Semantic Video Content Structuring and Visualization

Duan-Yu Chen†, Shu-Jiuan Lin and Suh-Yin Lee
Department of Computer Science and Information Engineering
National Chiao-Tung University, 1001 Ta-Hsueh Rd, Hsinchu, Taiwan
{dychen†, shujiuan, sylee}@csie.nctu.edu.tw
Telphone: 886-3-5731678, Fax: 886-3-5724176

Abstract

In this paper, we propose a novel approach to visualize video content based on shot description of motion activity and textual information of closed caption in MPEG-2 sports videos. In order to speed up in scene change detection, GOP-based approach first checks video streams GOP by GOP and then finds out the actual scene boundaries in the frame level. Segmented shots are described by the proposed object-based motion activity descriptor. The descriptor is computed based on the object 2D-histogram, in which long-term consistency of spatial-temporal relationship of moving objects within video shots is considered. Utilizing the characterized features of motion activity in video shots, video clips are recognized by the proposed algorithm of shot identification. Subsequently, the specific shots of interest are selected and the proposed mechanism of closed caption localization is exploited to detect captions in these shots. Moreover, the SOM (Self-Organization Map) based algorithm is designed as a filter to distinguish the superimposed closed captions from the high-textured background regions. Finally, we construct a sports video content visualization system and provide the table of video content composed of the hierarchical structure of story unit, consecutive shots and closed captions. Furthermore, summarization mechanism – the dynamic tree structure of video content is provided. The experimental results show the effectiveness of the proposed system and reveal the feasibility of the hierarchical structuring of video content.

Keywords: *Video segmentation, motion activity, shot identification, closed caption, video structuring*

† Please address all correspondence related to this manuscript to Duan-Yu Chen

1. Introduction

The tremendous growth in the amount of digital videos is driving the need for more effective methods to access and acquire desired video data. Advances in automatic content analysis and feature extraction improve capabilities for effective browsing, searching and filtering videos along with perceptual features and semantics. Content-based indexing provides users natural and friendly querying, searching, browsing and retrieving. In order to provide users more efficient and effective access methods, it is necessary to support high-level and semantic features for video content representation and indexing. The need of representation and indexing for high-level and semantic features motivates the emerging standard MPEG-7, formally called multimedia content description interface [22]. However, the methods that produce the desired features are non-normative part of MPEG-7 and are left open for research and future innovation. In order to present visual content in more compact forms and to support users to comprehend the video abstract, browse through video sequences and navigate from one clip to another, we need to structure videos by utilizing high-level and semantic features. In the research of video structuring, Lu and Tang [6] proposed a video-structuring scheme, which classifies and clusters sports video shots by low-level features, color and global motion information. Kwon et al [16] proposed a scene segmentation scheme based on adaptive weighting of color and motion features, and merge scene units by using the improved overlapping link scheme. Hanjalic and Lagendijk [15] segment movies into logical story units based on the global temporal consistency of its visual content characterized by the color feature - MPEG-DC sequences. Yeung and Yeo [17] proposed a time-constrained and MPEG DC based visual similarity clustering approach to segment a video into logical story units. These researches structure videos by using low-level features, like color and motion, and perform shots merging to generate logical story units based on visual similarity by applying some time-constrained mechanism.

However, video shots classification and scene merging based on object or event

information would be more semantically meaningful. Therefore, in this paper, we use the proposed object-based motion activity descriptor for MPEG-7 to classify sports video shots and to infer video events further. In addition, to present sports video content in a more compact form, we automatically localize the viewer-interested closed caption – the scoreboard in compressed domain for structuring videos semantically. Recently, increasing researches focus on feature extraction in compressed video domain, especially in MPEG format because many videos are already being stored in compressed form due to mature compression techniques. Edge features are directly extracted from MPEG compressed videos to detect scene change [5] and captions are processed and inserted into compressed video frames [7]. Features, like chrominance and shape are directly extracted from MPEG videos to detect face regions [1][3].

There is an increasing trend in the research of text caption localization to detect captions in video frames either in uncompressed domain [12-13], [18-21] or in compressed videos [2], [4], [14]. Li et al [12] exploited a neural network trained on texture features to find text regions and a text-region tracker was also proposed for moving text tracking. Shim et al [13] segment text areas by chaincodes in pixel domain and exploit temporal information to refine segmented text. Both Li et al [12] and Shim et al [13] involved in the examination of the similarity of text regions in terms of their positions, intensities and shape features. Chen and Zhang [18] detect text areas using vertical edge information followed by horizontal edge information and apply Bayesian based shape suppression technique for result refinement. Ohya et al [19] segment characters by local threshold setting and merge neighboring regions based on the similarity of gray levels. Kannangara et al [20] extract text in specific areas and proposed a method based on vertical projection profile to segment individual letters. Wu et al [21] segment text areas exploiting both multiscale texture segmentation and spatial cohesion constraint in pixel domain. Zhong et al [2] and Zhang and Chua [4] perform text localization in MPEG videos using DCT AC coefficients to obtain texture information in individual

I-frames. In addition, Zhang and Chua [4] identify text regions with size filter. Gargi et al [14] perform text detection by counting the number of intra-coded blocks in P and B frames based on the assumption that the background is static and a threshold for size of text segments is also predefined for noise filtering. From the previous researches, there are little efforts to localize text captions in compressed videos. Besides, text region localization using size filter may not work well especially for the cases that captions though are small in size but are very important and meaningful to viewers. For example, in sports videos, the size of scoreboard is generally very small but significant showing the competition going on as clearly as possible without interference. Moreover, once potential text regions are localized, few researches focus on captions filtering, for instance, to separate the superimposed captions from the high-textured regions of the background. Therefore, both text identification with no size constraint and captions filtering are our concerns.

