Available online at www.sciencedirect.com
SCIENCE DIRECT®
J. Vis. Commun. Image R. 16 (2005) 212–232
JOURNAL OF VISUAL Communication & IMAGE Representation
www.elsevier.com/locate/jvci

Robust video sequence retrieval using a novel object-based T2D-histogram descriptor

Duan-Yu Chen[a,*], Suh-Yin Lee[a], Hong-Yuan Mark Liao[b]

[a] *Department of Computer Science and Information Engineering, National Chiao-Tung University, 1001 Ta-Hsueh Road, Hsinchu, Taiwan, ROC*
[b] *Institute of Information Science, Academia Sinica, 128 Sinica Road, Sec 2, Nankang, Taipei 11529, Taiwan, ROC*

Received 10 June 2003; accepted 3 August 2004
Available online 8 October 2004

Abstract

Due to the tremendous growth in the number of digital videos, the development of video retrieval algorithms that can perform efficient and effective retrieval task is indispensable. In this paper, we propose a high-level motion activity descriptor, object-based transformed 2D-histogram (T2D-histogram), which exploits both spatial and temporal features to characterize video sequences in a semantics-based manner. The discrete cosine transform (DCT) is applied to convert the object-based 2D-histogram sequences from the time domain to the frequency domain. Using this transform, the original high-dimensional time domain features used to represent successive frames are significantly reduced to a set of low-dimensional features in frequency domain. The energy concentration property of DCT allows us to use only a few DCT coefficients to effectively capture the variations of moving objects. Having the efficient scheme for video representation, one can perform video retrieval in an accurate and efficient way.

Keywords: Motion activity; Spatio-temporal feature description; Video sequence matching; Discrete cosine transform; Video similarity measure

* Corresponding author. Fax: + 886 3 5724176.
E-mail addresses: dychen@csie.nctu.edu.tw (D.-Y. Chen), sylee@csie.nctu.edu.tw (S.-Y. Lee), liao@iis.sinica.edu.tw (Hong-Yuan Mark Liao).

1047-3203/$ - see front matter © 2004 Elsevier Inc. All rights reserved.
doi:10.1016/j.jvcir.2004.08.003

1. Introduction

The tremendous growth in the number of digital videos has become the main driving force for developing automatic video retrieval techniques. Among different types of tools that can push the advancement of retrieval techniques, an efficient automatic content analyzer that can help execute correct browsing, searching, and filtering of videos is a must. To achieve this goal, one has to make use of high-level semantic features to represent video contents. The need of representing high-level semantic features has motivated the emergence of MPEG-7, formally called the multimedia content description interface (Sikora, 2001). However, the methods that produce the specific features and the corresponding similarity measures represent the non-normative part of MPEG-7 and are still open for research and future innovation.

Usually, the high-level semantic features of video sequences can be inferred from low-level features. The low-level features can be color distribution, texture composition, motion intensity, and motion distribution. Among different types of features that can be extracted from a video, motion is considered as a very significant one due to its temporal nature. In the literature, Divakaran et al. (2000) used a region-based histogram to compute the spatial distribution of moving regions. The run-length descriptor in MPEG-7 (Jeannin and Divakaran, 2001) is used to reflect whether moving regions occurred in a frame. Aghbari et al. (1998) proposed a motion-location-based method to extract motion features from divided sub-fields. Peker et al. (2000) calculated the average motion vectors of a *P*-frames and those of a video sequence to be the overall motion features. In addition to the above mentioned local motion features, Wang et al. (2000) and Tang et al. (2000) proposed to use some global motion features to describe video content.

In contrast to the motion-based features of individual frames, another group of researchers proposed to use spatio-temporal features between successive frames because these types of features are more abundant in the amount of information. Wang et al. (2001) extracted features of color, edge and motion, and zmeasured the similarity between temporal patterns using the method of dynamic programming. Lin et al. (2001) characterized the temporal content variation in a shot using two descriptors—dominant color histograms of group of frames and spatial structure histograms of individual frames. Cheung and Zakhor (2001) utilized the HSV color histogram to represent the key-frames of video clips and designed a video signature clustering algorithm for detecting similarities between videos. Agnihotri and Dimitrova (2000) represented video segments by color super-histograms, which are used to compute color histograms for individual shots. Other works that fall into this category can be found in (Manjunath et al., 2001; Mohan, 1998; Roach et al., 2001; Yeung and Liu, 1995; Zhao et al., 2001).

There are several drawbacks associated with the key-frame-based matching process. First, the features selected from key-frames usually suffer from the high-dimensionality problem. Second, the features chosen from a key-frame is in fact local features. For a matching process that is targeting at measuring the similarity among a great number of video clips, the key-frame-based matching method is not really feasible because the information used to characterize the relationships among consecutive frames is not

taken into account. To overcome these drawbacks, we propose an object-based motion activity descriptor, which can exploit the spatio-temporal information of a video clip in the matching process. Basically, the proposed spatio-temporal features can support high-level semantic-based retrieval of videos in a very efficient manner. We make use of some spatio-temporal relationships among moving objects and then use them to support the retrieval task. In the retrieval process, we use the discrete cosine transform (DCT) to reduce the dimensionality of the extracted high-dimensional feature. Using DCT, we can maintain the local topology of a high-dimensional feature. In addition, the energy concentration property of DCT allows us to use only a few DCT coefficients to represent the moving objects and their variations. Therefore, the transformation can make an accurate and efficient retrieval process possible.

The rest of the paper is organized as follows. Section 2 presents an overview of the proposed scheme. Section 3 illustrates the methods used to characterize video segments. Section 4 describes the representation and matching of video sequences. Section 5 presents the experimental results. Section 6 draws conclusions and suggests avenues for future work.

2. Overview of the proposed scheme

In this section, we shall provide an overview of the proposed video retrieval system. Fig. 1 shows the flowchart of the proposed system. MPEG videos are efficiently

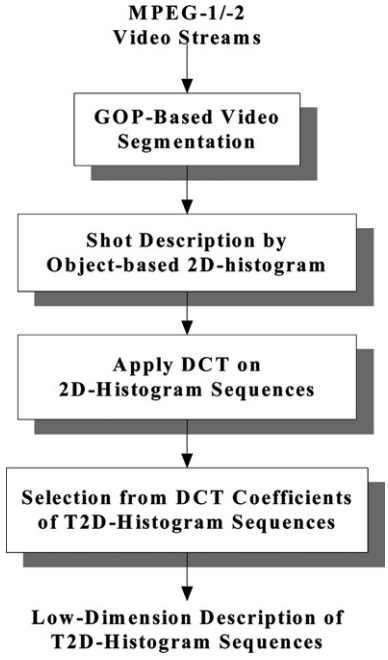

Fig. 1. An overview of extracting the proposed T2D-histogram descriptor—compressed videos are parsed semantically and represented by reduced low-dimensional DCT coefficients.

segmented into shots using our previously proposed GOP-based video segmentation algorithm (Lee et al., 2001). This video segmentation algorithm checks video streams GOP-by-GOP rather than frame-by-frame. The actual shot boundaries are then determined at the frame level. After the process of shot segmentation, the next step is to execute an algorithm, which can generate an object-based motion activity description. The motion activity descriptor is able to describe moving objects in compressed videos. The features used by this motion descriptor are statistically computed by spatial and temporal distributions along the horizontal and vertical directions, respectively. The function of the descriptor is basically an encoder, which can encode video contents into high-level relational features. In order for maintaining high-computational efficiency, we choose *P*-frames for motion activity analysis. Under these circumstances, a video clip can be represented by a set of motion activity descriptions of consecutive frames in the time domain. However, it is impractical to search a large video database using the time domain features. Therefore, we propose to apply DCT on the target frames and make them become lower dimensional in the frequency domain. Finally, we conduct an indexing process on the transformed DCT coefficients. As we mentioned before, due to the energy concentration property of DCT, we are able to represent the original moving objects in a most accurate and efficient way.

