行政院國家科學委員會專題研究計畫　成果報告

總計畫(3/3)

計畫類別：整合型計畫
計畫編號：NSC93-2213-E-002-101-
執行期間：93 年 08 月 01 日至 94 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：陳良基，李枝宏，馮世邁

報告類型：完整報告

處理方式：本計畫可公開查詢

中　華　民　國 94 年 5 月 12 日

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

計畫編號: NSC-93-2213-E-002-101
執行期限：93 年 8 月 1 日至 94 年 7 月 31 日

主持人：貝蘇章　台灣大學電機系教授

摘要

將對運動視訊分析和視訊物體擷取做一個概括性的介紹，我們將定義出有意義之棒球場景且解釋如何有效率地偵測這些場景，藉由這些已被偵測的場景做出一有限狀態流程且提出一濃縮球賽進行的方法，介紹一個半自動的視訊物體擷取方法，最後做一個總結且提出一些未來的工作方向。

關鍵字: 視訊分析、移動物體擷取、移動估測

Abstract

A general introduction to sports video analysis and video objects extraction. We will define semantic scenes in baseball, and explain how to detect these semantic scenes effectively. And then using these detected scenes to label each shot to model and to filter whole sports video will be presented. A semi-automatic video object extraction method will be described. Finally, conclusion and future work will be.

Keywords: Video analysis、moving object extraction、motion estimation

1. Introduction

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.

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.

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. Semantic scenes detection in baseball game videos

As the above section, this section proposed a method to detect and classify semantic scenes in baseball video programs. The same as tennis video, a baseball game also can be divided into innings in which several batters are at bats. In addition to this characteristic, there are also several fixed cameras located in the stand. Therefore, it also has periodic appearing scenes. In our approach, features specific to baseball broadcast, including field color distribution, edge orientation, player's location, and face detection, used to detect semantic scenes. We can roughly divide video scenes into several scenes, such as pitching scene, field scene and close-up scenes. However, if to analysis finer, we can use edge information to further divide the field scene into several particular scenes.

Figure 9 shows the whole flow chart of proposed method to detect semantic scenes in baseball video. We will describe every block thoroughly.

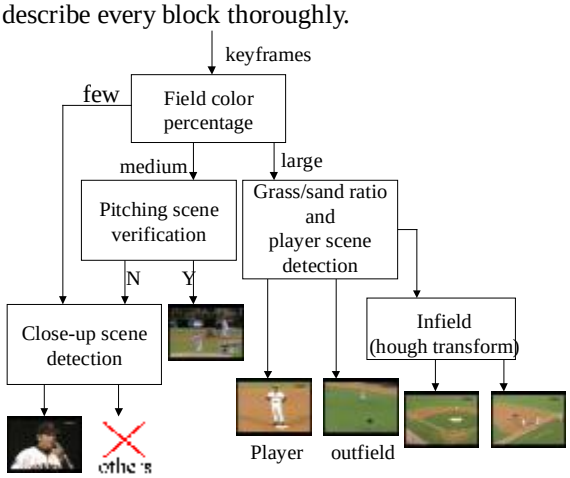

Fig.9 Block diagram of semantic scenes detection in baseball video

2.1 Field color percentage

After shot detection, every key frame first should be detected by field color distribution and percentage. But every game is held in different place and on different day, there must be some chromatic aberration. So we define a loose range about grass color and soil color:

grass color range：0.19<H<0.46、0.2<S<0.7、V>100

soil color range：0.06<H<0.15、0.25<S<0.8、V>100 …(3)

Through checking field color percentage, there will be following three situations: large, medium and small part of picture. And then we process these three situations respectively.

If the proportion of field color is medium (20%~45%), we suppose it may be pitching scene. We next do the pitching scene verification to this pitching scene candidate. This will be reserved for next sub-section explanation.

If the proportion of field color in the image is large (>=45%), it is probable an outfield scene or an infield scene because of the camera zooming-out. And then we will distinguish the candidates into outfield and infield by edge detection or grass-to-soil ratio.

