COMBINING CAPTION AND VISUAL FEATURES FOR SEMANTIC EVENT CLASSIFICATION OF BASEBALL VIDEO

Wen-Nung Lie and Sheng-Hsiung Shia

Department of Electrical Engineering, National Chung Cheng University
160, San-Hsing, Ming-Hsiung, Chia-Yi, 621, Taiwan, ROC.
E-mail: wnlie@ee.ccu.edu.tw

ABSTRACT

In baseball game, an event is defined as the portion of video clip between two pitches, and a play is defined as a batter finishing his plate appearance. A play is a concatenation of many events, and a baseball game is formed by a series of plays. In this paper, only the event happened in the last pitch of a plate appearance is detected. It is then semantically classified to represent the corresponding play by using an algorithm integrating caption rule-inference and visual feature analysis. Our proposed system is capable of classifying each baseball play into eleven semantic categories, which are popular and familiar to most of the audiences. In an experiment of 260 testing plays, the classification rate achieves up to 87%.

I. INTRODUCTION

With the rapid increase of many kinds of sports video, how to analyze and manage their contents becomes an important and interesting issue. Semantic classification that identifies meaningful events from a video sequence forms the kernel technology for applications in such as sports video understanding, summarization, and retrieval. For example, the highlights can be identified to form a personalized summary of the video so that users would save a lot of time to see what he/she wants to see.

Recently, many researches focus on content analysis of sports video. For example, Ekin et al. [1] proposed a system to automatically summarize a soccer video by using cinematic and object-based features; Leonardi [7] used camera motion and audio feature to semantically index a soccer video. In [2,3], baseball video structure analysis and semantic event detection were realized by using low-level features. Chang et al. [5] combined several low-level features, such as edges and object heights, to detect events and classify them into four kinds of highlights, such as homerun, nice catches, nice hits, and plays within the diamond. Lie et al. [6] computed a dense field of motion vectors from raw video data to derive camera motion parameters and used a neural network to classify baseball shots into “Non-hitting”, “Infield”, or “Outfield”. Basically, low-level visual features do not provide enough information to derive semantic meanings for human’s full understanding about the game process. The superimposed caption information (e.g., number of strikes and scores) was used in [4] for the understanding of a baseball game. However, they did not combine other features to get more semantic results. Here in this paper, both the caption and visual feature information are combined to semantically classify a baseball event/play into up to 11 categories, much closer to human’s comprehension about a baseball game.

A baseball sports video is featured of a well-defined structure that contains segments of pitches and batters. We define an event as the portion of video clip between two pitches, and a play as a batter finishing his plate appearance. A play is a concatenation of many events, and a baseball video is composed of a series of plays. Figure 1 illustrates the structure of a common baseball game event: pitching shot, critical shot, idle shots, caption change, and replay. For example, if a pitch is hit, the camera would be moved to catch the trajectory of the ball (i.e., critical shot). After that, it may come with several idle shots, e.g., audience overview or close-up of batter and coach, which are beyond observers’ interest. The superimposed captions in the score box will then be updated. If this event is especially interesting or controversial, the replay shots will be provided for more clarity.

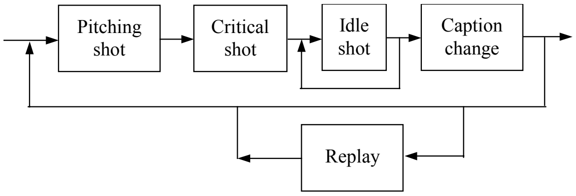

Fig.1 Structure of a baseball game event.

Hence, event detection is actually equal to identifying two contiguous pitching shots. In this paper, only the event happened in the last pitch of a plate appearance is detected and semantically classified to represent the corresponding play. After a pitching shot is identified, the superimposed video caption is detected and recognized to judge whether the current play ends. In case of that, the event is analyzed and classified by the following procedure. First, the recognized caption is inferred to find possible semantic categories of the play. Second, the video visual features are utilized to find out the play type out of “non-hitting”, “infield”, and “outfield”. Third, the above two information is combined to find the exact semantic meaning of the play (including 11 semantics such as strike out, walk, hit by pitch, infield ground out or fly out, infield hit, infield double play, infield hit with wild throw, outfield fly out, outfield hit, sacrifice fly, and homerun). Figure 2 illustrates the processing flow of our algorithm. Our system processes MPEG-4 baseball video. To get abundant information, MPEG-4 video clips are decoded to the YCrCb image domain for analysis, but the motion vector (MV) information is retained for further use.

