ISSN 1330-3651 (Print), ISSN 1848-6339 (Online) https://doi.org/10.17559/TV-20190506101459
Original scientific paper

ATSN: Attention-Based Temporal Segment Network for Action Recognition

Yun-lei SUN, Da-lin ZHANG

Abstract: In human action recognition, a reasonable video representation is still a problem to be solved. For humans, it is easy to focus on the prominent areas of the image in the video, focusing on the part of interest. Inspired by this, we proposed a deep Temporal Segment Network based on visual attention-ATSN. By lightly modifying the model structure, ATSN integrates the human attention mechanism into the Temporal Segment Networks, can effectively add a weight to the video representation features, pays attention to the beneficial regions in the features, and achieves more accurate action recognition. We conducted the Oilfield-7 dataset for human actions on the oilfield. The experimental results on HMDB51 and Oilfield-7 show that the ATSN had achieved excellent performance.

Keywords: action recognition; attention; Temporal Segment Network

1 INTRODUCTION

Human action recognition, a basic problem in computer vision, has attracted widespread attention in the industry. Action recognition in video is a key issue in high-level video understanding tasks. Although deep convolutional neural networks (CNNs) have achieved great success in image recognition tasks with their powerful performance [1-4], no progress has been made in action recognition tasks similar to image recognition. To some extent, action recognition in video has similar problems to object recognition in static images. Both tasks must deal with significant intra-class difference, clutter and occlusion of background [5]. However, the obvious difference is that video has an additional and more important time cue than image, which can help to obtain motion information and reliably identify multiple actions by motion information. Recent video representations for motion recognition are mainly based on two different CNN architectures: (1) 3D spatiotemporal convolution [6-7]; (2) two-stream architecture [8]. Although both have achieved good performance in action recognition, two-stream structure usually outperforms 3D spatiotemporal convolution by making use of the new ultra-deep architecture and the pre-training model of static image classification [9].

However, the main challenge in action recognition is still the lack of video representation. For humans, it is easy to focus on the prominent areas of the image in the video, then follow the parts we are interested in. However, the existing method of action recognition aggregate all the local features of each short fragment segmented from video to form global features averagely, and classify the action categories according to the global features of each fragment. Average aggregation is not an appropriate way. For each image frame in a short clip, it is not every pixel or area that can provide beneficial features. Some areas we should focus on (e.g. human motion, man-machine interaction) and some other areas (e.g. background, occlusion) should be consciously ignored.

Inspired by the above, we use the attention mechanism to highlight the salient areas in video. In this paper, a deep Temporal Segment Network [10] based on visual attention (ATSN) is proposed, which integrates the attention mechanism into the two-stream Temporal Segment Networks. The characteristics of attention mechanism enable us to locate the region of each image frame without supervision, assign weight to each region space, then aggregate the local spatial features according to the weighted sum. This extraordinary joint method is simple and effective, and can easily solve the problem of video presentation not outstanding. In order to verify this statement, we conducted a series of video-based action recognition experiments on the oilfield action data set. The results show that our ATSN model is effective.

The rest of this paper is as follows: We discuss the related work in Section 2, describe our visual attention-based deep Temporal Segment Network model in Section 3, give the experimental details in Section 4, and summarize the work in Section 5.

2 RELATED WORKS

Action recognition is a hot topic in visual applications, and its research progress is largely driven by the progress of image recognition methods. The purpose of action recognition is to identify single or multiple actions in each video, which is usually described as a simple classification problem. Before CNNs had achieved such great success, Ivan Laptev et al. proposed a method to extend the spatial pyramid to the spatiotemporal domain by using spatial-temporal features, to detect sparse spatiotemporal points of interest and to describe them by using local spatiotemporal features (including HOG and HOF), to encode them into bag of feature (BoF) and to classify actions with SVM [11]. In the following work, Heng Wang et al. expanded four feature descriptors (HOG3D, HOG/HOF, Cuboids, ESURF) to describe local spatiotemporal features [12]. Experiments show that the dense sampling of performing local features better than sparse interest of point detection. Afterwards, Heng Wang et al. proposed a dense trajectory algorithm for action recognition. By sampling dense points from image and tracking them according to the displacement information of dense optical flow field, the dense trajectory can cover the motion information in video [13]. The improved dense trajectory algorithm [14] achieves more prominent performance by eliminating background trajectory and distorted optical flow.

