IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received March 13, 2019, accepted March 24, 2019, date of publication April 11, 2019, date of current version May 13, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2910604

A Spatiotemporal Heterogeneous Two-Stream Network for Action Recognition

ENQING CHEN[1,2], (Member, IEEE), XUE BAI[1,2], LEI GAO[3], (Member, IEEE), HARON CHWEYA TINEGA[1,2], AND YINGQIANG DING[1,2]

[1]School of Information Engineering, Zhengzhou University, Zhengzhou 450001, China
[2]Industrial Technology Research Institute, Zhengzhou University, Zhengzhou 450001, China
[3]Department of Electrical and Computer Engineering, Ryerson University, Toronto, ON M5B 2K3, Canada

Corresponding author: Yingqiang Ding (dyq@zzu.edu.cn)

This work was supported in part by the Program of NSFC, under Grant U1804152 and Grant 61331021.

ABSTRACT The method based on the two-stream networks has achieved great success in video action recognition. However, most existing methods employ the same structure for both spatial and temporal networks, leading to unsatisfied performance. In this paper, we propose a spatiotemporal heterogeneous two-stream network, which employs two different network structures for spatial and temporal information, respectively. Specifically, the Residual network (ResNet) and BN-Inception are utilized as the base networks to present the spatiotemporal characteristics of different human actions. In addition, a segmental architecture is employed to model long-range temporal structure over video sequences to better distinguish the similar actions owning sub-action sharing phenomenon. Moreover, combined with the strategy of data augment, a modified cross-modal pre-training strategy is proposed and applied to the spatiotemporal heterogeneous network to improve the final performance of human actions recognition. The experiments on UCF101 and HMDB51 datasets demonstrate the proposed spatiotemporal heterogeneous two-stream network outperforms the spatiotemporal isomorphic networks and other related methods.

INDEX TERMS Action recognition, spatiotemporal heterogeneous, two-stream networks, ResNet, long-range temporal structure, training strategies.

I. INTRODUCTION

As one of the most popular research directions in the field of computer vision, human action recognition has attracted extensive attentions from research and industrial communities. It plays an important role in video surveillance, behavior analysis, smart home, video retrieval, human-computer intelligent interaction, etc. However, human action recognition is still facing major challenges due to various limitations such as viewpoint changes, background clutter, and different illumination conditions. In recent years, the use of deep Convolutional Networks (ConvNets) [1] resulted in a tremendous breakthrough on image and speech recognition, in terms of performance. Computer vision researchers have since then sought to transfer the use of ConvNets to human action recognition [2]–[23].

Compared to the success of the image field, deep learning develops relatively slowly in the field of video based action recognition. There are two main reasons. First, compared with the massive image datasets, the scale and diversity of video data are not enough. Thus, it is difficult to build a large-scale tagged video database for training depth networks. Second, compared to 2D images, videos contain additional temporal information introducing more complex analysis works than images.

To solve the aforementioned challenges, video action recognition based on deep ConvNets made many attempts in recent years and achieved rapid development. Karpathy et al. [2] compared several ConvNet architectures for action recognition, and the corresponding training process was carried out on a very large Sports-1M dataset. Tran et al. [3] introduced a method based on the 3-dimensional convolutional network for action recognition. Simonyan et al. [4] proposed a method based on a two-stream network leading to better performance. Although these methods utilized temporal information in the video to some extent, they only paid attentions to short-term movement changes, without capturing long-range temporal information of the video. To address

The associate editor coordinating the review of this manuscript and approving it for publication was Shuhan Shen.

2169-3536 © 2019 IEEE. Translations and content mining are permitted for academic research only.
Personal use is also permitted, but republication/redistribution requires IEEE permission.
See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

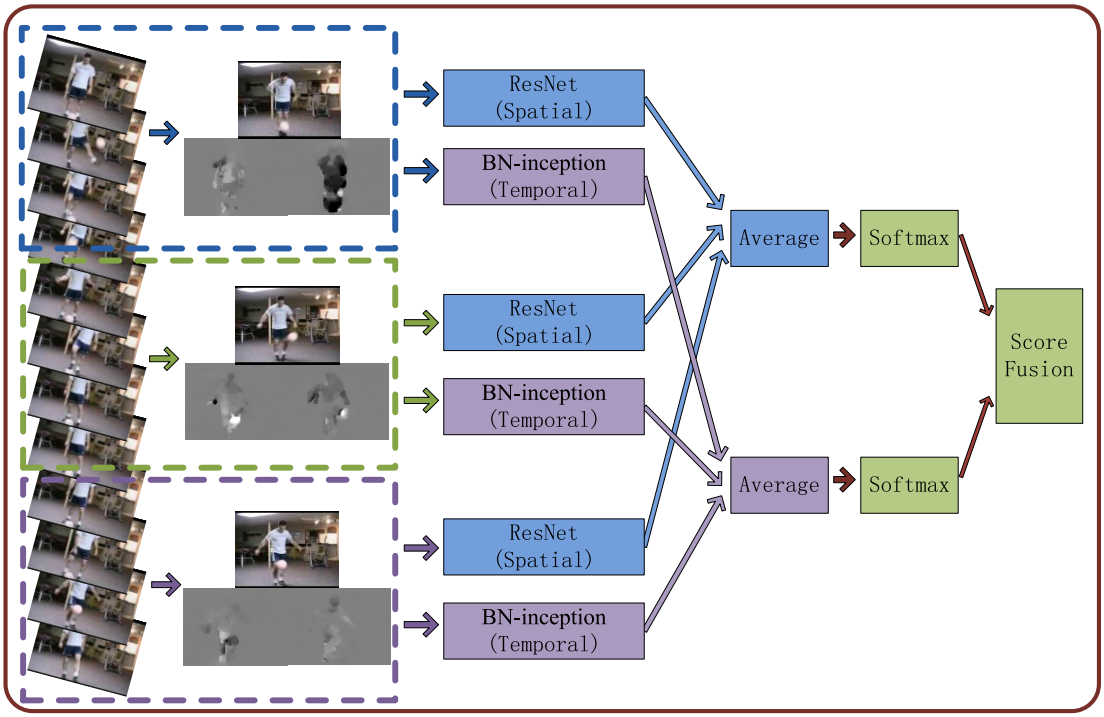

FIGURE 1. The spatiotemporal heterogeneous two-stream network based on long-range temporal structure modeling.

this issue, Wang et al. [5] proposed a Temporal Segment Network (TSN) to extract the long-range temporal information from the video data.

