.68	0.15	14	0.60	0.101
News 3	214	36	0.72	0.27	24	0.63	0.11	17	0.55	0.107
News 4	190	27	0.68	0.31	21	0.59	0.14	14	0.57	0.100
Average	*1889*	*162*	*0.665*	*0.086*	*124*	*0.643*	*0.0656*	*99*	*0.563*	*0.052*

Unfortunately, all these methods ignore the fact that video not only consists of shots and frames, it is also constructed with video groups and scenes that vary in semantic content and visual features. Hence, the video similarity evaluation should consider the similarity between groups and scenes.

A. Frame Level Similarity Evaluation

At the frame level, two types of visual features are extracted: 256-bin dimensional HSV color histogram and 10-bin dimensional tamura coarseness texture. Given frame F_i, we denote its normalized color histogram and texture as $H^l_{F_i}$, $T^k_{F_i}$, where $l \in [0, 255]$, $k \in [0, 9]$. Then the similarity between F_i and F_j is given by (29)

$$FmSim(F_i, F_j) = W_c \cdot \sum_{l=0}^{255} Min\left(H^l_{F_i}, H^l_{F_j}\right) + W_T \cdot \left(1 - \sqrt{\sum_{k=0}^{9}\left(T^k_{F_i} - T^k_{F_j}\right)^2}\right). \quad (29)$$

B. Shot-Level Similarity Evaluation

At the shot level, four kinds of low-level features were extracted: average color histogram, camera motion, key-frame information, and shot length. Given shot S_i and S_j, while calculating their similarity, both features match *degree* and match *order* are taken into account.

1) Average Color Histogram Matching: An average color histogram $\overline{H}_{S_i}$ (in HSV space) defined by (30) is used to describe the average color information of S_i, where M is the number of frames in S_i and $H^l_{F_k}$ is the color histogram of frame F_k in S_i. The average color histogram matching degree, $H(S_i, S_j)$, between S_i and S_j is determined by (31)

$$\overline{H}_{S_i} = \frac{\sum_{k=1; F_k \in S_i}^{M} H^l_{F_k}}{M}; \quad l = 0, \ldots, 255 \quad (30)$$

$$H(S_i, S_j) = \sum_{l=0}^{255} Min\left(\overline{H}^l_{S_i}, \overline{H}^l_{S_j}\right). \quad (31)$$

2) Shot-Length Matching: To measure the differences between the longer or shorter edition of similar shots, the length of the shot is considered as one feature. Length matching degree between S_i and S_j is determined by (32)

$$L(S_i, S_j) = 1 - \frac{\left|L_{S_i} - L_{S_j}\right|}{Max\left(L_{S_i}, L_{S_j}\right)} \quad (32)$$

where L_{S_i} and L_{S_j} are the frame numbers of the S_i and S_j, respectively.

3) Camera Motion Matching: Since the camera motion in a shot may imply some semantic information (especially within specific domains, such as sports videos [56]), a video shot similarity evaluation scheme based on camera motion will help construct a motion based video index structure [51] or retrieval system. In [53], various motion matching strategies have been proposed, where the motion activities from the global or small regions of each frame are used to facilitate content-based retrieval. However, these mechanisms only support the retrieval at the frame level, i.e., the query motions are from each single frame. To support motion retrieval at the shot (or even higher) level, we need to explore a new motion matching strategy. In MPEG-7 [59], the amount of camera motion in each frame has

TABLE III
CAMERA MOTION MATCHING DEGREE (R_C AND R_AC DENOTE ROLL_CLOCKWISE AND ROLL_ANTICLOCKWISE, RESPECTIVELY)

Camera Motion	Panning left	Panning right	Tilting up	Tilting down	Zoom in	Zoom out	R_C	R_AC	Still	Irregular
Panning left	1	0	0.2	0.2	0.5	0.5	0	0	0	0.1
Panning right	0	1	0.2	0.2	0.5	0.5	0	0	0	0.1
Tilting up	0.2	0.2	1	0	0.5	0.5	0	0	0	0.1
Tilting down	0.2	0.2	0	1	0.5	0.5	0	0	0	0.1
Zoom in	0.5	0.5	0.5	0.5	1	0	0	0	0	0.1
Zoom out	0.5	0.5	0.5	0.5	0	1	0	0	0	0.1
R_C	0	0	0	0	0	0	1	0	0.8	0.1
R_AC	0	0	0	0	0	0	0	1	0.8	0.1
Still	0	0	0	0	0	0	0.8	0.8	1	0.1
Irregular	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	1

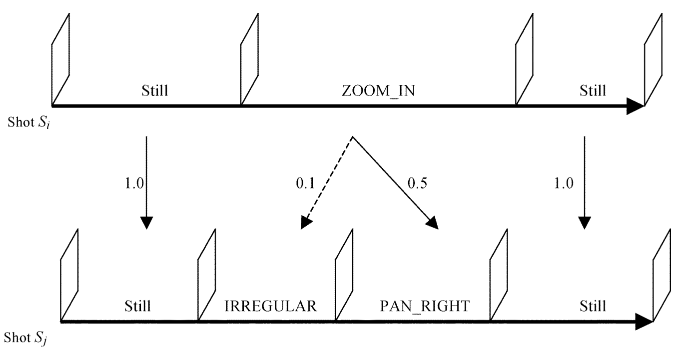


Fig. 10. Camera motion matching.

