Computer Vision and Image Understanding 113 (2009) 415–424

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

journal homepage: www.elsevier.com/locate/cviu

A framework for flexible summarization of racquet sports video using multiple modalities ☆

Chunxi Liu [a], Qingming Huang [a,b,*], Shuqiang Jiang [b], Liyuan Xing [c], Qixiang Ye [a], Wen Gao [d]

[a] Graduate University of Chinese Academy of Sciences, No. 19, Yuquan Road, Shijingshan District, Beijing 100049, PR China
[b] Key Lab of Intell. Info. Process., Inst. of Comput. Tech., Chinese Academy of Sciences, No. 6, Kexueyuan South Road Zhongguancun, Haidian District, Beijing 100190, PR China
[c] Centre of Quantifiable Quality of Service, Norwegian University of Science and Technology, O.S. Bragstads plass 2E, Trondheim N-7491, Norway
[d] Peking University, No. 5, Summer Palace Road, Haidian District, Beijing 100871, PR China

ARTICLE INFO

Article history:
Received 26 September 2007
Accepted 18 August 2008
Available online 29 August 2008

Keywords:
Sports video summarization
Scene segmentation
Temporal voting strategy
Highlight ranking

ABSTRACT

While most existing sports video research focuses on detecting event from soccer and baseball etc., little work has been contributed to flexible content summarization on racquet sports video, e.g. tennis, table tennis etc. By taking advantages of the periodicity of video shot content and audio keywords in the racquet sports video, we propose a novel flexible video content summarization framework. Our approach combines the structure event detection method with the highlight ranking algorithm. Firstly, unsupervised shot clustering and supervised audio classification are performed to obtain the visual and audio mid-level patterns respectively. Then, a temporal voting scheme for structure event detection is proposed by utilizing the correspondence between audio and video content. Finally, by using the affective features extracted from the detected events, a linear highlight model is adopted to rank the detected events in terms of their exciting degrees. Experimental results show that the proposed approach is effective.

1. Introduction

Sports video plays an important role in our daily life and has a wide range of audiences. On one hand, a large volume of sports video data are produced everyday. On the other hand, the sports video content is redundant and the highlight points in the video are sparse. Many people have no time to watch the whole game or just want to see the highlights. Therefore, from the users' perspective, it is necessary to develop a system to automatically analyze the sports video and generate highlight summarization for the audience to browse what they want.

Because of the great commercial potential behind sports video analysis, a lot of research work has been contributed to it. Existing work on sports video content summarization can be classified into two classes: event detection and highlight summarization. Single-modal features, including image/audio, and multi-modal features that combine image, audio as well as text are employed to deal with these tasks. In the following we will review the existing work based on these two tasks.

For event detection, a lot of research work has been proposed. We will review the existing work according to the used features, which range from single modality to multi-modality. Some previous work employed single-modal feature, such as image or audio, for event detection. For example, for the image modality, Gong et al. [1] used player, ball, line marks and motion features to detect special events in soccer program. Xie et al. [2] proposed a method to segment soccer video into play or break segments for content abstraction by using a Hidden Markov Model (HMM), where video dominant color and motion activity were extracted as low-level features. In [3], cinematic features such as shot type, replays and object features were integrated into a Bayesian Network classifier to identify goal event in broadcast soccer video. For the audio modality, Rui et al. [4] used announcers' speech pitch and baseball batting sound to detect exciting segments in baseball games. Xu et al. [5] built audio keywords for event detection in soccer video. Xiong et al. [6] proposed a unified framework to extract highlight from baseball, golf and soccer by detecting cheer and applause. The content of sports video is intrinsically multi-modal and each modality takes different role and can compensate the limitation of other modalities. Therefore, integrating multiple modalities in a framework is a direction for event detection in recent years and lots of multi-modal approaches have been proposed. Snoek and Worring [7] categorized multi-modal approaches into simultaneous or sequential in terms of content segmentation, statistical or knowledge-based in terms of classification method, iterated or non-iterated in terms of processing cycle. It is also mentioned that

☆ This work is supported by National Hi-Tech Development Program (863 Program) of China under Grant 2006AA01Z117, National Natural Science Foundation of China under Grant 60773136 and 60702035.

* Corresponding author. Address: Key Lab of Intell. Info. Process., Inst. of Comput. Tech., Chinese Academy of Sciences, No. 6, Kexueyuan South Road Zhongguancun, Haidian District, Beijing 100190, PR China.
E-mail addresses: cxliu@jdl.ac.cn (C. Liu), qmhuang@jdl.ac.cn (Q. Huang).

1077-3142/$ - see front matter © 2008 Elsevier Inc. All rights reserved.
doi:10.1016/j.cviu.2008.08.002

most integration methods reported are symmetric and non-iterated, but there are some dilemmas in selecting classification method. Gong et al. [8] proposed a unique framework based on maximum entropy model for detecting and classified baseball highlight by seamlessly integrating image, audio, and speech clues. Leonardi et al. [9] used audio–visual based controlled Markov chain models for salient event detection in soccer. Huang et al. [10] proposed four different HMMs for video scene classification. Petkovic et al. [11] inferred semantic by using DBN to fuse multimodality information. These statistical model based methods provide generic solution by depending on decision tools and statistical rules, but the limitation is that the domain knowledge is not properly used and the processing of training and testing is not intuitive. Thus, researchers seek help from knowledge-based method. Rui et al. [4] used the heuristic rule of weighted exciting speech and baseball hit for baseball highlight. Nepal et al. [12] extracted the scored events in basketball by using the heuristic rule of cheers, change in direction and scoreboard display.

All of the above methods are proposed for event detection. Most of these methods are applied to baseball and football etc., where the exciting events are rare since they happen casually in a match. Therefore, using these events for highlight content summarization is effective. However, racquet sports such as tennis and table tennis have more events than the sports mentioned above. It is hard to evaluate the exciting degrees of these events. Moreover, using all these events for content summarization is not suitable, especially for the application of mobile terminal which has limited processing and storage capability. Therefore, according to users' requirement and hitting the capability of the mobile terminal, a flexible content summarization will be a new incremental service mode.

