RGB Video Based Tennis Action Recognition Using a Deep Weighted Long Short-Term Memory

Jiaxin Cai, Xin Tang

***Abstract*—Action recognition has attracted increasing attention from RGB input in computer vision partially due to potential applications on somatic simulation and statistics of sport such as virtual tennis game and tennis techniques and tactics analysis by video. Recently, deep learning based methods have achieved promising performance for action recognition. In this paper, we propose weighted Long Short-Term Memory adopted with convolutional neural network representations for three dimensional tennis shots recognition. First, the local two-dimensional convolutional neural network spatial representations are extracted from each video frame individually using a pre-trained Inception network. Then, a weighted Long Short-Term Memory decoder is introduced to take the output state at time t and the historical embedding feature at time t-1 to generate feature vector using a score weighting scheme. Finally, we use the adopted CNN and weighted LSTM to map the original visual features into a vector space to generate the spatial-temporal semantical description of visual sequences and classify the action video content. Experiments on the benchmark demonstrate that our method using only simple raw RGB video can achieve better performance than the state-of-the-art baselines for tennis shot recognition.**

***Index Terms*—tennis game, action recognition, deep learning, Long Short-Term Memory, convolutional neural networks.**

I. Introduction

Action recognition technology are very useful in sports game analysis. Due to the great increasing of sport game videos, it is implementable to analysis the technology and movement of players using the collected video data. Many works have be devoted in vision-based sport action recognition in recent years. Several public video datasets such as UCF-Sport and Sports-1M are provided for this problem. This paper focuses on automatically recognizing the player shot of tennis game by machine learning technology using the original RGB videos. The research works in the area of tennis action recognition have been devoted . Zhu et al. [1], [2] used the support vector machine to classify tennis videos into left-swing and right-swing actions by optical flow based descriptor. Farajidavar et al. [3] employed the transfer learning to classify tennis videos into non-hit, hit and serve actions. Gourgari et al.[4] presented a tennis actions database called THETIS dataset. It consists of 12 fine-grained tennis shot acted by 55 different subjects multiple times at different scenes with dynamic background. The objective of analyzing the THETIS dataset is to classify the videos into the 12 pre-defined tennis actions from raw video data. Mora et al. [5] proposed a deep learning model for domain-specific tennis action recognition using RGB video content on the THETIS dataset.

J. Cai is with the School of Applied Mathematics, Xiamen University of Technology, Xiamen, P.R. China. e-mail: (caijiaxin@xmut.edu.cn).

Xin Tang is with College of Sciences, Huazhong Agricultural University, Wuhan, P.R. China. e-mail:(tangxin2@mail2.sysu.edu.cn). Corresponding author.

Recently, deep learning based methods have achieved promising performance for action recognition. Li et al. [6] presented a skeleton-based action recognition method using LSTM and CNN. Cheron et al. [7] proposed a pose-based CNN Features for action recognition. Gammulle et al. [8] presents a two-stream LSTM framework to fusing the deep networks human action recognition. Zhang et al. [9] use the multi-layer LSTM networks to learn the geometric features on skeleton information for action recognition. Liu et al. [10] proposed a spatio-temporal LSTM with trust gates for 3D Human Action Recognition. Zhu et al. [11] presented a co-occurrence skeleton feature learning based on regularized deep LSTM networks for human action recognition. Lee et al. [12] proposed an ensemble learned temporal sliding LSTM networks for skeleton-based action recognition Tsunoda et al. [13] used the hierarchical LSTM model for football action recognition. Song et al. [14] presented a spatio-temporal attention-based LSTM networks for recognizing and detecting 3D action Recognition and Detection. Liu [15] proposed a global context-aware attention LSTM networks for 3D action recognition.

In this paper, we propose a framework using convolutional neural networks with weighted Long Short-Term Memory networks for tennis action recognition. First, the local two-dimensional convolutional neural network spatial representations are extracted from each video frame individually using a pre-trained Inception network. Then, a weighted Long Short-Term Memory model is proposed to take the output state at time t and the historical embedding feature at time t-1 to generate feature vector using a score weighting scheme. Finally, the CNN model and the weighted LSTM model are adopted to map the raw RGB video into a vector space to generate the spatial-temporal semantical description and classify the action video content. Experiments on the benchmark demonstrate that our method outperforms the state-of-the-art baselines for tennis shot recognition using only raw RGB video.

