Fusion of Multiple Asynchronous Information Sources for Event Detection in Soccer Video

Huaxin XU, Tat-Hoe FONG, and Tat-Seng CHUA
School of Computing, National University of Singapore
xuhuaxin@comp.nus.edu.sg, zanefong07@hotmail.com, chuats@comp.nus.edu.sg

Abstract

Our previous research shows that the use of multiple sources of information based on intrinsic AV features and external knowledge helps to detect events in soccer video. To make the system scalable, we process each source of information independently before fusing the detection results. The fusion of results is vital to the success under this architecture. However, this fusion problem is unique in that the detection results in terms of likelihood values to be fused are asynchronous. Thus, the fusion scheme has to determine which likelihood values are corresponding as well as the final likelihood. This paper formulates three fusion schemes, namely, rule-based scheme, aggregation and Bayesian inference, and studies their properties. Our results show that Bayesian inference has the best capabilities to tackle asynchronism.

1. Introduction

Semantic analysis of video based on multi-source knowledge is a promising trend. For the task of event detection in soccer video, our previous work shows that it helps to integrate both AV features and external knowledge in performing the analysis [1]. The AV analysis has been found to be effective in detecting the break–draw–attack structures at high level [5], and certain visual oriented events such as *goal*, *corner-kick*, etc. The external knowledge, including match reports and game log, provides details of almost all important actions, however, with only approximate timings because of human logging errors. We call the events detected from AV analysis as video events, and those extracted from external knowledge such as game log as text events.

In order to make the system viable in cases when certain sources of knowledge is not available, such as the external knowledge, the system architecture is designed such that the AV features and external knowledge are processed independently, and the results of video and text events are fused at the end [1][5]. Under this architecture, the fusion of detection results is vital to the success of the whole system. This is because AV analysis is accurate in detecting events' boundaries, but poor in event type identification, while the reverse is true for the text events extracted from external knowledge analysis. Fusion enables us to take advantage of the strength of each method. Our earlier results demonstrate that both detection accuracy and coverage were improved by fusion [5].

Many papers have been dedicated to the fusion problem for multimedia analysis, especially in a multi-modal setting. Similar to the problem of sensor-fusion, fusion for multimedia analysis can be categorized into fusion of features and fusion of decisions. Fusion of features joins and transforms all relevant features before sending as input to the classifier, such as Fisher's Linear Discriminant algorithm [2]. Fusion of decisions fuses outputs of individual classifiers, such as stacked SVM and bagging algorithms [2]. A special fusion of decisions is the fusion of rank lists. Wu et al [4] proposed an optimal multi-modal fusion scheme by first identifying a number of independent modalities, which can be viewed as fusion of features, followed by fusion of multiple modalities, which can be viewed as fusion of decisions. Most of the existing multi-modal fusion schemes assume that the features or decisions to be fused refer to the same sample, i.e., synchronized if samples are drawn from the same sequential source. However, this assumption may not hold in multi-source fusion of temporal items. The items to be fused, such as the pair of video and text events, often have different timings. In general, video event timings are accurate, while the text events are often off-set or misaligned slightly. This is because of human errors in entering game log entries as the operator needs time to judge the outcome and/or type of the action. In view of this, the fusion scheme has to identify which video and text events are corresponding in determining the final likelihood.

In this paper, we formulate three fusion schemes, namely rule-based scheme, aggregation, and Bayesian inference. The rule-based scheme was first introduced in [5] and refined in [6]. It assumes that the temporal off-sets between the game log entries and the actual occurrences of actions cannot be numerically modeled. In contrast, the aggregation scheme models the off-sets

0-7803-9332-5/05/$20.00 ©2005 IEEE

to follow a probabilistic distribution; and Bayesian inference models the off-sets to be binary.

The contributions of this paper are in highlighting the property of asynchronism which makes multi-source fusion of temporal items challenging, and in devising three fusion schemes with emphasis on tackling asynchronism.

2. Problem description

The basic problem is to fuse video events derived from AV analysis and text events extracted from external knowledge such as the game log. Each video or text event is represented by a triple - (event type, start time, end time), and so is each of the fused results. The time line on which text events are recorded (called text time line) is often off-set, while the video time line is accurate. The event types considered are the intersection of video and text event types. In the case of soccer, the event types are *goal*, *save*, *shot-off-target*, *penalty-goal*, *corner-kick*, and *free-kick*.

3. Fusion methods

Figure 1 illustrates asynchronism of events on video and text time lines and how the three fusion schemes tackle asynchronism.