Hence, in this paper, in order to support high-level semantic retrieval of video content, we propose a novel approach that structures videos utilizing closed captions and object-based motion activity descriptors. The mechanism consists of four components: GOP-based video segmentation, shot identification, closed caption detection and video structuring. The rest of the paper is organized as follows. Section 2 presents the overview of the proposed scheme. Section 3 shows the GOP-based scene change detection and Section 4 describes the motion activity based shot identification. The component of closed caption localization is introduced in Section 5. Section 6 illustrates the experimental results and the conclusion and the future works are given in Section 7.

2. Overview of the Proposed Scheme

Fig. 1 shows the mechanism of the proposed system architecture. The testing videos are formatted in MPEG-2 and the volleyball sport is selected as the case study. First, video streams are segmented into shots by using our proposed GOP-based scene change detection [8]. Instead of frame by frame, this module of video segmentation checks video streams GOP

by GOP and then finds out the actual scene change boundaries in the frame level. The segmented shots are described and identified by utilizing the proposed motion activity descriptor in MPEG-7 [7]. Thus the type of each shot of the volleyball videos can be recognized and encoded in the descriptor. The structure of volleyball videos consists of various types of shots, "service" shot, "full-court view" shot and "close-up" shot and the service shots usually are the leading shots in a volleyball competition. Thus, our focus is to recognize these three kinds of shots and select service shots to localize the closed caption to support high-level video structuring. Therefore, the module of motion activity based shot identification is used to distinguish the types of shots and the shots of interest can be automatically recognized and selected for further analysis. Furthermore, the module of closed caption localization is designed based on the concept of SOM (Self-Organization Map) to localize the superimposed scoreboard, of which the caption size is fairly small. The text in the localized closed caption of scoreboard can be used to support high-level video structuring. Finally, the table of video content is created by the key frames that contain the scoreboard and also by the semantic shots identified by the motion activity descriptor.

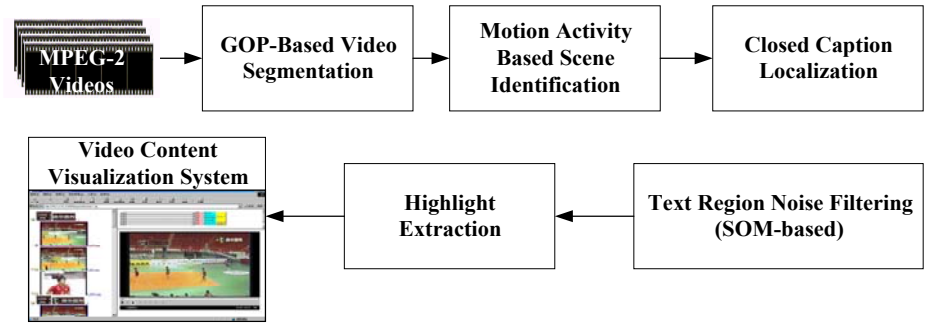


Fig.1. The system architecture of motion activity based video structuring

3. Scene Change Detection

Video data is segmented into meaningful clips to serve as logical units called "shots" or "scenes". In MPEG-II format [9], GOP layer is random accessed point and contains GOP header and a series of encoded pictures including I, P and B-frame. The size of a GOP is about 10 to 20 frames, which is usually less than the minimum duration of two consecutive

scene changes (about 20 frames) [10].

We first detect possible occurrences of scene change GOP by GOP (inter-GOP). The difference between each consecutive GOP-pair is computed by comparing the I-frames in each consecutive GOP-pair. If the difference of DC coefficients between these two I-frames is larger than the threshold, then there may have scene change in between these two GOPs. Hence, the GOP that contains the scene change frames is located. In the second step – intra GOP scene change detection, we further use the ratio of forward and backward motion vectors to locate the actual frame of scene change within a GOP. The experimental results on real long videos in [8] are encouraging and prove that the scene change detection is efficient for video segmentation.

4. Shot Identification

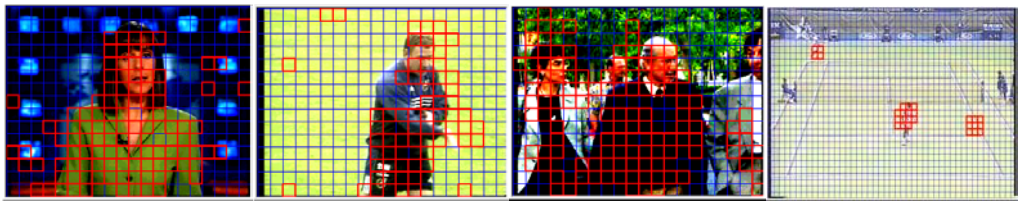
In this section, the approach of shot identification based on object motion activity is introduced. The method of detection of significant moving objects is illustrated in Subsection 4.1 and Subsection 4.2 shows the motion activity descriptor. Scene identification based on the descriptor is presented in Subsection 4.3.

4.1 Moving Object Detection

For the computation efficiency, only the motion vectors of P-frames are used for object detection since in general, in a video with 30 fps consecutive P-frames separated by two or three B-frames are still similar and would not vary too much. Therefore, it is sufficient to use the motion information of P-frames only to detect moving objects. However, the motion vectors of P-frames or B-frames via motion estimation in MPEG-2 may not exactly represent the actual motions in a frame. For a macroblock, a good match is found among its neighbors in the reference frame. However, this motion estimation does not mean that a macroblock does match exactly the correct position in its reference frame.

