A Novel Video Key-Frame-Extraction Algorithm Based on Perceived Motion Energy Model

Tianming Liu, Hong-Jiang Zhang, and Feihu Qi

***Abstract*—The key frame is a simple yet effective form of summarizing a long video sequence. The number of key frames used to abstract a shot should be compliant to visual content complexity within the shot and the placement of key frames should represent most salient visual content. Motion is the more salient feature in presenting actions or events in video and, thus, should be the feature to determine key frames. In this paper, we propose a triangle model of perceived motion energy (PME) to model motion patterns in video and a scheme to extract key frames based on this model. The frames at the turning point of the motion acceleration and motion deceleration are selected as key frames. The key-frame selection process is threshold free and fast and the extracted key frames are representative.**

***Index Terms*—Key-frame extraction, motion pattern, video summarization.**

I. Introduction

A VIDEO KEY frame is the frame that can represent the salient content of a video shot. Key frames provide a suitable abstraction for video indexing, browsing, and retrieval [1]. The use of key frames greatly reduces the amount of data required in video indexing and browsing and provides an organizational framework for dealing with video content [5]. Users can quickly browse over the video by viewing only a few highlighted key frames. In this paper, we focus on key-frame extraction within a given shot, instead of key frames at story or scene level. One or more key frames can be extracted from a single shot depending on the content complexity of the shot.

Current key-frame-extraction techniques can be classified according to their various measurements of visual content complexity of a video shot or sequence. The first way of measuring visual content complexity of a video shot is by sequential comparison of color. Zhang *et al.* [7] first selected key frames in a sequential fashion for each shot by computing color histogram difference between current frame and the last extracted key frame. This idea of sequential comparison is also extended by using the information of dominant or global motion in [7]. Gunsel and Tekalp [4] computed the discontinuity value between the current frame and the N previous frames by comparing the color histogram of current frame and the average color histogram of the previous N frames. The second way of measuring visual content complexity of a video shot or sequence is by color clustering. Hanjalic and Zhang [11] automatically find the optimal number of clusters by applying cluster-validity analysis. After the optimal number of clusters is found, each of the clusters is represented by one characteristic frame which minimizes the Euclidean distance between feature vectors of all cluster elements and the cluster centroid. Zhuang *et al.* [2] assigned a new frame to an existing cluster, if it is similar enough to the centroid of that cluster. If the computed similarity is lower than the pre-specified threshold, a new cluster is formed around the current frame. As a result, the number of key frames is determined by the number of clusters and the frame closest to the centroid of a key cluster is extracted as a key frame. The third way of measuring visual content complexity of a video shot or sequence is by the sum of frame-to-frame differences along a shot or a sequence. Hanjalic *et al.* [6] accumulated histogram difference of consecutive frames over the shot and over the entire sequence. Each shot of the sequence gets assigned one part of given N key frames according to the percentage of its difference accumulation of the total difference accumulation of the sequence. A numerical algorithm is used to distribute key frames in each shot. The fourth way of measuring visual content complexity of a video sequence is by occurring frequency of local motion or activity minima. Wolf [8] computed the optical flow for each frame and then used a simple motion metric to evaluate the changes in the optical flow along the sequence. Key frames are then found at places where the metric as a function of time has its local minima. Gresle and Huang [9] computed the intra and reference histograms and then compute an activity indicator. Based on the activity curve, the local minima are selected as the key frames.

In this paper, we measure visual content complexity of a shot by motion patterns. A motion pattern of a shot is usually composed of a motion acceleration process, followed by deceleration process. Such a motion pattern usually reflects an action in events. For example, there are usually the break-play-break sequences in sports video and static-camera pan-static sequences in news video and movie. Hence, the occurring frequency of motion patterns, namely, action events is a good indicator of visual content complexity of a shot.

To extract key frames based on motion patterns, we need first to build a motion model that reflects the motion activities in video shots, thus guiding the selection of key frames. Toward this objective, a triangle model of perceived motion energy (PME) has been developed. With this model, a video shot is segmented into subsegments of consecutive motion patterns in term of acceleration and decelerations. Key frames are extracted from these subsegments based on the proposed triangle model. The left-bottom vertex of the triangle represents the start point of the motion acceleration process, the right-bottom vertex represents the end point of the motion deceleration process, and the top vertex of the triangle represents the point of the maximum speed. In addition, the area of the triangle represents the accumulated PME within the pattern. Based on this model, we have

Manuscript received November 15, 2001; revised July 22, 2002 and March 27, 2003. This paper was recommended by Associate Editor S. Panchanathan.

