SCALABLE TEMPORAL INTEREST POINTS FOR ABSTRACTION AND CLASSIFICATION OF VIDEO EVENTS

Seung-Hoon Han, Member, IEEE, and In-So Kweon**, Member, IEEE*

* Digital Media R&D Center, Samsung Electronics Co., Korea
** Robotics and Computer Vision Laboratory, Division of Electrical Engineering, Dept. of EECS, Korea Advanced Institute of Science and Technology, Korea.

ABSTRACT

The image sequence of a static scene includes similar or redundant information over time. Hence, motion-discontinuous instants can efficiently characterize a video shot or event. However, such instants (key frames) are differently identified according to the change of velocity and acceleration of motion, and such scales of change might be different on each sequence of the same event. In this paper, we present a scalable video abstraction in which the key frames are obtained by the maximum curvature of camera motion at each temporal scale. The scalability means dealing with the velocity and acceleration change of motion. In the temporal neighborhood determined by the scale, the scene features (motion, color, and edge) can be used to index and classify the video events. Therefore those key frames provide temporal interest points (TIPs) for the abstraction and classification of video events.

1. INTRODUCTION

A video can be hierarchically abstracted according to its syntactic features and semantics. There exist video semantics such as event or scenario in the highest level. They are represented by syntactic features like editing effects (e.g. cut and dissolve) and key frames (or shots) in the lower level. The lowest level consists of scene features (motion, color, shape, and texture) of key frames or partitioned shots. Although extensive research has been performed on the video hierarchy, the work related to each level was studied independently one another. In other words, shot boundaries, key frames, and their features was not unified to describe higher semantic level.

Specifically, let us imagine a situation where camera motion characterizes a video event by an example of golf highlights. In a drive shot (full-swing) of a player, there exist the consecutive and different camera movements tracking the golf ball flying in the sky and then falling into the field. While a single frame (start or end) frame can not depict the whole event in this case, the key frames corresponding to such motion discontinuities could efficiently summarize the video event, and also scene features (motion, color, edge, and etc.) at those key frames can be used to represent the event. However, the key frames are differently identified according to the scale change of velocity and acceleration of motion, and the scales might be different on each sequence of the same event.

In this paper, we present a scalable video abstraction in which the key frames are obtained by computing motion curvature at each temporal scale. The motion curvature is obtained from motion parameters smoothing around different temporal widow sizes (scales). The key frames are corresponding to maximum curvature points at each scale, which provide efficient temporal interest points (TIPs) for the abstraction and indexing of video events.

Previous attempts for video abstraction are mainly based on color change [1], motion activity [3, 4], and unsupervised clustering [2]. Liu [3] extracted key frames corresponding to the start and end points of motion acceleration. However, this method can not provide key frames according to the scale change of velocity and acceleration of motion. In this paper, TIPs can be extracted according to the scale of motion. As a result, the small number of TIPs is produced if slow and gradual camera motion is focused. More TIPs are extracted when the detail change of motion is considered important.

To classify the events, we define the event descriptor using color, edge, and motion computed around TIPs. In experiments for clustering putting, drive shots and human walking scenes on golf sequences, we demonstrate the proposed TIPs efficiently represent and classify video events using those features.

2. TEMPORAL INTEREST POINTS (TIP)

By considering psychology experiments [5] on human perception of video events, we can know human tends to focus on the points corresponding to changes in speed and direction of motion. In contrast to traditional approaches using the whole spatio-temporal structure, the spatio-temporal interest points closely relate to human perception. For example, in a heading shoot of a soccer game, people

0-7803-9332-5/05/$20.00 ©2005 IEEE

are interested in the point when and where the head and the ball meet [6]. In complex and dynamic scenes, however, the spatio-temporal interest points proposed by Linderberg [6] are not efficient for video abstraction and indexing. To extract features well charactering the video events, we focus on only temporal interest points rather than spatio-temporal ones.

Fig. 1. Coastguard sequence with sequential camera motion, left panning, upward tilting, and right panning.

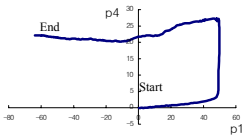


Fig. 2. p_1 and p_4 are the accumulated parameters over time. Two high curvature points are corresponding to one discontinuity between left panning and upward tilting, and the other between upward tilting and right panning in Fig. 1.