Hence, in order to achieve more robust analysis, it is necessary to eliminate noisy motion vectors before the process of motion vectors clustering. Motion vectors of relatively small or

approximately zero magnitude are recognized as noise and hence are not taken into account. On the contrary, motion vectors with larger magnitude are more reliable. In the consideration of low computation complexity, the average of motion vectors of inter-coded macroblocks is computed and selected as the threshold to filter out motion vectors of smaller magnitude or noise. While noisy motion vectors are filtered out, motion vectors of similar magnitude and direction are clustered into the same group (the object) by applying the region growing approach. Some examples of moving object detection in MPEG videos [7] are shown in Fig. 2 and each square represents a macroblock.



(a) (b) (c) (d)

Fig. 2. Demonstration of the moving objects detection (a) anchor person shot (b) football shot (c) walking person shot (d) tennis competition shot

Video shots shown in Fig. 2(a) to Fig. 2(c) are extracted from the MPEG-7 test data set and the shot of tennis competition is recorded from the TV-channel of Star-Sports. In Fig. 2, we can see that moving objects can be successfully detected. While some noise is still present in the result of object detection, we can eliminate the noise by object tracking in the forward and backward frames.

4.2 Motion Activity Descriptor – 2D Histogram

2D-histogram is computed for each P-frame. The horizontal axis of the X-histogram (Y-histogram) is the quantized X-coordinate (Y-coordinate) in a P-frame. In the experiments, the X- and Y-coordinates are quantized into *a* and *b* bins according to the aspect ratio of the frame.The object size is estimated before bin assignment. If the object size is larger than the predefined unit size (*frame-size/a*b*), the object is weighted and accumulated by Eq. (1). $Bin^{x}_{i,j}$ means the j^{th} bin of X-histogram in frame *i*. $Acc^{x}_{i,j,\alpha}$ means the accumulated

value of the j^{th} bin of *object* α in *frame i* for X-histogram, and *Obj* is the number of objects in frame *i*.

$$Bin_{i,j}^{x} = \sum_{\alpha=1}^{Obj} Acc_{i,j,\alpha}^{x}, \quad Acc_{i,j,\alpha}^{x} = \begin{cases} 1, & if\ object\ size \le \dfrac{1}{a*b}\ frame\ size \\ \dfrac{size\ of\ object\ \alpha}{frame\ size} * a * b, & otherwise \end{cases} \tag{1}$$

By utilizing the statistics of the 2D-histogram, spatial distribution of moving objects in each P-frame is characterized. In addition, spatial relationships within the moving objects are also approximately shown in the X-Y histogram pair since each moving object is assigned to the histogram bin according to the X-Y coordinate of its centroid position. Objects belong to the same coordinate interval are grouped into the same bins, and hence the distance between two object groups can be represented by the differences between the associated bins.

4.3 Shot Identification

Some representative frames of the shots of "service", "full-court view" and "close-up" are shown in Fig. 3. From the object-based X-histograms of the shots of service, full-court view and close-up individually shown from Fig. 3(d) to Fig. 3(f), we can observe that the characteristic of the service shots is that one or few objects appear in the left or the right side of the frame and more objects appear in the other side of the frame. In the shot of full-court view, generally the number of objects of the left part and the number of objects of the right part are balanced and the difference of the number of objects between them is relatively smaller than that of service shot. In the closed-up shots, there is a large object near the middle position of the frame. Therefore, based on the observation, we can distinguish these major shots of volleyball videos. In the algorithm, we use the X-histogram only to be the descriptor of video shots. Since these three kinds of shots can be distinguished according to the variation of the number of objects in the horizontal axis, i.e. in the X-coordinate.

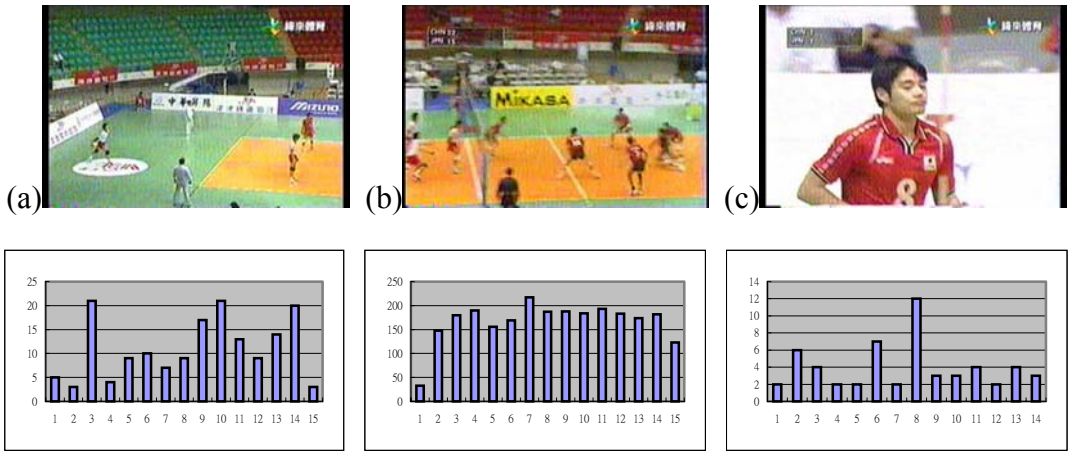


Fig. 3. Key frames of shots (a) Service (b) Full-court view (c) Closed-up

5. Closed Caption Localization

To correctly locate closed captions, we first compute the horizontal gradient energy to filter out some noise by using the DCT AC coefficients. The next step is to remove some noisy regions by the morphological operation. While the candidate caption regions are detected, we utilize the SOM-based algorithm to filter out non-caption regions. The details of the closed caption detection are described in Subsection 5.1 and the algorithm of SOM-based filtering is shown in Subsection 5.2.

