International Journal of Multimedia and Ubiquitous Engineering
Vol.10, No.10 (2015), pp.47-56
http://dx.doi.org/10.14257/ijmue.2015.10.10.05

Multi-Mode Semantic Cues Based on Hidden Conditional Random Field in Soccer Video

Yu Wang[1], Yu Cao[2], Miao Wang[3]and Gang Liu [4,*]

[1] *Capital Normal University, Beijing 100048, china*
[2]*Renmin University of China, Beijing 100048, china*
[3]*Harbin Sports Institute, Harbin 150000, china*
[4]*Physical Education Department, Harbin Engineering University, Harbin 150000, china*
13488850358@126.com

Abstract

A new framework based on multi- modal semantic clues and HCRF (Hidden Conditional Random Field) for soccer wonderful event detection. Through analysis of the structural semantics of the wonderful event videos, define nine kinds of multi- modal semantic clues to accurately describe the included semantic information of the wonderful events. After splitting the video clips into several physical shots, extract the multi- modal semantic clues from the key frame of each shot to get the feature vector of the current shots, and compose the observed sequence of the feature vectors of all shots in the test video clips. Using the above observed sequence as HCRF model input in the case of small-scale training samples, establish wonderful event detection HCRF model effectively. Experiments show that the average recall of this paper reaches 95.3%, the average precision rate reaches 96%, the performance of this paper is obviously superior to the contrast method.

Keywords: *pixel unmixing, end-member extraction, data field, weighted posterior probability support vector machine*

1. Introduction

To solve the automation problem of event detection, plenty of machine learning algorithms were widely applied, including Dynamic Bayesian Network (DBN) model, Hidden Markov Model (HMM), Conditional Random Fields model and Support Vector Machine (SVM) model [1-2]. However, those machine learning algorithm models have obvious shortcomings. To be specific, HMM requires possibly complete sample space. The model construction is complex [3-4]. The computational amount is huge. Moreover, it needs conditional independence assumption when creating the model, not describing the true structure of events [5-6]. The machine learning methods based on SVM take semantic event detection directly as feature classification problem to solve, not thoroughly utilizing semantic information, leading to bad performance of the detection. The machine learning strategies based on Conditional Random Fields (CRF) model can't describe interior structure and potential information of semantic events by defining hidden state variables when creating the event model, limiting the model's abilities to depict and recognize semantic events [7-8].

With the deeper investigation, Quattoni [9] introduced a brand new machine learning method [10-11], which is based on hidden conditional random fields (*i.e.* HCRF) model. The strategy incorporates merits of all algorithms mentioned above. It has been successfully applied in the field of gesture, voice and action recognition. The HCRF model makes full advantage of potential hidden state structure to discover effectively the

ISSN: 1975-0080 IJMUE

internal laws of semantic events. The manifested long-distance dependence and overlapping feature are more accordant with structural features of video signals. Meanwhile, according to video multi-granularity, through overall analysis and description of video semantic events in terms of video images and audios, as well as the fusion of multimodal semantic clues and HCRF model, a feasible new framework is developed to detect wonderful episodes in football videos [12].

2. Extraction of Multimode Semantic Clues

After analyzing the semantic structure of highlights in football match videos, it defines nine multi-pattern semantic clues, as to mine accurately the semantic information contained by such highlights and thus to express them roundly and clearly. Of the nine multimode semantic clues, the selection of scoring board SB, referee rate RR and frame motion FM are defined and retrieved by the following method:

2.1. Scoring Board (SB)

When there's foul, referee will show red card/yellow card. Then on the scoring board will show the name of the offender and information relating to the red/yellow card. In this case, SB can be regarded as an important semantic clue for red/yellow episode. Here we use Harris corner detection method to separate SB Board Area (BA), defining and giving Aspect Ratio of BA (AR_BA) and Area Ratio of BA (AreaR_BA). The design formula (1) (2) and quantitative rule of semantic shots is shown as (3):

$$AR\_BA = \frac{W\_BA}{H\_BA} \tag{1}$$

$$AreaR\_BA = \frac{AreaR\_BA}{p \times q} \tag{2}$$

$$SB = \begin{cases} 1, AR\_BA(i) < T_s, AreaR\_BA(i) < T_e \\ 0, else \end{cases} \tag{3}$$

$SB = 1$ means the scoreboard shots, otherwise is non scoreboard shots. Figure 1 (a) is a representative frame of the scoreboard shot.