As mentioned before, human pays attention to the instants corresponding to changes in speed and direction of motion. We are interested in the temporal interest points produced by camera motion. These temporal instants occur at camera motion discontinuities as shown in Fig. 1. The coastguard sequence includes sequential camera motion, left panning, upward tilting, and right panning. These discontinuities between camera operations can be found using global motion parameters. The six motion parameters (p_1, p_2,…,p_6) are required under affine camera. The velocity (u, v) are defined as

$$u = p_1 + p_2 x + p_3 y \tag{1}$$

$$v = p_4 + p_5 x + p_6 y \tag{2}$$

Accumulating these parameters over time produces a trajectory in the parameter space. Fig. 2 shows a trajectory resulted from the accumulation of the transition parameters (p_1, p_4) over time. Note that the high curvature points on the trajectory are corresponding to camera motion discontinuity and called temporal interest points (TIPs)

3. VIDOE ABSTRACTION BY TIPS

Now, we define the curvature of motion parameters to detect TIPs. The formal definition of the curvature of a curve is given by

$$\kappa = \left| \frac{dT}{ds} \right| \tag{3}$$

where T is the unit tangent vector and s is arc length. In case of two parameters, the trajectory is given by

$$\kappa(t) = \frac{\sqrt{A^2 + B^2 + C^2}}{\left((p_1')^2 + (p_4')^2 + (t')^2 \right)^{3/2}} \tag{4}$$

where

$$A = \begin{vmatrix} p_4' & t' \\ p_4'' & t'' \end{vmatrix}, \quad B = \begin{vmatrix} t' & p_1' \\ t'' & p_1'' \end{vmatrix}, \quad C = \begin{vmatrix} t' & p_4' \\ t'' & p_4'' \end{vmatrix} \quad . \tag{5}$$

The notation $|\cdot|$ denotes the determinant. A discrete approximation is used to compute the derivatives, for example, $p_1'(t) = p_1(t) - p_1(t-1)$, $p_4'(t) = p_4(t) - p_4(t-1)$, and time derivatives $t' = 1$, $t'' = 0$.

As shown in Fig. 2, the original trajectory is very noisy, which naturally requires smoothing the trajectory. Since the speed of motion is different among the events, we analyze the curvature in scale space. We compute the curvature at each scale (σ) ranging 1 to L. If G(t,σ) is a one-dimensional Gaussian kernel of width σ, then the smoothed trajectory (P_1(t),P_4(t)) is given by

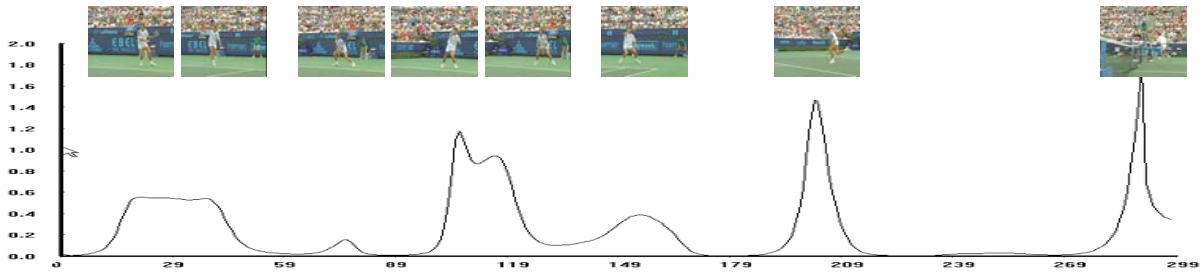
$$P_1(t) = p_1(t) * G(t,\sigma) \qquad P_4(t) = p_4(t) * G(t,\sigma) \tag{6}$$

and the smoothed velocity and acceleration of p_1(t) are

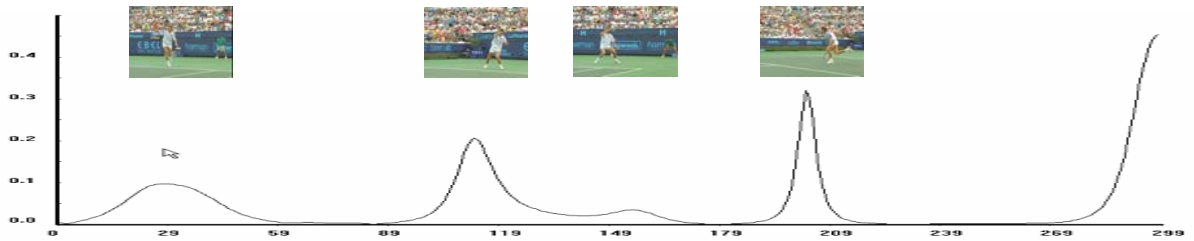
$$P_1'(t) = p_1(t) * \nabla G(t,\sigma) \qquad P_1''(t) = p_1(t) * \nabla^2 G(t,\sigma) \tag{7}$$

Fig. 3 shows the temporal curvature of the accumulated affine motion parameters for the Stefan sequence, which is computed at two different scales. The Stefan sequence includes sequential camera motion, right panning, left panning, holding, left panning, right panning, holding, right panning, and left panning with zooming. Note that maximum curvature points are corresponding to TIPs because the speed and direction of motion change at the instants. Also, the maximum curvatures points are different and characteristic at each scale. Therefore we can represent the video events in terms of TIPs in temporal scale space and their features.

