Journal of Zhejiang University SCIENCE A
ISSN 1009-3095 (Print); ISSN 1862-1775 (Online)
www.zju.edu.cn/jzus; www.springerlink.com
E-mail: jzus@zju.edu.cn

Sports video summarization and adaptation for application in mobile communication*

GAO Wen[1], HUANG Qing-ming[1,2], JIANG Shu-qiang[1], ZHANG Peng[1]

(*[1]Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100080, China*)

(*[2]Graduate School, Chinese Academy of Sciences, Beijing 100039, China*)

E-mail: wgao@jdl.ac.cn; qmhuang@jdl.ac.cn; sqjiang@jdl.ac.cn; peng.zhang@jdl.ac.cn

Received Dec. 14, 2005; revision accepted Feb. 25, 2006

Abstract: Sports video appeals to large audiences due to its high commercial potentials. Automatically extracting useful semantic information and generating highlight summary from sports video to facilitate users' accessing requirements is an important problem, especially in the forthcoming broadband mobile communication and the need for users to access their multimedia information of interest from anywhere at anytime with their most convenient digital equipments. In this paper, a system to generate highlight summaries oriented for mobile applications is introduced, which includes highlight extraction and video adaptation. In this system, several highlight extraction techniques are provided for field sports video and racket sports video by using multi-modal information. To enhance users' viewing experience and save bandwidth, 3D animation from highlight segment is also generated. As an important procedure to make video analysis results universally applicable, video transcoding techniques are applied to adapt the video for mobile communication environment and user preference. Experimental results are encouraging and show the advantage and feasibility of the system for multimedia content personalization, enhancement and adaptation to meet different user preference and network/device requirements.

Key words: Video analysis, Video transcoding, Highlight summary, Animation generation, Mobile communication
doi:10.1631/jzus.2006.A0819 **Document code:** A **CLC number:** TN919.8

INTRODUCTION

We are entering the era of broadband mobile communication with the upcoming 3G transmission systems, which will provide high-quality multimedia content service to the rapidly growing mobile market. The availability of reliable connections through various handheld mobile devices such as intelligent cell phone and PDA has brought users the possibility of accessing their video information of interest from anywhere at anytime by using their most convenient accessing mode. The problem of video service is that the content is huge and crowded with much redundant information so that it often takes a long time to browse the content from the beginning to the end. The video limitations lie in the cost of data transfer that is paid by the users and the displaying and power capability of the digital devices, as well as the users' patience to endure the longtime boring information that cannot draw any interest from them. In this case, transmission of meaningful abstract summaries extracted from continuous video streams occurred at the significant highlight events as well as adaptation for the network environment and users' preference becomes an important problem.

Sports video program always appeals to large audiences. Automatically extracting useful semantic information from sports video to facilitate users' accessing requirements is an important problem which has emerged as a hot research area recently due to its high commercial potentials especially with the upcoming popularity of broadband multimedia mo-

* Project supported by NEC Research of China (No. 0P2004001), "Science 100 Plan" of the Chinese Academy of Sciences (No. m2041), and the Natural Science Foundation (No. 4063041) of Beijing, China

bile communication. As a typical video application over mobile networks, most users may want to obtain the sports video highlights in a game within a limited time. Therefore, a system is desired to automatically generate highlight summary from sports video, and subsequently to transform the analysis results into certain visual format by implementing content-based adaptation methods.

As a novel component of the system, 3D transformation from broadcast video can also be incorporated to meet users' high-level requirement. As in broadcast video, only a single viewpoint is available to the users, rather than 3D viewing environments that we are more accustomed to. Automatically transforming 2D broadcast video into 3D animations can allow users to select the viewing effect from any viewpoint at any distance. This could not only enhance the viewing experience and enhance visual interest for users, but also save much bandwidth, as only motion parameters are transmitted with player models embedded in the digital equipment.

On the other hand, to meet the constraints of transmission bandwidth and terminal capabilities as well as user preference, transcoding is an important technique to obtain the desired video content. H.264/AVC is the newest video-coding standard designed for a variety of multimedia applications over hybrid network environments. Its high performance gives it a promising future in mobile video applications. However, as most multimedia resources are created and stored in MPEG-2 format, such as DVD and DVB, a flexible adaptation method from MPEG-2 bitstream to H.264/AVC bitstream is highly desirable for content based video applications over mobile networks. In general, format conversion and dynamic bit rate scaling are most important tools in video adaptation.

Recently researchers have paid attention to the emerging and promising field of video analysis and adaptation (Chang, 2003; Bertini *et al*., 2004; Vetro *et al*., 2003). Chang (2003) presented a general overview of video summarization and format transformation, but did not provide specific video processing methods. A performance measure to incorporate user preference and media annotation errors in the adaptation part was proposed by Bertini *et al*.(2004). The semantic annotation is used to generate highlight video clips. Video adaptation in (Vetro *et al*., 2003) is also used for surveillance purpose.