The rest of the paper is organized as follows. Section 2 describes the proposed method. Section 3 shows the experimental results and anylysis. Finally, section gives the conclusion and future work of the paper.

II. Method

A. Framework

In this paper, we adopt a improved Long Short-Term Memory networks accompanied by convolutional neural network for the tennis action recognition. First, the original RGB frame images of all videos are fed to the 2D convolutional neural networks for capturing the appearance and spatial information as the local feature. In this paper, we adopt the GoogLeNet Inception V3 as CNN model. The model is pre-trained on the Large Scale Visual Recognition Challenge 2012 (ILSVRC-2012) ImageNet. Then a Long Short-Term Memory decoder model is adopted as a complement to the CNN model to capture the contextual information in the temporal domain as the global feature. The output scores of the last CNN layer is fed to the LSTM model as the input of the first LSTM layer. The CNN features are suitable to feed into the LSTM model because of providing rich spatial information. Specially, we introduce an improved version of the Long Short-Term Memory decoder, called weighted Long Short-Term Memory, to describe the historical information . It employs a score weighting scheme to generate the iterated feature vector using the output state at time t and the historical embedding feature at time t-1. Finally, we use the output of weighted LSTM as the spatial-temporal semantical description of visual sequences. It is fed to a softmax layer for classifying the action video content.

B. RNN and LSTM

RNN is a powerful sequence model widely used in speech recognition and natural language processing. For a typical RNN model, the update equation at time t is described as follows.

$$h_t = g(b + Wh_{t-1} + Ux_t) \tag{1}$$

$$\hat{y}_t = softmax(c + Vh_t) \tag{2}$$

where h_t is the state of the tth neuron, and x_t is the observation at time t which is the input of the t-th neuron. In a typical RNN model, the state response h_t of the t-th neuron is determined by the previous neuron state h_{t-1} and the input x_t of the t-th neuron. $g(\cdot)$ is the activation function. Usually, the sigmoid function $\sigma(\cdot)$ or hyperbolic tangent function $tanh(\cdot)$ can be employed as the activation function. $\hat{y}_t$ denotes the prediction of label y at time t. It is estimated by the output response h_t of state neuron at t time using a softmax function. U, W and V are the weight matrices. b and c are the bias vectors.

Traditional RNNs is suitable for modeling the short-term dynamics but unable to capture the long-terms relations. LSTM is an improved version of RNN architecture for learning long-range dependencies and resolving the "vanishing gradient" problem. In the typical LSTM model, the t-th hidden neuron is the cell unit containing an input gate i_t, a forget gate f_t, output gate o_t, an internal memory cell state c_t and an output response h_t. The transition equations of LSTM at time t can be written as follows.

$$i_t = \sigma(U^i x_t + W^i h_{t-1} + P^i c_{t-1} + b^i) \tag{3}$$

$$f_t = \sigma(U^f x_t + W^f h_{t-1} + P^f c_{t-1} + b^f) \tag{4}$$

$$c_t = f_t \odot c_{t-1} + i_t \odot tanh(U^c x_t + W^c h_{t-1} + b^c) \tag{5}$$

$$o_t = \sigma(U^o x_t + W^o h_{t-1} + P^o c_t + b^o) \tag{6}$$

$$h_t = o_t \odot tanh(c_t) \tag{7}$$

where U^i, U^f, U^c, U^o, W^i, W^f, W^c, W^o, P^i, P^f and P^o are the weight matrices. b^i, b^f, b^c and b^o are the bias vectors. Operator $\odot$ indicates element-wise product.