2.2. Referee Ratio (RR)

Referee-relating information in the football videos can be used for detecting highlights, like red/yellow incidents. Here we employ the approach in [13] to detect referee' clothing, define the aspect ratio, area rate and Aspect Ratio of MBR (AR_MBR) of the minimum bounding rectangle (MBR) in referee's clothes Area Ratio of MBR (AreaR_MBR), which are calculated by and the quantitative rule is(4) (5) (6):

$$AR\_MBR = \frac{W\_MBR}{H\_MBR} \tag{4}$$

$$AreaR\_MBR = \frac{AreaR\_MBR}{p \times q} \tag{5}$$

$$PR = \begin{cases} 1, AR\_MBR(i) < T_u, AreaR\_MBR(i) < T_r \\ 0, else \end{cases} \tag{6}$$

$PR = 1$ means the Referee shots, otherwise is non Referee shots. Figure 1 (b) is a representative frame of the Referee shot.

2.3. Frame Motion (FM)

After one wonderful episode happens in the videos, there are slow-motion playback parts, of which the frame motion intensity is generally lower than normal motions. The frame motion FM is expressed as (7):

$$FM = \begin{cases} 1, m(i) = \dfrac{1}{p \times q \times | o_{\max} |} \sum_{t=1}^{p \times q} o_t(i) > T_m \\ 0, else \end{cases} \quad (7)$$

$FM = 1$ said that the slow shot, otherwise is normal shot.

(a) The Scoreboard Lens Representative Frame

(b) The Referee Shot Representation Frame

Figure1. The Scoreboard Shot, Representative Frame of Referee Shot

3. Detection of Highlights Based on Multimode Semantic Clues and HCRF Model

3.1. Definition of HCRF Model

According to the semantic structure of corner balls, penalty kicks and red/yellow incidents as well as the semantic information contained in the shot sequence, we build HCRF model for every event. The observation sequence and class tag are defined in the following. Regarding corner kicks, we perform physical shot segmentations and use the first frame of each shot as the key one, then fetch multimode semantic clues of every key frame. Those clues refer to field rate FR, player rate PR, midfield area MA, penalty area PA, corner area CA, and audio energy AE as well. They altogether form the characteristic vector $v = [FR, PR, MA, PA, CA, AE]$ of the shot, which is further defined as the observational value of the shot. Finally we get the observation sequence $[v_1, v_2, ..., v_{m'}]$ composed of m' shots of video fragments. Define HCRF model's category tag as $Y_v = \{0,1\}$, 0 for non-corner kick event, 1 for corner kick event.

Likewise, retrieve multimode semantic clues for the key frames of penalty kick events, including field rate FR, player rate PR, penalty area PA, midfield area MA and frame motion FM, which constitute the feature vector $t = \{RR, SB, FM\}$. Make t the

observational value of the shot. Then we get the observation sequence $[t_1, t_2, ..., t_{l'}]$ having l' shots of video clips. Define penalty kick event's classification tag as $Y_t = \{0,1\}$, 0 for non-penalty event, 1 for penalty event.

3.2. Training and Reasoning of HCRF Model

The proposed method based on multimode semantic clues and HCRF model is carried out by these steps: it shown in Figure 2

Step 1 Create training data set

Take corner kick event for example. Select artificially M_1 corner kick fragments and M_2 non-corner kick events to create training data set of them;

Step 2 Build HCRF model

As per different training data set, modify constantly hidden state number n and widow length ω. Use Quasi-Newton algorithm and data set training method to calculate the best parameter θ^* of HCRF mode and therefore get HCRF model of every event;

Step 3 Set up observation sequence

For one testing video clip, make physical shot segments. Then retrieve multimode semantic clues of key frames of each shot to acquire eigenvector of the current shot. Form the observation sequence with feature vectors of all shots in testing segments;

Step 4 Detect highlights

Use the acquired observation sequence as HCRF model's input of relative highlights. Apply confidence diffusion algorithm to estimate the logarithmic likelihood probability of videos corresponding to wonderful and not wonderful events' tags. Therefore, the classification tag with higher possibility is prediction tag of input observation sequence. The automatic detection of highlights is accordingly finished.