II. EVENT DETECTION

As shown in Fig.2, procedures of event detection include shot change and pitching shot detections. In baseball video,

0-7803-9332-5/05/$20.00 ©2005 IEEE

there are often two kinds of shot change. One is abrupt shot change, and the other is dissolved shot change. For different shot-change types, we take advantage of different methods for detection.

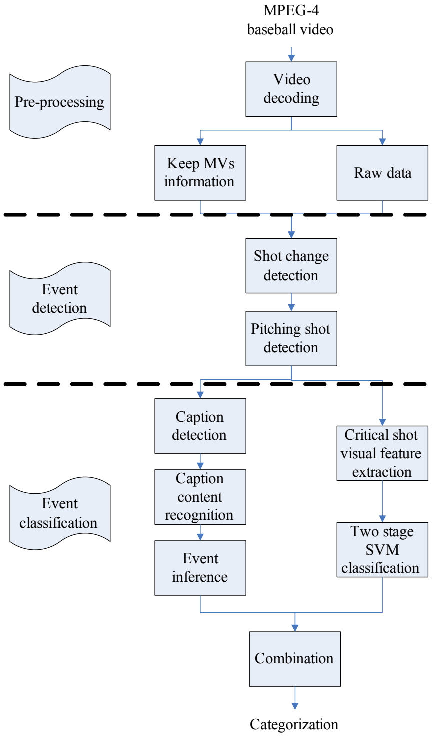

Fig.2 Processing flow of proposed algorithm

A. Shot change detection

Three methods are integrated for abrupt change detection: (A) color histogram difference, (B) linear regression analysis, and (C) field color difference.

For method (A), Haar-wavelets analysis [8] is used to decompose each R, G, or B color histogram to multiple levels, which are then matched between successive frames to indicate the probability of scene change. For method (B), the difference image between successive frames is calculated and a block of interest (BOI) [9] is found out therein. For two pairs of BOIs from 3 successive frames, a 2-D scatter plot is analyzed to obtain a regression line. Behaviors of this line indicate the chance of abrupt shot change. On the other hand, method (C) recognizes the fact of strong color correlation between baseball video shots. Hence, field-color (red for the sand and green for the grass) percentage difference can be used to complement methods (A) and (B). Any frame satisfying one of these three criteria would be recognized to be at an abrupt shot change.

For dissolved (gradual transitions) shot change detection, the method proposed by [10] is used to detect all kinds of dissolved shot change. This algorithm is advantageous of its capability of discriminating global motions caused by camera movement and local motions caused by object movement from a real dissolve.

B. Pitching shot detection

Five visual features, two from colors and three from motions, are extracted from each frame to recognize pitching frame via the SVM (support vector machine) classifier, after a shot change is identified. The selected features [13] are: 1) field color percentage, 2) dominant color around a pre-set pitching area, 3) relation of average MV magnitude between two pre-set areas (the area for the pitcher, catcher, and batter, and its complement), 4) direction of motion activity, and 5) intensity of motion activity. For shots containing above a certain percentage of pitching frames, they will be identified as pitching shots.

C. Identification of the last pitch in a play

To determine whether a pitch is the last one in a play, the information of superimposed caption is used. There will be eight information extracted from the superimposed caption: strike (*S*), ball (*B*), out (*O*), team scores (*Sc1* and *Sc2*), and status of bases (*1B*, *2B*, and *3B*). One criterion is used to judge the last pitch of a batter:

$$A_{current} - A_{previous} = 1, \qquad (1)$$

where $$A = Rn + Sc + O, \qquad (2)$$

Rn is the number of runners on bases, and $Sc=Sc1+Sc2$. If $A_{current}$ equals $A_{previous}$, a play is still in progress (i.e., in ball counting).

III. CAPTION RECOGNITION AND RULE INFERENCE

As stated in the last subsection, caption is important to the recognition of the last pitch in a play. Here, it is also crucial to find out possible semantic categorization for the considered play. In this paper, each baseball play would like to be classified into eleven semantic categories (see Table I).

Here, we would analyze the caption information to find the "possible" categories for an event. Figure 3 illustrates an inference tree based on the above-mentioned caption information. The leaf nodes marked with "Need to be further classified" mean ambiguity there. With the aid of other information (e.g., visual features), this ambiguity can be solved out. Note that we do not consider special situations such as fielding and running errors.

