MOTION REGION-BASED TRAJECTORY ANALYSIS AND RE-RANKING FOR VIDEO RETRIEVAL

Bailan Feng[1, 2], Juan Cao[1], Shouxun Lin[1], Yongdong Zhang[1], Kun Tao[1, 2]
[1]Laboratory of Advanced Computing Research, Institute of Computing Technology,
China Academy of Sciences, Beijing, China
[2]Graduate University of the Chinese Academy of Sciences, Beijing, China
{fengbailan, caojuan, sxlin, zhyd, ktao}@ict.ac.cn

ABSTRACT

Event-related query is playing a more and more important role in video retrieval. However, it is still a challenge to the existing video retrieval engines for lacking the effective motion analysis. In this paper, we propose a novel re-ranking scheme for video retrieval based on motion region trajectory analysis. By focusing on the changes of the primary moving regions, we construct an intuitive motion region-based trajectory descriptor (MRTD) to depict the shot activities. In the re-ranking phase, the proposed approach takes the MRTD as a motion cue and re-ranks the baseline results by motion-related query selection and MRTD-based weighting. We evaluate our method in TRECVID2007 and 2008 datasets, and observe consistent improvement over all the baselines, leading to a greatest performance gain of 42.9%, and an average gain of 17%. The experiments also show that the motion descriptor of MRTD is fruitful for a variety of features.

1. INTRODUCTION

With the rapid increase in the volume of digital videos, how to effectively and efficiently index and retrieval has become more and more urgent. Compared to the static image, video is a consecutive frame sequence and contains the rich motion information, which can cover spatial and temporal characteristics simultaneously. Some traditional motion-based video retrieval systems have been proposed over the past decade [1, 2] [3, 4]. These systems first utilize statistics to analyze the distribution of "motion points" and represent it with vectors (named as statistics-based motion descriptors), then apply the extracted feature vectors to video retrieval by means of certain distance measure [1, 2] or classifier [3, 4]. The advantage of these statistics-based approaches is the compact vector representation and the fast computation. However, one problem with the above approaches is the lack of capacity with respect to the characterization of relationship between temporal and spatial information.

An alternative use of motion information for video retrieval [5, 6] [7] is to extract the motion trajectories of target from motion vectors (termed target-based motion descriptors). There are two types of targets: the foreground object and the pixel-like point. For example, VideoQ [5] proposed grouping regions that are similar in low visual features to form a video object, and adopted the query-by-sketch (QBS) scheme to retrieve videos. Babu *et al.* [6] proposed a system to extract object-based and global features using motion vector information for video retrieval. The object segmentation was done by applying EM algorithm and the number of objects was determined by the K-means clustering procedure. Instead of object segmentation, Su *et al*. [7] used the local motion vectors directly across consecutive video frames to form motion flows, and triggered the retrieval phase by QBS scheme. The advantage of these target-based methods is the good observability, i.e. it can integrate temporal-spatial information well. However, its drawback is that, on one hand, the segmentation of video objects is still an open question for object-based trajectory method, on the other hand, for point-based trajectory method, the frame by frame tracking process is over precise and produces many disorderly motion curves which are less useful for retrieval. Otherwise, the matching strategy in video retrieval phase is rather unreasonable. It is not convenient for the user to provide a sketch of motion trajectory for retrieval, especially without having any semantic information.

To address the problems above, we propose a new semantic-level motion region-based trajectory descriptor (MRTD) and a novel layered re-ranking scheme to apply the MRTD into video retrieval. As shown in Figure 1, after shot segmentation, we first hierarchically extract two-level motion features, including a 17-D frame-level motion feature and a 38-D shot-level motion feature from compressed video units. Then a semantic-level MRTD is represented by quantizing the primary moving region's trajectory extracted from the above two-level motion features. Finally, we apply the proposed MRTD in a MRTD-based re-ranking scheme for video retrieval.

