2016 IEEE 28th International Conference on Tools with Artificial Intelligence

Automatic Soccer Video Event Detection Based On A Deep Neural Network Combined CNN and RNN

Haohao Jiang, Yao Lu, Jing Xue
School of Computer Science, Beijing Institute of Technology
Beijing Laboratory of Intelligent Information Technology, Beijing, China
vis_yl@bit.edu.cn

***Abstract*—Soccer video semantic analysis has attracted a lot of researchers in the last few years. Many methods of machine learning have been applied to this task and have achieved some positive results, but the neural network method has not yet been used to this task from now. Taking into account the advantages of Convolution Neural Network(CNN) in fully exploiting features and the ability of Recurrent Neural Network(RNN) in dealing with the temporal relation, we construct a deep neural network to detect soccer video event in this paper. First we determine the soccer video event boundary which we used Play-Break(PB) segment by the traditional method. Then we extract the semantic features of key frames from PB segment by pre-trained CNN, and at last use RNN to map the semantic features of PB to soccer event types, including goal, goal attempt, card and corner. Because there is no suitable and effective dataset, we classify soccer frame images into nine categories according to their different semantic views and then construct a dataset called Soccer Semantic Image Dataset(SSID) for training CNN. The sufficient experiments evaluated on 30 soccer match videos demonstrate the effectiveness of our method than state-of-art methods.**

***Keywords*-Video Semantic Analysis; Convolution Neural Network; Recurrent Neural Network;**

I. Introduction

Sports video analysis, especially soccer video, has attracted a large number of researchers because of the practicability and interestingness compared with the ordinary video. According to the proposed methods, soccer video event detection mainly depends on the theory of pattern recognition.

The method [5], [10], [17] based on pattern recognition mainly consists of four steps. First, the soccer video is divided into shots and extracted the key frames from shots. In the next step, the shot features are extracted, including low image features and middle features. The middle features, like the rate of glass and penalty area, are calculated from low features such as color, edge. After the last step, the event boundary is determined according to the features of shot. As the event cannot be fully described by a single shot, the method of PB segment was proposed by [18] based on the match state to determine the event boundary. Then the event features are extracted, most of the event features are calculated by the shot features. In the finally step, the classifier like HCRF[15], DBN[10], BN[17], HMM[11], EHMM[14], is used to identify the soccer events including shoot, corner, foul, goal etc. For the soccer video, some researchers also use rules[19] to detect exciting events.

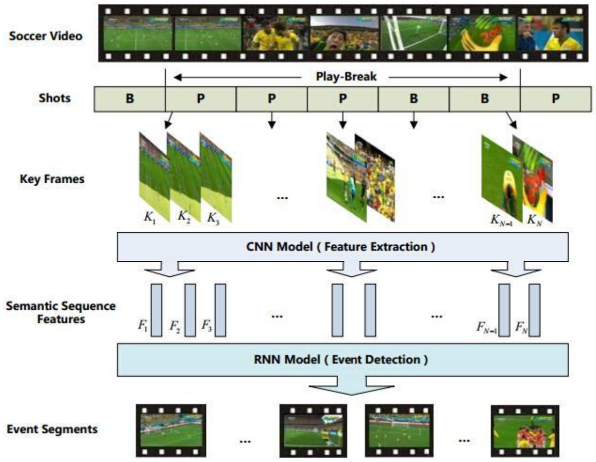
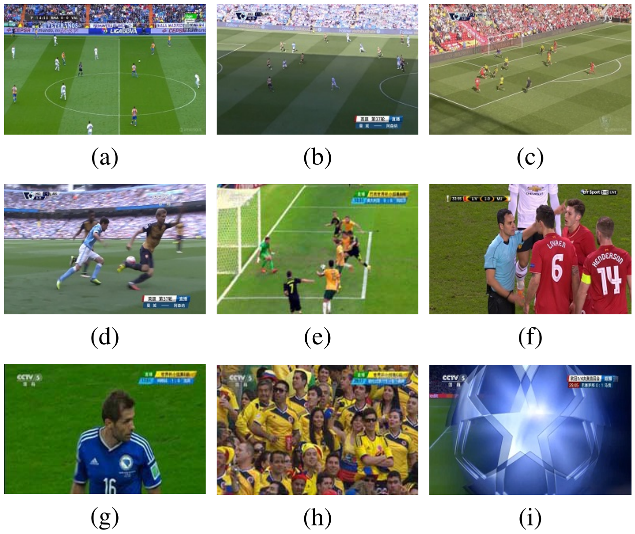