In this paper, we propose a network architecture based on a two-stream network for human action recognition. Inspired by the two-stream hypothesis [4], the two-stream network is designed with two identical stream structure to extract the spatial and temporal information of a video, respectively. Subsequently, the extracted information from the two streams are fused using the late fusion strategy [4] and each stream is implemented by an independent ConvNet. Specifically, the spatial stream works on a single RGB image, and the temporal stream takes a stack of consecutive optical flow fields as inputs. For the previous approaches based on two-stream networks [4]–[6], [16]–[19], both the spatial and temporal networks share the same network structure. Nevertheless, since human recognizing and understanding of appearance and motion are two completely different processes, a satisfied design of spatial and temporal networks should not be identical. To solve this issue, in this paper, a spatiotemporal heterogeneous two-stream network is proposed for human action recognition. Also, to extract the long-range temporal information from video sequences, the idea of video segmentation [5] is introduced to the proposed spatiotemporal heterogeneous network. The network architecture is shown in Fig. 1 on the next page. From Fig. 1, at first, a video is divided into three segments. Then the short snippets are randomly sampled form the corresponding segment, which will be utilized in the spatiotemporal heterogeneous two-stream network to capture the initial action category scores. The corresponding initial category scores are fused to obtain the final score, which is a video-level prediction. In addition, modified cross-modal pre-training and data augmentation technologies are also adopted in this architecture. It can exploit the massive tagged image data and avoid the limitation of the dataset in sample size and label noise. During the experiments, it is observed that the performance of the spatiotemporal heterogeneous two-stream network is better than the spatiotemporal isomorphic network.

The contributions of this paper are summarized as follows.

(a) A novel spatiotemporal heterogeneous two-stream architecture is proposed. The performance of spatiotemporal heterogeneous networks and spatiotemporal isomorphic networks is explored through related experiments. The appropriate network structures are obtained for spatial and temporal networks.

(b) Residual network (ResNet) and BN-Inception are introduced to the proposed spatiotemporal heterogeneous two-stream network to extract more spatiotemporal features.

(c) The strategy of video segmentation processing in TSN is introduced, which can leverage the long-range temporal information in the video.

(d) The modified cross-modal pre-training and data enhancement techniques are adopted to enable spatial and temporal networks to make the best use of the vast amount of tagged image data, avoiding sample size limitations.

Experimental results on standard video datasets UCF101 and HMDB51 demonstrate that the proposed method is superior to state-of-the-art methods proposed in [4]–[6].

II. RELATED WORKS

Recently, deep ConvNets have been widely used in the fields of speech recognition and image recognition, achieving competitive performance. Computer vision researchers seek to transfer the success of these deep ConNets from speech and image recognition to video action recognition. In deep learning, video action recognition based methods are mainly divided into three categories: 1) Action recognition based on the 3D convolutional network. 2) Action recognition based on the recurrent neural network. 3) Action recognition based on the two-stream network.

A. 3D CONVOLUTIONAL NETWORK

Tran et al. [3] constructed a network using 3D convolution and 3D Pooling to extend the convolution kernel to the time domain. The convolution is performed simultaneously in both the spatial domain and the temporal domain. Although both the spatial and temporal domains are taken into account by this method, the related computational cost and model storage are too huge. Qiu et al. [7] reformed the 3D convolution and proposed a Pseudo-3D Residual Net (P3D ResNet). They replaced $3 \times 3 \times 3$ convolutions with $1 \times 3 \times 3$ convolution filters and $3 \times 1 \times 1$ convolutions filters. The former is used to obtain the characteristics of the spatial dimension, and the latter is used to capture the characteristics of the temporal dimension information. Different from traditional 3D convolutional networks, this architecture successfully improves the performance on video recognition tasks. Diba et al. [8] proposed a ''Temporal 3D ConvNet'' (T3D) for action recognition which employed a 3D DesnseNet-based architecture and a new temporal layer ''Temporal Transition Layer'' (TTL) to simulate variable time convolution kernel depth. However, this method is challenged on the ground that it is only evaluated on several RGB frames, and ignored the temporal information which is crucial in the video analysis.

B. RECURRENT NEURAL NETWORKS

Motivated by the recent success of LSTM in sequence modeling [9], [10], more and more research is trying to employ LSTM in video action recognition. Donahue et al. [11] proposed a Long-term Recurrent Convolutional Neural Network (LRCN) which uses the CNN to extract the spatial characteristics and then feeds the characteristics into the LSTM network for extracting time information. Differential LSTM [12] added a new gating into LSTM to keep track of the derivatives of the memory states to discover patterns within salient motion patterns. Other researchers have applied the LSTM to motion analysis using human skeletal data. For instance, Zhu et al. [13] proposed a mixed-norm regularization term to the cost function of the deep LSTM network to introduce the co-occurrence characteristics of the joints in the actions into the LSTM network. Liu et al. [14] focused on the adjacent joints in the skeleton, which divided the body into smaller parts than the previous work. Later, a tree-based traversal approach is used to extend LSTM to the spatiotemporal domain. However, due to the additional parameters of LSTM and the differences between video and speech [15], LSTM has not yet shown its good capacity in video action recognition.