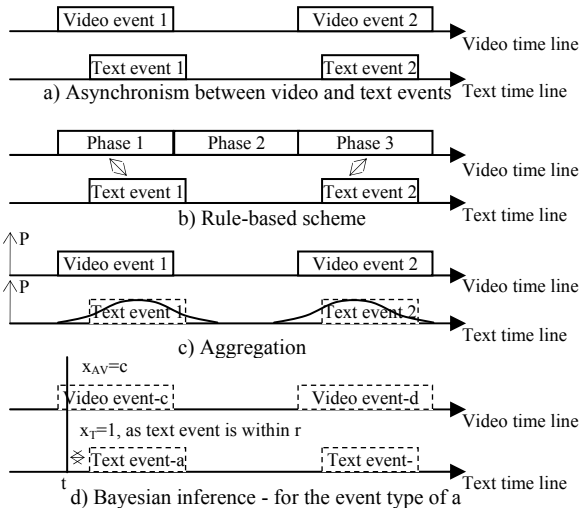

Figure 1 Asynchronism and fusion schemes

3.1. The rule-based scheme

The rule-based scheme follows the guideline of "identifying the pair of items first before fusing them". It is accomplished in three steps.

a) Aligning text events and phases. A phase is a *break*, a *draw* or an *attack* on the video time line, detected by the AV global analysis [5]. Given that there `is no modeling of text event off-sets, the alignment is sought by maximizing the number of matches between text events and phases. Here a *match between a text event and a phase* means that the phase conforms to the event modeling of the text event [5] (e.g., an *attack* followed by a *break* conforms to *goal*'s event modeling), are within a temporal range, and they occur in the same sequential temporal order. In view of temporally overlapping text events, a phase may match multiple text events. Thus, this problem is similar to the Longest Common Subsequence (LCS) matching problem and can be solved by dynamic programming technique.

b) Determining event type. Event type at each matched phase is determined based on text and video events' comparative accuracy and completeness [6].

c) Determining temporal boundaries. This step picks two boundaries of AV analysis units, which are nearest to the text event's boundaries respectively, and make them the boundaries of the final event [6].

3.2. Aggregation

The aggregation scheme models the temporal off-sets probabilistically and transforms the text events to a curve of likelihood on the video time line. It has three steps as explained below.

a) Modeling the distribution of actual occurrence of actions with respect to off-sets based on training data. This is similar to that of Yang et al [7] that modeled the probability of a face occurring along the time with respect to when the name is mentioned.

b) Computing the likelihoods of video and text events on the video time line. Having the distributions on the video time line that describe the start and end of the text event, the probability of any time point t being in an event occurrence is

$$P_{in}(t) = \int_{-\infty}^{t} D_s(x)dx \cdot \int_{t}^{+\infty} D_e(x)dx \quad (1)$$

where $D_s(x)$ and $D_e(x)$ are distributions of the start and end of the text event, respectively.

Likelihood of event type i at time point t generated by game log analysis is

$$P_{i-T}(t) = C_i P_{in}(t) \quad (2)$$

where C_i reflects the confidence of game log analysis on event type i. We take it to be the precision of game log analysis on event type i over the training set.

Suppose event type j is declared by AV analysis at time point t, the likelihood of event type i at time point t generated by AV analysis is

$$P_{i-AV}(t) = Confusion_{ji} \quad (3)$$

where $Confusion_{ji}$ is the element of confusion matrix that indicates the percentage of genuine type i samples among all samples detected to be of type j. Note that the confusion matrix includes the null event type.

c) Aggregating the likelihoods of video and text events. Let $P_{i-AV}(t)$, $P_{i-T}(t)$ and $P_i(t)$ denote the

likelihoods generated by AV analysis, game log analysis and the final likelihood. Several algorithms presented by [3] are investigated to fuse these events:

Min $$P_i(t) = \min(P_{i-AV}(t), P_{i-T}(t)) \quad (4)$$

Max $$P_i(t) = \max(P_{i-AV}(t), P_{i-T}(t)) \quad (5)$$

Product $$P_i(t) = P_{i-AV}(t) \cdot P_{i-T}(t) \quad (6)$$

Sum $$P_i(t) = w_{i-AV} P_{i-AV}(t) + w_{i-T} P_{i-T}(t) \quad (7)$$

where w_{i-AV} and w_{i-T} are determined empirically. If $P_i(t)$ is greater than a pre-defined threshold, the event occurrence at time point t is declared to be of type i as the fused result.