To satisfy this demand, recently highlight detection for flexible content summarization on sports videos has been proposed by building the exciting model of video scenes using affective features. Hanjalic [13] obtained the excitement time curve by linearly combining three excitement components, which are motion activity, density of cuts and sound energy. Xiong et al. [14] formed the plot of averaged relative entropy curve by probabilistic fusion of audio and motion. Then, a highlight is the time when there is a local maximum in the exciting degree curve. The advantages of this method are generality and flexibility. It mimics the changes in user's excitement by the content of the video such as domain-independent audio/visual properties and the editing scheme of the video rather than modeling domain-specific events. In this way, flexible length content summarization can be obtained according to the exciting degree curve. However, both of the above methods have difficulty in deciding the boundary of highlight since the calculated unit is frame [13] and window [14]. Zhu et al. [15] proposed a method for highlight generation based on human behavior analysis. In [16] Zhang et al. proposed a highlight detection method based on audio analysis and adapted HMM model. Further they proposed a highlight detection algorithm by fusing audio and visual information in [17]. A highlight ranking of broadcasting table tennis video based on ball and player tracking has been proposed in [18]. Researchers usually evaluate the experimental results by observing the local maxima of the exciting degree curve to see whether it is an exciting event or not. But until now there are no criteria that are able to tell to what extent the highlight reflect human perception and can guide for the excitement components (affective features) selecting and exciting degree (highlight) modeling.

In this paper, we focus on analyzing racquet sports video, which usually has the following features. The racquet sports happen in a periodic pattern and are score constrained. There are two players fighting with each other with a racquet. The main contribution of the paper lies in that we propose a novel framework for flexible summarization of racquet sports video using multiple modalities. Our goal is to segment a racquet sports video into rally and break events, and rank the rally events according to their exciting degrees. For the ranked highlights, flexible summarization can be obtained by arbitrarily selecting a highlight rank by the user. This framework combines the event detection with highlight detection and improves each method for better generic flexible summarization. We restrict the event detection in structure event, such as play and break, rather than specific event, such as ace, double fault, which can make the event detection more generic. We achieve the generality by unsupervised shot clustering and supervised audio classification, as well as temporal voting strategy. It is a middle-level feature, knowledge-based and context sensitive fusion method for structure event detection. Compared with the methods in [13,14], our method has clear boundary and can reduce the amount of content to be analyzed for subsequent processing. Moreover, in the highlight ranking, we adopt a subjective evaluation criterion for selecting affective features and modeling highlight to make the ranking reflect human perception more reasonable. We believe that the proposed flexible framework has a broad application in sports video summarization and also has great commercial potential in mobile terminal services.

The rest of the paper is organized as follows. In Section 2, the framework for racquet sports video summarization is introduced. The details of structure event detection and highlight ranking for flexible summarization are presented in Sections 3 and 4, respectively. Experimental results are illustrated in Section 5 and conclusion is provided in Section 6.

2. System overview

Unlike soccer and baseball, racquet sports have few distinct definitions of exciting events. We can define *shot on goal* as an exciting event in soccer but cannot define all of the *scoring* events as exciting ones in tennis. A racquet sports game consists of many play and break events. The play event in the racquet sports, which has a unique structure characteristic, is called *rally* in this paper. Each *rally* is a period during which the ball is in play. In general, after a *rally* the audience will express their feelings by cheering, applauding or keeping silence. In the broadcast sports video, announcers may comment on the rally. After a rally, there may be several seconds for rest, say there is a break event. The alternation of *rally* and *break* event as well as the alternation of audio from audience to ball impact make racquet sports video having a well defined temporal structure.

In this paper, we propose a system for racquet sports video summarization, which is shown in Fig. 1. Fig. 1A demonstrates the proposed process of the flexible racquet sports video summarization. Suppose the overall highlights rank is $R(0, 1, 2, \ldots, r, \ldots, R)$ with rank 0 being the most exciting and rank R the least exciting. If a user selects rank r for summarization, the system will produce the sum of rally events from rank 0 to rank r. Further, users can access any rally event by its index. The summarization result is valuable for some applications such as sports news material preparation, wireless video content transmission, etc.

The technical framework of the proposed system is shown in Fig. 1B. In general, a broadcast racquet sports video happens in a space constrained condition, which leads to the video content restricted in a given match. Therefore, we can group the video shots of similar visual content into the same cluster by using an unsupervised method. We use unsupervised instead of supervised video shot classification for rally event detection, because the scene in racquet sports varies a lot due to the significant difference in the color of playfield among games. The unsupervised method has good generality since no model needs to be trained with the change of sports video, which keeps the universality of the proposed framework. By our observation, in most racquet sports

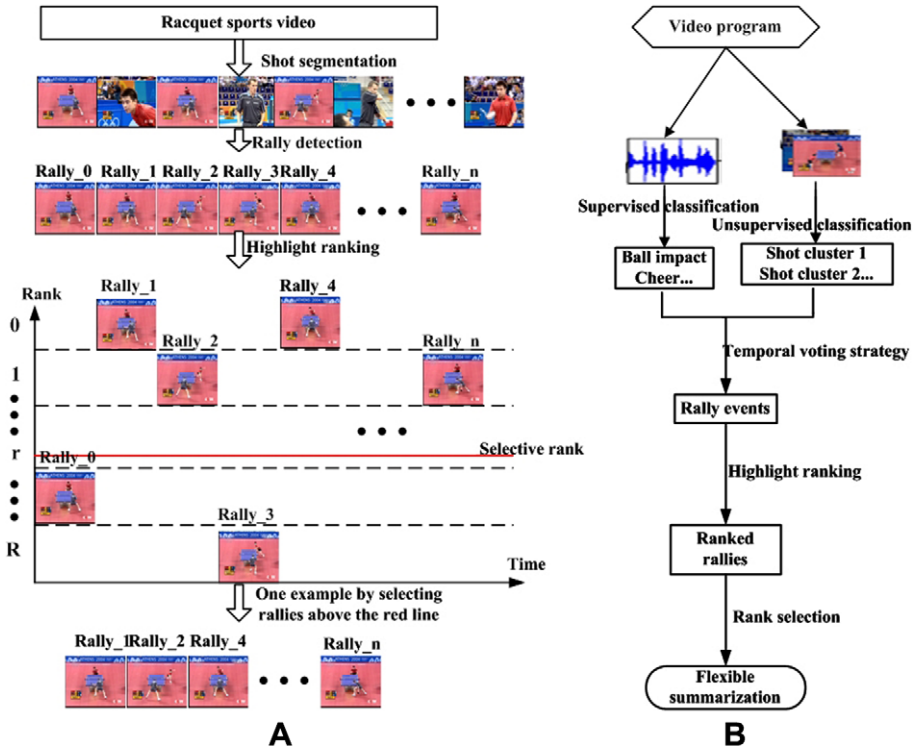

Fig. 1. Flexible video content summarization demonstration and system framework.

video, audio can be classified into four classes: silence, ball impact, cheer and speech. These audio types nearly invariable in racquet sports video and are robust and general clues for content analysis.

The video shot classification result from unsupervised method keeps no semantic meaning. Therefore, we propose a strategy to label the clustered video shot with semantic meaning by audio–video correspondence based on the fact that the audio information is more related to the state of a match. For example, the occurrence of *ball impact* usually indicates the ball is in play, and the occurrence of *cheer* often indicates the end of a *rally* or the happening of highlight. Audio symbols are distributed into clusters according to the temporal information and the meaning of a cluster can be determined by the voting result of audio symbols. For example, in racquet sports video, the *ball impact* can vote in the *rally* scene. By this strategy, both of the characteristics of the robustness and generalization of unsupervised method and the strong meaning of audio are utilized. Its obvious advantage is generality. The voting foundation is the time stamp, which is the basic information in audio/video stream. After voting, rally events are obtained by segmenting out the shots with their semantic meanings.