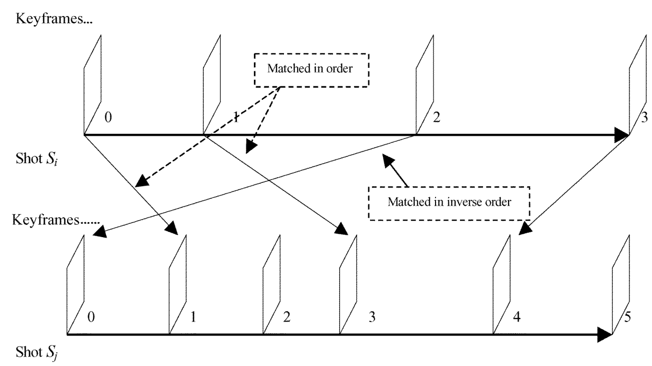


Fig. 11. Key-frame matching.

also been characterized, however, we believe when compared with quantitative motion matching, naïve users may be more concerned about qualitative matching (finding similar types of camera motions); therefore, we adopt a shot-level qualitative motion matching scheme.

Given shot S_i and S_j, if M_{S_i} is the number of camera motion types in S_i from start frame to end frame, we denote the shot with a fewer number of camera motion types by $\hat{S}^M_{i,j}$, and the other shot is denoted by $\widetilde{S}^M_{i,j}$. As Fig. 10 illustrates, $\hat{S}^M_{i,j} = S_i$, $\widetilde{S}^M_{i,j} = S_j$. We then will use $\hat{S}^M_{i,j}$ as the benchmark to find camera motion matching in $\widetilde{S}^M_{i,j}$.

- For each camera motion in $\hat{S}^M_{i,j}$, use Table III to find the closest matching motion in $\widetilde{S}^M_{i,j}$.
- If there is a motion in $\widetilde{S}^M_{i,j}$ that exactly matches the current camera motion in $\hat{S}^M_{i,j}$ (the matching degree is 1), the current matching process will stop.
- If there is no exact match for the current motion in $\hat{S}^M_{i,j}$, the camera motion in $\widetilde{S}^M_{i,j}$ which has the largest matching degree is treated as the match.
- If any motion in $\widetilde{S}^M_{i,j}$ has been exactly matched with the motion in $\hat{S}^M_{i,j}$, any other matching operation will start from the next motion in $\widetilde{S}^M_{i,j}$.
- For any motion in $\hat{S}^M_{i,j}$, start from the last exactly matched camera motion in $\widetilde{S}^M_{i,j}$ to seek the next match. If there is no more camera motion in $\widetilde{S}^M_{i,j}$, the algorithm is terminated.

After the matching process, (33) is used to get the uniform camera motion matching degree. L^+ and L^- are the number of camera motions matched "in order" and "in reverse order", as shown in Fig. 10:

$$M(S_i,S_j)=\frac{1+\alpha_M\frac{\sum_{k=1}^{L^+}D(k)}{Min(M_{S_i},M_{S_j})}-(1-\alpha_M)\cdot\frac{\sum_{k=1}^{L^-}D(k)}{Min(M_{S_i},M_{S_j})}}{2} \quad (33)$$

where $D(k)$ denotes the matching degree between matched camera motions in S_i and S_j (according to Table III), α_M denotes the weight of the *matching order* in similarity evaluation, we set $\alpha_M = 0.7$ in our system.

4) Key-Frame Matching: The key-frame matching degree between shots S_i and S_j. is given by (34), where NK_{S_i} is the key-frame number in shot S_i and $Min(NK_{S_i}, NK_{S_j})$ is the minimal key-frame number in S_i and S_j. α_F denotes the weight of the *matching order* in similarity evaluation, we set $\alpha_F = 0.8$ in our system. We denote the shot with minimal key-frame number as $\hat{S}^F_{i,j}$, and the other shot is denoted as $\widetilde{S}^F_{i,j}$. As shown in Fig. 11, $\hat{S}^F_{i,j} = S_i$, $\widetilde{S}^F_{i,j} = S_j$, F^+ and F^- are the numbers