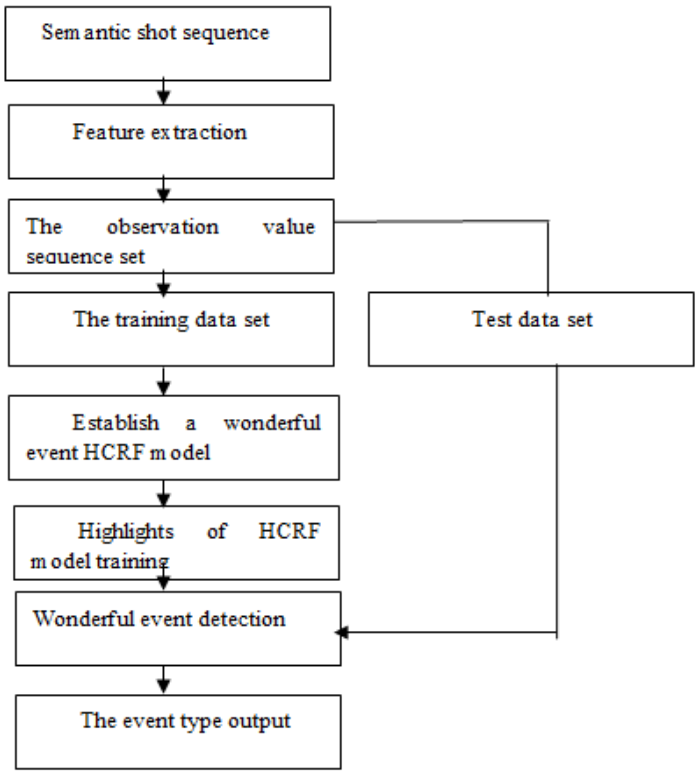


Figure 2. The Diagram of the Football Video Event Detection Algorithm

4. Experimental Analysis and Results

Experimental videos were collected from many sessions of 2010 South Africa FIFA, 2011 EPL and 2011 UEFA, MPEG format, 352 x 288 DPI, software environment Matlab R2008a. The empirical data include training data and testing data, of which training data for corner kicks have 20 corner kick fragments and 10 non-corner kick fragments; testing data have 30 corner kick clips and 20 non-corner kick clips; training data for penalty kicks have respectively 20 penalty and non-penalty kick fragments; testing data have 61 penalty kick fragments and 20 non-penalty kick fragments; for the red/yellow cards, training data have 20 red/yellow card clips and 10 non red/yellow card clips; testing data include 37 red/yellow card clips and 15 non red/yellow card clips. In the experiment, we use recall rate and precision rate to evaluate quantitatively the retrieval results of multimode semantic clues and detection results of above highlights. Due to the space here, Table1 only lists partial experimental video information about corner kick episodes.

Table 1. Experimental Video Information of the Corner Event

Video name	ID	Matches	Date of the match	Score	Video length
South Africa World Cup	F1	England VS USA	2010.6	1:1	106
	F2	Germany VS Australia	2010.6	4:0	102
	F3	Spain VS Switzerland	2010.6	0:1	107
	F4	Germany VS Argentina	2010.7	4:1	109
UEFA Champions League	U1	Real Madrid VS dynamo Zagreb	2011.11	6:2	95
	U2	Bayern Munich VS Villarreals	2011.11	3:1	106
	U3	Napoli VS Manchester City	2011.11	2:1	100
	U4	AC Milan VS Barcelona	2011.11	2:3	101
England Football Super League	E1	Chelsea VS Wigan	2012.4	2:1	109
	E2	Manchester United VS Queens Park Rangers	2012.4	2:0	96
	E3	A Senna VS Manchester City	2012.4	1:0	107
	E4	Tottenham VS Norwich	2012.4	1:2	102