3. Characterization of video segments

In this section, we shall describe how to characterize a video segment so that it can be used to perform efficient video retrieval. We shall describe how to detect moving objects in a video segment in Section 3.1 and then discuss how to describe motion activity of a video segment in Section 3.2.

3.1. Moving object detection

For computational efficiency, motion information in *P*-frames is used for the detection of moving objects. In general, consecutive *P*-frames separated by two or three *B*-frames are still similar and would not vary too much. Therefore, it is reasonable to use *P*-frames as targets for moving objects detection. On the other hand, since the motion vectors estimated in MPEG-2 videos may not be 100% correct, one has to remove the noisy part before they can be used. For those motion vectors that are small in magnitude, we consider they are noises and should be removed. For the sake of computation speed, the average of motion vectors in those inter-coded macroblocks is computed and selected as the threshold for noise removal. After noisy motion vectors are filtered out, the motion vectors with similar magnitude and direction are clustered into a group by applying a region growing process with an morphological operator of 2×2 macroblocks. Thus, moving areas with size smaller than 4 macroblocks would be recognized as noises and be removed. Fig. 2 illustrates some examples of moving object detection in MPEG videos.

In our previous works (Chen and Lee, 2001; Chen et al., 2002), we have successfully detected moving objects in several kinds of videos such as tennis, traffic

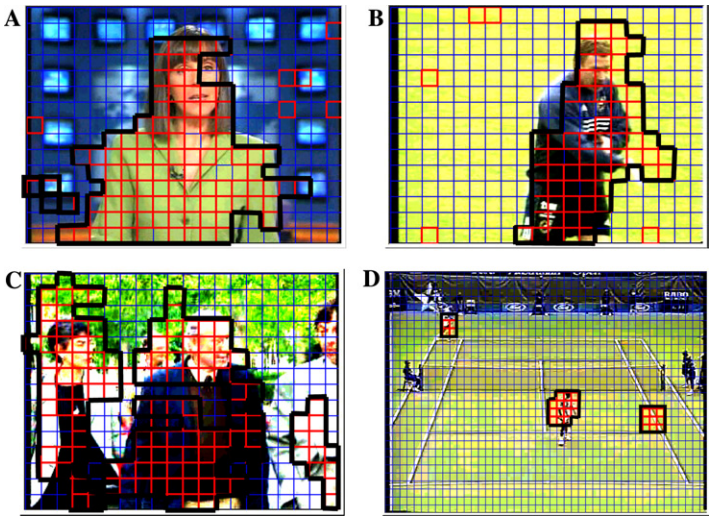

Fig. 2. Demonstration of moving object detection: (A) anchor person, (B) football, (C) walking person, and (D) tennis competition.

monitoring, news, and football. Moving objects can be detected with an over 90% success rate when the camera is stationary. When the camera moves, camera motion such as pan or tilt should be estimated in advance before detecting moving objects. In our previous work, the precision is about 83% when the camera moves. However, the recall is still higher than 90%. Examples of moving object detection using our previous algorithm are demonstrated in Fig. 2. Video shots shown in Figs. 2A–C are extracted from an MPEG-7 testing dataset, and the shot of tennis competition in Fig. 2D is recorded from the Star-Sports TV-channel. Based on the results shown in Fig. 2, it is obvious that all moving objects are successfully detected.

3.2. Describing motion activity in a video segment

In this section, we shall elaborate how to describe object-based motion activity in a video segment. After moving objects are detected, the spatial distribution of them is characterized using the statistics derived from the 2D-histogram. A 2D-histogram for each P-frames consists of an X-histogram and a Y-histogram. The horizontal axis of the X-histogram (Y-histogram) is the quantized X-coordinate (Y-coordinate) in a frame. The X- and Y-coordinates are quantized into β bins, which should be moderate and be adaptive to various content types of MPEG videos. Thus, β should be related to the frame resolution and the threshold of object size-based noise filtering, and is defined by

$$\beta = \min\left(\frac{R_{\text{row}}}{\sqrt{S}}, \frac{R_{\text{column}}}{\sqrt{S}}\right), \tag{1}$$

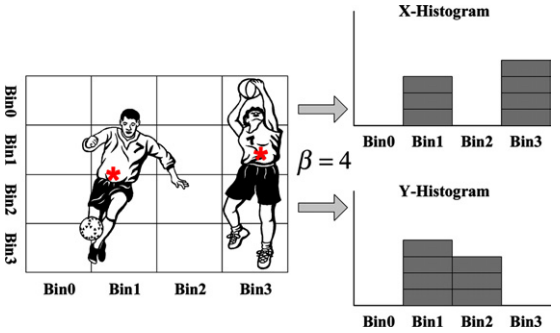

Fig. 3. Demonstration of the computation of 2D-histogram.

where $R_{\text{row}} \times R_{\text{column}}$ is the resolution of frame size in terms of macroblocks and S is the size of morphological operator in noise filtering. The decision of β will be verified by the simulated results in Section 5.3. Initially, the object size is estimated before bin assignment. If an object is larger than the predefined unit size (frame-size/β^2), then it is normalized and accumulated according to the following equation:

$$\text{Bin}_{i,j}^{X} = \sum_{r=1}^{\text{Obj}} \text{Acc}_{i,j,r}^{X}, \tag{2}$$

where

$$\text{Acc}_{i,j,r}^{X} = \begin{cases} 1 & \text{if object size} \leqslant \frac{1}{\beta^2} \text{ frame size}, \\ \frac{\text{size of object } \gamma}{\text{frame size}} * \beta^2 & \text{otherwise}, \end{cases}$$

where $\text{Bin}_{i,j}^{X}$ denotes the jth bin of an X-histogram in-frame i, $\text{Acc}_{i,j,r}^{X}$ means the accumulated value of the jth bin of object r in-frame i for an X-histogram, and Obj is the number of objects in-frame i. Fig. 3 shows how a 2D-histogram is computed, with the number of histogram bins set to four. In the example, two objects with sizes of three units and four units are present in the frame. To obtain the X-histogram, the size of each object is assigned to a histogram bin based on its centroid (indicated by the symbol "*") on the horizontal axis. For example, the football player of size three is assigned to Bin 1 and the basketball player of size four is assigned to Bin 3 in the X-histogram. Similarly, in the Y-histogram, Bin 2 is increased by 3 and that of Bin1 is increased by 4.

Using the proposed 2D-histogram, the spatial distributions among moving objects are approximately described since each moving object is assigned to the histogram bin based on its centroid. Objects that belong to the same coordinate interval are grouped into the same bin, and thus the distance between object groups can be specified as the distance between the associated bins.

4. Video sequence matching

After video segments are characterized by the descriptor of object-based 2D-histogram, temporal relationships among the moving objects have to be described.

To characterize the temporal relationships among moving objects, a few DCT coefficients of the transformed time sequence are used to represent the variations of original objects among consecutive frames. A brief review of DCT will be elaborated in Section 4.1. Section 4.2 will describe how to represent a video sequence. The similarity metric that can be used to measure the degree of similarity will be discussed in Section 4.3.

