Postprint of Human Action Recognition Model Based on Improved Deep Neural Network

Authors: He Bingqian, Wei Wei, Zhang Bin, Gao Lianxin, Song Yanbei

Date: 2018-09-12T00:00:00+00:00

Abstract

To address issues such as the requirement for fixed-length video segments and insufficient utilization of spatio-temporal information in existing human action recognition methods, we propose a deep neural network model combining spatio-temporal pyramids with attention mechanisms. The model integrates a 3D-CNN incorporating spatio-temporal pyramids and an LSTM model augmented with spatio-temporal attention mechanisms, enabling multi-scale processing of video segments and full exploitation of complex spatio-temporal action information. RGB images and optical flow fields are employed as spatial and temporal inputs, respectively, while fused features combining motion and appearance features from pyramid pooling layers serve as the fusion domain input. Finally, a decision-level fusion strategy is adopted to obtain the final action recognition results. Experiments on the UCF101 and HMDB51 datasets achieved recognition accuracies of 94.2% and 70.5%, respectively. Experimental results demonstrate that the improved network model achieves high recognition accuracy on video-based human action recognition tasks.

Full Text

Preamble

Improved Deep Convolutional Neural Network for Human Action Recognition

He Bingqian, Wei Wei, Zhang Bin, Gao Lianxin, Song Yanbei
(College of Computer Science & Technology, Chengdu University of Information Technology, Chengdu 610225, China)

Abstract: Existing human action recognition methods require fixed-length video segments as input and fail to fully utilize spatiotemporal information. To address these limitations, this paper proposes a deep neural network model that combines spatiotemporal pyramid pooling with an attention mechanism.

The improved architecture integrates a 3D-CNN incorporating spatiotemporal pyramids with an LSTM model enhanced with spatiotemporal attention, enabling multi-scale processing of video segments and comprehensive exploitation of complex spatiotemporal action information. RGB images and optical flow fields serve as inputs for the spatial and temporal domains, respectively, while fused features combining motion and appearance characteristics from the pyramid pooling layer constitute the fusion domain input. Final action recognition results are obtained through a decision-level fusion strategy. Experiments on the UCF101 and HMDB51 datasets achieve recognition accuracies of 94.2% and 70.5%, respectively. The experimental results demonstrate that the improved network model attains high recognition accuracy for video-based human action recognition tasks.

Keywords: action recognition; deep learning; spatiotemporal pyramid; attention mechanism; convolutional neural network

0 Introduction

Human action recognition has become one of the most active research areas in computer vision due to its wide applications in human-robot interaction, virtual reality, home security, and public safety. Current recognition algorithms can be broadly categorized into two classes: traditional hand-crafted feature-based methods and deep learning-based approaches. Deep learning methods have demonstrated significant advantages over traditional approaches on various challenging video datasets. Nevertheless, accurately distinguishing between different action categories remains challenging due to environmental factors such as illumination variations and occlusions, inter-class and intra-class differences in action categories, and the limited size of video datasets. These issues pose substantial challenges for robust feature extraction and action classification.

To overcome the limitation of convolutional neural networks being primarily applied to 2D images and to effectively incorporate motion information for video analysis, Ji et al. [14] proposed performing three-dimensional convolutions in CNN layers to capture discriminative features across both spatial and temporal dimensions. However, this model still fails to fully exploit spatiotemporal video features. Simonyan and Zisserman [6] introduced a two-stream convolutional network that better utilizes temporal information in video data. This architecture employs two convolutional networks that take RGB images and optical flow from video frames as inputs, respectively, extracting temporal and spatial features for action representation before fusing them for classification. While this model leverages spatiotemporal video sequence features to some extent, it may be insufficient for capturing complex spatiotemporal cues across different action categories since it only focuses on convolutional mappings at the current step [13].

