SOCCER HIGHLIGHT DETECTION USING TWO-DEPENDENCE BAYESIAN NETWORK

Jianguo Li, Tao Wang, Wei Hu, Mingliang Sun, Yimin Zhang

Intel China Research Center, Beijing, P.R. China, 100080
{jianguo.li, tao.wang, wei.hu, mingliang.sun, yimin.zhang}@intel.com

ABSTRACT

Soccer highlight detection is an active research topic in recent years. One of the difficult problems is how to effectively fuse multi-modality cues, i.e. audio, visual and textual information, to improve the detection performance. This paper proposes a novel two-dependence Bayesian network (2d-BN) based fusion approach to soccer highlight detection. 2d-BN is a particular Bayesian network which assumes that each variable depends on two other variables at most. Through this assumption, 2d-BN can not only characterize the relationships among features but also be trained efficiently. Extensive experiments demonstrate the effectiveness of the proposed method.

1. INTRODUCTION

With the great increasing of video contents, soccer highlight detection is becoming an active research topic in multimedia analysis. There is rich literature on highlight detection [2, 7, 3, 8]. Despite years of extensive research, highlight detection is still a challenging problem due to the semantic gap between low-level features and high-level semantic events. In general, there are three modalities in video production, i.e. the audio, visual and textual modality. How to fuse multimodal cues to robustly detect highlights is one of the most important problems.

Existing works of multimodal fusion can be categorized according to *fusion approach* (i.e. what kind of classifier is used), and *fusion strategy* (i.e. how to combine classifiers to obtain the final result). According to *fusion meta-approach*, existing works can be divided into rule based approaches and machine learning based approaches. Duan et al [2] used game-specific rules to classify events. Although rule system is intuitive to yield adequate results, it lacks of scalability and robustness. Therefore, most other works are based on machine learning techniques. Wang et al used SVM to detect events [7]. SVM is a good classifier particularly for small sample set. However, it may not sufficiently characterize the relations and temporal layout of features. Some researchers utilized Naive Bayesian classifier (NBC) to detect specific events [3]. NBC assumes that features are independent of each other, and thus neglects the important relations among features too. HMM based approach detects events by automatically determining video hidden states and state transitions [8]. However, this approach is less stable due to small training data set available and the complex parameter estimations procedure.

According to *fusion strategy*, there are two typical fusion frameworks, i.e. *flat fusion* and *stack fusion* framework. The former fuses modalities in the feature space, while the later fuses modalities in the semantic space. In detail, flat fusion first concatenates multimodal features into one single vector, and then adopts learning algorithms over the flat vector space to detect events. In contrast, stack fusion first utilizes supervised learning over each unimodal features to detect events, and then combines the detection result of each unimodal to make the final decision. Snoek [6] compared these two kinds of fusion frameworks on the semantic concept detection of TRECVID, and demonstrated that the stack fusion performs better and more robustly than the flat one.

In this paper, we propose a novel two-dependence Bayesian Network (2d-BN) based fusion approach to soccer highlight detection. 2d-BN is a particular Bayesian network which assumes that each variable depends on two variables at most. Under this assumption, 2d-BN is easy for training without losing the power of characterizing complex relations among features. To minimize the semantic gap between low-level features and high-level events, we take advantage of mid-level keyword representation framework [2]. The system first detects mid-level semantic keywords from low-level audio/visual features, and then 2d-BN infers events according to the multiple keyword streams in a stack fusion framework.

The rest of this paper is organized as follows. In Section 2, we present the system framework, and briefly describe how to generate mid-level keywords. In Section 3, we propose the 2d-BN algorithm, and the fusion framework. To evaluate the effectiveness of this approach, extensive experiments with over 12.6 hours of soccer videos are reported in section 4. Finally, concluding remarks are given in section 5.

2. SYSTEM OVERVIEW

2.1. Framework of highlight detection

We view a broadcast soccer video from the perspective of a program editor. Based on a predefined semantic intention, an

editor combines certain multimedia layout and content elements to express highlight events. Hence, highlight detection can be viewed as a reverse process of authoring [5].

Fig.1 illustrates our system framework. To minimize the semantic gap between low-level features and high-level events, we adopt the mid-level keyword representation [2]. The framework consists of three levels, i.e. low-level feature extraction, mid-level semantic keywords generation and high-level event detection. In processing, the low-level module first extracts audio/visual features from the video stream. Then the mid-level module uses the above audio/visual features to detect semantic keywords, e.g. view type, playing field position, excited speech, etc. Finally, the high-level module infers highlight events in the semantic space of these keyword streams.