Fig. 4 shows an example of the temporal curvature of camera motion parameters for a putting shot in a golf video where the camera is panning left to track the ball and zoom gradually. Hence, the camera motion



(a) Temporal scale = 4 (vetical axis: cuvature, and horizontal axis: frame number)



(b) Temporal scale = 12

Fig. 3. The temporal curvature of accumulated affine motion parameters computed at temporal scale 4 and 12. The peak points are TIPs and the overlaid images represent the key frames at the discontinuous moments.

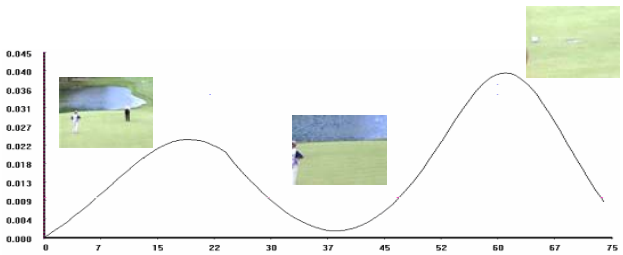


Fig. 4. An golf example of temporal curvature (at scale 14) of camera motion parameters for a putting shot, TIPs (peaks) and representative images between TIPs

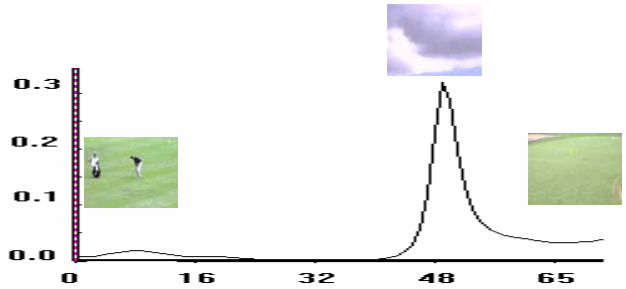


Fig. 5. An golf example of temporal curvature (at scale 8) of camera motion parameters in a drive shot, TIPs and their key frames

discontinuity produces two TIPs corresponding to panning and zooming. Fig. 5 shows the temporal curvature of camera motion for a drive shot where camera fast moves upward to track the ball in the sky and downward to follow the ball on the green. Therefore, two motion discontinuities lead to the same number of TIPs for a drive shot.

4. EXPERIMENTAL RESULTS ON GOLF EVENT CLASSIFICATION

To demonstrate that the proposed TIPs well characterize video events, we utilize visual cues observed from the golf videos, by which we classify them into drive, putting, and arbitrary walking scenes. From the previous section, we can know the camera motion in golf games is a very important cue for identifying a specific event such as drive or putting shots. In addition, the color of the field, sky and bunker is another characteristic to discriminate the events from other scenes. The position of players also can be used to resolve the ambiguity between the similar distributions of color in events and non-events. It is identified using edge distribution of players on the region of green. The information on motion, color and edge are extracted from the images around TIPs. Based on this property, we define the event descriptor as

$$\mathbf{E}(t_i, s) = [\mathbf{c}(t_i, s), \mathbf{e}(t_i, s), \mathbf{M}(t_i, s)] \quad (8)$$

where t_i is a TIP at scale $s \in [1,...,L]$, and $\mathbf{c}(t_i, s)$, $\mathbf{M}(t_i, s)$ and $\mathbf{e}(t_i, s)$ are color (R,G, and B), edge orientation histogram (0˚, 45˚, 90˚, and 135˚), and motion vectors (six affine motion parameters), which are averaged over about $2s$ frames around the TIP t_i. We obtain $\mathbf{c}(t_i, s)$ from the dominant (e.g., largest) regions

detected by color quantization (mean shift algorithm [7]), and $\mathbf{e}(t_i, s)$ from the edge map on the dominant regions. This situation is shown in Fig. 6 where the quantized image and edge map on a dominant region (on the green) are shown in the right column.

