EXPLICIT BASEBALL EVENT DETECTION BY COMBINING VISUAL AND SPEECH INFORMATION

[1]Wei-Ta Chu(朱威達) and [1,2]Ja-Ling Wu (吳家麟)

[1]Department of Computer Science and Information Engineering
[2]Graduate Institute of Networking and Multimedia
National Taiwan University, Taiwan
{wtchu,wjl}@cmlab.csie.ntu.edu.tw

ABSTRACT

To explicitly detect baseball events, a multimodal approach that combines the decisions of visual and speech event detection is proposed. We respectively perform event detection from visual and speech perspectives and estimate the corresponding confidences. By combining the decisions from different modalities, the detection performance increases by 8% ~ 20%, in terms of F1 metric. This work achieves explicit event detection in baseball videos and facilitates the development of realistic applications for management or entertainment.

1. INTRODUCTION

Sports event detection has attracted much attention because of fixed game structure and high potential commercial benefits. Many studies have been conducted on soccer, baseball, tennis games based on visual or aural information. For event detection, the most recent researches mainly focus on exploiting motion data for play-break analysis [1], extracting caption data and shot types for event inference [2, 3], or employing audio energy for highlight extraction [4]. Most works were conducted on single modality, while how to combine information from different modalities in order to make analytical results more explicit and useful is still not well studied.

In this paper, we propose an information fusion scheme that integrates visual and speech information to perform explicit baseball event detection. In our previous work [3], we employed caption data and shot transition information to infer what happened in baseball games. Based on official baseball rules and event models, the proposed system detects thirteen common events in baseball games. Although this system works well in most situations, its performance in discriminating confused events is still not good enough for baseball fans. The so-called confused events [3] mean the event pairs like ‘single’ and ‘walk’ that cannot be accurately discriminated by simply checking visual information (caption data).

Commentator’s speech, which completely states the game progress, plays an important role for audiences to realize the game status. Therefore, it’s attractive to exploit a speech recognition module and facilitate event detection through speech information. We apply a key-phrase spotting module that maps speech signal with limited number of key-phrases, which provide some clues to the occurrence of some effective events or actions, such as hit, out, and catch.

The main contribution of this paper is that we propose a fusion scheme, which combines the information derived from visual and speech modalities. We evaluate the performance with and without multimodal fusion and demonstrate the effectiveness of the proposed methods.

The rest of this paper is organized as follows. Section 2 describes the system overview. Event detection via visual and speech information is described in Section 3. Information fusion, including confidence evaluation and combination strategies, is addressed in Section 4. Section 5 provides the experimental results and Section 6 concludes this paper.

2. SYSTEM OVERVIEW

Figure 1 shows the system diagram, which consists of event detection and confidence calculation from visual and speech perspectives and the integrated decision module. From visual data, two components, including rule-based and model-based detection [3], identify what events occurring and where their boundaries are. Based on these event boundaries, a key-phrase spotting module is applied to spot what key-phrases the commentator has spoken, which may provide clues for identifying what really happened in specific intervals. The events detected from visual and speech data are described as *visual events* and *speech events* for convenience. The confidences of visual and speech events are estimated respectively to be the bases of integrated decision. Based on the strategy of combining classifier decisions [7], we find the consensus from two modalities and make an integrated decision.

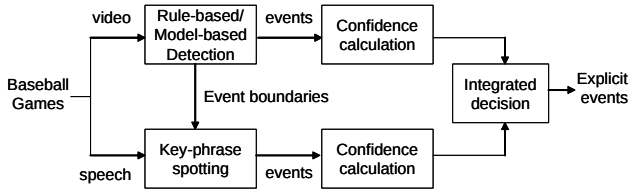

Figure 1. System block diagram.

3. EVENT DETECTION

This section briefly describes event detection from visual and speech data. Basically, the module for visual event detection is the principal part of this system. It identifies most events and their boundaries by exploiting superimposed caption and shot transition information. For the events that cannot be detected explicitly, such as single vs. walk and strikeout vs. infield/outfield out, speech event detection module is further applied.

3.1. Event Detection from Visual Information

In a broadcasting baseball video, it's easy to see that almost all events occur between two consecutive pitches. When the game proceeds, the information on the caption reflects game status to help viewers realizing game progress. These characteristics facilitate us to detect what event occurs by checking the caption information changes between two consecutive pitches.

Based on the changes of outs, scores, and base-occupation situation, thirteen events can be inferred, including single, double, triple, home run, stolen base, caught stealing, field out, strikeout, walk, sacrifice bunt, sacrifice fly, double play, and triple play. For example, as shown in Figure 2, no out and score for team 2, and one base is occupied in the *i*th pitch shot. In the (*i+1*)th pitch shot, the score increases by two, no out occurs, and no base is occupied. Because no out increases, we know the event occurred in this duration must be a hit. According to the base-occupation situation, we can infer that it should be a homerun rather than single, double, or triple.