The major contributions of this paper can be summarized as follows:

1. We propose a novel motion region-based trajectory descriptor (MRTD). Instead of considering the object segmentation and the disorderly motion curves integration, this MRTD takes the movement tendency of primary moving regions as a cue and contains certain semantic information. Otherwise, the vector representation of MRTD derived from quantization process is convenient for the usage of video retrieval.

2. We propose a MRTD-based video re-ranking scheme, which consists of adaptive motion-related query selection and MRTD-based weighting. Experimental results show that our scheme is fruitful for all the features and achieves an important gain in MAP and infAP.

The remainder of this paper is organized as follows. Section 2 presents two parts: first part describes the extraction and representation of the MRTD, and the second part is the MRTD-based re-ranking scheme in video retrieval. Section 3 shows the experiments over TRECVID benchmark, followed by conclusions in Section 4.

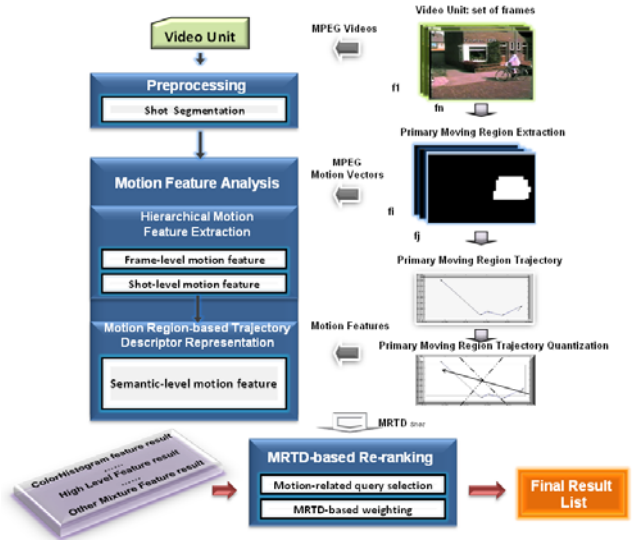

Figure 1: Framework for our motion region trajectory analysis and re-ranking scheme for video retrieval

2. MRTD-BASED RE-RANKING FOR VIDEO RETRIEVAL

2.1. Motion Feature Analysis

2.1.1 Hierarchical Motion Features Extraction

Motion information is both spatial and temporal simultaneously. To preserve these excellent characteristics, we generate two-level hierarchies of motion features, including a frame-level motion feature and a shot-level motion feature.

a) Frame-level motion feature (spatial characteristics)

Firstly, we extract the motion vectors from the compressed MPEG streams and filter the noisy macroblocks based on the three criteria in [8]. Then we estimate global camera motion with the simplified discriminative models mentioned in our previous work [8]. Finally, by subtracting the camera motion vector from the macroblock's motion vector, we get a 17-D frame-level motion feature V [8]. It consists of the scaled values of the center of gravity of the primary region (x_c, y_c), the primary region's area m_{00}, and rotation-, translation- and scale- invariant moments of the primary region u_{pq}.

b) Shot-level motion feature (temporal characteristics)

For preservation of temporal characteristics, we explore the motion vectors in a shot sequence and form a 38-D shot-level motion feature. We take the average of every item in the 17-D frame-level features denoted as $\bar{V}$ and the absolute values of standard deviation denoted as $\triangle\bar{V}$ in a shot sequence. Four other statistics, p_{null}, p_{still}, $p_{irregular}$ and $\bar{r}$ are added, which represent the percentage of no object involved frames, the percentage of still frames, the percentage of irregular frames and the average of all existing motion magnitudes respectively.