With the rise of deep learning, convolutional neural networks with powerful performance have been widely used in the field of action recognition. Andrej Karpathy et al. used deep convolution neural network to train the

Sports-1M data set and made an empirical evaluation of large-scale action classification [15]. Karen Simonyan and Andrew Zisserman proposed a dual-stream architecture. They input a video to obtain video image information and dense optical flow information respectively, train a CNNS for each stream to judge the action category, and finally fuse the action classification scores of the two branches. Christoph Feichtenhofer et al. improved the fusion method on the basis of the two-stream architecture. They advanced the original integration of the Softmax layer to the convolution layer, and further improved the performance of the spatial and temporal network in the last convolution layer [16]. Joe Yue-Hei Ng and others also studied the fusion method of dual-stream architecture [17]. They improved the temporal network by utilizing LSTM's powerful memory function for temporal information. According to the characteristics of video, Limin Wang et al. proposed the Temporal Segment Network (TSN), a novel framework for video-based action recognition. which is based on the idea of long-range temporal structure modelling. It combines a sparse temporal sampling strategy and video-level supervision to enable efficient and effective learning using the whole action video 10]. Du Tran et al.'s 3D convolution neural network (C3D) is another mainstream method for video-level action classification parallel to dual-stream architecture. Because 2D convolution cannot capture temporal information well in video, the proposed 3D convolution extends the original convolution layer and pooling layer to 3D convolution and 3D pooling. Video can be processed directly through 3D convolution. J. Carreira et al. proposed a new model combining 3D convolution and dual-stream network, called I3D, which can learn seamless spatial-temporal feature extractors from video. Moreover, I3D model is a general learning method for video representation [18].

Human visual attention has always been a concern in the field of computer vision. Xiaodi Hou et al. proposed an attention model based on continuous feature sampling to explain the concept of saliency feature using energy. In addition, the model can achieve attention selectivity between static and dynamic scenes [19]. Stefan Mathe and others have conducted a series of experiments on motion recognition in video, mainly on the relationship between human visual attention and computer vision [20]. Compared with the above work, our proposed deep TSN based on visual attention integrates human attention mechanism into the existing deep CNNs. Through lightweight modification of the model structure, the processed video representation features have local saliency.

3 DEEP TEMPORAL SEGMENT NETWORK BASED ON VISUAL ATTENTION

In this section, we will introduce our ATSNmodel. By lightly modifying the model structure, ATSN integrates the human attention mechanism into the TSN, so that the processed video representation features have local significance.

A recent standard approach to action recognition in video is to use the fusion of multiple information streams (RGB and optical flow), which can achieve significant performance [8, 15, 21]. Before that, we will introduce the infrastructure of our network, Temporal Segment Networks [10]. Finally, we describe the attention mechanism.

3.1 Infrastructure of the ATSN

The TSN is essentially a Two-Stream mode. The network part is composed of two-way CNN, which are spatial stream convnet and temporal stream convnet. Both networks use BN-Inception (the evolution of the GoogleNet network). Before the TSN was proposed, the two-streamConvolutional network could not model the video data of long-distance time structure. They could only deal with a single frame in the spatial network or a single stack in the temporal network, and could not effectively obtain the connection between the upper and lower contents in the temporal sequence. Temporal Segment Network obtains a series of short fragments from the whole video through a sparse sampling method, which can integrate the visual information of the whole video to classify the video level. Each fragment will give its own preliminary prediction of the action category, and from the "consensus" of these fragments get the prediction result of video level [10].

Specifically, given a video$V$, it is divided into $K\{S_1, S_2, \cdots, S_k\}$ segments at equal intervals. Then, the Temporal Segment Network models a series of short segments in the following way.

$$TSN(T_1, T_2, \cdots, T_k) = H(G(F(T_1; W), F(T_2; W), \cdots, F(T_k; W))) \quad (1)$$

where $(T_1, T_2, \cdots, T_k)$ denotes a sequence of segments, each segment T_k is randomly sampled from its corresponding segment S_k, $F(T_k; W)$ function means convolutional network with W as parameter acting on short segment T_k, $G()$ represents the segment consensus function, which combines the category score output of several short segments to obtain a consensus on the category judgment between them. Function $H()$ predicts the probability that the whole video belongs to each action category based on this consensus. In addition, the form of loss function Gabout consensus is:

$$L(y, G) = -\sum_{i=1}^{C} y_i \, (G_i - log \sum_{j=1}^{C} exp \, G_j) \quad (2)$$

where C is the number of action categories and y_i is the real label of class i.