C. THE TWO-STREAM NETWORK

Simonyan et al. [4] proposed a two-stream network, including RGB and optical flow channels together. It operates on the single RGB image and continuous optical flow frames respectively with the softmax scores from the temporal and spatial stream fused in the end. Wang et al. [16] proposed Trajectory-Pooled Deep-Convolutional Descriptors (TDDs), which combines two-stream network and trajectory features. It is a successful example of combining deep neural network and shallow local features. Feichtenhofer et al. [17] improved the two-stream network by exploring better methods to fuse spatial and temporal streams. They found that combining the two streams in the convolutional layer can better simulate the correlation of spatial and temporal stream rather than average the score of softmax layer. Wang et al. [5] introduced the Temporal Segmentation Network and further improved the performance by using multi-modality input. Wang et al. [18] proposed a novel spatiotemporal pyramid network, which combined the spatial and temporal features by introducing spatiotemporal compact bilinear operator and hierarchical fusion strategies. Feichtenhofer et al. [6] combined the two-stream network with the ResNet and extended ResNet from two-dimension to three-dimension. Furthermore, the residual connection is introduced from the motion stream to the appearance stream, which increases the interaction between two streams. Ji et al. [19] proposed an end-to-end architecture to handle with the joint actor-action semantic segmentation problem in a video. In terms of the part of action recognition, they employ a two-stream network with temporal aggregation. From the aforementioned algorithms, the two-stream network is efficient to be applied to video action recognition especially with limited training data.

Furthermore, there are some other types of methods besides the above three types of methods. For example, Zhu W et al. [20] improved the accuracy of action recognition by mining the key volume in the video. Sharma et al. [21] proposed the Visual Attention Model, which introduced the idea of attention mechanism into action recognition. Sengupta et al. [22] introduced a pillar network for action recognition by combining a four-stream convolutional neural network (2 ResNet and 2 Inception networks) with multi-kernels based support-vector-machines (SVM). Furthermore, Sengupta et al. [23] improved the pillar network, and proposed pillar networks++ by using a Gaussian Process classifier to improve the accuracy of action recognition. In this paper, our work is built with a spatiotemporal heterogeneous two-stream network, and some improvements are introduced to take advantages of longer-range temporal information and richer spatial information in the video.

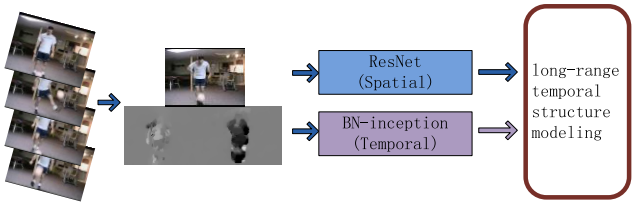

FIGURE 2. Spatiotemporal heterogeneous two-stream network.

III. TECHNICAL APPROACH

In this section, the proposed network architecture is presented. First, the spatiotemporal heterogeneous two-stream network architecture is proposed. Then the Residual network (ResNet) and BN-Inception are reviewed and introduced to the proposed spatiotemporal heterogeneous two-stream network as the base networks. After that, the idea of video segmentation in TSN is introduced to the spatiotemporal heterogeneous two-stream network to model the long-range temporal structure of video sequences. Finally, training strategies for optimizing the network are given.

A. SPATIOTEMPORAL HETEROGENEOUS TWO-STREAM NETWORKS

1) TWO-STREAM ARCHITECTURE

It is known that videos can be represented by spatial and temporal components, respectively. According to the two-stream hypothesis [24], the human visual cortex contains two channels of information: the ventral stream (corresponding to object recognition) and the dorsal stream (corresponding to motion recognition). Simonyan et al. [4] proposed a two-stream method to extract the spatial and temporal/motion information from a video. For the spatial information, it can be conveyed through RGB frames. Motion information, which is the changes between adjacent frames, can be expressed in the form of dense optical flow. Correspondingly, each stream is processed with an independent and identical deep ConvNet. Specifically, while the spatial stream is implemented with a single RGB image, the input of the temporal stream is the continuous optical flow frames with horizontal and vertical directions. The length (L) of the flow frames is fixed to 10 frames such that the total length of the flow frames with horizontal and vertical directions is set to $2L = 20$ frames [4]. Then, two streams are trained separately, and the scores of the softmax layer in two streams are fused together to obtain the final classification decision. Two different fusion methods including direct averaging and SVM are considered in [4] to handle the scores from the softmax levels.

In this subsection, a spatiotemporal heterogeneous two-stream network is proposed on the basis of the two-stream network. In the spatiotemporal heterogeneous network, different network architectures are applied to spatial and temporal streams shown in Fig. 2. From Fig. 2, there are two motivations for designing spatiotemporal heterogeneous two-stream networks. 1) when the spatial and temporal streams in the two-stream network share the same network structure (spatiotemporal isomorphism), a great number of redundant information will be generated when the two streams are merged. 2) since human recognizing and understanding of appearance and movement are two completely different processes, the network structure for spatial and temporal should be different. Through a great number of experiments, it is observed that the performance of spatiotemporal heterogeneous networks is better than spatiotemporal isomorphic networks.

B. BASE NETWORK

In section III.A, we have established a spatiotemporal heterogeneous two-stream network. It is known that a good video representation network should be able to extract more discriminant spatial and temporal information. Previous studies [25], [26] have shown that deeper CNNs can extract more discriminant information. Yu et al. [26] used deconvolution to realize feature visualization and made related analysis on visual information stored in different layers. Through comparison of CNNs with different depths, they found that deeper CNN is better at extracting the discriminant information, which improves the prediction performance. Furthermore, recent studies [27], [28] show that the depth of the network plays a vital role in visual representation. As a new deep neural networks model, Residual networks (ResNet) [25], [29] solves the problem of degradation [30] caused by the deepening of network layers effectively. Therefore, to explore the potential of the spatiotemporal heterogeneous two-stream network to the greatest extent, ResNet is introduced to the proposed spatiotemporal heterogeneous two-stream network as the base network to extract spatial and temporal features. Moreover, the BN-Inception network which improves network performance by increasing network depth and width is employed as the base network. In what follows, Residual networks and BN-Inception will be presented, respectively.