5.1 Closed Caption Detection

While service shots are identified, we apply the proposed closed caption detection to further localize the closed caption in these shots, like the scoreboard and the channel trademark. We use the DCT AC coefficients shown in Fig. 4 to compute the horizontal and vertical gradient energy. We use the horizontal AC coefficients from $AC_{0,1}$ to $AC_{0,7}$ to compute the horizontal gradient energy by Eq. (2). Based on the observation that some blank space usually appears between consecutive letters in closed captions, thus the variation of gradient energy in the horizontal direction would be larger than that in the vertical direction. Therefore, the horizontal gradient energy E_h for each 8 x 8 block is exploited as the filter for noise elimination. Moreover, if E_h of a block is greater than a predefined threshold, the

block is regarded as a potential caption block. Otherwise, if E_h of a block is smaller than the threshold, the block is removed.

$$E_h = \sum_{j=1}^{7} \left| AC_{0,j} \right| \tag{2}$$

However, different shots may have different lighting conditions which will reflect in the contrast in frames even over the whole shot. Furthermore, different contrast will affect the decision of the threshold and the result of closed caption detection might fail due to this reason. Therefore, we adopt adaptive threshold decision to overcome this problem. The threshold T is adaptable and is computed by Eq. (3), where γ is a factor that can be adjusted, $SVar_s$ represents the average of horizontal gradient energy of shot s, $FVar_{s,i}^{AC}$ means the horizontal gradient energy of frame i in shot s and AC_h is the horizontal DCT ac coefficient from $AC_{0,1}$ to $AC_{0,7}$. Based on the fact that higher value of $FVar_{s,i}^{AC}$ means the higher contrast of the frame I, and we can thus remove noisy regions more easily in the frame of higher gradient energy. Therefore, we set lower weight to the frame with higher contrast and set higher weight to that with lower contrast. By this way, we can remove most of the noisy regions and an example is demonstrated in Fig. 5(b).

$$T = \gamma \times SVar_s, \quad \gamma = \begin{cases} 3.2, when\ FVar_{s,i}^{AC} < SVar_s \\ 2.4, when\ FVar_{s,i}^{AC} \geq SVar \end{cases}, \quad SVar_s = \frac{1}{M} \sum_{i=1}^{M} FVar_{s,i}^{AC} \tag{3}$$

$$FVar_{s,i}^{AC} = \sum_{j=1}^{N} \sum AC_{h,j}^{2} / N - (\sum_{j=1}^{N} \sum AC_{h,j} / N)^2$$

DC	$AC_{0,1}$	$AC_{0,2}$	$AC_{0,3}$	$AC_{0,4}$	$AC_{0,5}$	$AC_{0,6}$	$AC_{0,7}$
$AC_{1,0}$							
$AC_{2,0}$							
$AC_{3,0}$							
$AC_{4,0}$							
$AC_{5,0}$							
$AC_{6,0}$							
$AC_{7,0}$							

