行政院國家科學委員會專題研究計畫　期中進度報告

總計畫(2/3)

計畫類別：整合型計畫

計畫編號：NSC92-2219-E-002-013-

執行期間：92年08月01日至93年07月31日

執行單位：國立臺灣大學電機工程學系暨研究所

計畫主持人：貝蘇章

共同主持人：李枝宏，陳良基，馮世邁

報告類型：精簡報告

處理方式：本計畫可公開查詢

中　華　民　國 93年5月4日

智慧型音視訊和傳輸技術及多媒體應用(II)–總計畫
Intelligent audio/video/transmission techniques and multimedia applications (II)

計畫編號: NSC-92-2219-E-002-013

執行期限：92 年 8 月 1 日至 93 年 7 月 31 日

主持人：貝蘇章　　台灣大學電機系教授

摘要

運動視訊在大部分的電視頻道是相當主要的節目，而且有許多的觀眾群。再加上運動視訊有很固定的內容架構且有其規則。因此，我們在此論文中提出一系列的方法去分析、索引、模組化網球運動視訊。此外，我們還會提出一視訊物體擷取的方法。這方法可以幫助我們追蹤感興趣的視訊物體不管是從靜態背景或動態背景的視訊中。

關鍵字: 視訊分析、移動物體擷取、移動估測

Abstract

Sports video is a major part in most broadcasting TV programs, and has large number of audience. Further, sports video usually has well-defined content structure and domain rule. Thus, we propose a series of methods to analysis, indexing, and modeling tennis game video in this paper. Besides, a video object extraction scheme is proposed in this paper. This scheme can help people to track the interesting video object whether from static background or moving background in the video.

Keywords: Video analysis、moving object extraction、motion estimation

1. Introduction

With the increasing amount of audio-visual information that is broadcasted or available in prerecorded media, people prefer to actively access information they are interested in. Therefore, Powerful tools for video indexing, retrieval and filtering are in great demands. A typical indexing and retrieval procedure of video contents is shown in Figure 1. First, an input video is segmented into spatial consistent units. Visual features are extracted from these segments to build indices and structure. And, people can use these indices and structure to retrieval and access the information they want. Within this procedure, extracting visual features to detect and classify semantic scenes plays an important role. Therefore, in this thesis, we will put emphasis on how to detect semantic scenes in sports videos.

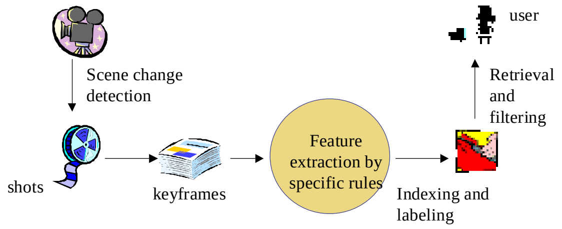


Fig.1 Procedure of video retrieval system

In this paper, we present baseball and tennis semantic scene detection methods. Combining with domain-specific knowledge, we index keyframes by low-level features such as color histogram, color projection and edge detection, etc. We try to get the best performance without motion information so that can approach real-time. Figure 2 expresses the skeleton outline of our method.

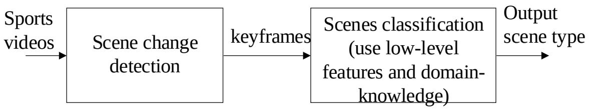

Sports videos
Scene change detection
keyframes
Scenes classification (use low-level features and domain-knowledge)
Output scene type

Fig.2 Skeleton outline of our semantic scenes detection method

Tennis and baseball, which are very popular sports nowadays, have well-defined content structure and domain rules. A tennis game is divided first into sets, then games and serves, as shown in Figure 3. A baseball game is divided into inning, then half inning, batter and pitches. In addition, there are a fixed number of cameras at almost fixed position. Therefore, there are some typical scenes in every sport video. In following section, we will define several semantic scenes we detect in our method. Given the detection results, useful applications such as events detection and structure summaries can be developed.