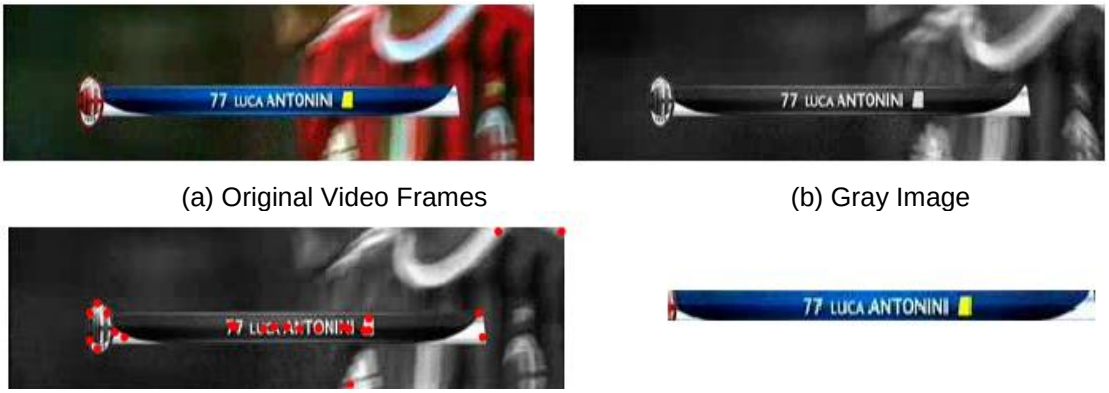
4.1. Experiment on Retrieving Multimode Semantic Clues

We defined nine multimode semantic clues, of which scoring board, referee rate and frame motions were chosen for the experiment. The retrieving results were described as follows:

4.1.1. Scoring Board (SB): Figure 3 displays the retrieval results of SB in the football videos. The picture3 (a) is the image formed after the lower half part of the original video frames was cut out. picture3(c) marked the corner location in the image and in red dot in picture 3(d) As seen picture 3, it located accurately the SB area in the image, laying good foundations for determining later the size and shape of SB area. Table2 gives the retrieval results of SB by our proposed method, suggesting that the method fetched effectively SB area, precision and recall rate reaching 89.71% and 96.83% separately.

4.1.2. Frame Motion (FM): Figure 4 graphs neighboring image frames and frame motions of both normal and slow motions in football videos. Figure 4 (a) (b) is image examples of adjacent frames chosen from normal and slow motions. The frame intensity of normal motions is obviously strong, semantically classifying well the normal and slow

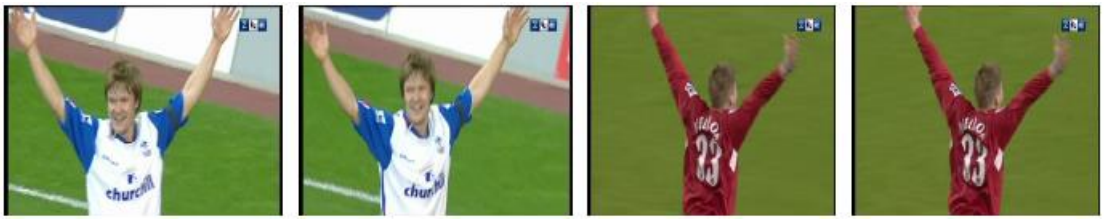
motions in the videos. Table2 also shows results of semantic shots classification by according to different kinds of frame motion clues. They imply that the proposed approach can discern effectively slow and normal motions, with precision rate 94.12% and recall rate 86.49%.



(a) Original Video Frames (b) Gray Image

(c) Corner Position Tagging (d) Results of Extracting Scoreboard Area

Figure 3. The Extraction Results of Scoreboard Area in Soccer Video



(a) The Normal Shot Adjacent Frame Image (b) The Slow Shot Adjacent Frame Image

Figure 4. Frame Motion Intensity and Frame Image of Normal Lens and the Slow Motion

4.1.3. Referee Rate (RR): Figure 5 is the experimental results of retrieving referee's original frames and dressing. Figure 5 (b) portrayed the retrieval results of referee's dresses. Noticeably, the method here realized the extraction of referee's clothes, with complete segmentation and good retrieval effects, meaningful to subsequently judge and determine effectively shots regarding referees. Table 2 is the results of referee rate by our method, which indicate that the method can effectively fetch SB area, precision and recall rate achieving respectively 98.04% and 83.33%.

