Spatial-Temporal Neural Networks for Action Recognition

Chao Jing, Ping Wei, Hongbin Sun, Nanning Zheng

Spatial-Temporal Neural Networks for Action Recognition

Chao Jing[1,2], Ping Wei[1*], Hongbin Sun[1], Ningnan Zheng[1]

[1]Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University
[2] State Key Laboratory of Mathematical Engineering and Advanced Computing

Abstract. Action recognition is an important yet challenging problem in many applications. Recently, neural network and deep learning approaches have been widely applied to action recognition and yielded impressive results. In this paper, we present a spatial-temporal neural network model to recognize human actions in videos. This network is composed of two connected structures. A two-stream-based network extracts appearance and optical flow features from video frames. This network characterizes spatial information of human actions in videos. A group of LSTM structures following the spatial network describe the temporal information of human actions. We test our model with data from two public datasets and the experimental results show that our method improves the action recognition accuracy compared to the baseline methods.
Keywords: Action Recognition, Spatial-Temporal Structure, LSTM.

1 Introduction

Action recognition is to predict an action category label for an input video. It is an important problem in many applications, such as video search, security surveillance, and human-machine interaction.

Recognizing actions in daily-activity videos is a challenging problem. First, some different action categories have similar appearance and motion features. For example, the actions *drinking* and *eating* are very similar in motion features. Second, motion noise in videos increases the difficulty of action recognition. Third, the unrelated background or scene features often make the model unable to capture the key information of action recognition.

In this paper, we present a spatial-temporal neural network model to recognize human actions in videos. This network is composed of two connected structures - the spatial structure and the temporal structure, as shown in Figure 1. The spatial structure is a Two-Stream Network [1], which extracts appearance and optical flow features from video frames. Following the spatial network, the temporal structure is a group of LSTM networks [23] which represent the temporal and transition information of human actions. With these two structures, our model can deeply mine and utilize the spatial and temporal features in videos for action recognition. We test our model with data from two challenging datasets - MSR DailyActivity 3D [2] and UCF101 [3]. The

* Corresponding author: pingwei@xjtu.edu.cn

experimental results show that our method improves the action recognition performance compared to other baseline methods.

1.1 Related Work

Traditional action recognition methods generally consist of two key parts: feature extraction and feature classification. For feature extraction, most approaches are based on appearance, geometric, or motion features of human bodies, such as skeleton features [4,5], optical flow features [6]. These methods extract features from human bodies and can achieve satisfactory results in most scenes. However, in complex scenes with cluttered backgrounds, it is difficult to compute the accurate positions of human body parts, the action recognition accuracy is drastically depressed. Similar to HOG (Histogram of oriented gradient) [7] and SIFT (Scale-invariant feature transform) [8] in images, multi-scale feature extraction algorithms with prior knowledge were proposed. For example, some approaches extracted action features around the spatial-temporal interest point [9-11]. In complex scenes or backgrounds, such kinds of methods have achieved impressive improvements in action recognition accuracy. With the extracted features, various classifiers are learned to recognize actions, such as Support Vector Machine (SVM) [12].

Recently, neural networks and deep learning techniques [13-15] have been widely used in action recognition and achieved impressive performance [16-21]. Compared with static image classification, the temporal components of videos provide additional and important recognition clues - motion information [1,22]. In the early stage, action recognition based on single CNNs model was adopted [16]. Although this method improves action recognition performance compared with traditional methods, the characteristics of time series was not deeply processed. Later, two-stream networks [1] which utilize appearance and optical flow CNNs have significantly improved action recognition performance compared with the single CNNs model. After that, the Long Short-Term Memory (LSTM) models [23] and other Recurrent Neural Network (RNN) models are applied to action recognition tasks [24-26]. LSTM and RNN models incorporate the temporal information of videos into spatial features and therefore remarkably improve the recognition accuracy compared with previous neural network architectures.

Inspired by those models, our spatial-temporal network model is a hybrid architecture of the two-stream network [1] and the LSTM network [23-25]. From data preprocessing to network structures, it extracts local and global features, combines multi-feature learning, and is consistent with the sequential-data-based action recognition.