Most existing CNN-based recognition models capture only short-term spatiotem-

poral features and cannot represent long-term variations. Experimental results from several studies [9,10,15,16] have demonstrated that recurrent neural networks (RNNs), particularly Long Short-Term Memory (LSTM) models [17], can effectively address this issue by modeling video sequences. However, in these models, LSTM inputs are high-level features extracted directly from fully connected layers of CNNs, which lack spatiotemporal detail. Current LSTM-based action recognition methods also suffer from losing important spatiotemporal cues, preventing models from obtaining sufficient spatiotemporal relationships for human actions, and most require manual preprocessing of video segments to handle fixed-length inputs.

To address these problems, this paper proposes a deep neural network model that combines spatiotemporal pyramid pooling with an attention mechanism (STPP and attention-mechanism network) based on the spatiotemporal two-stream convolutional network architecture. For tasks requiring direct processing of variable-length video segments, we introduce a simple modification to the original C3D network by adding spatiotemporal pyramid pooling after the final convolutional layer, enabling the model to generate fixed-length feature vectors. Since spatiotemporal pyramid pooling processes feature maps from multiple perspectives, the model can obtain deeper feature representations, thereby improving recognition accuracy. For capturing complex spatiotemporal cues between human actions, we design an LSTM model with added spatiotemporal attention mechanism. This model not only captures long-term temporal information but also captures complex spatiotemporal cues through the attention mechanism. Additionally, we incorporate a spatiotemporal feature fusion module to minimize the loss of important action features. Experimental results on the UCF101 and HMDB51 datasets demonstrate that the proposed model can effectively recognize human actions in videos.

---

2 Design of Deep Neural Network Model for Human Action Recognition Combining Spatiotemporal Pyramid and Attention Mechanism

2.1 Overall Framework

Deep learning has achieved remarkable success in image recognition, leading to extensive research and application of deep learning methods, particularly convolutional neural networks, in computer vision. Unlike static images, video sequences contain both appearance and motion information [18]. Consequently, recent studies have attempted to design CNN-based action recognition models that effectively utilize both appearance and motion information from video sequences. Karpathy et al. [19] compared three widely used CNN connection methods: late fusion, early fusion, and slow fusion, concluding that these approaches cannot fully exploit motion information and only provide modest improvements for individual frames. Tran et al. [20] trained a deeper CNN model

called C3D on UCF101 and Sports-1M, which approximates a 3D version of VGGnet [7] and includes 3D convolutional filters and 3D pooling layers operating on both temporal and spatial domains. Simonyan and Zisserman [6] proposed a two-stream convolutional neural network that trains a second CNN stream on optical flow from video frames, compensating for the inability of stacked RGB streams to fully utilize temporal information and providing performance gains for action recognition. This model has been widely adopted in many other action recognition methods [8,9,21,22].

However, the original two-stream convolutional neural network model has two main problems: (a) it cannot capture long-term temporal information as it only contains 10 consecutive optical flow frames; and (b) it trains spatial and temporal domains separately, with final predictions obtained by averaging the outputs of two classifiers, thus failing to effectively learn spatiotemporal relationships between temporal and spatial streams. To address these issues, Krishnan et al. [10] proposed an LSTM-based action recognition method to fuse features from longer-term video sequences. Wang et al. [23] introduced a segmental network architecture with sparse sampling to model long-term temporal structures. Feichtenhofer et al. [24] proposed a spatiotemporal fusion method by investigating various ways to combine networks in time and space, arguing that two-stream networks should be fused at the final convolutional layer. Although these methods improve upon the original two-stream convolutional neural network, they still suffer from losing important spatiotemporal cues, preventing models from obtaining sufficient spatiotemporal relationships for human actions and requiring manual preprocessing of video segments to fixed lengths.

Considering these issues, this paper proposes a human action recognition model based on a deep neural network combining spatiotemporal pyramid and attention mechanism, building upon Feichtenhofer et al. [24]. The model first obtains spatiotemporal convolutional feature maps from RGB images and optical flow of video sequences through 3D convolutional neural networks. It then employs spatiotemporal pyramid pooling (STPP) to aggregate local spatiotemporal information into fixed-length feature vectors. Effective spatiotemporal feature fusion is performed at the STPP layer through a spatiotemporal feature fusion strategy. Finally, the spatiotemporal features extracted from the spatiotemporal 3D two-stream network and the fused features are input into LSTM models with and without spatiotemporal attention mechanisms for modeling, respectively. The classification results from these models are fused to obtain the final human action classification. Experiments on the UCF101 and HMDB51 datasets demonstrate the effectiveness of the proposed model in recognizing human actions in videos.