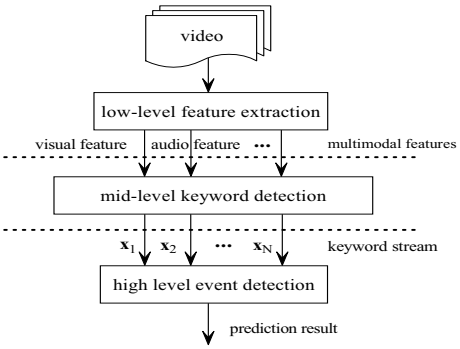


Fig. 1. Overall system framework.

2.2. Mid-level keyword generation

The mid-level module generates relevant semantic keywords from low-level audio/visual features. Details of keywords generation are described as following.

-$\mathbf{x}_1$ View type: By accumulating HSV color histogram, we get the dominant color to segment the playing field region. According to the area of playing field and the size of player, we then classify each shot into global view, medium view, close-up view and out of view [3, 2]. Fig. 2 shows examples of these view types.

-$\mathbf{x}_2$ Play-position : We classify the play-position of those global-view shots into five regions: left, mid-left, middle, mid-right and right as shown in fig.3. We first execute Hough transform to detect field boundary lines and the penalty box lines. Then a decision tree based classifier determines the play position according to lines' slope and position [3, 7].

-$\mathbf{x}_3$ Replay: It is an important cue for highlights, since a replay usually follows a highlight. At the beginning and ending of each reply, there is generally a logo flying in high speed. We detect these logos to identify replay by dynamic programming [10].

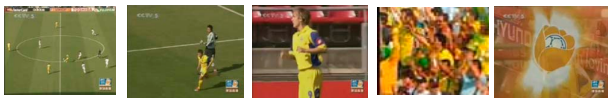
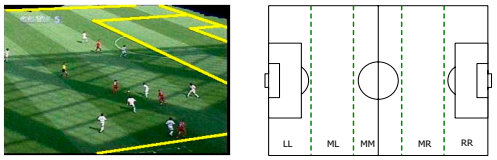


Fig. 2. From left to right, these are examples of global view, middle view, close-up view, out of view, and replay logo.



(a) (b)

Fig. 3. (a) Hough line detection on a segmented playing field (b) Five regions of playing field.

-$\mathbf{x}_{4,5}$ Audio keywords: There are some significant sounds that have strong relations to some soccer highlights such as goal, foul, etc. Our system detects two types of audio keywords: commentator's excited speech, and referee's whistle. Gauss mixture model (GMM) and SVM classifiers are used to detect the above two keywords respectively from low-level audio features including Mel frequency Cepstral coefficients (MFCC), linear prediction coefficient (LPC) and pitch [2, 9].

We utilize a post-processing to enhance the keyword detection precision. Since the keyword value will not change too fast, any sudden change in one keyword stream can be viewed as an error, and will be eliminated. After the post-processing, the high-level event detection module fuses the multiple keyword streams in the semantic space.

3. HIGHLIGHT DETECTION

Highlights in soccer videos are the special events that audiences are specially interested in, e.g. goals, shots, fouls and free-kicks etc. In this section, we first propose the 2d-BN algorithm, and then describe 2d-BN based fusion approach for soccer highlight detection.

3.1. Two-dependence Bayesian network

Given a m-dimensional feature vector $\mathbf{x} = [x_1, x_2, \cdots, x_m]^T$ and its class label y, the proposed 2d-BN assumes that each variable x_j depends on at most two variables, i.e. x_i and y. Given an observed feature vector $\mathbf{x}$, its label y is inferred by:

$$y^* = \arg\max_y P(y, \mathbf{x}) \tag{1}$$

We first present the inference of 2d-BN, and then handle the structure learning problem. Suppose x_i is given to be the parent of variables $\mathbf{x}_{\setminus i}$, where $\mathbf{x}_{\setminus i}$ indicates all elements in $\mathbf{x}$ except x_i. Under this assumption, the basic 2d-BN structure

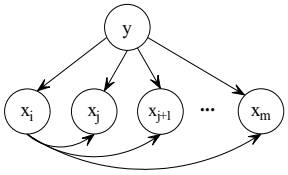


Fig. 4. The 2d-BN's Bayesian Network structure when assuming x_i to be the parent of variables $\mathbf{x}_{\setminus i}$

is shown in Fig.4, and using the product rule of probability, we have:

$$\begin{aligned} P(y,\mathbf{x}) &= P(x_i,y)P(\mathbf{x}_{\setminus i}|x_i,y) \\ &\cong P(x_i,y)\prod_{j=1,j\neq i}^{m} P(x_j|x_i,y) \triangleq P_i(y,\mathbf{x}) \quad (2) \end{aligned}$$

Since $P(x_j|x_i,y) = P(x_i,x_j|y)P(y)/P(x_i,y)$, we further have:

$$\begin{aligned} P_i(y,\mathbf{x}) &= P(x_i,y)\prod_{j=1,j\neq i}^{m}\frac{P(x_i,x_j|y)P(y)}{P(x_i,y)} \\ &= P(y)P(x_i|y)^{2-m}\prod_{j=1,j\neq i}^{m} P(x_i,x_j|y) \quad (3) \end{aligned}$$

