Event Detection in Baseball Videos Using Genetic Algorithm Optimization

Po-Chih Lin, Sheng Yang Li, and Chih-Yi Chiu,
Department of Computer Science and Information Engineering, National Chiayi University
sancitu@gmail.com, shengyang_li@hotmail.com, cychiu@mail.ncyu.ecu.tw

***Abstract*—We proposed a novel approach for event detection in baseball videos. Different from the previous approaches that employ image processing and machine learning techniques, we integrate the webcast text information to assist in video content analysis. In particular, the proposed approach applies a genetic algorithm to detect video events by associating the video content with the webcast text through unsupervised learning methods. Experimental results show that the proposed approach is robust against various styles of baseball videos.**

I. Introduction

Sport video analysis is a very important issue in multimedia research fields because of the increasing growth of video content. Related applications have entered our daily life. For example, with a TV-BOX and its functions, people can capture, browse, and annotate interesting segments for future purposes. However, this work is usually done with a great amount of human labor so far. In this study, we propose an automatic method that does not use much manpower and time to facilitate sports video analysis for baseball games.

Baseball games are prevalent in sports video content analysis. Current techniques are mainly base on image processing and machine learning. On the other hand, webcast text accompanied with baseball games conveys different materials from video content. The webcast text records the detailed game information. Therefore, we proposed a novel approach for video event detection that is accomplished by mapping video content and webcast text. A genetic algorithm-based optimization method is proposed to automatically construct the relation between video content and webcast text. Experimental results show that the proposed approach is robust against various styles of baseball videos.

The remainder of this paper is organized as follows. Section II reviews the recent work. An overview of the proposed approach is given in Section III. In Section IV and V, we detail our approach in two aspects, including content analysis and genetic algorithm optimization. Section VI demonstrates and discusses some experiment results. Conclusions are summarized in Section VII.

II. Related Work

Sports video content analysis and event detection for baseball videos is very active in multimedia content understanding. Related topics include shot classification [5], highlight extraction [1], and event detection [3]. Although comprehensive audiovisual features and complicate learning models were investigated in the above studies, they seldom considered the utilization of webcast text in analyzing baseball video content.

On the other hand, integrating webcast text with video content has attracted increasing attention in the analysis of soccer, American football, and basketball games. Xu and Chua [7] identified event boundaries in video by matching the timestamps extracted from video content and webcast text. Xu *et al*. [6] conducted a conditional random field model for the matching. Although these studies demonstrate the good effect of multimodal fusion of video content and webcast text, their frameworks are infeasible for baseball videos due to their different natures.

III. The Proposed Approach

The structure of a baseball video is composed of a number of half-innings. For each half-inning, it can be partitioned into several video segments of two types, namely, the *pitch segment* and the *event segment*. The pitch segment displays the pitcher act of throwing a baseball toward the home plate to start play, and its subsequent event segment shows the batter performance as a baseball event for the pitch. Note that a batter usually occupies more than one pitch/event segments. The last event segment of the batter is called the *at-bat event*, which is the main body recorded in webcast text.

The basic principle of the proposed approach is to assess batter similarities in adjacent pitch segments. If the similarity is low enough, it can be regarded the two adjacent pitch segments belong to two different batters; the latter pitch segment is denoted as a batter change point. Consequently, the video content and webcast text can be associated as the event annotation.

The proposed approach consists of two main parts, including content analysis and genetic algorithm optimization. Suppose that the video content $\mathbb{V}_\Phi$ and webcast text $\mathbb{W}_\Phi$ of the half-inning Φ are given. We first analyze $\mathbb{V}_\Phi$ to extract batters. Each batter has his personal appearance and battering posture, forming a unique characteristic to distinguish him from other batters. In this study, HoG proposed by Dalal and Triggs [4] is used to detect batter figures in video frames, and the SIFT-Bag descriptor proposed by Zhou *et al*. [8] is employed to characterize the detected batter image.

In the genetic algorithm optimization part, we fuse the clues extracted from video content $\mathbb{V}_\Phi$ and webcast text $\mathbb{W}_\Phi$ to infer the best solution of the mapping between $\mathbb{V}_\Phi$ and $\mathbb{W}_\Phi$.

This research was supported in part by the National Science Council of Taiwan under Grant: NSC99-2221-E-415-011.