In this paper, a system to generate highlight summaries and adapt them for mobile communication applications is introduced. The architecture of the proposed system is illustrated in Fig.1. Several highlight extraction techniques are provided in this system for field sports video and racket sports video by using multi-modal information. To extract important events from field sports video, an incremental learning method is employed to obtain better results. By fully taking advantages of the periodicity of video scenes in racquet sports video, a flexible video content summarization solution combining structure parsing module and linear highlights ranking model is proposed. Generating 3D animation from highlight clips by using computer vision and computer graphics techniques are also developed to enhance users' viewing experience and save bandwidth. Video transcoding is an important procedure to make the video analysis results applicable for mobile communication environment and user preference. In particular, two video transcoding techniques are integrated into our video adaptation engine: MPEG-2 to H.264/AVC format conversion with spatial resolution reduction transcoding and dynamic bit rate reduction transcoding for H.264/AVC streams. Results of the proposed system both in video analysis stage and video adaptation stage could satisfy the requirements

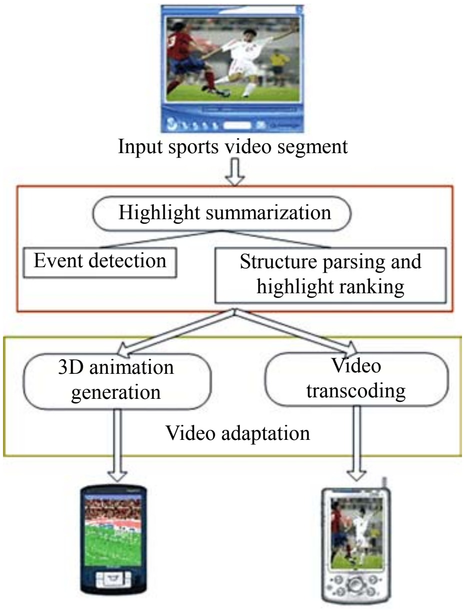


Fig.1 A general scheme of the proposed system

for mobile communication environment and user preference. Experimental results are encouraging and show the advantage and feasibility of the system for multimedia content summarization, enhancement and adaptation.

The organization of the paper is as follows. In Section 2, techniques of highlight summarization from sports video are provided. Three dimensional animation generation and video transcoding are discussed in Sections 3 and 4, respectively. Experimental results are presented and discussed in Section 5. The paper is concluded in Section 6.

HIGHLIGHT SUMMARY EXTRACTION

Many kinds of sports types can be roughly classified into two categories: field sports video (e.g., soccer, baseball, football, etc.) and racquet sports video (e.g. tennis, table tennis, badminton, etc.). Most research works focus on the former category and little attention is paid to the latter, which contains repetitive video and audio structure throughout a match game so that exact highlights are difficult to define. Solutions proposed to extract highlight summaries include single-modal features (Gong *et al*., 1995; Xie *et al*., 2003; Rui *et al*., 2000; Ekin *et al*., 2003) which include visual/audio, and multi-modal features (Snoek and Worring, 2005; Kijak *et al*., 2003; Leonardi *et al*., 2004) which usually combine visual, audio and text information to implement their work. These systems are mostly designed for special exciting event detection by integrating domain knowledge, and their expansibility needs to be improved. Especially these methods are hard to extend to periodic and structural sports games such as racquet sports. In this section, highlight extraction methods are proposed for not only field sports, but also for racquet sports.

Event detection

A new method for event detection in broadcast field video based on mid-level visual description and incremental SVM learning is investigated by taking soccer video as example. First, for each video frame, the playfield area is classified based on low-level features, and then frames are hierarchically classified into defined views. Finally, view label, camera motion and shot boundary descriptions together with temporal relationship are fed into an SVM classifier to identify events. In our implementation, “shoot on goal”, “placed kick”, “break by offence” are selected for experiments.

To segment the playfield, training models are automatically built by Gaussian Mixture Models (GMMs) (Liu *et al*., 2005a; 2005b). The model is updated online by an incremental procedure. Then three kinds of mid-level descriptions are extracted to represent the soccer video content, which is an effective way to the event detection problem. For each soccer frame, we can semantically assign a view label to them by a hierarchical classification procedure. The views are: “goal mouth view”, “corner view”, “middle field view”, “player close-up view” and “out-field view”. For each frame we extract low-level features of playfield ratio, projection profile of non playfield, and shape for classification. SVM classifiers are employed to perform the classification task on extracted features. An incremental scheme is adopted to improve the extensibility.

For each soccer view, three kinds of mid-level descriptors: “view label”, “camera motion” and “shot length”, are extracted. The values of “Zoom” (z), “Pan” (p) and “Tilt” (t) of each video frame are obtained by a standard global motion estimation algorithm.

On the frame labelling result, camera motion in a labelled view can be described by binary features as: (1) MS: More than 80 percent of the frames are salient 0/1; (2) SL: Existing from salient to large pan/tilt/zoom 0/1; (3) LM: Existing large pan/tilt/zoom 0/1.