3.2 ATSN Model Architecture

The ATSN assigns larger weights to the features extracted from the spatial network and the temporal network in the two-stream, which makes it easy to locate the regions of interest and thus can classify them more accurately. The structure is shown in Fig. 1. Our ATSN is modified on the basis of TSN. We connect the attention model to the feature extracted from the last convolutional layer of the spatial network and the temporal network respectively. Then we send the weighted feature to the full-connected layer and softmax to predict the class probability of the two-stream network. We also combine the results of

spatial stream and temporal stream before the final video category is judged.

Given a complete video$V$, we process it into a series of segments $S_i(i = 1,2,\dots k)$, k is the equal number of the entire video, each segment contains one RGB image and two optical flow graphs. Convolutional neural network (CNNs) extracts the global visual features $F_{RGB} = (F_1, F_2, F_3, \dots F_L)$ of RGB and the global visual features $F_{OF} = (F_1, F_2, F_3, \dots F_L)$ of optical flow graphs. L indicates that each image is divided into L blocks, and each region is a vector of m dimensions. Features F_{attRGB} and F_{attOF} are obtained after the attention mechanism processing, and then the category scores $C_{Si}$and C_{Ti} of each fragment S_i in the dual-stream network are obtained. After the consensus function $G()$, the two-stream results are sent to the *Softmax* function to calculate the probability, and then a complete video classification result W is obtained.

The workflow can be summarized as the following consensus:

$$F_{attRGB} = f(F_{RGB}) \quad (3)$$
$$F_{attOF} = f(F_{OF}) \quad (4)$$
$$g_S = G\left(\sum_{i=1}^{k} C_{Si}\right) \quad (5)$$
$$g_T = G\left(\sum_{i=1}^{k} C_{Ti}\right) \quad (6)$$
$$W = Softmax(g_S, g_T) \quad (7)$$

Formulas (3) and (4) are the attention features F_{RGB} and F_{OF} obtained by distributing the regional spatial weights of the features F_{attRGB} and F_{attOF} with the attention model respectively. Formulas (5) and (6) use consensus functions to sum the scores of all segments belonging to the same category in spatial flow and temporal flow to get g_S and $g_T$respectively. Formula (7) is the classification result W of the whole video obtained by integrating the scores of two-stream networks.

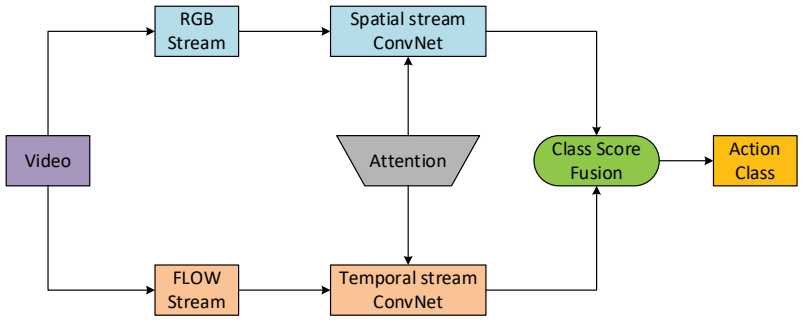

Figure 1 ATSN model structure. Based on the dual-flow model architecture, it can be divided into spatial stream network and temporal stream network, and the attention mechanism is added to the two-stream network. The classification score of the two-stream network output is fused, and the whole video action classification is finally obtained.

3.3 Attention Model

The ATSN's attention model attaches a weight between 0 and 1 to the eigenvector output from the last convolutional layer to focus on the salient area of the image. The structure of the model is shown in Fig. 2. The spatial stream feature F_{RGB}^t and the temporal stream feature F_{OF}^t extracted by convolutional neural network are vectors of $L \times m$ dimension, i.e. there are L regions in the image, each region is represented by m dimension feature vectors:

$$F_{RGB/OF}^t = \{F_1^t, F_2^t, F_3^t, \dots F_L^t\}, F_i \in R^m, t = (1,2,\dots k) \quad (8)$$

Among them, R^m represents the feature representation of m dimension, F^trepresents the i^{th} image region, and F^t is the feature representation of the video segment centering by time t. For each image area, the attention function O_{att}generates the attention weight α_i^t of the corresponding video sampling segment t according to the eigenvectors F_{RGB}^t and F_{OF}^t:

$$\alpha_i^t = O_{att}\left(F_{RGB/OF}^t\right) \quad (9)$$

Normalization:

$$\alpha_n^t = \frac{exp(\alpha_i^t)}{\sum_{n=1}^{L} exp(\alpha_n^t)} \quad (10)$$

Among them, α_n^t represents the weight of the n th image area in the attention model in the video segment t.

