VIDEO RETRIEVAL FROM COMPRESSED VIDEOS

1. M. NAGARAJU 2. M. THANOJ 3. P. MANIKANTA 4. T.K. PRANEETH 5. S. BHAVANI
1,3,4,5.Assistant Professor, IT Dept, Gudlavalleru Engineering College, Gudlavalleru
2. HOD, CSE Dept, Andhra Loyola Institute of Engineering and Technology, Vijayawada

Abstract

This paper proposes a thorough scheme, by virtue of camera zooming descriptor with two-level threshold, to automatically retrieve close-ups directly from moving picture experts group (MPEG) compressed videos based on camera motion analysis. A new algorithm for fast camera motion estimation in compressed domain is presented. In the retrieval process, camera-motion-based semantic retrieval is built. To improve the coverage of the proposed scheme, close-up retrieval in all kinds of videos is investigated. Extensive experiments illustrate that the proposed scheme provides promising retrieval results under real-time and automatic application scenario.

Keywords

Camera motion analysis, close-up retrieval, moving picture experts group (MPEG) compressed videos

I. INTRODUCTION

Moving picture experts group (MPEG) video compression techniques are already well developed and widely deployed, and the goal of effective retrieval directly in compressed domain has yet to be realized. However, most of the existing digital video processing techniques present difficulties to extract any meaningful features for further analysis from MPEG compressed videos. To extract relevant features, the content should in principle be decoded first, since this operation is time consuming, especially when a large video database should be processed, features extraction directly in compressed domain would be particularly interesting by providing fast and reliable information retrieval and analysis tools. While considerable research has been conducted to achieve image or video analysis, such as content feature extraction at low levels[1, 2], semantic feature extraction at high levels[3, 4], and video indexing/retrieval[5−11], only few solutions have been given to this challenging task with limited decoding of MPEG video streams. To this end, our work in this paper focuses on automatically retrieving the high-level semantic concept, i. e., close up, directly in MPEG compressed domain based on camera motion analysis. In order to reduce the semantic gap, we propose a thorough scheme with two stages: camera motion feature extraction directly in MPEG compressed domain, and then semantic retrieval for close-ups via exploiting camera zooming descriptor with two-level threshold, but without human intervention, especially under real-time application scenario, and worthy of investigation.

Camera motion is an important feature in video analysis. Various camera operations used in video production have been described[12−16]. In many applications, however, only pan, tilt, and zoom parameters are considered. Furthermore, models of camera motion can be used to detect important events. Herein, a new algorithm for fast camera motion estimation in compressed domain is presented.

Previous research on close-up detection is mostly in sports video[17−19]. To improve the coverage of the proposed scheme, we investigate close-up retrieval extensively in all kinds of videos. The basic principle is to extract zooming descriptor directly in compressed domain, and when its value is larger than a certain two-level threshold, the frame is retrieved as a close-up.

The remainder of this paper is organized as follows. Section II introduces the new algorithm for fast camera motion estimation in compressed domain. Section III presents automatic close-up retrieval from MPEG compressed videos. Section IV contains the experimental results and evaluations. Section V provides conclusions.

II. FAST CAMERA MOTION ESTIMATION INCOMPRESSED DOMAIN

In the past few decades, significant research has been carried out to estimate the camera motion parameters in pixel domain. The general principle for all the work can be described as follows. Assuming that camera is undergoing rotation and zoom but no translation, the change of image intensity between neighboring frames can be modeled by the following 6-parameter projective transformation:

$$x' = \frac{p_1 x + p_2 y + p_3}{p_5 x + p_6 y + 1}$$
$$y' = \frac{-p_2 x + p_1 y + p_4}{p_5 x + p_6 y + 1} \quad (1)$$

where (*x*, *y*) and (*x*_, *y*_) are the image coordinates of corresponding points in two neighboring frames and [*p*1…,*p*6] are parameters of camera motion.

Reference [16] put forward a compressed domain parameter estimation under the assumption that from one frame to the next: 1) we can set *p*5 = *p*6 = 0, i. e., that perspective distortion effects are minimal; 2) we can set *p*2 = 0, i. e., that the camera does not rotate about the axis of the camera lens. In this case, the 6-parameter model can be approximated by the three-parameter transformation:

$$\begin{pmatrix} x' \\ y' \end{pmatrix} = \begin{pmatrix} p_1 & 0 \\ 0 & p_1 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} p_3 \\ p_4 \end{pmatrix}. \quad (2)$$

s is referred to as the camera zoom factor, *fα* as the camera pan rate, and *fβ* as the camera tilt rate, respectively. A zooming action takes place by a change of camera focal length from *f* to *f'*, which is characterized by *s* = *f/f'*_. As *p*1 = *s*, *p*3 = −*sfα*, and *p*4 = *sfβ*, (2) can be rearranged into

$$\begin{pmatrix} x' \\ y' \end{pmatrix} = s\left[\begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} -f\alpha \\ f\beta \end{pmatrix}\right]. \quad (3)$$