We treat the mapping as an optimization problem and seek for the best solution by the genetic algorithm. An objective function is formulated to consider the similarity measures with respect to the batter image similarity, video length estimation, change of left/right handed batters, and event likelihood. The genetic algorithm searches for the solution that best fits the objective function through a mimicked evolution process. Finally, the mapping is expressed as

$$\{f_{\Phi}(PS_i, ES_i) = WT_m\}$$

where i = 1, 2, ... , N and m = 1, 2, ... , M. The webcast text item WT_m for the mth batter can be tagged to the ith video sequence (PS_i, ES_i). For simplicity, we write $PS_i \rightarrow WT_m$ for the mapping hereafter.

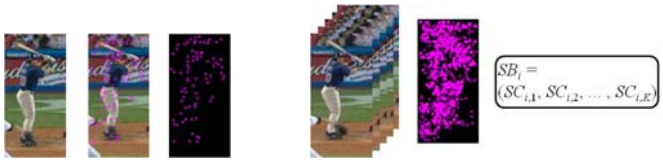
IV. Content Analysis

Given a half-inning of a baseball video, we first apply an unsupervised method based on Markov random walk to detect pitch segments [2]. Then, based on the detected pitch segments, we extract the batter information from them. The objective is to represent each batter with a discriminative feature, so that we can categorize pitch segments according to their respective batters. The extraction includes two steps, namely, batter detection and feature extraction. A human detection method based on histogram of oriented gradients (HOG) [4] is used as a batter detector. We use the batter detector to locate batters in pitch segments $PS_i \in \Phi$, i = 1, 2, ... , N, where N is the number of the detected pitch segments in Φ.

Denote the detected batter image sets in PS_i as BI_i. For each of the batter image sets, a SIFT-based feature is employed to characterize. Since the SIFT-based feature is a sparse representation of image keypoints, it would be more discriminative for low-resolution batter images than the HoG descriptor. In this study, we employ a modified version of the SIFT-Bag descriptor [8] to represent the batter feature as follows. We employ k-means clustering to partition the SIFT feature space at the global clustering step. Let SK_i be a set of SIFT feature vectors extracted from batter image set BI_i. The training corpus is all SIFT feature vectors in the half-inning Φ, i.e., $\{SK_i \mid i = 1, 2, \ldots, N\}$. k-means clustering is applied on the training corpus to obtain K global cluster centers, denoted as $\{GC_k \mid k = 1, 2, \ldots, K\}$. At the specialized clustering step, we adapt the batter image set based on the global cluster centers. Each SIFT keypoint of SK_i is assigned to its nearest global cluster centers. The specialized cluster center $SC_{i,k}$ is obtained:

$$SC_{i,k} = cent(\{SK_{i,k}\}),$$

where $SK_{i,k}$ is the subset of SK_i whose elements are assigned to GC_k, and $cent(\cdot)$ returns the centroid of the set, i.e., the average of the set. Consequently, the ith batter image set BI_i is characterized by the SIFT-Bag descriptor with K specialized cluster centers, denoted as $SB_i = (SC_{i,1}, SC_{i,2}, \ldots, SC_{i,K})$. Figure 1 shows a SIFT-Bag example of a batter image set.



(a) **(b)**

Figure 1. A SIFT-Bag example: **(a)** SIFT keypoints of a batter image; **(b)** the SIFT-Bag descriptor SB_i of the ith batter image set BI_i.

V. Genetic Algorithm Optimization

We propose to use the genetic algorithm (GA), a probabilistic global search technique, to find the solution of matching $\mathbb{V}_{\Phi}$ and $\mathbb{W}_{\Phi}$. In GA, the solution domain is encoded as chromosomes, and their fitness scores are evaluated by an objective function. By mimicking the nature evolution of selection, mutation, and crossover, GA assumes a population of chromosomes will evolve toward the best fitness, i.e., the best solution.

We define the representation scheme of chromosomes in this study. Let $A \in \{1, \ldots, M\}$ be a alphabet set, where M is the number of batters in the half-inning Φ (recorded in webcast text). Each chromosome $x \in A^N$ is an alphabet string of length N, where N is the number of pitch segments in Φ (detected from video content). x_i, the ith alphabet of x, represents the batter index of the ith pitch segment PS_i. x is a sorted string, i.e., $x_i \le x_j$ for $i < j$. We take an example to illustrate the representation scheme. Suppose there are three batters and nine pitch segments in Φ; we have A = {1, 2, 3} and N = 9. A chromosome "111223333" represents the candidate solution of the mapping relation: $\{PS_1, PS_2, PS_3\} \rightarrow WT_1$, $\{PS_4, PS_5\} \rightarrow WT_2$, and $\{PS_6, PS_7, PS_8, PS_9\} \rightarrow WT_3$.