2.1.2 Motion Region-Based Trajectory Descriptor (MRTD) Representation

The motion descriptor of MRTD can be regarded as an unit of computation which is further extracted from continuous feature vectors. Different from continuous feature vectors, MRTD is discrete and more intuitive, containing certain semantic information. In this section we represent the constructing process of our MRTD as follows:

Step 1: Noisy frame filtration. From repeated observation of the contents of every frame, frames which present many disorderly and small areas or stationary pictures are usually the ones causing interruptions in the target tracking, besides, due to the unreliable shot segmentation, the marginal frames may bring into irrelevant content which disturbs the track of the primary moving region in a video unit. Therefore, we need to remove these noisy frames before "linking" trajectories. Given an original shot $Shot_i = \{Frame_1^i, Frame_2^i, \cdots Frame_n^i\}$, and $Frame_j^i = <x_j^i, y_j^i, Area_j^i>$ is one frame of the shot, where x, y and $Area$ are the scaled values of the center of gravity of the primary region and the scaled values of the primary region's area respectively, and i and j are the index of shots and frames respectively. We calculate it as follows:

$$Shot_i^{Key} = \{Frame_j^i \mid \alpha_{M\arg inFilter} < j < n - \alpha_{M\arg inFilter}\} \quad (1)$$

$$Shot_i^{Key} = \{Frame_j^i \mid Area_j^i \geq \beta_{AreaFilter}\} \quad (2)$$

Where $\alpha_{MarginFilter}$ denotes the number of marginally noisy frames, and $\beta_{AreaFilter}$ denotes certain values of the moving region's area. In our method, we empirically set $\alpha_{MarginFilter} = 2$, and $\beta_{AreaFilter}$ takes the average of m_{00} in a shot sequence derived from the 38-D shot-level feature automatically.

Step 2: Primary moving region trajectory representation. After eliminating the noisy frames, we link the centroids of the primary regions of all remainder frames $Shot_i^{Key}$ to form a trajectory of primary moving regions, denoted as

$$Trajectory_i = \{<x_j^i, y_j^i> \mid x_j^i, y_j^i \in Frame_j^i \,\&\, Frame_j^i \in Shot_i^{Key}\} \quad (3)$$

Step 3: Trajectory quantization. The necessity of the trajectory quantization is that on one hand, it is convenient for the similarity matching of trajectory with vector-like representation, especially suitable for more practical query-by-example (QBE) retrieval scheme. On the other hand, it can simultaneously alleviate the error of trajectory extraction to some extent. Specifically, the trajectory quantization is completed using the following operations,

$$\delta_i^{LeastSquaresFit} = \frac{\left|Shot_i^{Key}\right| * \sum(x_j^i * y_j^i) - \sum x_j^i * \sum y_j^i}{\left|Shot_i^{Key}\right| * \sum x_j^{i\,2} - \sum x_j^i * \sum x_j^i} \quad (4)$$

where $|Shot_i^{Key}|$ denotes the frame number of $Shot_i^{Key}$. Then values of the MRTD are determined by the sign function (5)

$$\text{sign}(Shot_i) = \begin{cases} Still, \left|Shot_i^{Key}\right| \leq 1 \\ HorizontalMotion, (-1 \leq \delta_i^{LeastSquaresFit} \leq 1) \,\&\, \left|Shot_i^{Key}\right| > 1 \\ VerticalMotion, (\delta_i^{LeastSquaresFit} > 1 \,\|\, \delta_i^{LeastSquaresFit} < -1) \,\&\, \left|Shot_i^{Key}\right| > 1 \end{cases} \quad (5)$$

2.2. MRTD-Based Re-Ranking

2.2.1 Motion-Related Query Selection

Motion information impacts different queries in different ways. It is important for query of "a person walking or riding a bicycle" while is useless for that of "a street market scene". Therefore, we propose a motion-related query selection mechanism that can remain adaptive to the characteristics of the motion for each query. We obtain the query type by calculating the proportion of each MRTD's value and taking out the portion of the value which satisfies a certain threshold with respect to type. The threshold is empirically set to meet the requirement that the value of maximum

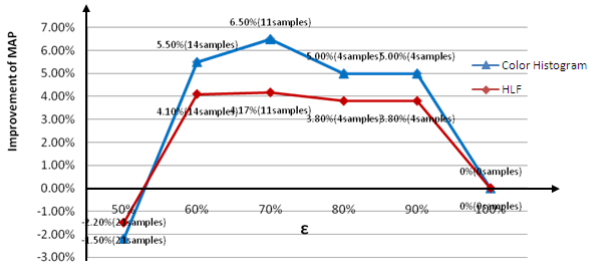

Figure 2: Relationship between ε and improvement of MAP of two typical features

proportion is taken and the number of its corresponding type should be empirically more than ε of total in query examples simultaneously.