We visualize in Fig. 7 the features (motion, color, and edge) around TIPs in test golf clips (Showdown 2000 match). The axes represent the three coefficients obtained from principal component analysis (PCA) applied to the event vector defined in (8). Note that drive, putting, and walking scenes are distinguishable from each other since the adopted features around TIPs represent well the events.

We classify the events by k-NN algorithm and evaluate the performance while changing the combination of the adopted features, motion, color, and edge, as shown in Table 1. The test sequences are captured from the match "Showdown 2000" and including 13 drive, 17 putting shots, and 21 walking scenes. The precision is the fraction of the correctly retrieved events among the all events to which an input event is matched up to rank 13.

Note that the best result is resulted when the three features are simultaneously used to classify the golf events, drive, putting shots and walking scenes.

Table 1. Precision values according to the adopted features (motion (M), color (C), and edge (E))

Precision	M	C	E	M+C	E+M	E+C	M+C+E
Drive	0.46	0.38	0.38	0.42	0.58	0.43	**0.61**
Putting	0.28	0.33	0.25	0.34	0.32	0.47	**0.47**
Walking	0.69	0.52	0.69	0.71	0.73	0.70	**0.73**

10. CONCLUSION

In this paper, we proposed a scalable video summarization where key frames or TIPs (the maxima curvature points of camera motion parameters) are obtained at each temporal scale. As a result, the small number of TIPs is produced if slow and gradual camera motion is favored, while more TIPs are extracted when the detail change of motion is considered important. We also showed that the golf events (putting, drive shots and human walking scenes) can be effectively classified using the features around TIPs. To enhance the classification precision, more sophisticated features will be studied in the future.

11. REFERENCES

[1] H. Zhang, J. Y. A. Wang, and Y. Altunbasak, "An integrated system for content-based video retrieval and browsing", Pattern Recognition, vol. 30, pp. 643-648, 1997.

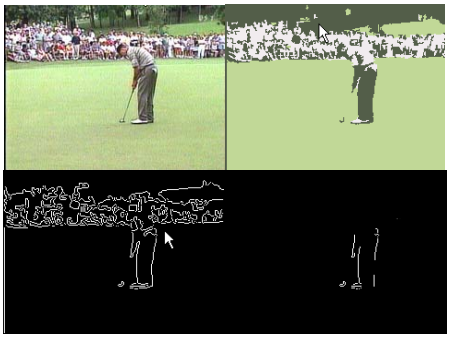


Fig. 6. Original and color-quantized images, edge map on whole regions and a dominant region, arranged from left (top) to right (bottom)

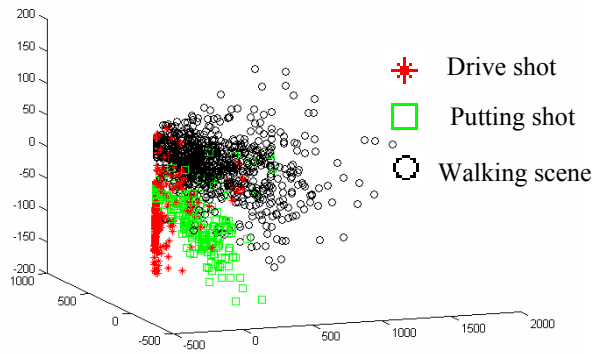


Fig. 7. Feature space spanned by the first three coefficients obtained from PCA applied to the feature vectors (motion, color, and edge)

[2] Y. Zhuang, Y. Rui, T. S. Huang, and S. Mehrotra, "Adaptive key frame extraction using unsupervised clustering", ICIP 1998, pp. 886-870, 1998.

[3] T. M. Liu, H. J. Zhang, F.H. Qi, "A novel video key-frame-extraction algorithm based on perceived motion energy model", vol. 13, Issue 10, pp. 1006-1013, Oct. 2003.

[4] A. Divakaran, R. Regunathan, and K. A. Peker, "Video summarization using descriptors of motion activity: A motion Activity Based Approach to Key-Frame Extraction from Video shots", Journal of Electronic Imaging, vol. 10, pp 909-916, Oct. 2001.

[5] Zacks, J., Tversky, B., "Event structure in perception and cognition," Psychological Bulletin, 127(1), pp 3-21, 2001.

[6] Ivan Laptev and Tony Lindeberg, "Space-time interest points", IEEE International Conference of Computer Vision (ICCV), 2003.

[7] D. Comaniciu and P. Meer, "Robust analysis of feature spaces: Color image segmentation", Proc. 1997 IEEE Conf. Computer Vision and Pattern Recognition, pp. 750-755, June 1997.