Figure 1. The flowchart of our approach

Although some effective methods have been proposed, the neural network method has never been appiled to soccer video event detection from now. The neural network method has made a lot of research achievements in recent years and the CNN has shown its great advantages on image classification[13] and object detection[7]. Some researchers have explored the way to use CNN to solve video analysis problem[6], [20], and proved the better effect in image processing than the traditional methods. As a method of evolution for neural network in dealing with temporal relation, the RNN has been widely used in Natural Language Processing(NLP)[16]. And to solve the exploding gradients problem [1] of the standard RNN model, some effective models like LTSM[8], [9], GRU[3] have been proposed. [12] has analyzed the advantages and disadvantages of different RNN models.

In the semantic analysis of traditional soccer video methods, the dimension of event features is relatively small, and one of the important reasons is that a separate algorithm is needed in the process of extracting each feature, which is not robust caused by illumination, visual angle and so on.

2375-0197/16 $31.00 © 2016 IEEE
DOI 10.1109/ICTAI.2016.78



(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

Figure 2. Examples of nine soccer image semantic views

And small features can not reflect the time relationship of events sufficiently, which is an important factor of forming an event. In this paper, we propose an effect and efficient framework to automatically detect the soccer video events, and the flowchart is shown in Fig 1. The main contributions of our work are: (1)based on the construction of SSID, we introduce the CNN to extract the semantic features of soccer video event automatically; (2)we use the RNN to fully dig the time relationship of semantic sequence features to realize the transition from low-level features to high-level event concept; (3)through the analysis of experimental data, we show the effective and accuracy of the proposed framework compared with previous methods in soccer video event detection.

The rest of our paper is organized as follows. In Section II, we describe the approach to use CNN in extracting the semantic features of the events of soccer video, and illustrate the situation of SSID. Section III presents the RNN structures in detail, we experiment with different RNN models and compare to the traditional methods in soccer video event detection in Section IV. Finally, Section V summarizes our main work and the direction of future work.

II. Semantic Sequence Features

A. Soccer Semantic Image Dataset

Most low-level and middle-level features are extracted from key frames in the traditional methods of soccer video event detection like[5], [10], [17], so it is essential to extract the appropriate and discriminative features of the image in soccer video event detection. The CNN has been proved more effective and automatical than the traditional methods in in image feature extraction. For avoiding the effect caused by insufficient dataset, [6] regard ImageNet [4] as the auxiliary dataset to obtained the initial parameters by pre-training the CNN network and then adjust the parameters to adapt to the new task of video event detection by fine-tuning.

Because different fine-tuning methods have great effect on the final results [2] and motivated by [6], [20], we establish our dataset to pre-train our CNN model by dividing the images into nine semantic views according to the potential semantic content of soccer video frames.

According to the focus position of the playground, we divide the Long View into three views, including Midfield Long View, Defend Long View and Penalty Long View, which is shown in Fig 2 (a)-(c). The midfield of playground is the area of match begining, so we divided this long view to Midfield Long View. For the soccer match, penalty area is the main area of exciting events, so we need Penalty Long View to determine this focus position.

For the Medium View, we also classify into three views containing Player Medium View, Goalkeeper Medium View and Referee Medium View based on the the identity of character. If a referee appears on the Medium View like Fig 2 (d), we called the view as Referee Medium View, the same to Goalkeeper Medium View in Fig 2 (e). And if only the players appears like Fig 2 (f), we called it Player Medium View. One reason of the appearance of referee as one type of our Medium View is that referee generally not appear in the view, unless there is some special situations, such as a foul, corner or red card event incident. And we regard Goalkeeper Medium view as one important view for that a goalkeeper in Medium View often appears when the player shoots or be ready to shoot.