In the training phase of 2d-BN, $P(y)$ is estimated using Laplace estimation [1]: $P(y=c)=\frac{\#(y=c)+1}{N+C}$, where $\#(y=c)$ is the number of instances satisfying $y=c$, N is the total number of training instances, and C is the number of classes. $P(x_i|y)$ is estimated using M-estimation [1]: $P(x_i=a|y=c)=\frac{\#(x_i=a,y=c)+k\times p}{\#(y=c)+k}$, where $p=P(x_i=a)$ and $k=2$. $P(x_i,x_j|y)$ is estimated using M-estimation which is similar to $P(x_i|y)$, and thus omitted here.

Now we deal with the structure learning problem, i.e. determining which feature x_i should be the parent node. In literature, many score functions have been proposed for learning Bayesian network structure [4]. For 2d-BN, we use the mutual information $I(x_i,y)$ as the criterion:

$$I(x_i,y)=\sum_{x_i,y}P(x_i,y)\log\frac{P(x_i,y)}{P(x_i)P(y)} \quad (4)$$

Generally, the larger the value $I(x_i,y)$ is, the stronger the dependence between x_i and y. Hence the variable x_i with the largest mutual information score should be selected as the parent node besides y.

Since 2d-BN is just an approximation to the exact Bayesian Network structure, selecting out only one variable x_i as parent node of $\mathbf{x}_{\setminus i}$ may yield large variance and bias for the probability estimation. In practice, we select those nodes whose mutual information satisfies $I(x_i,y)>\epsilon$, as the parent nodes. Hence we can obtain many joint probability estimates $P_i(y,\mathbf{x})$. The final joint probability is robustly estimated by the geometric average as:

$$P(y,\mathbf{x})=\Big[\prod_{I(x_i,y)>\epsilon}P_i(y,\mathbf{x})\Big]^{1/K} \quad (5)$$

where K is the number of variables satisfying the constraint $I(x_i,y)>\epsilon$. The geometric average is analytically tractable, and can yield a much more robust probability estimation. It is not hard to prove that 2d-BN is optimal for the Bayesian inference with pairwise dependent variables.

By the two-dependence assumption, 2d-BN is very easy to train. In the training phase, we only need to estimate the conditional probability tables of $P(x_i,x_j|y)$ and $P(x_i|y)$, and calculate the mutual information $I(x_i,y)$. Given n training samples, it is easy to find out that 2d-BN's training complexity is $O(m^2n)$. In the detection phase, we just look up those trained probability tables and infer class label via Eqn(5) and Eqn(1). Hence 2d-BN can run very efficiently.

3.2. Multimodal fusion using 2d-BN

Since broadcast video generally uses a close-up view or a replay as break to emphasize the exciting event, we first filter multiple keyword streams to find candidate highlight regions, i.e. *play-break units*, using the algorithm described in [3]. Then each keyword stream of the play-break unit is time sampled and represented by a m-dimensional feature vector $\mathbf{x}_k=[x_{k,1},x_{k,2},\cdots,x_{k,m}]^T$, $k=1,2,\cdots,N$, N=5 keywords,and m= 40 time slices, which are then input into the 2d-BN based fusion framework for event detection.

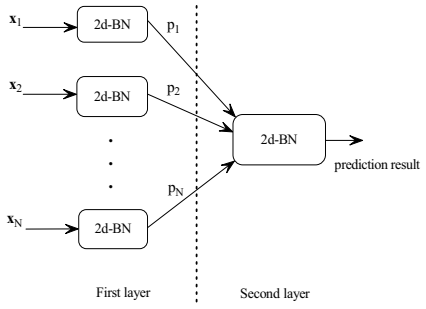


Fig. 5. 2d-BN based stack fusion framework.

Our 2d-BN based fusion approach is a two-layer stack fusion framework as illustrated in Fig. 5. The first layer characterizes the intra-keyword relations in the feature space, while the second layer characterizes the inter-keyword relations in the semantic event space. In detail, for each keyword represented by $\mathbf{x}_k$, we learn a 2d-BN model over it. Each 2d-BN in the first layer will produce a probabilistic score $p_k=P(y,\mathbf{x}_k)$, which indicates the probability that the current video segment belongs to a specific event y. For the convenience of learning the second layer's 2d-BN model, each probabilistic score p_k is quantized into five discrete values. Then the sec-

ond layer's 2d-BN infers the event from the combined probabilistic score vector $\mathbf{p} = [p_1, p_2, \cdots, p_N]^T$.