C. Weighted LSTM

The typical LSTM use the last layer to encode the input sequence. When it is adopted for action video recognition using image sequences, the last frame of the video will have the most influence on the action class recognition result. And the previous frames have a minimal impact on the classification due to the forgetting effect. However, a human action is often defined by the whole movement process. To better describe the human action, we embed the historical information of human movement to the LSTM model for building a global feature. We propose an improved version of the Long Short-Term Memory decoder, called weighted Long Short-Term Memory, to describe the historical information. In the weighted LSTM model, a historical state l_t is introduced at time t. It is generated by a score weighting scheme using the output response h_t at time t and the historical state at time t-1. The update equation of historical state l_t is expressed as follows.

$$l_t = \begin{cases} \alpha_t h_t + (1-\alpha_t) l_{t-1}, & if \quad \epsilon_{h_t} \geq \epsilon_{h_{t-1}} \\ \sum_{k=1}^{t} \omega_k^t h_k, & if \quad \epsilon_{h_t} < \epsilon_{h_{t-1}} \end{cases} \tag{8}$$

where α_t is the weight controlling a balance between the response h_t and the last historical state l_{t-1}. It is calculated by the following formula:

$$\alpha_t = \frac{1}{2} \ln(\frac{\epsilon_{h_{t-1}}}{\epsilon_{h_t}}) \tag{9}$$

where ϵ_{h_t} denote the loss between the training label y and the estimated label $\hat{y}_t$ at time t using the softmax function on $c + Vh_t$.

ω_k^t denotes the weight of response h_k. It is calculated by:

$$\omega_k^t = \begin{cases} 0, & if \quad k \leq \tau \\ \frac{1}{t-\tau}, & if \quad k > \tau \end{cases} \tag{10}$$

where τ is the parameter controlling the forgetting effect.

Finally, a softmax layer is employed to provide the estimated label $\hat{y}$ of action video is determined by the last historical state l_T:

$$\hat{y} = softmax(d + Ql_T) \tag{11}$$

where d and Q is the bias vector and weight matrix of the softmax layer respectively.

D. Implementation Details

TensorFlow and Python are employed as the deep learning platform. An NVIDIA GTX1080Ti GPU is adopted to run the experiments. We use a five-layer deep LSTM in the stacked network. Each layer of LSTM has 30 hidden neuron. The probability of dropout is set as 0.5. The initial learning rate is set as 0.001 on LSTM model. The learning rate is set by exponentially decaying with a base of 0.96 every 100 000 steps during the training process. The regularization value of the LSTM model is set as 0.004. The batch size fed to the model is set as 32. The Adam Optimizer is used to trained the network for LSTM model.

III. Experimental results

A. General action recognition

1) Dataset: We use the HMDB51 dataset [16] to show the effectiveness of the proposed weighted LSTM for general action recognition tasks. The HMDB51 dataset is a common used dataset including 6849 videos. It consists of 51 human action categories including facial actions and body movement. The original RGB videos of HMDB51 dataset are used to test the proposed framework.

2) Results: The five-fold cross-validation strategy is used to split the dataset to a training set and a test set. We compared the proposed WLSTM to the typical LSTM. The parameter τ of WLSTM is fine tuned from 2 to 5. Table 1 shows the accuracy values of the methods. As is shown in the table, the average accuracy in prediction of the WLSTM model achieve a best accuracy 0.73 when the parameter τ is set as 2. Furthermore, the average accuracy of the WLSTM model when the parameter τ is set as 3 is 0.62. When the parameter τ is set as 4, the WLSTM model achieve an accuracy of 0.63. However, when τ is set as 5, the accuracy of WLSTM decline to 0.62. It is seen that setting the parameter τ to a medium number according to the time period in the WTSM model is better. Meanwhile, the average accuracy of the LSTM model is 0.47. The experimental results shows that the proposed WLSTM model outperforms the LSTM model and has higher recognition rate. We also compared the proposed model to other state-of-the-art models using RGB data, such as Spatial stream ConvNet [17], Soft attention model [18], Composite LSTM [19] and Mora's model [5]. As shown in Table 1, the correct recognition rate of our method exceeds other existing models using RGB data. We also compare our model using RGB data to other methods using different information. They includes: the two-stream ConvNet using optical flow information [17], Wang's model using Fisher Vectors [20], Wang's model using a combination of HOG, HOF and MBH [20]. The experimental results show that our model achieves a high accuracy using the very simple RGB data. It will results in better results if other information such as optical flow data, skeletons data and depth information are added to the input of our proposed model.