4.1. Discrete cosine transform

The discrete cosine transform (DCT) is a powerful tool that has been extensively used in many data compression applications. The DCT of a finite length sequence often has its coefficients more highly concentrated at low indices than other transforms do (Oppenheim and Schafer, 1999). It has been proven in Shi and Sun (2000) that the approximation capability of DCT is much better than that of other approximation methods. Therefore, we shall use the DCT to characterize the temporal variations among moving objects in a video sequence.

4.2. Representation of video sequences

In this section, we shall describe how to characterize the temporal variations among moving objects exploiting the DCT. The algorithm that can be exploited to generate video sequence representation is as follows:

Video sequence representation algorithm
Input: Consecutive P-frames $\{P1, P2, P3, \ldots, PN\}$
Output: Sequences of representative DCT coefficients $[Z_{f,j}]$, where $f \in [1, \alpha]$ and $j \in [1, \beta]$
Procedure:

1. For each P-frames P_i,
   Detect moving objects by clustering macroblocks that have similar motion vector magnitudes and similar motion directions.
2. For each object $\mathrm{Obj}_{i,r}$, where i and r denote the rth object in the ith P-frames;
   Compute the centroid and the object size in the unit of macroblocks.
3. Set the number of histogram bins to β.
4. For each P-frames P_i,
   Compute the X-histogram and the Y-histogram according to the horizontal and vertical position of the objects, respectively.
5. For each sequence of histogram bins $[\mathrm{Bin}^Z_{t,j}]$, where $t \in [1, N]$, $j \in [1, \beta]$, and $Z \in \{X, Y\}$
   Compute the transformed sequence $[Z_{f,j}]$ using the discrete cosine transform

$$Z_{f,j} = C(f) \sum_{t=1}^{N} \mathrm{Bin}^z_{t,j} \cos\left(\frac{(2t+1) f \pi}{2N}\right), \quad \text{where } f \in [1, N].$$

6. Set the number of DCT coefficients to α.
7. For β transformed sequences $[Z_{f,j}]$ of DCT coefficients,
   Select the DC coefficient and $(\alpha - 1)$ AC coefficients to represent a transformed sequence.
8. Generate the β reduced low-dimensional sequences $[Z_{f,j}]$, where $f \in [1, \alpha]$ and $j \in [1, \beta]$.

Fig. 4 is the graphical representation of the above algorithm. For each *P*-frames, the feature of the object-based motion activity is described by a 2D-histogram, in which the spatial distribution of moving objects in horizontal and vertical direction are characterized by the bin values of the *X*-histogram and the *Y*-histogram, respectively. Therefore, a video sequence can be represented by a sequence of 2D-histogram with $2N\beta$ dimensions, where N is the number of *P*-frames in a video sequence and β is the number of bins in *X*-histogram and *Y*-histogram. To reduce the dimensionality of the feature space, DCT is exploited to transform the 2D-histogram of the original video sequence into the frequency domain. The value of the *j*th

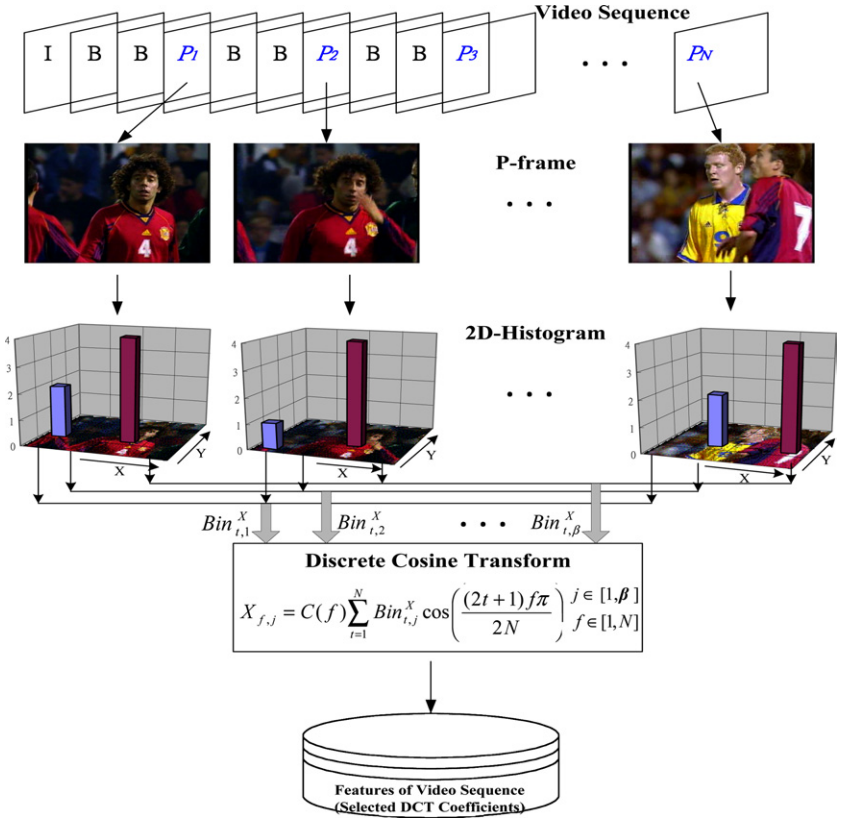

Fig. 4. Video sequences are characterized by the object-based T2D-histogram descriptor and further represented by reduced low-dimensional DCT coefficients.

bin $\mathrm{Bin}^X_{i,j}$ of X-histogram ($\mathrm{Bin}^Y_{i,j}$ of Y-histogram) in the ith P-frames is considered to be a signal in time i, and thus the corresponding jth X-histogram bin in the consecutive N P-frames is regarded as a time signal $x_j = [\mathrm{Bin}^X_{t,j}]$ ($y_j = [\mathrm{Bin}^Y_{t,j}]$ of the Y-histogram), where $t = 1, 2, 3, \ldots, N$. The N-point DCT of a signal x_j is defined as a sequence $\boldsymbol{X} = [X_{f,j}]$, $f = 1, 2, 3, \ldots, N$ as follows:

$$\begin{aligned} X_{f,j} &= C(f)\sum_{t=1}^{N}\mathrm{Bin}^X_{t,j}\cos\left(\frac{(2t+1)f\pi}{2N}\right), \\ C(0) &= \sqrt{\frac{1}{N}} \quad \text{and} \quad C(f) = \sqrt{\frac{2}{N}}, \quad f = 1, 2, \ldots, N-1, \end{aligned} \tag{3}$$

where N is the number of P-frames and $j \in [1,\beta]$. Eq. (3) indicates that a video sequence is represented by β sequences of DCT coefficients restricted by the number of bins in the histogram. It means that temporal variations among original objects in the successive P-frames are characterized by β sequences of DCT coefficients in frequency domain.

It is well known that the first few low-frequency AC terms together with the DC term will suffice for the need. Therefore, for easy computation we only choose these terms to represent a video sequence instead of selecting all coefficients. However, to select an appropriate amount of AC coefficients is always a crucial issue. Since the selection of coefficients is an ill-posed problem, we shall discuss this problem in the experiments.