In the last situation, if the image isn't correct image by above verifications or its proportion of filed color is small, we will use face detection to determine whether it is close-up scene or not.

2.2 Pitching scene verification

When we found the image is pitching scene candidate by calculating field color percentage, we should use some features about space distribution to check further more. Therefore, we first build a binary image by assigning field color to 1-pixel and non-field color to 0-pixel (shown in Figure 10(b)). And then we project the binary image by horizontal and vertical direction so as to get two histograms (Figure 10(c)~(d)).

In the horizontal projection histogram P_H, we can get that the field in pitching scene almost distribute in the bottom of the image. On the other hand, in the vertical projection histogram P_V, the pitcher covers the field so that there is a valley in the left side.

$$\frac{sum(P_H(1:Height/2))}{sum(P_H)} < T \qquad (6)$$

$$m < \#\{c \mid P_v(c) < 0.8 * mean(P_v(c)), 1 < c < Width/2\} < M$$

where T, m, M are the thresholds we define.

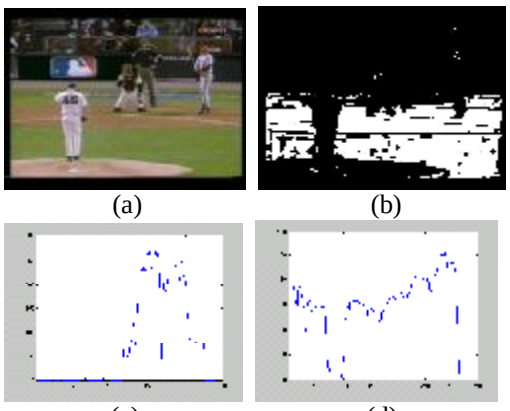

Fig.10 Pitching scene verification: (a) original pitching scene; (b) the result of binarizing (a) with field color; (c) histogram of horizontal projection; (d) histogram of vertical projection

2.3 Close-up scene detection

The close-up scene detection in baseball is the same as in tennis. The method can be referred to in sub-section 3.2.

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

Fig.11 Examples of figure scene

2.4 Player scene detection

This player scene is mainly aimed at those images whose lead role is a figure but background is composed of lots of field components (shown in Figure 11). For this reason, it can be taken as a scene between infield or outfield and close-up scene. In Figure 11(c)(d), we can found that, however, it is can be detected as a close-up scene, but it is difficult to detect the player's face due to the background soil. Thus, we judge this kind of scene to be figure scene so as to make up for the errors due to categorizing to other scenes.

The background of this scene always is field. So we can find that if field color percentage is large and there are some big concaves in the horizontal projection diagram we will say the keyframe represents player scene.

2.5 Infield and outfield scene detection

When the batter hits the ball, the camera will track the trajectory of the ball. Usually, the ball will fall in the outfield, or roll on the infield. Therefore, these two kinds of scenes are quite important. It is obvious that the ratio of grass to soil is key information to make out these two scenes.

First, we can filter out field scene candidates in the previous processing block. And make sure the frame not belong to the player scene or other scenes. If the proportion of field color in the image is large (>=45%), it is probable an outfield scene, or it is an infield scene. Following.

3. Sports videos analysis and structuring

We proposed an efficient method to detect semantic scenes in 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.

3.1 Analysis and structuring in baseball

A baseball game is more complex than a tennis game, because the tennis court is simpler and smaller, and the number of baseball players is much more than tennis. So, the structure of baseball video must be more complicated. We will analysis the structure of baseball as follows.