Figure 4 demonstrates an example of baseball caption. It is observed that the position and content layout of a caption are normally fixed. Hence, a pattern template is designed for each TV program for caption detection. If the caption template is matched in this frame, 8 sub-images could be clipped out for further processing and recognition. The captions 1*B*, 2*B*, and 3*B*, are recognized by using their color and size information. The other five are recognized by using a series of binary thresholding, size normalization (to 32x32 pixels), and OCR technique.

Table I. Eleven categories for baseball plays

Play	Hitting	Infield	Infield ground out or fly out
			Infield hit
			Infield double play
			Infield hit with wild throw
		Outfield	Outfield fly out
			Outfield hit
			Sacrifice fly
			Homerun
	Non-hitting	Strike out	
		Walk	
		Hit by pitch	

Image features adopted for caption OCR recognition include "Contour", "Spatial distribution", and "Central projection transformation" [13]. A vector of 256 features is formed for each normalized binary sub-image. The training and recognition algorithm is based on the subspace projection method [11].

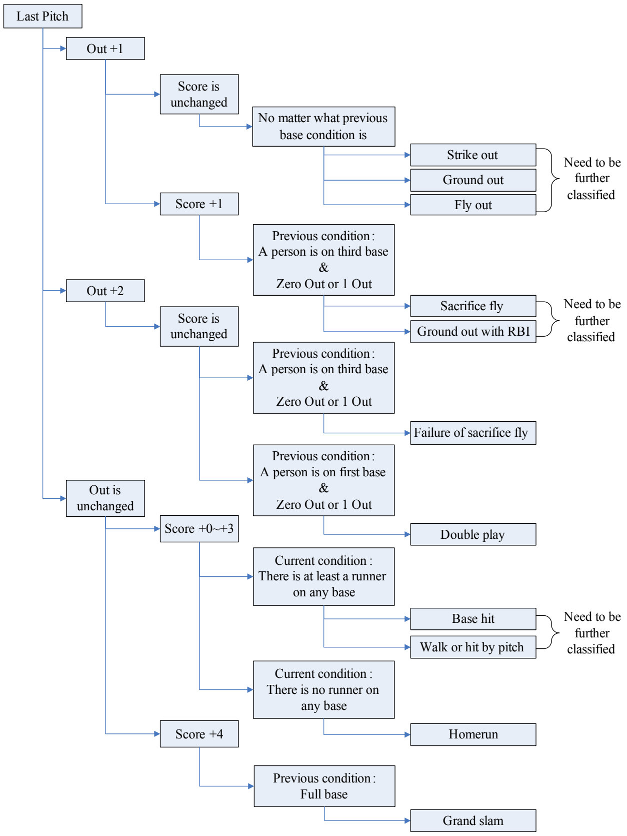
Last Pitch
Out +1
Score is unchanged
No matter what previous base condition is
Strike out
Ground out
Fly out
Need to be further classified
Score +1
Previous condition : A person is on third base & Zero Out or 1 Out
Out +2
Sacrifice fly
Ground out with RBI
Need to be further classified
Score is unchanged
Previous condition : A person is on third base & Zero Out or 1 Out
Failure of sacrifice fly
Previous condition : A person is on first base & Zero Out or 1 Out
Out is unchanged
Score +0~+3
Double play
Current condition : There is at least a runner on any base
Base hit
Walk or hit by pitch
Need to be further classified
Current condition : There is no runner on any base
Score +4
Homerun
Previous condition : Full base
Grand slam

Fig. 3 The inference tree based on 8 caption information.

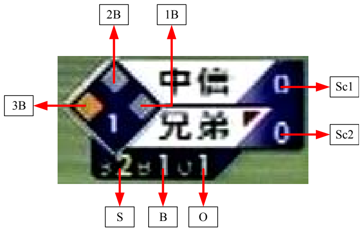

Fig. 4 An example of baseball caption.

IV. VISUAL FEATURE ANALYSIS TO RESOLVE AMBIGUITY

According to Fig.3, it is clear that there are still some ambiguities after rule inferences. By inspection, it was found that all the ambiguities come from three cases: "Non-hitting", "Infield", and "Outfield". Our strategy is to extract visual features from the critical shot that immediately follows the pitching shot for resolving these ambiguities so that an exact play category can be found.

