# Motion Feature Extraction Scheme for Content-based Video Retrieval

Chuan Wu*, Yuwen He, Li Zhao, Yuzhuo Zhong

Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China

## ABSTRACT

This paper proposes the extraction scheme of global motion and object trajectory in a video shot for content-based video retrieval. Motion is the key feature representing temporal information of videos. And it is more objective and consistent compared to other features such as color, texture, etc. Efficient motion feature extraction is an important step for content-based video retrieval. Some approaches have been taken to extract camera motion and motion activity in video sequences. When dealing with the problem of object tracking, algorithms are always proposed on the basis of known object region in the frames. In this paper, a whole picture of the motion information in the video shot has been achieved through analyzing motion of background and foreground respectively and automatically. 6-parameter affine model is utilized as the motion model of background motion, and a fast and robust global motion estimation algorithm is developed to estimate the parameters of the motion model. The object region is obtained by means of global motion compensation between two consecutive frames. Then the center of object region is calculated and tracked to get the object motion trajectory in the video sequence. Global motion and object trajectory are described with MPEG-7 parametric motion and motion trajectory descriptors and valid similar measures are defined for the two descriptors. Experimental results indicate that our proposed scheme is reliable and efficient.

**Keywords:** Video retrieval, motion feature extraction, MPEG-7, motion estimation, object tracking

## 1. INTRODUCTION

Efficient and quick retrieval in a large-scale multimedia database is one emergent problem for content-based multimedia retrieval applications nowadays. A standard description of the media content is one important requirement for content reuse. The corresponding standard named as MPEG-7 has been worked on by ISO SC29 WG11 group, which will provide a standard description for multimedia content data. The description, along with the multimedia content, supports quick retrieval and easy access to the content in which the user is interested. The main elements of MPEG-7 standard are Descriptors ( D ), Description Schemes ( DS ), a Description Definition Language ( DDL ) and System Tools. MPEG-7 defines color, texture, shape and motion descriptors for description of low level features. While the definition of descriptors is specified within the scope of MPEG-7, the extraction of features and the search engine are not inside the scope of MPEG-7. Motion feature is significant for the description of video content. Video retrieval based on motion features is one important part of retrieval applications in video database. For instance, when browsing the video obtained by surveillance system or watching sports programs, the user always has the need to find out the object moving in some special direction. In our multi-feature MPEG-7 video retrieval project, the part of motion feature

* Correspondence: email: wuchuan00@mails.tsinghua.edu.cn; phone: +86-10-62786910; fax: +86-10-62771138



Storage and Retrieval for Media Databases 2002, Minerva M. Yeung, Chung-Sheng Li, Rainer W. Lienhart, Editors, Proceedings of SPIE Vol. 4676 (2002)

extraction, description and retrieval is of much importance. Automatic extraction algorithms of motion descriptors will be much useful.

Some related work has been done in the aspect of extraction of motion descriptors. Jeannin et. al.[1] proposed their algorithms for extraction of camera motion descriptor and motion trajectory descriptor. In their algorithm, the extraction of motion trajectory descriptor was based on the assumption that the object was already segmented correctly and they didn't deal with the problem of object segmentation. Kang et. al. [2] proposed their algorithm on compressed domain data, and only did their work on camera motion analysis. Divakaran et. al. [3] focused on motion activity descriptor, which described the activity in a video sequence in a whole. However, more accurate and complete motion information will be obtained if foreground area can be segmented automatically from the frames. Because in many situations motion of background and motion of foreground represent different semantic information. Based on the information, some useful and high-level information can be understood of the video content. Then content-based retrieval will become more efficient. In this paper, we try to extract more high-level information in this way. We further investigate algorithms to automatically extract global motion parameters and motion trajectory parameters from compressed domain, which are effective for the description of background and foreground motion respectively. The extraction algorithm of parametric motion descriptor is based on our former global motion estimation algorithm proposed for sprite coding used in MPEG-4 [4]. Based on the shots segmented correctly, we estimate the global motion with 6-parameter affine model. We propose the algorithm of global motion compensation to exclude the background from a frame and get the region of the object. Then we track the object automatically and create motion trajectory descriptor. Similarity measures of the two descriptors are defined for retrieval. In order to test the effect of proposed algorithms, some experiments are made based on a motion-based video retrieval system.

The rest of this paper is organized as follows. In section 2, we describe the automatic extraction algorithms of the two motion features. In section 3, similarity measures are defined. Experimental results are presented in section 4. Conclusions are given in section 5.