2 Spatial-Temporal Neural Network Model

A video frame containing appearance and geometric information of human actions is the smallest feature unit of the video sequence [24-28]. The temporal and motion information between successive frames is also essential for distinguishing different actions. Inspired by the previous convolutional network and LSTM methods

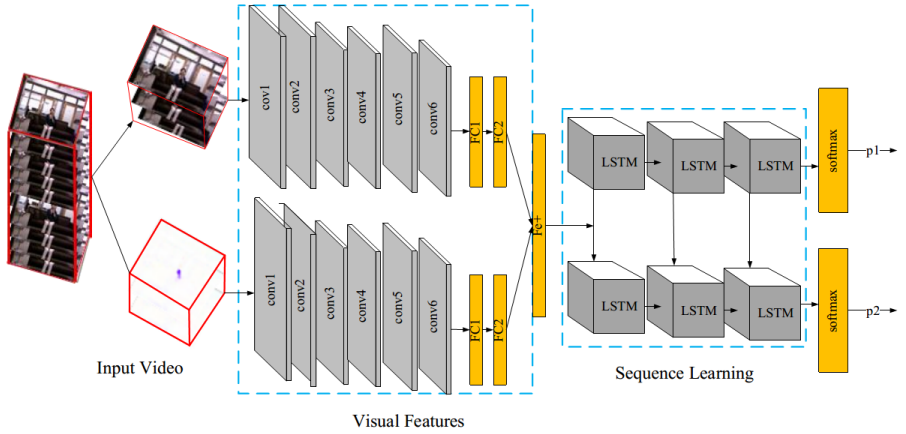


Fig. 1. Illustration of our spatial-temporal network model. The left side is a two-stream model and the right side with LSTM describes the temporal information of human actions.

[1, 23-25, 29], we present a spatial-temporal neural network model to jointly describe the spatial information in single frames and the temporal information between successive frames, as shown in Figure 1.

This network is composed of two connected structures - the spatial structure and the temporal structure, as shown in Figure 1. The spatial structure is a two-stream network [1], which extracts appearance and optical flow features from video frames. Following the spatial network, the temporal structure is a group of LSTM networks [25] which describe the temporal and motion information of human actions. With these two structures, our model can deeply mine and utilize the spatial and temporal features in videos for action recognition.

For a complete video, we first carry out the frame pre-processing (Section 3). One stream of the spatial structure extracts RGB features and another stream extracts optical flow features. The main structures of each stream are 6 convolutional layers and two dense layers [29]. The size of each channel input to the first convolutional layer (conv1) is 227*227*8. Through the feature extraction of the convolutional layer (conv1 to conv6), the dimension of our FC1 and FC2 layers is 4096. We use two layers of dense layer to prevent over fitting. Through this part, we obtain the feature in Fe+ layer with dimension of 4096*2L.

The temporal structure is a sequential 6 layers LSTM network. By dynamically inputting the obtained Fe+ features into the sequence learning module, it can learn the temporal feature of the video sequence.

3 Data Processing

Data preprocessing is the process to convert the original video sequence into the actual input data of the network model. In this section, we will address the problem of multi-frame inputs and introduce how to calculate multi-frame optical flows.

3.1 RGB Multi Frame Sequence Input Processing

We adopt a pre-processing method to take into account the feature expression of single frame RGB data and multi-frame RGB data. The basic idea is that for a complete video sequence, we take into account the connectivity of a video segment when we divide video into multiple segments. In the process, every 8 frames are used as a unit fragment. A sliding window is defined with 4 frames per step. It slides from the video start to the end. This process method is shown in algorithm 1.

Algorithm 1. Multi-frame preprocessing algorithm.

Multi-frame preprocessing algorithm.

for $i = 1; i < L; i++$ do
$N++;$
end for
for $i = 1; i < N; i++$ **do**
  for $w = 1; w < 8; w++$ **do**
    $F(f_1, f_2, \dots, f_n);$
  end for
end for
while$(4n + 8 \leq N)$ *and* $(n \geq 2)$ **do**
  $F_n \in [4n, 4n + 8];$
end while

3.2 Multi Optical Flow Calculation

In this section, we introduce how to calculate optical flow features in videos. Optical flow is an important feature of videos and it is widely used in the task of action recognition [22,30,1]. The optical flow contains the motion information of targets and is used to describes the video frame changes. In video sequence data, the instantaneous speeds of pixels can be used to characterize the correlations between pixel sequences in time domain, as shown in Figure 2.

We adopt the similar method with two-stream network [1] to calculate the optical flow featues in videos. For a video segment with L frames, we extract the optical flow information along the X and Y axes in each two adjacent frames. Then the optical flow feature of the video segment is an encapsulation of all the frame optical flow features. It is a vector with a dimension of w*h*2L [1], where w*h is the dimension of the single optical flow.