Fig. 4. DCT ac coefficients used in text caption detection

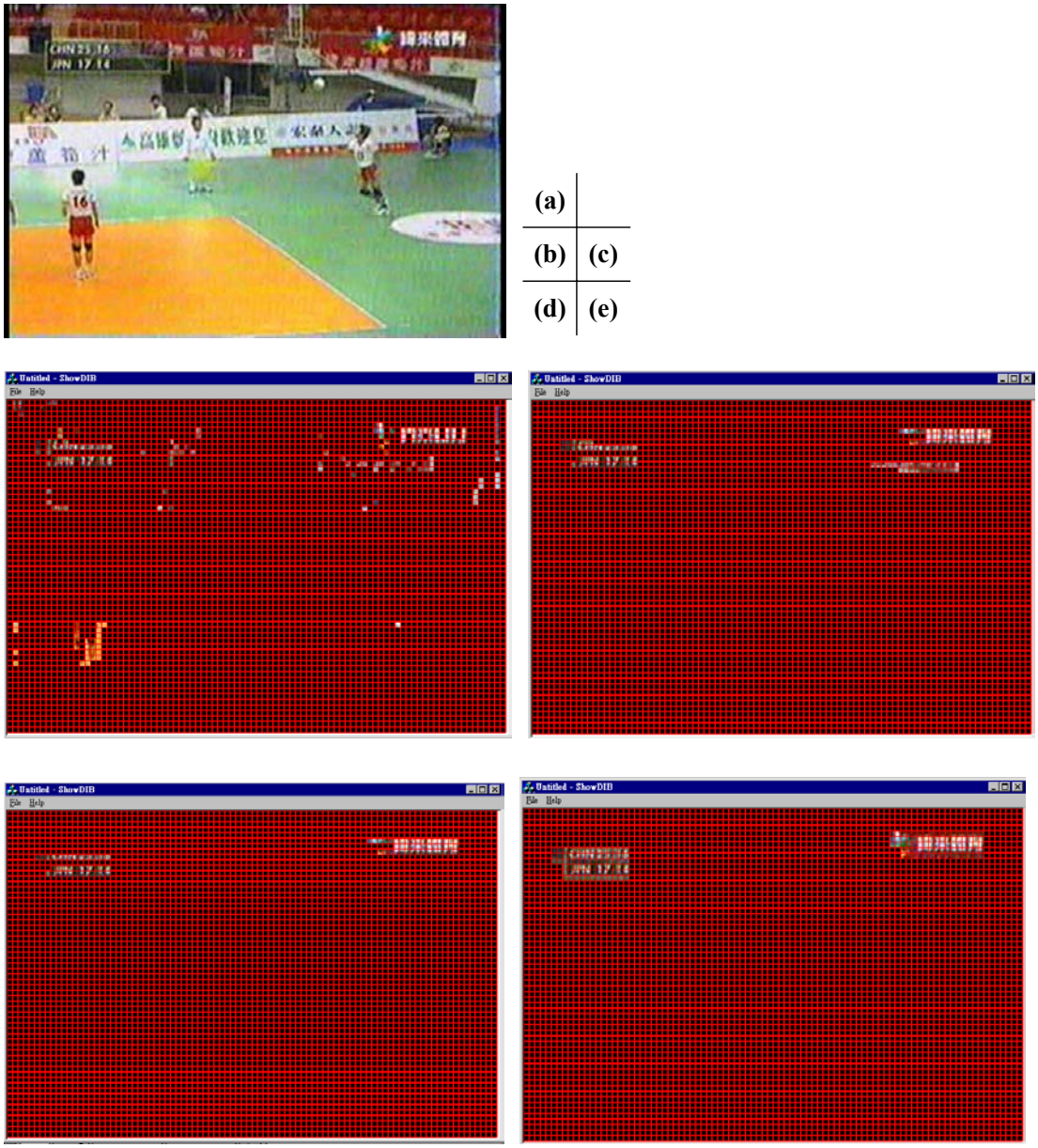


Fig. 5. Demonstration of the closed caption localization (a) Original I-frame (b) Result after filtering by horizontal gradient energy (c) Result after morphological operation (d) Result after filtering by SOM-based algorithm (e) Result after dilation

After eliminating most of the noisy regions, there still have many small separated regions in which they are either very close or faraway. Some regions are supposed to be connected, like the scoreboard and the channel trademark. Hence, we need to perform the task of regions merging and remove some isolated ones. Therefore, a morphological operator 1x3 blocks is used to merge the regions that the distance in between is smaller than 3 blocks and furthermore the regions of size smaller than 3 blocks are eliminated. The result of applying

morphological operation is shown in Fig. 5(c) and we can see that many small and isolated regions are filtered out and the caption regions are merged together. However, some background regions that have large horizontal gradient energy are still present after morphological operation. Hence, we propose an algorithm that is based on the concept of SOM (Self-Organization Map) to further differentiate the foreground captions and background high textured regions.

5.2 SOM-Based Noise Filtering

The Self-Organizing Map based algorithm [11] has been applied to the segmentation and recognition of textures and is well suited to the task of texture classification. In order to further differentiate the foreground captions from background high textured regions, a SOM-based noise-filtering algorithm is proposed. The details of the algorithm are described as follows.

SOM-Based Noise Filtering Algorithm

Input: Candidate regions after morphological operation $\Psi = \{ R_1, R_2, \ldots, R_n \}$

Output: Closed caption regions

1. Initially, set threshold $T = 70$ and cluster number $j=0$.
2. For each candidate region R_i, compute the average horizontal-vertical gradient energy E_i that is weighted by w_h and w_v. Here we set w_h to 0.6 and w_v to 0.4. n is the number of regions in Ψ.

$$E_i = \frac{1}{n}\sum_{j=1}^{n}\left(w_h \sum_{u=1}^{7}\left|AC_{0,u}\right| + w_v \sum_{v=1}^{7}\left|AC_{v,0}\right| \right) \quad (4)$$

3. For each region $R_i \in \Psi$

   If $i = 1$, $j=j+1$, assign R_i to cluster C_j

   Else if there is a cluster C such that $D_k \leq T$ and D_k is minimal among $\{ D_k \}$, where $k \in [1,j]$ and D_k is defined in Eq. (5)

   assign R_i to C

$$D_k = \frac{2}{|C_k|(|C_k|-1)} \sum_{i=1}^{|C_k|} \sum_{j=i+1}^{|C_k|} |E_i - E_j| \quad (5)$$

Else

j=j+1, create a new cluster C_j and assign R_i to C_j

4. Set $T = T - \delta$

   Select the cluster C_k (say C_{high}) that has the largest average gradient energy $E_{avg,k}$ computed by Eq. (6)

$$E_{avg,k} = \frac{1}{|C_k|} \sum_{i=1}^{|C_k|} E_i \quad (6)$$

5. If D_{high} of C_{high} is greater than *T*, then reset Ψ = C_{high}.

   Go to step 3.

   Else

   Go to step 6.

6. The cluster C_{high} is the set of closed captions.

In the algorithm, we set more weight to the horizontal DCT AC coefficients because closed captions generally appear in rectangular form and the AC energy in the horizontal direction would be larger than that in the vertical direction due to the reason that letters of each word are fairly close and the distance between two rows of text is relatively large. Furthermore, the SOM-based candidate region clustering is iterated until the gradient energy $E_{avg,k}$ of the cluster C_{high} is smaller than the threshold *T*. Based on the experiments, *T* is set to 70 and δ is set to 11 empirically, and the number of iterations is about 2 or 3. By this method, we can automatically find out the set of closed captions and this method is based on the fact that closed captions are the foreground and are superimposed after filming. Therefore, the closed captions are clearer and its gradient energy would be larger than the background. After the

step of SOM-based noise filtering, each closed caption region is dilated by one block row. The result is shown in Fig. 5(e) and we can see that regions belonging to the same closed caption are merged.