4.3. Choice of similarity measure

A very important property of Parseval's theorem is that the Euclidean distance between DCT transformed signals is able to maintain the local topology. Therefore, for matching between video sequences we employ the modified Euclidean distance as the metric. Let $[W^X_f]$ and $[H^X_f]$ be two finite point sets of X-histogram ($[W^Y_f]$ and $[H^Y_f]$ of the Y-histogram). Then the modified Euclidean distance between two video sequences w and h is defined as

$$\begin{aligned} \mathrm{Dist}_X(w,h) &= \mathrm{Min}\begin{pmatrix} \mathrm{Dist}_X(W,H), \mathrm{Dist}_X(W,\mathrm{shr}(1,H)), \\ \mathrm{Dist}_X(W,\mathrm{shr}(2,H)), \ldots, \mathrm{Dist}_X(W,\mathrm{shr}(\beta-1,H)) \end{pmatrix} \\ \mathrm{Dist}_Y(w,h) &= \mathrm{Min}\begin{pmatrix} \mathrm{Dist}_Y(W,H), \mathrm{Dist}_Y(W,\mathrm{shr}(1,H)), \\ \mathrm{Dist}_Y(W,\mathrm{shr}(2,H)), \ldots, \mathrm{Dist}_Y(W,\mathrm{shr}(\beta-1,H)) \end{pmatrix}, \end{aligned} \tag{4}$$

where $\mathrm{Dist}_X(W,H) = \sum_{j=1}^{\beta}\sum_{f=1}^{\alpha}(W^X_{f,j} - H^X_{f,j})^2$, $\mathrm{Dist}_Y(W,H) = \sum_{j=1}^{\beta}\sum_{f=1}^{\alpha}(W^Y_{f,j} - H^Y_{f,j})^2$ and W and H are the transformed signals of w and h, respectively. In Eq. (4), j denotes the jth histogram bin, f represents the fth coefficient and α denotes the number of selected DCT coefficients. $\mathrm{shr}(n,H)$ is a bin-rotating function which rotates the β histogram bins to the right n times in a cyclic way. For example, $\mathrm{shr}(1,H)$ shifts the first $(\beta-1)$ bins one time to the right and the last bin rotates from the βth bin to the 1st bin. Using the distance metric with function $\mathrm{shr}(n,H)$, two video sequences will be regarded as similar when they are spatially and tempo-

rally similar. If the function $\mathrm{shr}(n, H)$ were not employed in the distance function, a shot A with objects poisoned in the left and a shot B with objects positioned in the right would be regarded as dissimilar because the peak bins of shots A and B are in the left and right, respectively, and thereby the distance between A and B would be very large.

To further address the overall moving trend of objects within a video sequence, $\mathrm{Dist}_X(w,h)$ and $\mathrm{Dist}_Y(w,h)$ are weighted adaptively based on the average motion vector magnitudes derived from the x- and y-directions. Under these circumstances, the total distance $\mathrm{Dist}_{\mathrm{total}}(w,h)$ between two video sequences w and h can be defined as

$$\mathrm{Dist}_{\mathrm{total}}(w,h) = WT_H \cdot \mathrm{Dist}_X(w,h) + WT_V \cdot \mathrm{Dist}_Y(w,h), \tag{5}$$

$$WT_H = \frac{1}{N}\sum_{i=1}^{N} \frac{MV_{i,H}}{MV_{i,H} + MV_{i,V}}, \quad WT_V = 1 - WT_H,$$

where WT_H is the weight of the X-histogram (WT_V of Y-histogram), N is the number of P-frames, and $MV_{i,H}$ and $MV_{i,V}$ are the average motion vector magnitudes of the X- and Y-component, respectively, of the inter-coded macroblocks in the ith P-frames. The reason why the analysis on object motion is split into two independent directions is as follows. It is well known that a camera would normally pan or tilt to catch moving objects in a scene. This act will in fact result in the situation that the global motion is mainly horizontal (vertical) when most active regions move in the horizontal (vertical) direction. Therefore, it is feasible to use the dominant moving trend to measure the video similarity. For example, we can discriminate between baseball and football videos using the above mentioned similarity metric because most players in a baseball game run vertically and the camera tilts to track them or the baseball, while players in a football game primarily run horizontally and the camera pans to track significant events.

5. Experimental results and discussions

To show the effectiveness of the proposed method, we simulated the color video sequence matching algorithm by MPEG-7 test dataset (ISO/IEC JTC1/SC29/WG11/N2466, 1998), which includes various programs such as documentaries, news, sports, entertainment, education, scenery, interview, etc., and consists of 1173 shots. In the test dataset, the degree of strength of the motions in these shots ranged from low, medium to high, and the size of moving objects were classified as either small, medium or large. The anchorperson shots and interview shots (API shots) are typical low activity shots with small-range motions of mouth and head. The close-up tracking shots (CUT shots) are medium or large activity shots with medium or large-area moving foreground objects. The walking person shots (WP shots) are typical medium activity shots with medium or large motion areas. The aims of the experiments were to (1) evaluate the retrieval performance using different number of DCT coefficients; (2) analyze the degree of accuracy when distinct number of histogram bins

was used in the retrieval process; and (3) evaluate the retrieval performance of the proposed object-based motion activity descriptor. To evaluate the performance of the above three issues, precision and recall were used as the metrics to measure the performance of the proposed retrieval system. Recall and precision were defined as follows:

$$\begin{aligned} \text{Recall} &= \frac{\|\text{Retrieve}(q) \cap \text{Relevant}(q)\|}{\|\text{Relevant}(q)\|}, \\ \text{Precision} &= \frac{\|\text{Retrieve}(q) \cap \text{Relevant}(q)\|}{\|\text{Retrieve}(q)\|}, \end{aligned} \tag{6}$$

where "Retrieve(q)" means the retrieved video sequences that corresponded to a query sequence q; "Relevant(q)" denotes all video sequences in the database that were relevant to a query sequence q and $\|\cdot\|$ indicates the cardinality of the set. Recall was defined as the ratio of the number of retrieved relevant video sequences to the total number of relevant video sequences in the video database, and Precision was defined as the ratio of the number of retrieved relevant video sequences to the total number of retrieved video sequences. In the following subsections, we shall elaborate on how to determine some important thresholds that will be used in the experiments and report the retrieval performance of the proposed system.

5.1. Selecting appropriate number of DCT coefficients

In the experiments, we used four shot classes to test the performance of our algorithms. Among these test videos, the shots of the close-up tracking (CUT) and the walking person (WP) were with high degree of motion. The shots covered in the bicycle racing (BR) and the anchor person (API) were with medium degree of motion and low degree of motion, respectively. Figs. 5A–D show the examples of these four shot types, with key-frames sampled per 40 frames. To evaluate the effect when different number of DCT coefficients was used in the retrieval process, the number of DCT coefficients, α, including the DC and the first $(\alpha - 1)$ AC coefficients, was varied and tested under the condition that the number of histogram bins, β, was set to 8. β was set to 8 because in the test dataset the resolution of frame size in terms of macroblocks was 20×15 in SIF 320×240 format. The descriptors D, the X-histogram, the Y-histogram, the 2D-histogram and the weighted 2D-histogram were independently used.