3.3. Bayesian inference

Instead of modeling the off-sets in probabilistic distribution, this scheme only differentiates if the off-set is under a maximum allowable value. To determine whether a particular event type p is present at time point t, there is a binary hypotheses: H_0 - not present, and H_1 - present. Most likely hypothesis is $\arg\max_{i\in\{0,1\}} P(x_{AV}, x_T \mid H_i) \cdot P(H_i)$, where $x_{AV} \in \{null, e_1, ..., e_N\}$ (N - number of event types) indicates the video event type declared at time point t, and $x_T \in \{0,1\}$ indicates whether p is detected in the span of $[t-r, t+r]$ on the text time line, where r is the maximum allowable off-set. x_{AV} and x_T are from two independent analysis, hence

$$\arg\max_{i\in\{0,1\}} P(x_{AV}, x_T \mid H_i) \cdot P(H_i) = \arg\max_{i\in\{0,1\}} P(x_{AV} \mid H_i) \cdot P(x_T \mid H_i) \cdot P(H_i) \quad (8)$$

$P(H_i)$, $P(x_{AV} \mid H_i)$ and $P(x_T \mid H_i)$, $i = 0, 1$ are obtained from the training set.

4. Experiments and Discussion

To test the effectiveness of the three fusion schemes, we use 5 English Premier League (EPL) matches for training and another 5 for testing. Table 1 shows the results of individual AV/text analysis and fused results by the schemes. As events only have approximate boundaries, an event is regarded as correct if it has the correct event type, and the start/end time is within a tolerance range from the ground truth.

Comparing the results of individual AV/text analysis vs. fused results, we observe that: a) Fused results are generally better than video events in both recall and precision, regardless of the fusion scheme used. This is attributed to the introduction of reliable textual information that is independent from AV signals that helps to constrain the search space. b) As compared to text events, fused results have higher recall, thanks to video events contributing boundaries that are more accurate than those of text events, but have slightly lower precision, due to errors in AV analysis and fusion process. In general, fused results are better than results of individual analysis.

Table 1 Event detection results after fusion

*Max is used as the combining algorithm for *save*, and Sum is used as the combining algorithm for the other event types.
% *Save* and *shot-off-target* combined for video events.

		Video events	Text events	Rule-based	Aggregation*	Bayesian inference
Goal	Recall	0.88	0.71	1.0	1.0	1.0
	Precision	0.71	1.0	1.0	1.0	1.0
Save	Recall	%Recall: 0.63, Precision: 0.62	0.77	0.9	0.83	0.93
	Precision		1.0	0.92	0.92	0.88
Shot-off-target	Recall		0.66	0.91	0.82	0.89
	Precision		1.0	0.89	0.80	0.84
Penalty-goal	Recall	1.0	1.0	1.0	1.0	1.0
	Precision	0.25	1.0	1.0	1.0	1.0
Corner-kick	Recall	0.68	0.73	0.89	0.65	0.89
	Precision	0.81	1.0	0.89	0.81	0.85
Free-kick	Recall	0.60	0.67	0.87	0.80	0.87
	Precision	0.53	1.0	0.93	0.86	0.87

We further examine the results by each fusion scheme. Aggregation has poorer recall or precision in some event types than the rule-based scheme, namely, the recall in *save*, *shot-off-target*, and *corner-kick*, and the precision in *shot-off-target* and *corner-kick*. A major cause of this is that the variance in off-sets for certain temporal entity, such as the event start, end or center time, is quite large for some event types. The variance in off-sets for a temporal entity can be defined as $\theta = (O_{\max} - O_{\min})/S$, where $O_{\max}$ and $O_{\min}$ are the largest and smallest off-sets of the respective temporal entity, and S the average temporal span of the event type which is used for normalization. Larger θ means more randomness in off-sets. It is found that some event types have large θ for event start or end (up to 3). For these event types, there would be large overlap between the ranges of start and end time of the same event. In this case, the probability $P_{in}(t)$ stays small during the whole duration that the event is likely to occur. Smaller $P_{in}(t)$ in turn makes the combined probability $P_i(t)$ small. Therefore, it becomes difficult to define a threshold that distinctly separates segments of positive and negative instances.

Table 2 Performance of different combining algorithms for the aggregation scheme

+R stands for recall; -P stands for precision.

Event type	θ	Min		Max		Product		Sum	
		R+	P-	R	P	R	P	R	P
Goal	0.61	0.81	1.0	1.0	0.89	0.81	1.0	1.0	1.0
Save	0.49	0.43	0.94	0.83	0.92	0.32	1.0	0.83	0.67
Shot-off-target	0.74	0.47	1.0	0.80	0.69	0.38	0.85	0.82	0.80
Penalty-goal	1.12	0.67	1.0	1.0	0.75	0.67	1.0	1.0	1.0
Corner-kick	2.93	0.49	1.0	0.63	0.54	0.37	0.79	0.65	0.80
Free-kick	1.45	0.58	1.0	0.76	0.54	0.45	0.83	0.85	0.90

The variance in off-set not only affects the overall performance of aggregation for each event type, it also has an impact on the relative performance among combining algorithms. Table 2 shows that event types with small θ, such as *save*, favor Max, while event types with relatively larger θ, favor Sum. The relative performance of the combining algorithms varies as θ varies, making it difficult to choose an algorithm that yields the best results for all event types. Note that here we use θ for event center to describe the off-set of an event as a whole.

