IEICE TRANS. INF. & SYST., VOL.E92–D, NO.1 JANUARY 2009

LETTER

Global Motion Representation of Video Shot Based on Vector Quantization Index Histogram

Fa-Xin YU[†], **Zhe-Ming LU**[††a)], **Zhen LI**[†††], *Nonmembers*, *and* **Hao LUO**[††††], *Member*

SUMMARY In this Letter, we propose a novel method of low-level global motion feature description based on Vector Quantization (VQ) index histograms of motion feature vectors (MFVVQIH) for the purpose of video shot retrieval. The contribution lies in three aspects: first, we use VQ to eliminate singular points in the motion feature vector space; second, we utilize the global motion feature vector index histogram of a video shot as the global motion signature; third, video shot retrieval based on index histograms instead of original motion feature vectors guarantees the low computation complexity, and thus assures a real-time video shot retrieval. Experimental results show that the proposed scheme has high accuracy and low computation complexity.

key words: *global motion representation, video shot, vector quantization index histogram*

1. Introduction

The rapid development of digital multimedia and network technologies in recent years has resulted in huge audiovisual data over the Internet or via CD-ROM. We need to ensure that the techniques for organizing audiovisual data stay in tune with the tremendous amounts of information. Compared with other media formats, video has become more and more popular due to its large capacity of storing information and intuitive visual effects. Thus, with the arrival of video on demand, there is an urgent need for effective video feature representation and retrieval methods. Various content-based video representation schemes have been raised in the literature, where colour-based, texture-based, motion-based [1]–[4] and spatio-temporal relationship-based [5] features are extracted from spatial-temporal, transformed and/or compressed domains to represent the video content. Moreover, video retrieval can be performed at different levels, such as shot-based [5], object-based [6], keyframe-based [7], hierarchical [8] and semantic [9] levels, while the proposed algorithm in this Letter falls in the category of shot-based methods.

Manuscript received July 2, 2008.
Manuscript revised August 24, 2008.

[†]The author is with Institute of Astronautic Electronic Engineering, School of Aeronautics and Astronautics, Zhejiang University, Hangzhou 310012, P. R. China.
[††]The author is with Media Processing and Communication Lab, School of Information Science and Technology, Sun Yat-Sen University, Guangzhou 510275, P. R. China.
[†††]The author is with School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore.
[††††]The author is with the Department of Automatic Test and Control, Harbin Institute of Technology, Harbin, China.
a) E-mail: luzhem@mail.sysu.edu.cn
DOI: 10.1587/transinf.E92.D.90

Motion characterization plays a crucial role in video indexing. An effective way of characterizing camera motion facilitates the video representation, indexing and retrieval tasks. Reference [1] proposes a hybrid motion-based video retrieval system through trajectory matching. This hybrid method includes a sketch-based scheme and a string-based one to analyze and index a trajectory with more syntactic meanings. Reference [2] utilizes motion vectors embedded in MPEG bitstreams to generate so-called "motion flows" for video retrieval. Reference [3] presents a motion trajectory-based compact indexing and efficient video retrieval mechanism, representing trajectories as temporal ordering of sub-trajectories. Reference [4] presents a global motion feature-based retrieval scheme that adopts the least square estimation to remove the singular and noisy points. In this Letter, we employ the VQ technique to reduce the noisy and singular points in the motion feature vector space while decreasing the computation complexity. VQ is an effective lossy coding technique and can be defined as a mapping from k dimension Euclidean space R^k to its limited subset (codebook) $C = \left\{\boldsymbol{y}_1, \boldsymbol{y}_2, \cdots, \boldsymbol{y}_N \middle| \boldsymbol{y}_i \in R^k\right\}$, that is, $Q : R^k \to C$. This mapping satisfies both $Q\left(\boldsymbol{x} \middle| \boldsymbol{x} \in R^k\right) = \boldsymbol{y}_p$ and $d\left(\boldsymbol{x}, \boldsymbol{y}_p\right) = \min\limits_{1 \le i \le N} d(\boldsymbol{x}, \boldsymbol{y}_i)$, where $\boldsymbol{x}$ and $\boldsymbol{y}_p$ are k dimensional vectors and $\boldsymbol{x}$ is then represented by the codeword index p. The distortion between the input vector $\boldsymbol{x}$ and each codeword $\boldsymbol{y}_i$ is often measured by the squared Euclidean distance $d^2(\boldsymbol{x}, \boldsymbol{y}_i) = \sum_{l=1}^{k} (x_l - y_{il})^2$. VQ index histogram technique has been already used in image retrieval application [10]; however, it has not been applied to video retrieval yet. In this Letter, we use the histogram of motion feature vector indices as the signature of each video shot for the purpose of video retrieval, where we measure the similarity between two histograms by L^2 norm.