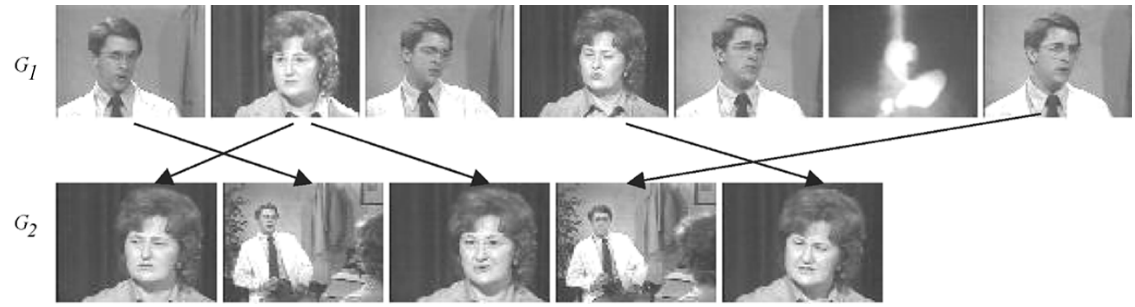


Fig. 12. Group similarity evaluation (the arrows indicate the most similar shots between G_1 and G_2).

of key-frames matched in order and reverse order respectively. $FmSim(k)$ is the similarity between matched key-frames in S_i and S_j which is given in (29). Then, the key-frame matching strategy can be expressed as follows.

- For each key-frame in $\hat{S}^F_{i,j}$, the most similar key-frame in $\widetilde{S}^F_{i,j}$ is selected as the matched frame.
- If any key-frame in $\widetilde{S}^F_{i,j}$ has been matched, it will never be used to match with another key-frame.
- After all key-frames in $\hat{S}^F_{i,j}$ have been matched, (34) is used to get the matching degree

$$F(S_i,S_j) = \frac{1+\alpha_F \frac{\sum_{k=1}^{F^+} FmSim(k)}{Min\left(NK_{S_i},NK_{S_j}\right)} - (1-\alpha_F)\cdot\frac{\sum_{k=1}^{F^-} FmSim(k)}{Min\left(NK_{S_i},NK_{S_j}\right)}}{2}. \tag{34}$$

Based on the four types of shot features and their matching degrees, the similarity between S_i and S_j is computed as the weighted sum of the four matching degrees, as shown in (35), where W_H, W_S, W_F, W_L are the user-specified weights for each of the features

$$StSim(S_i,S_j) = W_H\cdot H(S_i,S_j) + W_M\cdot M(S_i,S_j) + W_F\cdot F(S_i,S_j) + W_L\cdot L(S_i,S_j). \tag{35}$$

C. Group-Level Similarity Evaluation

Based on (35), given a shot S_i and a group G_j, the similarity between them is defined with (36)

$$StGpSim(S_i,G_j) = \underset{S_j\in G_j}{Max}\left\{StSim(S_i,S_j)\right\}. \tag{36}$$

This implies that the similarity between S_i and G_j is the similarity between S_i and the most similar shot in G_j.

In general, when we compare similarity between two groups using the human eye, we usually take the group with fewer shot numbers as the benchmark, and then find whether there are any shots in the other group similar enough to shots in benchmark group. If most shots in the benchmark group were similar enough to the other group, they are treated as similar, as shown in Fig. 12. Therefore, given group G_i and G_j, assume $\hat{G}_{i,j}$ indicates the group with fewer shot numbers, and $\widetilde{G}_{i,j}$ denotes the other group. Suppose $NT(x)$ denotes the number of shot in group x, then, the similarity between G_i and G_j is given by (37)

$$GpSim(G_i,G_j) = \frac{1}{NT(\hat{G}_{i,j})}\sum_{i=1;S_i\in\hat{G}_{i,j}}^{NT(\hat{G}_{i,j})} StGpSim(S_i,\widetilde{G}_{i,j}). \tag{37}$$

Hence, the similarity between group G_i and G_j is the average similarity between shots in the benchmark group and their most similar shots in the other group.

D. Scene-Level Similarity Evaluation

A video scene consists of visually similar groups, so given two scenes SE_i and SE_j, the similarity between them is derived from the similarity among the groups they contain. Assume G_i and G_j are the representative groups in scenes SE_i and SE_j, then the similarity between SE_i and SE_j is given by (38)

$$SeSim(SE_i,SE_j) = \underset{G_i=Select\mathrm{Rep}Group(SE_i);G_j=Select\mathrm{Rep}Group(SE_j)}{GpSim(G_i,G_j)}. \tag{38}$$

That is, the similarity between two scenes is the similarity between their representative groups.

E. Video Level Similarity Evaluation

Assuming $NS(x)$ indicates the number of scenes in video x. Then, based on video similarity evaluation at the scene level, (39), is used to evaluate the distance between two videos V_i and V_j

$$VdSim(V_i,V_j) = \underset{SE_k\in V_i,k=1,\ldots,NS(V_i);SE_l\in V_j,l=1,\ldots,NS(V_j)}{Max}\left\{SeSim(SE_k,SE_l)\right\}. \tag{39}$$

That is, the distance between V_i and V_j is the distance between the most similar scenes among them. Hence, if videos V_i and V_j were very similar to each other, the similarity evaluated from (39) would be large; however, if V_i and V_j are not similar to each other, their similarity may also be relatively large, since they may contain just one similar scene. Hence, (39) is utilized as the first step for video similarity evaluation to find those relatively similar videos, and then the similarity evaluation strategy at the scene, group and shot levels is utilized to refine the retrieval results.