After labeling every shot with specific types, how to filter and summarize the sports video is also a critical issue. It will be also discussed in this paper, and we design some models to fit different sports.

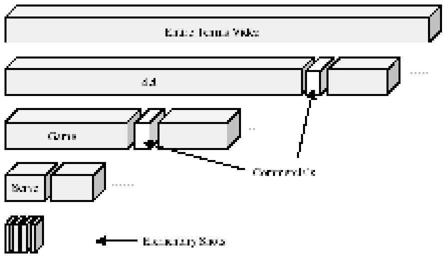


Fig.3 Typical content structure in tennis video [1]

Recently, video content analysis is an important and challenging problem in view of the increasing amount of digital video content available. Existing video content analysis methods may be classified into the following three categories: (1) syntactic structurization of video, (2) video classification, (3) extraction of semantics. The work in the first category can be as basic as detecting abrupt video scene changes and selecting the first frame of a scene as a representative frame, i.e., keyframe [2][3]. However, to deliver meaningful representation to the user, the segments should be further analyzed to extract information that truly represents the content of the video. And some researchers use various methods such as story units or stratification approach in order to model video [4][5]. Story units or stratification approach usually is used for unknown video such as movies. In sports videos, we needn't use story units because sports videos almost have well-defined structures and fixed view types. The work in the second category tries to classify video sequences into certain categories such as news, sports, action movies, etc [6][7]. Video classification is probably needed to help users find what they are looking for and provide some cues in a finer level to analyze video content.

Our method belongs to the third categories. This category is always specific to particular domain. There have been research activities for sports videos analysis [8][9][10]. Many people have incorporated motion information, embedded in MPEG compressed bit-streams or extracted directly from image, for game structure analysis. Since the motion information hardly provides much insight of what is really going on in the games and sometimes require more computational complexity, we temporarily do not adopt the motion information. H. Pan et al. extracted game highlights based on detection of replay or slow-motion [11]. The method did not have the ability to give semantic meanings of the events in replay or slow-motion. Chang and Zhong present an effective framework for scene detection and structure analysis for sports videos [2][12][13][14].

The framework in our paper adopts some concepts from the method proposed by Chang. However, we don't use video object tracking in object level verification in order to reduce the complexity, yet we use object-location to verify which view type the keyframe belongs to. Further, we define other dominant scenes in tennis/baseball videos so that we can find shot transition rules to construct finite states model to fit different sports models, and summarize the important parts of the sports video by the shot transition rules.

The remainder part of this paper is organized as follows. In section 2, we briefly introduce the method to detect scene change. And the algorithm of semantic scenes detection in tennis and baseball will be presented in section 3 and 4. In section 5, structuring and filtering the sports videos will be described here. And several experimental results are shown in section 6. Finally, give a conclusion and future works in section 7.

2. Scene change detection

Because the video is too long, annotation of its content can benefit from segmenting the video into smaller units. Therefore a powerful scene change detection algorithm is required in order to characterize video sequences completely for content-based video indexing and retrieval. The video will be partitioned into shots, each video shot representing a meaningful event or a continuous sequence of action. Once shots have been identified, key frame for each shot must be selected

In our experiment, we use IBM VideoAnnEx Annotation Tool [15], which assists authors in the task of annotating video sequences with MPEG-7 metadata, to execute the action of scene change detection and key frame extraction. The IBM VideoAnnEx Annotation Tool performs the shot detection algorithm, which is based on the multiple timescale differencing of the color histogram. It segments our video content into shorter shots, where scene cuts, dissolves, and fades are effectively detected. Because each video shot and key frame can be described and retrieved independently of each other, the next step is to define several typical scenes in sports videos and bring up some ideas to detect these scenes.

3. Semantic scenes detection in tennis game videos

In this section we will introduce a unique scheme to detect scenes in tennis. First, we should know the structure and content of a typical tennis video program, and must thoroughly trace the common shooting pattern of the cameramen. In temporal domain, a tennis game is divided first into sets, then games and serves (as shown in Figure 3). In addition to standard division, we can find that there are always some typical scenes repeating periodically, such as serving scene, close-up scene, whole-body scene. Except these three dominant scenes, we group the rest into "others". After extracting keyframes from the video, we will quantize these keyframes into 9 colors (quantized in HSV domain). Figure 4 depicts the pre-processing in order to get dominant color and reference histogram. Figure 5 shows the architecture for scenes detection. Following, every block will be explained in detail.