At last we classify the rest into three important views, including Closeup View, Outfield View, Logo View, like Fig 2 (g)-(i). Logo View represents replay, generally we can see this view only when the wonderful event occurs.

To build the Soccer Semantic Image Dataset, we download 3500 minutes soccer match videos from the Internet and each video is coded by MPEG-4 with 50 frames per second. And then we extract the frames through a fixed interval T_v and classify manually those frame images into different views we proposed above. The parameter we set is T_v=200, and there are about 52500 images in our SSID dataset.

B. Training CNN Model

Some effective CNN structures has been proposed in the last research, like [13], which was mainly used for image classification. [2] proposed a developed CNN model for video analysis, and motivited by which, we proposed a kind of CNN model for extracting the semantic feature of soccer video event automatically, which is shown in Figure 3. We realize the CNN model by using the MatConvNet toolbox.

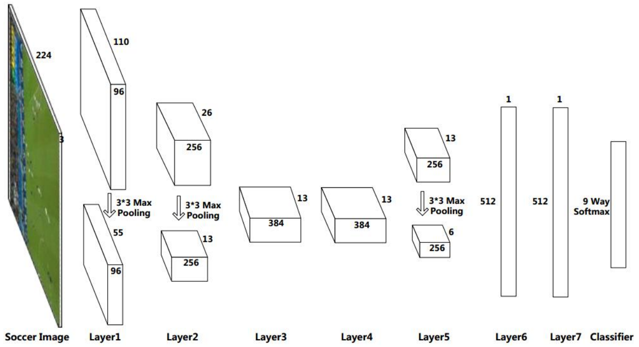


Figure 3. The structure of CNN for semantic analysic

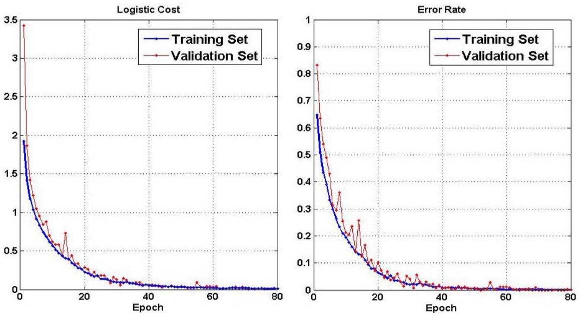


Figure 4. The training and validating process of CNN model

The network is consisted of five convolution layers and two full-connected layers, and the output of the last full-connected layer is fed into a 9 way softmax classifier. The size of convolution filter we used respectively is 7*7*96, 5*5*256, 3*3*384, 3*3*384, 3*3*256 and the stride is 2, 2, 1, 1, 1. Because the number of categories of soccer image views is not as large as the ImageNet[4], so we choose the size of full-connnected layer is 512. Like [13], the Rectified Linear Units(ReLU) is used as the nonlinear activation function. In order to avoid over-fitting, the first two convolution layers and the last convolution layer are pooled by a 3*3 max filter, and we also use a dropout layer with the rate of 0.5 at the full-connected layer.

For training the soccer images, firstly we resize it to 256*256 pixels and then subtract the mean of the image as the pre-processing, and then randomly extract 224*224 patches as the input to train our network. In the training process, we set the learning rate is 0.001 and the momentum is 0.9. And in our experiments, we use 80% of the SSID as the training set, and the remaining as the validation set. The training and validating process is shown in Fig 4.

C. The Extraction of Sequence Features

For a soccer video, the way to get shots we used is the HSV color histogram difference between adjacent frames. That is, a frame is regarded as shot boundary when the adjacent difference is bigger than a threshold, and then we can get the shots based on those frames.