VI. Joint Content Hierarchy for Progressive Video Access

With the constructed video content hierarchy and the video similarity assessment at various levels, our video browsing

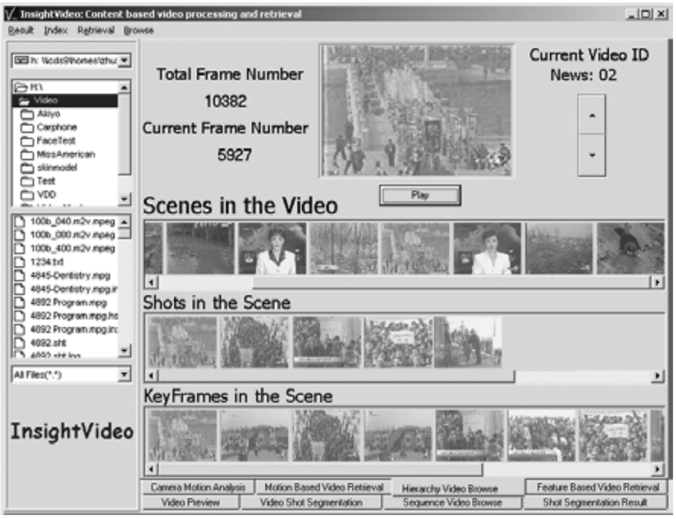


Fig. 13. Hierarchical video content browsing.

Fig. 14. Video retrieval interface.

Fig. 15. Content hierarchy of the retrieved result.

and retrieval can be integrated with great benefit to both. The user can also refine his query by progressively executing the browsing and retrieval process. For example, the user executes the retrieval at the video level, and then adopts (39) to find similar video sequences. Since semantic and low-level features in the query video sequence may vary, it is difficult to tell which part the user is mostly interested in. Hence, by hierarchically browsing the content of the retrieved video sequences, the user can refine his query by selecting a scene (or a group) as the query. Iterative execution operations guide the user in finding the unit he/she is most interested in. In general, the progressive video content access strategy of InsightVideo is executed as follows.

1) A hierarchical video browsing interface is first utilized to help the user gain an overview of the video or video database, as shown in Fig. 13. During video browsing, the user may select any video unit as the query to retrieve from the database. That is, either the key-frame, group, scene, or even the whole video may be selected as the query.
2) The user can also submit an example that is not in the database as the query. In that case, the video analysis strategies are used to construct its content hierarchy. The hierarchical browsing interface is utilized to help the user browse the content table of the query example and refine the query.
3) After the user has selected the query example, the system will utilize the similarity evaluation scheme that corresponds to the same level as the query instance to find similar instances, and present the results to the user, as shown in Fig. 14. Users can click the "Up" or "Down" buttons to view other retrieved units.
4) The user may also browse the content hierarchy of the retrieved video unit by double clicking. Then, Fig. 15 will show the hierarchical content structure of the selected video unit. The first row shows the summary of the current video, and all other rows illustrate the scene information in the video (each row represents one scene). The row with the magnifier icon image on the left indicates that it was ranked as one of the retrieved results. The user can click the magnifier icon image to browse more details in the unit. Then, the user may select any unit in the current interface as the new query. In this way, the retrieval and hierarchical browsing capabilities are integrated to benefit each other in helping the user accessing the video content and refining his/her query efficiently.

TABLE IV
SYSTEM RETRIEVAL PERFORMANCE AT SHOT LEVEL (FROM TOP 20 RESULTS)

Content of the video database	Shots in the video	Recall	Precision
Film Abstracts	526	0.62	0.65
News Programs	771	0.85	0.78
Medical Videos	1286	0.81	0.74
Average Performance	*2583*	*0.768*	*0.726*

TABLE V
SYSTEM RETRIEVAL PERFORMANCE AT VIDEO LEVEL

Videos	Query 1		Query 2		Query 3		Query 4		Query 5		Query 6		Query7		Query8	
Medical videos	2	1	3	2	3	1	4	2	1	3	1	4	6	8	1	3
News program	6	7	5	3	7	11	8	3	7	2	4	7	8	13	4	12
Film Abstract	4	3	2	6	1	5	4	1	6	2	3	1	6	4	3	2
Average	4.27															

5) Iteratively execute steps 3 and 4 until the user finds the satisfactory results or halts the current retrieval operation at any time.

A. System Performance

Two types of video retrieval results, retrieval at the shot level and video level, are executed in our system. The experimental results are shown in Table IV and Table V, respectively. Our video database consists of 16 videos (about 6 h) from various sources (five film abstracts, four news, and seven medical videos). All videos were processed with the techniques described above to extract their feature and content table. Then, InsightVideo was used to hierarchical browse and progressively retrieve videos from the database.