## 2. MOTION FEATURE EXTRACTION ALGORITHM

### 2.1 EXTRACTION OF PARAMETRIC MOTION

In MPEG-7 [6], parametric motion descriptor is defined which represents the global motion information in video sequences with 2D motion models. Global motion is the movement of background in a frame sequence and it is mainly caused by camera motion. Global motion information represents the temporal relations in video sequences. Compared with other video features, it can represent the high-level semantic information better. And it is important for motion-based object segmentation, object tracking, mosaicing, etc. Motion-based video retrieval can be implemented by parametric motion descriptor on the basis of appropriately defined similarity measure between motion models.

**[Extraction]**

In large-scale video databases, the videos are mostly MPEG-1, MPEG-2 compressed video streams. We estimate the global motion of video shots using DC images obtained by partly decoding the compressed video stream. A DC image is composed of the DC coefficients acquired by DCT transform on $8 \times 8$ blocks in a frame. It is a miniature of $1/8 \times 1/8$

the size of original frame and represents the color distribution of the original picture. So we use DC images for two reasons: 1) It can save much of the time taken by complete decoding of the compressed video streams and processing a full image. 2) Data in a DC image is the average of each block, so it has the effect of smoothing, depressing the influence of noises in a frame. In the description of our global motion estimation algorithm hereinafter, "image" represents the DC image of a frame, "a pixel" represents a point in a DC image, and "the pixel's position" is represented by the position of the pixel at the top left corner of the 8×8 block in the original frame.

We take six-parameter affine model as the model of our global motion estimation. We use $[x_t, y_t]^{\mathrm{T}}$ to represent the position of the pixel in current image. $[x_{t-1}, y_{t-1}]^{\mathrm{T}}$ is the corresponding position in the previous image. The relations between them can be show as Equation 1:

$$\begin{cases} x_{t-1} = ax_t + by_t + c \\ y_{t-1} = dx_t + ey_t + f \end{cases}. \tag{1}$$

Let $I(x, y)$ represent current image and $I'(x', y')$ represent the previous image. $\theta = (a, b, c, d, e, f)^{\mathrm{T}}$ is the parameter vector. We define energy function as

$$R(\theta) = \sum_{x_t, y_t} w[\mathrm{I}'(x'_{t-1}, y'_{t-1}) - \mathrm{I}(x_t, y_t)]^2. \tag{2}$$

The goal of global motion estimation is to get the optimum parameter vector $\theta$ which minimizes $R(\theta)$. Gause-Newton and Levenberg-Marquardet iterative methods can be used to solve such optimum problems. When we get the parameter vector $\theta_k$ at step k, we expand the energy function $R(\theta)$ at $\theta_k$ in Tailor series and drop the second-order terms:

$$R(\theta) \approx R(\theta_k) + g_k^{\mathrm{T}}(\Delta\theta_k) + \frac{1}{2}(\Delta\theta_k)^{\mathrm{T}}\mathrm{H}_k(\Delta\theta_k) \quad . \tag{3}$$

In the above equation, $g_k$ is gradient matrix and $\mathrm{H}_k$ is Hessian matrix. They are defined as follows:

$$g_k = J_k^{\mathrm{T}} W \gamma_k, \mathrm{H}_k = J_k^{\mathrm{T}} W J_k + \sum_i \gamma_i w_i H_{ik}. \tag{4}$$

$\gamma_k = [r_1 r_2 \cdots r_N]^{\mathrm{T}}$ represents the residuals at $\theta_k$. $J_k = \partial\gamma / \partial\theta$. The weight matrix $W$ is a diagonal matrix. $H_{ik}$ is the Hessian matrix of $\gamma_i$. If $\gamma$ is small, we can get the approximation:

$$\mathrm{H}_k \approx J_k^{\mathrm{T}} W J_k . \tag{5}$$

Let $\partial R(\theta)/\partial\theta = 0$, then we get

$$J_k^{\mathrm{T}} W J_k (\Delta\theta_k) = -J_k^{\mathrm{T}} W \gamma_k . \tag{6}$$

We can get the increment of $\theta_k$ from the above equation. Then the parameter vector at next step can be got: $\theta_{k+1} = \theta_k + \Delta\theta_k$. By this iterative calculation, we can finally get an optimum parameter vector.