The rally events are further represented by affective features. In this paper, six affective features are extracted from audio and image modalities. Based on these features, a linear regression model is adopted to rank the rally events. In order to evaluate the effectiveness of the features and the ranking method, a subjective evaluation criterion is adopted.

In the following, we will provide a detailed description of the framework by taking tennis and table tennis as examples.

3. Structure events detection

In this section, the visual and audio representations of video shots are processed by unsupervised and supervised methods respectively. Then a temporal voting strategy is proposed to obtain the rallies.

3.1. Unsupervised video shot classification

Unsupervised video shot classification for sports video analysis has been adopted in [19,20]. In [19] an iterative filtering shot clustering method is proposed. However their results only obtained few clusters, which is not so discriminative. *K*-mean clustering is adopted in [20]. However, in clustering the most important and difficult task is to determine the cluster number. Duan et al. [21] proposed a unified framework for semantic shot classification. The cluster number is pre-defined according to semantic. In our approach, an unsupervised hierarchical clustering method of our previous work [22] is adopted, where a merging stop criterion for clustering has been proposed. Let's define J_{ratio} as the total scene cluster scatter, which describes the ratio of intra-cluster scatter to inter-cluster scatter of the scenes in the clustering. Supposing the scene number is K_1 in the clustering process, J_{ratio} is calculated as

$$J_{\text{ratio}} = \frac{\sum_{c=0}^{K_1} J_w^c}{J_{\text{inter}}} = \frac{\sum_{c=0}^{K}\sum_{i=0}^{N_c}\|\vec{s}_i^c - \vec{s}_{\text{mean}}^c\|}{\sum_{i=0}^{N}\|\vec{s}_i - \vec{s}_{\text{mean}}\|} \quad (1)$$

where J_{inter} is the total inter-cluster scatter of the initial scene sequence, J_w^c is the intra-cluster scatter of scene cluster c. N is the total scene number in the initial scene sequence. N_c is the shot number of scene cluster c. $\|\bullet\|$ represents the Euclidean distance. $\vec{s}_i^c$ ($\vec{s}_i$) denotes shot i in scene cluster c (the initial scene sequence), and $\vec{s}_{\text{mean}}^c$ ($\vec{s}_{\text{mean}}$) denotes the mean feature value of shots in scene c (the initial scene sequence).

At the beginning of a clustering procedure, the intra-cluster scatter of all initial scenes is set as 0 and J_{ratio} as 0.0. With the increasing of intra-cluster scatter when two scenes are clustered into one, J_{ratio} is increasing. If all the scenes (shots) are clustered into one scene, J_{ratio} reaches its maximum 1.0. The smaller J_{ratio} is, the more similar the shots within each scene cluster are. Actually, it is expected that both J_{ratio} and the scene number are

small. As a tradeoff between J_{ratio} and the scene number, we try to find a condition where $J_{ratio} + k_l$ is the smallest. Here $k_l = K_l/N$ is the ratio of the scene number to the total number of scenes in the initial scene sequence in the classification process. More details about the algorithm can be found in [22].

3.2. Supervised audio classification

For audio classification, four groups of clip-level features including both time-domain and frequency-domain ones are extracted based on frame-by-frame audio data. They are listed in Table 1. Particularly, the four sub-bands cover the frequency interval of 0 Hz-fs/8 Hz, fs/8 Hz-fs/4 Hz, fs/4 Hz-3 * fs/8 Hz and 3 * fs/8 Hz-fs/2 Hz. All these features are commonly used for audio classification and the detail calculation of these features can be found in [23–25].

Totally, we extract 55 features from a 1-s clip. Although all these features can be used to distinguish audio, some features may be more effective than others. Using the most powerful subset of the features will reduce the time for feature extraction and classification. Furthermore, the existing research has shown that when the number of training samples is limited, using a large feature set may decrease the generality of a classifier. Thus, a forward searching algorithm [26] is adopted to perform the feature selection task before the features are fed into the classifier. A short description of this feature selection algorithm is as below:

Firstly, the feature set F is divided into selected feature set F_S and unselected feature set F_U, and then the feature is selected one by one using the following procedure:

(1) Set F_S = empty and $F_U = F$;
(2) Label all of the features in F_U untested;
(3) Find the feature f which achieves the highest classification accuracy among all the features in F_U when combining with the feature in set F_S and test in the data set, then move f from F_U to F_S;
(4) If feature set F_U is not empty then goto Eq. (3), else the procedure exits.

Fig. 2 shows the performance curve in the feature selection process. We find that with the increasing of feature dimension, the accuracy increases sharply at first but slightly decreases when the number passes certain value (17 dimensions), which is selected as the feature dimension. The selected features are listed in the last column of Table 1. In the following, audio classification is performed on the 17 selected features.

We use Support Vector Machine (SVM) as the classifier for its good performance on small training set [27]. The kernel function for SVM is RBF (Radial Basis Function) with parameter gamma = 1/dim.

Table 1
Extracted and selected audio clip features

Frame level features	Mean value	Standard deviation	Low ratio	High ratio	Difference[a]	Selected features (17)
ZCR	1	2	3	4		1, 2, 4
STE	5	6	7		8, 9	6, 8, 9
Pitch	10					
Brightness	11	12				11
Bandwidth	13	14				
Spectrum flux (15)						15
Sub-band power	16–19	20–23				
LPCC	24–31	32–39				25, 26, 34, 36
MFCC	40–47	48–55				45, 46, 47, 54, 55

[a] Difference means the mean value and standard deviation of different ZCR.

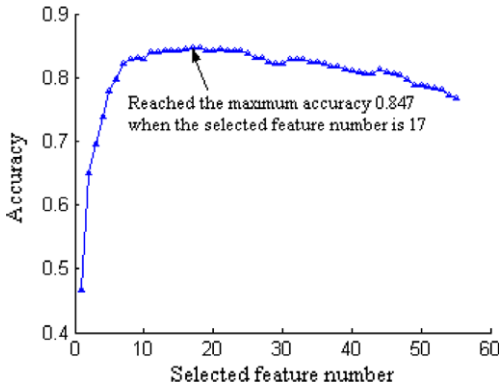

Fig. 2. The feature selection by forward searching method.