T. Liu was with Microsoft Research Asia, Beijing, China. He is now with the University of Pennsylvania, Philadelphia, PA 19104 USA.

H.-J. Zhang is with Microsoft Research Asia, Beijing 100080, China.

F. Qi is with Shanghai Jiaotong University, Shanghai 200030, China.

Digital Object Identifier 10.1109/TCSVT.2003.816521

proposed to select the frames at top vertexes of the triangles, which is the turning point from the accelerating to decelerating motion, as key frames. Our reasoning is that the turning point of motion acceleration to deceleration usually represents the most salient point of an action, and one can infer the movement within the acceleration to deceleration process from the turning point. In this way, the whole video sequence is segmented into meaningful actions represented by the triangle model of PME, and each of the actions will be captured by its corresponding key frame. If there are no motion patterns in a shot, the first frames of the shot detected by the color-histogram-based algorithm are selected as the only key frames because one key frame is sufficient to represent a static shot. If the maximum allowed number of key frames for a sequence is regulated, each shot gets assigned a number of given key frames according to the percentage share of its motion pattern series in the total pattern series of the entire sequence. That is, the motion pattern series of each shot are sorted by their accumulated PME, and patterns with larger accumulated energy are selected. Hence, an advantageous feature of the proposed algorithm is that it is threshold free in the key-frame selection process.

The remainder of the paper is organized as follows. Section II details the PME feature and the triangle model. Section III describes the key-frame selection algorithm. Experiment results are shown in Section IV, and Section V concludes the paper.

II. PME Feature and Triangle Model

To extract key frames based on motion patterns, a triangle model of PME has been developed to represent the motion activities in video shots. Compared to the optical flow in [8] and the frame difference in [6], PME is a combined metric of motion intensity and motion characteristics with more emphasis on dominant motion. With this triangle model, a video shot is segmented into subsegments of different motion patterns in terms of acceleration and decelerations. Consequently, the accumulated PME along a subsegment reflects its relative salience of visual action and can be used as the criterion for sorting importance of motion patterns.

A. PME

To simplify the motion analysis process, we choose to extract motion data to build the PME model directly from MPEG video streams. In the MPEG stream, there are two motion vectors in each macro block of the B frame for motion compensation, often referred as motion vector field (MVF). Since the magnitude of a motion vector reflects motion velocity of a macro block, it can be used to compute the energy of motion of frame. Although the angle of a motion vector is not reliable to represent motion direction of a macro block, the spatial consistency of angles of motion vectors does reflect the intensity of global motion. The spatial motion consistency can be obtained by calculating the percentage of dominant motion direction in an entire frame. The more consistent the angles are, the higher the intensity of global motion is.

Then, average magnitude $\mathrm{Mag}(t)$ of motion vectors in the entire frame is calculated as

$$\mathrm{Mag}(t) = \frac{\left(\frac{\sum \mathrm{Mix}FEn_{i,j}(t)}{N} + \frac{\sum \mathrm{Mix}BEn_{i,j}(t)}{N}\right)}{2} \quad (1)$$

where $\mathrm{Mix}FEn_{i,j}(t)$ represents forward motion vectors and $\mathrm{Mix}BEn_{i,j}(t)$ represents backward motion vectors. N is the number of macro blocks in the frame. Computing $\mathrm{Mix}FEn_{i,j}(t)$ and $\mathrm{Mix}BEn_{i,j}(t)$ is similar to computing $\mathrm{Mix}En_{i,j}(t)$ in [12].

The percentage of dominant motion direction $\alpha(t)$ is defined as

$$\alpha(t) = \frac{\max\left(AH(t,k), k \in [1,n]\right)}{\sum_{k=1}^{n} AH(t,k)}. \quad (2)$$

The angle in 2π is quantized into n angle ranges. Then, the number of angles in each range is accumulated over the whole forward motion vectors to form an angle histogram with n bins, denoted by $AH(t,k)$ and $k \in [1,n]$. So, $\max(AH(t,k))$ is the dominant direction bin among all motion directions. n is set at 16 throughout this work.

The PME of a B frame is computed as

$$\mathrm{PME}(t) = \mathrm{Mag}(t) \times \alpha(t). \quad (3)$$

The right-hand side item $\alpha(t)$ represents the percentage of dominant motion direction. We can see that PME is a combined metric of motion intensity and the kind of motion with more emphasis on dominant video motion. So, the accumulated PME within a subsegment reflects its relative salience of visual action content.