In our algorithm, we calculate one parameter vector for every pair of adjacent DC images in one video shot. Since matching by the 6 parameters in the global motion model directly does not have much meaning in a retrieval application, we instead calculate the horizontal translation velocity $T_x$, vertical translation velocity $T_y$, angle of rotation $\theta$ and scale $s$ of global motion from every parameter vector of the global motion model. We transform the six-parameter affine model in Equation 1 into the following:

$$\begin{bmatrix} x_{t-1} \\ y_{t-1} \end{bmatrix} = sR \begin{bmatrix} x_t \\ y_t \end{bmatrix} + \begin{bmatrix} T_x \\ T_y \end{bmatrix} . \tag{7}$$

In Equation 7, $s$ is the scale. $T_x$ is the horizontal translation velocity. $T_y$ is the vertical translation velocity. And R is the rotation matrix, defined as

$$R = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}. \tag{8}$$

in which $\theta$ is the angle of rotation. We then calculate the new transformed parameter vector $(T_x, T_y, s, \theta)^T$ in Equation 7 from $(a, b, c, d, e, f)^{\mathrm{T}}$ in Equation 1.

We select the histogram space of global motion parameters as the feature space and cluster the feature points in it. We use the parameter vector of the point at the center of one cluster as the cluster's representation. Thus we get a set of parameter vectors as representations for this shot. We describe this set of parameter vectors with MPEG-7 parametric

motion descriptor. They form the description of the shot's global motion. In our implementation, the parametric description of one video shot contains 3-4 parameter vectors.

## 2.2 EXTRACTION OF MOTION TRAJECTORY

MPEG-7's motion trajectory descriptor includes a list of keypoints and a set of interpolating functions that describe the trajectory of object between two keypoints. The interpolating function used is first order interpolating or second order interpolating function. The extraction of motion trajectory is significant for object-based video retrieval. In a given context with certain priori knowledge, it can be much useful in many applications. For example, the detection of an object with dangerous moving trajectory in a surveillance system, the retrieval of special action in a sports game, etc.

**[Extraction]**

Our algorithm here is also implemented on DC images and based on the assumption that the shots are segmented correctly. Before tracking of object in a shot, we should first detect the object region. After we get global motion parameters of the background, the background can be excluded from the image by global motion compensation. In the different map between each pair of frames, object region are left with high remainder value. Thus we get the rough region of the object. The object region can be further refined with morphological image processing methods. We then project the differences in a difference map along the horizontal axis and the vertical axis. We get two 1-D histograms along the two axes. A simple case with one moving object in the video sequence is illustrated in Figure 1.

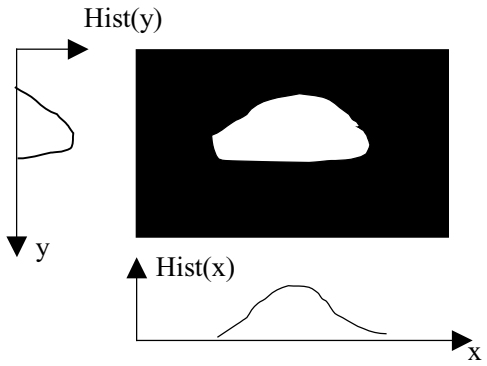


Figure 1: Difference Projection Along x and y Axes

We calculate the position of the moving object in a frame with the statistical results in each histogram. We compute the mean of the sample distribution in the histogram along x axis as the x coordinate of the center of the moving object, and the mean of sample distribution in the histogram along y axis as the y coordinate.

When tracking the object, we calculate the center of the object region with above algorithm for each frame. We then select the set of keypoints to be described with MPEG-7 motion trajectory descriptor from the set of centers. The trajectory we extract is a 2-D trajectory. And the algorithm is applied to each dimension (x and y) respectively. We begin with the interval containing the first three centers in the frame sequence. When current interval contains N

centers, we calculate a second order interpolating function as in Equation 9 to approximate object trajectory on this interval.

$$f(t) = f_a + v_a(t - t_a) + \tfrac{1}{2} a_a (t - t_a)^2 . \tag{9}$$

Then the next center is taken into this interval. We keep the interpolating function as the same and calculate the approximation error. If the error is smaller than the threshold, the new center is added to this interval and the process repeats. Otherwise, the N center interval is kept and the first and last centers are taken as the keypoints to be described in MPEG-7 motion trajectory descriptor, together with the second order interpolating function. The process then starts again with the following three centers. Finally, we can get the motion trajectory descriptor of the shot.

## 3. SIMILARITY MATCHING METHODS

We implement and test our algorithms with query-by-example mode. We extract the descriptors of the example selected by the user, and then match them with descriptors of shots in the database.