While executing video retrieval at the shot level, two factors, *precision* and *recall* are defined to evaluate the efficiency of the system. *Precision* specifies the ratio of the number of relevant shots to the total number of returned shots. *Recall* specifies the ratio of the number of relevant video sequences found to the total number of relevant video sequences in the database. While evaluating the similarity between shots, we set W_F, W_M, W_H, and W_L equal to 0.5, 0.1, 0.3, and 0.1 respectively. That is, we put heavy emphasis on the matching of visual features. During the retrieval process, the relevant shots retrieved in the top 20 are returned as the results. Performance is measured by comparing results produced by the assessment strategy on the five queries for each type of video against human relevance judgment. From Table IV, we can see that our system achieved rather good performance (76.8% in recall and 72.6% in precision) on different kinds of video. However, as content and background of film abstracts are more complex than other videos in terms of camera motion and content changes, the performance results of the film abstracts are somewhat worse. A more reliable and efficient method may be needed for film evaluation.

Another experiment is executed to evaluate the efficiency of the video similarity evaluation model at the video level. In this experiment, each of 16 videos in the database is first manually separated into three nearly equal clips (the entire video database contains $16 \times 3 = 48$ clips, no scene overlaps with the manually segmented boundaries). Then, one clip is randomly selected from the database as the query, and results are retrieved from the database. The ranks of the other two clips that are in the same video as the query are used to evaluate the system performance.

We randomly select 24 retrieval results (with eight for each), and show them in Table IV. To evaluate the similarity between the videos, we set W_F, W_M, W_H, and W_L equal to 0.4, 0.0, 0.5, and 0.1, respectively. From Table V, we see that our retrieval strategy achieved relatively good results, the average location of the retrieved clips that are in the same video as the query is 4.27 (out of 47 clips, since the query clip is excluded from the database). However, we find the retrieval results for News are worse than for the other two types of video. This is because News programs are usually different from general video data: in common videos, a similar scene may be shown repetitively in the video, however, in a news program, most story units are reported only once, hence, the three clips of the same News video may have large variety in content and visual features. This can cause our system to falsely locate the related clips with the query example.

VII. CONCLUSION

In this paper, we have presented the InsightVideo system, which constructs the video content hierarchy for efficient video content access. A progressive video retrieval scheme was proposed which seamlessly integrates hierarchical video browsing and retrieval. To create the video content table, several video analysis techniques were introduced: 1) a statistical information based camera motion classification method; 2) a compressed domain key-frame extraction method which is based on detected camera motions in each shot; and 3) video group and scene detection and scene clustering strategies which organize the video content hierarchy. Video similarity assessment at different levels (frame, shot, group, scene, video) was addressed. Based on constructed video content hierarchy and the video similarity evaluation strategy, the hierarchical video browsing and retrieval are seamlessly integrated together.

Unlike most other video retrieval systems that execute video retrieval at the shot level, the retrieval results of the InsightVideo are the units likely related to the query example in both low-level features and semantics. In contrast to other video browsing systems, we joined the hierarchical video browsing with video retrieval. This benefits both processes, and produces progressive video retrieval system. The features that distinguish our system from others are the following: 1) the integration of several novel video processing techniques which improve existing algorithms in important ways; 2) the construction of a video content hierarchy which allows the hierarchical video content browsing and summarization to be executed directly; 3) address of the video similarity at different granularities to support the retrieval at various content levels; and 4) a progressive retrieval which integrates the video browsing and retrieval processes, and allowsusers to shrink and refine queries efficiently.

ACKNOWLEDGMENT

The authors would like to thank the anonymous reviewers for their constructive comments on an earlier version of this paper. They also thank Dr. X. Wu at the University of Vermont for his valuable comments in revising the paper.

REFERENCES

[1] H. J. Zhang, A. Kantankanhalli, and S. W. Smoliar, "Automatic partitioning of full-motion video," *ACM Multimedia Syst.*, vol. 1, no. 1, 1993.

[2] H. J. Zhang, C. Y. Low, S. W. Smoliar, and D. Zhong, "Video parsing, retrieval and browsing: an integrated and content-based solution," in *Proc. ACM Multimedia Conf.*, San Francisco, CA, 1995.

[3] M. M. Yeung and B.-L. Yeo, "Video visualization for compact presentation and fast browsing of pictorial content," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 7, pp. 771–785, 1997.

[4] J. R. Smith, "VideoZoom spatio-temporal video browser," *IEEE Trans. Multimedia*, vol. 1, no. Jun., pp. 157–171, 1999.

[5] Y. A. Aslandogan and C. T. Yu, "Techniques and systems for image and video retrieval," *IEEE Trans. Knowl. Data Eng.*, vol. 11, no. 1, pp. 56–63, Jan. 1999.

[6] H. Wactlar, "Informedia—Search and summarization in the video medium," in *Proc. Imagina Conf.*, Monaco, 2000.

[7] S. F. Chang, W. Chen, H. J. Meng, H. Sundaram, and D. Zhong, "VideoQ: an automated content based video search system using visual cues," in *Proc. ACM Multimedia Conf.*, Seattle, WA, 1997, pp. 313–324.

[8] A. Hampapur, A. Gupta, B. Horowitz, C.-F. Shu, C. Fuller, J. Bach, M. Gorkani, and R. Jain, "Virage video engine," in *Proc. SPIE*, vol. 3022, San Jose, CA, 1997.

