IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received April 9, 2020, accepted April 30, 2020, date of publication May 8, 2020, date of current version May 19, 2020.

*Digital Object Identifier 10.1109/ACCESS.2020.2993227*

# Action Recognition Based on Two-Stream Convolutional Networks With Long-Short-Term Spatiotemporal Features

**YANQIN WAN, ZUJUN YU, YAO WANG, AND XINGXIN LI**
School of Mechanical, Electronic and Control Engineering, Beijing Jiaotong University, Beijing 100044, China
Key Laboratory of Vehicle Advanced Manufacturing, Measuring and Control Technology, Beijing Jiaotong University, Beijing 100044, China

Corresponding author: Yao Wang (wangyao@bjtu.edu.cn)

This work was supported in part by the National Key Research and Development Program of China under Grant 2016YFB1200401.

**ABSTRACT** Human action recognition is an important research topic in the field of computer vision due to its application values. Recently, a variety of approaches based on deep learning features have been proposed due to the effectiveness of deep neural networks. But most of these approaches are not able to fully extract spatiotemporal features from videos, because of the lack of consideration of the diversity of scales in temporal domain. In this paper, we propose a two-stream convolutional network with long-short-term spatiotemporal features (LSF CNN) for human action recognition task. The network is mainly composed of two subnetworks. One is long-term spatiotemporal features extraction network (LT-Net) that takes the stacked RGB images as inputs. Another one is short-term spatiotemporal features extraction network (ST-Net) that takes the optical flow as input, which is estimated from two adjacent frames. The two-scale spatiotemporal features are fused in the fully-connected layer and fed into the linear support vector machine (SVM). We also propose a new expression for optical flow field, which is proved to have better performance than traditional expression in action recognition problem. With two-stream architecture, the network can fully learn deep features in both spatial and temporal domains. The experimental results on HMDB51 and UCF101 datasets indicated that the proposed approach improves the action recognition accuracy by using the long-short-term spatiotemporal information.

**INDEX TERMS** Action Recognition, convolutional networks, optical flow, spatiotemporal features.

## I. INTRODUCTION

Human action recognition, which classifies the human behaviors in videos, is one of the most important active areas in computer vision. It has high application value in the field of video surveillance [1], [2], virtual reality [3], intelligent human computer interface [4], etc. Although extensive researches have been done on this topic, video-based action recognition is still a challenging issue due to the complex background, object's appearance and the diversity of behavior types. Video can be regarded as composed of image sequences, including temporal domain information and spatial domain information. Therefore, the main difficulties of action recognition are the diversity of behavior scales in temporal domain and the moving object appearance in spatial domain. Compared with image recognition, due to the introduction of time dimension, the intra class diversity of behavior samples are more abundant. Feature extraction of video data is very difficult, especially the feature extraction in temporal domain because of different action lengths.

Spatiotemporal feature extraction is an essential and core step for visual recognition task. Several classic methods attempted to extract more local features from the local space-time cube. These approaches normally constructed feature descriptors using video patches. These descriptors [5]–[9] have been proven to be able to represent video information effectively. Recently, the Convolutional Neural Networks (CNN) [10] based approaches have been widely used for computer vision, such as image classification [11], [12], image segmentation [13], face recognition [14], [15], foreground detection [16], target tracking [17], [18], etc. CNNs have been proven that they can efficiently extract spatial

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

features from static image. However, traditional CNNs are not able to extract motion information [19], [20], which is important in video analysis. Consequently, 2D-CNN has some limitations in dealing with the problem of video information recognition. Many CNN-based algorithms [21]–[29] for extracting spatiotemporal features have been proposed. These algorithms have demonstrated that deep learning methods can be used for video recognition task. Reference [21] presented a 3D CNNs structure to capture spatiotemporal features from videos. Some algorithms [22], [23] have been developed to compute the optical flow field as the motion features in videos. Although some action recognition approaches took motion information into account, they usually took the single scale optical flow image or single scale image sequence as the input sample. These approaches do not take the diversity of behavior scales in temporal domain account into. Therefore, most of these works have limited ability to capture more discriminative information for video recognition.