TABLE I
ACCURACY COMPARISON OF METHODS ON HMDB51 DATASET

Method	Accuracy
WLSTM ($\tau = 2$)	0.73
WLSTM ($\tau = 3$)	0.62
WLSTM ($\tau = 4$)	0.63
WLSTM ($\tau = 5$)	0.62
LSTM	0.47
Spatial stream ConvNet [17]	0.41
Soft attention model [18]	0.41
Composite LSTM [19]	0.44
Mora et al. [5]	0.43
Two-stream ConvNet using optical flow [17]	0.59
Wang et al. (using Fisher Vectors) [20]	0.53
Wang et al. (using a combination of HOG, HOF and MBH) [20]	0.60

B. Tennis action recognition

1) Dataset: The Three Dimensional Tennis Shots (THETIS) dataset [4] is used for evaluating the proposed method. This dataset contains 12 basic tennis actions each of which are acted repeatedly by 31 amateurs and 24 experienced players. The 12 tennis actions performed by actors are: Backhand with two hands, Backhand, Backhand slice, Backhand volley, Forehand flat, Forehand open stands, Forehand slice, Forehand volley, Service flat, Service kick, Service slice and Smash. Each actor repeats each tennis action 3 to 4 times. There are totally 8734 videos of the AVI format with a total duration of 7 hours and 15 minutes. Specially, a set of 1980 RGB viodeos of the AVI format are provided. There are two different indoor backgrounds. The backgrounds contain different scenes in which multiple persons pass or play basketball. The length of video sequences also varies. Although the depth, skeleton and silhouettes videos are also given, we only use the RGB videos to perform the tennis action recognition experiments.

2) Results: The five-fold cross-validation strategy is used to split the dataset to a training set and a test set. We compared the proposed WLSTM to the typical LSTM. The parameter τ of WLSTM is fine tuned from 2 to 5. Table 2 shows the accuracy values of the methods. As is shown in the table, the average accuracy in prediction of the WLSTM model achieve a best accuracy 0.74 when the parameter τ is set as 3. Furthermore, the average accuracy of the WLSTM model when the parameter τ is set as 2 is 0.70. When the parameter τ is set as 4, the WLSTM model achieve an accuracy of 0.71. However, when τ is set as 5, the accuracy of WLSTM decline to 0.63. It is seen that setting the parameter τ to a medium number according to the time period in the WTSM model is better. Meanwhile, the average accuracy of the LSTM model is 0.56. The experimental results shows that the proposed WLSTM model outperforms the LSTM model in the accuracy of perdition.

In [5], the authors perform experiments using the RGB videos on the THETIS dataset and get an average accuracy of 0.47 for tennis action recognition using a leave-one-out

TABLE II
ACCURACY COMPARISON OF METHODS ON THETIS DATASET

Method	Accuracy
WLSTM ($\tau = 2$)	0.70
WLSTM ($\tau = 3$)	0.74
WLSTM ($\tau = 4$)	0.71
WLSTM ($\tau = 5$)	0.63
LSTM	0.56
Mora et al. [5]	0.47
Gourgari et al. (using depth videos)[4]	0.6
Gourgari et al. (using 3D skeletons)[4]	0.54

strategy. Compared to their experiments, we employ less training samples and obtain a better performance under a more difficult condition. In [4] , the authors perform tennis action classification experiments using depth videos and 3D skeletons separately on the THETIS dataset and obtain 0.60 and 0.54 accuracy respectively. Compared to their results, our experiment use the raw images alone, which is a more challenge task. Experimental results prove that our proposed method is competitive.

IV. CONCLUSION

This paper proposes a weighted Long Short-Term Memory for modeling the tennis action image sequence. A score weighting strategy is introduced to the Long Short-Term Memory networks to model the historical information. The proposed weighted LSTM is used for tennis action recaption on the CNNfeatures of local frame images of videos. Experimental results on the tennis action dataset demonstrate that our method is effective. Our future works include the improvement of the weighting strategy and more experiments on real tennis game videos.