[9] A. Pentland, R. Picard, and S. Sclaroff, "Photobook: content-based manipulation of image database," *Int. J. Comput. Vis.*, vol. 18, pp. 233–254, 1996.

[10] J. R. Smith and S. F. Chang, "VisualSEEk: a fully automated content-based image query system," in *Proc. ACM Multimedia Conf.*, Boston, MA, 1996, pp. 87–98.

[11] V. E. Ogle and M. Stonebraker, "Chabot: Retrieval from a relational database of image," *Computer*, vol. 28, no. 9, pp. 40–48, Sep. 1995.

[12] C. Faloutsos, W. Equitz, M. Flickner, W. Niblack, S. Petkovic, and R. Barber, "Efficient and effective querying by image content," *J. Intell. Inf. Syst.*, vol. 3, pp. 231–262, 1994.

[13] H. Jiang and A. Elmagarmid, "WVTDB-A semantic content-based video database system on the world wide web," *IEEE Trans. Knowl. Data Eng.*, vol. 10, no. 6, pp. 947–966, Nov. 1998.

[14] A. Elmagarmid *et al.*, *Video Data Bases: Issues, Products and Applications*. Norwell, MA: Kluwer, Mar. 1997.

[15] S. F. Chang, J. Smith, and J. Meng, "Efficient techniques for feature based image/video access and manipulation," in *Proc. 33rd Ann. Clinic on Library Application of Data Processing Image Access and Retrieval*, Urbana-Champaign, IL, Mar. 1996.

[16] M. La Cascia and E. Ardizzone, "JACOB: Just a content-based query system for video database," in *Proc. ICASSP*, Atlanta, GA, 1996.

[17] J. R. Smith and S. F. Chang, "Searching for Images and Video on the World-Wide Web," Columbia Univ., New York, CTR Tech. Rep. 459-96-25, Aut. 1996.

[18] D. W. Howard, K. Takeo, A. S. Michael, and M. S. Scott, "Intelligent access to digital video: informedia project," *IEEE Computer*, vol. 29, no. 5, pp. 46–52, May 1996.

[19] B. L. Yeo and B. D. Liu, "Rapid scene analysis on compressed video," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 5, no. 6, pp. 533–544, Dec. 1995.

[20] J. Fan, W. Aref, A. Elmagarmid, M. Hacid, M. Marzouk, and X. Zhu, "MultiView: multilevel video content representation and retrieval," *J. Elect. Imaging*, vol. 10, no. 4, pp. 895–908, Oct. 2001.

[21] R. Szeliski, "Video mosaics for virtual environments," *IEEE Comput. Graph. Appl.*, vol. 16, no. 2, pp. 22–30, Mar. 1996.

[22] C. W. Ngo, T. C. Pong, H. J. Zhang, and R. T. Chin, "Motion characterization by temporal slice analysis," in *Proc. Computer Vision and Pattern Recognition*, vol. 2, Hilton Head, SC, 2000, pp. 768–773.

[23] W. Xiong and J. C.-M. Lee, "Efficient scene change detection and camera motion annotation for video classification," *Comput. Vis. Image Understanding*, vol. 71, pp. 166–181, 1998.

[24] M. V. Srinivasan, S. Venkatesh, and R. Hosie, "Qualitative estimation of camera motion parameters from video sequences," *Pattern Recognit.*, vol. 30, no. 4, pp. 593–606, Apr. 1997.

[25] R. C. Jain, "Direct computation of the focus of expansion," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. PAMI-5, no. 1, pp. 58–64, Jan. 1983.

[26] Y. P. Tan, D. D. Saur, S. R. Kulkami, and P. J. Ramadge, "Rapid estimation of camera motion from compressed video with application to video annotation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 10, no. 1, pp. 133–146, Feb. 2000.

[27] V. Kobla, D. S. Doermann, and A. Rosenfeld, "Compressed Domain Video Segmentation,", CfAR Tech. Rep. CAR-TR-839 (CS-TR-3688), 1996.

[28] C. Dorai and V. Kobla, "Perceived visual motion descriptors from MPEG-2 for content-based HDTV annotation and retrieval," in *Proc. IEEE 3rd Workshop on Multimedia Signal Processing*, Copenhagen, Denmark, 1999, pp. 147–152.

[29] A. Hanjalic and H. J. Zhang, "An integrated scheme for automated video abstraction based on unsupervised cluster-validity analysis," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 9, no. 8, pp. 1280–1289, Dec. 1999.

[30] H. J. Zhang, J. Wu, D. Zhong, and S. W. Smoliar, "An integrated system for content-based video retrieval and browsing," *Pattern Recognit.*, vol. 30, no. 4, pp. 643–658, Apr. 1997.

[31] Y. Zhang, Y. Rui, T. S. Huang, and S. Mehrotra, "Adaptive key frame extraction using unsupervised clustering," in *Proc. ICIP*, vol. 1, Chicago, IL, 1998, pp. 866–870.