The parameter selection of ε is basically an ill-posed problem and we believe the best way for deriving an appropriate ε value is through a large number of experiments. Here we use two features with wide differences to conduct an experiment for the optimal value of ε. Fig. 2 shows two different curves. These two curves indicate the improvement of MAP derived by using different sample proportions corresponding to Color Histogram and HLF. According to the figure, we can see that the highest improvements of MAP both appear when ε=0.7, therefore we suggest using ε=0.7 for all subsequent experiments. Subsequently, if no such threshold is satisfied, we denote the query type as "Unknown" and don't apply motion information to it.

2.2.2 MRTD-Based Weighting

The process of constructing MRTD has integrated spatial and temporal characteristics by incorporating hierarchical motion features, and the MRTD itself explicitly reflects directional information. During the MRTD-based weighting phase, motion intensity information is added as a coefficient to regulate initial scores, and is calculated as follows.

$$Score' = Score + MotionScore \times Coefficient \quad (6)$$

Where *MotionScore* measures the similarity of MRTD between queries and shots in the database and *Coefficient* is determined by $1-(\rho - min(\rho))/(max(\rho)-min(\rho))$, where $\rho = P_{null}+P_{still}$ is extracted from the shot-level motion feature. Sorting the updated scores and we could get the final re-ranking list.

3. EXPERIMENTS

We conduct the experiments on automatic search tasks using the TRECVID 07-08 datasets [9]. TRECVID provides 24 and 48 queries for 2007 and 2008 tasks respectively, and each query includes 3 to 11 video samples.

3.1. Motion-Related Query Selection

We extract each video sample's trajectory of primary moving region by using the method introduced in Sec.2.1, and quantize the trajectory to MRTD. Using the motion-related query selection mechanism, we automatically select 11 and 21 motion-related queries from the 2007 and 2008 query sets respectively, accounting for 45.8% and 43.8% in total. The selected queries from the 2007 query set and their corresponding MRTD values are shown in Table 1.

In Table 1, queries such as "a person walking or riding a bicycle", "a street protest or parade" and "a train in motion" have the attribute of horizontal motion. Likewise, queries such as "a street market scene", "3 or more people sitting at a table" and "one or more people playing musical instruments" have the attribute of still cues. These automatic analysis results are in accord with the subjective judgments. However, subjects like "a bridge" and "a door being opened" are judged as "Horizontal Motion", which seems to be inconsistent with common intuition. By analyzing the dataset, we find that there are high proportions of objects with horizontal motion in the samples of the two queries. For example, there are ships moving across the bridge, and persons moving across the opened door. More important is that, this inconsistentness has the statistical stability, and does not hurt the retrieval performance.

Table 1: Details of selected queries by motion-related query selection from TRECVID 2007 Search Task

Query ID	Concent	MRTD value
Topic0198	a door being opened	Horizontal Motion
Topic0199	a person walking or riding a bicycle	Horizontal Motion
Topic0203	a street market scene	Still
Topic0204	a street protest or parade	Horizontal Motion
Topic0205	a train in motion	Horizontal Motion
Topic0209	3 or more people sitting at a table	Still
Topic0210	one or more people walking with one or more dogs	Horizontal Motion
Topic0211	sheep or goats	Still
Topic0216	a bridge	Horizontal Motion
Topic0217	a road taken from a moving vehicle through the front windshield	Horizontal Motion
Topic0218	one or more people playing musical instruments	Still

3.2. Motion Fusion and Re-Ranking

To demonstrate the effectiveness of our proposed video re-ranking scheme based on motion region trajectory analysis, we extract the following static features or visual descriptors from each keyframe in each shot: Color Histogram (CH), Color Correlogram (CC), Edge Histogram (EH), Texture Cooccurence (TC), Texture Wavelet (TW), SIFT and SIFT_LDA [10].