The network framework is illustrated in Figure 1 [Figure 1: see original paper]. The model comprises three main modules: (1) a spatiotemporal two-stream 3D convolutional neural network incorporating spatiotemporal pyramid pooling; (2) spatial and temporal domain feature fusion; and (3) long short-term memory with spatiotemporal attention mechanism.

For the first module, we adopt the spatiotemporal two-stream model from [6] and the C3D network structure from [20], modifying them to form our spatiotemporal two-stream 3D convolutional neural network module. Both temporal and spatial deep convolutional networks consist of 5 convolutional layer groups, 4 max-pooling layers, 1 spatiotemporal pyramid pooling layer, and 2 fully connected layers. Specifically, the final max-pooling layer of the original C3D network is replaced with spatiotemporal pyramid pooling. STPP not only handles variable input sizes but also extracts deeper features through multi-perspective feature extraction, thereby improving recognition accuracy. The network comprises 5 convolutional layer groups with filter numbers of 64, 128, 256, 512, and 512 from groups 1 to 5, respectively, and two fully connected layers with 4096 units each. Based on experimental results from [19] showing that 3×3×3 kernel size is optimal for all convolutional layers, we adopt 3×3×3 kernels with a stride of 1×1×1 in this module. For max-pooling layers, the first layer uses a kernel size of 2×2×1, while the remaining three use 2×2×2. The first module connects directly to the third module.

The second module fuses spatiotemporal features extracted from the two streams and connects to a standard LSTM model without spatiotemporal attention mechanism in the third module. This module operates at the STPP layer of the first module. The third module is an LSTM model with added attention mechanism. As a recurrent neural network, LSTM can capture long-term spatiotemporal dependencies by preserving temporal sequence information while effectively avoiding gradient vanishing. Compared to the original LSTM, this module can capture more complex spatiotemporal cues, thereby improving recognition accuracy. Overall, the proposed network framework incorporates both feature-level and decision-level fusion, enabling more accurate human action recognition through these two fusion approaches.

After pre-training and fine-tuning on ImageNet, video sequences of RGB images and optical flow frames are input into the model. Two 3D convolutional neural networks are trained to extract temporal and spatial stream features, which are then processed by spatiotemporal pyramid pooling to obtain fixed-length feature vectors. Deep features of video frames are extracted through two fully connected layers. Spatiotemporal features extracted from the STPP layer are fused using a spatiotemporal feature fusion strategy. Finally, an LSTM model with spatiotemporal attention mechanism models the spatiotemporal features to obtain classification results.

2.2 Spatiotemporal Pyramid Pooling

To process video sequences of arbitrary size and length with our model, we employ spatiotemporal pyramid pooling (STPP) to generate fixed-length feature vectors. Since the STPP layer can extract features from convolutional feature maps at multiple perspectives, it enhances human action recognition performance. This layer can accept video sequences of any size and length. Assuming input RGB and optical flow image sequences have size $l \times h \times w$, where l is

the length (number of frames), while the final convolutional feature map has size $T \times H \times W$, where T is the temporal size of pooling cubes and h, H and w, W are the heights and widths of frames, respectively. We concentrate the response values from each spatiotemporal cube input to STPP through max pooling operations. Unlike standard sliding window pooling in [20], STPP' s sliding window size is dynamically adjusted within a given pooling level.

Let $\{p_t^s, p_s^s\}$ denote the spatiotemporal pooling levels, where p_t^s is the temporal pooling level and p_s^s is the spatial pooling level. Since the temporal scale of each video sequence is smaller than its spatial scale, we set $p_t^s = 1$. When $p_s^s = \{4, 2, 1\}$, each input video clip generates a fixed-length descriptor, enabling STPP to form fixed-length feature vectors by aggregating local spatiotemporal information.