[32] W. Wolf, "Key frame selection by motion analysis," in *Proc. IEEE ICASSP*, Atlanta, GA, 1996, pp. 1228–1231.

[33] A. Pentland, R. Picard, G. Davenport, and K. Haase, "Video and image semantics: advanced tools for telecommunications," *IEEE Multimedia Mag.*, vol. 1, no. 2, pp. 73–75, Jun. 1994.

[34] Y. Avrithis, N. Doulamis, A. Doulamis, and S. Kollias, "Efficient content representation in MPEG video database," in *Proc. IEEE Workshop on Content Based Access of Image and Video Database*, Santa Barbara, CA, 1998, pp. 91–95.

[35] B. Shahraray and D. C. Gibbon, "Automatic generation of pictoral transcripts of video programs," in *Proc. IS&T/SPIE Digital Video Compression: Algorithms and Technologies*, San Jose, CA, 1995, pp. 512–519.

[36] T. Lin and H. J. Zhang, "Automatic video scene extraction by shot grouping," in *Proc. ICPR*, Barcelona, Spain, 2000.

[37] Y. Rui, T. Huang, and S. Mehrotra, "Constructing table-of-content for video," *ACM Multimedia Syst. J., Special Issue Multimedia Systems on Video Libraries*, vol. 7, no. 5, pp. 359–368, Sep. 1999.

[38] M. M. Yeung and B. L. Yeo, "Time-constrained clustering for segmentation of video into story units," in *Proc. ICPR'96*, Vienna, Austria, 1996.

[39] W. Zhou, A. Vellaikal, and C. J. Kuo, "Rule-based video classification system for basketball video indexing," in *Proc. ACM Multimedia Conference Workshop*, Los Angeles, CA, 2001.

[40] D. Swanberg, C. F. Shu, and R. Jain, "Knowledge-guided parsing in video database," in *Proc. IS&T/SPIE Storage and Retrieval for Image and Video Databases II*, San Jose, CA, 1993, pp. 13–24.

[41] X. Zhu, L. Wu, X. Xue, X. Lu, and J. Fan, "Automatic scene detection in news program by integrating visual feature and rules," in *Proc. 2nd IEEE PCM Conf., LNCS2195*, Beijing, China, 2001, pp. 837–842.

[42] D. Zhong, H. J. Zhang, and S. F. Chang, "Clustering Methods for Video Browsing and Annotation," Columbia Univ., New York, Tech. Rep., 1997.

[43] D. A. Adjeroh, M. C. Lee, and I. King, "A distance measure for video sequence similarity matching," in *Proc. Int. Workshop on Multimedia Database Management Systems*, Dayton, OH, 1998.

[44] N. Dimitrova and M. Abdel-Mottaled, "Content based video retrieval by example video clip," in *Proc. SPIE*, vol. 3022, San Jose, CA, 1998.

[45] R. Lienhart, W. Effelsberg, and R. Jain, "VisualGREP: a systematic method to compare and retrieval video sequences," in *Proc. SPIE*, vol. 3312, San Jose, CA, 1997, pp. 271–282.

[46] L. Zhao, W. Qi, S. Z. Li, S. Yang, and H. Zhang, "Key-frame extraction and shot retrieval using Nearest Feature Line (NFL)," in *Proc. ACM Multimedia Conference Workshops*, Los Angeles, CA, 2000, pp. 217–220.

[47] S. Li and J. Liu, "Face recognition based on nearest linear combinations," *IEEE Trans. Neural Netw.*, vol. 10, no. 2, pp. 439–443, Mar. 1999.

[48] X. Liu, Y. Zhuang, and Y. Pan, "A new approach to retrieve video by example video clip," in *Proc. ACM Multimedia Conf.*, Orlando, FL, 1999, pp. 41–44.

[49] A. K. Jain, *Algorithms for Clustering Data*. Englewood Cliffs, NJ: Prentice-Hall, 1998.

[50] A. Hauptmann and M. Witbrock, "Story segmentation and detection of commercials in broadcast news video," in *ADL-98 Advances in Digital Libraries Conf.*, Santa Barbara, CA, Apr. 22–24, 1998.

[51] C. Ngo, T. Pong, and H. Zhang, "On clustering and retrieval of video shots," in *Proc. 9th ACM Multimedia Conf.*, Ottawa, ON, Canada, 2001, pp. 51–60.

[52] J. Fan, J. Yu, G. Fujita, T. Onoye, L. Wu, and I. Shirakawa, "Spatiotemporal segmentation for compact video representation," *Signal Processing: Image Commun.*, vol. 16, pp. 553–566, 2001.

[53] E. Ardizzone, M. Cascia, A. Avanzato, and A. Bruna, "Video indexing using MPEG motion compensation vectors," in *Proc. IEEE Conf. on Multimedia Computing and Systems*, Florence, Italy, 1999.

[54] A. Divakaran, R. Radhakrishnan, and K. Peker, "Motion activity-based extraction of key-frame from video shots," in *Proc. IEEE ICIP Conf.*, Rochester, NY, 2002.