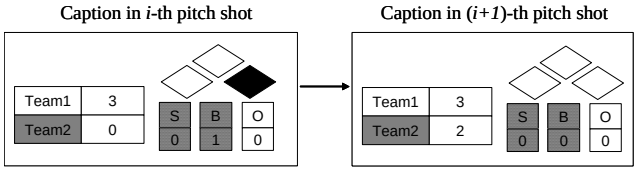

Figure 2. An example of rule-based detection.

By rule-based detection method, which comprehensively exploits official baseball rules, we can detect most events in baseball games. However, some event pairs, such as single vs. walk and strikeout vs. field out, cannot be explicitly discriminated because they cause the same information changes on the caption. It's critical to explicitly discriminate these events because: 1) outs frequently occur in games (around two third of events are outs if the team batting average is lower than 0.33); 2) strikeout and single respectively are important performance metrics for pitchers and batters. They have drastically different meanings to field out and walk.

To deal with this flaw, we included a model-based detection method that builds classifiers on the basis of shot transition information. With the aids of various broadcasting characteristics, such as more pitches in walk and strikeout cases and pitch-field pattern occurs in single and field out cases, the built classifiers tend to discriminate them to facilitate explicit event detection [3].

With the variety of broadcasting styles and different pitching strategies in different teams, the proposed model-based detection approach doesn't have promising results in all games. Therefore, we look for additional helps from speech information. As illustrated in Figure 3, the events with bold type have been explicitly determined by the rule-based method, while the other *confused* events should be further examined by the model-based detection method and speech event detection module.

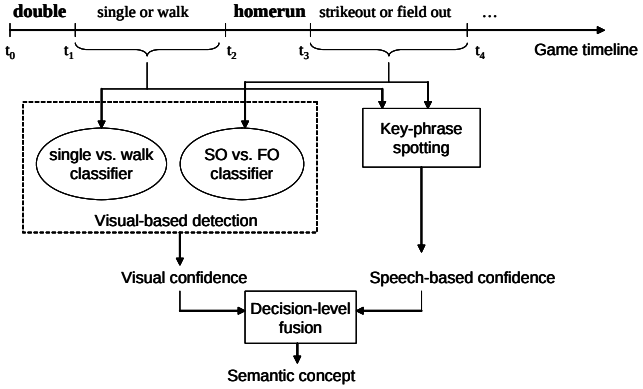

Figure 3. Examples of visual and speech event detection.

3.2. Event Detection from Speech Information

Because the commentator not only speaks the events just occurred but also review the player's past performance or chat with others, recognizing all his speech derives too many noises in event detection. In this work, we mainly appeal to key-phrase spotting module for facilitating confused event discrimination. Only the durations of the occurrence of confused events, such as ranges t_1 to t_2 and t_3 to t_4 in Figure 3, are applied with key-phrase spotting.

We exploit the key-phrase spotting system developed in [5], which is capable of extracting salient key-phrase fragments from an input utterance in real-time. High degree vocabulary flexibility and recognition accuracy can be achieved for any designate task. In this work, we define the mapping between baseball events and commonly used key-phrase in advance, as shown in Table 1. A specific speech event is identified if one or more of its corresponding key-phrases are recognized in the designated duration. For example, if the phrase 'touch out' is recognized in the case

of ‘strikeout vs. field out’ confusion, the occurred event claimed by speech information is “field out” rather than ‘strikeout.’

Table 1. Mapping between events and conventional key-phrases (in Mandarin Chinese).

Events	Corresponding Key-phrases
Single	R_1={安打(hit), 一壘安打(single)}
Walk	R_2={觸身球(hit by pitch), 保送(walk), 四壞球(four balls)}
Strikeout	R_3={三振(strikeout), 三振出局(strikeout)}
Field out	R_4={刺殺(‘touch out’ or ‘out before reaching bases’), 接殺(catch out)}

Although the key-phrase spotting module is now only applied to recognize Chinese, it is capable to be extended to other languages. The same framework, including visual and speech events detection, is general for any baseball games.

4. INFORMATION FUSION

After detecting events from visual and speech data, the problem narrows to making the final decision according to the detected results. It’s a trivial task if both the opinions from video and speech are identical. For example, ‘strikeout’ is surely the final answer if both visual and speech events are claimed as ‘strikeout.’ However, because both visual and speech event detection modules are not perfect, it’s often that the opinions from different modalities conflict. Therefore, we define and evaluate the confidence of two opinions and make the final decision.

4.1. Confidence of Visual-based Detection

In constructing two classifiers that discriminate single from walk and strikeout from field out, visual information including pitch-field pattern, field shot duration, motion, and etc. are used as the feature vectors [3]. K-nearest neighbor modeling is used to construct these classifiers. We derive the posterior probabilities to be the confidence of visual events.