1) RESIDUAL NETWORKS

To extract more discriminant information, ResNet [25], [29] with deep layers is employed as one of the base networks. As the number of network layers deepens, the degradation problem will occur. To solve this problem, He et al. [25] proposed the residual network. Instead of directly fitting a desired underlying mapping $H(x)$, they trained a deep network by fitting a residual mapping $F(x) := H(x) - x$ [25]. The residual units are defined as [25]:

$$x_{l+1} = \sigma(x_l + F(x_l; W_l)). \tag{1}$$

where x_l and x_{l+1} are the input and output of the l-th layer. $F(x_l; W_l)$ is a nonlinear residual mapping represented by convolutional filter weights $W_l = \{W_{l,k}|_{1<k<K}\}$ with $K \in \{2, 3\}$, and σ represents the ReLU function [31]. The main advantage of the residual unit is that the shortcut connection allows the signal to propagate directly from the first layer to any layers in the network, breaking the convention that the output of (n-1)-th layer of traditional neural networks can

only be used as input for the n-th layer. Therefore, the gradient can skip the intermediate layer and propagate directly from the loss layer to any shallow layer, avoiding the problem of gradient explosion and disappearance. More importantly, the shortcut connection does not introduce extra parameters and computational complexity at all. ResNet adopts batch normalization (BN) [32] after each convolution but before activation layer. It not only solves the problem of covariate shift but also speeds up the convergence of the network [25]. At the ending of the network, a global average pooling layer and softmax layer are adopted together instead of the fully connected layers plus softmax. It reduces the number of parameters effectively. In addition, the algorithm utilizes the bottleneck structure to reduce the computational complexity and guarantee the final performance.

2) BN-INCEPTION

The BN-Inception [32] network is also employed as the base network, which is an improvement on GoogleLeNet [28]. Unlike the traditional ConvNet, the biggest change of GoogleLeNet is to put forward the Inception structure and construct the network through the superposition of Inception modules. It improves the network performance by increasing width and depth of the network. Sergey et al. [32] further improved the architecture of GoogleLeNet and proposed BN-Inception network. The main contribution of the BN-Inception is to replace the 5×5 convolution in the previous inception module with two 3×3 convolutions. It not only reduces the number of parameters but also establishes more nonlinear transformations, enhancing the network's ability to learn features. On the other hand, by adding the BN layer, the output of each layer is normalized to the normal distribution of N(0,1), reducing the Internal Covariate Shift.

Since ResNet can extract more features by increasing the number of layers [25] and the BN-Inception network can improve network performance by increasing depth and width [32], ResNet and BN-Inception are taken as the base network of spatiotemporal heterogeneous two-stream network in this paper. Then the appropriate network structure is achieved for the temporal and spatial networks through experiments to maximize the potential of the spatiotemporal heterogeneous two-stream network. For the original two-stream network [4], the base network is VGG-M-2048 [33], and the two streams share the same network structure. Feichtenhofer et al. [17] replaced VGG-M-2048 with VGG-16 [27], which further improved the performance. To extract more spatial and temporal information, ResNet and BN-Inception are taken as the base network to construct a deeper heterogeneous two-stream network in this paper. Compared with VGG network, ResNet has fewer filters and lower computational complexity. Although the depth of the ResNet is increased, the computational complexity of the ResNet-152 (11.3 billion FLOPs) is still less than that of VGG-16 (15.3 billion FLOPs) and VGG-19 (19.6 billion FLOPs). The computational complexity of ResNet-50 and ResNet-101 is only 3.8 billion FLOPs and 7.6 billion FLOPs [25]. Besides, compared with the 3D two-stream residual network in [6], although the residual network used in this paper is two-dimensional, the performance is comparable to that in [6], and the number of parameters is less than the number in the 3D residual network. Specifically, the total number of model parameters of the ST-ResNet∗ model (using 50-layer ResNet) trained on split1 of UCF101 is 234M [6] while the parameters of the spatial and temporal network are 182M in this paper.

C. MODELING LONG-RANGE TEMPORAL STRUCTURE

For the general two-stream network, there is an obvious challenge that it is difficult to model a long-range temporal structure. The main reason is that the traditional two-stream network [4] only works on a single frame (spatial network) or a stack of frames (temporal network). Therefore, it cannot extract the long-range temporal information effectively. Nevertheless, it is known that the long-range temporal information in the video plays more important roles for action recognition [34], [35]. For example, basketball shooting and basketball dunk may be similar to each other in a short time, but existing considerable differences in a long-range period. Therefore, it will result in misjudgments only with a small piece of a video, leading to unsatisfied performance. Inspired by the TSN [5], the idea of video segmentation is adopted to improve the performance of the proposed spatiotemporal heterogeneous two-stream network and extract the long-range temporal information from video sequences.

First, a video is divided into three segments at equal intervals [5], expressed as $\{y_1, y_2, y_3\}$. Then, we randomly sample short snippets $\{x_1, x_2, x_3\}$ from the corresponding segment, and the short snippets are sent to the spatiotemporal heterogeneous two-stream network to obtain the initial action category scores. After that, the initial category scores are fused by averaging to obtain category consensus among short snippets. Finally, based on this consensus, the softmax function is utilized to predict the probability to each category. The modeling of the short snippets is written as follows,

$$f(x_1, x_2, x_3) = S(G(F(x_1; \boldsymbol{W}), F(x_2; \boldsymbol{W}), F(x_3; \boldsymbol{W}))). \quad (2)$$

where $F(x_1; \boldsymbol{W})$ represents a ConvNet function with a parameter$W$, with G indicating averaging, and S being the softmax function.