Shot length is meaningful for event detection. We combine the color histogram, edge distribution features and corner points features for shot boundary detection by an HMM shot detection model.

Based on the obtained mid-level descriptors and their temporal relationship, event detection by SVM-based incremental learning is carried out. In the soccer game, exciting events such as “shoot on goal” rarely occurs which makes it difficult to label a large number of training samples of them, so statistical model is not reliable for the events of small samples. Therefore, SVM classifier is employed for event detection task. An SVM-based incremental learning method is also applied to improve the extensibility of event detection models. Incremental learning can be found in (Cauwenberghs and Poggio, 2001).

Structure parsing and highlight ranking

A general method of highlights ranking for sports video is described in Fig.2. In the following, we will provide a detailed description of the process by taking racquet sports as examples.

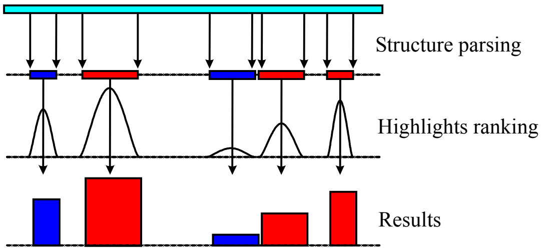


Fig.2 Framework of highlight ranking

Racquet sport is a typical sports type. They have well defined temporal structure because of the alteration of rally and break scenes and alteration of audio patterns. A novel temporal voting scheme is proposed for structure parsing to detect rallies by making use of the periodicity in audio and video information. First, we use a hierarchal merging method to group similar shots together. Each shot is represented by 5 key frames. Color histogram in HSV color space and edge histogram are employed as the low-level features of these frames. To determine the number of clusters, we propose a method to determine the merging-stop criteria by analyzing the *J* value which is defined as the ratio of the Inter class variance and Intra class variance on low-level visual features.

Audio types in most racquet sports can be classified into silence, ball impact, cheer, excited speech and plain speech. Supervised audio classification is composed of feature extraction, selection and classifier design. Then, a forward search algorithm is used to perform the feature selection task before they are fed into the SVM classifier. Compared with other classifiers, SVM is easier to train, needs fewer training samples and has better generalization ability.

The video scene classification resulting from unsupervised method has no semantic meaning. Therefore, a strategy is proposed to assign the clustered video scene with the semantic meaning by audio-video correspondence based on the fact that the audio information is more related to the state of a match. The temporal fusion method first applies audio symbols to find the clusters according to the temporal relationship between the audio and the shot in a cluster. Then the voting strategy makes the cluster obtain its meaning by the voted audio symbol with the highest voting value. Both the characteristics of the robustness and generalization of the unsupervised method and the definite meaning of audio are fully utilized in this part. The obvious advantage is its generality, since the voting foundation is the time stamp which is the basic information in the audio/video stream. After the voting procedure, video scenes are segmented with semantic meanings, and rally clips are detected. These results are used to further rank highlight and generate video summaries.

We propose a subjective criterion that can tell to what extent the highlights reflect human perception and can guide excitement components (affective features) selecting and exciting degree (highlights) modelling. Affective features are selected to express the highlights, which includes motion, cheers and pitch-related features, and shot length, respectively. Based on the observation in our application, six affective features are extracted. 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. Let us define the highlights rank of each event as an integer *r* which is 0 to *R*′, where highlight rank *R*′ is automatically decided in optimal quantization process in order to minimize the error between the continuous absolute value and the discrete relative level. Suppose that the segmented structure type (event) is *M*, the average continuous absolute value of event *m* is H_m in terms of subjective perception, and its corresponding discrete relative level of event *m* is $H_m(r)$. The formulation is as follows:

$$H_m(r) = r,\ \text{if}\ r/R \le H_m < (r+1)/R, \tag{1}$$

$$err_R = \sum_{m=1}^{M} | H_m(r) - H_m |R, \tag{2}$$

$$R' = \arg\min_R (err_R), \tag{3}$$

where 0≤*r*<*R*. As long as *R*′ is obtained, the continuous absolute value can be converted to the discrete relative level with the lowest quantization error.

Suppose that the test video contains *M*′ events and that the regression value is C_m, and $C_m(r)$ is the corresponding discrete relative level after quantization process with the R'. Then, based on the ground truth $H_m(r)$ and the highlight rank R', we can evaluate the highlights ranking result $C_m(r)$ by computing:

$$\text{Affective accuracy} = \frac{1}{M'}\sum_{m=0}^{M'}\frac{R' - |H_m(r) - C_m(r)|}{R'} = \frac{1}{M'}\sum_{m=0}^{M'}[1 - |H'_m(r) - C'_{\mathrm{m}}(r)|], \quad (4)$$

where $|H'_m(r) - C'_m(r)|$ represents the relative bias between highlights ranked by human and computer, so the change of R' which is selected according to 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.(4) shows that the accuracy is obtained by averaging the human-computer rank bias.