3.3. Temporal voting strategy for semantic video scene segmentation

In the previous steps, video shots have been grouped into clusters. However, we do not know the exact semantic meaning of these clusters. For temporal voting strategy, it means that the audio clips are marked with their classification labels and they are in a time sequence. At the same time, the video shot is aligned with these audio clips in time. The voting strategy is to find the best semantic of each shot cluster by using the audio clip labels. We seek help from the inevitable relationship between audio and visual information based on the fact that the sound classes have strong semantic meaning and the scene clusters have reliable temporal boundaries.

By observation, we find that for each of the video shot clusters, there exists a sound class coupling with it best. Assume that there are K sound classes and N video scene clusters in one video. Let m_{kn} be the number of time slots and belong to the k-th sound class that is distributed in the n-th scene cluster. Suppose that the k-th sound class has M time slots and t_{ki} is within the time-domain of the i-th time slot. Then

$$m_{kn} = \sum_{i=0}^{M} \delta[f(t_{ki}) - n] \tag{2}$$

$$n' = \arg_n \max(m_{kn}) \tag{3}$$

where $f(t)$ stands for the result of video clustering, and $f(t) \in [1, 2, ..., N]$, So n' is the exact match of k. For example, when the k-th sound class is the ball *impact* in racquet sports, the corresponding n'-th scene cluster is the rally scene. Eq. (2) describes the temporal fusion method with which the audio symbols find the clusters, according to temporal relationship between the audio and shot in cluster. Eq. (3) illustrates the voting strategy that the cluster obtains its meaning by the most voted audio symbol. As shown in Fig. 3, the audio symbols are distributed into the clusters by Eq. (2), and the meaning (color) of cluster is obtained (drawn) by the vote in sound symbol (color). In this work the interested video shot cluster is the rally cluster.

4. Highlight ranking for flexible summarization

In this section, we first extract some commonly used affective features and then adopt a subjective evaluation criterion for highlight modeling.

4.1. Affective feature extracting

Hanjalic and Xu [28] defined the affective content of video as the intensity and type of feeling or emotion which is contained

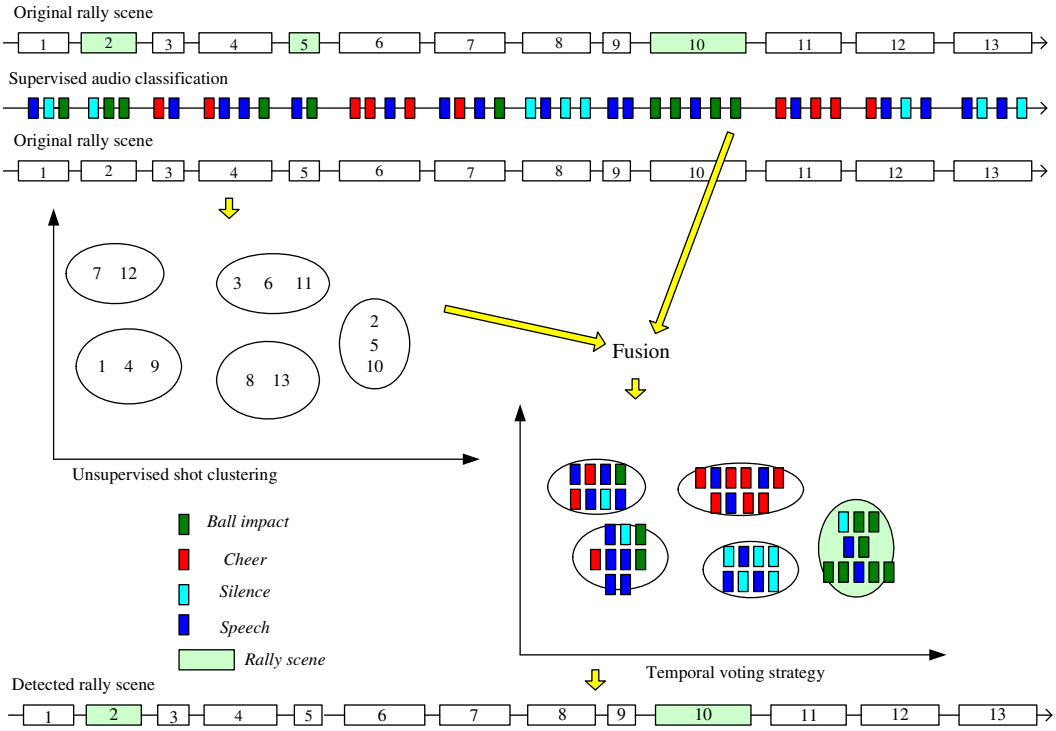

Fig. 3. Temporal voting strategy for semantic scene segmentation.

in video and expected to arise in users while watching the video. The features representing these affective contents are called as affective features. However, the affective content of a video does not necessarily correspond to affective response of a particular user to this content, and it is difficult to arrange a large number of the audiences to watch the video and record or measure the affect. Alternatively, researchers [13,14,24] use the response of audience and commentator in the video instead, because there exist abundant audience and professional commentator in broadcast sports video. In addition, both the visual [13,14,29,30] characteristic in event itself and editing [13,29] characteristic by the editor are adopted.

These above affective features can be coarsely classified into three categories: visual, audio and editing features. In [13,14,29,30], the commonly used affective visual feature is motion. It captures the pace of action [14] and shows a significant impact on individual affective response [29]. MPEG-7 intensity of motion activity descriptor [14,30] and average magnitude of all motion vectors [13] are adopted. These two motion activity descriptors as well as seven other features are evaluated in [29], and the results show that MPEG-7 motion vector is one of the best performing descriptors. The commonly used affective audio features are cheers and pitch-related features [13,14,29], which are considered as having strong relationship with the affective content of the video. For example, in sports video, the occurrence of audience *cheer* and the *pitch* improvement of the commentator imply highlight scenes. The longer duration and higher average energy of *cheer*, as well as the longer duration and higher average pitch of *excited speech* are, the more exciting the event is. The effective affective editing feature is the variation of shot length [13,29], and in our approach we use the duration of *event* instead, where the event duration in racquet sports video is from the beginning to the end of a rally.

By these observations, we totally extract six affective features from each event for highlight ranking. They are

a. *MPEG-7 motion vector* average
b. *cheer* duration
c. *cheer* average energy
d. *excited speech* duration
e. *excited speech* average pitch
f. *event* duration

4.2. Highlight modeling and ranking for video summarization

The goal of highlight modeling is to build the relationship between affective features and exciting degrees. Compared with the method in [29], which uses the criteria of comparability, compatibility and smoothness for establishing the highlight model, we adopt a subjective evaluation criterion, which is inspired by the pair-wise comparison method [31] and has been proved to be effective in [14], to help to establish the model which reflects human perception more reasonably.