Figs. 6A–D show the retrieval performance using four different types of shots, CUT, BR, WP, and API, respectively. The four curves shown in the figures corresponded to four descriptors, which had distinct number of DCT coefficients ($\alpha = 1$, 2, 3, and 5). The horizontal axis denotes recall and the vertical axis denotes precision. Table 1 compared the performance among distinct settings of α. "Rank" refers to the order of retrieval performance of recall–precision pairs and the first two ranks were listed for each descriptor measured by using different setting of α. The retrieval performance in the recall–precision pair with $\alpha = 2$ in the CU and BR shots was better than that obtained with other settings. Although the setting of $\alpha = 1$ yielded better retrieval than $\alpha = 2$ in the WP shot, the performance obtained by set-

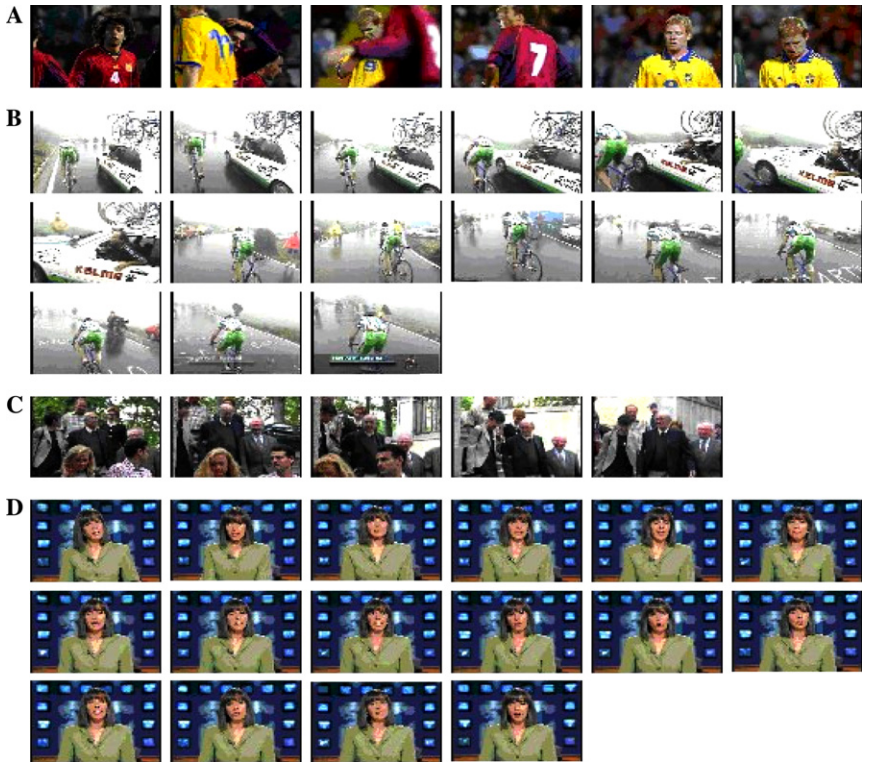

Fig. 5. Examples of the close-up tracking (CUT), bicycle racing (BR), walking person (WP), and anchor person and interview (API) shots.

ting $\alpha = 2$ was still in the second best. For the API shots, the setting $\alpha = 5$ was the best in terms of retrieval and the settings $\alpha = 3$ and $\alpha = 2$ were the second best as shown in Figs. 6A and B and C and D, respectively.

To evaluate the overall performance obtained using different numbers of DCT coefficients, the retrieval performance $P_{\lambda_{\mathrm{NDC}}}$ for different α was determined by

$$P_{\lambda_{\mathrm{NDC}}} = \sum_{i=1}^{|D|} \sum_{j=1}^{|\mathrm{Clips}|} \frac{\rho}{\mathrm{Rank}_{i,j}^{\lambda_{\mathrm{NDC}}}}, \tag{7}$$

where "NDC" denotes the "Number of DCT Coefficients"; ρ is the total number of different α settings in the experiment and $\mathrm{Rank}_{i,j}^{\lambda_{\mathrm{NDC}}}$ is the ranking of the retrieval performance for the shot of type j with $\alpha = \lambda_{\mathrm{NDC}}$, using descriptor i. When $P_{\lambda_{\mathrm{NDC}}}$ was larger, the performance obtained with $\alpha = \lambda_{\mathrm{NDC}}$ was better. From the curves shown in Figs. 6A–D, it is clear that P_2 can be computed and its value was larger than other P values. This outcome means when $\alpha = 2$, the retrieval result was the best. Hence, the experimental results imply that two DCT coefficients are enough for similarity measurement of video segments. This indicates the DC coefficient and the lowest-frequency AC coefficient will suffice.

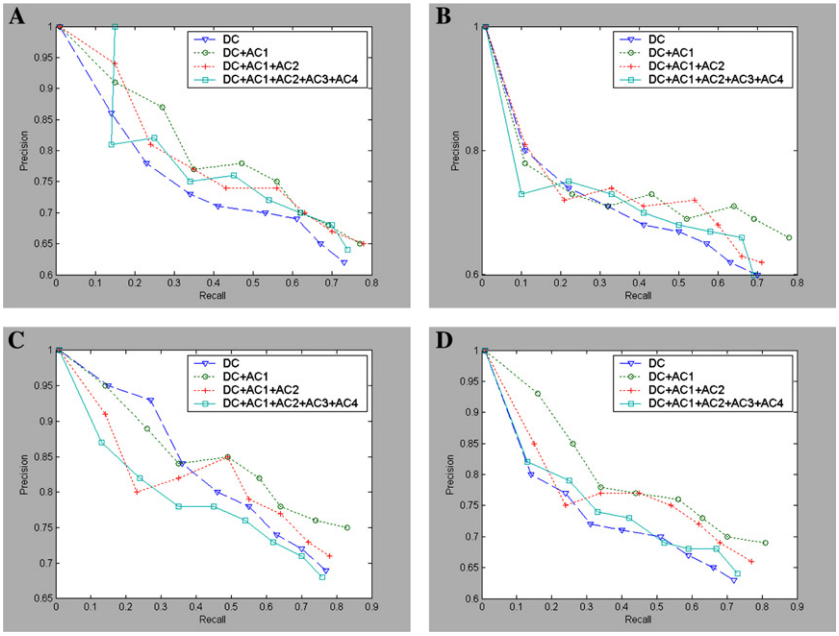

Fig. 6. Average retrieval performance with different descriptors ($\beta = 8$, $\alpha \in [1,5]$): (A) X-histogram, (B) Y-histogram, (C) 2D-histogram, and (D) weighted 2D-histogram.

Table 1
Performance using distinct α and four feature descriptors ($\beta = 8$)

Descriptor		Shot type			
		Close-up tracking (CUT)	Bicycle racing (BR)	Walking person (WP)	Anchor person (API)
X-histogram	Rank #1	2	2	1	5
	Rank #2	3	3	2	3
Y-histogram	Rank #1	2	2	1	5
	Rank #2	1	3	2	3
2D-histogram	Rank #1	2	2	1	5
	Rank #2	3	3	2	2
Weighted 2D-histogram	Rank #1	2	2	1	5
	Rank #2	3	3	2	2

5.2. Choosing an appropriate motion activity descriptor

To determine an appropriate motion activity descriptor, we changed the value of β from 4 to 10, each time with an increment of 2. Figs. 7A–D show, respectively, the performance of the recall–precision pair corresponding to $\beta = 4$, 6, 8, and 10. Table 2 illustrates the performance calculated by using four different number of histogram