Multi-bag SVM [11] is used to produce the initial results based on these features. Then we refine the baselines through the proposed MRTD-based re-ranking scheme. The performance is measured by the widely used Average Precision (AP) [9], and the results are shown in Table 2 and Table 3.

From Table 2, we find that the proposed re-ranking scheme based on MRTD outperforms initial benchmarks in all commonly used low-level features, and the improvements are obvious, with gain about 10%-22% and 6%-43% in different datasets respectively. Although the absolute values of above features are low, the improvement of our method is objectively remarkable, and we also find the robustness of our method to be suitable for various low-level features.

In addition, we evaluate the impact of MRTD and re-ranking scheme on some mixed features and high-level concept scores (HLF). The results are shown in Table 3.

From Table 3, we see once again the robustness of our feature and method for more complex features. Especially, the 3rd row of

Table 2: Evaluation for fusion of MRTD and various low-level features on TRECVID07-08 datasets

Visual Features	TRECVID2007			TRECVID2008		
	Queries-selected MAP /Queries-all MAP	MAP with MRTD Merged over selected/all	Improvement over selected/all	Queries-selected infAP /Queries-all infAP	infAP with MRTD Merged over selected/all	Improvement over selected/all
CH (166D)	0.0030/0.0059	0.0037/0.0063	**21.30%**/6.50%	0.0031/0.0039	0.0040/0.0043	**29.03%**/10.26%
EH (320D)	0.0107/0.0125	0.0123/0.0132	**13.90%**/5.60%	0.0036/0.0051	0.0038/0.0052	**5.56%**/1.96%
CC (166D)	0.0049/0.0081	0.0058/0.0086	**19.07%**/6.17%	0.0073/0.0050	0.0085/0.0054	**16.44%**/8.00%
TC (96D)	0.0088/0.0078	0.0098/0.0082	**9.94%**/5.13%	0.0014/0.0027	0.0020/0.0029	**42.86%**/7.41%
TW (108D)	0.0018/0.0015	0.0023/0.0017	**22.39%**/13.33%	0.0010/0.0008	0.0012/0.0009	**20.00%**/12.50%
SIFT(828D)	0.0070/0.0095	0.0081/0.0102	**13.74%**/7.37%			
SIFT_LDA(50D)	0.0094/0.0146	0.0115/0.0160	**22.34%**/9.59%	0.0069/0.0098	0.0080/0.0109	**15.94%**/11.22%

Table 3: Evaluation for fusion of MRTD and other complex features on TRECVID07-08 datasets

Mid- & Mixed-Features	TRECVID2007			TRECVID2008		
	Queries-selected MAP /Queries-all MAP	MAP with MRTD Merged over selected/all	Improvement over selected/all	Queries-selected infAP /Queries-all infAP	infAP with MRTD Merged over selected/all	Improvement over selected/all
CH+EH (486D)	0.0114/0.0136	0.0125/0.0142	**10.40%**/4.41%	0.0122/0.0151	0.0149/0.0162	**22.13%**/7.28%
HLF (374D)	0.0346/0.0585	0.0400/**0.0609**	**15.30%**/4.17%	0.0193/0.0279	0.0222/**0.0306**	**15.0%**/9.67%
Best Run	0.0357/0.0671	0.0404/**0.0692**	**13.14%**/3.13%	0.0361/0.0629	0.0403/**0.0669**	**11.63%**/6.36%

left side is our best run (0.0671) in the TRECVID07 search task, which ranked second among all of the 81 participant submissions. Our method improves it to 0.0692 again successfully. The right side of Table 3 shows the performance of our method in the TRECVID08 search task. After an improvement of 11.63%, we get our best run (0.0669).