In order to compare the subjective (Human) value and the measured (Computer) value, the general method is to gain the average error. The smaller the average error is, the better the measured one can describe the subjective perception. Most recently, Peker and Divakaran [29] proposed a novel pair-wise comparison method to have a more detailed analysis on individual clips. However, they referred to the fact that the conclusion of the pair-wise comparison is in accordance with that of the average error method. In addition, it is not feasible to quantitatively observe the effect of each affective feature and highlight model on individual event. Therefore, a subjective evaluation criterion that improves the average error method is adopted in this paper. The main improvement lies in introducing the highlight rank R' into the average error method. Instead of using fixed six levels in [29], there are no fixed levels here, which are hard to set up beforehand in our application. Subjects are left free to give value to each event between 0 and 1 according to its exciting degree. The highlight rank R' is automatically decided in an optimal quantization process. The R' can reduce

the gap between the continuous absolute value and the discrete relative level.

Firstly, we define the continuous absolute value of each event m to be r'_m which is between 0 and 1. Its corresponding discrete relative level is integer r_m, which is between 0 and R'. Especially, the highlight rank R' is automatically decided in optimal quantization process in order to minimize the error between the continuous absolute value r' and the discrete relative level r. Suppose that the segmented events number is M, then

$$Q(r'_m) = r_m, \quad \text{if } r_m/R \leqslant r'_m < (r_m+1)/R \tag{4}$$

$$\text{err}_R = \sum_{m=1}^{M} |r_m - r'_m| * R \tag{5}$$

$$R' = \arg_R \min(\text{err}_R) \tag{6}$$

As long as R' is obtained, the continuous absolute value can be converted to the discrete relative level with the lowest quantization error.

Secondly, suppose that the test video contains M' events and the regression value is c'_m, and $Q(c'_m)$ is the corresponding discrete relative level after quantization process with R'. Then, based on the ground truth $Q(r'_m)$ and the highlight rank R', we can evaluate the highlights ranking result $Q(c'_m)$ by computer as

$$\begin{aligned}\text{affective accuracy} &= \frac{1}{M'}\sum_{m=0}^{M'} \frac{R' - |Q(r'_m) - Q(c'_m)|}{R'} \\ &= \frac{1}{M'}\sum_{m=0}^{M'} 1 - |Q'(r'_m) - Q'(c'_m)|\end{aligned} \tag{7}$$

where $|Q'(r'_m) - Q'(c'_m)|$ represents the relative bias between highlight ranked by human and computer, so the change of R' which is selected according to the optimal quantization principle will not affect the accuracy. The difference of 1% in accuracy means a difference of 1% in relative bias. If the accuracy is 80%, there is 20% difference between human rank and computer rank relatively. Evaluation criterion Eq. (7) shows that the accuracy is obtained by averaging the human–computer rank bias. So the more effective features and reasonable highlight model are, the higher the affective accuracy is.

Generally, lots of signal systems in the real world are non-linear, especially in terms of system with subjective human perception. However, by experiments we find that in this ranking task, the performance of a simple linear model is close to that of the complex non-linear model. Thus different from the non-linear method in [15,22], in order to reduce the training complexity of the system, we adopt linear model for highlight ranking. After the highlight ranking procedure, flexible video content summarization can be obtained by just inputting an exciting rank r. The system will deliver all of the segmented scenes whose rank values are lower than r (which means the delivered scenes are more exciting than that of rank r) for summarization as demonstrated in Fig. 1.

5. Experiments

We prepare 5552 s tennis videos and 6451 s table tennis videos for experiments. The tennis videos are from four different living broadcast programs of French Open 2005 (Table 2 T1–T4) and the table tennis videos are from four different living broadcast programs of Athens Olympic 2004 (Table 2 P1–P4). Video information is listed in Table 2. These videos are in MPEG-2 format with 352 × 288 resolution. The audio signal is sampled at 44,100 Hz and 16 bits/sample.

5.1. Unsupervised shot classification and supervised audio classification

The most important task in unsupervised video shot classification is to correctly clustering the shots of similar visual content into the same scene cluster. As for the experimental videos, the determined cluster numbers are listed in Table 3.

It can be seen from the table that the cluster number varies a little in the same type of sport. This is because that the video edition for the same type of sports game is mature. It also can be seen that the cluster number of tennis is often larger than that of the table tennis. This is because that tennis is more complicated than table tennis and the camera number in tennis is more than that in table tennis, which makes the views in tennis more various. Fig. 4 demonstrates some of the scene clustering results by the method.

In audio classification, the selected audio features are firstly extracted from each of the audio clips with 1 s duration. The duration of audio clips is determined by performance-oriented test which compares the classification performance of duration from 0.1 to 2 s. Then these audio clips are classified into one of the pre-defined audio classes: *ball impact*, *cheer*, *silence* or *speech*. For different genres of sports video we train different classifier models. To evaluate the classification results, we refer to the human labeled ground truth. Audio clips from videos T1(P1) and T2(P2) in tennis (table tennis) are selected for audio training, and audio clips from T3(P3) and T4(P4) for testing. The classification result is listed in Tables 4 and 5, respectively. The *recall* rate and *precision* is defined as in [18].

In the experiment, we find that audio is often mixed with each other in real video, which will aggravate difficulty in audio classification. Our experiments also show that the audio classification is general for different racquet sports.

5.2. Rally events detection by temporal voting strategy

We mainly test *rally* event detection performance, because they are basic units for highlight ranking and it is important to correctly extract them. The detection results are listed in Tables 6 and 7. As for table tennis video, the final rally events detection precision is 91.6% with a 93.2% recall rate, which is an encouraging result for real applications.

To prove the advantage of audio–visual combination method for event detection, event detection results by pure audio are presented in the tables for comparison. It can be seen from the tables that the *rally* events detected by audio–visual combination are much better than that by audio information only. Pure visual information can be used for event detection but cannot assign semantic meaning to segmented scenes as stated above. The main reason why we do not use supervised video shot classification for *rally* event detection is that the scene in racquet sports varies a lot from game to game.