The final loss function of category consensus $\mathbf{g} = G(F(x_1; \mathbf{W}), F(x_2; \mathbf{W}), F(x_3; \mathbf{W}))$ is defined as follows,

$$L(y, \mathbf{g}) = -\sum_{i=1}^{n} y_i (g_i - \log \sum_{j=1}^{n} \exp g_j). \quad (3)$$

where n denotes the number of action categories and y_i denotes the real label of class i. $g_i = G(F_i(x_1; \mathbf{W}), F_i(x_2; \mathbf{W}), F_i(x_3; \mathbf{W}))$ is the category consensus score of class i, which is obtained by averaging the scores of the same category for the three short snippets. In this paper, three short snippets are used to jointly optimize the model parameter $\mathbf{W}$. In the process of back-propagation, the gradient of $\mathbf{W}$ with respect

to the loss value L can be derived as follows,

$$\frac{\partial L(y, \mathbf{g})}{\partial \mathbf{W}} = \frac{\partial L}{\partial \mathbf{g}} \sum_{k=1}^{3} \frac{\partial G}{\partial F(x_k)} \frac{\partial F(x_k)}{\partial \mathbf{W}}. \quad (4)$$

Then, the related model parameters are learned by mini-batch stochastic gradient descent. As can be seen in Eq. 4, the category consensus **g** of three short snippets is employed to update the parameters. With this optimization method, the video-level model parameters can be learned to obtain the long-range temporal information.

D. NETWORK TRAINING STRATEGIES

Since the sample size of datasets used for action recognition is too small and the label noise is too large, it easily leads to over-fitting problem when training deep ConvNets. In this paper, a modified cross-modal pre-training and data enhancement strategies are employed together to solve the aforementioned challenges in video action recognition.

1) MODIFIED CROSS-MODAL PRE-TRAINING

Previous works [4], [5], [36] show that using the pre-training model on ImageNet [37] to initialize the deep ConvNet with insufficient dataset can improve the accuracy and accelerate convergence effectively. Since the spatial network takes RGB images as input, the pre-training model on ImageNet [37] can be adopted directly to initialize the spatial network. However, since the input of the temporal network is a stacked optical flow frame. The distribution of optical flow and RGB images and the content expressed by them exhibits a wide gap. Therefore, the temporal network cannot directly be pre-trained on the image dataset. In this paper, we put forward a modified cross-modal pre-training to initialize the temporal network. The detailed steps are described as follows. First, the optical flow field is mapped by a linear transformation [5] so that the range of the optical flow field is the same as the RGB image. Because the RGB image has three channels, but the inputs of temporal network in this paper are 10 optical flow frames (including horizontal and vertical directions). To adapt to the number of input channels in the temporal network, the weight of the first convolutional layer of the RGB model needs to be modified. Previous methods mostly average the weight of RGB channel and then copy the weight to make it adapt to the number of channels in the temporal network. Different from previous methods, the modified cross-modal pre-training in this paper firstly trains the temporal network from scratch and then takes the weight of its first convolutional layer to replace the first convolutional layer weight of the RGB model. Finally, the modified weights are used to initialize the temporal network. The comparison is given in TABLE 1.

TABLE 1. Comparison between modified cross-modal pre-training and the method in [5].

Question Description	**Spatial Network**: The input is an RGB frame (possessing three channels), and the kernel size of the first convolution layer Conv1 is (64, 3, 7, 7). **Temporal Network**: The input in this paper is 2L=10 optical flow frames (including x, y directions), and the kernel_size of the first convolution layer Conv1 is (64, 10, 7, 7). **The problem to be solved**: Spatial networks can be initialized directly using the pre-training model on ImageNet. To enable the temporal network to use the pre-training model, the first convolution layer conv1 of the pre-training model needs to be modified to fit the dimension of the first convolution layer conv1 of the temporal network.	
Method	The method in [5]	The proposed method
Step1	Average the parameters of the 3 channels of conv1	Fine tune the parameters of conv1, and obtain the kernel_size (64,10,7,7)
Step2	The average value is saved to the new convolution kernel conv1*, and the kernel_size is expanded to (64,10,7,7).	The conv1 parameters obtained from the fine-tuning are saved to the new convolution kernel conv1*, and the final kernel_size is (64,10,7,7).

2) DATA AUGMENTATION

When the size of the dataset is inadequate, the data augmentation technology is often employed to increase the diversity of samples, preventing over-fitting. Horizontal flipping, corner cropping, and scale jittering [27] are used to expand the dataset. For corner cropping, we extract the area from the corner or center of the image, thus avoiding the default focus on the center of the image. Scale jittering fixes the size of the input image or optical flow field as 256×340. The width and height of the cropped area are randomly selected from {256, 224, 192, 168}. Finally, these cropped areas are resized to 224×224 for training. Random horizontal flip is used in all our training steps.

IV. EXPERIMENTS

In this section, we first briefly introduce the datasets in our experiments. Then, the performance of spatiotemporal isomorphism and spatiotemporal heterogeneous networks is evaluated, respectively. After that, effectiveness of the modified cross-modal pre-training strategy proposed in III.D is verified. Experimental results and analysis are given in the last subsection.