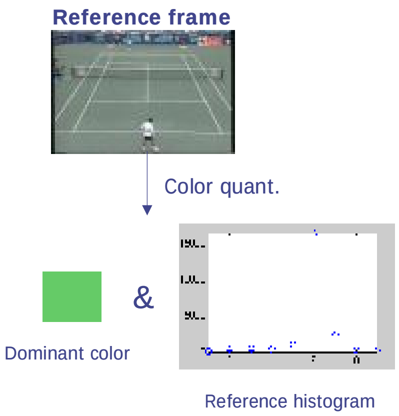


Fig.4 Preprocessing of semantic scenes detection in tennis broadcasting

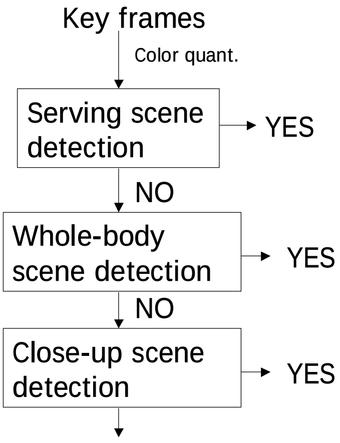


Fig.5 Block diagram of semantic scenes detection in ten-

nis broadcasting

3.1 Serving scene detection

When we get a tennis video, we extract the key-frames in the video. After we get these key frames, the first step of the process is to detect whether the key frame represent the serving scene or not. To detect serving scene in video, we quantize the key frame into 9 colors image and compute it histogram H_j. After we get histogram of input image, we calculate the distance $D_{\mathrm{inter}}(H,K)$ of input histogram H_j and reference histogram K_j by using histogram intersection.

$$d_{\mathrm{inter}}(H,K)=1-\sum_{j=1}^{N}\min(h_j,k_j) \qquad (1)$$

There will be three situations:

$$\begin{cases} D>0.3, & \text{It is not the serving scene} \\ D<0.1, & \text{It is a serving scene} \\ 0.1<D\le 0.3, & \text{Do object - level verification} \end{cases} \qquad (2)$$

If D is among 0.1-0.3, we should further do roughly object-level verification.

Color histogram is a global feature that can be computed and compared faster than real time. However, with color feature only, the detection accuracy is not high enough. Many close-up scenes of playgrounds and replay scenes are likely to be detected as false positives. To improve detection accuracy, the important region extraction can produce localized spatial features.

In tennis serving scene, the player closer to the camera are always in the bottom of the key frame. Figure 6 shows a simple region segmentation example. Because the court in tennis is almost uniform color and there must be a player in the bottom, we can extract a region in the bottom of frame. Figure 6(b) shows a binary image where 1-pixel represent court color (dominant color in Figure 4) and 0-pixel represent non-court color. In our experiment, because of reduction computation time, this process is performed on the down-sampled images so that the court lines are nearly no preserved due to down-sampling of frame size. The down-sampling rate is used in our experiment is 4, both horizontally and vertically, which results in images with size 90*60. And then we use a rectangular block (4*10) to filter non-court color pixel in the bottom of image. Figure 6(c) shows the result after applying block filtering. If there is only a region in the bottom of image and its size is in a proper range (60< *Region_size* < 500 in our experiment), we say it is a correct result.

After serving scene detection, if not a serving scene, the key frame will go to the next block.

3.2 Close-up scene detection

A close-up scene always target on one's face. Therefore, face detection is a key point in this block. First, seeing Figure 7, we will find that the face in close-up scene is almost located in red-block and it occupies a quite percentage of area in red-block, because the cameraman always wants to catch the player's facial expression in close-up scene.