Since MPEG has already exploited inter-frame redundancy via motion estimation and compensation, the above parameters can be quickly estimated via *P*-frames, in which correspondence can be approximated by the coordinates of corresponding macro-blocks from image point (*x*, *y*) to (*x*_, *y*_), connected by motion vector *MV*, where *MVx* = *x'*− *x*, *MVy* = *y'*− *y*.

Let (*i*0, *j*0) represent the image coordinates (i. e., standard image coordinates) of the center of the image, and (*ik*, *jk*) present the row and column coordinates of the center of the *k*-th inter coded macro-block in the current *P*-frame. Then, with respect to image-centered Cartesian axes, the center of the *k*-th inter coded macro-block has coordinates:

$$\begin{pmatrix} x'_k \\ y'_k \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}\left[\begin{pmatrix} i_k \\ j_k \end{pmatrix} - \begin{pmatrix} i_0 \\ j_0 \end{pmatrix}\right] =$$
$$\begin{pmatrix} j_k - j_0 \\ i_0 - i_k \end{pmatrix} \quad (4)$$

and its corresponding motion vector in consistent units is (MV_{yk}, $-MV_{xk}$).

This macro-block is matched with the point in the previous anchor frame that has image-centered Cartesian coordinates:

$$\begin{pmatrix} x_k \\ y_k \end{pmatrix} = \begin{pmatrix} j_k - j_0 - MV_{xk} \\ i_0 - i_k - MV_{yk} \end{pmatrix}. \quad (5)$$

Hence, we take image point from (*x*, *y*) to (*x'*, *y'*) in each inter coded macro block as a sample of the unknown projective transformation. These samples can then be used to form a linear-in-the-parameters least-squares regression problem that can be solved to determine estimates of the unknown parameters of the projective transformation. In this case, the cost to be minimized is

$$Q = \sum_{k=1}^{N}\left\|\begin{pmatrix} x'_k \\ y'_k \end{pmatrix} - s\begin{pmatrix} x_k \\ y_k \end{pmatrix} - \begin{pmatrix} -sf\alpha \\ sf\beta \end{pmatrix}\right\| =$$
$$\sum_{k=1}^{N}\left\|\begin{pmatrix} j_k - j_0 \\ i_0 - i_k \end{pmatrix} - s\begin{pmatrix} j_k - j_0 - MV_{xk} \\ i_0 - i_k - MV_{yk} \end{pmatrix} - \quad (6)$$
$$\begin{pmatrix} -sf\alpha \\ sf\beta \end{pmatrix}\right\|.$$

Finally, after a series of mathematical manipulation, the camera zoom parameter *s* can be estimated via the following equation:

$$s^* = \frac{\sum_{k=1}^{N}\left[\Delta i_k(\Delta i_k + \Delta MV_{yk}) + \Delta j_k(\Delta j_k - \Delta MV_{xk})\right]}{\sum_{k=1}^{N}\left[(\Delta i_k + \Delta MV_{yk})^2 + (\Delta j_k - \Delta MV_{xk})^2\right]} \quad (7)$$

where

$$\Delta i_k = i_k - \frac{1}{N}\sum_{k=1}^{N} i_k$$

$$\Delta j_k = j_k - \frac{1}{N}\sum_{k=1}^{N} j_k$$

$$\Delta MV_{xk} = MV_{xk} - \frac{1}{N}\sum_{k=1}^{N} MV_{xk}$$

$$\Delta MV_{yk} = MV_{yk} - \frac{1}{N}\sum_{k=1}^{N} MV_{yk}.$$

In an MPEG video, all of the above required data can be obtained directly: (*ik*, *jk*) are the image coordinates of the center of the *k*-th inter coded macro-block, (MV_{xk},MV_{yk}) is derived from its corresponding motion vector, and *N* is the number of inter coded macro blocks.