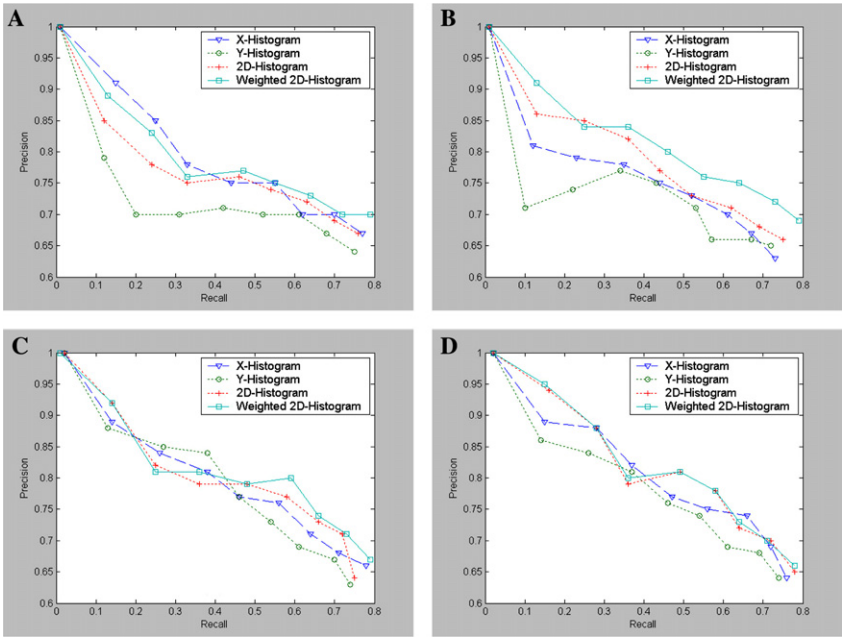

Fig. 7. Average retrieval performance ($\alpha = 2$) with different number of bins (β): (A) $\beta = 4$, (B) $\beta = 6$, (C) $\beta = 8$, and (D) $\beta = 10$.

Table 2
The performance obtained of four descriptors with different $\beta(\alpha = 2)$

β Setting		Shot type			
		Close-up tracking (CUT)	Bicycle racing (BR)	Walking person (WP)	Anchor person (API)
$\beta = 4$	Rank #1	X	X	W-2D	W-2D
	Rank #2	W-2D	W-2D	X	2D
$\beta = 6$	Rank #1	W-2D	Y	X	W-2D
	Rank #2	X	W-2D	W-2D	2D
$\beta = 8$	Rank #1	X	W-2D	W-2D	W-2D
	Rank #2	W-2D	2D	2D	2D
$\beta = 10$	Rank #1	W-2D	W-2D	W-2D	X
	Rank #2	2D	2D	2D	2-2D

X, *X*-histogram; Y, *Y*-histogram; 2D, 2D-histogram; and W-2D, Weighted 2D-histogram.

bins ($\beta = 4$, 6, 8, and 10). In most cases, the descriptor adopted weighted 2D-histogram outperformed other types of descriptors. To quantitatively compute the performance, we used a metric, P_{λ_D} , to measure the retrieval results

$$P_{\lambda_D} = \sum_{i=1}^{|\beta|} \sum_{j=1}^{|\text{Clips}|} \frac{|D|}{\text{Rank}_{i,j}^{\lambda_D}}, \tag{8}$$

where $|\beta|$ denotes the total number of distinct settings of β; $|D|$ represents the number of testing descriptors; $\text{Rank}_{i,j}^{\lambda_D}$ is the retrieval performance raking of the shot of type j with the ith β parameter setting and the descriptor $D = \lambda_D$. Based on the results calculated by Eq. (8), we chose the weighted 2D-histogram descriptor as the motion activity descriptor for all the experiments conducted in this work.

5.3. Determining the best number of histogram bins

In this section, we shall verify the decision of the number of histogram bins β. Therefore, we evaluated the performance by using different number of histogram bins, which ranged from 4, 6, 8 to 10. The recall–precision pair corresponding to each β setting was depicted in Fig. 8, and the ranking of retrieval performance for each shot type was illustrated in Table 3.

It is obvious that the retrieval performance at $\beta = 8$ decided by Eq. (1) was better than other settings and the worst case was when $\beta = 4$. The experimental results

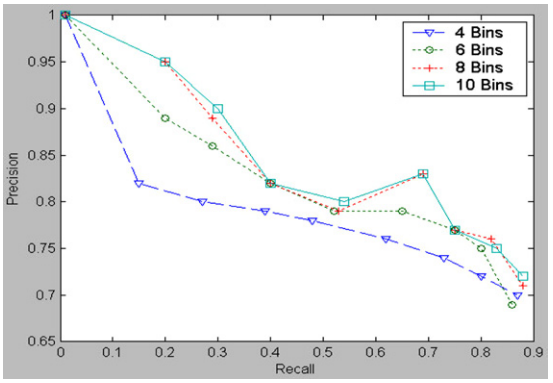

Fig. 8. Average retrieval performance with parameters: $\alpha = 2$, D: weighted 2D-histogram, $\beta \in \{4, 6, 8, 10\}$.

Table 3
Comparison of performance using different numbers of histogram bins (β)

Performance	Shot type			
	Close-up tracking (CUT)	Bicycle racing (BR)	Walking person (WP)	Anchor person (API)
Rank #1	6	8	8	8
Rank #2	10	10	10	10
Rank #3	8	6	6	6
Rank #4	4	4	4	4

reveal that the number of histogram bins should be moderate, because fewer histogram bins correspond to a less precise description of the variation in spatial distribution. In contrast, when the number of histogram bins was too large, the descriptor would be extremely responsive to the slight changes. Under this circumstance, the distance obtained from excessive number of bins between two similar shots is relatively high such that these two shots would be regarded as dissimilar.

5.4. Evaluation of retrieval performance

After the number of DCT coefficients, the number of histogram bins and the descriptor type are determined, we shall evaluate the overall retrieving accuracy of the proposed system. The ground truth and the overall performance corresponding to the four shot classes are shown in Table 4. In the experiment, each shot in these four classes was used as a query shot. The top 30 similar shots were returned as a query result for evaluating retrieval performance. Finally, the respective average recall and precision for each class were computed. The recall of these four kinds of shots exceeded 80% in which the recall of BR, CUT, and API were higher than 86%. The worst result was obtained by testing the API shots, with the precision of 78%. On the other hand, although the precision of the API shots was under 80%, the precision of the CUT, BR, and WP all exceeded 80%. From Table 4, the overall average recall and average precision were 86 and 81%, respectively.

For performance comparison, we have performed the same experiments using the algorithms of motion-based run-length descriptor (RLD) and shot activity histogram (SAH) provided by MPEG-7 (ISO/IEC JTC1/SC29/WG11/N4547, 2001). Fig. 9 shows the precision versus recall performance of RLD, SAH, and T2D-histogram. The T2D-histogram descriptor had performance gain over RLD of 45% in API shots, 30% in the CUT shots, 34% in the WP shots, and 35% in the BR shots. Also, the T2D-histogram had performance gain over SAH of 11% in the API shots, 7% in the CUT shots, 20% in the WP shots, and 21% in the BR shots. In average, the T2D-histogram descriptor had 37% and 15% performance gains over the RLD and SAH, respectively. The experimental results using extensive test videos show that the proposed T2D-histogram outperforms RLD and SAH in MPEG-7 in the performance of video similarity retrieval.