To initialize GA, we randomly select a set of chromosomes as the first population. The evolution operations are applied on the population and suitable chromosomes are selected to produce the next population. The selection is based on an objective function $f(x)$ that evaluates the fitness of chromosome x. The roulette-wheel scheme is employed to select chromosomes the population into a mating pool with probabilities proportional to their fitness, and the one-point crossover operation is used to generate offspring chromosomes from the mating pool. The mutation operation randomly changes a chromosome's alphabet in the mating pool with a predefined probability. In addition, we apply the elitism mechanism that copies the best-so-far chromosomes into the new population.

A critical issue of GA is the objective function $f(x)$. In this study, we define $f(x)$ as a blending function that fuses multimodal information:

$$f(x) = \alpha \cdot BS_x + \beta \cdot VL_x + \gamma \cdot LR_x + \rho \cdot EL_x.$$

BS_x, VL_x, LR_x, and EL_x are the measures of chromosome x with respect to batter image similarity, video length estimation, change of left/right handed batters, and event likelihood, respectively; α, β, γ, and ρ are the corresponding coefficients in the interval [0, 1], and $\alpha + \beta + \gamma + \rho = 1$. These measures are detailed in the following subsections.

A. Batter Image Similarity

This measure evaluates the intra-similarity of the batter image sets that belong to the same cluster. The cluster is labeled by the chromosome rather than generated by HAC. Let $\varphi_{x,m} = \{i \mid x_i = m\}$ be the index cluster of alphabet m and $\varphi_x = \{\varphi_{x,m} \mid m = 1, 2, \dots, M\}$ for the chromosome x. We define the batter image similarity for the chromosome as

$$BS_x = 1 - \frac{\sum_{m=1}^{M} \sum_{i \in \varphi_{x,m}} \left\| SB_i - SB_{\varphi_{x,m}} \right\|_1}{\sum_{m=1}^{M} \sum_{i \in \varphi_{x,m}} \left\| SB_i - SB_{\varphi_x} \right\|_1}.$$

In the above equation, the numerator is the total intra-distance between each batter image set and its corresponding cluster, and the denominator is a normalized term that makes the fraction between [0, 1]. Note that we subtract the fraction from one to recast the distance metric as a similarity metric.

B. Video Length Estimation

This measure is based on an intuitive assumption that the pitch count of a batter is proportional to the video length occupied by the batter. The pitch count is the number of pitches thrown by the pitcher; related information is recorded in webcast text. Let τ be total pitch count in Φ and τ_m be the pitch count for the mth batter; they are recorded in webcast text. We calculate the pitch segment length of the mth batter according to chromosome x's label

$$T_{\varphi_{x,m}} = \sum_{i \in \varphi_{x,m}} duration(PS_i),$$

where $duration(PS_i)$ returns the time duration (in seconds) of pitch segment PS_i. The video length similarity is defined by

$$VL_x = 1 - \frac{\sum_{m=1}^{M} \left\| \frac{\tau_m}{\tau} T - T_{\varphi_{x,m}} \right\|_1}{2 \cdot T},$$

where $T = \sum_{i=1}^{N} duration(PS_i)$, and $\frac{\tau_m}{\tau}$ is the pitch count ratio of the mth batter. We use the ratio to estimate the pitch segment length of the mth batter. The numerator of the above equation sums up all batters' differences between two pitch segment lengths: one is estimated from chromosome x, and the other is from the pitch count ratio; the denominator is a normalized term.

C. Change of Left/Right Batters

A baseball player is an either left or right handed batter. The batting order of a team lineup is usually interlaced with left and right handed batters. We leverage the change of left/right handed batters in a half-inning to investigate the interrelationship among video content, webcast text, and chromosomes. First, we define two sets of the change points U and V for the half-inning Φ. U represents the change points specified by the chromosome and webcast text expressed as

$$U = \{u | u = max(\varphi_{x,m}) \; if \; bat_m \neq bat_{m+1}$$

where $m = 1, 2, \dots, M-1$, bat_m represents the mth batter is left or right handed recorded in webcast text. V is obtained according to the classification result of batter images

$$V = \{v | v = i \; if \; classifyBat(BI_i) \neq classifyBat(BI_{i+1})\}$$

$i = 1, 2, \dots, N-1$, where $classifyBat(BI_i)$ is a classifier that returns left-handed "L" or right-handed "R" for the i-th batter image set BI_i. The classifier is trained through a SVM. We then measure the similarity between U and V by the *Jaccard coefficient*

$$LR_x = \frac{|U \cap V|}{|U \cup V|}$$

where $|\cdot|$ returns the cardinality of the set.