[55] ——, "Video summarization using descriptors of motion activity: a motion activity based approach to key-frame extraction from video shots," *J. Electron. Imag.*, vol. 10, no. 4, pp. 909–916, 2001.

[56] M. Lazarescu, S. Venkatesh, and G. West, "On the automatic indexing for cricket using camera motion parameters," in *Proc. IEEE ICME Conf.*, Laussane, Switzerland, 2002.

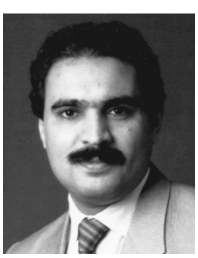
[57] S. Jeannin and A. Divakaran, "MPEG-7 visual motion descriptor," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 11, no. 6, pp. 720–724, Jun. 2001.

[58] B. S. Manjunath, P. Salembier, and T. Sikora, *Introduction to MPEG-7: Multimedia Content Description Interface*. New York: Wiley, Jun. 2002.

[59] "The MPEG-7 visual part of the XM 4.0, ISO/IEC MPEG99/W3068,", HI, Dec. 1999.

Xingquan Zhu received the M.S. degree in communication and electronic systems from Xidian University, Xian, China, in 1998, and the Ph.D degree in computer science from Fudan University, Shanghai, China, in 2001.

He previously spent about six months with Microsoft Research Asia, Beijing, China, where he worked on relevance feedback for image indexing and retrieval. During 2001 to 2002, he was a Postdoctoral Associate at the Department of Computer Science, Purdue University, West Lafayette, IN. He is currently a Research Assistant Professor at the Department of Computer Science, University of Vermont, Burlington. His research interests include data mining, and machine learning.



Ahmed K. Elmagarmid (SM'93) received the B.S. degree in computer science from the University of Dayton, Dayton, OH, and the M.S. and Ph.D. degrees from The Ohio State University in 1977, 1981, and 1985, respectively.

He is the Director of the Indiana Center for Database Systems (ICDS), Purdue University, West Lafayette, IN, and the newly formed Indiana Telemedicine Incubator. He was Chief Scientist for Hewlett-Packard from 2001 to 2003 while on leave from Purdue. He has served widely as an industry consultant and/or adviser to Telcordia, Harris, IBM, MCC, UniSql, MDL, BNR, etc. He served as a faculty member at the Pennsylvania State University from 1985–1988 and has been with the Department of Computer Science at Purdue University since 1988. His research interests are in the areas of video databases, multidatabases, data quality, and their applications in Telemedicine and digital government. He is the author of several books in databases and multimedia.

Dr. Elmagarmid received a Presidential Young Investigator Award from the National Science Foundation, and Distinguished Alumni awards from The Ohio State University and the University of Dayton in 1988, 1993, and 1995, respectively. He is the Editor-in-Chief of *Distributed and Parallel Databases: An International Journal* and of the book series on Advances in Database Systems, and serves on the Editorial Boards of the IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, *Information Sciences*, and the *Journal of Communications Systems*. He has previously served on the Editorial Boards of IEEE TRANSACTIONS ON COMPUTERS and the *IEEE Data Engineering Bulletin*. He is on the steering committees for the IEEE International Conference on Data Engineering, and the IEEE Symposium on Research Issues in Data Engineering and served on the organization committees of several international conferences.

Xiangyang Xue received the B.S., M.S., and Ph.D. degrees in communication engineering from Xidian University, Xian, China in 1989, 1992, and 1995, respectively.

Since 1995, he has been with the Department of Computer Science and Engineering, Fudan University, Shanghai, China, where he is currently a Professor. His research interests include multimedia information process and retrieval, image and video coding, and video streaming.

Lide Wu (SM'88) graduated from the Department of Mathematics, Fudan University, Shanghai, China.

He was with the Department of Mathematics, Fudan University, until 1975. Since then, he has been with the Department of Computer Science, Fudan University, where he is currently a Chair Professor. He was the Chairman of the department during 1982–1985. His research interests include image processing, video processing, computer vision, pattern recognition, and Chinese text processing. He is the author or co-author of more than ten books and more than 200 papers.

Mr. Wu is Vice Chairman of the Technical Committee of Artificial Intelligence and Pattern Recognition of the Chinese Computer Society and a member of the ACM and the New York Academy of Sciences.

Ann Christine Catlin received the B.S. degree in mathematics from Seton Hall University, South Orange, NJ, in 1977 and the M.S. degree in mathematics from Notre Dame University, Notre Dame, IN, in 1980.

Between 1980 and 1986, she managed engineering groups responsible for the design, simulation, and analysis of networks at AT&T Information Systems and Applied Data Research. She was the Director of Research and Development at Network Synergies Inc. from 1986 to 1989. Since 1991, she has worked in the Computer Science Department, Purdue University, West Lafayette, IN, leading research, design, and development efforts on projects in problem-solving environments, performance evaluation of parallel architectures, web-based systems, knowledge discovery in databases, and knowledge bases. She has more than 30 publications, including journal papers, conference proceedings, and book chapters.