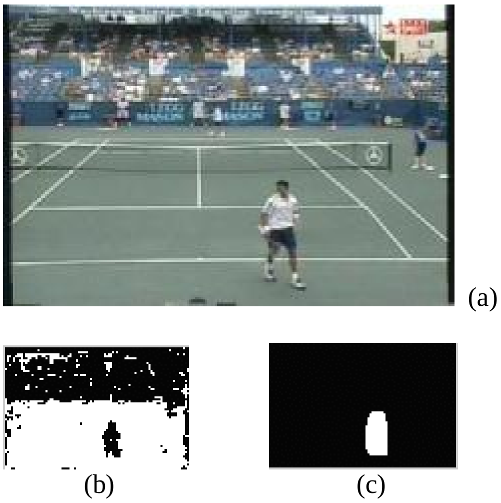

(a)
(b)
(c)

Fig.6 Object level verification in serving scene detection: (a) serving scene; (b) binarize (a) by dominant color and non-dominant color (down-sampled by 4); (c) after block filtering

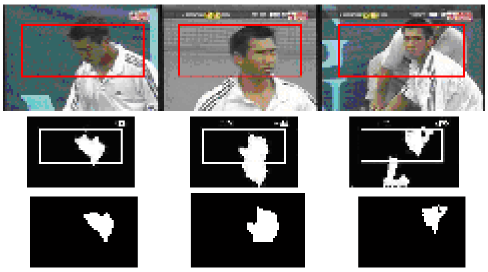


Fig.7 Illustration of face detection: first row: close-up scene; second row: depicts the skin color distribution; third row: depicts the region considered as face region.

According to [16], a model of fleshtones as defined by ranges in hue and saturation. Skin color can be segmented out of an image by looking for hues that are between 0 and 50 degrees and saturation between 0.23 and 0.68. However, we can't use general methods to detect face in close-up scene. Because sometimes the player wear a cap or they don't just right face the camera, we can hardly use region's shape to determine whether it is face or not.

Therefore, we won't use many constraints on detection face in close-up scene in tennis program. We already know the cameraman's filming manner, so we can exploit it and add the model of fleshtones to detect face so as to detect close-up scene effectively.

Briefly, we divide into two steps. The first step is to label skin color to 1-pixel, or label to 0-pixel. The second step is to find the largest region in our defined "red-block" by using region labeling. And the size of the face region candidate is in a proper range, we say it is a correct result. Otherwise, the keyframe will be checked in the next block.

3.3 Whole-body scene detection

Besides above two scenes, there is another constantly appearing scene we call it whole-body scene. It often appears before the serving view to show the player's serving pose. Because player must play tennis in the court, we can find two particular features. One feature is that there are some field distribute in the low part of image. Another one is that the player's feet cause one gap in the court region.

We first use the dominant color to detect court region, and do the dilation to eliminate some spots caused by court lines (Figure 8(a)~(b)). And then we histogram these court candidate pixels by horizontal projection, shown in Figure 8(c):

$$P_H(r) = \#\{c \mid (r,c) \in R\} \quad R: \text{court region} \tag{3}$$

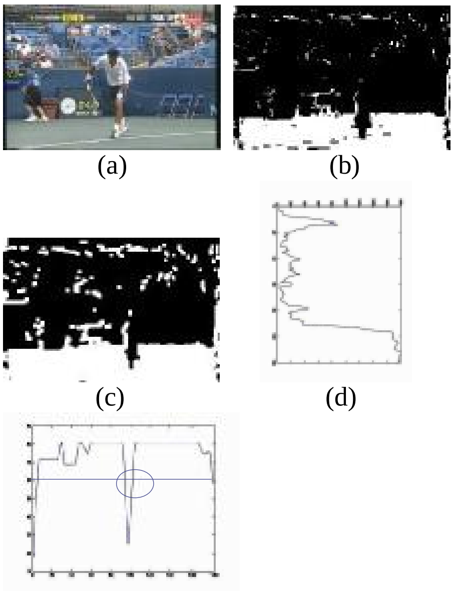
We will bottom-up check this horizontal projection histogram (Figure 8(d)). If the numbers of pixels greater than threshold we say this row is part of court, otherwise the court region just under this row. And if the row number fit the general range, we will check the feet's location.