III. CLOSE-UP RETRIEVAL

General observations reveal that close-ups are dominated by large foreground areas inside the video frames. Most of the close-ups can be detected by measuring the proportion of its foreground size to the frame size. To improve the coverage of close-up detection and implement it extensively, the camera motion descriptor is utilized, i. e., zooming descriptor with two-level threshold to automatically retrieve close-ups. The zoom factor modifies the perspective projection, not the displacement of world points relative to the camera reference frame. The parameter *s* indicates the zooming effect of the camera, where $s > 1$ represents zoom in, $s < 1$ represents zoom out, and $s = 1$ represents no zoom. Let the class labels for threshold be denoted as *THS* and *THM*; the former represents the threshold for a single detected zoom-in frame, and the latter represents the threshold for continuous multiple detected zoom-in frames. Then, the threshold operation is as follows: 1) If only one single zoom-in frame is detected, then the threshold is set to *THS*. Moreover, on the condition that its zoom factor *s* has a higher value than *THS*, this frame is retrieved as a close-up; 2) If continuous multiple zoom-in frames are detected, then the threshold is set to *THM*. Meanwhile, the maximum zoom factor $s_{\max}$ and the minimum zoom factor $s_{\min}$ for these continuous frames are obtained. If the ratio of $s_{\max}$ to $s_{\min}$ is larger than *THM*, these frames are retrieved as close-ups.
The above can be described as

$$C_F = \begin{cases} F_S, & \text{if } s > TH_S \\ F_M, & \text{if } \dfrac{s_{\max}}{s_{\min}} > TH_M \end{cases} \tag{8}$$

where *FS* denotes a single detected zoom-in frame, *FM* denotes continuously multiple detected zoom-in frames, and *CF* denotes retrieved close-ups. The proposed automatic close-up retrieval directly in MPEG compressed domain using zoom-in descriptor with two-level threshold is summarized in Fig. 1. It is well known that a limited number of samples cannot cover a wide variety of videos. The threshold determination issue is basically an ill-posed problem by nature. Therefore, a better solution to this problem is to use statistics to estimate a value, or to be determined empirically. After beginning with using a good theoretical foundation, the proposed approach is able to better retrieve close-ups using much fewer heuristics than conventional methods require. As for our close-up retrieval scheme, *THS* is set to 1.001, and *THM* is set to IV.

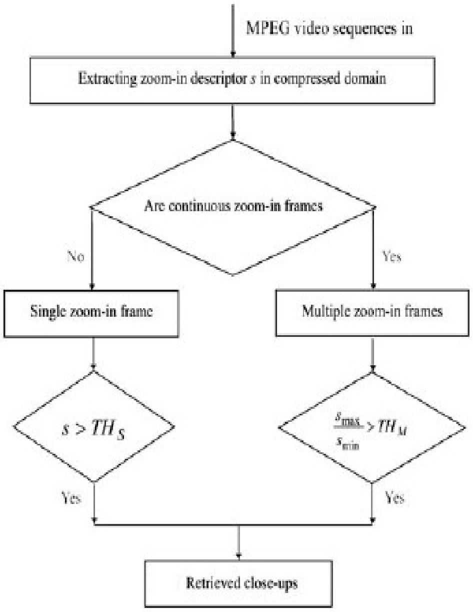

MPEG video sequences in
Extracting zoom-in descriptor s in compressed domain
Are continuous zoom-in frames
No
Yes
Single zoom-in frame
Multiple zoom-in frames
$s > TH_S$
$\frac{s_{\max}}{s_{\min}} > TH_M$
Yes
Yes
Retrieved close-ups

Fig. 1 Overview of the proposed close-up retrieval

IV. EXPERIMENTAL RESULTS AND EVALUATIONS

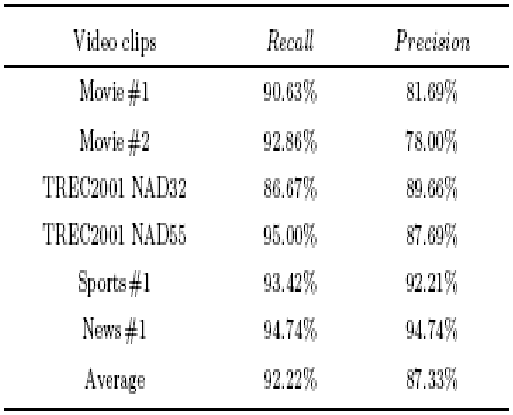
To evaluate the performance, the proposed automatic close-up retrieval scheme are tested on a database with different MPEG video clips, including documentary videos grabbled from well-known TREC2001 video sequences, movies, sports, and news. These video clips are chosen due to the complexity of extensive graphical effects. Videos are captured at a speed of 30 fps, a frame size of 352×240, and stored in MPEG format. The ground truth for the numbers of close-ups in these video clips are determined manually. In order to assess the accuracy, a statistical performance measurement for each video clip is implemented. The values for the quality are defined as recall and precision:

$$Recall = \frac{D}{D + MD} \qquad (9)$$

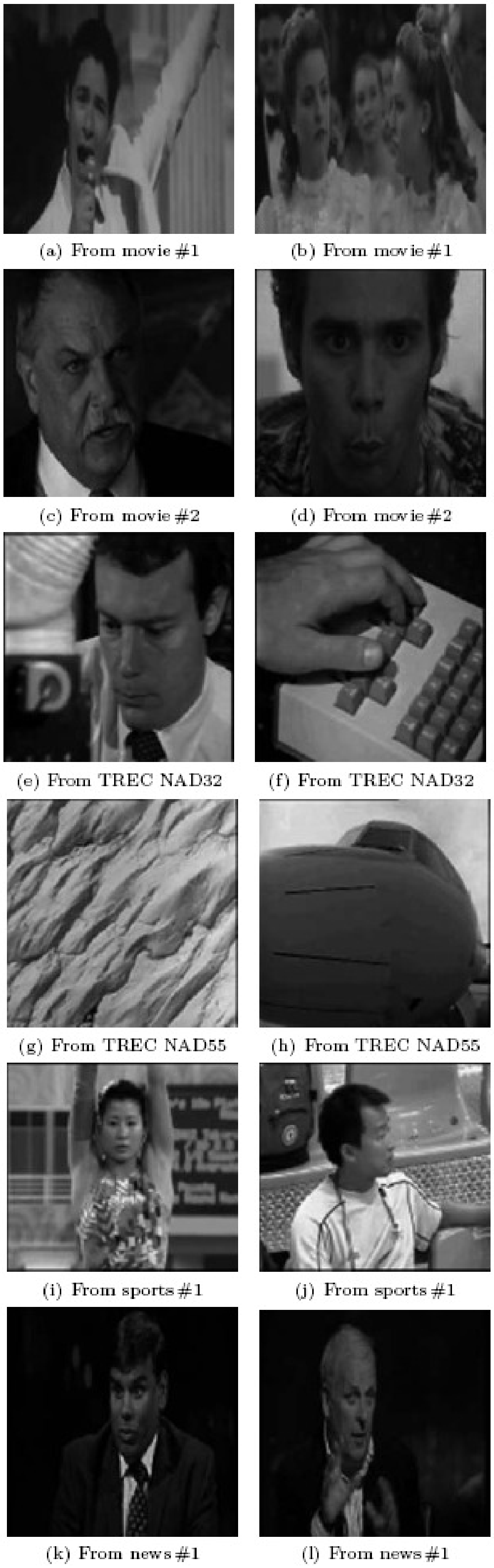
$$Precision = \frac{D}{D + FA} \qquad (10)$$

where *D* is the number of correct retrievals, *MD* is the number of missed retrievals, and *FA* is the number of false alarm. *Recall* measures the ratio of correct retrievals to the ground truth in a video clip, while *Precision* measures the ratio of correct retrievals to the total retrievals by algorithm. Based on this evaluation metric, we conduct experiments using the above test database, and the corresponding recall and precision rates for close-up retrieval are listed in Table 1. From Table 1, it can be seen that the proposed close-up retrieval algorithm achieves on average 92.22% recall rate with 87.33% precision rate. The results demonstrate that the proposed scheme is computationally efficient and consistent, also achieve superior performances in terms of both recall and precision rates. The close-ups of persons and objects can both be retrieved correctly and effectively. The samples of retrieval results of close-ups from test video clips are shown in Fig. 2. Furthermore, we analyze the experimental data in detail, and obtain that the camera zoom factor *s* does not increase gradually, but fluctuates up and down, when the camera continues zooming in.

Table 1 Summary of experimental results for close-up retrieval



Video clips	*Recall*	*Precision*
Movie #1	90.63%	81.69%
Movie #2	92.86%	78.00%
TREC2001 NAD32	86.67%	89.66%
TREC2001 NAD55	95.00%	87.69%
Sports #1	93.42%	92.21%
News #1	94.74%	94.74%
Average	92.22%	87.33%



(a) From movie #1 (b) From movie #1

(c) From movie #2 (d) From movie #2

(e) From TREC NAD32 (f) From TREC NAD32

(g) From TREC NAD55 (h) From TREC NAD55

(i) From sports #1 (j) From sports #1

(k) From news #1 (l) From news #1

Fig. 2 Samples of close-up retrieval results from the test database.

V. CONCLUSION

The main contributions of this paper are summarized as follows. By exploiting the new fast camera motion estimation in compressed domain, close-up retrieval using zooming descriptor with two-level threshold is built. The whole process is under real-time application scenario and without human intervention. The usability and efficiency of the proposed scheme are demonstrated through extensive experiments. It is shown that the computational complexity and the retrieval performance are well balanced in the proposed scheme. Close-ups of persons and objects can both be retrieved correctly and effectively.