Let the feature vector from visual data be $\mathbf{x}_1$, and K_1 (K_2) be the number of patterns among $\mathbf{x}_1$’s K nearest neighbors that belong to class C_1 (C_2). The estimated posterior probabilities [6] are given by

$$P(C_1|\mathbf{x}_1)=\frac{K_1}{K} \text{ and } P(C_2|\mathbf{x}_1)=\frac{K_2}{K}, \quad (1)$$

where $K_1+K_2=K$, and thus $P(C_1|\mathbf{x}_1)=1-P(C_2|\mathbf{x}_1)$.

With the K-nearest neighbor classifier that classifies classes C_1 and C_2, a test vector $\mathbf{x}_1$ is assigned to class C_1 if $K_1>K_2$, with the confidence value $P(C_1|\mathbf{x}_1)$.

4.2. Confidence of Speech-based Detection

The confidence of speech event is represented by “the posterior probability of the event C_i occurs given the recognized key-phrases.” Similar to visual-based detection, the recognized key-phrases are viewed as feature vectors. In the case of ‘single vs. walk’ confusion, the feature vector from speech data $\mathbf{x}_2$ may be constructed only by the key-phrases relevant to single ($\mathbf{x}_2=R_1$), only by the key-phrases relevant to walk ($\mathbf{x}_2=R_2$), or both ($\mathbf{x}_2=R_1,R_2$). Considering these three cases, the posterior probabilities are estimated as:

Case 1:

$$P(C_1|\mathbf{x}_2=R_1)=\frac{\#(C_1)}{\#(\text{only the key-phrases in } R_1 \text{ are recognized})},$$

$$P(C_2|\mathbf{x}_2=R_1)=\frac{\#(C_2)}{\#(\text{only the key-phrases in } R_1 \text{ are recognized})}.$$

Case 2:

$$P(C_1|\mathbf{x}_2=R_2)=\frac{\#(C_1)}{\#(\text{only the key-phrases in } R_2 \text{ are recognized})},$$

$$P(C_2|\mathbf{x}_2=R_2)=\frac{\#(C_2)}{\#(\text{only the key-phrases in } R_2 \text{ are recognized})}.$$

Case 3:

$$P(C_1|\mathbf{x}_2=R_1,R_2)=\frac{\#(C_1)}{\#(\text{key-phrases in } R_1 \text{ and } R_2 \text{ are recognized})},$$

$$P(C_2|\mathbf{x}_2=R_1,R_2)=\frac{\#(C_2)}{\#(\text{key-phrases in } R_1 \text{ and } R_2 \text{ are recognized})}.$$

The notation #(.) denotes the number of a specific situation. Based on this estimation method, we evaluate the posterior probability of a speech event given the recognized key-phrases. Note that if no key-phrase in R_1 or R_2 is recognized, it means that no contribution can be derived from speech event detection, and the discrimination work is done by visual-based detection only.

The case of discriminating strikeout and field out is done by considering key-phrases in R_3 and R_4. In the experiments, these probabilities were estimated based on the results of speech event detection from five games.

4.3. Combining Visual and Speech Opinions

In the duration where events C_1 and C_2 (single and walk, for example) cannot be explicitly discriminated, assume that the event C_1 is detected from visual information, with confidence $P(C_1|\mathbf{x}_1)$. However, the event detected from speech information is C_2, with confidence $P(C_2|\mathbf{x}_2)$. These two opinions compete and we have to make the final decision by checking their confidence values. To combine the opinions from different classifiers, Kittler et al. [7] describe the theoretical framework of different combining strategies. On the basis of the features from visual and speech data $Z=(\mathbf{x}_1,\mathbf{x}_2)$, we apply the sum rule to combine visual and speech opinions as follows:

$$\text{assign } Z\rightarrow C_j \text{ if } \sum_{i=1}^{2}P(C_j|\mathbf{x}_i)=\max_{k=1}^{2}\sum_{i=1}^{2}P(C_k|\mathbf{x}_i). \quad (2)$$

Although Kittler et al. proposed five rules (sum, product, max, min, and majority vote) for combining classifiers, we have similar performances by applying

different rules. The experimental results shown in the next section are all based on the sum rule.

5. EXPERIMENTAL RESULTS

The discrimination performance is evaluated for three baseball games, which are totally nine hours in length and consist of 228 plays. The performance of three phases including visual event only, speech event only, and integrated decisions are demonstrated in Figure 4 (single vs. walk) and Figure 5 (strikeout vs. field out). F1 metrics, which jointly consider precision and recall, are illustrated.

