

Real-time and Automatic Close-up Retrieval from Compressed Videos

Ying Weng and Jianmin Jiang

School of Informatics
University of Bradford, UK
{Y.Weng, J.Jiang1}@bradford.ac.uk

Abstract—In this paper, we propose a thorough scheme, by virtue of camera zooming descriptor with two-level threshold, to automatically retrieve close-ups directly from MPEG compressed videos based on camera motion analysis. In the retrieval process, we build camera-motion-based semantic retrieval. To improve the coverage of the proposed scheme, we investigate close-up retrieval in all kinds of videos. Extensive experiments illustrate that the proposed scheme provides promising retrieval results under real-time and automatic application scenario.

Keywords—camera motion analysis; MPEG compressed videos; close-up retrieval

I. INTRODUCTION

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 [2], [3], semantic feature extraction at high levels [4], [5], and video indexing/retrieval [6]-[12], 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. Here, 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 presented [14]-[18]. In many applications, however, only pan, tilt, and zoom

parameters are considered. Furthermore, models of camera motion can be used to detect important events.

Previous research on close-up detection is mostly in sports video [19]-[21]. 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 larger than a certain two-level threshold, the frame is retrieved as a close-up.

The remainder of the paper is organized as follows. Section II introduces 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 IN COMPRESSED 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 neighbouring frames can be modelled by the following 6-parameter projective transformation:

$$\begin{aligned}x' &= \frac{p_1x + p_2y + p_3}{p_5x + p_6y + 1} \\y' &= \frac{-p_2x + p_1y + p_4}{p_5x + p_6y + 1}\end{aligned}\quad (1)$$

where (x, y) and (x', y') are the image coordinates of corresponding points in two neighbouring frames and p_1, \dots, p_6 are parameters of the camera motion.

Reference [18] put forward a compressed domain parameter estimation under the assumption that from one frame to the next: (i) We can set $p_5 = p_6 = 0$, i.e., that perspective distortion effects are minimal; (ii) 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, $f\alpha$ as the camera pan rate, and $f\beta$ 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 $p_1 = s$, $p_3 = -sf\alpha$ and $p_4 = sf\beta$, (2) can be rearranged into:

$$\begin{pmatrix} x' \\ y' \end{pmatrix} = s \begin{pmatrix} x \\ y \end{pmatrix} + b, \quad b = \begin{pmatrix} q_3 \\ q_4 \end{pmatrix}. \quad (3)$$

where $q_3 = -f\alpha$ and $q_4 = f\beta$.

Since MPEG 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 macroblocks from image point $W_{\Delta}(x, y)$ to $W'_{\Delta}(x', y')$ (connected by motion

vector): $w_k = \begin{pmatrix} x_k \\ y_k \end{pmatrix} \Leftrightarrow w'_k = \begin{pmatrix} x'_k \\ y'_k \end{pmatrix}$. Therefore, s and b can be estimated by finding the values that minimize:

$$J(s, b) \triangleq \sum_{k=1}^N \|w'_k - s(w_k + b)\|^2. \quad (4)$$

where N is the number of intercoded macroblocks.

Finally, after series of mathematical manipulation, it is concluded in [18] that the zoom parameter s can be estimated via the following equation:

$$s^* = \frac{\sum_{k=1}^N (w'_k - \bar{w}')^T (w_k - \bar{w})}{\sum_{k=1}^N \|w_k - \bar{w}\|^2}. \quad (5)$$

where $\bar{w} = \frac{1}{N} \sum_{k=1}^N w_k$ and $\bar{w}' = \frac{1}{N} \sum_{k=1}^N w'_k$.

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 utilised, 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 thresholding be denoted as TH_S and TH_M ; 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 thresholding operation is as follows: (i) If only one single zoom-in frame is detected, then the

threshold is set to TH_S . Moreover, on the condition that its zoom factor s has a higher value than TH_S , this frame is retrieved as a close-up; (ii) If continuous multiple zoom-in frames are detected, then the threshold is set to TH_M . 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 TH_M , these frames are retrieved as close-ups.

The above can be described as:

$$C_F = \begin{cases} F_S, & \text{if } s > TH_S \\ F_M, & \text{if } \frac{s_{\max}}{s_{\min}} > TH_M \end{cases}. \quad (6)$$

Where F_S denotes a single detected zoom-in frame, F_M denotes continuously multiple detected zoom-in frames, and C_F 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, our approach is able to better retrieve close-ups using much fewer heuristics than conventional methods require. As for our close-up retrieval scheme, TH_S is set to 1.001, and TH_M is set to 4.

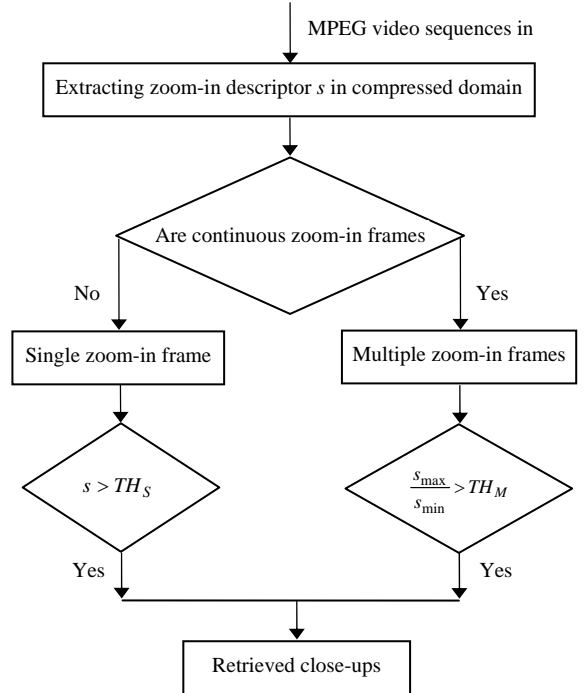


Figure 1 Overview of the proposed close-up retrieval.