For parametric motion descriptor, we define similarity measure on the basis of four parameters $(T_x, T_y, s, \theta)^T$, calculated from the 6 parameters of affine motion model obtained by global motion estimation. The similarity measure is defined as follows,

$$M(shot1, shot2) = \frac{W_X M_X(shot1, shot2) + W_Y M_Y(shot1, shot2) + W_\theta M_\theta(shot1, shot2) + W_S M_S(shot1, shot2)}{W_X + W_Y + W_\theta + W_S} \tag{10}$$

.

$$M_X(shot1, shot2) = \sum_i (T_{Xi1} - T_{Xi2})^2 . \tag{11}$$

$$M_Y(shot1, shot2) = \sum_i (T_{Yi1} - T_{Yi2})^2 . \tag{12}$$

$$M_\theta(shot1, shot2) = \sum_i (\theta_{i1} - \theta_{i2})^2 . \tag{13}$$

$$M_S(shot1, shot2) = \sum_i (S_{i1} - S_{i2})^2 . \tag{14}$$

$i = 0,1,...,m$ . $m$ is the number of representative global motion parameter vectors for the shot. $W_X, W_Y, W_\theta, W_S$ are the weights between 0 and 1, selected according to different applications. In our implementation, the weights can be adjusted by users in the query interface. For example, if the user cares more for querying video shots with dominant

horizontal motion than other motion, he can give $W_X$ a larger weight approximating 1 while giving $W_Y, W_\theta, W_S$ the weights near 0.

We define the following similarity measure for motion trajectory descriptor:

$$M(shot1, shot2) = \frac{W_p M_p(shot1, shot2) + W_s M_s(shot1, shot2) + W_a M_a(shot1, shot2)}{W_p + W_s + W_a}. \quad (15)$$

$$M_p(shot1, shot2) = \sum_i ((x_{i1} - x_{i2})^2 + (y_{i1} - y_{i2})^2). \quad (16)$$

$$M_s(shot1, shot2) = \sum_i ((v_{xi1} - v_{xi2})^2 + (v_{yi1} - v_{yi2})^2). \quad (17)$$

$$M_a(shot1, shot2) = \sum_i ((a_{xi1} - a_{xi2})^2 + (a_{yi1} - a_{yi2})^2). \quad (18)$$

$i = 0,1,...,n$. $n$ is the number of keypoints chosen to describe the trajectory. $W_p, W_s, W_a$ are the weights between 0 and 1. $M_p$ is the Euclidian measure between position vectors of the keypoints in the motion trajectory descriptor. $M_s$ is the Euclidian measure of velocity vectors of keypoints. And $M_a$ is that of acceleration vectors.

The similarity measures of parametric motion descriptor and motion trajectory descriptor can be combined to form one similarity measure. We first normalize each similarity measure to the same dynamic range of 0 to 1. Then we calculate the linear combination of the two similarity measures as our final similarity measure.

$$S = \alpha_1 M_1 + \alpha_2 M_2. \quad (19)$$

$\alpha_1, \alpha_2$ are the weights between 0 and 1. $M_1$, $M_2$ are calculated as in Equation 10 and 15 respectively.

## 4. EXPERIMENTAL RESULTS

Our experiments are made on a motion-based video retrieval system as shown in Figure 2. In the user interface, users can choose to retrieve video shots with parametric motion feature or object trajectory feature. Or they can use both features at the same time. The users can also adjust the weights in similarity measures according to their own needs.