Considering the experimental results above, we can conclude that our proposed motion descriptor of MRTD and MRTD-based re-ranking scheme is effective and reliable.

4. CONCLUSION

In this paper, we have presented a novel motion region-based trajectory descriptor (MRTD) which could depict the movement tendency of primary moving region in videos. Unlike conventional object-based and point-based trajectory construction methods, it is more operable and robust for video retrieval. Subsequently, an effective MRTD-based re-ranking scheme is investigated. Our ideas have been successfully tested on the TRECVID07-08 datasets, and experiments show that our method can contribute an effective supplement to video retrieval, offering on average 17% (best 42.9%) improvement over various baseline results. Our future work will focus on seeking more effective motion descriptors, as well as incrementally investigating the usage of these motion descriptors in video retrieval.

5. ACKNOWLEDGEMENT

This work is supported by National Basic Research Program of China (973 Program, 2007CB311100), National High Technology and Research Development Program of China (863 Program, 2007AA01Z416), National Nature Science Foundation of China (60873165, 60802028), Beijing New Star Project on Science & Technology (2007B071), and the Co-building Program of Beijing Municipal Education Commission.

6. REFERENCES

[1] R. Fablet, and P. Bouthemy, and P. Perez, "Nonparametric motion characterization using causal probabilistic models for video indexing and retrieval", *IEEE Trans. Image Process.*, vol. 11, no. 4，pp. 393-407, Apt. 2002.

[2] Y. –F. Ma and H. –J. Zhang, "Motion texture: A new motion based video representation", in *Proc. 16th Int. Conf. Pattern Recognation*, vol.2, pp.548-551, Aug. 11-15, 2002.

[3] Feng Wang, Yu-Gang Jiang, Chong-Wah Ngo, "Event-based Semantic Detection Using Motion Relativity and Visual Relatedness", *IEEE Int. Conf. on Multimedia(MM)*, 2008.

[4] Alexander Haubold, Milind Naphade, "Classification of Video Events using 4-dimensional time-compressed Motion Features", *ACM Int. Conf. on Image and Video Retrieval(CIVR)*, 2007.

[5] S. F. Chang, W. Chen, H. J. Meng, H. Sundaram, and D. Zhong, "A fully automated content-based video search engine supporting spatiotemporal queries," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 8, no.5, pp. 602-615, Sep.1998.

[6] R. Venkatesh Babu and K.R. Ramakrishman, "Compressed domain video retrieval using object and global motion descriptors," *Multimedia Tools and Applic.*, vol.32, no.1, pp.93-113, Jan.2007.

[7] Chih-Wen Su, Hong-Yuan Mark Liao, Hsiao-Rong Tyan, Chia-Wen Lin, Duan-Yu Chen and Kuo-Chin Fan, "Motion Flow-Based Video Retrieval", *IEEE Transactions on Multimedia*, Aug.2007.

[8] Tao Kun, Wu Si, Lin Shouxun, Zhang Yongdong, "Research on Panorama Composition Technique of Sports Video", *Journal of computer-aided design and computer graphics*, Nov. 2005, 17(11).

[9] Smeaton, A. F., Over, P., and Kraaij, W., " Evaluation campaigns and TRECVid", In *Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval*, ACM Press, pp.321-330, 2006.

[10] J. Cao, T. Xia, J. Li, Y. Zhang and S. Tang, "A Density-Based Method for Adaptive LDA Model Selection ", *Neurocomputing*, vol.72, pp.1775-1781, 2008.

[11] J. Tesic, A. Natsev, L. Xie, and J.R. Smith, "Data modeling strategies for imbalanced learning in visual search", *ACM International Conference on Multimedia and Expo(ICME)*, 2007.