5.3. Highlight ranking for flexible summarization

In order to evaluate the highlight ranking results, the ground truths are required. The experimental data are the same as the

Table 2
Videos information

Video	Tennis videos				Table tennis videos			
	T1	T2	T3	T4	P1	P2	P3	P4
Duration	29:27	26:53	9:40	26:32	27:40	33:57	11:08	34:46

Table 3
Cluster number determination results in tennis and table tennis videos

Video	Tennis videos				Table tennis videos			
	T1	T2	T3	T4	P1	P2	P3	P4
Cluster number	12	13	11	8	8	6	5	7

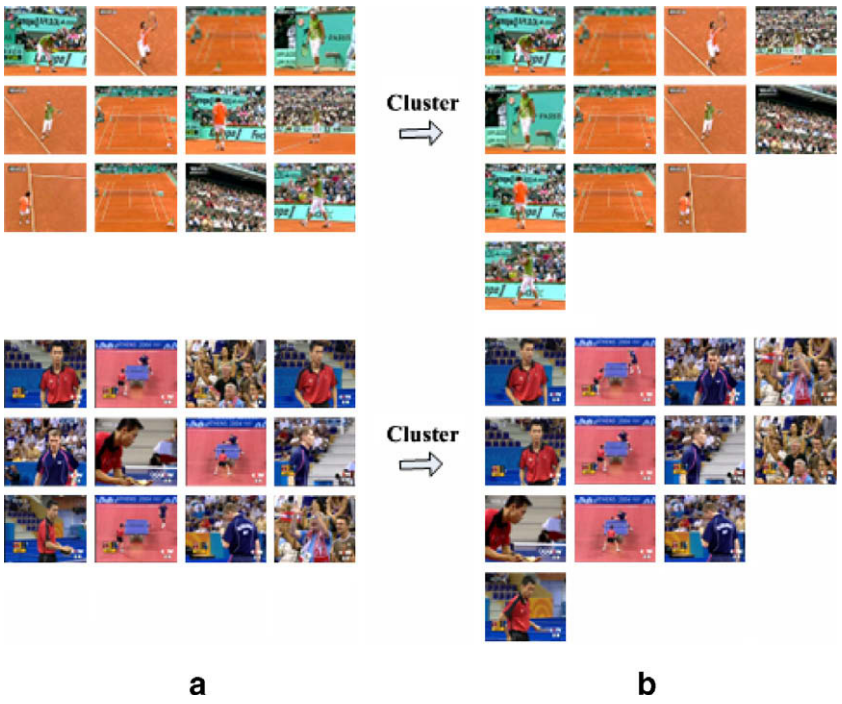

Fig. 4. Video shot classification result, (a) the key frames of shot in temporal sequence, (b) the clustering result where key frames in the same column represent the shots in same cluster.

Table 4
Audio classification results in tennis videos

Audio type	Total number	Discrimination results				Recall (%)	Precision (%)
		(1)	(2)	(3)	(4)		
(1)	484	415	0	25	34	85.7	65.4
(2)	159	29	116	3	21	73.0	92.1
(3)	587	83	0	499	5	85.0	81.0
(4)	944	128	10	89	717	76.0	92.3

(1), *ball impact*; (2), *cheer*; (3), *silence*; (4), *speech*.

Table 5
Audio classification results in table tennis videos

Audio type	Total number	Discrimination results				Recall (%)	Precision (%)
		(1)	(2)	(3)	(4)		
(1)	211	150	11	9	41	71.1	63.0
(2)	748	75	616	24	43	82.4	75.6
(3)	148	1	8	122	17	82.4	67.4
(4)	1457	12	180	26	1239	85.0	92.5

(1), *ball impact*; (2), *cheer*; (3), *silence*; (4), *speech*.

Table 6
Rally event detection results in tennis videos

Video clip	By audio information only		By audio/visual combination	
	Precision (%)	Recall (%)	Precision (%)	Recall (%)
T3	57.1	88.9	94.1	88.9
T4	55.1	75.4	78.0	80.7
Average	56.1	82.2	86.0	84.8

Table 7
Rally event detection results in table tennis videos

Video clip	By audio information only		By audio/visual combination	
	Precision (%)	Recall (%)	Precision (%)	Recall (%)
P3	73.3	39.3	89.7	92.9
P4	81.6	43.0	93.5	93.5
Average	77.5	41.0	91.6	93.2

above experiment. Event is the *rally* in tennis and table tennis video. The details of each video are listed in Table 8.

Since highlight is a subjective concept, a psychophysical experiment is designed. If the exciting degree of each event is only set by one person, it will limit the reliability of the statistical results. Therefore, six subjects, five males and one female aging from 23 to 29, are invited to give the exciting degree for each event. The six subjects are naïve to the purpose of the experiment and the

Table 8
Experimental video description

Video	Tennis videos				Table tennis videos			
	T1	T2	T3	T4	P1	P2	P3	P4
Duration	29:27	26:53	9:40	26:32	27:40	33:57	11:08	34:46
Event number	83	79	30	98	82	61	21	70

only thing they know and should do is to provide value for each event between 0 and 1 according to its exciting degree. The subjects are free to present exact definition and scale of the highlight. It will not arouse confusion, since people are good at comparison, especially in a continuous match. And they are able to automatically adjust the exciting degree value to a reasonable state according to the total video. Then the ground truth is defined as the mean values of the subjects for each event. Tables 9 and 10 show the average deviation of each subject on tennis and table tennis video, respectively. It can be seen that the mean of the subjects' deviation is 0.21 and 0.22, which are the guideline in the latter evaluation experiments.

5.3.1. Features selecting and highlight modeling

Based on the proposed evaluation criterion Eq. (7), we use the same forward search algorithm as in Section 3.2 to evaluate the affective features. The SVM cross validation is performed with three sets of randomly selected data to avoid the circular problem of training and testing on the same set. Kernel function for SVM is RBF (Radial Basis Function) with parameter gamma = 1/dim. The value of R' in function Eq. (6) is 10 based on the data set.

Fig. 5 shows the accuracy on different feature number. As shown in Fig. 5, the differences of minimum and maximum are 4% in tennis and 3% in table tennis, respectively. There is no much change in affective accuracy with the addition of one more affective feature. It means that these affective features are quite correlated. As shown in Fig. 6, it makes out the relativity again since one feature alone is able to reflect the exciting degree to a large extent and their ability has no much difference. Furthermore, It can be seen that feature *d* is the domain feature both in Figs. 5 and 6, but the combination of *a*, *b*, *d* and *f* gets maximum affective accuracy in Fig. 6. So we conclude that the combination of affective features of *a*, *b*, *d* and *f* is reasonable. Based on this experimental result, we can make the conclusion that the commonly used affective features are correlated and the combination of *a*, *b*, *d* and *f* is effective.

We need to get to the bottom of whether the linear regression model is effective for our highlight ranking. A non-linear model (SVM regression model) is selected for comparison. The reason why we adopt SVM regression for comparison is that it has the advantages of kernel-based learning method, such as requiring fewer training samples and having better generalization ability even for sparse data distribution [27].

With the selected affective features *a*, *b*, *d* and *f* being fed into the regression model, the comparison results of non-linear and linear regression are listed in Tables 11 and 12. It can be seen that there is little improvement by using non-linear regression (SVM regression). Therefore, in order to reduce the training complexity of the system, we adopt linear model for highlight ranking.