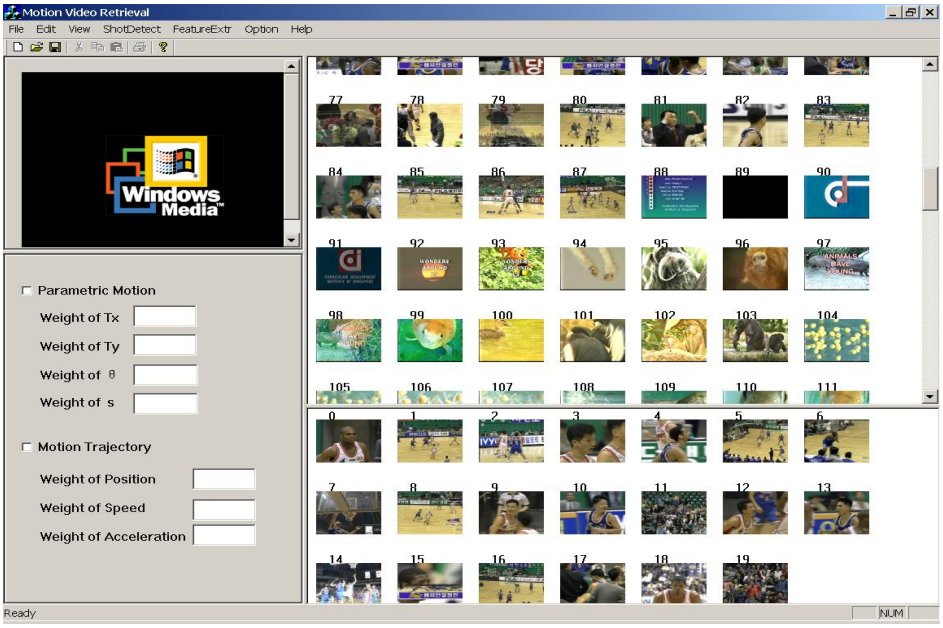


Figure 2: A Snapshot of Motion-based Video Retrieval System Interface

In our experiments to test the effect of our motion descriptor extraction algorithm, we construct a shot database of 358 shots from MPEG-7 test data set. The shots are taken from 40-minute video sequences, including basketball, golf, animals, news, etc. In order to test the efficiency of parametric motion feature extraction scheme, we choose different query shots representative of the following camera motion respectively, panning, tracking, booming, zooming and slight motion. In order to test the efficiency of object motion trajectory extraction scheme, we choose different query shots with certain object moving trajectory. Queries are also made by combining the two descriptors. The retrieved shots are ranked according to their similarity to the query shots. Parts of the experimental results are illustrated in Figure 3, Figure 4 and Figure 5. We display the best 5 matches for each query shot.

In Figure 3, we query shots of "tracking left".

Query Shot Five Best Matches

Figure 3: An Example of "Tracking" Query

In Figure 4, the main camera motion in the query shot is zooming.

Query Shot Five Best Matches

Figure 4: An Example of "Zooming" Query

In Figure 5, the object in the query shot is moving right.


Figure 5: An Example of "Moving Right" Query

The query experiments give good results. So in general, our parametric motion descriptor and motion trajectory descriptor extraction algorithms could work very well.

## 5. CONCLUSIONS AND FUTURE WORK

Experimental results indicate that our algorithm is effective. In this paper, we focus our research on motion analysis, which is a most consistent feature among all low-level features for videos. The motion retrieval system proposed in the paper is a practical attempt in motion-based video retrieval. And motion of background and foreground can be extracted with our algorithm. We can further take some methods to accelerate the convergence of our motion estimation algorithms. In our object's tracking algorithm, we get the region of object motion by wiping off the background in a frame. Algorithm to automatically obtain the precise contour of the object is to be developed in the future. We will continue our research in these aspects. Our retrieval is implemented by query-by-example and low-level motion features. For practical video retrieval system, it should support multi-modal and multi-feature retrieval. Thus all these work will be extended to develop retrieval applications based on high-level semantic query combined with multiple features.

## REFERENCES

1. S. Jeannin, B. Mory, "Video Motion Representation for Improved Content Access", *IEEE Transaction on Consumer Electronics*, Vol. 46, No. 3, pp. 645-655, August 2000
2. H.B. Kang, "Spatio-Temporal Feature Extraction from Compressed Video Data", *TENCON 99, Proceedings of the IEEE Region 10 Conference* , Vol. 2, pp. 1339-1342, 1999
3. A. Divakaran, A. Vetro, "Video Browsing System Based on Compressed Domain Feature Extraction", *IEEE Transaction on Consumer Electronics*, Vol. 46, No. 3, pp. 637-644, August 2000
4. Y.W. He, Y.Z. Zhong, S.Q. Yang, "Fast Approach of Sprite Coding for Video Content", *proceedings of the SPIE International Symposium on Information Technologies 2000(ISIT 2000)*, Boston Massachusetts, USA, November 2000
5. Y.W. He, L. Zhao, S.Q. Yang, Y.Z. Zhong, "Region-based Tracking in Video Sequences Using Planar Perspective Models", *proceedings of the 3rd International Conference on Multimodal Interfaces(ICMI 2000)*, Beijing, October 2000
6. Y.W. He, "Global Motion Estimation Algorithm and Its Application in Video Coding", Master's Thesis, Tsinghua University, P. R. China, October 2000

7. T. Vlachos, “Simple Method for Estimation of Global Motion Parameters Using Sparse Translational Motion Vector Fields”, *Electronics Letters*, Vol. 34, No. 1, pp. 60-62, January 1998
8. ISO/IEC JTC1/SC29/WG11 N3752, “Overview of the MPEG-7 Standard”, La Baule, October 2000
9. ISO/IEC JTC1/SC29/WG11/N3705, “Text of ISO/IEC 15938-5/CD Information Technology-Multimedia Content Description Interface-Part 5 Multimedia Description Schemes”, La Baule, October 2000