Characteristics $F_{attRGB/OF}$ after attention model processing:

$$F_{attRGB/OF} = \sum_{n=1}^{L} \alpha_n^t F_{RGB/OF} \quad (11)$$

The ATSN then feeds $F_{attRGB/OF}$ into the full-connected layer. Networks that incorporate attention mechanisms can still optimize learning through standard back-propagation.

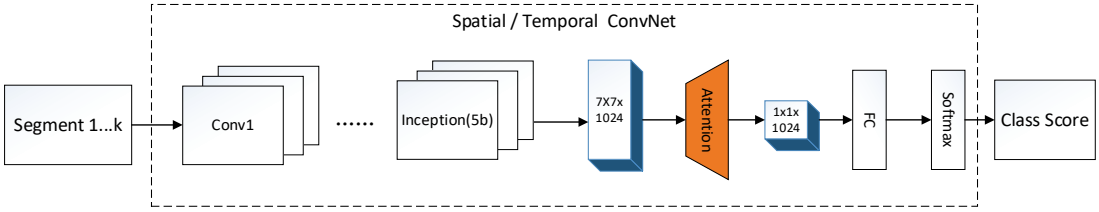

Figure 2 Our ATSN network architecture. The video clips are input into the network, and the spatial stream and temporal stream are convoluted separately. The virtual frames in the graph represent the same Attention processing for the spatial stream and temporal stream. The output score is the separate score of the two-stream network, which provides the basis for the subsequent fusion.

4 EXPERIMENTS

In this section, we first introduce our own oilfield video data set, and then evaluate the performance of ATSN on this data set. In addition, we also carry out validity experiments on the Temporal Segment Network which has achieved good performance in the action recognition task, that is, the improved basic network of ATSN. Finally, we compare the experimental results with some methods and visualize our attention maps.

4.1 Oilfield-7 Dataset

The behaviours in action recognition can be divided into two categories, one is a simple rule behaviour with restricted categories, such as walking, running, waving, bending, jumping, and so on. The other type is specific behaviours in specific scenarios, such as human body action on the oil field.

UCF101 [22] is an action recognition data set of realistic action videos, collected from YouTube, having 101 action categories. The action categories can be divided into five types: 1) Human-Object Interaction, 2) Body-Motion Only, 3) Human-Human Interaction, 4) Playing Musical Instruments, 5) Sports.

HMDB51 [23] – About 2GB for a total of 7,000 clips distributed in 51 action classes. The dataset contains 6849 clips divided into 51 action categories, each containing a minimum of 101 clips. The actions categories can be grouped in five types: 1) General facial actions, 2) Facial actions with object manipulation, 3) General body movements, 4) Body movements with object interaction, 5) Body movements for human interaction.

Oilfield-7 dataset includes seven action categories: Pump Clean, Rig Walk, Room Operate, Site Walk, Room Stand, Tank Construction, Tank Walk. The Oilfield-7 dataset is a data set for a specific scenario, contains human actions on the oilfield. In contrast, the HMDB51 and UCF101 data sets are human behaviours in a common scenario. Specific scenes and common scenarios are the biggest differences between them.

The Oilfield-7 dataset contains video clips, averaging frames per video. According to the standard evaluation criterion of data set [24], we divide the data into three kinds of training and testing data, and the classification performance is the average recognition accuracy generated by the three kinds of data partitioning. In addition, when editing our data set, we ensure that only one action occurs in a video, and we discard multiple actions in a video.