It also can be seen that the affective accuracy reaches around 80.0% in terms of the ground truth and evaluation criteria. We must make it clear that 82.0% (79.3%) affective accuracy is a con-

Table 9
Each subject's deviation on tennis video

Subject number	1	2	3	4	5	6	Average
Average deviation of each event	0.23	0.24	0.17	0.20	0.20	0.21	0.21

Table 10
Each subject's deviation on table tennis video

Subject number	1	2	3	4	5	6	Average
Average deviation of each event	0.25	0.21	0.22	0.24	0.19	0.21	0.22

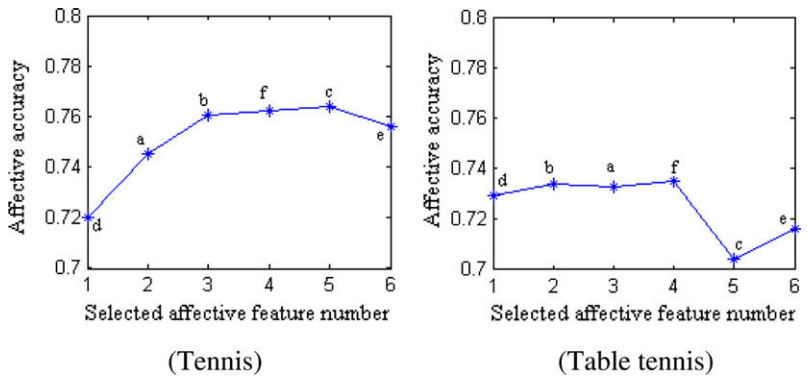

Fig. 5. The affective feature selection process.

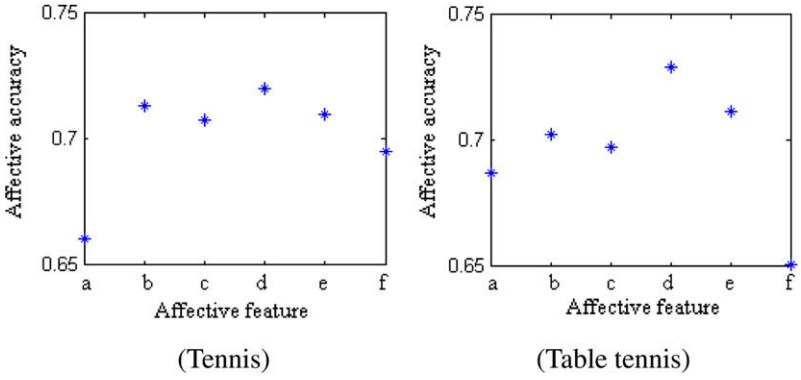

Fig. 6. Single affective feature.

Table 11
Highlights ranking results in tennis videos

Train data	Test data	Affective accuracy (%)	
		SVM regression	Linear regression
T1	T2, T3, T4	83.3	81.8
T4	T1, T2, T3	82.5	83.4
T1, T2	T3, T4	79.5	79.0
T3, T4	T1, T2	84.2	83.7
Average		82.4	82.0

Table 12
Highlights ranking results in table tennis videos

Train data	Test data	Affective accuracy (%)	
		SVM regression	Linear regression
P1	P2, P3, P4	79.3	77.9
P4	P1, P2, P3	78.9	77.0
P1, P2	P3, P4	85.2	83.1
P3, P4	P1, P2	80.6	79.0
Average		81.0	79.3

siderable highlight ranking result since it is obtained fully automatically by computer. This result shows that the determined affective features can reflect human perception to a large extent. Furthermore, it shows that under some conditions computer can learn from human perception for automatic video content understanding.

5.3.2. Flexible summarization by highlight ranking

As demonstrated in Fig. 1, users can choose to observe the most exciting parts by browsing these scenes. If they want to obtain more exciting video content, they can observe it by selecting rank 2, rank 3 and so on. Flexible summarization can be easily gained by selecting the highlights rank or specifying its duration. Fig. 7 illustrates the summarization result by highlight ranking. It can be observed that both "game point" and "match point" are either in first rank or second rank. Please pay attention to the score in bracket in Fig. 7.

6. Conclusions and further work

In this paper, we have presented a flexible racquet sports video summarization framework by fusing multiple modalities. The temporal voting strategy and highlight ranking enable the scheme to work properly on different racquet sports videos. Temporal voting strategy is a fusion strategy of unsupervised shot clustering and supervised audio classification, thus both of the characteristics of generalization of unsupervised method and the semantic meaning of audio are used. A subjective evaluation criterion for sports highlight is adopted. It is a guide for affective features selection and highlight modeling. Four affective features are selected and the linear highlight model is proved to be reasonable. The experimental results on racquet sports video are encouraging. The output of the proposed system provides the user an effective way for flexible summarization, which may have many potential applications for wireless video browsing, video retrieval and video-on-demand accessing, etc.

Currently, the scheme is only tested on racquet sports video, while we believe that the framework is also applicable to more types of sports, which have similar specific correspondence between sound and scene with periodicity. This kind of sports should have periodic pattern and specific correspondence between sound and scene with periodicity. Therefore, we can use time voting strategy for play events detection. Meanwhile, the highlight ranking for play events is generic to any sports video. Thus the only condition is that the sport has specific sound for play cluster recognition. In

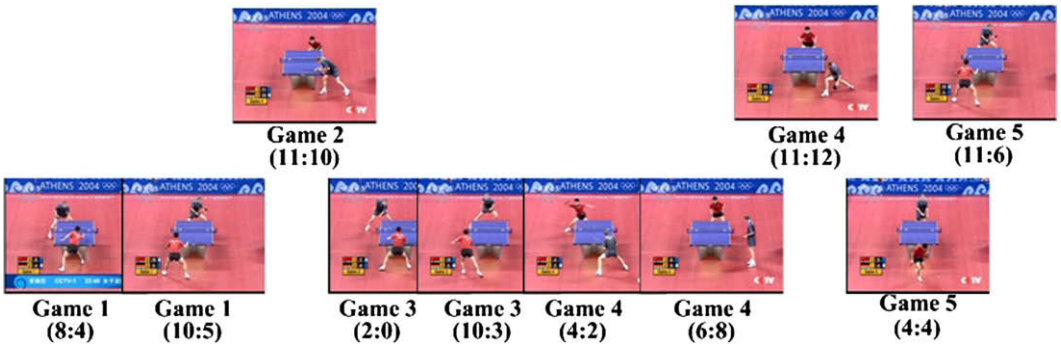

Fig. 7. The $r = 0$ (top) and $r = 1$ (bottom) summarization of table tennis P1.

future work, the framework may be extended to other sports videos.

Acknowledgments

This work is supported partly by Science100 Plan of Chinese Academy of Sciences: 99T3002T03. We would also thank the anonymous reviewers for their valuable comments.