For visual feature analysis, we decode the MPEG-4 baseball video to the YCrCb format to obtain color information, and simultaneously keep MVs to derive motion information. Based on these visual features, a two-stage kernel-SVM classify-cation method is adopted (Fig. 5). First, each frame in the critical shot is classified into two categories: "Hitting" and "Non-hitting". Second, the "Hitting" type is further classified into "Infield" and "Outfield". The critical shot, or the play, is classified into three categories according to the majority of classification for the frames therein. The main reason for this two-stage classification is the difficulty in finding visual features that are capable of distinguishing three categories effectively. It would be easier to find suitable features at these two separate stages.

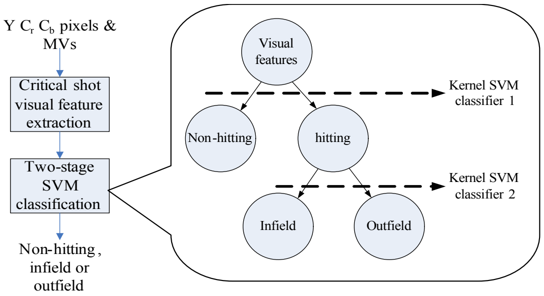
Y C$_r$ C$_b$ pixels & MVs
Critical shot visual feature extraction
Two-stage SVM classification
Non-hitting , infield or outfield
Visual features
Kernel SVM classifier 1
Non-hitting
hitting
Kernel SVM classifier 2
Infield
Outfield

Fig. 5 Two-stage classification based on visual features.

Visual features adopted at the first stage include: 1) field color percentage in a pre-defined region, denoted by CF_1, 2) dominant color percentage in another pre-defined region, denoted by CF_2, 3) range of dominant color projection, denoted by CF_3, and 4) the ZOOM status [12]. Especially, these four features are defined according to some a priori knowledge about "Hitting" and "Non-hitting".

Visual features adopted at the second stage mostly come from some image computations [13], including: 1) the row that clearly separates the audience platform and the outfield grass, and 2) sand to grass ratio (SGR, [3]).

V. EXPERIMENTAL RESULTS

The baseball videos were recorded from TV programs. A total of 305 MPEG-4 video clips were collected, each is of 720x480 pixels frame size, 30 fps, 450~1800 frames, 4~9 shots, and multiple plays.

A. Shot change detection

An amount of 94 out of 305 video clips are randomly chosen. There are 478 abrupt shot changes and 5 dissolved shot changes therein. The result is listed in Table II. About the dissolved shot change, all 5 are detected, but also 5 other false alarms are resulted.

Table II. Precision and recall rates for abrupt shot change detection

	Method A	Method B	Method C	Combination
Detection	386	430	339	478
Missing	92	48	139	0
False alarm	15	3	3	21
Precision	96.25%	99.31%	99.12%	95.79%
Recall	80.75%	89.96%	70.92%	100%

B. Pitching shot detection

An amount of 1334 frames (567 for pitching and 767 for non-pitching) from 15 video clips were used for SVM training.

For testing, 50 out of 305 video clips were randomly selected, resulting in 342 testing shots. Only 4 errors were found, implying a success rate of 338/342 = 98.83%.

On applying the above methods to all 305 clips, except that 45 out of them are used for training, there are totally 22 clips misclassified in shot change detection or pitching shot detection. Hence, the success rate of play detection is 238/260 = 91.54%. Only these 238 clips go into further classification.

C. *Caption recognition*

The caption information is extracted and recognized every 20 frames. Only characters of "0"~"9" need to be recognized. Only 10 video clips are randomly chosen and tested to make statistics. By experiments, the rejection rate is equal to 159/1100 = 14.45% (due to poor image quality) and the character recognition rate is (1100-159-8)/(1100-159) = 99.15%. However, a single error in caption recognition may lead to an error in the semantic tree inference for a clip. Among the 238 video clips, there are 15 clips failing to correctly infer the "possible" event categories by using the caption information. Hence, the raw precision rate is only 223/238=93.7%. To improve the performance, two criteria [13] considering context dependencies are used to correct inference errors. After that, the precision rate is improved to 235/238 = 98.74%.

From Fig.3, it is observed that three kinds of plays, such as {Sacrifice fly, Double plays, Homerun}, can be obtained from caption inference only and need no visual feature classification. According to experiments, 11 out of the 235 video clips belong to these three cases. Hence, only 224 of them need to be further analyzed based on visual information.

D. *Visual feature classification*

Polynomial kernel SVM classifier was used to distinguish "Hitting" and "Non-hitting" plays. A total of 704 sample vectors extracted from 15 "Non-hitting" shots and 1004 sample vectors extracted from 30 "Hitting" shots are used for training.

