Efficient Motion-Vector-Based Video Search Using Query by Clip

Chuan-Yu Cho, Ya-Ting Chuang, Pei-Chi Chu, Shih-Yu Huang*, and Jia-Shung Wang

Department of Computer Science, National Tsing Hua University, Hsinchu, Taiwan
*Department of Computer Science, Ming Chuan University, Taoyuan, Taiwan

ABSTRACT

In this paper, a simple and efficient motion-vector-based search engine supporting query by clip is presented for indexing and retrieving of videos. We propose a technique to filter and reconstruct noisy motion vectors to improve the fidelity of the motion activity. In addition to the motion features of intensity and direction, a new motion descriptor named representative spectrum extracted by a VQ expressed the spatial-temporal distribution of motion activity is introduced for video indexing. A dynamic programming scheme is also utilized to measure the similarity of two videos based on the concept of common subsequence. Furthermore, a 2-step search of similar videos is employed to reduce the computational complexity. The experimental results indicate that the proposed search engine performs well and the obtained recall and precision values are also high enough.

1. INTRODUCTION

Owing the rapid development of multimedia techniques especially on the MPEG-1/2 compression standards [1-2], the volume of multimedia contents is growing larger and wider today. However, it’s difficult to manage, find, select and filter what is needed on the huge volume of MPEG-1/2 contents. MPEG-7 [3], a standard for describing features of multimedia content, is introduced for efficient content indexing and retrieval applications. Much works [4] has been done on indexing of still image features such as color, texture, shape, etc. However, motion has rich meaning in video sequences. For example, in basketball, the motion trajectory of sportsmen allows the discrimination of “a left to right offense.” Cooperated with indexing based on still image features, motion-based indexing has been shown to significantly improvement of video browsing and retrieval system [5].

MPEG-7 introduces four motion descriptors (camera motion, motion trajectory, parametric motion, and motion activity) to support a wide range of applications. However, some limitations still exist in the motion-based video retrieval system. The first is that the motion descriptors of MPEG-7 are not friendly to a common user. It is difficult to utilize the motion descriptors to precisely describe what is needed. The second is that it is hard to extract the motion descriptors except the motion activity which could extract from the motion vectors on the MPEG-1/2 bitstreams. In the previous researches [6-7], algorithms of shot segmentation, key frame selection, object detection and tracing are widely employed to extract the camera motion, motion trajectory, and parametric motion either in the image or in the motion vector domain. However, lots of computations are necessary for these algorithms and the extracted information has wide variations. Different techniques may give different results, resulting in different motion descriptors.

In this paper, a friendly and efficient motion-vector-based video search engine is proposed for video retrieval systems. The basic idea is not to explicitly describe the motion descriptors. We allow queries involving a video clip (a few seconds of video which includes a set of shots describing a particular event). The motion information is implicitly represented by a series of motion activities. The example-based approach [8-9] is a good alternative because the users could not always say clearly what they want. It is natural and friendly to users. Since we only want to find what is similar to the query video clip, it is unnecessary to know the exact content in the sequence. No complicated algorithms of shot segmentation, key frame selection, object detection and tracing are employed by the proposed search engine. In this paper, a 2-step search is employed to reduce the computational complexity to retrieve the similar video segment. Based on the common subsequence concept, a new similarity measurement solved by a dynamic programming technique is also proposed.

The other characteristic of the proposed search engine is that only the motion vectors in the compressed domain are utilized to extract the descriptor of motion activity. This concept is also presented on [10-11] but most of them are focused on the design of intensity and direction descriptors. In this paper, a new motion descriptor named representative spectrum extracted by a vector quantization scheme [12] is introduced to describe the spatial-temporal distribution of motion vectors. Because motion vectors are originally designed...
for the coding purposes, motion activity of videos cannot directly be generated from the motion vectors. To improve the fidelity of motion activity, a technique is employed to filtering and reconstructing noisy vectors.