2.3 Spatiotemporal Feature Fusion

For video-based human action recognition, extracted features include both static visual and dynamic motion characteristics. Appropriate and effective feature fusion methods can leverage correlations between these two feature types to generate more diverse mixed features. Based on [24], we propose a spatiotemporal feature fusion framework.

For the t-th video segment input to the model, we obtain two STPP features at the first module' s STPP layer, denoted as X_t^a and X_t^m, where X_t^a represents RGB features (appearance features) and X_t^m represents optical flow features (motion features) of the t-th segment. We employ early fusion (element-wise concatenation) to fuse these STPP features, generating a new fused feature X_t^f. The mixed feature then passes through a fully connected layer with 4096 units before connecting to the third module, where a long short-term memory model models and classifies the fused features.

2.4 LSTM Model with Spatiotemporal Attention Mechanism

In this module, we design an LSTM model with spatiotemporal attention mechanism (S-P attention-mechanism LSTM) to model the deep features obtained earlier. As a recurrent neural network, LSTM can capture long-term spatiotemporal dependencies by preserving temporal sequence information. Unlike original RNNs, LSTM avoids gradient vanishing after backpropagation training. Video sequences for human action recognition contain numerous spatiotemporal cues. Directly inputting features from the first module' s fully connected layer into LSTM would be insufficient for capturing complex spatiotemporal cues of different actions. Therefore, to capture more useful features, we add spatiotemporal attention to the basic LSTM model.

A standard LSTM unit is shown in Figure 2 [Figure 2: see original paper], where $*$ represents either a or m. We describe high-dimensional features from the first module' s fully connected layer as X_t^a and X_t^m, representing appearance and motion features of the t-th video segment, respectively. Fused features from

the second module' s fully connected layer are described as X_t^f. i_t^*, f_t^*, and o_t^* represent input, forget, and output gates, respectively, while g_t^*, c_t^*, h_t^*, and Y_t^* represent memory modulation states, cell states (memory states), hidden states, and outputs. For fused features X_t^f, we input them into a standard LSTM, implemented as follows:

$$
\begin{aligned}
i_t^f &= \sigma(W_{xi}^f X_t^f + W_{hi}^f h_{t-1}^f + b_i^f) \\
f_t^f &= \sigma(W_{xf}^f X_t^f + W_{hf}^f h_{t-1}^f + b_f^f) \\
o_t^f &= \sigma(W_{xo}^f X_t^f + W_{ho}^f h_{t-1}^f + b_o^f) \\
g_t^f &= \tanh(W_{xg}^f X_t^f + W_{hg}^f h_{t-1}^f + b_g^f) \\
c_t^f &= f_t^f \odot c_{t-1}^f + i_t^f \odot g_t^f \\
h_t^f &= o_t^f \odot \tanh(c_t^f)
\end{aligned}
$$

where $\sigma(\cdot)$ and $\tanh(\cdot)$ denote the sigmoid and hyperbolic tangent activation functions, respectively, and $\odot$ represents the Hadamard product (element-wise multiplication).

2.5 Spatiotemporal Attention Mechanism of LSTM

The spatiotemporal attention mechanism model for LSTM is shown in Figure 3 [Figure 3: see original paper]. The attention mechanism operates simultaneously on both spatial and temporal domains. Spatial domain input is X_t^*, while temporal domain input is h_{t-1}^*. To avoid repetitive description, we uniformly represent the module' s input as X_t^*, where * represents either a or m. To identify feature vectors with important descriptive significance in the t-th video segment, we first perform spatial attention computation for each stream. The computational process is as follows.