Finally, based on the subjective evaluation criterion, we use a forward search algorithm (Jain, 2001) to select the affective features and nonlinear as well as linear regression to model highlights for gaining further insight. The experiment in tennis video shows that commonly used affective features are correlated; and that the highlights model is approximately linear.

3D ANIMATION GENERATION: HIGHLIGHT SEGMENT ENHANCEMENT

Just as can be seen from the previous section, highlight summarization can compress the original video to a large extent. It is well known that only a single viewpoint is available to the viewers at any time in sports video program, although many cameras are deployed around the playfield. In this section, a solution technique that can generate 3D animation from the original soccer video clip is proposed. It allows users to watch the game from arbitrary point of view to enhance the viewing experience.

Some related work exists in (Bebie and Bieri, 1998; Yu *et al*., 2004). The most related work is that of (Matsui *et al*., 1998). Their system can also generate animation from broadcast soccer video. However, our work differs from theirs as we adopt a new camera calibration method which can obtain camera calibration parameters even when the corresponding points are insufficient. In addition, our system can obtain the ball's 3D positions under certain assumptions. Fig.3 provides a flowchart of the proposed solution. It consists of two main parts: 3D information extraction and 3D animation generation. 3D information extraction includes camera calibration, multiple objects detection and tracking, and player and ball's 3D position estimation. 3D animation generation includes playfield modelling and player modelling.

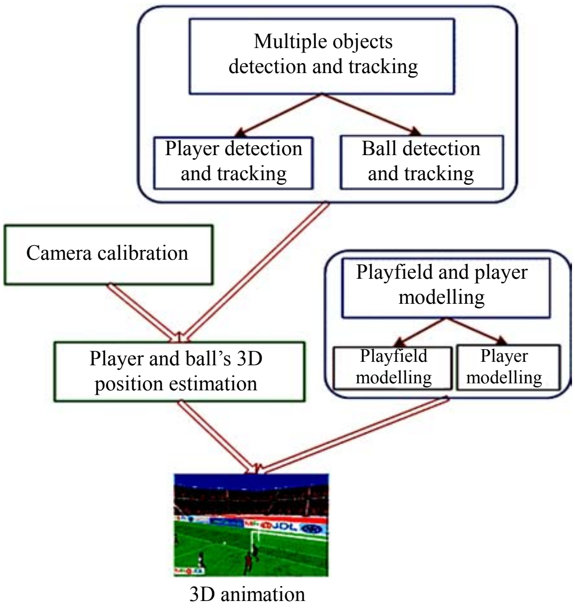

Multiple objects detection and tracking
Player detection and tracking
Ball detection and tracking
Camera calibration
Playfield and player modelling
Player and ball's 3D position estimation
Playfield modelling
Player modelling
3D animation

Fig.3 Flowchart of 3D animation generation

The demo of 3D animation generation can be accessed from http://www.jdl.ac.cn/en/project/mrhomepage/demo.htm#video2cartoon.

3D information extraction

In order to obtain player's and ball's 3D position, camera must be calibrated for each frame. Since the playfield is a plane, calibration is reduced to estimating the transformation between the playfield and its image, which is a 3 by 3 matrix called homography. Here a newly developed method (Liu *et al*., 2005b) is adopted. If there are more than four corresponding points, with no three of them collinear, the homography can be estimated directly. For the frames with insufficient corresponding points, global motion estimation (GME) and the calibrated frames are used to compute their homographies. Multiple objects detection and tracking approach based on support vector regression (SVR) particle filter is adopted. SVR particle filter is an improved particle filter that can achieve better performance with small particle set and enhance the efficiency of the tracking system. Object detection based on playfield detection is also combined into the tracking framework, which makes the tracking algorithm fully automatic.

It is reasonable to assume that the players move on the ground of the playfield in most cases, therefore, the 3D position can be obtained by homography. The bottom line midpoint of the minimal bounding box of the player is regarded as his image position. Under the assumption that the ball follows a parabolic trajectory in the air, the 3D position can be obtained by finding the intersection point of a line and a plane. The line is determined by the camera position (obtained by self-calibration) and the shadow position (the corresponding world position of the ball's image) of the ball on the playfield. The plane is perpendicular to the playfield and contains the parabolic trajectory of the ball.

3D information generation

Playfield modelling is done according to the Laws of the Game of FIFA. To enhance the viewing experience, the playfield is not only texture-mapped using the playfield detection result, but also decorated with some auxiliary materials, such as billboards, racetrack, auditoria and so on. The player model is built according to H-anim1.1, which is the specification for a standard humanoid, and hence capable of complex actions.

The key issue of 3D animation generation is the players' motion. Player motion database is built in advance, including some simple actions such as running, walking, etc. A preprocessing procedure is carried out to smooth the players' trajectories on the playfield, and after that, players' velocities are calculated to control the players' motion direction and motion type (e.g. running, walking, etc.). However, the players' poses can hardly to be recovered, much information is lost during imaging.