Table 4
Retrieval performance using the T2D-histogram descriptor

Performance	Clips			
	Close-up tracking (CUT)	Bicycle racing (BR)	Walking person (WP)	Anchor person (API)
Ground-truth video shots	162	47	239	152
Recall	88%	87%	80%	86%
Precision	80%	84%	81%	78%
	Average recall 86%		Average precision 81%	

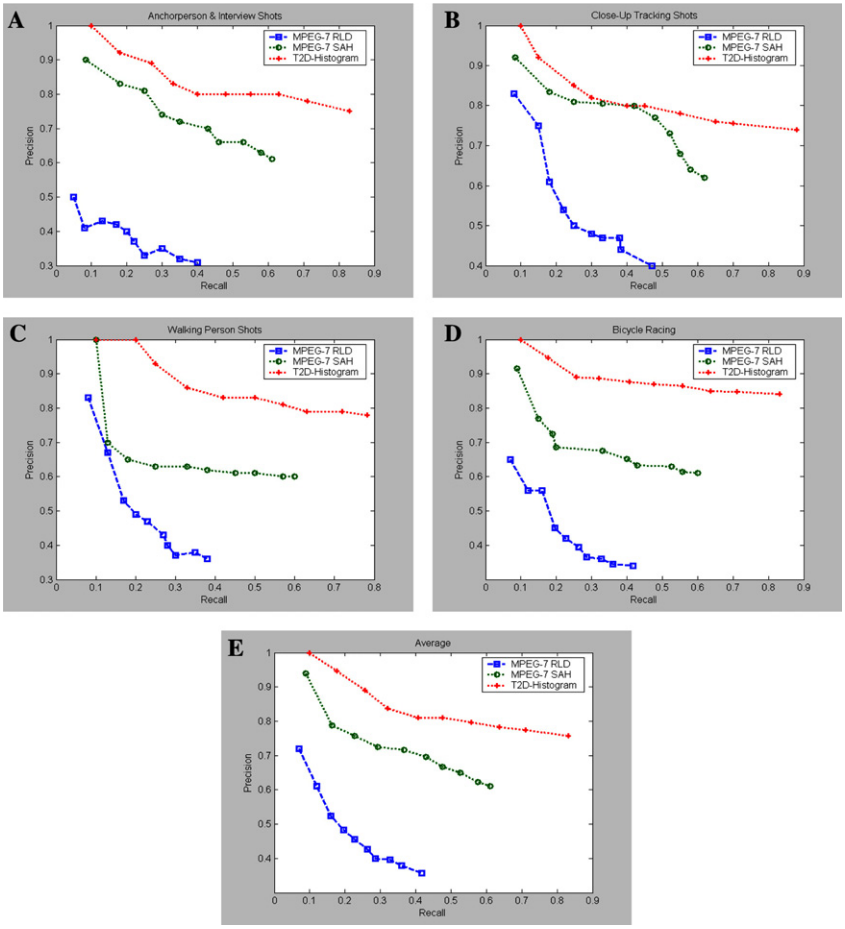

Fig. 9. Retrieval performance of the four shot classes: (A) API shots, (B) CUT shots, (C) WP shots, (D) BR shots, and (E) average.

Examples of the query results were demonstrated in Figs. 10–13, in which the top 20 similar shots for CUT, BR, WP, and API shots were listed, respectively. In Fig. 10, most retrieved shots included large objects with significant motion belonged to the CUT shots. However, due to camera motion, some shots were mistakenly detected. For example, the full-court shots of the football game like (4), (8), and (12) of Fig. 10 were retrieved due to the panning effect of the camera. As to the relevant shots, it is worth noticing that the major objects in these shots, such as (3), (18), and (19) of Fig. 10, had similar size with the object covered in the query although they had different colors. The reason why these shots could still be detected was due to their similarity with the objects in the query visually and semantically. When comparing with color-based methods such as color histogram, these shots with distinct dominant colors but semantically related cannot be retrieved.

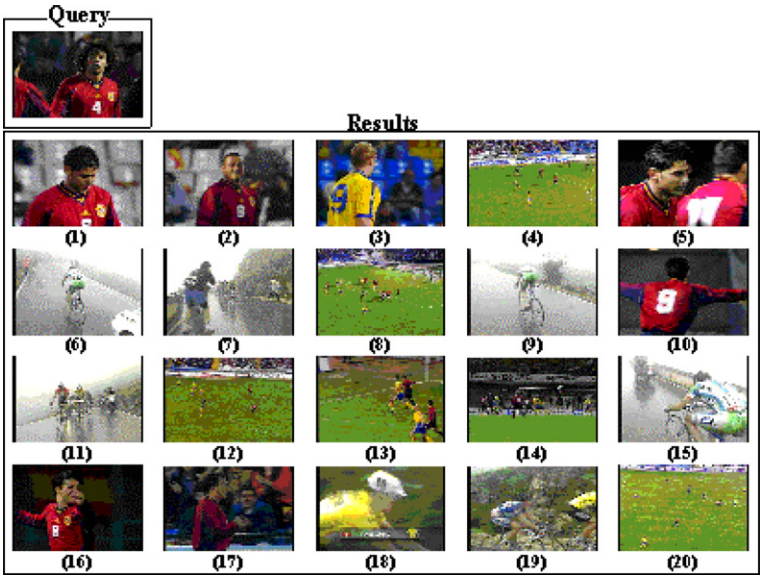

Fig. 10. Demonstration of the query result for a CUT shot.

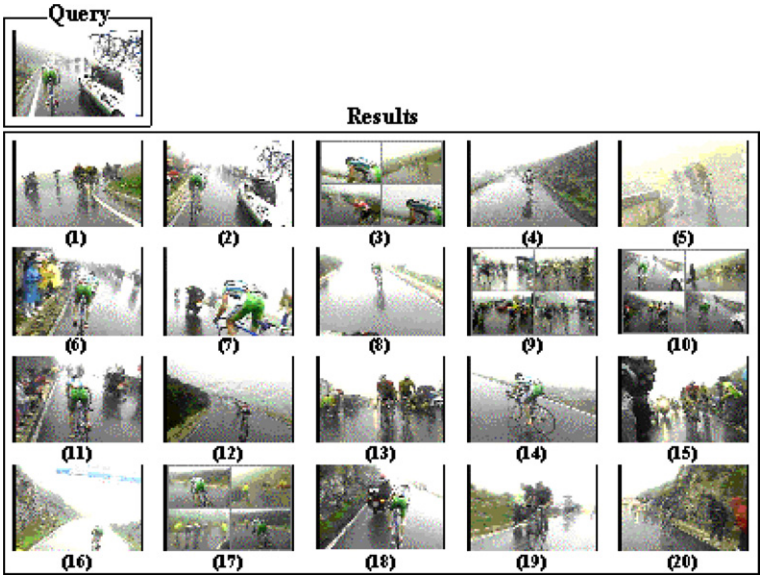

Fig. 11. Demonstration of the query result for a BR shot.

In Fig. 11, the retrieval performance of the BR shot was quite good and most retrieved video segments were similar to the query due to the particular motion of the rider(s). In Fig. 12, most retrieved video segments had a few medium-size moving

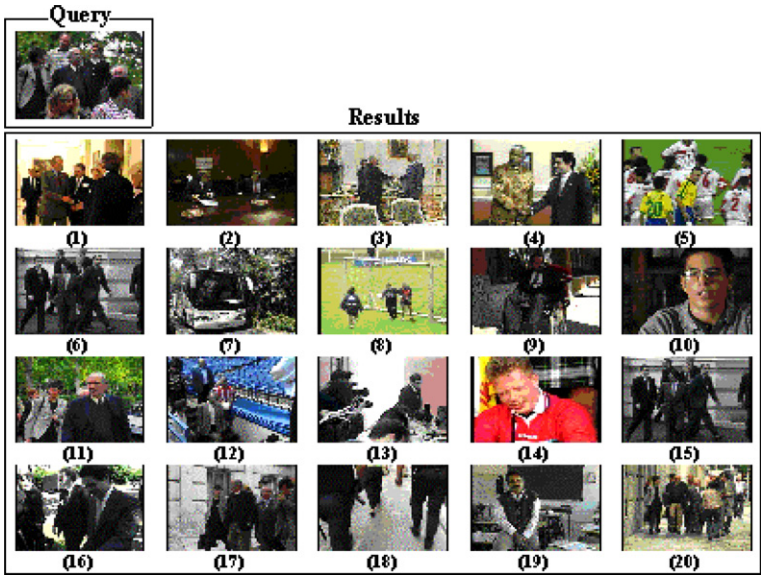

Fig. 12. Demonstration of the query result for a WP shot.