Taking the previous hidden state h_{t-1}^* of an LSTM unit as an example, we first calculate the spatial attention probability $\alpha_{t,nk}^*$ of the k-th feature vector on the n-th feature vector in the t-th video segment using equations (7) and (8):

$$
\alpha_{t,nk}^* = \frac{\exp(w_\alpha^T \tanh(W_{h\alpha}^* h_{t-1}^* + W_{X\alpha}^* X_{t,nk}^* + b_\alpha^*))}{\sum_{l=1}^{L} \exp(w_\alpha^T \tanh(W_{h\alpha}^* h_{t-1}^* + W_{X\alpha}^* X_{t,nl}^* + b_\alpha^*))}
$$

where $W_{h\alpha}^*$, $W_{X\alpha}^*$, w_α, and b_α^* are weight matrices and bias vectors of the spatial attention mechanism, L is the number of frames in the t-th video segment, and $\alpha_{t,nk}^*$ represents unnormalized attention probabilities. We then obtain the spatial feature vector $L_{t,n}^*$ for the n-th feature vector using equation (9):

$$
L_{t,n}^* = \sum_{k=1}^{L} \alpha_{t,nk}^* X_{t,nk}^*
$$

After obtaining the spatially-weighted feature vector $L^*_{t,n}$, we compute temporal attention. Similar to spatial attention calculation, we first compute temporal attention probabilities $\beta^*_{t,n}$ as follows:

$$\beta^*_{t,n} = \frac{\exp(w^T_\beta \tanh(W^*_{h\beta} h^*_{t-1} + W^*_{L\beta} L^*_{t,n} + b^*_\beta))}{\sum_{j=1}^{T} \exp(w^T_\beta \tanh(W^*_{h\beta} h^*_{t-1} + W^*_{L\beta} L^*_{t,j} + b^*_\beta))}$$

where $W^*_{h\beta}$, $W^*_{L\beta}$, w_β, and b^*_β are weight matrices and bias vectors of the temporal attention mechanism, and T is the total number of feature vectors in the t-th video segment. $\beta^*_{t,n}$ reflects the temporal importance of the n-th feature vector for the t-th video segment. The final important spatiotemporal context feature Φ^*_t captured by spatiotemporal attention is calculated according to equation (12):

$$\Phi^*_t = \sum_{n=1}^{T} \beta^*_{t,n} L^*_{t,n}$$

Since the context feature Φ^*_t is closely related to current step predictions, we use it as additional input to the LSTM model alongside the original feature vector X^*_t. The specific calculation formulas are as follows:

$$\begin{aligned}
i^*_t &= \sigma(W^*_{xi} X^*_t + W^*_{\Phi i} \Phi^*_t + W^*_{hi} h^*_{t-1} + b^*_i) \\
f^*_t &= \sigma(W^*_{xf} X^*_t + W^*_{\Phi f} \Phi^*_t + W^*_{hf} h^*_{t-1} + b^*_f) \\
o^*_t &= \sigma(W^*_{xo} X^*_t + W^*_{\Phi o} \Phi^*_t + W^*_{ho} h^*_{t-1} + b^*_o) \\
g^*_t &= \tanh(W^*_{xg} X^*_t + W^*_{\Phi g} \Phi^*_t + W^*_{hg} h^*_{t-1} + b^*_g) \\
c^*_t &= f^*_t \odot c^*_{t-1} + i^*_t \odot g^*_t \\
h^*_t &= o^*_t \odot \tanh(c^*_t)
\end{aligned}$$

3 Experiments

3.1 Dataset and Evaluation Metrics

We conduct experiments on two publicly available video action recognition datasets: UCF101 and HMDB51. UCF101 contains 13,320 video clips across 101 action categories, covering a wide range of human actions such as applying makeup, typing, blowing hair, horse riding, and high jumping. Most videos in this dataset are captured in unconstrained real-world environments, resulting in low resolution and environmental factors such as illumination variations and occlusions. HMDB51 contains 6,766 video clips across 51 action categories, mostly derived from movie clips with relatively low resolution. Major action categories include kissing, hugging, horse riding, and shooting.

In our experiments, we split both datasets into three partitions for training and testing. Each partition of UCF101 contains 9,500 video sequences, while HMDB51 contains 3,700 video clips. Since our network model comprises three streams (temporal, spatial, and fusion), for each dataset partition we linearly weight the results from these three base classifiers to obtain the final action recognition accuracy for that partition. The linear weighting uses adaptive dynamic weights calculated from the base classifiers' performance on the test set. After obtaining the final recognition accuracy for each of the three partitions, we compute their linear weighted average to obtain the dataset's overall recognition accuracy, which serves as our evaluation metric.