5.3 Highlight Extraction

In order to allow for quick understanding of the underlying story of the video program, we generate the video summary that is an audiovisual abstract of the video program. The structure of the video summary we provide is in hierarchical structure so that coarse-to-fine navigation is possible in order to access more detailed contents. In sports videos, we find that creating a video summary using low-level audiovisual features such as color is less semantic and less meaningful to human perception. In our paper, we create the video summary of highlights automatically, using the high-level semantic feature, assisted by content analysis and the highlight detection rule, respectively. The highlight detection rule proposed is to compute priority values of all the story units based on motion and provide users with the dynamic tree structure of video content according to the input.

According to our novel approach, we divide one point game into one service shot, several full-court view shots, and several close-up shots. Thus, we name the basic unit consisting of service hot, full-court view shots and close-up shots as "one story unit". We propose a quantitative measure of significance of a story unit based on object motion and camera motion. The priority **P** of the *n*th story unit is represented as Eq.(7).

$$P[n] = \left[W_F \times \sum_{1}^{n_F} MotionActivity_F \times L_F + W_C \times \sum_{1}^{n_C} L_C + (1 - W_F - W_C) \times L_S \right] \times \frac{3}{4} + \frac{Length_n}{Length_{Total}} \times \frac{1}{4} \quad (7)$$

$$MotionActivity_F = W_{object} \times \left| MV_{object} \right| + W_{camera} \times \left(W_{pan} \times \left| pan \right| + W_{zoom} \times \left| zoom \right| + W_{tilt} \times \left| tilt \right| \right) \quad (8)$$

In Eq. (7), $P[n]$ is the priority value of the *n*th story unit, W_F is the weight of full-court view shots, W_C is the weight of close-up shots, n_F is the number of full-court view shots in the *n*th story unit, n_C is the number of close-up shots in the *n*th story unit, $MotionActivity_F$ is the motion activity of one full-court view shot in the *n*th story unit based on object motion and

camera motion, L_F is the normalized full-court view shot's length - the number of P-frames, L_C is the normalized close-up shot's length, L_S is the normalized service shot's length, $Length_n$ is the number of total P-frames of the nth story unit, and $Length_{Total}$ is the number of total P-frames of a video sequence. In Eq.(8), W_{object} is the weight of object motion, W_{camera} is the weight of camera motion, W_{pan} is the weight of the pan camera motion, W_{zoom} is the weight of the zoom camera motion, W_{tilt} is the weight of tilt camera motion, MV_{object} is the average motion vector of objects in the full-court view shot, and the values of pan, zoom and tilt are obtained by utilizing these first three terms of Φ' in the affine motion model.

6. Experimental Results and Analysis

In the experiment, we record the volleyball videos from the TV channel of VL Sports and encode in the MPEG-2 format in which GOP structure is IBBPBBPBBPBBPBB and frame rate is 30 fps. The length of the video is about one hour and we obtain 163 shots of service, competition of the full-court view and closed-up. To measure the performance of the proposed scheme, we evaluate precision and recall for the approach of shot identification and the algorithm of closed caption detection. Table 1 shows the experimental result of the shot identification and we can see that the precision of all the three kinds of shots is higher than 92%. Moreover, the recall value of the type of close-up is up to 98%. The recall value of the type of full-court is just 87% due to the reason that the camera would zoom in to capture the scene in which players spike near the net. In this case, the scene would consist of a large portion of the net and an object of large size would be detected and thus the clip is regarded as a close-up shot. In addition, while a team is in the state of defense, several players may run to save the ball. In this case, the number of objects in the left and right parts may not be balanced and this shot would be classified as a service shot. Although the recall value of the shot of full court is not higher than 90%, the overall accuracy of shot identification is still very good. In Table 2, the result of closed caption localization is presented. There are 98 closed captions containing the scoreboard and the trademark in the testing video and 107

potential captions are detected in which 98 localized regions are the real closed captions. We can see that the recall value is up to 100% and the precision is about 92%. The number of false detection is 9 and this is because the background may consist of some advertisement page whose gradient energy is relatively high compared with the scoreboard and the channel trademark. In this case, the potential text region is assigned to the same cluster as closed caption since its gradient energy is very similar to the energy of the scoreboard.

Table1. Result of shot identification

Ground Truth	Number of Detection	Number of Correct Detection	Number of False Detection	Number of Miss Detection	Precision	Recall
Closed-up 58	62	57	5	1	92%	98%
Service 53	52	49	3	4	94%	92%
Full Court 52	49	45	4	7	92%	87%

Table 2. Result of Closed Caption Localization

Ground Truth	Number of Detection	Number of Correct Detection	Precision	Recall
98	107	98	91.59%	100%

The initial graphical user interface of the video browsing system is shown in Fig.6. The table of video content is provided in which the scoreboard of each game point is placed firstly and the representative frames of the three types of shots are shown in the order of service shot, full-court view shot and close-up shot, respectively. In the right top area, users can select which type of shots they are interested in and the percentage that they want to browse. Based on the semantic high-level video structuring, users can realize the overall view of the competition by the textual information in the scoreboard, select the desired point to watch and thus can effectively browse through the video sequences. In addition, while users are interested in the events of smash, defense or offense, they can select the shots of full court view. Fig.7 presents the full-court view shots while users select the percentage that they want to browse in the field of “FullCourt” and click the option “Submit”. Moreover, while users

want to look for some favorite players, they can watch through the shots of close-up view. Fig.8 shows the close-up shots while users select the first fifty-quantile of all the close-up shots.