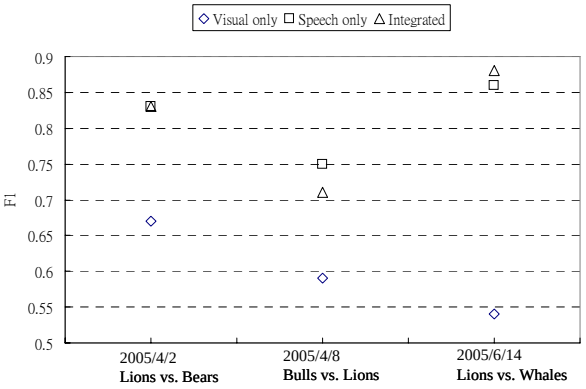

Figure 4. Discrimination performance of single vs. walk, where $F1 = \frac{2 \times Pr \times Re}{Pr + Re}$.

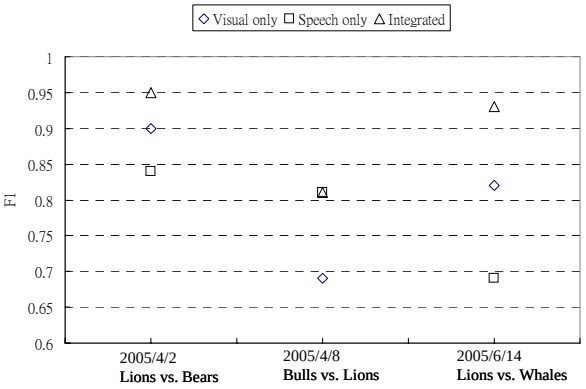

Figure 5. Dis. performance of strikeout vs. field out

In Figure 4, we see that combining two classifiers using the proposed fusion scheme outperforms single classifiers, except for the case of 'Bulls vs. Lions.' The cause of this exception lies on some extremely abnormal broadcasting situations or shot classification errors, which make the evaluation of visual event detection unreliable. Figure 5 shows the performance of discriminating strikeout from field out, where the enhancement of modality fusion is significant in two of the three games. The game of 'Bulls vs. Lions' relatively has worse performance because of lower character recognition accuracy caused by poorer video quality.

The overall performance of event detection (totally thirteen different types of events) is listed in Table 2. With the help of speech information, the detection performance increase (in terms of F1) by 8% ~ 20% relative to the visual only method [3].

6. CONCLUSION

We have presented a multimodal event detection method for broadcasting baseball videos. Based on the event boundaries determined by visual information, key-phrase spotting is applied to detect speech information. After estimating the confidences of event detection, two opinions from different modalities make a consensus and generate the final decision. Experimental results show that the proposed fusion scheme outperforms the single-modality-based approaches.

Table 2. Overall performance of event detection

Games	Decision	Precision / Recall	F1
Lions vs. Bears	Visual	0.88 / 0.82	0.85
	Visual + speech	0.96 / 0.89	0.92
Bulls vs. Lions	Visual	0.76 / 0.68	0.70
	Visual + speech	0.85 / 0.74	0.79
Lions vs. Whales	Visual	0.77 / 0.73	0.75
	Visual + speech	0.93 / 0.88	0.90

7. ACKNOWLEDGE

The authors would like to thank Prof. Lin-Shan Lee and his research team for providing their key-phrase spotting module. This work was partially supported by the National Science Council and the Ministry of Education of ROC under the contract No. NSC94-2752-E-002-006-PAE, NSC94-2622-E-002-024, and NSC94-2213-E-002-078.

8. REFERENCES

[1] L. Xie, P. Xu, S.-F. Chang, A. Divakaran, and H. Sun, "Structure Analysis of Soccer Video with Domain Knowledge and Hidden Markov Models," *Pattern Recognition*, vol. 25, no. 7, pp. 767-775, 2004.

[2] D. Zhang and S.-F. Chang, "Event Detection in Baseball Video Using Superimposed Caption Information," *Proc. of ACM Multimedia Conference*, pp. 315-318, 2002.

[3] W.-T. Chu and J.-L. Wu, "Integration of Rule-based and Model-based Methods for Baseball Event Detection," *Proc. of ICME*, 2005.

[4] Y. Rui, A. Gupta, and A. Acero, "Automatically Extracting Highlights for TV Baseball Programs," *Proc. of ACM Multimedia Conference*, pp. 105-115, 2002.

[5] B. Chen, H.-M. Wang, L.-F. Chien, and L.-S. Lee, "A*-Admissible Key-Phrase Spotting with Sub-Syllable Level Utterance Verification," *Proc. of ICSLP*, 1998.

[6] T.M. Cover and P.E. Hart, "Nearest Neighbor Pattern Classification," *IEEE Transactions on Information Theory*, vol. 13, pp. 21-27, 1967.

[7] J. Kittler, M. Hatef, R. Duin, and J. Mataz, "On combining classifiers," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, pp. 226-239, 1998.