12-2002

Motion retrieval by temporal slices analysis

Chong-Wah NGO

Chong-wah NGO
Singapore Management University, cwngo@smu.edu.sg

Hong-Jiang ZHANG

Motion Retrieval by Temporal Slices Analysis

Chong-Wah Ngo
Department of Computer Science,
City University of Hong Kong,
cwngo@cs.cityu.edu.hk

Ting-Chuen Pong
Department of Computer Science
Hong Kong University of Science & Technology
tcpong@cs.ust.hk

Hong-Jiang Zhang
Microsoft Research Aisa
Beijing 100080, PRC
hjzhang@microsoft.com.

Abstract

In this paper, we investigate video shots retrieval based on the analysis of temporal slice images. Temporal slices are a set of $2D$ images extracted along the time dimension of image sequences. They encode rich set of motion clues for shot similarity measure. Because motion is depicted as texture orientation in temporal slices, we utilize various texture features such as tensor histogram, Gabor feature, and the statistical feature of co-occurrence matrix extracted directly from slices for motion description and retrieval. In this way, motion retrieval can be treated in a similar way as texture retrieval problem. Experimental results indicate that the features extracted from slices perform satisfactorily in the sport video domain and are, in general, superior to the histogram of MPEG motion vector.

1. Introduction

Video retrieval techniques, to date, are mostly extended directly or indirectly from image retrieval techniques. Examples include first selecting keyframes from shots and then extracting image features such as color and texture features from those keyframes for indexing and retrieval. The success from such extension, however, is limited since the spatio-temporal relationship among video frames is not fully exploited. Recently more works have been dedicated to address this problem [3, 5, 6, 16], more specifically, to utilize motion information for retrieval.

To date, motion features that have been used for retrieval include the motion trajectories of objects [5], principle components of MPEG motion vectors [16], and temporal texture [6]. In this paper, we propose new ways of computing and extracting temporal texture and demonstrate their retrieval effectiveness and efficiency particularly for sport videos. Temporal texture was primarily proposed by Polana & Nelson to describe the dynamic of temporal events [13]. As image texture, temporal texture can be modeled as co-occurrence matrix [3, 13], autoregressive model [17], wold decomposition [10] and Gibbs random field [6]. Except Fablet & Bouthemy who shown the effectiveness of temporal texture for video retrieval [6], this feature is mostly utilized for recognizing complex dynamic motion (e.g., rivers and crowds [3, 13, 17]) and detecting periodic motion (e.g., walking and swimming [10]).

For most approaches [6, 13], the input to temporal texture is optical flow or normal flow field. In other words, motion information need to be explicitly computed before the generation of temporal texture. Consequently, the effectiveness of the computed temporal feature is dependent on the reliability of input motion information. Unfortunately, motion information such as optical flow is not only computationally expensive but also noise sensitive. Our proposed methods, with contrary to these approaches, computes temporal texture by taking the gray-level information of temporal slices as input. Temporal slices encode rich set of motion clues as the oriented texture patterns that have been vividly exploited for video partitioning [14], motion characterization and segmentation [15]. In this paper, we further propose methods to extract and represent these texture patterns as tensor histogram, Gabor feature, and co-occurrence matrix for motion retrieval.

We focus our attention for sport videos since motion features are essential cues for characterizing various sport events. While other visual cues such as color can also be used, they are not as generally discriminative as motion cues [16]. Sport videos are usually captured by several fixed cameras that are mounted in the stand. These camera motion are mostly regular and driven by the pace of sport games or the events that are taken place on spot. When coupling with the object motion of a particular event in sport videos, a unique texture pattern can always be observed in temporal slices.

2. Patterns in Temporal Slices

Temporal slices are a set of $2D$ images in an image volume with one dimension in t, and the other in x or y, for instance. Previous works on the analysis of temporal slices

Proceedings of the 16 th International Conference on Pattern Recognition (ICPR'02)
1051-4651/02 $17.00 © 2002 IEEE

for computational vision tasks include visual motion model [1, 8, 9, 18] and epipolar plane image analysis [2].

Figure 1 shows the patterns of various activities in the horizontal ($x-t$ dimensions) and vertical ($y-t$ dimensions) slices. It is worthwhile to observe the following details:

- For diving, since the motion proceeds in vertical direction, the vertical slices depict camera tilting while the horizontal slices explore panoramic information. A full court advance in basketball videos, on the other hand, has the horizontal slices depict camera panning while the vertical slices explore panoramic information.
- The periodic motion in the boat-race shot is indicated in the horizontal slices.
- The camera motion which tracks a flying hammer in a parabolic-like direction is depicted in the slanted lines of vertical slices.