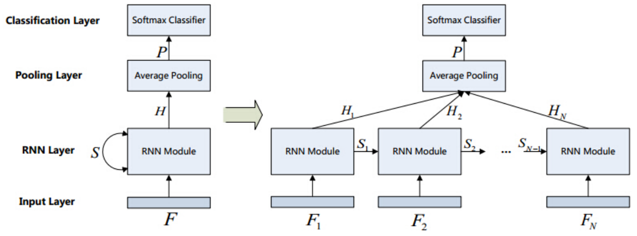


Figure 5. The structure of RNN for event detection (right is obtained from left order by time.)

A single shot can not describe the soccer event clearly, so we use PB as the soccer event boundary and use the method in [17] to determine the location of PB. For each PB segment, we extract N key frames as its description. We propose a new approach to avoid the lose of shot boundary information causing by the key frames averagely selected from the entire PB, such as the replay frame. If $P = (S_1, S_2, ..., S_t)$ indicates that the PB segment P is composed of t shots. Let $N_{S_1}, N_{S_2}, ..., N_{S_t}$ be the numbers of frames in shots. For shot S_i, we extract n_i key frames and n_i is calculated as:

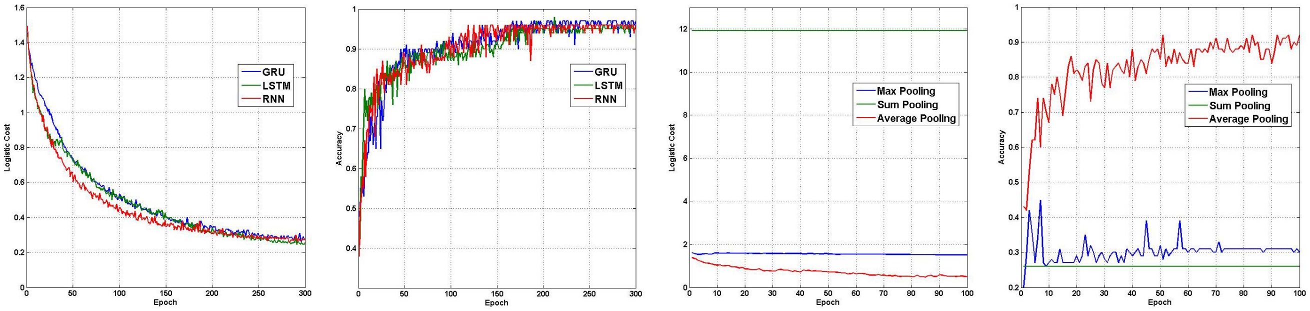
$$n_i = (N - 2 * t) * \frac{N_{S_i}}{\sum\limits_{j=1}^{t} N_{S_j}} + 2 \qquad (1 \leq i \leq t)$$

For a shot S_i, we choose the first frame, the last one and the frames at regular intervals from the rest as its key frames.

To extract the semantic feature, based on the trained CNN model, we remove the classification layer and add a normalized layer after the last full-connected layer. Denote a key frame sequence order by time by $K = (K_1, K_2, ..., K_N)$, and for each frame $K_j(1 < j < N)$, when we put it into the model, the corresponding output is the semantic features denoted by $F_j = (f_1, f_2, ..., f_{512}) \in R^{512}$. And the semantic sequence feature by $F = (F_1, F_2, ..., F_N)^T$.

III. Soccer Event Detection

Due to the effect of the time dimension, it is difficult to transform the features of continuous frame images into event features. Therefore, in semantic sequence feature, instead of directly removing the time dimension like [6], we use RNN to solve the temporal relation between frame sequences and event types. The details of RNN structure is realized by using the Keras toolbox, which is shown in Figure 5. Our RNN structure has four layers, including input layer, RNN layer, pooling layer and classification layer and the number of hidden units in each RNN model is 128. The output of RNN layer is a sequence, and to get a better result, we do extra fine tuning by a average pooling layer. We denote the output of RNN model at time i by $H_i = (h_{i1}, h_{i2}, ..., h_{in})^T \in R^n$, and the result of average