Bayesian inference has recall and precision rates comparable to those of the rule-based scheme. This is because Bayesian inference is less influenced by large θ (this property will be discussed later). The slightly lower precision of Bayesian inference is because the relative locations of video events are not differentiated with regards to text events (i.e., before or after), as long as they are within the maximum allowable off-set. Thus, multiple video events could be regarded as positive instances, while only one is correct.

The rule-based scheme generally produces the best recall and precision among the three schemes, thanks to reliable detection of phases and text events. Errors mainly arise from missing or wrong phases and misalignment between text events and phases.

To compare the sensitivity of aggregation and Bayesian inference to θ, we conduct further experiments with varying θ for event center. Larger θ for event center implies larger θ for event start/end as well. To vary θ to a desired value, we multiply the off-set of event center by a factor, which stretches (or shrinks) $(O_{\max} - O_{\min})$ while keeping the span of events unchanged. The factor used is event type-specific, determined by the desired θ, S - the average temporal span of the event type, and R_0 - the original $(O_{\max} - O_{\min})$. Figure 2 depicts the changes in precision/recall rates in response to θ for aggregation and Bayesian inference methods. It also shows how the best-performing threshold for aggregation varies with changing θ.

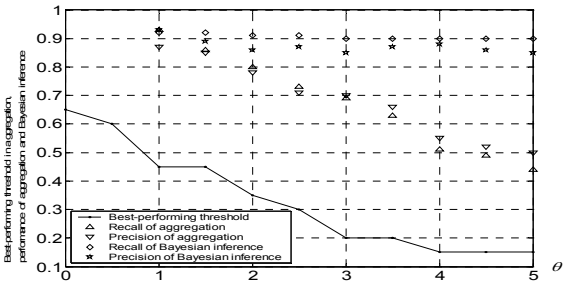

Figure 2 Sensitivity of performance of aggregation and Bayesian inference algorithms to θ

We can see from Figure 2 that as θ grows, the best-performing threshold for aggregation decreases, and positive and negative instances become more difficult to separate. Consequently, its precision and recall rates decline. In contrast, the precision and recall rates of Bayesian inference stay almost constant, which implies that Bayesian inference is less sensitive to large θ.

5. Conclusion

This paper highlights the problem of multi-source fusion, with emphasis on the challenging property of asynchronism. To tackle this problem, the paper presents three schemes and studies their properties. The findings are: a) aggregation is sensitive to large variance in off-sets, while Bayesian inference is not; and b) combining algorithms used in aggregation are inconsistent in relative performance. Both weaknesses of aggregation may be caused by detailed probabilistic modeling of off-sets, which does not exhibit strong and consistent patterns. Although the rule-based scheme performs best in fusing text events extracted from game log analysis and video events, it has poor adaptability. This is because accurate and complete text events are required in the alignment of text events and phases and in determining event type. Thus the rule-based scheme is not recommended. Bayesian inference has precision/recall rates close to the rule-based scheme, good adaptability and is insensitive to the variance in off-sets. Therefore, Bayesian inference is advised for this type of fusion problem.

6. References

[1] T.H. Fong, "Summarization and Retrieval of Football Highlights", Honours Year Project Report, National University of Singapore, 2004.

[2] A. Hauptman, R.V. Baron, M.-Y. Chen, M. Christel, P. Duygulu, C. Huang, R. Jin, W-.H. Lin, T.Ng, N. Moraveji, N. Papernick, C.G.M. Snoek, G.Tzanetakis, J.Yang, R.Yan, and H.D. Wactlar, "Informedia at TRECVID 2003: Analyzing and Searching Broadcast News Video", TRECVID2003 papers, Gaithersburg, November 2003.

[3] B. Tseng, C-Y. Lin, M. Naphade, A.Natsev and J. Smith, "Normalized Classifier Fusion for Semantic Visual Concept Detection", 2003 International Conference on Image Processing, Barcelona, September 2003.

[4] Y. Wu, E. Chang, K.C.-C. Chang, and J.R. Smith, "Optimal Multimodal Fusion for Multimedia Data Analysis", ACM Multimedia 2004, New York, October 2004.

[5] H. Xu, T.-S. Chua, "The Fusion of Audio-Visual Features and External Knowledge for Event Detection in Team Sports Video", the 6th ACM SIGMM International Workshop on Multimedia Information Retrieval, New York, October 2004.

[6] H. Xu, T.-S. Chua, "Detecting Events in Teams Sports Video", the 8th International Workshop on Advanced Image Technology, Jeju, January 2005.

[7] J.Yang, M-Y. Chen, and A. Hauptmann, "Finding Person X: Correlating Names with Visual Appearances", the 3rd International Conference on Image and Video Retrieval, Dublin, July 2004.