Fig.14 Two typical starts of play: (a) pitching view; (b) base-stealing

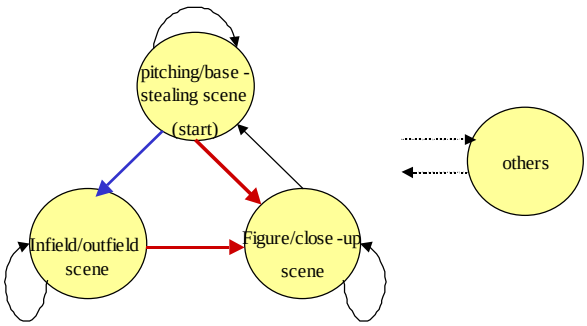

Fig.15 The model of baseball broadcasting videos
Blue arrow means the play continuous
Red arrow means the play is end

A play typically starts with a pitch. A pitching scene is usually captured behind the pitcher. This is because it is much easier to follow the movements of all the parties involved (the pitcher, the batter, the catcher, and the umpire) from this viewpoint than from any other angle. Thus, a play typically starts with a frame like those shown in Figure 14. Figure 14(b) is a special case "base-stealing". How the current play will end depends on the pitching result. For example, if the batter does not swing after the pitcher throwing, then the pitcher will prepare for the second pitch. If the time until the next pitch is too long, there will usually be a scene change and camera may be shooting some less important scene such as the players' rest space, pitcher's close-up, batter's close-up or some other less important scene until to next pitch. If, however, the batter hits the ball, then the scene will be switched to the camera that is shooting at the flying ball (almost always resulting in a frame containing the field). There may be several switches of scenes until shooting on a player. After that, the current play ends, and another start of play (pitching scene) occurs. Therefore, we can sum up two points: (1) a play usually starts with a pitching scene; (2) after the play starts, if after a scene change the camera is shooting the field, then the current play should continue;

otherwise, the current play ends when switched to a player scene. Figure 15 is illustrated the model of baseball video and the model of plays and non-play.

4. Experimental result

4.1 Semantic scenes detection

In this section, we describe the experimental results of our scenes detection system. The test videos are recorded from TV broadcasting by VHS, and they are digitized to MPEG-1 format. The frame size of test videos is 360*240, and frame rate is 30Hz.

In tennis broadcasting, we test several tennis clips (the total time is almost half an hour, except commercials), and it consists of 184 shots. Table 1 shows the result of scenes detection in tennis. We can find that the performance of the serving scene is very good. Comparing to serving scene, the other two types of scene is very good in precision but not good enough in recall. It is because that the serving scene is more stable than close-up scene and whole-body scene, and the global features of serving scene is more identical than the other views.

In baseball broadcasting, we test several baseball video clips, which contain CPBL and MLB. The total shots are 200. The test result is shown in Table 2-2.

	Precision	Recall
Serve	98%	96%
Close-up	96%	78%
Whole-body	98%	77%

Table 1 The detection results of Tennis videos

	Precision	Recall
Pitching	94.2%	94.2%
Close-up/figure	85.7%	60%
Infield	92.2%	90.2%
Outfield	90%	85%

Table 2The detection results of baseball videos

Roughly speaking, the detection results in tennis and baseball videos are satisfying. However, the result in the detection scenes that aren't shot by the specific camera is not good enough. So, more effective method should be proposed in this point.

Most important of all, the philosophy in our proposed method is simple and efficient. So we first use the concept of shots detection to divide a long sports video into lots of division. And we use low-level feature to analyze the keyframes so as to make every division semantic. Without motion information and object tracking, we can fast detect the semantically dominant scenes using low-level features and co-operating with domain models.

4.2 Filtering and summarizing

In tennis broadcasting summarization, we use four shorts clips in this experiment. And we test two cases in summarizing tennis video respectively. The results are respectively shown in Table 3 and Table 4. In Case 1, the process is very simple that only to pick up the serving scene. Therefore, the detection ratio is very high. In Case 2, we want to add the serving action before the serving scene. However, there will be some lost scenes and false alarms. Because, in Chapter 2, we can find that our method to detect close-up and whole-body scene is not good enough. In addition, a few parts of close-up/whole-body scenes before serving view are not the serving action or not the preparing action. Therefore, the result in Case 2 is little worse than Case 1.