The system is implemented based on OpenGL in Visual C++ 6.0. It is not fully automatic, mainly because of the confusion before a goal that leads to the tracking failure, and the difficulty to automatically determine the ball's starting and ending points on the playfield. These cases can be resolved by human intervention. In the prototype, the users can control a virtual camera to pan, tilt, zoom and change viewpoint, and can also roam around the playfield using the direction key and the mouse.

VIDEO TRANSCODING

In this section, two transcoding techniques are implemented in our video adaptation engine. One is to convert an MPEG-2 bitstream into an H.264/AVC bitstream with only half spatial resolution and the other is H.264 dynamic bit rate reduction transcoder to meet the bandwidth variation in mobile network environment. Details of the two transcoder are introduced below.

MPEG-2 to H.264/AVC format convert transcoding with spatial resolution reduction

In the literature, several spatial resolution transcoding methods are all designed for fixed-size block-based motion estimation. They have been mainly focused on mode decision, motion vectors mapping, re-estimation and refinement (Bjork and Christopoulos, 1998; Shanableh and Ghanbari, 2000). However, it is not suitable for the MPEG-2 to H.264/AVC transcoding system because of the variable block-size motion estimation feature adopted in H.264/AVC. In this section, we propose a new method to convert an MPEG-2 stream into an H.264/AVC stream with half of the spatial resolution. This method is a solution to reduce the complexity of mode decision and motion estimation, as this is a key technique for transcoding. Compared with related works, the feature of variable block-size motion estimation is fully exploited to obtain an adaptive transcoding method suitable for the H.264/AVC standard in our method. A new hybrid algorithm for downscale transcoding from MPEG-2 to H.264/AVC is proposed.

In this hybrid system, we first fully decode the MPEG-2 bit stream and record the coding type, motion vectors and residuals of every pre-encoded macroblock. After downscaling by factor 2, four pre-encoded macroblocks turn into four 8×8 blocks with each 8×8 block having a motion vector. Based on the four 8×8 blocks, we perform merge or split operation and make full use of the multiple modes feature in motion estimation in the H.264/AVC standard. Inter_16×16, inter_16×8, inter_8×16 and inter_8×8 are used in this hybrid architecture. Furthermore, if the inter_8×8 is selected, all the sub-block modes such as 8×4, 4×8 and 4×4, are considered. Details of merge and split techniques were proposed in (Hu *et al*., 2005).

Bit rate reduction transcoding for H.264/AVC streams

Bit rate reduction is one of the basic problems which aim to reduce the bit rate while maintaining low complexity and achieving the highest quality. In the literature, there are two architectures, the open-loop (Sun *et al*., 1996) and the close-loop (Assunçno and Ghanbari, 1996) for solving the bit-rate reduction transcoding problem.

However, both solutions assume that there is no mode variation when the target QP varies. This is true for MPEG-2 and H.263, because there are few mode options in them. As H.264/AVC appears, the problem becomes different as there are much more mode options in it. Reusing the mode information from the input streams will cause severe additional performance degradation because of the mismatch of prediction mode. In this section, we propose a new transcoding architecture which reduces the complexity by restricting mode searching range by limited R-D optimization (L-RDO). The key issues of the method are listed as follows: (1) Mode selection in Intra frames: intra frames support 4×4 (I4MB), 16×16 (I16MB), and I_PCM. I4MB supports 9 directions and I16MB supports 4 directions. To achieve best performance, we use exhaustive searching instead of other fast intra mode decision algorithms. (2) Inter mode selection: as we assume that the input streams are coded optimally, the information extracted from input streams can be used in our proposed L-RDO in the transcoder. (3) Prediction directions in B-frames: In B slices, four different types of inter-picture prediction are supported: foreward, backward, bi-predictive, and direct mode. As the target bit rate changes, these inter prediction types can be reusable. (4) Multi-reference frames: H.264 allows multiple reference frames for prediction. A maximum of five reference frames may be used for prediction. Each block in a macroblock can be predicted by different reference frames. (5) Motion vector refinement: Motion estimation is the most computationally heavy module. A high performance transcoder requires a simple but effective MV mapping scheme based on resolving the above four problems.

In our proposed solution, computational complexity is reduced by skipping unnecessary mode and reference frame selection as well as effective MV mapping. If a precise R-D model can be established to further restrict mode searching range, the complexity can be reduced further. Furthermore, fast mode decision algorithms and motion estimation algorithms can be adopted in this work to reduce the complexity further. Another advantage of our proposed transcoding architecture is its complexity scalability in practical applications. Trade-offs between quality and complexity can be easily achieved.

EXPERIMENTAL RESULTS

Event detection

Totally 15 soccer video segments with about 400 min from 2004 Europe Cup and England soccer matches are used for our experiments, about half of the video content are used as training set. All the videos are compressed in MPEG-1 with 25 fps and frame resolution of 352×288. Table 1 reporting the experimental result of view classification shows that when training examples and test examples are extracted from the same video set, the classification accuracy is high. When we change the test set, the classification result will drop a lot, while the classification accuracy will increase by introducing the incremental learning procedure.