(a) (b) (c) (d)

Figure 6. The validating process result of different RNN structures (a)the logistic cost of different RNN models (b)the accuracy of different RNN models (c)the logistic cost of different pooling methods (d)the accuracy of different pooling methods

Table I
CONFUSION MATRIX OF OUR METHOD

Predicted Actual	Corner	Goal	Goal Attempt	Card	Missed
Corner	48	3	1	0	2
Goal	2	45	2	0	0
Goal Attempt	2	3	66	1	2
Card	0	0	1	13	2
Recall(%)	88.88	91.84	89.19	81.25	-
Precision(%)	94.11	88.23	94.29	92.86	-

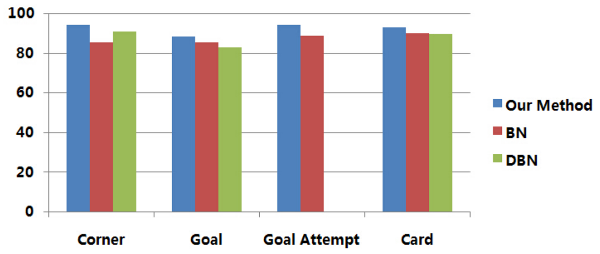


Figure 7. Comparison of our method with BN [17] and DBN [10]

pooling layer by $P = (p_1, p_2, ..., p_n)^T \in R^n$. P can be represented as:

$$p_i = \frac{1}{N} * \sum_{j=1}^{N} h_{ji}$$

At last, we use softmax function as the final loss function. And to avoid over fitting, we add a dropout layer with a rate of 0.5 after the pooling layer. And to optimize the model we use the Stochastic Gradient Descent(SGD) method in the training process, and the initial learning rate we set is 0.001.

IV. EXPERIMENTS

In our experiments, we detect four typical soccer events to verify the effectiveness of our approach in soccer event detection, including goal, goal attempt, corner and card. Due to the lack of open and high quality soccer video event dataset, we manually built and tagged our own dataset, including 30 soccer match videos gathered from FIFA world cup(WC), UEFA champion league, and England premier league(EPL). For 20 match videos of these, we cut into PB segments, labeled the shot boundaries and the event types for training the RNN structure. And the remaining 10 match videos is for testing our approach.

In our experiment, the number of key frames extracted from PB segment is 256. For the softmax classifier in RNN structure, it is a 4 way classifier and use one-hot encoding different events. In order to determine the optimal RNN structure, we use three different models in our experiment, including standard RNN model, LSTM[8], [9] and GRU[3]. The result of training process is shown in Fig 6 (a)-(b). From the Fig 6 (b) we can see that LSTM achieved the best accuracy at the beginning than the other two RNN structures for the ability to solve the problem of exploding gradients [1], and the accuracy remained stable after the 180th epoch. Therefore in the same sitution, LSTM can work better than the others. We also do experiments to test the effects of different pooling methods to do the fine tuning based on LSTM model, including max pooling, sum pooling and average pooling, the result of which is shown in Fig 6 (c)-(d). And from the result curves we can clearly see that the max pooling and the sum pooling do not have the ability to detect soccer event, but the average pooling can do well.

We used LSTM as the RNN model and the final result is shown in Table I. We compared our method with two state-of-art approaches BN[10] and DBN[17] on precision. The result in Fig 7 indicates the proposed approach outperform the two other approaches(the precision of DBN in detecting goal attempt event is null because of unable to do it). In the stage of extraction event features, a separate extraction algorithm is used for each feature in BN and DBN. In our experiment, we find some features are not robust like highPlayerInPB [17]. As time goes by, the state of event changes. The value of time dimension is ignored in the traditional methods like BN and DBN. In the method we proposed, time dimension is successfully introduced by RNN and achieved better results.

V. Conclusion and Future Work

By combining CNN and RNN, we construct a deep neural network for soccer video event detection and achieved better effect than traditional methods in this paper. The problem of the feature extraction algorithm is not robust can be solved by CNN and RNN use time relation to solve the task of soccer event detection. For future work, we are ready to use this method for the detection of common complex events.