The rest of this paper is organized as follows. Section 2 presents the proposed video search engine. Section 3 describes the experimental results. The conclusions are given in Section 4.

**2. THE PROPOSED SEARCH ENGINE**

The structure of the proposed video search engine is depicted in Figure 1. A huge volume of MPEG-1/2 videos are stored in the video database. The corresponding motion activities of videos are extracted by motion vectors in the compressed bitstreams and stored in the motion activity database. The search engine allows queries by video clip because of the difficulty to explicitly describe the motion descriptors. Similarly, the motion activities of the queried clip are first extracted from motion vectors on compressed domain. Based on a similarity measurement, the extracted motion activities are then matched with the activities stored in the motion activity database. k retrievals with largest similarity would be generated by a query.

![Figure 1. Structure of the search engine.](image)

Intensity, direction, spatial distribution, and temporal distribution are the four descriptors of motion activity in MPEG-7. The proposed video search engine compresses the above descriptors into intensity spectrum, direction spectrum and intensity-direction spectrum for compresses the above descriptors into intensity spectrum, activity database. The search engine designed for the coding purposes, motion activity of videos are stored in the video database. The search area of motion estimation is limited within ±15 pels. A 2-dimensional vector quantization (VQ) is employed to quantize all points into a few representational points (codevectors). The number of codeword is set to be 16 and the Linde-Buzo-Gray (LBG) algorithm is used to design the codebook with the training set consists about 300 frames. Every scene generates an individual codebook and more codevectors can be used if the motion activity is more complex. A representative value of a frame, \( R_i \), is defined as the sequence of the number of the quantized motion vectors in the order of codebook, i.e.,

\[
R_i = (r_{i0}, r_{i1}, \ldots, r_{i31}),
\]

where \( r_j \) is the number of the \( j^{th} \) quantized motion vector. The representative spectrum of a video clip is denoted as the histogram of the corresponding representative values. The spectrum can tell us the variation of major motion activity by the time.

To retrieve the similar ones in the video database of a query clip is the key function in a video search engine. The formulation of similar match problem can be addressed as follows. Let the queried clip \( Q \) be denoted as \( Q = \{Q_1, Q_2, \ldots, Q_m\} \), where \( Q_i \) denotes the \( i^{th} \) frame in \( Q \) and \( m \) is its length, and the source video \( S \) be denoted as \( S = \{S_1, S_2, \ldots, S_n\} \), where \( S_i \) denotes the \( i^{th} \) frame in \( S \) and \( n \) is its length. The problem of exact match can be formulated as to find the clip \( S' \) with \( D(Q, S') \) smaller than a preset threshold, where \( S' = \{S_{i0}, \ldots, S_{im}\} \) is a video segment with \( m \) successive frames from the source video and \( D(Q, S') \) is the distortion between the clip \( Q \) and \( S' \). There are multiple retrievals in a query and the retrievals may be longer or shorter than the queried clip. If the length difference of the two clips is too large, we conclude they are not similar. Therefore, the length of the similar segment, \( S'^{kl} = \{S_i, \ldots, S_{j}\} \), is restricted in the range of half to 1.5 times of the queried clip length. The similar retrieval is to find the first \( k \) clips \( S' \) with \( k \) smallest distortions \( D(Q, S'^{kl}) \),
the dissimilarity value between the query clip \( Q \) and the clip \( S^j_n \). In the case of distortions with the same value, the clip which length is close to the length of \( Q \) is chosen.

Dynamic programming approach is utilized to compute \( D(Q,S^j_n) \). Let \( D(m,n) \) be the distortion of optimal matching of the first \( m \) frames in \( Q \) and the first \( n \) frames in \( S^j \) with \( D(0,0)=0 \), \( D(0,n)=\infty \) and \( D(m,0)=\infty \). Then, \( D(m,n) \) can be derived by the following:

\[
D(m,n) = \min \begin{cases} 
D(m-1,n) + d(Q_m, S^j_n) \\
D(m-1,n-1) + d(Q_m, S^j_n) \\
D(m,n-1) + d(Q_m, S^j_n) \\
\end{cases}
\]

where \( d(Q_m, S^j_n) \) is the distortion of the \( m \)th frame in \( Q \) and the \( n \)th frame in \( S^j \) and could be computed in the domains of intensity, direction, or representative.