Table 1 View classification result

Sequence	View classification accuracy (%)
Without changing video set	92.5
Changing video set	83.4
Incremented by 10 frames	86.7
Incremented by 20 frames	90.1
Incremented by 30 frames	91.0

We select three most representative exciting events for experiment. It can be found that the classification result is improved a lot with an incremental learning procedure, which can ensure that the method can be easily extended to different styles of soccer video with little human labor. The detailed result can be found in Table 2.

Table 2 Event classification result

Sequence	Event detection accuracy (%)
Without changing video set	84.6
Before incremental learning	71.1
Incremented by 10 frames	76.4
Incremented by 20 frames	78.4
Incremented by 30 frames	78.9

Highlight ranking

We prepare 10447 s tennis videos for experiments. The tennis videos are from 6 different live broadcast programs of French Open 2005. Video information is listed in Table 3.

Table 3 Data on tennis videos

Video	*a*	*b*	*c*	*d*	*e*	*f*
Length	18:31	31:54	37:08	17:24	29:45	39:25

1. Rally detection by temporal voting strategy

Videos from *a* to *f* in tennis are selected for rally scenes segmentation experiment. We mainly test rally scenes segmentation performance because they are basic units for highlights ranking procedure and it is important to correctly segment them. The segmentation results are listed in Table 4. As for table tennis video, the final rally scene segmentation precision is 89.0% with 85.4% recall rate which is an encouraging result for real applications.

Table 4 Rally segmentation results in tennis videos

Video clip	By audio information only		By audio/visual combination	
	Precision (%)	Recall (%)	Precision (%)	Recall (%)
c	77.3	80.6	93.1	85.1
d	75.9	74.8	84.7	89.4
e	72.4	77.6	85.8	86.5
f	68.8	70.1	94.4	81.3
g	69.7	78.4	89.6	86.7
Average	72.0	77.1	89.0	85.4

To prove the advantages of audio-visual combination method for scene segmentation, scene segmentation results by pure audio are presented in Table 4 for comparison. It can be seen from the table that the "rally" scenes segmented by audio-visual combination are much better than those by audio information only. Pure visual information can be used for scene segmentation but cannot assign semantic meaning to segmented scenes as stated above.

2. Highlights ranking for summarization

We need to get to the bottom of whether the linear regression model is effective for our highlights ranking task. A nonlinear model (SVM regression model) is selected for comparison. The reason 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.

The comparison results of nonlinear and linear regression are listed in Table 5 after the selected affective features *a*, *b*, *d* and *f* are fed into the regression model. It can be seen that there is nearly no improvement by using nonlinear regression (SVM regression). We can then conclude that highlights ranking problem can be modelled as a linear regression problem.

Table 5 Highlights ranking results in tennis videos

Training data	Test data	Affective accuracy (%)	
		SVM regression	Linear regression
c	*d*, *e*, *f*	81.5	83.1
d	*c*, *e*, *f*	85.7	83.5
c, *d*	*e*, *f*	82.2	85.4
e, *f*	*c*, *d*	83.6	81.4
Average		83.3	83.4

It can also be seen that the affective accuracy reaches to around 83.0% in terms of the ground truth and evaluation criteria. We must make it clear that 83.3% (83.4%) affective accuracy is a marvelous highlights ranking result since it was 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 in some special conditions a computer can learn from human perception for automatic video content understanding.

Video transcoding

This section presents the experimental results of video transcoding of format conversion and dynamic bit rate reduction. We first provide comparison of the performance of MPEG-2 to H.264/AVC in different ways, and then demonstrate the results of dynamic bit rate reduction by a cascaded decoder-encoder with RDO enabled, and another reusing the MB mode of input stream for comparison. All the experiments are based on JM9.4.

1. Result of proposed MPEG-2 to H.264/AVC with spatial resolution downscaling adaptation

In this simulation, the input MPEG-2 bitstreams have CIF resolution (352×288) at 2 Mbps. Each group of pictures (GOP) contains 10 frames just including I- and P-frames (B-frames can be processed similarly as P-frames). The target H.264/AVC bitstream has QCIF resolution (176×144). We use several scenes with different motion activity as our test sequence. For each sequence, four transcoders are simulated which are different at the re-encoding process.

(1) Full mode: Encode with all the supported modes in H.264/AVC and do an exhaustive search for motion estimation.

(2) Only inter_16×16: Besides the intra and skipped type in H.264/AVC, only inter_16×16 mode is selected for encoding the inter frame. Also do an exhaustive search for motion estimation.

(3) Only inter_8×8: Besides the intra and skipped type in H.264/AVC, only inter_8×8 mode is selected for encoding the inter frame. Also do an exhaustive search for motion estimation.

(4) Proposed method: Our proposed transcoding architecture.