Fig. 2. Illustration of optical flow in human actions. (a) and (b) show a pair of successive video frames with human body motion. (c) shows the motion area.

4 Experiments

We use the action recognition accuracy to evaluate the performance of different methods. The action recognition accuracy is defined as the ratio of correctly labeled video numbers to all testing video numbers. We train the model under the caffe framework [31] and use the hardware CUDA plus GPU to deal with the floating-point matrix operation of the network. We test the models on the MSR DailyActivity 3D dataset [2] and the data samples from the UCF101 datase [3].

For the convolutional network component, we use video frames and optical flows to fine-tune a pre-trained AlexNet model [32]. We set the learning batch size as 32. The learning rate starts at 0.001 and is divided by 10 after every 30k iterations. For all experimental settings, we set the dropout regularization ratio as 0.5 to reduce complex co-adaptations of neurons in nets.

For the LSTM part, the output of Fe+ is used as the input to the LSTM. The momentum and weight decay are set as 0.9 and 0.0005, respectively. The learning rate starts at 0.01 and is divided by 10 after every 30k iterations. The output dimension of the softmax layer is 16.

4.1 Action Recognition on MSR DailyActivity 3D Dataset

The MSR DailyActivity 3D dataset [2] is captured using a Kinect camera. There are 16 action classes: *drinking water, eating, reading, calling, writing on paper, using notebooks, vacuuming, waking up, sitting, throwing paper, playing games, lying on the sofa, walking, playing guitar, standing up, sitting down*. There are ten subjects in total and two types of actions in each subject. One type action is at a standing position and one at a sitting position. The depth frames, the 3D skeletons of human bodies, and the RGB frames are recorded.

We compare our method with seven other approaches: Dynamic Temporal Warping [33], Actionlet Ensemble on Joint Features [34], HDMM+3ConvEets [35], 4DH [36], 4DHOI [36], Proposed method with Spatial Structure, and Proposed method with Temporal Structure. Proposed method with Spatial Structure only uses the Two-Stream Network component, and Proposed method with Temporal Structure uses the

LSTM component. Table 1shows the overall action recognition accuracy comparison, and Figure 3 (a) shows the accuracy of each action category.

Our method achieves an accuracy of 0.87, which outperforms other baseline approaches. Table 1 also shows that our method outperforms the spatial structure method and temporal structure method by a considerable margin, which proves the effectiveness of joint spatial-temporal network.

Table 2. Action recognition comparison on MSR DailyActivity 3D Dataset.

Method	Accuracy
Dynamic Temporal Warping [23]	0.54
Actionlet Ensemble on Joint Features [24]	0.74
HDMM+3ConvEets [25]	0.81
4DH [26]	0.74
4DHOI [26]	0.80
Spatial Structure Method	0.74
Temporal Structure Method	0.78
Proposed Method (Spatial + Temporal)	**0.87**

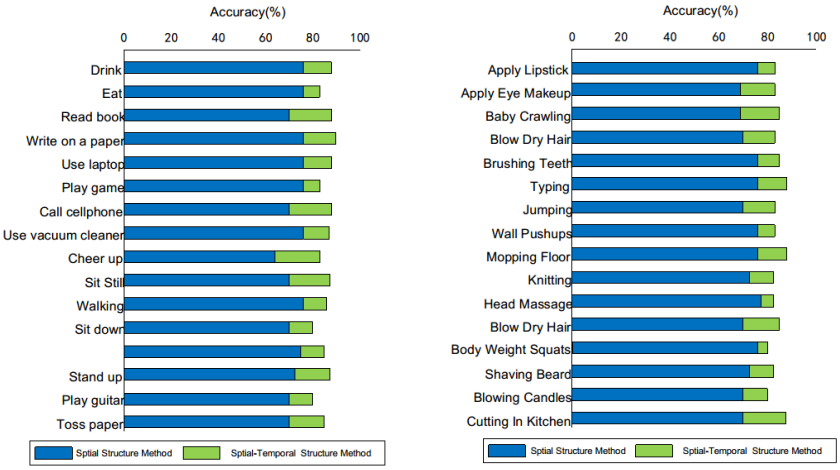
4.2 Action Recognition on UCF101 Dataset