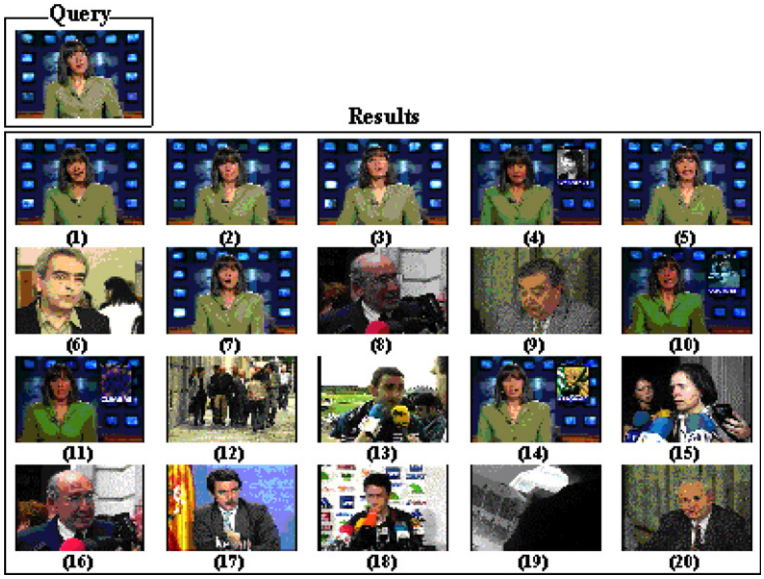

Fig. 13. Demonstration of the query result for an API shot.

objects. Some video segments were mistakenly detected, such as (10) and (14) of Fig. 12. These shots were retrieved due to the reason that the complex background was detected as several medium-size objects with a moving camera. In Fig. 13, most

retrieved video segments included one large object with low motion, and so interview shots were also retrieved such as the shots (6), (8), (13), (16), and (20) of Fig. 13. An example of false detection can be found in (12) of Fig. 13, wherein some medium-size objects moved near to each other and so were incorrectly detected as a single large moving object.

6. Conclusions

A novel framework of high-level video representation for video sequence matching has been proposed in this work. The proposed framework has two special features: (1) the proposed descriptor of object-based T2D-histogram has exploited both spatial and temporal features of moving objects and characterized video sequences in a semantics-based manner; (2) the dimensionality of feature space has been reduced using DCT while characterizing the temporal variations among moving objects. Experimental results obtained using the extensive test dataset of MPEG-7 have demonstrated that a few DCT coefficients could suffice for representing a video sequence and also shown that the proposed T2D-histogram descriptor was quite robust. Using this novel motion activity descriptor of object-based T2D-histogram, one can perform video retrieval in an accurate and efficient way.

References

Aghbari, Z., Kaneko, K., Makinouchi, A., 1998. A motion-location based indexing method for retrieving MPEG videos. In: Proceedings of the Ninth International Workshop on Database and Expert Systems Applications, pp. 102–107.

Agnihotri, L., Dimitrova, N., 2000. Video clustering using superhistograms in large archives. In: Proceedings of the Fourth International Conference on Visual Information Systems, Lyon, France, pp. 62–73.

Chen, D.Y., Lee, S.Y., 2001. Motion-based semantic event detection for video content descriptions in MPEG-7. In: Proceedings of the second IEEE Pacific-Rim Conference on Multimedia, pp. 110–117.

Chen, D.Y., Chen, H.T., Lee, S.Y., 2002. Motion activity based semantic video similarity retrieval. In: Proceedings of IEEE Third Pacific Rim Conference on Multimedia, Hsinchu, Taiwan, pp. 319–327.

Cheung, S.S., Zakhor, A., 2001. Video similarity detection with video signature clustering. In: Proceedings of the IEEE International Conference Image Processing, vol. 2, pp. 649–652.

Divakaran, A., Peker, K., Sun, H., 2000. A region based descriptor for spatial distribution of motion activity for compressed video. In: Proceedings of the International Conference Image Processing, vol. 2, pp. 287–290.

ISO/IEC JTC1/SC29/WG11/N2466, 1998. Licensing Agreement for the MPEG-7 Content Set, Atlantic City, USA, 1998.

ISO/IEC JTC1/SC29/WG11/N4547, 2001. Extraction and Use of MPEG-7 Descriptions, Pattaya.

Jeannin, S., Divakaran, A., 2001. MPEG-7 visual motion descriptors. IEEE Trans. Circuits Syst. Video Technol. 11 (6), 720–724.

Lee, S.Y., Lian, J.L., Chen, D.Y., 2001. Video summary and browsing based on story-unit for video-on-demand service. In: Proceedings of the Third International Conference on Information, Communications and Signal Processing, Singapore.

Lin, T., Ngo, C.W., Zhang, H.J., Shi, Q.Y., 2001. Integrating color and spatial features for content-based video retrieval. In: Proceedings of the IEEE International Conference Image Processing, vol. 2, pp. 592–595.

Manjunath, B.S., Ohm, J.R., Vasudevan, V.V., Yamada, A., 2001. Color and texture descriptors. IEEE Trans. Circuits Syst. Video Technol. 11 (6), 703–715.

Mohan, R., 1998. Video sequence matching. IEEE Int. Conf. Acoustics Speech Signal Process. 6, 3697–3700.

Oppenheim, A.V., Schafer, R.W., 1999. Discrete-Time Signal Processing. Prentice-Hall, New Jersey.

Peker, K.A., Alatan, A.A., Akansu, A.N., 2000. Low-level motion activity features for semantic characterization of video. In: Proceedings of the IEEE International Conference Image Processing, vol. 2, pp. 801–804.

Roach, M., Mason, J.S., Pawlewski, M., 2001. Motion-based classification of cartoons. In: Proceedings of the International Symposium on Intelligent Multimedia, Video and Speech Processing, Hong Kong, pp. 146–149.

Shi, Y.Q., Sun, H., 2000. Image and Video Compression for Multimedia Engineering. CRC Press, New York.

Sikora, T., 2001. The MPEG-7 visual standard for content description—an overview. IEEE Trans. Circuits Syst. Video Technol. 11 (6), 696–702.

Tang, Y.P., Saur, D.D., Kulkarni, S.R., Ramadge, P.J., 2000. Rapid estimation of camera motion from compressed video with application to video annotation. IEEE Trans. Circuits Syst. Video Technol. 10 (1), 133–146.

Wang, R., Naphade, M.R., Huang, T.S., 2001. Video retrieval and relevance feedback in the context of a post-integration model. In: Proceedings of the IEEE Fourth Workshop on Multimedia Signal Processing, pp. 33–38.

Wang, R., Zhang, H.J., Zhang, Y.Q., 2000. A confidence measure based moving object extraction system built for compressed domain. In: Proceedings of the IEEE International Symposium on Circuits and Systems, vol. 5, pp. 21–24.

Yeung, M.M., Liu, B., 1995. Efficient matching and clustering of video shots. In: Proceedings of the IEEE International Conference Image Processing, vol. 1, pp. 338–341.

Zhao, L., Qi, W., Li, S.Z., Yang, S.Q., Zhang, .J., 2001. Content-based retrieval of video shot using the improved nearest feature line method. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, vol. 3, pp. 1625–1628.