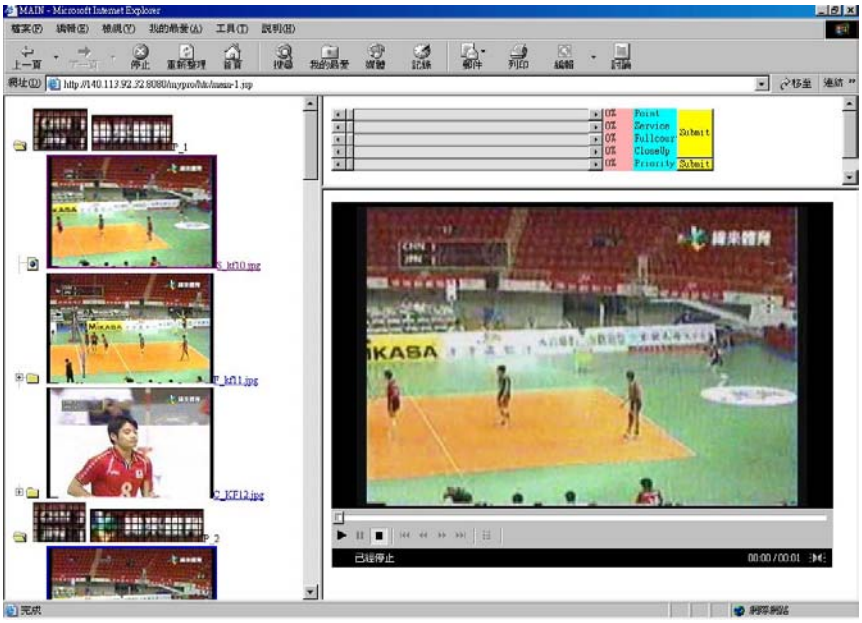


Fig. 6. Video structure of caption frames, shots of service, full-court view, and closed-up

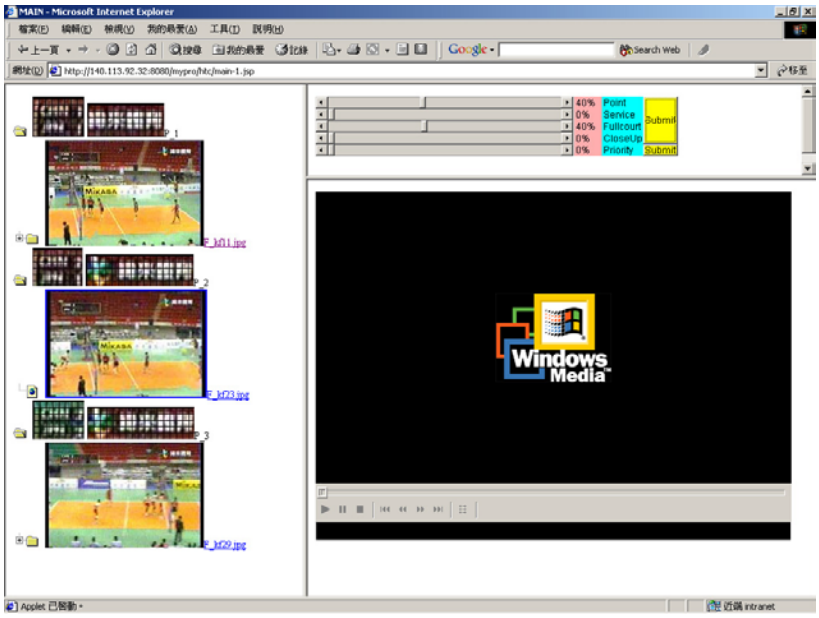


Fig. 7. The interface shows the full-court view shots, which are in the first forty-quantile in the left area

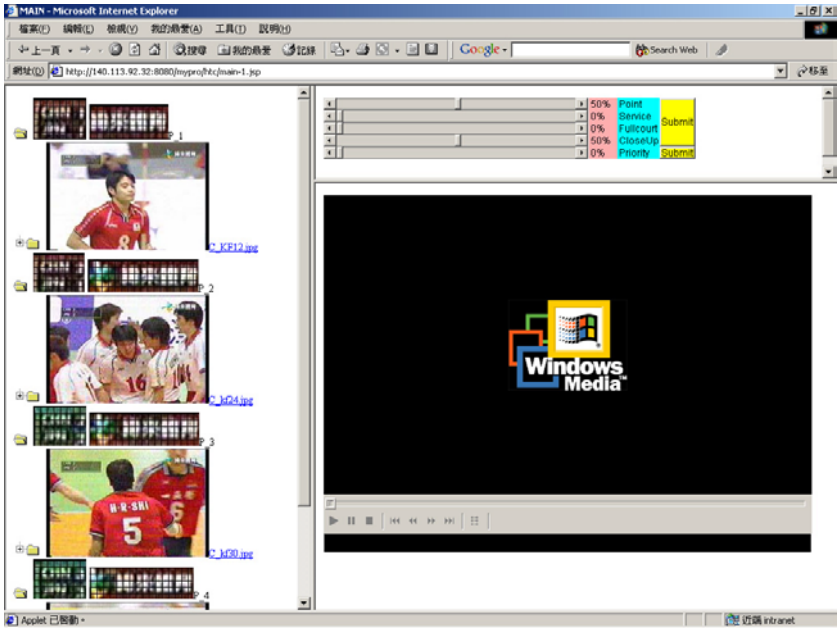


Fig. 8. The interface shows the close-up shots, which are in the first fifty-quantile in the left area

7. Conclusion and Future Work

In this paper, we propose a novel approach to structure volleyball videos in MPEG compressed domain and construct the table of video content by utilizing the textual information of the localized scoreboard and the shot description of motion activity. We use the approach of GOP-based scene change detection to efficiently segment videos into shots and these shots are further described by the object-based motion activity descriptor. Experimental results show that the proposed scheme performs well to recognize several kinds of shots of volleyball videos. In addition, we design an algorithm to localize the closed caption in shots of interest and to effectively filter out high-textured background regions. From the user interface, based on the table of content created by closed captions and semantically meaningful shots, we can support users to browse videos in more compact form and to acquire the desired data more efficiently.