2. Global Motion Feature Extraction

We define the global motion to be the motion vector of the background of a video shot. To obtain the motion feature vectors from a video shot, we should obtain motion vectors between adjacent frames of the video shot first. A motion vector is used to describe the displacement between the current macroblock and the best matched macroblock in the subsequent frame. In this Letter, we divide each frame into non-overlapping macroblocks of size 16×16, and then employ Three Step Search (TSS) [11] to obtain the motion vectors, since TTS is a simple and robust method for block

Copyright © 2009 The Institute of Electronics, Information and Communication Engineers

matching with low computation complexity.

For each frame in a video shot, we collect all motion vectors of macroblocks to calculate the 7-dimensional feature vector $\boldsymbol{x} = (x_1, x_2, \ldots x_7)$, which include the numbers of the macroblocks moving up, down, left and right, the number of the macroblocks remaining stable, the overall horizontal translation, and the overall vertical translation in each frame, respectively. In fact, this 7-D feature is used to represent the short time global motion between two adjacent frames. Because two shots may have different frame sizes, shots with similar content could have different values of $x_i, i = 1, 2, \ldots, 7$. Thus, we perform Gaussian normalization on the 7-D features:

$$y_i = [x_i - M_i(X)]/\sqrt{V_i(X)} \tag{1}$$

where X is the set of 7-D feature vectors, $\boldsymbol{x} \in X$ represents the original feature vector and x_i is the i-th component, $M_i(X)$ and $V_i(X)$ denotes the mean and variance values over all the i-th components of vectors in X respectively, and $\boldsymbol{y}$ represents the normalized feature vector, $i = 1, 2, \ldots, 7$.

3. Proposed Video Shot Retrieval Scheme

Based on global motion feature vectors obtained above, a simple video retrieval method may be to match the average global motion feature vector of a video shot between the query video shot and any video shot from the video database. However, although the block-matching method can obtain correct motion vectors in most macroblock positions, if there is no distinct texture in some macroblocks, we cannot avoid generating some random error motion vectors that are the noisy points in the motion vector space. Thus the falsely estimated feature vectors will affect the average feature vector severely, resulting in low retrieval precision.

In this Letter, we utilize the VQ technique to reduce the effect of the singular and noisy motion vectors to the motion feature vectors. We assume that the foreground objects and indistinct texture regions account for little proportion of a video shot, thus after performing vector quantization on the feature vector space, motion feature vectors corresponding to the statistics of the background motion vectors in a video shot will be in some clusters containing relatively more motion feature vectors. As a result, VQ index histogram of the global motion feature vectors of a video shot can be used as an effective signature. Here, in order to calculate the distance between VQ index histograms, we should define a distance function $d(H_1, H_2)$ of the VQ index histograms as follows:

$$d_H = \frac{1}{K}\sum_{k=1}^{K}(h_{1,k} - h_{2,k})^2 \tag{2}$$

where K is the number of the codewords in the codebook, namely, the number of bins in a histogram.

In this Letter, the codebook is generated from the training set containing all motion feature vectors obtained from all shots in the database based on the well-known LBG algorithm. To reduce the computation complexity in the codeword searching process, the Hadamard transform partial distortion search (HTPDS) [12] is adopted. This algorithm requires the vector dimension to be 2^n (zero padding is required if the dimension is not 2^n).

The advantages of our video retrieval algorithm may include: (1) The algorithm concentrates the feature data, so it can reduce the computation complexity and the affection of the singular points; (2) The algorithm is based on the statistical characteristics of the motion feature vectors including some semantic meanings, e.g., in sports video, each type of global motion corresponds to a certain type of event to some extent; (3) The retrieval method can be performed real-time because each video shot is represented by a histogram of a fixed small size.

4. Experimental Results and Discussions

Our experiments were performed on a database containing 127 video shots that were segmented by a specific shot detection software from the soccer video "The best 30 goals of Ronaldinho" (12:23 minutes and 30.0 fps) downloaded from the Internet. The repetitive shots and shots of extremely short time are removed and a shot in the database has 151.8 frames on average. To evaluate the proposed method, all the video shots were manually classified into 6 classes: 35 shots with leftward motion, 35 shots with rightward motion, 4 shots with upward motion, 5 shots with downward motion, 8 shots with no obvious motion and 40 not relevant shots including zooming in, zooming out and motion of large area foreground. Obviously, the classification is performed solely to enable the evaluation. In a pre-processing step, the codebook is generated from all the feature vectors of the 87 shots in the first 5 classes. Then, a signature was calculated offline for every shot and stored in the database.