References

[1] . Bengio, Y., . Simard, P., and . Frasconi, P., "Learning long-term dependencies with gradient descent is difficult." *IEEE Transactions on Neural Networks*, vol. 5, no. 5, pp. 157–66, 1994.

[2] Y. Bian, Y. Dong, H. Bai, B. Liu, K. Wang, and Y. Liu, "Reducing structure of deep convolutional neural networks for huawei accurate and fast mobile video annotation challenge," in *2014 IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*, 2014, pp. 1–6.

[3] K. Cho, B. Van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," *Eprint Arxiv*, 2014.

[4] J. Deng, W. Dong, R. Socher, L. J. Li, K. Li, and F. F. Li, "Imagenet: A large-scale hierarchical image database." in *Proceedings / CVPR, IEEE Computer Society Conference on Computer Vision and Pattern Recognition. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255.

[5] A. Ekin, A. M. Tekalp, and R. Mehrotra, "Automatic soccer video analysis and summarization," *Image Processing, IEEE Transactions on*, vol. 12, no. 7, pp. 796–807, 2003.

[6] C. Gan, N. Wang, Y. Yang, D. Yeung, and A. G. Hauptmann, "Devnet: A deep event network for multimedia event detection and evidence recounting," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, 2015, pp. 2568–2577.

[7] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, 2014, pp. 580–587.

[8] A. Graves, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[9] ——, "Supervised sequence labelling with recurrent neural networks," *Studies in Computational Intelligence*, vol. 385, 2012.

[10] C.-L. Huang, H.-C. Shih, and C.-Y. Chao, "Semantic analysis of soccer video using dynamic bayesian network," *Multimedia, IEEE Transactions on*, vol. 8, no. 4, pp. 749–760, 2006.

[11] G. Jin, L. Tao, and G. Xu, *Hidden Markov Model Based Events Detection in Soccer Video*. Springer Berlin Heidelberg, 2004.

[12] R. Jozefowicz, W. Zaremba, and I. Sutskever, "An empirical exploration of recurrent network architectures," in *Proceedings of the 32nd International Conference on Machine Learning (ICML-15)*, D. Blei and F. Bach, Eds. JMLR Workshop and Conference Proceedings, 2015, pp. 2342–2350.

[13] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in Neural Information Processing Systems*, vol. 25, p. 2012, 2012.

[14] X. Qian, G. Liu, H. Wang, Z. Li, and Z. Wang, "Soccer video event detection by fusing middle level visual semantics of an event clip," *Lecture Notes in Computer Science*, vol. 6298, pp. 439–451, 2010.

[15] X. Qian, G. Liu, Z. Wang, Z. Li, and H. Wang, "Highlight events detection in soccer video using hcrf," in *Proceedings of the Second International Conference on Internet Multimedia Computing and Service*. ACM, 2010, pp. 171–174.

[16] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," *Advances in Neural Information Processing Systems*, vol. 4, pp. 3104–3112, 2014.

[17] M. Tavassolipour, M. Karimian, and S. Kasaei, "Event detection and summarization in soccer videos using bayesian network and copula." *IEEE Trans. Circuits Syst. Video Techn.*, vol. 24, no. 2, pp. 291–304, 2014.

[18] D. Tjondronegoro, Y. P. P. Chen, and B. Pham, "The power of play-break for automatic detection and browsing of self-consumable sport video highlights," *Mir 04 Proceedings of Acm Sigmm International Workshop on Multimedia Information Retrieval*, pp. 267 – 274, 2004.

[19] .W. and Y. P. P. Chen, "Knowledge-discounted event detection in sports video," *IEEE Transactions on Systems Man & Cybernetics Part A Systems & Humans*, vol. 40, no. 5, pp. 1009 – 1024, 2010.

[20] Z. Xu, Y. Yang, and A. G. Hauptmann, "A discriminative cnn video representation for event detection," *Eprint Arxiv*, 2014.