Threshold-Based Camera Motion Characterization of MPEG Video

Jae-Gon Kim, Hyun Sung Chang, Jinwoong Kim, and Hyung-Myung Kim

ABSTRACT—We propose an efficient scheme for camera motion characterization in MPEG-compressed video. The proposed scheme detects six types of basic camera motions through threshold-based qualitative interpretation, in which fixed thresholds are applied to motion model parameters estimated from MPEG motion vectors (MVs). The efficiency and robustness of the scheme are validated by the experiment with real compressed video sequences.

Keywords—Video indexing, camera motion, motion vector field, affine motion model.

I. Introduction

Camera motion is a distinct feature that essentially characterizes video content in the context of content-based video representation [1], [2]. Most existing methods for camera motion characterization are based on the analysis of optical flow [1]. In the case of MPEG video, a few methods that directly use MPEG motion vectors (MVs) have been proposed to save the expensive computation of full decompression and optical flow estimation [3], [4]. These previous methods classify the limited set of camera motion types and only deal with limited MPEG coding formats with specific group-of-picture structures, picture coding types, and so on. Our proposed scheme detects six types of basic camera motions directly from an MPEG-1 or MPEG-2 video regardless of its coding format, except in the case of I-picture. In camera motion characterization, we have developed a novel qualitative interpretation method—based on the direct thresholding of motion model parameters—to detect basic camera motions and segment a video sequence into the sub-shots that keep a particular type of camera motion. Unlike the previous work [5] which relied on a statistical approach to avoid an unstable threshold selection, we easily set stable thresholds as fixed values that do not depend on the video content from the practical observation of the induced flow field synthesized by the associated parameters.

Manuscript received July 5, 2003.

This work was supported in part by the Ministry of Information and Communication of Korea.

Jae-Gon Kim (phone: +82 42 860 4980, email: jgkim@etri.re.kr), Hyun Sung Chang (email: chs@etri.re.kr), and Jinwoong Kim (email: jwkim@etri.re.kr) are with Digital Broadcasting Research Division, ETRI, Daejeon, Korea.

Hyung-Myung Kim (email: hmkim@csplab.kaist.ac.kr) is with the Department of Electrical Engineering and Computer Science, Korea Advanced Institute of Science and Technology (KAIST), Daejeon, Korea.

II. MV Field Generation and Motion Model Fitting

We generate an MV field (MVF) for every frame as an alternative to the optical flow field by normalizing the types of MVs extracted from the MPEG video. The normalization converts every MV into a forward-predicted MV between consecutive frames regardless of the picture coding type or the prediction mode as follows [6]: i) convert two field-MVs to a frame-MV for the case of the field prediction in MPEG-2; ii) map a backward-predicted MV or bidirectionally-predicted MVs into a forward-predicted one; and iii) estimate MVs for I-frames by interpolating the MVs of the two nearest P-frames. We filter out the suspected noisy components, especially in a nearly uniform area, by applying median filters to the magnitudes of the MV.

We choose the affine motion model to fit the generated MVF in the consideration that it is more resilient to the noisy and sparse MVF conditions since it is a lower order model and therefore less sensitive to noisy data while preserving most of the basic camera motion types. In the affine motion model, the MV(u, v) of a macroblock centered at (x, y) pixel-point is expressed as:

$$\begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} a_2 & a_3 \\ a_5 & a_6 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} a_1 \\ a_4 \end{pmatrix},$$

where $\Phi = (a_1, a_2, a_3, a_4, a_5, a_6)$ is the parameter vector estimated from the MVF of a given frame by using the least square method. In this process, we remove the outlier vectors and iterate the regression process until the estimates of the six parameters converge.

III. Qualitative Interpretation of Camera Motion

The estimated parameters vector can be expressed in another form more directly related to the physical camera motion:

$$\Phi' = (pan, tilt, div, rot, hyp),$$

where
$$pan = a_1, \qquad tilt = a_4,$$
$$div = \frac{1}{2}(a_2 + a_6), \qquad rot = \frac{1}{2}(a_5 - a_3),$$
$$hyp = \frac{1}{4}(|a_2 - a_6| + |a_3 + a_5|).$$

The terms *pan*, *tilt*, *div* and *rot* represent components of the MVF induced by the following camera operations: pan or horizontal tracking, tilt or vertical tracking, zoom or forward/backward tracking, and rotation. On the other hand, *hyp* represents a hyperbolic flow that is likely observed when object motion is dominant over global motion. We classify this kind of complex motion as a type of object motion under which the dominant object motion or any undefined complicated motion falls.