Decision fusion combines results from multiple base classifiers according to specific rules to produce a global result, eliminating information deficiencies in individual decisions or between decisions, thereby improving the reliability and stability of the final result [25]. Our network structure contains three parts: a fusion stream after feature fusion at the CNN' s STPP layer, and two branches that retain temporal and spatial streams after feature fusion with added attention mechanisms to capture complex spatiotemporal cues for refining the fusion stream' s recognition results. Therefore, for each dataset partition, the network produces three base classifier results, which are combined through decision fusion to obtain the final classification output.

Let C_j represent the final fused classification result, with the fusion rule formulated as:

$$C_j = \arg\max_c \sum_{i=1}^{3} \omega_i \cdot p_i(X_j|c)$$

where X_j is the source feature of the j-th base classifier, $p_i(X_j|c)$ is the confidence produced by the classifier when classifying each category c ($c = 1, 2, \ldots, N$), and ω_i represents the fusion weight, which is the classification accuracy of each base classifier (i.e., individual classification accuracy). Thus, we obtain source classification results from temporal, spatial, and fusion domain base classifiers, then fuse them using equation (19) to obtain the recognition result for each dataset partition.

3.2 Pre-training

Compared to image datasets, human action recognition datasets are relatively small, making deep neural networks prone to overfitting. Therefore, we pre-train our model. For the spatial domain network with RGB image inputs, we directly use the ImageNet dataset [26] for pre-training. Training images are from an augmented training set with random position cropping, resized to 224×224. For the temporal stream network with optical flow inputs, we use action video optical flow data extracted from TV-L1 [27]. To ensure the same range as RGB data, we linearly transform the optical flow data to the $[0, 255]$ interval. We then

average the filters of the first layer from the pre-trained spatial stream network across channels and replicate the averaged data 20 times as initialization values for the temporal network.

3.3 Experimental Results and Analysis

Experiments are conducted on a TensorFlow platform built on a Linux system. Deep neural networks are prone to overfitting, so we set dropout rates to 0.7 and 0.8 for spatial and temporal streams, respectively. The initial learning rate for the spatial domain is 10^{-3}, reduced to 10^{-4} after 15,000 iterations, with training stopping after 30,000 iterations. The temporal domain initial learning rate is 10^{-3}, reduced to one-tenth every 20,000 iterations after the first 20,000 iterations, with a maximum of 80,000 iterations.

Motion and appearance features of video sequences are extracted through our model's first module. Considering that different pooling levels in the STPP layer affect action recognition differently, we design comparative experiments with various pooling levels. These experiments evaluate action recognition accuracy using only the two-stream 3D convolutional neural network of the first module. We consider two STPP pooling levels: $\{4,4,1\}\times\{4,4,1\}\times\{1,1,1\}$ and $\{4,2,1\}\times\{4,2,1\}\times\{1,1,1\}$, denoted as STPP-1 and STPP-2, respectively. Experiments are conducted on the first split of UCF101. As shown in Table 1 , when the STPP pooling level is $\{4,2,1\}\times\{4,2,1\}\times\{1,1,1\}$, action recognition accuracy surpasses both STPP-1 and standard max pooling. Therefore, in subsequent experiments, STPP pooling levels are set to this configuration. Table 1 also shows that temporal domain recognition rates are higher than spatial domain rates under the same network structure, indicating that motion information better represents human action information than appearance information.

Table 1: Comparison of Action Recognition Accuracy Under Different STPP Pooling Levels

Method	Spatial Domain	Temporal Domain
Max pooling	82.76%	85.74%
STPP-1	82.18%	85.78%
STPP-2	87.26%	89.91%

Table 2 presents action recognition accuracy results with and without the spatiotemporal attention mechanism in the LSTM model of our third module. These results are obtained by weighted averaging across three splits of both datasets. Table 2 shows that using the attention-enhanced LSTM model in both temporal and spatial domains yields higher accuracy than without attention, proving that the spatiotemporal attention mechanism is more effective for human action recognition.