(a) Referee Shot Original Image Frame (b) Referee Clothes Extraction Results

Figure 5. The Original Image Frame of Referee Shot and the Referee Shots Extraction Results

Table 2. Experiment Results of Multi-Pattern Semantic Cue Extraction

Semantic cue Extract type	Shot type	Actual number	Correct number	False number	Missing number	Precision%	Recall%
Scoreboard	Scoreboard shot	63	61	7	2	89.6%	96.5%
Frame motion intensity	Slow motion	34	32	5	2	94.6%	86.6%
Referee ratio	Referee shot	60	50	1	10	98.6%	83.6%

4.2. Experiment on Detecting Wonderful Events

To detect a few wonderful events with the HCRF model, we selected parameters like: hidden state number n (1-3) and window length ω (0,-2); where ω=0 means considering only the current observed value to predict the present hidden state; ω =1 means considering the current and also the previous and posterior observations, and the like. For each group of parameter setting, when the well trained HCRF model is utilized to detect highlights, the classification tag with bigger probability is the detection result of input testing videos. Table3 lists the experimental results of detecting some highlights by our method.

Table 3. The Experimental Results of More Exciting Event Detection

Parameters of HCRF model		Corner		Penalty		Red yellow card	
		Recall	Precision	Recall	Precision	Recall	Precision
n=1	$\omega=0$	93.75	73.17	100	85.92	81.08	73.17
	$\omega=1$	93.75	73.17	100	85.92	81.08	73.17
	$\omega=2$	93.75	73.17	100	85.92	81.08	73.17
	$\omega=0$	96.77	90.91	98.39	95.31	75.68	93.33
n=2	$\omega=1$	96.77	100	93.89	89.71	78.38	93.55
	$\omega=2$	96.77	100	96.83	88.41	72.97	96.43
	$\omega=0$	96.77	90.91	98.39	95.31	83.78	93.94
n=3	$\omega=1$	96.77	100	100	93.58	89.19	94.29
	$\omega=2$	96.77	90.91	100	88.41	89.19	91.67

From the Table 3, we learnt that the detection effect was affected a lot by the model parameter. We take corner kick event for instance:

(1) When the parameter n=1, hidden state number is smaller, not completely expressing the inner structure of video sequence; under the circumstances, no matter what window length ω is selected, precision and recall ratio won't change and the detection performance can't be improved;

(2) When hidden state number n=2, HCRF model's expressive capability was enhanced, describing better the internal laws of events and the overall detection performance becoming the optimal; recall and precision rate at respectively 96.77% and 90.91%; when n=2 is fixed and window length ω increased, HCRF model considered fully the dependence relationship between neighboring local observations when predicting the current hidden state, improving the detection performance, recall and precision rate at respectively 96.77% and 100%;

(3) When the model parameter n=3, window length ω enlarged from 0 to 1, precision rate appreciated from 90.91% to 100%;

(4) If we continue increasing window length ω, model structure became complicated, HCRF model's predictive ability was weakened owing to the increased number of neighboring local observations, precision rate down from 100% to 90.91%. Based on the above analysis, we conclude that the more the hidden states are, the bigger the window length is, the more complicated the HCRF model becomes; with time going by, the detection performance is declined instead of improved. Therefore, when the model parameter is n=2, $\omega=1, n=2, \omega=2, n=3, \omega=2$ the performance of detecting corner kick events achieves the best. In practice, in consideration of some factors like model complexity and time consumption, we choose $n=2, \omega=1$ as the optimal parameter of the HCRF model. Similarly, we get the optimal parameter $n=3, \omega=1$of HCRF model for detecting penalty and red/yellow card episodes.