REFERENCES

[1] J. Jiang, Y. Weng, P. J. Li. Dominant Colour Extraction in DCT Domain. *Image and Vision Computing Journal*, vol. 24, no. 12, pp. 1269–1277, 2006.

[2] J. Jiang, Y. Weng. Video Extraction for Fast Content Access to MPEG Compressed Videos. *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 14, no. 5, pp. 595–605, 2004.

[3] J. Vendrig, M. Worring. Systematic Evaluation of Logical Story Unit Segmentation. *IEEE Transactions on Multimedia*, vol. 4, no. 4, pp. 492–499, 2002.

[4] H. W. Agius, M. C. Angelides. Modeling Content for Semantic-level Querying of Multimedia. *Multimedia Tools and Applications*, vol. 15, no. 1, pp. 5–37, 2001.

[5] T. Athanasiadis, P. Mylonas, Y. Avrithis, S. Kollias. Semantic Image Segmentation and Object Labeling. *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 17, no. 3, pp. 298–312, 2007.

[6] D. Djordjevic, E. Izquierdo. An Object- and User-driven System for Semantic-based Image Annotation and Retrieval. *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 17, no. 3, pp. 313–323, 2007.

[7] D. Vallet, P. Castells, M. Fernandez, P. Mylonas, Y. Avrithis. Personalized Content Retrieval in Context Using Ontological Knowledge. *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 17, no. 3, pp. 336–346, 2007.

[8] J. W. Hsieh, S. L. Yu, Y. S. Chen. Motion-based Video Retrieval by Trajectory Matching. *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 16, no. 3, pp. 396–409, 2006.

[9] K. W. Sze, K. M. Lam, G. Qiu. A New Key Frame Representation for Video Segment Retrieval. *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 15, no.

9, pp. 1148–1155, 2005. [10] F. Jing, M. Li, H. J. Zhang, B. Zhang. Relevance Feedback in Region-based Image Retrieval. *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 14, no. 5, pp. 672–681, 2004.

[11] S. Antani, R. Kasturi, R. Jain. A Survey on the Use of Pattern Recognition Methods for Abstraction, Indexing and Retrieval of Images and Video. *Pattern Recognition*, vol. 35, no. 4, pp. 945–965, 2002.

[12] Y. H. Ho, C.W. Lin, J. F. Chen, H. Y. M. Liao. Fast Coarseto-fine Video Retrieval Using Shot-level Spatio-temporal Statistics. *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 16, no. 5, pp. 642–648, 2006.

[13] D. Farin, P. H. N. De. With. Enabling Arbitrary Rotational Camera Motion Using Multisprites with Minimum Coding Cost. *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 16, no. 4, pp. 492–506, 2006.

[14] Y. Su, M. T. Sun, V. Hsu. Global Motion Estimation from Coarsely Sampled Motion Vector Field and the Applications. *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 15, no. 2, pp. 232–242, 2005.

[15] J. C. Huang, W. S. Hsieh. Automatic Feature-based Global Motion Estimation in Video Sequences. *IEEE Transactions on Consumer Electronics*, vol. 50, no. 3, pp. 911–915, 2004.

[16] Y. P. Tan, D. D. Saur, S. R. Kulkarni, P. J. Ramadge. Rapid Estimation of Camera Motion from Compressed Video with Application to Video Annotation. *IEEE Transactions on*

Circuits and Systems for Video Technology, vol. 10, no. 1, pp. 133–146, 2000.

[17] L. Liu, X. Ye, M. Yao, S. Zhang. A Semantic Description Scheme of Soccer Video Based on MPEG-7. In *Proceedings of the 5th Pacific Rim Conference on Multimedia, Lecture Notes in Computer Science*, Springer-Verlag, Tokyo, Japan, vol. 3332, pp. 298–305, 2004.

[18] D. W. Tjondronegoro, Y. P. Chen, B. Pham. Classification of Self-consumable Highlights for Soccer Video Summaries. In *Proceedings of the 5th IEEE International Conference on Multimedia and Expo*, IEEE Press, Taipei, PRC, vol. 1, pp. 579–582 , 2004.

[19] G. Jin, L. Tao, G. Xu. Hidden Markov Model Based Events Detection in Soccer Video. In *Proceedings of International Conference of Image Analysis and Recognition, Lecture Notes in Computer Science*, Springer-Verlag, Porto, Portugal, vol. 3211, pp. 605–612, 2004.