References

[1] Y. Gong, T.S. Lim, H.C. Chua, Automatic parsing of TV soccer programs, in: Proc. Int. Conf. Multimedia Computing and Systems, Washington, DC, USA, May 1995, pp. 167–170.
[2] L. Xie, S.-F. Chang, A. Divakaran, H. Sun, Structure analysis of soccer video with hidden Markov models, in: Proc. Int. Conf. Acoustic, Speech and Signal Processing, Orlando, USA, May 2002, vol. 4, pp. 4096–4099.
[3] A. Ekin, A.M. Tekalp, R. Mehrotra, Automatic soccer video analysis and summarization, IEEE Transaction on Image Processing 12 (7) (2003) 796–807.
[4] Y. Rui, A. Gupta, A. Acero, Automatically extracting highlights for TV baseball programs, in: Proc. ACM Multimedia, Marina del Rey, CA, USA, October 2000, pp. 105–115.
[5] M. Xu, L.-Y. Duan, C.-S. Xu, Q. Tian, A fusion scheme of visual and auditory modalities for event detection in sports video, in: Proc. Int. Conf. Acoustic, Speech and Signal Processing, Hong Kong, China, April 2003, vol. 3, pp. 189–192.
[6] Z. Xiong, R. Radhakrishnan, A. Divakaran, T.S. Huang, Audio events detection based highlights extraction from baseball, golf and soccer games in a unified framework, in: Proc. Int. Conf. Acoustic, Speech and Signal Processing, Hong Kong, China, April 2003, vol. 5, pp. 632–635.
[7] C.G.M. Snoek, M. Worring, Multimodal video indexing: a review of the state-of-the-art, Multimedia Tools and Applications 25 (1) (2005) 5–35.
[8] Y. Gong, M. Han, W. Hua, W. Xu, Maximum entropy model-based baseball highlight detection and classification, Computer Vision and Image Understanding 96 (2) (2004) 181–199.
[9] R. Leonardi, P. Migliorati, M. Prandini, Semantic indexing of soccer audio–visual sequences: a multimodal approach based on controlled Markov chains, IEEE Transaction on Circuits System and Video Technology 14 (5) (2004) 634–643.
[10] J. Huang, Z. Liu, Y. Wang, Y. Chen, E.K. Wong, Integration of multimodal features for video scene classification based on HMM, in: Proc. Int. Workshop on Multimedia Signal Processing, Denmark, September 1999, pp. 53–58.
[11] M. Petkovic, V. Mihajlovic, W. Jonker, S. Kajan, Multimodal extraction of highlights from TV formula 1 programs, in: Proc. Int. Conf. Multimedia and Expo, Switzerland, August 2002, vol. 1, pp. 817–820.
[12] S. Nepal, U. Srinivasan, G. Reynolds, Automatic detection of 'goal' segments in basketball videos, in: Proc. ACM Multimedia, Ottawa, Canada, October 2001, pp. 261–269.
[13] A. Hanjalic, Generic approach to highlights extraction from a sports video, in: Proc. Int. Conf. Image Processing, Spain, September 2003, vol. 1, pp. 1–4.
[14] Z. Xiong, R. Radhakrishnan, A. Divakaran, Generation of sports highlights using motion activity in combination with a common audio feature extraction framework, in: Proc. Int. Conf. Image Processing, Catalonia, Spain, September 2003, vol. 1, pp. 14–17.
[15] G. Zhu, Q.-M. Huang, C.-S. Xu, L.-Y. Xing, W. Gao, H.-X. Yao, Human behavior analysis for highlight ranking in broadcast racket sports video, IEEE Transaction on Multimedia 9 (6) (2007) 1167–1182.
[16] B. Zhang, W.-B. Dou, L.-M. Chen, Audio content-based highlight detection using adaptive Hidden Markov Model, in: Proc. Int. Conf. Intelligent Systems Design and Applications, Shandong, China, vol. 1, 2006, pp. 798–803.
[17] B. Zhang, W. Chen, W.-B. Dou, J. Yu, L.-M. Chen, Content-based table tennis games highlight detection utilizing audiovisual clues, in: Proc. Int. Conf. Image and Graphics, Sichuan, China, 2007, pp. 833–838.
[18] W. Chen, Y.-J. Zhang, Tracking ball and players with applications to highlight ranking of broadcasting table tennis video, in: Proc. Int. Conf. Computational Engineering in Systems Applications, Beijing, China, October 2006, vol. 2, pp. 1896–1903.
[19] H. Lu, Y.-P. Tan, Unsupervised clustering of dominant scenes in sports video, Pattern Recognition Letters 24 (15) (2003) 2651–2662.
[20] T. Mei, Y.-F. Ma, H.-Q. Zhou, W.-Y. Ma, H.-J. Zhang, Sports video mining with mosaic, in: Proc. Int. Conf. Multi-media Modeling, Melbourne, Australia, January 2005, pp. 107–114.
[21] L.Y. Duan, M. Xu, Q. Tian, C.S. Xu, J.S. Jin, A unified framework for semantic shot classification in sports video, IEEE Transactions on Multimedia 7 (6) (2005) 1066–1083.
[22] W.-G. Zhang, Q.-X Ye, L.-Y. Xing. Q.-M. Huang, W. Gao, Unsupervised sports video scene clustering and its applications to story units detection, in: Proc. SPIE. Visual Communications and Image Processing, Beijing, China, July 2005, vol. 5960, pp. 446–455.
[23] L. Lu, H.-J. Zhang, S.-Z. Li, Content-based audio classification and segmentation by using support vector machines, Multimedia Systems 8 (6) (2003) 482–492.
[24] T. Zhang, C.-C.J. Kuo, Audio content analysis for online audiovisual data segmentation and classification, IEEE Transaction on Speech and Audio Processing 9 (4) (2001) 441–457.
[25] Y. Wang, Z. Liu, J. Huang, Multimedia content analysis using audio and visual information, IEEE Transaction on Signal Processing Magazine 17 (4) (2000) 12–36.
[26] A.-K. Jain, Statistical pattern recognition: a review, IEEE Transaction on Pattern Analysis and Machine Intelligence 22 (1) (2001) 4–37.
[27] B. Scholkopf, C. Burges, A.-J. Smola, Advances in kernel methods: support vector machine, MIT Press, 1999.
[28] A. Hanjalic, L.-Q. Xu, Affective video content representation and modeling, IEEE Transaction on Multimedia 17 (1) (2005) 143–154.
[29] K.-A. Peker, A. Divakaran, Framework for measurement of the intensity of motion activity of video segments, in: Proc. SPIE Internet Multimedia Management Systems III, July 2002, vol. 4862, pp. 126–137.
[30] K.-A. Peker, R. Cabasson, A. Divakaran, Rapid generation of sports video highlights using the MPEG-7 motion activity descriptor, in: Proc. SPIE Storage and Retrieval for Media Databases, San Jose, CA, January 2002, vol. 4676, pp. 318–323.
[31] K.A. Peker, A. Divakaran, Framework for measurement of the intensity of motion activity of video segments, Journal of Visual Communication and Image Representation 15 (3) (2004) 265–284.