In our scheme, we divide camera motion types into two sub-groups: one consisting of zoom and rotation represented by linear parameters, and the other consisting of pan and tilt represented by translational parameters. Then, we sequentially detect all camera motion types through a two-step detection (frame-level and segment-level) in the following order: zoom and rotation, pan and tilt, object motion, and static motion. In this way, we can easily detect the more dominant camera motion in cases when different types are combined together. For example, a zoom is detected first as a dominant motion when it is combined with a pan, based on the assumption that a zoom is generally more distinct and more significant in terms of semantic meaning. The detection of a dominant camera motion in the same sub-group is decided by comparing the mean value of the magnitudes of the associated parameters.

In the frame-level detection, we apply thresholds of T_{lin} and T_{tr} to the magnitudes (absolute values) of the linear motion parameters (*div*, *rot*, *hyp*) and translational motion parameters (*pan*, *tilt*), respectively. These thresholds can be set as fixed values in the sense that the magnitude of flow needed to be observed as one of the defined types of camera motion does not depend on the underlying content. Therefore, they are set to the minimum values (T_{lin}=0.45 × 1/*frame_rate* and T_{tr}=30 × 1/*frame_rate*) that make perceptible specific flow patterns by observation of the synthesized flow fields with the associated values. In the segment-level detection, we finally detect the segments that have specific camera motion types by thresholding on the segment-duration in which all the frames, based on the frame-level detection, are detected as the same type. We also easily set a threshold T_{temp} as the minimum segment-duration by observing that a camera motion is generally maintained for at least a half second. As a result, most of the segments are detected as one of the camera motion types while the others are not labeled as any type. In the sub-shot segmentation, a full video sequence is partitioned into the sub-shots of a homogeneous camera motion by merging these residual segments into the detected neighboring segments according to a simple rule derived from the observation.

IV. Experimental Results

We validated the proposed scheme by experimenting on two 5-min MPEG-1 video sequences coming from the MPEG-7 test set: golf and soccer videos. Both of them contain very complicated motions including large object motions as well as various types of camera motions. Figure 1 shows an example of a generated MVF overlaid on a corresponding frame.

Fig. 1. A generated MVF overlaid on a corresponding frame, frame 2,222 of a golf video.

To evaluate the performance of the MVF generation, we compared the motion parameters estimated from the MVF to those estimated from the flow field, which were derived by the block matching algorithm between consecutive frames, as shown in Fig 2. The estimates of the two methods differ slightly but each

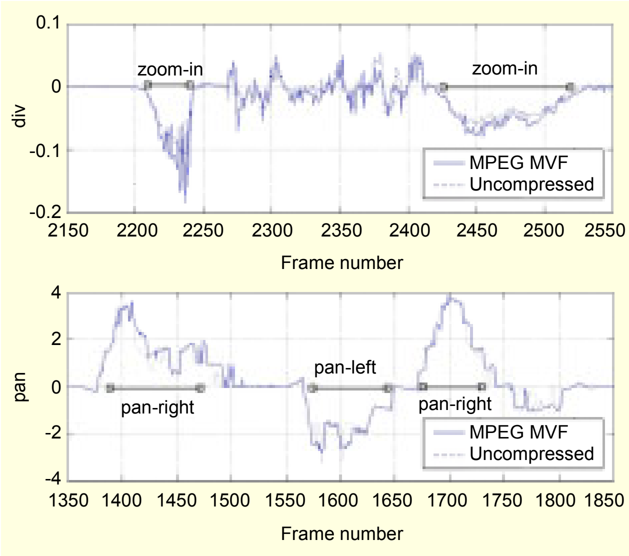

Fig. 2. The estimated *div* and *pan* obtained by using the MVF from MPEG MVs and the flow field estimated from uncompressed domain (golf video), respectively.

has similar temporal evolution results, giving them similar detection results as well.

To verify the qualitative interpretation, we show an example of zoom detection applied to the golf sequence in Fig. 3. We exclude the frame inducing hyperbolic component that has different signs of the parameters a_2 and a_6 as shown in Fig. 3(b).