To calculate $\mathrm{Mag}(t)$, or more specifically, the average forward and backward motion vectors of each B frame in (1) $\mathrm{Mix}FEn_{i,j}(t)$ and $\mathrm{Mix}BEn_{i,j}(t)$, a set of filtering steps are applied to remove noises and atypical vectors in a MVF because such noises usually result in inaccurate energy accumulation. First, a spatial filtering process is applied to the MVF to remove atypical vectors resulting from the inaccurate blocking matching process. The spatial filter we used is a modified median filter. The elements in the filter window at a macro block $MB_{i,j}$ (either forward or backward) are denoted by $\Omega_{i,j}$ in MVF, where W_s is the width of window. The filtered magnitude of motion vector is computed by

$$\mathrm{Mag}_{(i,j)} = \begin{cases} \mathrm{Mag}_{i,j}, & \text{if } \mathrm{Mag}_{i,j} \le \text{Max 4th}\left(\mathrm{Mag}_k\right) \\ \text{Max 4th}\left(\mathrm{Mag}_{i,j}\right), & \text{if } \mathrm{Mag}_{i,j} > \text{Max 4th}\left(\mathrm{Mag}_k\right) \end{cases} \quad (4)$$

where $k \in \Omega_{i,j}$ and the function Max 4th(Mag_k) returns the fourth value in the descending sorted list of magnitude elements $\Omega_{i,j}$ in the filter window. The choice of returning the fourth value is determined from experiments.

After the spatial filtering, a temporal filter is applied to further filter out noise in obtaining the final motion magnitude at each macro block position (i,j). This temporal filtering process adopts an alpha-trimmed filter within a temporal window of size W_t. In this filtering process, all of the magnitudes in the temporal window are sorted first. After the values at two ends of sorted list are trimmed, the rest of the magnitudes are averaged to form mixture energy $\mathrm{Min}En_{i,j}$, which includes the energy of both object and camera motion, as

$$\mathrm{Mix}En_{i,j} = \frac{1}{\left(M - 2 \times \lfloor \alpha M \rfloor \times W_t^2\right)} \sum_{m=\lfloor \alpha M \rfloor + 1}^{M - \lfloor \alpha M \rfloor} \mathrm{Mag}_{i,j}(m) \quad (5)$$

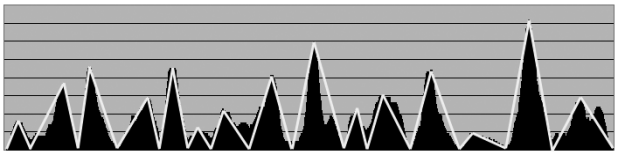

Fig. 1. Example of a sequence of motion triangles.

where M is the total number of magnitudes in the window, $\lfloor \alpha M \rfloor$ equals the largest integer not greater than αM, and $\text{Mag}_{i,j}(m)$ is the magnitudes value in the sorted list. The trimming parameter α ($0 \leq \alpha \leq 0.5$) controls the number of data samples excluded from the accumulating computation. More detail on the motivation and design of the filtering processing can be found in [12].

B. Triangle Model

The PME value is calculated for each B frame because this is enough to capture the motion of video sequence and accurate enough for the application of selecting key frames. Now, the original video sequence is represented by the PME value sequence PME(t). Before temporally segmenting a sequence into successive subsegments, PME(t) is filtered by averaging PME values within a window of T_0 to smooth out potential noises in the PME value sequence. Then, the pattern of an acceleration process and deceleration process is modeled by a triangle.

The triangle model is used to segment video sequence into successive segments and represent each of the segments. Fig. 1 shows an example. The left-bottom vertex of a triangle represents the start point of the segment, and its PME value is zero. The right-bottom vertex of the triangle represents the end point of the segment, and its PME value is also zero. The top vertex of the triangle represents the maximum PME value of the segment. That is, for segment i, the triangle model is represented by $(ts_i, te_i, tp_i, \text{PME}_i, \text{AP}_i)$, where ts_i is the start point, te_i is the end point, tp_i is the point of peak motion, PME_i is the peak PME value of the segment, and AP_i is the accumulated PME obtained simply by summing up all PME values within the subsegment. Obviously, we have $\text{PME}(ts_i) = \text{PME}(te_i) = 0$. A special triangle model $(ts_i, te_i, tp_i, 0, 0)$ is used for successive zeros.