As seen in the examples, the diversities of texture patterns are encoded in both horizontal and vertical slices by different sport activities. Intuitively, these patterns can be extracted directly for motion retrieval. In this paper, unless being stated, all slices are used for feature extraction.

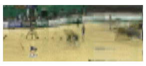
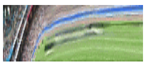
Activity	*Horizontal Slice*	*Vertical Slice*	*Video frame*
Dive			
Full court advance			
Boat−race (periodic)			
Hammer flying			

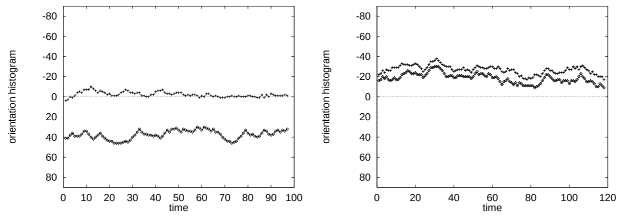
Figure 1. Patterns in both horizontal and vertical slices.

3. Feature Extraction

For computational and storage efficiency, all the features are extracted from the temporal slices that are obtained directly from the compressed video domain.

3.1. Tensor Histogram

A $2D$ tensor histogram $M(\phi, t)$, with one dimension as an $1D$ orientation histogram and the other dimension as time, can be employed to model the distribution of texture orientations in slices [15]. This distribution inherently reflects the motion trajectories in an image sequence, two examples are given in Figure 2. The trajectories in the figure are the histogram peaks tracked over time. In Figure 2(a) one trajectory indicates a non-stationary background, and the other indicates objects moving to the right; in Figure 2(b) two trajectories progress in a similar manner, they correspond to parallax motion (or camera panning).



(a) moving object (b) parallax panning

Figure 2. Motion trajectories in the tensor histograms.

For motion retrieval, a $1D$ tensor histogram $\mathcal{M}(k)$ is computed directly by

$$\mathcal{M}(k) = \frac{1}{n}\{\sum_{\phi}\sum_{t} M(\phi, t)\} \quad \forall_{\phi}\{\mathcal{Q}(\phi = k)\} \tag{1}$$

where $\mathcal{Q}(\phi)$ is a quantization function, and $k = \{1, 2, \ldots, 8\}$ represents a quantized level. The histogram is uniformly quantized into 8 bins with each bin has a range $\frac{\pi}{8}$. The computed motion features describe the motion intensity and direction of shots. In our experiment, the tensor histograms of both horizontal and vertical slices are used for feature computation. As a result, the total feature vector length is 16.

3.2. Gabor Feature

Gabor feature is frequently used for browsing and retrieval of texture images, and have been shown to give good results [11]. A Gabor filter $g(x, t)$ can be written as

$$G(x, t) = (\frac{1}{2\pi\sigma_x\sigma_t})\exp\{-\frac{1}{2}(\frac{x^2}{\sigma_x^2} + \frac{t^2}{\sigma_t^2})\}\exp\{2\pi jWx\} \tag{2}$$

where σ_x and σ_t are smoothing parameters, $j = \sqrt{-1}$, $W = \sqrt{u^2 + v^2}$ and (u, v) is the center of the desired frequency. A self-similar filter $G_{\theta\mathcal{S}}(x, t)$ can be obtained by the appropriate rotation θ and scaling $\mathcal{S}$ of $G(x, t)$ [4, 11]. The Gabor filtered image of a slice $\mathbf{H}$ is

$$\hat{\mathbf{H}}_{\theta\mathcal{S}} = \mathbf{H} * G_{\theta\mathcal{S}} \tag{3}$$

where $*$ is a convolution operator. A feature vector is constructed by using the mean $\mu_{\theta\mathcal{S}}$ and the standard deviation $\sigma_{\theta\mathcal{S}}$ of all $\hat{\mathbf{H}}_{\theta\mathcal{S}}$ as components. In the experiment,

Proceedings of the 16 th International Conference on Pattern Recognition (ICPR'02)
1051-4651/02 $17.00 © 2002 IEEE

$\theta = 6$ and $\mathcal{S} = 2$. The resulting feature vector has length $6 \times 2 \times 2 \times 2 = 48$ in the following form

$$[\underbrace{\mu_{00}, \sigma_{00}, \mu_{01}, \sigma_{01}, \ldots \mu_{51}, \sigma_{51}}_{\text{for horizontal slices}}, \underbrace{\mu_{00}, \sigma_{00}, \mu_{01}, \sigma_{01}, \ldots \mu_{51}, \sigma_{51}}_{\text{for vertical slices}}]$$

3.3. Co-occurrence Matrix

Gray level co-occurrence matrix is frequently utilized to describe image texture [7]. The co-occurrence matrix of a slice can be represented by $P(i, j; d, \theta)$. It specifies the frequencies of two neighboring pixels separated by distance d at orientation θ in the temporal slices, one with gray level i and the other with gray level j.