In order to validate the adaption of the proposed solution to any game videos, we randomly chose several sessions of EPL and UEFA football videos. For the space limit, Table 4 shows partially experimental results of corner kick events. We see that the proposed method has good recall and precision rate, which is on average 95.2% and 97.2% for UEFA, and on average 93.3% and 100% for EPL. The detection performance is superior over the comparative method. Apparently, the proposed algorithm has certain adaptability to any game video

Table 4. Adaptability Test for Soccer Video

Video name	ID	Corner count	Correct number	Missing number	False number	Recall %	Precision %
UEFA Champions League	U1	8	7	1	0	87.5%	100%
	U2	8	8	0	1	100%	88.9%
	U3	15	14	1	0	93.3%	100%
	U4	5	5	0	0	100%	100%
England Football Super League	E1	7	6	1	0	85.7%	100%
	E2	4	4	0	0	100%	100%
	E3	5	5	0	0	100%	100%
	E4	8	7	1	0	87.5%	100%

6. Conclusion

This paper presents a new framework for soccer video highlights multimodal cues and detection based on HCRF model. Firstly, the fusion of audio and video features, constructed middle level semantic space using multi modal semantic clues, make up the semantic gap from the low-level features to high-level semantics. Secondly, the multi-pattern semantic cues to form the feature vector as the observation sequence of HCRF model. Finally, in small training sample spaces, and effective constructs HCRF model of soccer video highlights, realizes the automatic detection of exciting events.

Acknowledgements

This work was supported by The Fundamental Research Funds for the Central Universities. No. HEUCF151601.

References

[1] Quattoni A., Wang S. and Morency L. P., "Hidden conditional random fields", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 29, no. 10, **(2007)**, pp. 1848-1852.
[2] Yedidia J. S., Freeman W. T. and Weiss Y., "Generalized belief propagation", Advances in Neural Information Processing Systems, **(2001)**, pp. 689-695.
[3] M. D. Liwei and M. J. Chan, "Soccer video highlights the fusion of HCRF and AAM detection", Journal of computer research and development, vol. 01, **(2014)**, pp. 225-236.
[4] D. Xiping, L. Jiafeng, W. Jianhua and T. dragon, "A semantic level collaborative text image recognition method", Journal of Harbin Institute of Technology, vol. 03, **(2014)**, pp. 49-53.
[5] Y. Minghao, T. Jianhua and L. Hao, "Nest forest at multi-channel man-machine dialogue system for natural interaction", Computer science, vol. 10, **(2014)**, pp. 12-18+35.
[6] W. Lian. "And realize multimode teaching video semantic analysis", Nanjing University of Science and Technology, **(2014)**.
[7] Tian, "Study on construction of spatial knowledge obviously multi modal based on information fusion", Huazhong Normal University, **(2014)**.
[8] H. Yucheng, Y. Junqing, H. Xianqiang, H. Yunfeng and Tao, "User preference mining pipe in the engine of the soccer video search", China Journal of image and graphics, vol. 04, **(2014)**, pp. 622-629.
[9] Y. Junqing, Z. Qiang, W. Zengkai and H. Yunfeng, "Using the playback scene and emotion encouragement detection in soccer video highlights", Chinese Journal of computers, vol. 06, **(2014)**, pp. 1268-1280.
[10] Z. Yanjiao, "Regional map of target detection of video abstract Gauss", Hebei Normal University, **(2014)**.
[11] L. Yafei, "Study on detection of pedestrian tracking and abnormal motion video surveillance", China Jiliang University, **(2014)**.
[12] S. Chenhan, "Methods and annotation of video structure extraction", Computer knowledge and technology, vol. 26, **(2014)**, pp. 6178-6180.
[13] Tung T. P., Tuyet T. and Viet H. V., "Event Retrieval in Soccer Video from course to fine based on Multi- modal Approach", IEEE International Conference on Computing and Communication Technologies, Research, Innovation, and Vision for the Future, **(2010)**, pp. 1-4.

Author

Yu Wang, He received his B.S degree from Beijing Sports University, and received his M.S degree from Hebei Normal University. He is a lecturer in Capital normal university. His research interests include Sports teaching and training.