Table 2: Action Recognition Accuracy Comparison of LSTM Model With and Without Attention Mechanism

Attention Mechanism	UCF101 (%)	HMDB51 (%)
Without attention	89.73%	67.95%
With attention	91.02%	68.13%
Without attention (fusion)	92.52%	68.16%
With attention (fusion)	93.57%	70.52%

Table 3 shows the action recognition accuracy of our deep neural network model combining spatiotemporal pyramid and attention mechanism. For each dataset split, recognition accuracy is obtained through decision-level fusion that linearly weights base classifier results from temporal, spatial, and fusion domains. The final accuracy for each dataset is then computed by linearly averaging results from three splits.

Table 3: Human Action Recognition Accuracy of Our Model

Dataset Split	UCF101 (%)	HMDB51 (%)
Split1	93.95%	69.16%
Split2	94.67%	71.08%
Split3	94.13%	70.86%
Average	**94.21%**	**70.50%**

We compare our method with several typical deep learning methods and network models in recent action recognition research on UCF101 and HMDB51 datasets. These methods include the two-stream convolutional network [6], the C3D network [20] that trains deeper CNNs, the spatiotemporal fusion network [24] based on two-stream VGG, and the multi-level pyramid fusion model [28] built upon [24]. As shown in Table 4 , our proposed method achieves more accurate recognition of human actions in video sequences compared to recent classical algorithms.

Table 4: Action Recognition Accuracy of Different Methods on UCF101 and HMDB51 Datasets

Method	UCF101 (%)	HMDB51 (%)
Two-stream [6]	88.00%	59.40%
C3D [20]	85.80%	54.90%
Two-stream VGG [24]	92.70%	65.10%
SPN-VGG-16 [28]	93.40%	68.90%
Our method	**94.21%**	**70.50%**

4 Conclusion

Deep learning methods have been widely applied to various pattern recognition tasks. For human action recognition, this paper proposes an improved deep neural network model combining spatiotemporal pyramid pooling and attention mechanism, constructing a spatiotemporal two-stream deep neural network architecture. After pre-training and fine-tuning on ImageNet, we apply our model to UCF101 and HMDB51 datasets, achieving final recognition accuracies of 94.2% and 70.5% through fusing spatiotemporal and fusion streams. Experiments demonstrate that our improved deep learning model can effectively recognize human actions in these datasets. However, applying it to real-world commercial applications remains challenging. Future work will focus on developing robust algorithms for videos with significant environmental factors or noise.

References

[1] Mur O, Frigola M, Casals A. Modelling daily actions through hand-based spatio-temporal features [C]// Proc of International Conference on Advanced Robotics. Piscataway, NJ: IEEE Press, 2015: 478-483.

[2] Liu Fang, Xu Xiangmin, Qiu Shuoyang, et al. Simple to complex transfer learning for action recognition [J]. IEEE Trans on Image Processing, 2016, 25 (2): 949-960.

[3] Uddin A, Joolee J B, Alam A, et al. Human action recognition using adaptive local motion descriptor in Spark [J]. IEEE Access, 2017, 5: 21157-21167.

[4] Huang Xiaohui, Dong Chaojun. The depth map denoising and spatio-temporal feature extraction for human action recognition [J]. Modern Industrial Economy and Informationization, 2017, 2017 (5): 64-68.

[5] Zhang Jie, Wu Jiangzhang, Tang Jiali, et al. Human action recognition method based on spatio-temporal image segmentation and interactive area detection [J]. Application Research of Computers, 2017, 34 (1): 302-305.

[6] Simonyan K, Zisserman A. Two-Stream convolutional networks for action recognition in videos [J]. Advances in Neural Information Processing Systems, 2014, 1 (4): 568-576.

[7] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition [EB/OL]. (2015-04-10). https://arxiv.org/abs/1409.1556.

[8] Chéron G, Laptev I, Schmid C. P-CNN: pose-based CNN features for action recognition [C]// Proc of IEEE International Conference on Computer Vision. Piscataway, NJ: IEEE Press, 2015: 3218-3226.