To demonstrate the efficiency of the proposed signatures, we compare the proposed MFVVQIH method with the basic method of matching the average motion feature vector of each video shot (AMFV). To compare the performance more reasonably, every shot in the database is selected as the query video shot for each test. For every query shot, we perform the retrieval process based on the prestored signature to return the ranked 127 similar video shots. For each number of returned shots (from 1 to 127), we average the recall and precision value over all the 87 tests, as shown in Fig. 1. From the figure, we know that we can obtain much better recall and precision performance with the proposed MFVVQIH method than with the basic AMFV method. Even though the number of entries in the signature is set to be only 7, that is to say, the storage and computational costs are the same as AMFV, MFVVQIH still outperforms AMFV in precision and recall. Besides the superiority of the retrieval performances, the computation complexity is kept very low. For each video shot query task, only KN multiplications, $(2K-1)N$ additions and sorting of N similarity values are required, where K is the number of en-

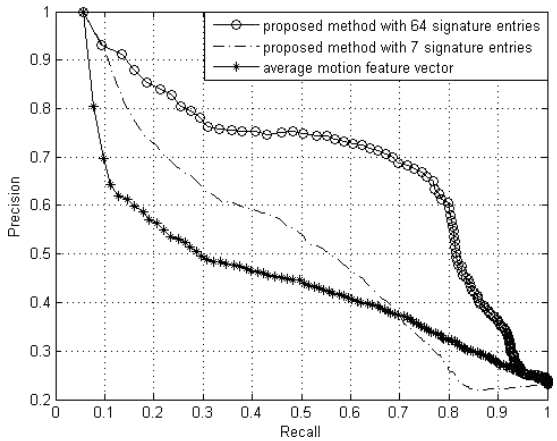

Fig. 1 The average recall and precision values.

tries in the signature and N is the total number of video shots in the database.

5. Conclusions

A video retrieval method based on matching the VQ index histogram calculated from short time global motion feature vectors of a shot is proposed. The experimental results demonstrate that the proposed algorithm can effectively represent the global motion of a video shot with different camera motion velocities and different shot lengths. The drawback of the proposed scheme may be that: 1) camera zooming that is also a global motion is not considered; 2) MFVVQIH proposed here can also be used as input of sophisticated classifiers to improve video retrieval performances. These are our future work.

References

[1] 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, 2006.

[2] C.W. Su, H.Y.M. Liao, H.R. Tyan, C.W. Lin, D.Y. Chen, and K.C. Fan, "Motion flow-based video retrieval," IEEE Trans. Multimed., vol.9, no.6, pp.1193–1201, 2007.

[3] F.I. Bashir, A.A. Khokhar, and D. Schonfeld, "Real-time motion trajectory-based indexing and retrieval of video sequences," IEEE Trans. Multimed., vol.9, no.1, pp.58–65, 2007.

[4] T.L. Yu and S.J. Zhang, "Video retrieval based on the global motion information," Acta Electronica Sinica, vol.29, no.Z1, pp.1794–1798, 2001.

[5] 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, 2006.

[6] H.J. Li and K.N. Nhan, "Automatic video segmentation and tracking for content-based applications," IEEE Commun. Mag., vol.45, no.1, pp.27–33, 2007.

[7] K.W. Sze, K.M. Lam, and G.P. Qiu, "A new key frame representation for video segment retrieval," IEEE Trans. Circuits Syst. Video Technol., vol.15, no.9, pp.1148–1155, 2005.

[8] X.Q. Zhu, A.K. Elmagarmid, X.Y. Xue, L.D. Wu, and A.C. Catlin, "InsightVideo: Toward hierarchical video content organization for efficient browsing, summarization and retrieval," IEEE Trans. Multimed., vol.7, no.4, pp.648–666, 2005.

[9] W.M. Hu, D. Xie, Z.Y. Fu, W.R. Zeng, and S. Maybank, "Semantic-based surveillance video retrieval," IEEE Trans. Image Process., vol.16, no.4, pp.1168–1181, 2007.

[10] Z.M. Lu and H. Burkhardt, "Colour image retrieval based on DCT-domain vector quantization index histograms," Electron. Lett., vol.41, p.29, Aug. 2005.

[11] T. Koga, K. Linuma, A. Hirano, Y. Lijima, and T. Ishiguro, "Motion compensated interframe coding for video conferencing," Proc. National Telecommunications Conference, pp.C9.6.1–C.9.6.5, New Orleans, LA, 1981.

[12] Z.M. Lu, J.S. Pan, and S.H. Sun, "Efficient codevector search algorithm based on Hadamard transform," Electron. Lett., vol.36, pp.1364–1365, 2000.