Response time of a query is another important issue in the design of a video search engine. Although the representatives of motion can provide more accurate information than the intensity or direction of motion, the required computations to evaluate the distortions in the representative domain are larger than that in the other domains. 16 bins of distortions should be computed when the representative features are utilized. A strategy called 2-step search is employed to deduce the computations. In the first step, intensity and/or direction features are employed to eliminate a large portion of unqualified candidates. The remaining candidates are refined by the representative feature in the second step. The order of utilized motion features is based on the corresponding importance for the queries. For instance, suppose what we want is to find the "fast-break" clip in a basketball sequence, the intensity is the key feature and should be employed in the first step search. Then, the direction or representative features could be utilized in the next steps. In the other example, if we want to find the "left to right offense", the direction feature is more important than the intensity feature.

3. EXPERIMENTAL RESULTS

The experimental sequences come from a 5 minutes basketball video recorded from TV and a 5 minutes video of "The leopard family" from the Discovery channel. They are MPEG-1 format. Four video segments are selected from the two sequences to be the queried clips in our experiment.

1. "A left to right offense" in the basketball game, about 7 seconds.
2. "An offense under the basket" in the basketball game, about 4 seconds.
3. "An animal walking from right to left" in the video of "the leopard family" from the Discovery channel, about 6 seconds.
4. "A hunting event" in the video of "the leopard family" from the Discovery channel, about 17 seconds.

In the experiment of similar match, we want to retrieve the video segments similar to the queried video clip. The top 3 similar match for the first two queried clips are given in Figure 2. The queried clips of "a left to right offense" and "an animal walking from right to left" use direction feature as the key feature in our proposed 2-step search, whereas "an offense under the basket" and "a hunting event" use intensity feature as the key feature. We use 5 frames to represent each video clip.

In order to evaluate the reliability and performance of our strategy, ground truth is manually analyzed by the two video sequences of basketball game and the leopard family. The performance is measured by precision and recall, where recall is the ratio of the number of similar segments detected correctly over the actual number of similar segments and precision is the ratio of the number of similar segments detected correctly over the total number of similar segments detected correctly or incorrectly. Based on the human's relevance judgment, the recall and precision values for the four queries are presented in Table 1. In general, the results are good except the query of "an offense under the basket." This is because motion intensity is not a good feature in the basketball game. Many segments have similar high motion intensity of an offense under the basket.

![Table 1. The recall and precision values.](image)

<table>
<thead>
<tr>
<th>The content of queried clip</th>
<th>Precision</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Left to right offense</td>
<td>5/6</td>
<td>5/5</td>
</tr>
<tr>
<td>An animal walking from right to left</td>
<td>6/6</td>
<td>6/7</td>
</tr>
<tr>
<td>An offense under the basket</td>
<td>4/7</td>
<td></td>
</tr>
<tr>
<td>A hunting event</td>
<td>4/4</td>
<td>4/4</td>
</tr>
</tbody>
</table>

4. CONCLUSIONS

Feature extraction and similarity measurement are the key researching topics on the problem of video indexing and retrieval. In this paper, a new motion descriptor named representative spectrum is proposed for spatial-temporal distribution of motion activity. Because the feature extraction based on a VQ scheme is in the compressed domain, it's extremely simple and fast. In addition, a dynamic programming measurement based on the concept of common subsequence is presented to extract the similarity of between two video clips. A filtering and reconstructing technique is introduced to remove the noisy motion vectors to improve the fidelity of motion activity. Furthermore, a 2-step search is proposed to reduce the computational complexity. Our experiments show promising results in some constrained domains such as basketball game.

References:


Figure 2. The retrievals of a left to right offense and an animal walking from right to left.