4. EXPERIMENTS

To demonstrate the effectiveness of the proposed approach, extensive experiments were conducted on eight soccer matches summing up to 12.6 hours of video. Five matches are used as training data, and the others as testing data. The ground truth is labeled manually.

We define semantic events "goal", "shot", "foul", "free kick" and "corner kick" as highlights, and all others as non-highlights. Table 1 summarizes the highlight detection performance by different methods. From Table 1, it is obvious that the proposed 2d-BN based approach performs better than SVM and Naive Bayesian (NBC), since SVM and NBC do not explicitly model any feature relations or temporal layout. Further for the comparison between the stack and the flat fusion framework, the stack fusion framework outperforms the flat one, which is consistent with the conclusion in [6].

Table 1. Comparison results for highlight detection

method	Detect	Miss	False	Precision	Recall
stack 2d-BN	186	19	19	90.7%	90.7%
stack SVM	184	21	28	86.8%	89.7%
stack NBC	175	30	26	87.1%	85.4%
flat 2d-BN	183	22	23	88.8%	89.3%
flat SVM	179	26	27	86.9%	87.3%
flat NBC	169	36	27	86.2%	82.4%

After detecting the highlight, we further classify the highlights into specific events by the two-layer 2d-BN based approach. Table 2 shows the event classification results. Following observations can be made from Table 2:

- The goal detection performance is perfect with nearly 92% in precision and 100% in recall due to the discriminative pattern generally with close-up view, penalty box, replay and excited speech etc.
- The performance of shot detection is good (about 80%). The false alarms and miss fault are mainly due to the confusion that shot and free-kick may happen simultaneously.
- The performance of foul/free-kick is similar to that of shot detection. This may be due to two facts. First, we currently do not treat offside as foul, which has similar features to that of foul. Second, there are some backfield fouls, which do not have discriminative features to distinguish it from other events.

5. CONCLUSIONS

In this paper, we propose a two-dependence Bayesian Network (2d-BN) based multi-modality fusion approach and apply it to soccer highlight detection. By decomposing a complex joint probability into the product of a series of simple two-dependence conditional probabilities, 2d-BN is much easier to train and can quickly infer a joint probability for event detection. To minimize the semantic gap, a mid-level semantic keyword module is utilized for soccer highlight detection. The system first extracts mid-level keyword streams from low-level features, and then employs 2d-BN to infer highlights in a stack fusion framework. Extensive experiments show that the 2d-BN based fusion approach is a fruitful direction to explore robust event detection systems.

Table 2. Specific event detection results using 2d-BN

Event	Detect	Miss	False	Precision	Recall
goal	11	0	1	91.7%	100%
shot	46	11	14	76.7%	80.7%
foul/free	70	13	19	78.6%	84.3%

6. ACKNOWLEDGEMENT

The authors are thankful to YangBo, WangFei, Sun Yi, Prof. Sun Lifeng, and Prof. Ou Zhijian of Dept. CS. and EE. of Tsinghua university for the research on the mid-level audio-visual keywords detection.

7. REFERENCES

[1] B. Cestnik. Estimating probabilities: A crucial task in machine learning. In *Proc. 9th European Conf. on Artificial Intelligence*, pages 147–149, 1990.

[2] L. Duan, M. Xu, T.-S. Chua, Q. Tian, and C. Xu. A mid-level representation framework for semantic sports video analysis. In *ACM Multimedia Conference*, 2003.

[3] A. Ekin, A. M. Tekalp, and R. Mehrotr. automatic soccer video analysis and summarization. *IEEE Trans. on Image processing*, 12(7):796–807, 2003.

[4] N. Friedman, D. Geiger, and M. Goldszmidt. Bayesian network classifiers. *Machine Learning*, 29(2):131–163, 1997.

[5] C. G. Snoek and M. Worring. Multimedia event-based video indexing using time intervals. *IEEE Trans on Multimedia*, 7(4):638–647, 2005.

[6] C. G. Snoek, M. Worring, and A. Smeulders. Early versus late fusion in sematic video analysis. In *ACM Multimedia Conference*, pages 399–402, 2005.

[7] J. Wang, C. Xu, E.Chng, K. Wan, and Q. Tian. Automatic replay generation for soccer video broadcasting. In *ACM Multimedia Conference*, 2004.

[8] L. Xie, S.-F. Chang, A. Divakaran, and H. Sun. Structure analysis of soccer video with hidden markov models. *Proc. ICASSP*, 4:4096–4099, 2002.

[9] M. Xu, N. Maddage, C.Xu, M. Kankanhalli, and Q.Tian. Creating audio keywords for event detection in soccer video. In *IEEE ICME 2003*, volume 2, pages 281–284, 2003.

[10] X. Yang, P. Xue, and Q.Tian. Repeated video clip identification system. In *ACM Multimedia 2005*, pages 227–228, 2005.