In our experiment, $d = \{1, 2, 3, 4, 5\}$ and $\theta = \{-45°, 0°, 45°\}$. The co-occurrence matrices of horizontal and vertical slices are computed, summed and normalized separately, hence, there are thirty matrices used to model the spatial relationships of slices. The smoothness $Sm(d, \theta)$ and contrast features $Con(d, \theta)$ are then computed from these matrices by

$$Sm(d,\theta) = \sum_i \sum_j P^2(i,j;d,\theta) \tag{4}$$

$$Con(d,\theta) = \sum_i \sum_j (i-j)^2 P(i,j;d,\theta) \tag{5}$$

The resulting feature vector has length $15 \times 2 \times 2 = 60$.

4. Distance Measure

Let $\mathcal{F}$ and $\mathcal{F}'$ represent the feature vectors of two shots. Each vector is composed of n components. For tensor histogram, we use L_1 norm to measure the feature distance by

$$D(\mathcal{F}, \mathcal{F}') = \frac{1}{Z(\mathcal{F}, \mathcal{F}')} \{\sum_{i=1}^{n} |\mathcal{F}(i) - \mathcal{F}'(i)|\} \tag{6}$$

where $Z(\mathcal{F}, \mathcal{F}') = \sum_{i=1}^{n} \mathcal{F}(i) + \sum_{i=1}^{n} \mathcal{F}'(i)$ is a normalizing function. For Gabor and co-occurrence matrix, however, the range of different feature components can significantly vary. Hence, the distance measure

$$D(\mathcal{F}, \mathcal{F}') = \sum_{i=1}^{n} |\frac{\mathcal{F}(i) - \mathcal{F}'(i)}{\alpha(i)}| \tag{7}$$

is used where $\alpha(i)$ is the standard deviation of the i^{th} feature component over the entire database.

5. Experiments

We conduct experiments on basketball ($18,000$ frames with 76 shots), soccer ($100,000$ frames with 404 shots), and TV sport video ($37,000$ frames with 180 shots) databases. We adopt average normalized modified retrieval rank (ANMRR) developed by MPEG Video Group for performance evaluation [12]. The values of ANMRR range between $[0, 1]$. A low value of ANMRR reveals high retrieval rate and good ranking capability.

Besides comparing the performance among the features extracted from temporal slices, we contrast their retrieval accuracy with MPEG motion vectors (MPEG MV). Only motion vectors from P-frames are used and they are represented by a histogram that is composed of eight bins. Each bin corresponds to one of the eight neighborhood directions in the discrete space. To take motion intensity into account, each bin contains the total length, instead of frequency, of the motion vectors having same direction. Similar to tensor histogram, L_1 norm is used for distance measure.

5.1. Retrieval Accuracy

In the basketball database, the shots are categorized into full-court-advance (FCA), close-up shots of player, penalty shots, shooting shots, and audience scene. The close-up of players are further classified into players moving to the left, players moving to the right, and players with no motion. In this database, twenty queries that are manually checked to have good answers are picked for testing. The categorization of the soccer database is similar to the basketball database. The shots in this database are classified into bird views, medium shots, close-up shots of players, shooting scenes and audience scenes. A total of fifty three queries from these categories are selected for testing. The sport database contains a diversity of sport games including diving, golf and race. We categorize the shots according to the type of sport games. Some games are further categorized into bird view or close-up shots. The close-up shots are also categorized into tracking or stationary shots. A total of 124 shots from these categories are selected for testing. These categorizations are based not only on the semantic events of various sport videos, but also mainly based on the motion content of shots. The retrieval performance is given in Table 1. Experimental results indicate that tensor histogram outperforms other approaches in the three databases.

5.2. Speed Efficiency

Table 2 compares the performance efficiency in term of the feature vector length and the feature extraction time (second per image frame). When all horizontal and vertical slices are used for feature extraction, approaches based on temporal slices are not as efficient as MPEG MV. Nevertheless, the extraction time can be significantly speed up if only a subset of slices is processed. Table 3 compares the speed of tensor histogram when the number of slices is

Table 1. Retrieval accuracy

Approach	ANMRR Basketball	Soccer	Sport
Tensor histogram	0.399*	0.393*	0.456*
Gabor	0.431	0.430	0.481
Co-occurrence	0.543	0.492	0.577
MPEG MV	0.498	0.590	0.557

*The mark * indicates the best performance.*

Table 2. Speed efficiency

Approach	Feature Vector Length	Extraction time (sec)
Tensor histogram	16	0.072
Gabor	48	0.791
Co-occurrence	60	0.130
MPEG MV	8	0.017*

*The mark * indicates the best performance.*

uniformly sampled and recursively reduced by half. The reduction of slices not only increases the efficiency but even improve the retrieval accuracy. The improvement may due to the elimination of some slices that contains image noise and poor texture information during feature extraction. It should be noted that, for all the tested database, when only two slices are used, the processing speed is comparative to motion vector histogram while the retrieval accuracy is still superior to all other features.