To segment a shot sequence is to detect triangle patterns in the PME sequence. The PME value of start point and that of end point of a segment are both zero. Thus, a simple search process is used to find the triangle patterns. However, when motion continues for a long time, the triangle will become less accurate. Fig. 2 shows an example. Therefore, a splitting process is performed before the triangle pattern search process. To split long continuous motion, splitting boundaries are to be found first. For a particular point $(t, \text{PME}(t))$, if

$$\text{PME}(t) = \min(\text{PME}(t-T), \ldots, \text{PME}(t-i), \ldots \text{PME}(t+i), \ldots, \text{PME}(t+T)))$$

and

$$\text{PME}(t+j) > 0 \quad j \in [-T, T]$$

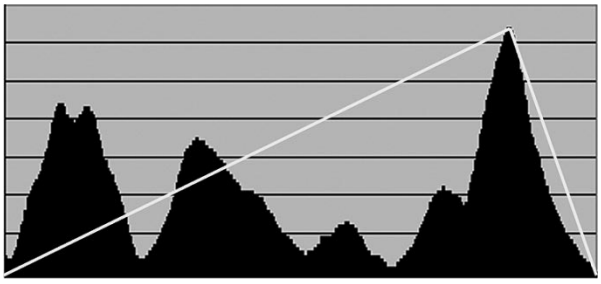

Fig. 2. Original triangle.

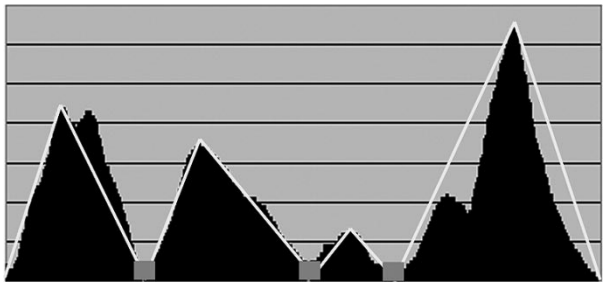

Fig. 3. Split triangles.

then, PME(t) is set to 0. Thus, $(t, \text{PME}(t))$ now becomes a spitting boundary. That is, some local minimums of the PME sequence are set as splitting boundaries. Fig. 3 shows the splitting results of Fig. 2. The three grey blocks show three splitting boundaries, which are local minimums of the original PME sequence. As a result, the large triangle in Fig. 3 is split into four small triangles. T_0 and T are set at 20 and 100, as these values give good results in our experiments.

Figs. 4 and 5 show the PME sequences of two example segments, respectively. We can see that the motion pattern typically composed of an acceleration process and a deceleration process is repeated during the segment. Using the triangle model is simple, yet effective to represent such motion patterns. A motion pattern usually corresponds to an action event. Hence, the occurring frequency of action event is a good metric for measuring visual content complexity of a shot. The more action events exist in a shot means the shot is more complex and more key frames are needed to abstract the shot.

We have evaluated how well the triangle model describes motion acceleration and deceleration processes in typical video sequences by analyzing the difference ratio between the area of the triangle and the accumulated PME in the motion acceleration and deceleration process. That is, for a triangle model $(ts_i, te_i, tp_i, PME_i, \text{AP}_i)$, the area AR_i of the triangle and accumulated PME in a motion acceleration and deceleration process are

$$\text{AR}_i = \frac{1}{2} \times (te_i - ts_i) \times \text{PME}_i \text{ and } AC_i = \sum_{j=ts_i}^{j=te_i} \text{PME}(j)$$

respectively. The difference ratio between the area of the triangle and the accumulated PME is

$$p = \frac{||\text{AR}_i - AC_i||}{AC_i.}$$

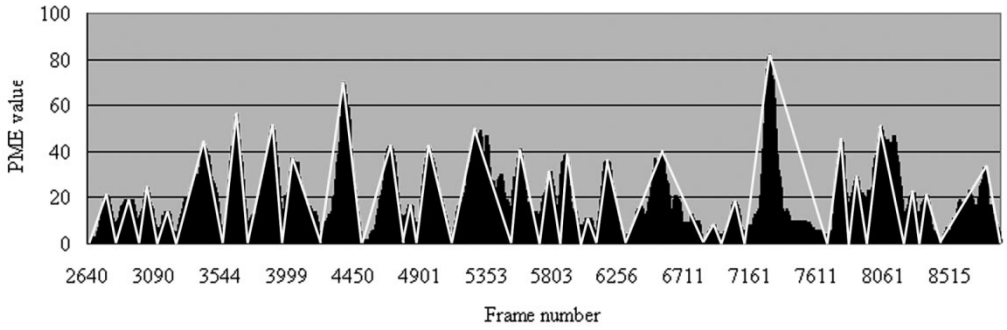

Fig. 4. PME sequence of the first example segment.