UCF101 [3] is a large-scale dataset of realistic action videos. It has 101 action categories. UCF101 dataset has the largest diversity in terms of actions and with large variations in camera motion, object appearances, poses, scales, viewpoints, cluttered backgrounds, illumination conditions, etc. It is one of the most challenging datasets to date. We choose 16 categories of indoor actions in the dataset for our experiments. These 16 categories are: *Apply Eye Makeup, Apply Lipstick, Baby Crawling, Blow Dry Hair, Brushing Teeth, Typing, Jumping Jack, Wall Pushups, Mopping Floor, Knitting, Head Massage, Blow Dry Hair, Body Weight Squats, Shaving Beard, Blowing Candles, Cutting In Kitchen*. Each category consists of 25 groups and each group has 4 videos, with a total of 100 videos per category. The videos from the same group may share some common features, such as similar background, similar viewpoint, etc.

We compare our approach with the spatial structure method of two-stream network and the temporal structure method of LSTM. Table 2 shows the accuracy comparison and Figure 3 (b) shows the accuracy of each action class. Our method achieves an accuracy of 0.85. The results show that our method outperforms the comparison methods by a large margin, which proves the strength and effectiveness of our method.

Table 2. Action recognition comparison with UCF101 data.

Method	Accuracy
Spatial Structure Method	0.72
Temporal Structure Method	0.75
Proposed Method (Spatial + Temporal)	**0.85**



(a) MSR DailyActivity3D (b) UCF101

Fig. 3. Accuracy of each action of MSR DailyActivity 3D and UCF101 recognition on the network.

5 Conclusions

This paper presents a spatial-temporal neural network model to recognize human actions in videos. Our model jointly uses temporal and spatial dimension features of video sequences. With spatial and temporal structures, our model can deeply mine and utilize the spatial and temporal features in videos for action recognition. We test our model on two challenging datasets. The experimental results show that our methods improve the performance compared to other baseline methods.

Our future work will focus on complex neural network models on action recognition and video understanding.

Acknowledgement

This work is supported by National Natural Science Foundation of China 61503297, National Key Research and Development Program of China 2016YFB1000903, National Natural Science Foundation of China 61790563, and the Open Project Program of State Key Laboratory of Mathematical Engineering and Advanced Computing.

References

1. Simonyan, K., Zisserman, A.: Two-stream convolutional networks for action recognition in videos. In: Advances in Neural Information Processing Systems 1(4) (2014) 568–576
2. Wang, J., Liu, Z., Wu, Y., Yuan, J.: Learning actionlet ensemble for 3d human action recognition. In: IEEE Conference on Computer Vision and Pattern Recognition. (2012) 1290–1297
3. Soomro, K., Zamir, A.R., Shah, M.: Ucf101: A dataset of 101 human actions classes from videos in the wild. In: Computer Science (2012)
4. Fujiyoshi, H., Lipton, A.J.: Real-time human motion analysis by image skeletonization. In:Applications of Computer Vision, 1998. WACV'98. Proceedings., Fourth IEEE Workshop on. (2002) 15
5. Wei, P., Zheng, N., Zhao, Y., Zhu, S.C.: Concurrent action detection with structural prediction. In: International Conference on Computer Vision. (2013) 3136–3143
6. Chaudhry, R., Ravichandran, A., Hager, G., Vidal, R.: Histograms of oriented optical flow and binet-cauchy kernels on nonlinear dynamical systems for the recognition of human actions. In: Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on. (2009) 1932–1939
7. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on. (2005) 886–893
8. Lowe, D.G.: Object recognition from local scale-invariant features. In: Proceedings of the International Conference on Computer Vision. (1999)
9. Sch, C., Laptev, I., Caputo, B.: Recognizing human actions: A local svm approach. In: International Conference on Pattern Recognition. (2004) 32–36
10. Wang, H., Kläser, A., Schmid, C., Liu, C.L.: Action recognition by dense trajectories. In: IEEE Conference on Computer Vision and Pattern Recognition. (2011) 3169–3176
11. Kläser, A., Marszalek, M., Schmid, C.: A spatio-temporal descriptor based on 3d-gradients. In: The British Machine Vision Conference2008. (2008)
12. Cortes, C., Vapnik, V.: Support-vector networks. Machine Learning 20(3) (1995) 273–297
13. Schmidhuber, J.: Deep learning in neural networks: an overview. In: Neural networks: The Official Journal of the International Neural Network Society 61 (2014) 85
14. Goodfellow, I., Bengio, Y., Courville, A.: Deep Learning. MIT Press (2016)
15. Lecun, Y., Bengio, Y., Hinton, G.: Deep learning. Nature 521 (5 2015) 436–444
16. Ji, S., Xu, W., Yang, M., Yu, K.: 3d convolutional neural networks for human action recognition. In: IEEE Transactions on Pattern Analysis and Machine Intelligence 35(1) (2012) 221–231
17. Chen, X., Weng, J., Lu, W., Xu, J., Weng, J.: Deep manifold learning combined with convolutional neural networks for action recognition. IEEE Transactions on Neural Networks & Learning Systems (99) (2017) 1–15
18. Li, C., Sun, S., Min, X., Lin, W., Nie, B., Zhang, X.: End-to-end learning of deep convolutional neural network for 3d human action recognition. In: IEEE International Conference on Multimedia & Expo Workshops. (2017) 609–612
19. Rahmani, H., Mian, A., Shah, M.: Learning a deep model for human action recognition from novel viewpoints. IEEE Transactions on Pattern Analysis and Machine Intelligence 40(3) (2018) 667 - 681
20. Husain, F., Dellen, B., Torras, C.: Action recognition based on efficient deep feature learning in the spatio-temporal domain. IEEE Robotics & Automation Letters 1(2) (2016) 984–991