4.2 Experiment Details

Temporal Segment Network [10] is the most prominent two-stream model recently used for action recognition from video. It obtains 69.4% and 93.5% accuracy on two large action data sets, HMDB51 [23] and UCF101 [22], respectively. In this experiment, we use TSN to train the Oilfield-7 data set to extract the features of video segments. For spatial network and temporal network, the deep convolutional network structure we choose is BN-Inception [25], which is consistent with the setting of temporal network [10]. Because of the small number of data sets used for action recognition, there is a risk of over-fitting, so we enhanced the data. We adjust the input RBG and optical flow graphs to 256×340, and cut the width and height randomly from the set {256, 224, 192, 168}, and then adjust them to 224×224 uniformly as the input of the network. The optical flow graph is obtained by TV-L1 optical flow algorithm [26]. According to the previous work [8, 21], the 25 frame RGB or optical flow stack is selected from each video in the test. For each frame/stack, data enhancement is achieved by cutting four corners and one center. The learning of network parameters is carried out by using small batch stochastic gradient descent algorithm. The deep convolution neural network for feature extraction is pre-trained on ImageNet [27]. In the process of network training, the learning rate is 10^{-3} and the dropout rate is 0.8. When the final score of video category enters the fusion, the weight of spatial flow is set to 1 and the weight of temporal flow is set to 0.5.

4.3 Results and Analysis

In our experiment, two variants of ATSN are compared. One is the baseline deep convolution neural network BN-Inception, which removes attention and TSN structure; the other is the TSN_BN-Inception which removes attention. As shown in Tab. 1, our method shows better advantages. When attention is introduced to process long-distance time-structured video, higher attention can be paid to the local area of the video clip, so the feature representation of the video clip can better reflect the category of the video.

Table 1 Variant Model Comparison on Oilfield-7 Data Set (Split1)

Models (AUC)	Spatial net	Temporal net	Two-stream net
BN-Inception	0.896	0.536	0.873
BN-Inception_TSN	0.916	0.571	0.896
ATSN	**0.923**	**0.582**	**0.908**

We have carried out experiments on three kinds of partitioning data, each part shows the accuracy of the fusion of two streams. After that, the final result of our comparison is the average accuracy of the three parts, as shown in Tab. 2. Compared with the other two variants, ATSN shows the best performance. Compared with BN-Inception, the average accuracy increased by 2.3%, and 1.4% compared with BN-Inception_TSN.

Table 2 Performance comparison of three models on Oilfield-7 dataset

Models (AUC)	Split 1	Split 2	Split 3	Average
BN-Inception	0.873	0.876	0.871	0.873
BN-Inception_TSN	0.896	0.895	0.896	0.896
ATSN	**0.908**	**0.917**	**0.905**	**0.910**

In the Oilfield-7 data set, we use three models to test and use mAP evaluation indicators. The results are shown in Tab. 3. Our approach, ATSN, performs best in mAP than the other two. However, in the "Room Operate" and "Tank Construction" categories of action, BN-Inception_TSN showed better results than ATSN, because the human action performance in these two categories was not obvious, the attention was more concentrated in the background, and lost the attention to the action, resulting in negative effects which reduced the accuracy. In order to better understand the saliency of the network to the local area of the image in the learning process, we can visualize some attention maps, as shown in Fig. 3.

Table 3 Comparison of different methods on Oilfield-7 data set (abbreviations for the first line represent seven categories)

Models (AP)	PC	RW	RO	SW	RS	TC	TW	mAP
BN-Inception	0.594	0.745	0.683	0.838	0.786	0.674	0.738	0.723
BN-Inception_TSN	0.656	0.779	**0.766**	0.885	0.828	**0.796**	0.835	0.792
ATSN	**0.696**	**0.825**	0.716	**0.936**	**0.866**	0.737	**0.868**	**0.806**