6. Conclusion

Three new temporal texture features based on the analysis of temporal slices have been presented and applied to motion retrieval. Among the proposed features, tensor histogram is empirically found to be superior to other features in term of retrieval accuracy and speed efficiency. Furthermore, the feature computational time of tensor histogram can be as fast as the histogram of MPEG motion vector by reducing the number of slices being processed without significantly degrading the retrieval performance.

Table 3. Performance of tensor histogram

Number of slices	Extract time (s)	ANMRR Basketball	Soccer	Sport
All (74 slices)	0.072	0.399	0.393	0.456
Half	0.042	0.399	0.392*	0.453
One-third	0.033	0.397	0.392*	0.450
One-quarter	0.028	0.392*	0.394	0.448*
Two	0.015*	0.416	0.416	0.471

*The mark * indicates the best performance.*

Acknowledgments

This work is supported in part by RGC Grants HKUST661/95E, HKUST6072/97E, SSRI99/00.EG11 and DAG01/02.EG16.

References

[1] E. H. Adelson & J. Bergen, “Spatiotemporal Energy Models for the Perception of Motion”, *Journal of Optical Society of America*, 2(2):284-299, Feb 1985.

[2] R. C. Bolles & H. H. Baker, “Epipolar Plane Image Analysis: An Approach to determining Structure from Motion”, *Int. J. of Computer Vision*, 1(1):7-55, 1987.

[3] P. Bouthemy & R. Fablet, “Motion Characterization from Temporal Cooccurrences of Local Motion-based Measures for Video Indexing”, *Int. Conf. on Pattern Recognition*, 1998.

[4] A. C. Bovic, M. Clark & E. S. Geisler, “Multichannel Texture Analysis Using Localized Spatial Filters”, *IEEE Trans. on PAMI*, 12(1):55-73, Jan 1990.

[5] S. F. Chang *et. al.*, “A Fully Automatic Content-based Video Search Engine Supporting Multi-object Spatio-temporal queries”, *IEEE Trans. on Circuits and Systems for Video Technology*, 8(5):602-615, 1998.

[6] R. Fablet *et. al.*, “Statistical Motion-based Video Indexing and Retrieval”, *Int. Conf. on Content-based Multimedia Info. Access*, pp. 602-619, 2000.

[7] R. M. Haralick *et. al.*, “Textural Features for Image Classification”, *IEEE Trans. on Systems, Man, and Cybernetics*, 3(6):610-621, Nov 1973.

[8] D. J. Heeger, “Model for the Extraction of Image Flow”, *Journal of Optical Society of America* 4(8):1455-1471, Aug 1987.

[9] B. Jähne, “Spatio-temporal Image Processing: Theory and Scientific Applications”, *Springer Verlag*, 1991.

[10] F. Liu & R.W. Picard, “Finding Periodicity in Space and Time”, *ICCV*, pp. 376-383, 1998.

[11] B. S. Manjunath & W.Y. Ma, “Texture Features for Browsing and Retrieval of Image Data”, *IEEE Trans. on PAMI*, 18(8):837-842, Aug, 1996.

[12] MPEG Video Group, “Description of Core Experiments for MPEG-7 colour/texture descriptors”, *ISO/MPEGJTC1/SC29/WG11 MPEG98/M2819*, 1999.

[13] R. Nelson and R. Polana, “Qualitative Recognition of Motion using Temporal Texture”, *CVGIP: Image Understanding*, 56(1):78-99, July, 1992.

[14] C. W. Ngo *et. al.*, “Video Partitioning through Temporal Slices Analysis”, *IEEE Trans. on Circuits and Systems for Video Technology*, 11(8):941-953, Aug 2001.

[15] C. W. Ngo *et. al.*, “Motion Characterization by Temporal Slice Analysis”, *Computer Vision and Pattern Recognition*, vol. 2, pp. 768-773, 2000.

[16] E. Sahouria & A. Zakhor, “Content Analysis of Video Using Principle Components”, *IEEE Trans. on CSVT*, 9(8):1290-1298, Dec 1999.

[17] M. O. Szummer, *Temporal Texture Modeling*, MIT, 1995.

[18] A. B. Watson and A. J. Ahumada, “Model of Human Visual-Motion Sensing”, *Journal of of Optical Society of America* 2(2):322-341, Feb 1985.