A. DATASETS

To verify the effectiveness of the proposed approach, the network architecture is evaluated on UCF101 [38] and HMDB51 [39] datasets. UCF101 and HMDB51 are two challenging action datasets, which contain several complex actions. There are many uncontrolled scene changes in the videos. The UCF101 dataset is composed of fully annotated video clips from YouTube, including 101 types of actions and 13320 video clips. Each video clip lasts 3-10 seconds with an average of 100-300 frames. The HMDB51 dataset comprises 6766 video clips, which covers 51 action categories. All the videos are collected from various sources, mostly from movies, and a small part from YouTube and Google videos. For both datasets, we follow the provided evaluation protocol and use standard training/testing splits. We first explore and evaluate our proposed architecture on split1 of the UCF-101

TABLE 2. The performance of spatiotemporal heterogeneous and spatiotemporal isomorphism networks on UCF101 (split1).

Network Architectures	Spatial	Temporal	Two-Stream
spatio_ResNet50+temporal_ResNet50	83.6%	84.6%	92.9%
spatio_ResNet101+temporal_ResNet50	83.4%	84.6%	93.6%
spatio_ResNet152+temporal_ResNet50	85.5%	84.6%	93.1%
spatio_ResNet101+temporal_BN-Inception	83.4%	88.1%	94.3%

dataset. To compare with state-of-the- art methods, the average recognition accuracy on three splits of both UCF-101 and HMDB-51 are reported.

B. SETTING UP

1) TRAINING

The mini-batch stochastic gradient descent algorithm is used to learn the weights of the network. The batch size is set to 256 and the L2 norm of the gradient is limited to 40. In addition, the weight decay and the momentum are set to 0.0005 and 0.9, respectively. The network weights are initialized by pre-training on ImageNet, and the temporal network is initialized by the method described in Section III. C. For the spatial network, the base learning rate is set to 0.0001. The learning rate is reduced by 10 times in every 15000 iterations, and the training stops at 36000 iterations. In terms of the time network, the learning rate is initialized to 0.0001, which is decreased by 10 times after 20000 and 32000 iterations. The maximum iteration is set to 40000. The method of [40] is used to calculate the optical flows, and all experiments are implemented based on the Caffe platform [41].

2) TESTING

During the testing, we followed the test scheme of the original two-stream ConvNets [4]. 25 RGB frames or optical flow stacks are sampled from the action video in an equal time interval. For every sampled frame, 10 ConvNet inputs are obtained by cropping four corners, one center, and their horizontal flipping. In this paper, the weighted averaging is used to fuse spatial and temporal networks. The weight ratio of the spatial network and the temporal network is set to 1:1.5, which refers to the setting of weight ratio in [5]. Moreover, we find that 1:1.5 is the best weight ratio by experiments.

The related codes have been released at https://github.com/baixuexue/Spatiotemporal_Heterogeneous_Two-stream_Network.

C. EXPLORATION AND EVALUATION

In this subsection, we first evaluate the proposed network structure and compare the performance of spatiotemporal heterogeneous with spatiotemporal isomorphic networks. Then, the modified cross-modal pre-training strategy proposed in III.D is evaluated in experiments, demonstrating the effectiveness of the proposed method. All experiments in this subsection are performed on UCF-101 (split1) dataset.

This paper divides the spatiotemporal heterogeneity into two types, including the same type networks of different depths, and the different types of networks. ResNet-50, ResNet-101, ResNet-152 [25] and BN-Inception [32] are employed for the testing. Experimental results are tabulated in TABLE 2. The performance on three different network architectures is compared as following: (1) Spatial and temporal networks with the same structure. (2) Spatial and temporal networks with different depths networks but the same structure. (3) Spatial and temporal networks with different network structures. During experiments, we found that the performance of temporal network using ResNet-50 is better than ResNet-101 and ResNet-152. From TABLE 2, it is observed that the performance of spatial and temporal networks with the same structure but different depths is better than the spatiotemporal isomorphic network. From the fused results of two-stream, ResNet-101 is the best choice for the spatial network. When the ResNet-101 is chosen as a spatial network and a different structure BN-Inception is selected as a temporal network, it achieves 94.31% accuracy on split1 of UCF101. Experiments show that the performance of spatiotemporal heterogeneous networks is better than that of spatiotemporal isomorphic networks.

TABLE 3. Evaluation of different training strategies for temporal network on the UCF101 dataset (split1).

Training strategy	ResNet50	BN-Inception
From Scratch	79.8%	81.7%
Pre-training strategy in [5]	83.4%	86.6%
Modified cross-modal pre-training (ours)	84.6%	88.1%

To demonstrate the effectiveness of the modified cross-modal pre-training strategy proposed in III.D, experiments are carried out on ResNet50 and BN-Inception respectively with different strategies. Specifically, three cases are presented as follows: (1) training temporal network from scratch (2) using the method in [5] to train temporal network (3) using our modified cross-modal pre-training strategy to train temporal network. In this paper, we train the temporal network in these three cases, and the accuracy of the temporal network is reported, respectively. The experiments are performed on UCF-101 (split1) dataset and the results are summarized in TABLE 3. From TABLE 3, the strategy of using pre-training to initialize a deep ConvNet can bring great accuracy improvement than training from scratch. In addition, by comparing the last two rows of TABLE 3, it can be found that the strategy of using the modified cross-modal pre-training method can improve the accuracy by 1.2% and 1.5%

TABLE 4. recognition accuracy over 3 splits of UCF101 and HMDB51.

Method	UCF101	HMDB51
Improved dense trajectories(IDT)[42]	85.9%	57.2%
Two-Stream[4]	88.0%	59.4%
Two-stream Fusion [17]	92.5%	65.4%
TDD[16]	90.3%	63.2%
RNN+FV[43]	88.0%	54.3%
C3D [3]	85.2%	-
ActionVLAD [44]	92.7%	66.9%
Factorized ConvNet [45]	88.1%	59.1%
P3D ResNet + IDT[7]	93.7%	-
TSN (2 modalities)[5]	94.0%	68.5%
T3D+TSN[8]	93.2%	63.5%
ST-ResNet*[6]	93.4%	66.4%
Ours（spatiotemporal isomorphism）	93.9%	65.3%
Ours（spatiotemporal heterogeneity）	94.4%	67.2%

on ResNet50 and BN-Inception in [5], respectively. Experimental results demonstrate the effectiveness of the modified cross-modal pre-training strategy proposed in this paper.