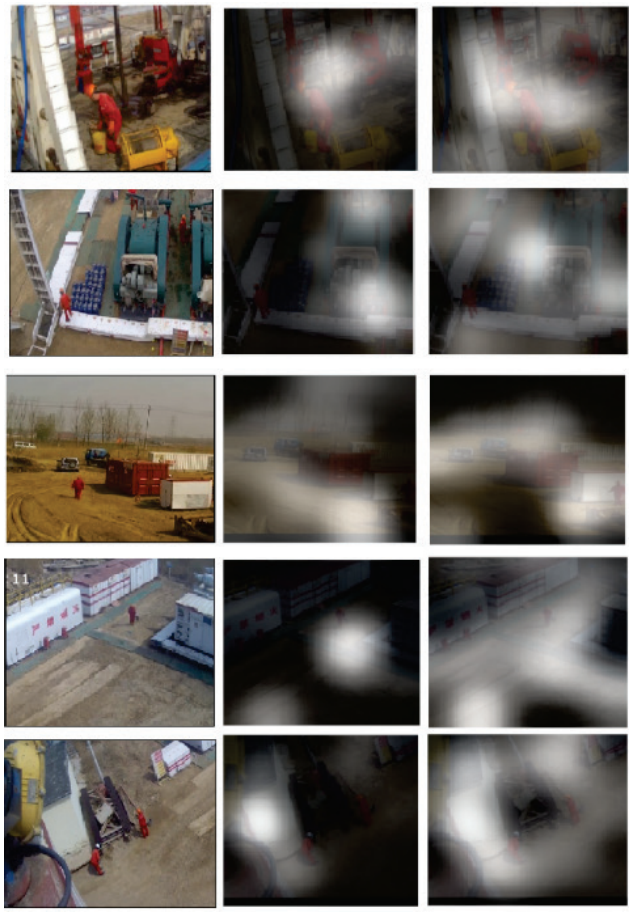

Figure 3 Automated attention visual image in Oilfield-7. The first column represents the original image extracted from the video, the second column is the most accurate result after attention, and the third column represents the broadest effect of attention. For example, for Tank Walk (Line 4), we can focus on and narrow the focus to the people walking in the field.

5 CONCLUSIONS

This paper presents a method of human action recognition based on deep convolution of visual attention, which we call ATSN. This method uses the attention mechanism to have a significant understanding of global information in image, focusing on local areas to obtain information, and realizes video classification more accurately and efficiently. Experiments on our Oilfield-7 dataset show that ATSN achieves higher accuracy of action recognition than basic deep convolutional network BN-inception and temporal network TSN, which proves the effectiveness of attention. There are some shortcomings in this method. In the process of network training of temporal stream, the accuracy of this method is not high. Because of the small human action range in Oilfield-7 data set, the information extracted from optical flow graph loses most of the motion information, which results in low recognition accuracy. In ATSN, the two-stream convolution network uses consensus function to achieve feature fusion. Noise labels exist in some segments of video, which affects video classification. In the next work, we will explore a fragment feature aggregation approach to replace consensus approach, and further study its impact on action recognition tasks.

This paper proposes a deep temporal network based on visual attention to identify human behavior, which we call ATSN. The method utilizes the attention mechanism to understand the global information in the image, focuses on the local area to obtain information, and realizes video classification more accurately and efficiently. Experiments on the Oilfield-7 dataset show that ATSN achieves higher accuracy of action recognition than the basic deep convolutional network BN-inception and the Temporal Segment Network, which proves the effectiveness of the attention mechanism. It should be pointed out that ATSN does not get high precision in the network training process of temporal flows. The reason is that in the Oilfield-7 dataset, the human motion amplitude is too small, and the optical flow graph information extracted by the network loses most of the motion information, resulting in lower recognition accuracy. The two-stream convolutional network in ATSN is implemented by the consensus function when performing feature fusion. Some segments in the video have noise tags, which affects video classification. In the next work, we will explore a way to segment the feature aggregation to replace the consensus method, and further study the impact of the aggregation mode on the action recognition task.

Acknowledgements

This work was supported in part by the Fundamental Research Funds for the Central Universities (Grant No.18CX02019A).

6 REFERENCES

[1] Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun. (2016). Deep Residual Learning for Image Recognition. *CVPR2016*. https://doi.org/10.1109/CVPR.2016.90

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun. (2016). Identity Mappings in Deep Residual Networks. *ECCV2016*. https://doi.org/10.1007/978-3-319-46493-0_38

[3] Simonyan, K. & Zisserman, A. (2015). Very Deep Convolutional Networks for Large-scale Image Recognition. *ICLR2015*.

[4] Szegedy, C., Ioffe, S., & Vanhoucke, V. (2017). Inception-v4, Inception-ResNet and the Image of Residual Connections on Learning. *AAAI2017*.

[5] Nguyen, T. V., Song, Z., & Yan, S. (2015). STAP: Spatial-Temporal Attention-Aware Pooling for Action Recognition. *IEEE2015*. https://doi.org/10.1109/TCSVT.2014.2333151

[6] Tran, D., Bourdev, L., Fergus, R., Torresani, L., & Paluri, M. (2015). Learning Spatiotemporal Features with 3D Convolutional Networks. *ICCV2015*. https://doi.org/10.1109/ICCV.2015.510

[7] Ji, S., Xu, W., & Yu, K. (2013). 3D Convolutional Neural Networks for Human Action Recognition. *IEEE2013*. https://doi.org/10.1109/TPAMI.2012.59

[8] Simonyan, K. & Zisserman, A. (2014). Two-Stream Convolutional Networks for Action Recognition in Videos. In *NIPS2014*, 568-576.