For the 224 video clips left, there are 59 "Non-hitting" and 165 "Hitting" plays. Only 2 "Non-hitting" and 1 "Hitting" plays are wrongly classified, getting a success rate of 221/224=98.66%.

For the second stage classification (into "Infield" and "Outfield"), a Gaussian radial basis kernel SVM classifier was used. In the 164 "Hitting" clips left, there are 92 "Infield" and 72 "Outfield" plays. With 30 extra video clips (15 "Infield" and 15 "Outfield") used for training, classification of the above-mentioned 164 clips leads to failures of 1 "Infield" and 5 "Outfield" plays, getting a success rate of 158/164 = 96.3%.

In summary, there are totally 9 plays wrongly classified in the two-stage classification process, getting a success rate of 215/224= 95.98%. In other words, visual feature analysis leads to a classification success rate of 95.98%.

E. *Summary for system semantic classification*

Table III shows the success rates for each stage of classification. According to the statistics made above, there are totally 22+3+9=34 failures among 260 video clips (45 out of 405 clips are used for training), getting a semantic classification rate of 226/260=86.92%.

VI. CONCLUSIONS

In this paper, an algorithm integrating caption rule inference and visual feature analysis to achieve high-level semantic play classification of baseball video is proposed. Our proposed system is capable of providing 11 semantic categories that are popular and familiar to audiences in the baseball game. By experiments, our system has a success rate of up to 86.9% and is promising to applications of baseball video retrieval and summarization.

Table III. Performance of each stage of classification.

	success rate
Shot change & pitching shot detection	91.54%
Classification after caption inference	98.74%
Classification based on visual features	95.98%
Overall semantic classification	86.92%

REFERENCES

[1] Ahmet Ekin, A. Murat Tekalp, and Rajiv Mehrotra, "Automatic soccer video analysis and summarization," *IEEE Trans. on Image Process.*, Vol. 12, No. 7, pp. 796-807, 2003

[2] Di Zhong and Shih-Fu Chang,, "Structure analysis of sports video using domain models," *Proc. of IEEE Int'l Conf. on Multimedia and Expo*, pp.713-716, 2001

[3] Soo-Chang Pei and Fan Chen, "Semantic scenes detection and classification in sports videos," *IPPR Conf. on Computer Vision, Graphics and Image Processing*, Taiwan, pp.210-217, 2003

[4] D. Zhang, S.-F. Chang, "Event detection in baseball video using superimposed caption recognition," *Proc. of ACM Int'l Conf. on Multimedia*, pp.315-318, 2002

[5] Peng Chang, Mei Han and Yihong Gong, "Extract highlights from baseball game video with hidden Markov models," *Proc. Of IEEE Int'l Conf. on Image Processing*, pp.22-25, 2002

[6] Wen-Nung Lie, Ting-Chih Lin, and Sheng-Hsiung Hsia, "Motion-based event detection and semantic classification for baseball sport videos," *Proc. of IEEE Int'l Conf. on Multimedia and Expo*, June 2004.

[7] Riccardo Leonardi, Pierangelo Migliorati, and Maria Prandini, "Semantic indexing of soccer audio-visual sequences: a multimodal approach based on controlled markov chains," *IEEE Trans. on Circuits and Syst. for Video Technol.*, Vol. 14, No. 5, pp. 634–643, 2004

[8] Jyrki Korpi-Anttila, "Automatic color enhancement and scene change detection of digital video," *Licentiate thesis, Helsinki University of Technology,* 2002

[9] Seung-Hoon Han and In-So Kweon, "Detecting cuts and dissolves through linear regression analysis," *Electronics Letters*, Vol. 39, pp.1579-1581, 2003

[10] C. W. Su, H. R. Tyan, H. Y. Mark Liao, and L. H. Chen, "A motion-tolerant dissolve detection algorithm," *Proc. of IEEE Int'l Conf. on Multimedia and Expo*, Vol. 2, pp. 26-29, 2002

[11] Dr. E. Oja, "Subspace Method of Pattern Recognition," John Wiley & SONS INC., New York, 1983.

[12] Wen-Nung Lie and Chun-Ming Lai, "News video summarization based on spatial and motion feature analysis," *Proc. Of Pacific-Rim Conference on Multimedia*, Tokyo, Japan, Vol. II, pp.246-255, 2004

[13] Sheng-Hsiung Shia, *Combining caption and visual features for semantic event detection and classification of baseball video*, Master Thesis, National Chung Cheng University, Chia-Yi, Taiwan, ROC., July 2004.