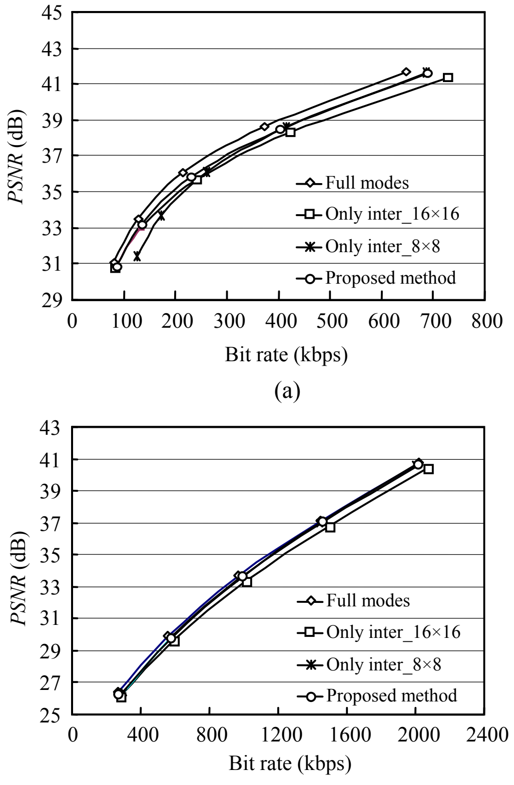
Fig.4 shows the R-D performance of the four sequences. It can be seen from Fig.4 that our proposed method has steady performance not only at low bit rate but also at high bit rate. So we can conclude that, for some sequences with low motion activity, the merge procedure will take effect. While the same image quality is maintained, many bits are saved from motion vectors during re-encoding. On the other hand, for some sequences with high motion activity, the partition procedure will make more accurate prediction with small block-size.

In the meantime, our proposed method has a very low computational complexity and is more practical for real-time transcoding as shown in Fig.5.

2. Results of dynamic bit reduction for H.264/AVC video streams

To compare the performances, all the input streams are pre-coded at 30 fps, 1 Mbps for CIF sequences, with M=3, N=30, and have 5 reference frames. Bit rate of the output streams are transcoded to 768, 576, 384, 256, 192, 128, and 96 kbps, respectively.

Table 6 lists partial results for a set of testing sequences selected to represent a bit-rate scaling transcoding application. The result of our transcoding scheme is close to that of the cascaded decoder-



(a)

(b)

Fig.4 R-D curves of different transcoding methods. (a) Scene-1 with low motion activity; (b) Scene-2 with high motion activity

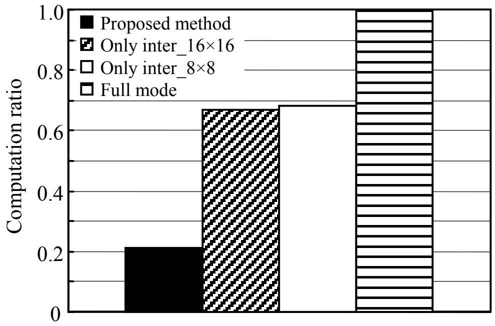


Fig.5 Average computation complexity

Table 6 Experimental results on test sequences

Test sequence	Cascade		Mode reuse		Transcoding without RC		Transcoding with RC	
	Rate (kbps)	*PSNR* (dB)	Rate (kbps)	*PSNR* (dB)	Rate (kbps)	*PSNR* (dB)	Rate (kbps)	*PSNR* (dB)
Shot 1	384.98	39.06	384.94	37.08	385.04	38.19	385.17	38.15
	256.76	35.82	256.99	34.10	256.89	35.43	257.01	35.38
	128.62	32.23	129.26	28.63	128.65	31.34	128.60	31.35
Shot 2	386.86	29.97	386.80	28.81	387.15	30.47	387.07	30.46
	257.38	27.68	279.36	26.80	258.24	27.81	257.37	27.72
	137.26	24.60	178.84	24.23	141.48	24.41	141.38	24.41
Shot 3	514.50	44.04	514.96	41.66	514.43	44.27	514.21	44.29
	386.69	42.24	386.40	40.34	386.00	42.40	386.18	42.39
	257.50	39.74	258.48	38.17	257.44	39.88	257.21	39.91

encoder with RDO, and re-using mode without RDO has more than 1 dB loss. These results also prove that our rate control algorithm is effective.

CONCLUSION

The bandwidth in mobile communication is relatively limited and costly. Most people only care about the highlight scene. As the highlight segment length is very short compared with the original video, bandwidth and cost can be saved through transmitting the highlight segment only. In this system, we provide sports video summarization and adaptation techniques for mobile applications. Several highlight extraction techniques are provided both for field sports video and racket sports video by using multi-modal information. This could save users' viewing time and bandwidth by taking out only the most important video segment. To enhance users' viewing experience and save bandwidth, 3D animation from highlight segment is also generated. On the other hand, video transcoding techniques to adapt the video for mobile communication environment and users' preference are discussed in detail. These techniques are very promising when applied in the cell phone video services, especially for users' requirement to access their multimedia content of interest at the desired time and place by using the most convenient solutions.