21. Mora, S.V., Knottenbelt, W.J.: Deep learning for domain-specific action recognition in tennis. In: Computer Vision and Pattern Recognition Workshops. (2017) 170–178
22. Papenberg, N., Bruhn, A., Brox, T., Didas, S., Weickert, J.: Highly accurate optic flow computation with theoretically justified warping. In: International Journal of Computer Vision 67(2) (2006) 141–158
23. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Computation 9(8) (1997) 1735–1780
24. Donahue, J., Hendricks, L.A., Guadarrama, S., Rohrbach, M., Venugopalan, S., Darrell, T., Saenko, K.: Long-term recurrent convolutional networks for visual recognition and description. In: Computer Vision and Pattern Recognition. (2015) 677–691
25. Baccouche, M., Mamalet, F., Wolf, C., Garcia, C., Baskurt, A.: Sequential deep learning for human action recognition. In: Human Behavior Understanding, Springer Berlin Heidelberg (2011) 29–39
26. Ng, Y.H., Hausknecht, M., Vijayanarasimhan, S., Vinyals, O., Monga, R., Toderici, G.: Beyond short snippets: deep networks for video classification. 16 (2015) 4694–4702
27. Graves, A.: Supervised sequence labelling with recurrent neural networks. In: Springer Berlin Heidelberg (2012)
28. Barbu, A., Bridge, A., Burchill, Z., Coroian, D., Dickinson, S., Fidler, S., Michaux, A., Mussman, S., Narayanaswamy, S., Salvi, D., Schmidt, L., Shangguan, J., Siskind, J.M., Waggoner, J., Wang, S., Wei, J., Yin, Y., Zhang, Z.: Video in sentences out. 1401 (2012) 274–283
29. Yuan, Z.W., Zhang, J.: Feature extraction and image retrieval based on alexnet. In: Eighth International Conference on Digital Image Processing. (2016)
30. Baker, S., Roth, S., Scharstein, D., Black, M.J., Lewis, J.P., Szeliski, R.: A database and evaluation methodology for optical flow. In: IEEE International Conference on Computer Vision. (2007) 1–31
31. Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., Darrell, T.: Caffe: Convolutional architecture for fast feature embedding. (2014) 675–678
32. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems 25. (2012) 1097–1105
33. Müller, M., Röder, T.: Motion templates for automatic classification and retrieval of motion capture data. In: ACM Siggraph/eurographics Symposium on Computer Animation, SCA 2006, Vienna, Austria, September. (2006) 137–146
34. Wang, J., Liu, Z., Wu, Y., Yuan, J.: Learning actionlet ensemble for 3d human action recognition. In: IEEE Transactions on Pattern Analysis & Machine Intelligence 36(5) (2014) 914
35. Wang, P., Li, W., Gao, Z., Zhang, J., Tang, C., Ogunbona, P.: Deep convolutional neural networks for action recognition using depth map sequences. In: Computer Science (2015)
36. Wei, P., Zhao, Y., Zheng, N., Zhu, S.C.: Modeling 4d human-object interactions for joint event segmentation, recognition, and object localization. In: IEEE Transactions on Pattern Analysis and Machine Intelligence 39(6) (2017) 1165–1179