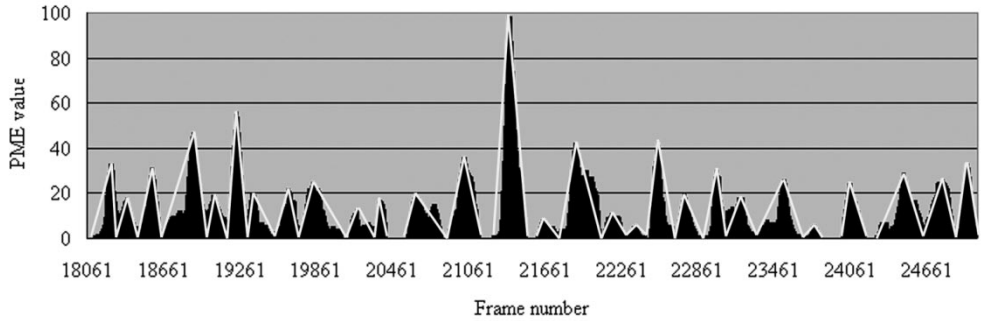

Fig. 5. PME sequence of the second example segment.

Fig. 6. Key frames selected by triangle model for Basketball.

In our test dataset of about 10 h of MPEG video, the difference ratio p over 2579 triangle motion patterns (not including zero motion patterns) is 0.15. It can be seen that the triangle model describes the motion acceleration and deceleration process quite well, and this validates the triangle model.

III. Key-Frame Extraction

As discussed in Section II, since a motion pattern usually corresponds to an action, we extract one key frame for each motion pattern and the turning point of the motion acceleration and deceleration is selected as the key frame. That is, the top vertex of the triangle is selected as the key frame. The advantages of the proposed algorithm are twofold. First, the triangle model of PME segments the whole video sequence into meaningful action events. Each of the action events will be represented by its corresponding key frame such that visual action content within video sequence is fully captured. Second, the turning point of motion acceleration and deceleration usually

Fig. 6. *(Continued)* Key frames selected by triangle model for Basketball.

represents the most salient point of an action event and one can infer the movement within the acceleration process and deceleration process given the turning point. Figs. 6 and 7 show the key frames selected by the triangle model for Basketball (MPEG-7 CD 26 KBS) and Soccer (MPEG-7 CD 28, Samsung), respectively. It can be seen that the key frames capture salient action events in the two sequences.

Before applying the triangle model of PME to detect key frames, a video sequence is segmented into shot. The twin-comparison method [10] is used for this process. For those shots that have motion patterns, the key frames selected by the triangle model are key frames for these shots. For those shots with no motion pattern, we select the first frame as a key frame. Usually, the first frame is usually enough to represent a static shot. Figs. 8 and 9 shows the key frames selected by this rule for Basketball and Soccer, respectively.

If the maximum allowed number of key frames for a sequence is regulated, each shot of the sequence will be

Fig. 7. Key frames selected by triangle model for Soccer.

Fig. 8. Key frames by selecting the first frames of shots for Basketball.

assigned one part of given N key frames according to the percentage share of its motion patterns in the total patterns of the sequence, denoted by N_i. For each shot, motion patterns are sorted by the accumulation of PME, and top N_i patterns

Fig. 9. Key frames by selecting the first frames of shots for Soccer.

TABLE I
EVALUATION RESULTS

	Good	Acceptable	Bad	Total	Key-frame /Shot	Key-frame Percentage
Home video	200 (85.5%)	31 (13.2%)	3 (1.3%)	234	1.8	0.18%
Sports video	110 (82.1%)	23 (17.2%)	1 (0.7%)	134	2.9	0.56%
News video	140 (77.0%)	40 (22.0%)	2 (1.0%)	182	1.6	0.35%
Entertainment	84 (79.2%)	20 (18.9%)	2 (1.9%)	106	2.3	1.00%

are selected as key frames. The first frame of a static shot is selected as the only key frame.

IV. EVALUATION

We have evaluated the performance of the proposed key-frame extraction algorithms by subjective user studies. As there are no benchmarking or ground truth results for key-frame extraction algorithms so far, we do not perform any comparison between this algorithm and others. Testers are asked to give subjective scores to the key-frame extraction results shot by shot. Ten testers from our research lab are involved in the evaluation. The testers are asked to give scores based on their satisfaction to how well the key frames capture the salient content of a shot. The following three-level scales for rating the satisfaction are used: 3. Good, 2. Acceptable, 1. Bad.