D. COMPARISON WITH STATE-OF-THE-ART

In this subsection, the optimal accuracy obtained in the experiment is compared with state-of-the-art methods. The average recognition accuracies over the three splits of UCF-101 and HMDB-51 are reported, respectively. The experimental results are summarized in TABLE 4. From TABLE 4, for UCF 101 dataset, the proposed method outperforms other state-of-the-art methods. Moreover, compared with the original two-stream method [4] and ST-ResNet* [6], the accuracy has been improved by 6.4% and 1.0%, respectively. For HMDB51 dataset, the proposed method is also competitive to the optimal method. Again, compared with the original two-stream method [4] and ST-ResNet* [6], the accuracy has been improved by 7.8% and 0.8%, respectively.

Experimental results demonstrate the effectiveness of our proposed spatiotemporal heterogeneous two-stream network based on long-range temporal structure modeling. More importantly, the performance of the spatiotemporal heterogeneous two-stream network is improved to a certain extent compared with spatiotemporal isomorphism two-stream network, improved by 0.5% and 1.9% on UCF101 and HMDB51, respectively.

V. CONCLUSION

In this paper, a spatiotemporal heterogeneous two-stream network for human motion recognition is proposed. Since human recognizing and understanding of appearance and motion are two completely different processes, we improve the existing two-stream method and design different network structures to extract spatial and temporal information. The performance of spatiotemporal isomorphic and spatiotemporal heterogeneous two-stream networks is explored through experiments. During the experiments, it is found that the performance of spatiotemporal heterogeneous networks is better than that of spatiotemporal isomorphic networks. To release the potential of the spatiotemporal heterogeneous network to a larger extent, ResNets and BN-Inception are employed as the basic networks, extracting more appearance and motion characteristics. Furthermore, a long-range temporal structure is built to extract long-range temporal information of the video. Finally, a modified cross-modal pre-training and data augment strategies are employed to improve the final performance. Throughout the end-to-end training, the system has brought performance improvement on both HMDB51 and UCF101 datasets.

REFERENCES

[1] Y. LeCun *et al.*, "Backpropagation applied to handwritten zip code recognition," *Neural Comput.*, vol. 1, no. 4, pp. 541–551, 1989.
[2] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1725–1732.
[3] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3D convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 4489–4497.
[4] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 568–576.
[5] L. Wang *et al.*, "Temporal segment networks: Towards good practices for deep action recognition," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, Sep. 2016, pp. 20–36.
[6] C. Feichtenhofer, A. Pinz, and R. Wildes, "Spatiotemporal residual networks for video action recognition," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 3468–3476.
[7] Z. Qiu, T. Yao, and T. Mei, "Learning spatio-temporal representation with pseudo-3D residual networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 5533–5541.
[8] A. Diba *et al.* (Nov. 2017). "Temporal 3D ConvNets: New architecture and transfer learning for video classification." [Online]. Available: https://arxiv.org/abs/1711.08200?context=cs
[9] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. (Apr. 2015). "Show and tell: A neural image caption generator." [Online]. Available: https://arxiv.org/abs/1411.4555
[10] A. Graves, N. Jaitly, and A.-R. Mohamed, "Hybrid speech recognition with deep bidirectional LSTM," in *Proc. IEEE Workshop Autom. Speech Recognit. Understand.*, Dec. 2013, pp. 273–278.
[11] J. Donahue *et al.*, "Long-term recurrent convolutional networks for visual recognition and description," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 2625–2634.
[12] V. Veeriah, N. Zhuang, and G.-J. Qi, "Differential recurrent neural networks for action recognition," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 4041–4049.
[13] W. Zhu *et al.*, "Co-occurrence feature learning for skeleton based action recognition using regularized deep LSTM networks," in *Proc. AAAI*, 2016, vol. 2, no. 5, p. 6.
[14] J. Liu, A. Shahroudy, D. Xu, and G. Wang, "Spatio-temporal LSTM with trust gates for 3D human action recognition," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, Sep. 2016, pp. 816–833.
[15] C.-Y. Ma, M.-H. Chen, Z. Kira, and G. AlRegib. (Mar. 2017). "TS-LSTM and temporal-inception: Exploiting spatiotemporal dynamics for activity recognition." [Online]. Available: https://arxiv.org/abs/1703.10667
[16] L. Wang, Y. Qiao, and X. Tang, "Action recognition with trajectory-pooled deep-convolutional descriptors," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 4305–4314.
[17] C. Feichtenhofer, A. Pinz, and A. Zisserman, "Convolutional two-stream network fusion for video action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 1933–1941.
[18] Y. Wang, M. Long, J. Wang, and P. S. Yu, "Spatiotemporal pyramid network for video action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 1529–1538.