ACKNOWLEDGMENT

The authors would like to thank to the Xiamen University of Technology High Level Talents Project (No.YKJ15018R) and the National Natural Science Foundation of P. R. China (No.61602148). The authors also gratefully acknowledge the helpful comments and suggestions of the reviewers, which have improved the presentation.

REFERENCES

[1] Zhu, Guangyu, et al. "Action recognition in broadcast tennis video using optical flow and support vector machine." International Conference on Pattern Recognition IEEE, 2006:251-254.
[2] Zhu, Guangyu, et al. "Player action recognition in broadcast tennis video with applications to semantic analysis of sports game." ACM International Conference on Multimedia ACM, 2006:431-440.
[3] Farajidavar, Nazli, et al. "Transductive transfer learning for action recognition in tennis games." IEEE International Conference on Computer Vision Workshops IEEE, 2011:1548-1553.
[4] Gourgari, Sofia, et al. "THETIS: Three dimensional tennis shots a human action dataset." 13.4(2013):676-681.
[5] Mora, Silvia Vinyes, and W. J. Knottenbelt. "Deep Learning for Domain-Specific Action Recognition in Tennis." Computer Vision and Pattern Recognition Workshops IEEE, 2017:170-178.
[6] Li N C, Wang N P, Wang N S, et al. "Skeleton-based action recognition using LSTM and CNN." IEEE International Conference on Multimedia and Expo Workshops IEEE Computer Society, 2017:585-590.
[7] Cheron, Guilhem, I. Laptev, and C. Schmid. "P-CNN: Pose-Based CNN Features for Action Recognition." (2015):3218-3226.
[8] Gammulle, Harshala, et al. "Two Stream LSTM: A Deep Fusion Framework for Human Action Recognition." arXiv. (2017).
[9] Zhang, Songyang, X. Liu, and J. Xiao. "On Geometric Features for Skeleton-Based Action Recognition Using Multilayer LSTM Networks." Applications of Computer Vision IEEE, 2017.
[10] Liu, Jun, et al. "Skeleton-Based Action Recognition Using Spatio-Temporal LSTM Network with Trust Gates." IEEE Transactions on Pattern Analysis and Machine Intelligence PP.99(2017):1-1.
[11] Zhu, Wentao, Lan C, Xing J, et al. "Co-occurrence feature learning for skeleton based action recognition using regularized deep LSTM networks." Thirtieth AAAI Conference on Artificial Intelligence AAAI Press, 2016:3697-3703.
[12] Lee, Inwoong, et al. "Ensemble Deep Learning for Skeleton-Based Action Recognition Using Temporal Sliding LSTM Networks." IEEE International Conference on Computer Vision IEEE Computer Society, 2017:1012-1020.
[13] Tsunoda, Takamasa, et al. "Football Action Recognition Using Hierarchical LSTM." IEEE Conference on Computer Vision and Pattern Recognition Workshops IEEE, 2017:155-163.
[14] Song, S., et al. "Spatio-Temporal Attention-Based LSTM Networks for 3D Action Recognition and Detection. " IEEE Transactions on Image Processing PP.99(2018):1-1.
[15] Liu, Jun, et al. "Global Context-Aware Attention LSTM Networks for 3D Action Recognition." IEEE Conference on Computer Vision and Pattern Recognition IEEE Computer Society, 2017:3671-3680.
[16] Kuehne, H., et al. "HMDB: A large video database for human motion recognition." IEEE International Conference on Computer Vision, ICCV 2011, Barcelona, Spain, November DBLP, 2012:2556-2563.
[17] Simonyan, Karen, and A. Zisserman. "Two-stream convolutional networks for action recognition in videos." International Conference on Neural Information Processing Systems MIT Press, 2014:568-576.
[18] Sharma, Shikhar, R. Kiros, and R. Salakhutdinov. "Action Recognition using Visual Attention." Computer Science (2016).
[19] Soomro, Khurram, and A. R. Zamir. Action Recognition in Realistic Sports Videos. Computer Vision in Sports. Springer International Publishing, 2014:408-411.
[20] Wang, Heng, et al. "A Robust and Efficient Video Representation for Action Recognition." International Journal of Computer Vision 119.3(2016):219-238.