In the future, with the successful identification of shots in volleyball game in this paper and effective classification of video shots of MPEG-7 test data set in our previous research, we would like to apply the proposed system architecture of motion activity based shots identification/classification to other videos, like movies, documentaries and some other sports

videos. In addition, we will investigate in the video OCR to recognize the localized closed caption to support automatic meta-data generation, like the name of the team in sports video or the name of the leading character in movies or an important person of other kinds of videos. Besides, we can also support semantic event detection and description for automatic descriptor and description scheme production in MPEG-7.

References

[1] H. Wang and S. F. Chang, "A Highly Efficient System for Automatic Face Region Detection in MPEG Video," IEEE Transactions on Circuits and Systems for Video Technology, Vol. 7, No. 4, pp. 615-628, Aug. 1997.

[2] Y. Zhong, H. Zhang and A. K. Jain, "Automatic Caption Localization in Compressed Video," IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 22, No. 4, pp. 385-392, Apr. 2000.

[3] H. Luo and A. Eleftheriadis, "On Face Detection in the Compressed Domain," Proc. ACM Multimedia Conference, pp. 285-294, 2000.

[4] Y. Zhang and T. S. Chua, "Detection of Text Captions in Compressed Domain Video," Proc. ACM Multimedia Workshop, CA, pp. 201-204, USA, 2000.

[5] S. W. Lee, Y. M. Kim and S. W. Choi, "Fast Scene Change Detection using Direct Feature Extraction from MPEG Compressed Videos," IEEE Transactions on Multimedia, Vol. 2, No. 4, pp. 240-254, Dec. 2000.

[6] H. Lu and Y. P. Tan, "Sports Video Analysis and Structuring," Proc. IEEE 4th Workshop on Multimedia Signal Processing, pp.45-50, 2001.

[7] D. Y. Chen and S. Y. Lee, "Object-Based Motion Activity Description in MPEG-7 for MPEG Compressed Video," Proc. the 5th World Multi-conference on Systemics, Cybernetics and Informatics (SCI 2001), Vol. 6, pp. 252-255, July 2001.

[8] S. Y. Lee, J. L. Lian and D. Y. Chen, "Video Summary and Browsing Based on Story-Unit for Video-on-Demand Service," Proc. 3rd International Conference on Information, Communications and Signal Processing, Singapore, Oct. 2001.

[9] J. L. Mitchell, W. B. Pennebaker, Chad E.Fogg, and Didier J. LeGall, "MPEG VIDEO COMPRESSION STANDARD," Chapman&Hall, NY, USA, 1997.

[10] J. Meng, Y. Juan and S.F. Chang, "Scene Change Detection in a MPEG Compressed Video Sequence," Proc. IS&T/SPIE, Vol. 2419, pp.14-25, 1995.

[11] T. Kohonen, "The Self-Organizing Map," Proceedings of IEEE, 78: 1464-1480, 1990.

[12] H. Li, D. Doermann and O. Kia, "Automatic Text Detection and Tracking in Digital Video," IEEE Transactions on Image Processing, Vol. 9, No. 1, pp. 147-156, Jan. 2000.

[13] J. C. Shim, C. Dorai and R. Bollee, "Automatic Text Extraction from Video for Content-Based Annotation and Retrieval," Proc. 14th International Conference on Pattern Recognition, pp. 618-620, 1998.

[14] U. Gargi, S. Antani and R. Kasturi, "Indexing Text Events in Digital Video Databases," Proc. 14th International Conference on Pattern Recognition, pp. 916-918, 1998.

[15] A. Hanjalic and R. L. Lagendijk, "Automated High-Level Movie Segmentation for Advanced Video-Retrieval Systems," IEEE Transactions on Circuits and Systems for Video Technology, Vol. 9, No. 4, pp. 580-588, June 1999.

[16] Y. M. Kwon, C. J. Song and I. J. Kim, "A New Approach for High Level Video Structuring," Proc. IEEE International Conference on Multimedia and Expo., Vol. 2, pp. 773-776, 2000.

[17] M. M. Yeung and B. L. Yeo, "Video Visualization for Compact Presentation and Fast Browsing of Pictorial Content," IEEE Transactions on Circuits and Systems for Video Technology, Vol. 7, No. 5, pp. 771-785, Oct. 1997.

[18] X. Chen and H. Zhang, "Text Area Detection from Video Frames," Proc. 2nd IEEE Pacific Rim Conference on Multimedia, pp. 222-228, Oct. 2001.

[19] J. Ohya, A. Shio and S. Akamastsu, "Recognizing Characters in Scene Images," IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 16, No. 2, pp. 214-220, February 1994.

[20] S. Kannangara, E. Asbun, R. X. Browning and E. J. Delp, "The Use of Nonlinear Filtering in Automatic Video Title Capture," Proc. IEEE/EURASIP Workshop on Nonlinear Signal and Image Processing, 1997.

[21] V. Wu, R. Manmatha and E. M. Riseman, "TextFinder: An Automatic System to Detect and Recognize Text in Images," IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 21, No. 11, pp. 1224-1229, November 1999.

[22] ISO/IEC JTC1/SC29/WG11/N3913, "Study of CD 15938-3 MPEG-7 Multimedia Content Description Interface – Part 3 Visual," Pisa, January 2001.