[19] J. Ji, S. Buch, A. Soto, and J. C. Niebles, "End-to-end joint semantic segmentation of actors and actions in video," in *Proc. Eur. Conf. Comput. Vis.*, Sep. 2018, pp. 702–717.
[20] W. Zhu, J. Hu, G. Sun, X. Cao, and Y. Qiao, "A key volume mining deep framework for action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 1991–1999.
[21] S. Sharma, R. Kiros, and R. Salakhutdinov. (Feb. 2016). "Action recognition using visual attention." [Online]. Available: https://arxiv.org/abs/1511.04119
[22] B. Sengupta and Y. Qian. (Jul. 2017). "Pillar networks for action recognition." [Online]. Available: https://arxiv.org/abs/1707.06923v1
[23] B. Sengupta and Y. Qian. (Aug. 2017). "Pillar networks++: Distributed non-parametric deep and wide networks." [Online]. Available: https://arxiv.org/abs/1708.06250
[24] M. A. Goodale and A. D. Milner, "Separate visual pathways for perception and action," *Trends Neurosci.*, vol. 15, no. 1, pp. 20–25, 1992.
[25] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 770–778.
[26] W. Yu, K. Yang, Y. Bai, T. Xiao, H. Yao, and Y. Rui, "Visualizing and comparing AlexNet and VGG using deconvolutional layers," in *Proc. 33rd Int. Conf. Mach. Learn.*. Jun. 2016, pp. 1–7.
[27] K. Simonyan and A. Zisserman. (Sep. 2014). "Very deep convolutional networks for large-scale image recognition." [Online]. Available: https://arxiv.org/abs/1409.1556
[28] C. Szegedy *et al.* (Sep. 2014). "Going deeper with convolutions." [Online]. Available: https://arxiv.org/abs/1409.4842
[29] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, Sep. 2016, pp. 630–645.
[30] R. K. Srivastava, K. Greff, and J. Schmidhuber. (May 2015). "Highway networks." [Online]. Available: https://arxiv.org/abs/1505.00387
[31] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks." in *Proc. Adv. Neural Inf. Process. Syst.* 2012, pp. 1097–1105.
[32] S. Ioffe and C. Szegedy. (Feb. 2015). "Batch normalization: Accelerating deep network training by reducing internal covariate shift." [Online]. Available: https://arxiv.org/abs/1502.03167
[33] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman. (May 2014). "Return of the devil in the details: Delving deep into convolutional nets." [Online]. Available: https://arxiv.org/abs/1405.3531
[34] J. C. Niebles, C.-W. Chen, and L. Fei-Fei, "Modeling temporal structure of decomposable motion segments for activity classification," in *Proc. Eur. Conf. Comput. Vis.* Berlin, Germany: Springer, 2010, pp. 392–405.
[35] L. Wang, Y. Qiao, and X. Tang, "Latent hierarchical model of temporal structure for complex activity classification," *IEEE Trans. Image Process.*, vol. 23, no. 2, pp. 810–822, Feb. 2014.
[36] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 10, pp. 1345–1359, Oct. 2010.
[37] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.
[38] K. Soomro, A. R. Zamir, and M. Shah. (Dec. 2012). "UCF101: A dataset of 101 human actions classes from videos in the wild." [Online]. Available: https://arxiv.org/abs/1212.0402
[39] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, "HMDB: A large video database for human motion recognition," in *Proc. Int. Conf. Comput. Vis.*, Nov. 2011, pp. 2556–2563.
[40] C. Zach, T. Pock, and H. Bischof, "A duality based approach for real-time TV-L^1 optical flow," in *Proc. Joint Pattern Recognit. Symp.* Berlin, Germany: Springer, Sep. 2007, pp. 214–223.
[41] Y. Jia *et al.*, "Caffe: Convolutional architecture for fast feature embedding," in *Proc. 22nd ACM Int. Conf. Multimedia*, Nov. 2014, pp. 675–678.
[42] H. Wang and C. Schmid, "Action recognition with improved trajectories," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 3551–3558.
[43] G. Lev, G. Sadeh, B. Klein, and L. Wolf, "RNN Fisher vectors for action recognition and image annotation," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, Sep. 2016, pp. 833–850.
[44] R. Girdhar, D. Ramanan, A. Gupta, J. Sivic, and B. Russell, "ActionVLAD: Learning spatio-temporal aggregation for action classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 971–980.
[45] L. Sun, K. Jia, D.-Y. Yeung, and B. E. Shi, "Human action recognition using factorized spatio-temporal convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 4597–4605.

ENQING CHEN (S'06–M'11) received the B.E. and M.S. degrees from Zhengzhou University, in 2000 and 2003, respectively, and the Ph.D. degree in communication and information system from the Beijing Institute of Technology, China, in 2007. Since 2007, he has been with the School of Information Engineering, Zhengzhou University, where he is currently a Professor. In 2015, he was a Visiting Scholar with the Department of Electrical and Computer Engineering, Ryerson University, Canada. His research interests are in the areas of machine learning and signal processing, including image processing, computer vision, multimedia signal processing, and optimization algorithms.

XUE BAI received the bachelor's degree in communication engineering from the South-Central University for Nationalities, in 2014. She is currently pursuing the master's degree with the School of Information Engineering, Zhengzhou University. Her research interests include action recognition and machine learning.

LEI GAO (S'10–M'18) received the Ph.D. degree in electrical and computer engineering from Ryerson University, Toronto, Canada, in 2017, where he is currently with the Department of Electrical and Computer Engineering. His research interests include multimedia signal processing, pattern recognition, machine learning, and information fusion. He is a recipient of the 2018 International Symposium on Multimedia Best Student Paper Award (coauthor); a Visiting Fellowship from the Microsoft Research Asia, in 2016; and the 2015 IEEE International Conference on Image Processing Top 10% Papers Award.

HARON CHWEYA TINEGA received the B.Sc. degree from Karnataka University, India, in 2007, and the M.Sc. degree from Bharathidasan University, India, in 2009. He is currently pursuing the Ph.D. degree with Zhengzhou University, China. His research interest is in the areas of machine learning, such as computer vision and image processing.

YINGQIANG DING received the B.E. degrees from Zhengzhou University, in 2003, and the M.S. and Ph.D. degrees from Tianjin University, China, in 2006 and 2009, respectively. Since 2009, he has been with the School of Information Engineering, Zhengzhou University. His research interests are in the areas of machine learning, computer vision, and optimization algorithms.