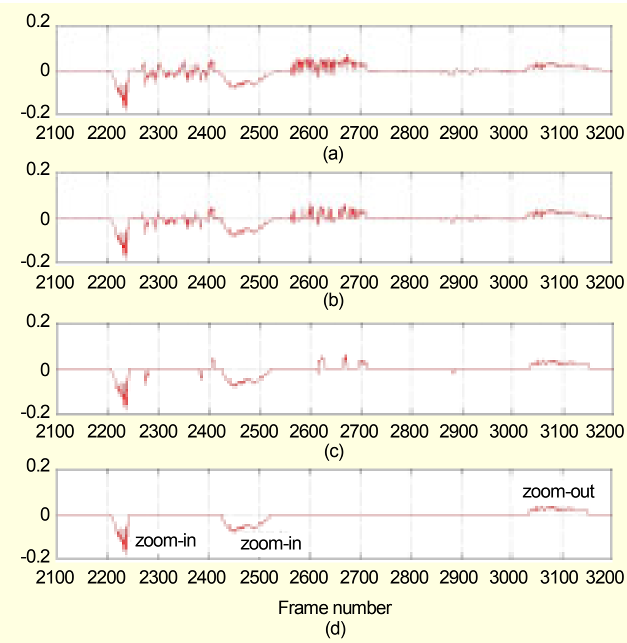

Fig. 3. Zoom detection in the golf video: (a) the estimated *div*, (b) the result of the sign validation, (c) the result of the frame-level detection, and (d) the result of the segment-level detection.

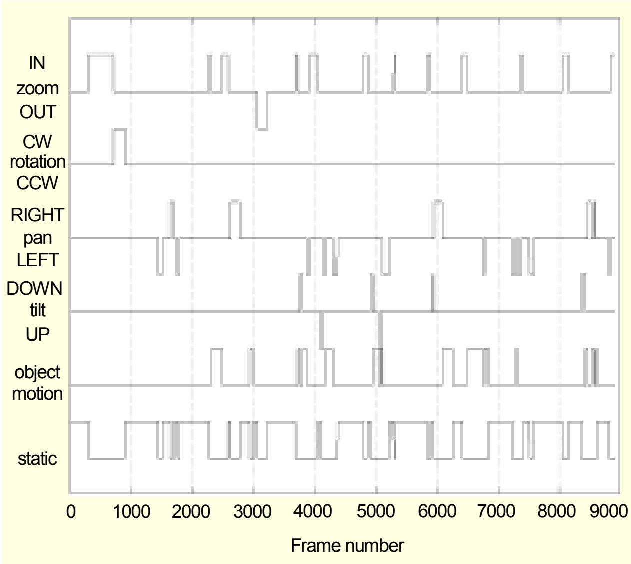

Fig. 4. The results of the sub-shot segmentation in the golf video.

Figure 4 illustrates the sub-shot segmentation results. It turns out that the results are almost in accordance with those given by the manual segmentation. The results of zoom detection are shown in Table 1, where the number of detected zoom-segments is compared with the ground truth. These show that the proposed method yields fairly good performance. For the soccer video, two zoom-segments are missed due to the unreliable MVs in the frame-repeated slow motion scenes.

Table 1. Performance of zoom detection.

Sequences	Zoom-segments	Correctly detected	Missed	Falsely detected
Golf	13	13	0	0
Soccer	19	17	2	2

V. Conclusions

We described an efficient and robust scheme for the camera motion characterization in MPEG video. Our scheme includes the generation of an MVF from the MPEG MVs and the threshold-based qualitative interpretation of the camera motions. The results on real video sequences show that the scheme works well in the detection of the well-known basic camera motions and the sub-shot segmentation. Further applications of the proposed scheme will include semantic event detection, summarization, and so on.

References

[1] M.V. Srinivasan, S. Venkatesh, and R. Hosie, "Qualitative

Estimation of Camera Motion Parameters from Video Sequences," *Pattern Recognit*., vol. 30, no. 4, Apr. 1997, pp. 593-606.

[2] J.-G. Kim, H.S. Chang, Y.-t. Kim, K. Kang M. Kim, J. Kim, and H.-M. Kim, "Multimodal Approach for Summarizing and Indexing News Video," *ETRI J*., vol. 24, no. 1, Feb. 2002, pp. 1-11.

[3] Y.-P. Tan, D.D. Saur, S.R. Kulkarni, 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, Feb. 2000, pp. 133-146.

[4] A. Benzougar, P. Bouthemy, and R. Fablet, "MRF-Based Moving Object Detection from MPEG Coded Video," *Proc. Int'l Conf. Image Processing*, Oct. 2001, pp. 402-405.

[5] P. Bouthemy, M. Gelgon, and F. Ganansia, "A Unified Approach to Shot Change Detection and Camera Motion Characterization," *IEEE Trans. Circuits Syst. Video Technol*., vol. 9, no. 7, Oct. 1999, pp. 1030-1044.

[6] J.-G. Kim, K.-W. Lee, J. Kim, and H.-M. Kim, "Extraction of Moving Objects from MPEG-Compressed Video for Object-Based Indexing," *Proc. 1st European Workshop on CBMI*, Oct. 1999, pp. 131-139.