	Input duration /output duration	Compressed ratio	Total shots	Total play	Miss shot	False alarm shots	Detection ratio
Clip 1	5'40'' / 1'51''	3.06 : 1	49	15	1	0	93.3%
Clip 2	6'37'' / 2'53''	2.3 : 1	46	18	1	1	94.4%
Clip 3	4'49'' / 1'57''	2.47 : 1	34	12	1	0	91.7%
Clip 4	8'42'' / 3'23''	2.57 : 1	55	18	1	0	94.4%
Total	25'48'' / 10'04''	2.56 : 1	184	63	4	1	93.7%

Table 3 Test result of case1 in summarizing tennis video

	Input duration /output duration	Compressed ratio	Total shot	Total play	Miss shot	False alarm shots	Detection ratio
Clip 1	5'40'' / 2'07''	2.67 : 1	49	15	3	1	88.5%
Clip 2	6'37'' / 3'43''	1.78 : 1	46	18	4	6	84%
Clip 3	4'49'' / 2'36''	1.85 : 1	34	12	1	5	93.8%
Clip 4	8'42'' / 4'42''	1.85 : 1	55	18	2	5	92.3%
Total	25'48'' / 13'08''	1.96 : 1	184	63	10	17	89.7%

Table 4 Test result of case2 in summarizing tennis video

p.s. 1. Compressed ratio = input duration / output duration
2. Detection rate = correct detected shots / sum of shot in play
3. In case 2, a play may consist of a shot or two shots.

We use three baseball video clips, which are captured from ESPN TV programs. The rule of summarizing the baseball video has been mentioned in Section 5.2. We concisely describe the rule again as follows:

1) Search pitching scene or base-stealing scene, and collect it in the summarized video.

2) Check the next scene after pitching scene or base-stealing scene. Pick up the infield scene or outfield scene until detecting the figure/close-up scene.

By observation the result (shown in Table 5), we can understand the compressed ratio is very surprising. And the detection ratio is also quite high. There are some recessive mistakes because the motion of sports video is too fast and scenes cut is too frequently so that some scenes change cannot find out by the IBM's annotation tools.

	Input duration /output duration	Compressed ratio	Total shot	Total play	Miss shot	False alarm shots	Detection ratio
Clip 1	7'14" / 2'08"	3.4 : 1	94	23	4	5	87.9 %
Clip 2	10'52 / 2'37"	4.15 : 1	133	25	4	1	87.5 %
Clip 3	11'14"/ 2'32"	4.43 : 1	129	22	4	4	89.2 %
Total	29'20"/ 7'07"	4.12 : 1	356	70	12	10	88.2 %

Table 5 Test result of summarizing baseball video (p.s. play may consist of more than one shot)

5. Conclusion and future work

Semantic scenes detection and structuring in tennis and baseball are presented in this paper and their experimental results are already shown in previous section.

We first introduce our method of the semantic scenes detection in tennis and baseball. It combines low-level features and domain-specific knowledge. Thus, we reduce lots of computation but get a convincing result. In this method, we define several semantically dominant scenes to describe the whole sports video so that we can obtain an initial description of every shot. We also give a further discuss in infield shot, which is subdivided into more situations.

Most important of all, after indexing every shot, we can realize the transition between these scenes by modeling a state transition model. And we also propose how to summarize the important parts of the sports videos.

However, the features we extraction in system we proposed to detect semantic scenes are low-level features such as the field color distribution, field color percentage, color histogram similarity, color-based object location verification, etc. Although these features are effective owing to combining with domain-specific rules, the accuracy of some scenes is not high enough. Thus, low complexity method to obtain global motion and analysis object tracking must be researched in the future.

Moreover, we should structure more models to detect semantic scenes in different kinds of sports videos, and combine our system with video classification system. We can build an interactive sports scenes retrieval system by then.

Finally, there is a concept should be discussed. In the future work, we should add the viewpoint of commentator. By voice recognition, and co-operate with our semantic scenes detection. The system will be more humanization.

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.