Second, we will check the feet's location. Using vertical projection, shown in Figure 8(e):

$$P_V(c) = \#\{r \mid (r,c) \in R\} \quad R: \text{court region} \tag{4}$$

We will find a concave in the vertical projection histogram in the whole-body scene, and record the indices whose accumulative pixels don't reach the threshold. Following, calculating the variance of these indices, if the variance is too large, we consider it not the player's feet or it maybe two people in the court.

$$VAR(index\_feet) < T_{\text{var}}$$
$$index\_feet = \{c \mid P_V(c) < 0.75 * \max(P_V(c))\} \tag{5}$$



(a) (b)

(c) (d)

(e)

Fig.8 Whole-body scene detection: (a) whole-body scene; (b) binarized image by dominant color; (c) the result of dilating (b); (d) illustration of bottom-up check; (e) illustration of feet's location check.

Through above processing blocks, if the keyframe doesn't belong to those three types of views, we will classify it into 'others'. It might be the audience view or some scenes that shoot in abnormal angle.

4. Sports videos analysis and structuring

In section 3 and section 4, we proposed an efficient method to detect semantic scenes in tennis and baseball. These semantic scenes are the most representative elements in sports videos. After labeling every keyframe as one type of scenes, we can easily and clearly search the specific type of scenes we want by the indexing. For example, in a tennis video, after scene change detection and semantic scene type detection, every shots will be given a label (e.g. "S C W C S C W S S C C S… ", a series of labels like above. If we want to search close-up scenes, all the shots labeling with "C" will be picked up.). Moreover, in many types of sports broadcasting, not only in tennis and in baseball, although a typical game lasts a few hours, only part of the time is important in general. Therefore, we analyze and exposition a tennis or baseball game by one scene or the combination of several scenes, called it "events". Although the editing patterns and the course of sports videos are similar, yet every researcher analyzed and structured sports videos in different kinds of view. Therefore, we also bring out our opinion on sports videos analysis and structure, and give several state diagrams about different events.

4.1 Analysis and structuring in tennis

In section 3, we already classified all the shots of tennis video into three dominant semantic scenes (serving scene, close-up scene, whole-body scene). If the shots did not belong to these three, they would classify to "other scene". Consequently, most shots belong to first three and just a few shots belong to "others".

As everyone know, a regular play in tennis start with serving action, and then the scene is switched to court view. After the ball flies out of bounds or net ball, the scene will be switched to close-up view or whole-body view until next serving action. Therefore, we will establish a state diagram for tennis video, shown in Figure 12. However, it is possible to categorize all shots into these three scene types due to some reasons such as audience view, other abnormal-shooting view, or some mistakes in detection. We will discuss these details in the final section.

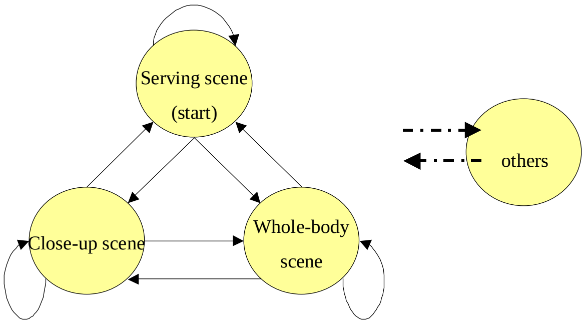


Fig.12 The model of tennis broadcasting videos

Besides constructing a model for tennis video, we can summarize a tennis video in coarse level by our shots indexing. We will introduce two cases to summarize the important parts of a tennis video as following.

Case 1: All court view

This case is very simple; Generally speaking, we pick all serving scenes (serving scenes is the same with the shots of court view) from the tennis video clip. We just reserve the shots that the ball is in play except those less important such as idle time the ball is not in play, time-out, or audience view, etc. Figure 13 roughly shows what would be reserved and what would be discarded.