[9] Srivastava N, Mansimov E, Salakhutdinov R. Unsupervised learning of video representations using LSTMs [C]// Proc of the 32nd International Conference

on Machine Learning. [S. l.]: International Machine Learning Society (IMLS), 2015: 843-852.

[10] Krishnan K, Prabhu N, Babu R V. ARRNET: Action recognition through recurrent neural networks [C]// Proc of International Conference on Signal Processing and Communications. 2016: 1-5.

[11] Wang Limin, Xiong Yuanjun, Wang Zhe, et al. Temporal segment networks: Towards good practices for deep action recognition [C]// Proc of European Conference on Computer Vision. Berlin: Springer, 2016: 20-36.

[12] Kar A, Rai N, Sikka K, et al. AdaScan: adaptive scan pooling in deep convolutional neural networks for human action recognition in videos [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2017: 5699-5708.

[13] Du Wenbin, Wang Yali, Qiao Yu. Recurrent spatial-temporal attention network for action recognition in videos [J]. IEEE Trans on Image Processing, 2017, 27 (3): 1347-1360.

[14] Ji Shuiwang, Xu Wei, Yang Ming, et al. 3D convolutional neural networks for human action recognition [J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2013, 35 (1): 221-231.

[15] Veeriah V, Zhuang Naifan, Qi Guojun. Differential recurrent neural networks for action recognition [C]// Proc of IEEE International Conference on Computer Vision. Piscataway, NJ: IEEE Press, 2015: 4041-4049.

[16] Ordóñez F J, Roggen D. Deep convolutional and LSTM recurrent neural networks for multimodal wearable activity recognition [J]. Sensors, 2016, 16 (1): 115-140.

[17] Hochreiter S, Schmidhuber J. Long short-term memory [J]. Neural Computation, 1997, 9 (8): 1735-1780.

[18] Chen Shengdi, Wei Wei, He Bingqian, et al. Human action recognition based on improved deep convolutional neural networks [J/OL]. Application Research of Computers, 2019, 36 (4). (2018-02-09) [2018-08-23]. http://www.arocmag.com/article/02-2019-04-054.html.

[19] Karpathy A, Toderici G, Shetty S, et al. Large-scale video classification with convolutional neural networks [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2014: 1725-1732.

[20] Du Tran, Bourdev L, Rob Fergus, et al. Learning spatiotemporal features with 3D convolutional networks [C]// Proc of IEEE International Conference on Computer Vision. Piscataway, NJ: IEEE Press, 2015: 4489-4497.

[21] Sun Lin, Jia Kui, Yeung D Y, et al. Human action recognition using factorized spatio-Temporal convolutional networks [C]// Proc of IEEE International Conference on Computer Vision. Piscataway, NJ: IEEE Press, 2015: 4597-4605.

[22] Liu Li, Shao Lin, Li Xuelong, et al. Learning spatio-temporal representations for action recognition: a genetic programming approach [J]. IEEE Trans on Cybernetics, 2016, 46 (1): 158-170.

[23] Wang Miao, Sun Jifeng, Yu Jialin, et al. Human action recognition based on feature level fusion and random projection [C]// Proc of the 5th International Conference on Computer Science and Network Technology. Piscataway, NJ: IEEE Press, 2016: 767-770.

[24] Feichtenhofer C, Pinz A, Zisserman A. Convolutional two-stream network fusion for video action recognition [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2016: 1933-1941.

[25] Zhang Wenyu. Research on belief function based decision fusion for wireless sensor networks [D]. Beijing: Beijing Jiaotong University, 2016.

[26] Deng Jia, Dong Wei, Socher R, et al. ImageNet: A large-scale hierarchical image database [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2009: 248-255.

[27] Pérez J S. TV-L1 optical flow estimation [J]. Image Processing on Line, 2013, 2 (4): 137-150.

[28] Yu Yunbo, Long Mingsheng, Wang Jianmin, et al. Spatiotemporal pyramid network for video action recognition [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2017: 2097-2106.

Note: Figure translations are in progress. See original paper for figures.