D. Event Likelihood

The last measure takes baseball events into consideration. In the half-inning Φ, each batter's performance that is shown in video content and recorded in webcast text can be classified into one of several predefined baseball events, such as homerun, strikeout, walk, etc. The relation between video content and webcast text is thus connected through their associated baseball events. Let ES_i be the following video segment of the ith pitch segment PS_i. The measure EL_x is defined as

$$EL_x = \frac{\sum_{m=1}^{M} classifyEvent_{WT_m}(ES_{max(\varphi_{x,m})})}{M}$$

$ES_{max(\varphi_{x,m})}$ is the last video segment of the mth batter that shows the batter's final performance. Function $classifyEvent_{WT_m}(ES_i)$ returns the likelihood of the event recorded in WT_m for the given video segment ES_i. In this study, baseball events are categorized to three types: non-hitting, infield, and outfield. Three event-category classifiers are modeled by the SIFT-Bag Kernel [8]. For an event tagged to a batter in the webcast text, his at-bat event segment, which is represented as a SIFG-Bag descriptor, is fed to the event's corresponding classifier. The output likelihood is in the interval [0, 1].

VI. Experimental Results

To evaluate the proposed approach, we compiled a video dataset for use in several experiments. Dozens of baseball games played in the MLB 2008 regular season were recorded from TV broadcasting as our baseball video dataset; they vary in teams, stadiums, and broadcasting channels. The video dataset were transformed to 720×480 frame pixels and 15 frames per second. Meanwhile, the corresponding webcast text for each recorded games were download from the MLB official site (http://www.mlb.com).

To train a HOG-based batter detector, we extracted 192 left and right handed batter images, each of which is 64×128 pixels, from our video dataset; they were used for the positive examples. The negative examples are generated from the INRIA dataset, with total 1179 64×128 image blocks. In addition, 165 hard examples were added to the negative examples in the second round training. We then trained two other classifiers to classify left and right handed batters by using 6 cluster centers.

The three event categories of baseball events are listed in Table I. To train the classifiers of the three event categories, the training data were compiled by collecting event segments from our video dataset. Each event segment was represented as a SIFT-Bag descriptor with 256 cluster centers. We trained

a SIFT-Bag kernel through a SVM for each event-category classifier.

We selected six half-innings of three games from our video dataset for this experiment. The configuration of the genetic algorithm was set by the chromosome population size 32, crossover probability 0.75, mutation probability 0.0075, and maximal evolution iterations 20. We further list five combinations of measures in the genetic algorithm: batter image similarity only (BS), video length estimation only (VL), change of left/right handed batters only (LR), event likelihood only (EL), and all measures combined (ALL). The quadruple parameters (α, β, γ, ρ) for GAO's combinations are BS: (1, 0, 0, 0), VL: (0, 1, 0, 0), LR: (0, 0, 1, 0), VL: (0, 0, 0, 1), and ALL: (0.05, 0.35, 0.3, 0.3).

Given the video content and webcast text of a half-inning as input, the proposed methods will yield an alphabet string as output, i.e., the mapping $x^* = \{x_i^* | i \in \{1,2,\ldots,N\}, x_i^* \in 1,2,\ldots,M$, where x_i^* is the batter index of the ith pitch segment, N is the number of pitch segments, and M is the number of batters in the half-inning. Let $\hat{x}$ be the ground truth of the half-inning. The similarity between $\hat{x}$ and x_i^* is defined based on the Hamming distance

$$HS(\hat{x}, x^*) = 1 - \frac{\sum_{i=1}^{N} (\hat{x}_i \oplus x_i^*)}{N}$$

where $\oplus$ is the XOR logical operator that returns zero if the two operands are the same; else returns one.