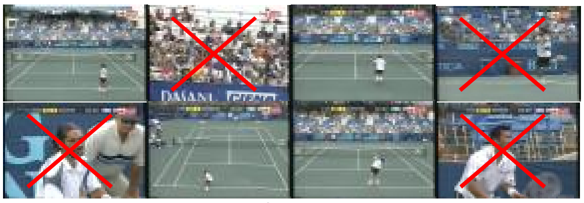


Fig.13 A tennis clip. Each frame represents a shot, and the frame painted with a cross would be discarded in Case 1.

Case 2: Serving action view + court view

In some cases, the serving action is an attractive part in the tennis video. Therefore, in this case, we try to add the serving action before the scene is switched to court view.

It is also uncomplicated just to check whether the scene before serving scene is close-up or whole-body scene or not. If correct, we will include the shot to the summarized video. However, the close-up scene or whole-body scene before the serving scene is not completely the serving action, while it may be the preparing action of the receiver. No matter what kinds of situation, it is still an important view in tennis video, and it will make every start of summarized play smoother.

Reference

[1] D. Zhong and S.F. Chang, "Structure Analysis of Sports Video Using Domain Models," *IEEE conference on Multimedia and Exhibition*, Japan, Aug. 2001.

[2] L. Zhao et al., "Key-frame Extraction and Shot Retrieval Using Nearest Feature Line (NFL)," *ACM Multimedia Workshops 2000,* pp. 217-220.

[3] M. Ahmed et al., "Key Frame Extraction and Indexing for Multimedia Databases," *Vision Interface '99*, Canada, 19-21 May, pp. 506-511.

[4] M.M. Yeung and B.L. Yeo, "Time-constrained Clustering for Segmentation of Video into Story Units," *Proceedings of the 13th International Conference on Pattern Recognition*, 1996, Volume: 3, Aug, pp. 375-380.

[5] T. Chua et al., "Stratification Approach to Modeling Video", *Multimedia Tools and Applications*, 16, 2002, pp. 79-97.

[6] N. Dimitrova et al., "Color SuperHistograms for Video Representation," *International Conference on Image Processing*, 1999, Volume 3, pp. 314-318.

[7] N. Vasconcelos and A. Lippman, "A Bayesian Framework for Semantic Content Characterization," *Proc. Computer Vision and Pattern Recognition*, 1998, pp. 566-571.

[8] W. Hua, M. Han and Y. Gong, "Baseball Scene Classification Using Multimedia Features," *IEEE International Conference on Multimedia and Expo*, Volume 1, 2002, pp. 821-824.

[9] H. Lu and Y.P. Tan, "Sports Video Analysis and Structuring," *IEEE Fourth Workshop on Multimedia Signal Processing*, 2001, pp.45-50.

[10] B. Li and M. I. Sezan, "Event Detection and Summarization in Sports Video," *IEEE Workshop on Content-Based Access of Image and Video Libraries*, 2001, pp. 132-138.

[11] H. Pan, B. Li and M.I. Sezan, "Automatic Detection of Replay Segments in Broadcast Sports Programs by Detection of Logos in Scene Transitions," *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2002, Volume 4, pp. IV-3385-IV-3388.

[12] D. Zhong, R. Kumar and S.F. Chang, "Real-Time Personalized Sports Video Filtering and Summarization," *ACM Multimedia*, 2001, pp. 623-625.

[13] P. Xu et al., "Algorithms and System for High-Level Structure Analysis and Event Detection in Soccer Video," *IEEE Conference on Multimedia and Exhibition*, Japan, Aug. 2001, pp. 928-931.

[14]S.F. Chang, D. Zhong and R. Kumar, "Real-Time Content-Based Adaptive Streaming of Sports Video," *IEEE Workshop on Content-Based Access to Video/Image Library*, Hawaii, Dec. 2001, pp. 139-146.

[15] http://www.research.ibm.com/VideoAnnEx/

[16] K. Sobottka and I. Pitas, "Looking for Faces and Facial Features in Color Images," *Pattern Recognition and Image Analysis: Advances in Mathematical Theory and Applications, Russian Academy of Sciences*, Volume 7, No. 1, 1997.