The test data library is composed of about 3-h video sequences of various video types including home video, sports video, news video, and entertainment video. Each original sequence is segmented into consecutive shots and key frames are extracted using the proposed algorithm. In the test interface, the original shot clips are showed in the left window, and the key frames are listed in the right window such that testers can clearly see how well the key frames capture the salient content of a shot. The shot clips and their key frames are shown in sequential order along the temporal axis. The evaluation results are shown in Table I. Each row of the table shows the results for a different video type. The column of Total shows the total number of test shots for each video type. The column of Good, Acceptable, and Bad are the numbers of shots that testers give the rate Good, Acceptable, and Bad, respectively. The percentages over total test shots are also shown in these columns. The column of key-frame/shot represents the averaged number of key frames per shot, and the column of key-frame percentage stands for the percentage of key frames over total video frames.

We can see from Table I that the percentages of key frames in sports and entertainment video are several times higher than those of news and home video, and the key-frame/shot values of sports and entertainment video are also larger than those of news and home video. This matches the fact that there is usually much more action in sports and entertainment video than in home and news video. This illustrates that our algorithm truly measures visual content complexity of a shot by action events and determines the number of key frames to abstract a video by action events. We can also see that the percentage of Bad rate is very low and the percentage of Good rate is fairly high. This means that the number of and where to place to key frames in our algorithm can capture the salient visual content within a video sequence.

V. CONCLUSION

In this paper, we have presented a novel key-frame-extraction approach that combines motion-based temporal segmentation and color-based shot detection. The turning point of motion acceleration and deceleration of each motion pattern is selected as a key frame. If a shot is static, the first frame of the video shot is selected as a key frame. With this approach, both the number of key frames and the location of the key frames in a given video are determined automatically by the perceived motion patterns of the video. The proposed approach is threshold free and also fast since motion information in MPEG video can be directly utilized in the motion analysis. Our future work to improve the proposed algorithm includes the integration of color-change analysis and audio cues.

REFERENCES

[1] P. Aigrain, H. Zhang, and D. Petkovic, "Content-based representation and retrieval of visual media: A state-of-the-art review," *Multimedia Tools Applicat.*, vol. 3, pp. 179–202, 1996.
[2] Y. Zhuang, Y. Rui, T. S. Huang, and S. Mehrotra, "Adaptive key-frame extraction using unsupervised clustering," in *Proc. IEEE Int. Conf. Image Processing*, Chicago, IL, Oct. 1998, pp. 886–870.
[3] Y. Rui, T. S. Huang, and S. Mehrotra, "Exploring video structures beyond the shots," in *Proc. IEEE Conf. Multimedia Computing Systems*, Austin, TX, June–July 28–1, 1998, pp. 237–240.
[4] B. Gunsel and A. M. Tekalp, "Content-based video abstraction," in *Proc. IEEE Int. Conf. Image Processing*, Chicago, IL, 1998, pp. 128–132.
[5] H. J. Zhang, J. Y. A. Wang, and Y. Altunbasak, "Content-based video retrieval and compression: A united solution," in *Proc. IEEE Int. Conf. Image Processing*, vol. 1, Oct. 1997, pp. 13–16.
[6] A. Hanjalic, R. L. Lagendijk, and J. Biemond, "A new key-frame allocation method for representing stored video-streams," in *Proc. 1st Int. Workshop Image Databases Multi Media Search*, Amsterdam, The Netherlands, 1996, pp. 64–67.
[7] 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, 1997.
[8] W. Wolf, "Key frame selection by motion analysis," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Proc.*, vol. 2, May 1996, pp. 1228–1231.

[9] P. O. Gresle and T. S. Huang, "Gisting of video documents: A key frames selection algorithm using relative activity measure," in *Proc. 2nd Int. Conf. Visual Information Systems*, 1997, pp. 279–286.

[10] H. J. Zhang, A. Kankanhalli, and S. W. Smoliar, "Automatic partitioning of full-motion video," *ACM Multimedia Syst.*, vol. 1, no. 1, pp. 10–28, 1993.

[11] 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, pp. 1280–1289, Dec. 1999.

[12] Y. Ma and H. J. Zhang, "A new perceived motion based shot content representation," in *Proc. IEEE Int. Conf. Image Processing*, Thessaloniki, Greece, Oct. 7–10, 2001, pp. 426–429.