The experiment results are listed in Table II. The three games are played by different teams and broadcasted by different channels. In the second column, G, N, and M represents the number of ground-truth pitch segments, the number of detected pitch segments, and the number of individual batters, respectively. Overall speaking, the VL and LR measures outperform the BS and EL measures. We consider that the BS measure is formulated based on SIFT-Bag descriptors; it would be sensitive to abrupt visual changes. For the EL measure, we find that there are other video segment types such as close-ups and audience views. These types do not belong to any of the three event categories listed in Table I. Their likelihoods of the three event classifiers might be incorrect, making the EL measure abnormal in some cases. We summarize that each GAO's measure has its unique power to assess an interesting cue derived from video content and webcast text. Their combination would complement each other to deal with a variety of video content styles in baseball games. Our viewpoint is obviously supported by the experiment result of the ALL measure, which demonstrates the highest accuracy rate and the most robust performance in general.

VII. Conclusion

Conventional baseball annotation approaches do not utilize webcast text well, whereas existing multimodal fusion frameworks do not fit baseball video properly. To address the above issues, we present a novel approach for baseball video event detection by mapping video content and webcast text. The mapping problem is conquered by the proposed genetic algorithm optimization, which integrates interesting properties extracted from both low-level video content and high-level webcast text. Experiments on various styles of baseball video content demonstrate a robust result, manifesting that the proposed approach is widely adaptable and mostly automated in baseball event detection.

Table I. The categorization of baseball events

Category	Non-hitting	Infield	Outfield
Event	*Strike out, walk, hit by pitch*	*Infield ground out, infield fly out, infield hit, double play*	*Outfield fly out, outfield hit, homerun*

Table II. The accuracy of the proposed genetic algorithm optimization.

Game	**Inning (*G* / *N* / *M*)**	**BS**	**VL**	**LR**	**EL**	**ALL**
2008/06/24 (MY9) NYY vs. PIT	2 top (13/13/4)	0.85	0.85	1.00	0.85	1.00
	5 top (10/10/3)	0.70	0.80	0.90	0.90	0.90
2008/07/05 (FOX) BOS vs. NYY	4 bot. (16/16/4)	0.60	1.00	0.88	0.88	1.00
	5 top (12/12/4)	0.33	0.67	0.75	0.33	0.83
2008/08/25 (FSN) LAD vs. PHI	4 top (16/16/5)	0.31	0.75	0.69	0.75	0.88
	4 bot. (8/8/3)	0.38	1.00	0.75	0.75	1.00

References

[1] C. C. Cheng and C. T. Hsu, "Fusion of audio and motion information on HMM-based highlight extraction for baseball games," *IEEE Transactions on Multimedia*, Vol. 8, No. 3, pp. 585-599, 2006.

[2] C. Y. Chiu, P. C. Lin, W. M. Chang, H. M. Wang, and S. N. Yang, "Detecting pitching frames in baseball game video using Markov random walk," In *Proceedings of IEEE International Conference on Image Processing (ICIP)*, pp. 1493-1496, Hong Kong, China, Sep. 26-29, 2010.

[3] W. T. Chu and J. L. Wu, "Explicit semantic events detection and development of realistic applications for broadcasting baseball videos," *Multimedia Tools and Applications*, Vol. 38, No. 1, pp. 27-50, 2008.

[4] N. Dalal, and B. Triggs, "Histograms of oriented gradients for human detection," In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 886-893, San Diego, USA, Jun. 20-26, 2005.

[5] Y. H. Huang and L. H. Tung, "Semantic scene detection system for baseball videos based on the MPEG-7 specification," In *Proceedings of ACM Symposium on Applied Computing (SAC)*, pp. 941-947, Sierre, Switzerland, Mar. 22-26, 2010.

[6] C. Xu, Y. F. Zhang, G. Zhu, Y. Rui, H. Lu, and Q. Huang, "Using webcast text for semantic event detection in broadcast sports video," *IEEE Transactions on Multimedia*, Vol. 10, No. 7, pp. 1342-1355, 2008.

[7] H. Xu and T. S. Chua, "Fusion of AV features and external information sources for event detection in team sports video," *ACM Transactions on Multimedia Computing, Communications and Applications*, Vol.2, No. 1, pp. 44-67, 2006.

[8] X. Zhou, X. Zhuang, S. Yan, S. Chang, M. Hasegawa-Johnson, and T. S. Huang, "SIFT-Bag kernel for video event analysis," In *Proceedings of ACM International Conference on Multimedia (ACM-MM)*, pp. 229-238, Vancouver, Canada, Oct. 26-31, 2008.