ACKNOWLEDGEMENT

The authors sincerely thank Liyuan Xing, Dawei Liang, Bo Hu, Junbiao Pang and Qixiang Ye for their contributions to this paper.

References

Assunçno, P., Ghanbari, M., 1996. Post-processing of MPEG-2 Coded Video for Transmission at Lower Bit-rates. Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing. Atlanta, GA, p.1998-2001.

Bebie, T., Bieri, H., 1998. SoccerMan-reconstructing Soccer Games from Video Sequences. Proc. of ICIP, p.898-902.

Bertini, M., Cucchiara, R., Bimbo, A.D., Prati, A., 2004. Content-based Video Adaptation with User's Preference. IEEE International Conference on Multimedia and Expo (ICME).

Bjork, N., Christopoulos, C., 1998. Transcoder architectures for video coding. *IEEE Trans. Consumer Electron.*, **44**(1): 88-98. [doi:10.1109/30.663734]

Cauwenberghs, G., Poggio, T., 2001. Incremental and Decremental Support Vector Machine Learning, Advances in Neural Information Processing Systems. MIT Press, Cambridge, MA, **13**:409-415.

Chang, S.F., 2003. Content-Based Video Summarization and Adaptation for Ubiquitous Media Access. Proceedings of the 12th International Conference on Image Analysis and Processing (ICIAP'03).

Ekin, A., Tekalp, A.M., Mehrotra, R., 2003. Automatic soccer video analysis and summarization. *IEEE Trans. Image Processing*, **12**(7):796-807. [doi:10.1109/TIP.2003.812758]

Gong, Y., Lim, T.S., Chua, H.C., 1995. Automatic Parsing of TV Soccer Programs. Proc. IEEE Int. Conf. on Multimedia Computing and Systems.

Hu, B., Zhang, P., Huang, Q., Gao, W., 2005. Reducing Spatial Resolution for MPEG-2 to H.264/AVC Transcoding. Proc. on Pacific-Rim Conference on Multimedia, **II**:830-840.

Jain, A.K., 2001. Statistical pattern recognition: a review. *IEEE Trans. on PAMI*, **2**:4-37.

Kijak, E., Gravier, G., Gros, P., Oisel, L., Bimbot, F., 2003. HMM Based Structuring of Tennis Videos Using Visual and Audio Cues. Proc. IEEE Int. Conf. Multimedia and Expo.

Leonardi, R., Migliorati, P., Prandini, M., 2004. Semantic indexing of soccer audio-visual sequences: a multimodal approach based on controlled Markov chains. *IEEE Trans.*

Circuits Syst. Video Techn., **14**(5):634-643. [doi:10.1109/TCSVT.2004.826751]

Liu, Y., Jiang, S.Q., Ye, Q.X., Gao, W., Huang, Q.M., 2005a. Playfield Detection Using Adaptive GMM and Its Application. ICASSP2005. Philadelphia, PA, USA.

Liu, Y., Huang, Q., Ye, Q., Gao, W., 2005b. A New Method to Calculate the Camera Focusing Area and Player Position on Playfield in Soccer Video. Proc. of VCIP.

Matsui, K., Iwase, M., Agata, M., Tanaka, T., Ohnishi, N., 1998. Soccer Image Sequence Computed by a Virtual Camera. Proc. of CVPR, p.860-865.

Rui, Y., Gupta, A., Acero, A., 2000. Automatically Extracting Highlights for TV Baseball Programs. Proc. the Eighth ACM Int. Conf. Multimedia, p.105-115.

Shanableh, T., Ghanbari, M., 2000. Heterogeneous video transcoding to lower spatio-temporal resolutions and different encoding formats. *IEEE Trans. Multimedia*, **2**(2):101-110. [doi:10.1109/6046.845014]

Snoek, C.G.M., Worring, M., 2005. Multimodal video indexing: a review of the state-of-the-art. *Multimedia Tools and Applications*, **25**(7):767-775.

Sun, H., Kwok, W., Zdepski, J., 1996. Architectures for MPEG compressed bitstream scaling. *IEEE Trans. Circuits Syst. Video Technol.*, **6**(2):191-199. [doi:10.1109/76.488826]

Vetro, A., Haga, T., Sumi, K., Sun, H.F., 2003. Object-Based Coding for Long-Term Archive of Surveillance Video. Proceedings of International Conference on Multimedia and Expo.

Xie, L., Chang, S.F., Divakaran, A., Sun, H., 2003. Structure analysis of soccer video with domain knowledge and hidden Markov models. *Pattern Recognition Letters*, **24**(15):767-775.

Yu, X., Yan, X., Hay, T.S., Leong, H.W., 2004. 3D Reconstruction and Enrichment of Broadcast Soccer Video. Proc. of ACM Multimedia.