IV. EXPERIMENTAL RESULTS AND EVALUATIONS

To evaluate the performance, our proposed automatic close-up retrieval scheme are tested on a database with different MPEG video clips, including documentary videos grabbed 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 30fps, 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:

$$Recall = \frac{D}{D + MD} \quad (7)$$

$$Precision = \frac{D}{D + FA} \quad (8)$$

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 I.

From Table I, 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 our 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 we obtain that the zoom factor s does not increase gradually, but fluctuates up and down, when the camera continues zooming in.

TABLE I. 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%



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

V. CONCLUSIONS

The main contributions of this paper are summarized as follows. Via exploiting fast camera motion estimation in compressed domain, we build close-up retrieval using zooming descriptor with two-level threshold. 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. The close-ups of persons and objects can both be retrieved correctly and effectively.

ACKNOWLEDGMENT

The authors would like to acknowledge the financial support under European IST FP6 Research Programme as funded for the Integrated Project: LIVE (Contract No. IST - 4 - 027312).

REFERENCES

- [1] J. Jiang, Y. Weng, B. Guo, and Y. Feng, "Robust-to-rotation texture descriptor for image retrieval in wavelets domain," *Journal of Electronic Imaging*, vol. 15, no. 1, pp. 013013, 1–14, Jan. – Mar. 2006.
- [2] J. Jiang, Y. Weng, and P. Li, "Dominant colour extraction in DCT domain," *Image & Vision Computing Journal*, Elsevier, vol. 24, no. 12, pp. 1269–1277, May 2006.
- [3] J. Jiang and Y. Weng, "Video extraction for fast content access to MPEG compressed videos," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 14, no. 5, pp. 595–605, May 2004.
- [4] J. Vendrig and M. Worring, "Systematic evaluation of logical story unit segmentation," *IEEE Trans. Multimedia*, vol. 4, no. 4, pp. 492–499, Dec. 2002.
- [5] H. W. Agius and M. C. Angelides, "Modeling content for semantic-level querying of multimedia," *Multimedia Tools Applicat.*, vol. 15, no. 1, pp. 5–37, Sep. 2001.
- [6] T. Athanasiadis, P. Mylonas, Y. Avrithis, and S. Kollias, "Semantic image segmentation and object labeling," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 17, no. 3, pp. 298–312, Mar. 2007.
- [7] D. Djordjevic and E. Izquierdo, "An object- and user-driven system for semantic-based image annotation and retrieval," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 17, no. 3, pp. 313–323, Mar. 2007.
- [8] D. Vallet, P. Castells, M. Fernandez, P. Mylonas, and Y. Avrithis, "Personalized content retrieval in context using ontological knowledge," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 17, no. 3, pp. 336–346, Mar. 2007.
- [9] J. W. Hsieh, S. L. Yu, and Y. S. Chen, "Motion-based video retrieval by trajectory matching," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 16, no. 3, pp. 396–409, Mar. 2006.
- [10] K. W. Sze, K. M. Lam, and G. Qiu, "A new key frame representation for video segment retrieval," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 15, no. 9, pp. 1148–1155, Sep. 2005.
- [11] F. Jing, M. Li, H. J. Zhang, and B. Zhang, "Relevance feedback in region-based image retrieval," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 14, no. 5, pp. 672–681, May 2004.
- [12] S. Antani, R. Kasturi, and R. Jain, "A survey on the use of pattern recognition methods for abstraction, indexing and retrieval of images and video," *Pattern Recognit.*, vol. 35, pp. 945–965, 2002.
- [13] Y. H. Moon, D. W. Ki, and J. H. Kim, "A hybrid motion-vector coding scheme based on an estimation of the locality for motion-vector difference," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 16, no. 4, pp. 492–506, Apr. 2006.
- [14] Y. H. Ho, C. W. Lin, J. F. Chen, and H. Y. M. Liao, "Fast coarse-to-fine video retrieval using shot-level spatio-temporal statistics," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 16, no. 5, pp. 642–648, May 2006.
- [15] D. Farin and P. H. N. De. With, "Enabling arbitrary rotational camera motion using multisprites with minimum coding cost," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 16, no. 4, pp. 492–506, Apr. 2006.
- [16] Y. Su, M. T. Sun, and V. Hsu, "Global motion estimation from coarsely sampled motion vector field and the applications," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 15, no. 2, pp. 232–242, Feb. 2005.
- [17] J. -C. Huang and W. -S. Hsieh, "Automatic feature-based global motion estimation in video sequences," *IEEE Trans. Consumer Electronics*, vol. 50, no. 3, pp. 911–915, Aug. 2004.
- [18] 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, pp. 133–146, Feb. 2000.
- [19] L. Liu, X. Ye, M. Yao, and S. Zhang, "A semantic description scheme of soccer video based on MPEG-7," *Lecture Notes in Computer Science*, vol. 3332, pp. 298–305, 2004.
- [20] D. W. Tjondronegoro, Y. P. Chen, and B. Pham, "Classification of self-consumable highlights for soccer video summaries," *Proc. IEEE International Conference on Multimedia and Expo*, vol. 1, pp. 27–30, Jun. 2004.
- [21] G. Jin, L. Tao, and G. Xu, "Hidden Markov model based events detection in soccer video," *Lecture Notes in Computer Science*, vol. 3211, pp. 605–612, 2004.