[9] Girdhar, R., Ramanan, D., Gupta, A., Sivic, J., & Russell, B. (2017). ActionVLAD: Learning spatio-temporal aggregation for action classification. *CVPR2017*. https://doi.org/10.1109/CVPR.2017.337
[10] Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X., Van Gool, L. (2016). Temporal Segment Networks: Towards Good Practices for Deep Action Recognition. *ECCV2016*. https://doi.org/10.1007/978-3-319-46484-8_2
[11] Laptev, I., Marszalek, M., Schmid, C., & Rozenfeld, B. (2008). Learning realistic human actions from movies. *IEEE2008*. https://doi.org/10.1109/CVPR.2008.4587756
[12] Wang, H., Ullah, M. M., Klaser, A., Laptev, I., & Schmid, C. (2009). Evaluation of local spatio-temporal features for action recognition. *BMVC2009*. https://doi.org/10.5244/C.23.124
[13] Wang, H., Klaser, A., Schmid, C., & Cheng-Lin, L. (2011). Action Recognition by Dense Trajectories. *CVPR2011*. https://doi.org/10.1109/CVPR.2011.5995407
[14] Wang, H. & Schmid, C. (2013). Action Recognition with Improved Trajectories. *ICCV2013*. https://doi.org/10.1109/ICCV.2013.441
[15] Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., & Fei-Fei, L. (2014). Large-scale Video Classification with Convolutional Neural Networks. *IEEE2014*. https://doi.org/10.1109/CVPR.2014.223
[16] Feichtenhofer, C., Pinz, A., & Zisserman, A. (2016). Convolutional Two-Stream Network Fusion for Video Action Recognition. *IEEE2016*. https://doi.org/10.1109/CVPR.2016.213
[17] Yue-Heing, J., Hausknecht, M., Vijayanarasimhan, S., Vinyals, O., Monga, R., & Toderici, G. (2015). Beyond Short Snippets: Deep Networks for Video Classification. *CVPR2015*.
[18] Carreira, J. & Zisserman, A. (2017). Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. *CVPR2017*. https://doi.org/10.1109/CVPR.2017.502
[19] Hou, X. & Zhang, L. (2008). Dynamic Visual Attention: Searching for coding length increments. *NIPS2008*.
[20] Mathe, S. & Sminchisescu, C. (2012). Dynamic Eye Movement Datasets and Learnt Saliency Models for Visual Action Recognition. *ECCV2012*. https://doi.org/10.1007/978-3-642-33709-3_60
[21] Wang, L., Xiong, Y., Wang, Z., & Qiao, Y. (2015). Towards Good Practices for Very Deep Two-Stream ConvNets. arXiv: 1507.02159.
[22] Soomro, K., Zamir, A. R., & Shah, M. (2012). UCF101: A Dataset of 101 Human Actions Classes From Videos in the Wild. CoRR, abs1212.0402.
[23] Kuehne, H., Jhuang, H., Garrote, E., Poggio, T., & Serre, T. (2011). HMDB: A Large Video Database for Human Motion Recognition. *ICCV2011*. https://doi.org/10.1109/ICCV.2011.6126543
[24] Idrees, H., Zamir, A. R., Jiang, Y.-G., Gorban, A., Laptev, I., Sukthankar, R., & Shah, M. (2016). The THUMOS Challenge on Action Recognition for Videos "in the Wild". arXiv: 1604.06182.
[25] Ioffe, S. & Szegedy, C. (2015). Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. ICML2015.
[26] Wedel, A., Pock, T., Zach, C., Bischof, H., & Cremers, D. (2009). *An Improved Algorithm for TV-L1 Optical Flow*. Springer. https://doi.org/10.1007/978-3-642-03061-1_2
[27] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). ImageNet: A Large-Scale Hierarchical Image Database. *CVPR2009*. https://doi.org/10.1109/CVPR.2009.5206848

Contact information:

Yun-lei SUN, Lecturer
College of Computer & Communication Engineering,
China University of Petroleum (East China),
Qingdao, 266580, China
E-mail: sunyunlei@upc.edu.cn

Da-lin ZHANG, Associate Professor
(Corresponding author)
National Research Center of Railway Safety Assessment,
Beijing Jiaotong University,
Beijing, 100044, China
E-mail: dalin@bjtu.edu.cn