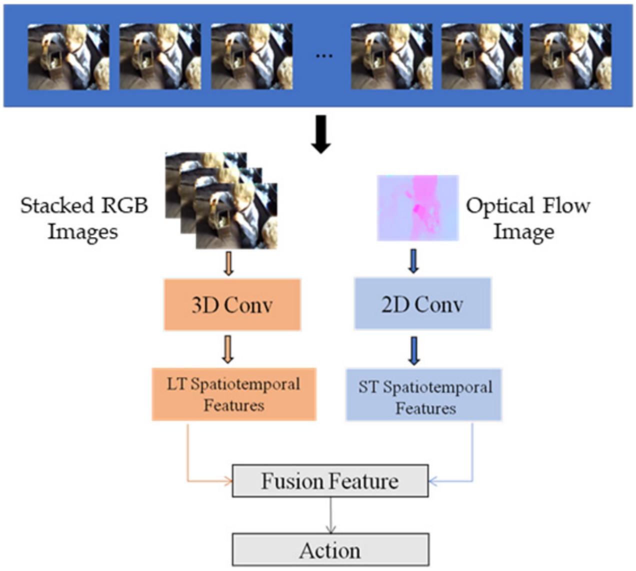


**FIGURE 1. Flowchart of our proposed approach for action recognition.**

In this paper, we propose a novel action recognition approach with long-short-term spatiotemporal features extraction convolutional network (LSF CNN), which is based on 3D convolutional network and 2D convolutional network. The proposed network uses a two-stream network, which is composed of two subnetworks. One is long-term spatiotemporal features extraction network (LT-Net), and another one is short-term spatiotemporal features extraction network (ST-Net). Firstly, a sequence of images sampled from video are taken as inputs of LT-Net to capture long-term spatiotemporal features. Then an optical flow image obtained from two adjacent images is taken as input of ST-Net to capture short-term spatiotemporal features. Finally, these two-scale spatiotemporal features are concatenated in the fully-connected layer and fed into an SVM [30]. Figure 1 shows the flowchart of the proposed approach for action recognition. Our key contributions of this work are summarized as follows:

- We propose a novel two-stream convolution network structure for human action recognition. The network can learn both short-term spatiotemporal features and long-term spatiotemporal features, which are important for video action recognition task. By introducing two-scale spatiotemporal features, the network can extract more comprehensive information in both spatial and temporal effectively. Therefore, the proposed two-stream network structure has good performance for action recognition in real scenes.
- We present an extensive experimental analysis on public datasets to demonstrates the effectiveness of our approach over state-of-the-art.
- We propose a new expression of optical flow image which is like RGB image consist of three-channel data to improve the effectiveness of the network.
- The proposed LSF CNN is a novel framework that uses a two-stream structure to build a convolutional network from video data to action classification results. It also can be used as a common feature extractor to complete other visual recognition task.

The remaining parts of this paper are organized as follows. In Section 2, we introduce the related researches. Section 3 thoroughly explains the proposed approach for action recognition. Experiment and analysis are described in Section 4. In the end, Section 5 presents our conclusions.

## II. RELATED WORK

Action recognition has been studied for decades. Many methods for human action recognition have been proposed. But it is still a challenging task due to the large number of action categories, different action lengths, different object's appearance, complex background motion, etc. Most of these methods are based on spatiotemporal feature extraction. In this section, we briefly review the related works based on spatiotemporal features that most relevant to our work.

Extracting motion and spatial information has been studied for decades, which is still a challenge due to the motion of cameras, different viewpoint, changing light, noise, etc. Early, studies on human action recognition focused on designing handcrafted descriptors to represent video information. Among them, most methods are based on spatiotemporal point detection, such as dense trajectories [31], [32]. Peng *et al.* [33] introduced bag of visual words for video representation. Reference [34] explored learning the adjacent frames to represent a video in a single dynamic image used for video analysis. Lately, [35] combined foreground trajectory with traditional features descriptors and obtained better performance than traditional dense trajectory method. Most of the handcrafted action video representations were obtained by tracking the motion points throughout the entire video. Therefore, these methods are sensitive to noise and have limited capability to obtain more discriminative information in real scenes. Furthermore, it is difficult to design an outstanding handcrafted descriptor to represent complex video data.

Encouraged by the great success of CNN model for image recognition [36]–[40], there are a lot of attempts to employ deep learning methods for action recognition. Ji *et al.* [21] extended the 2D CNN for images to 3D CNN by convoluting the local space-time of multi-frame images. This method was a good attempt of deep learning model in the field of behavior recognition and obtained an excellent achievement in real scene datasets. Then, Du *et al.* [29] modified traditional 2D kernels and introduce the 3D CNNs for spatiotemporal features. One issue with 3D CNNs model is that it takes a single scale image sequence as the input of network to capture the single scale spatiotemporal features. Karpathy *et al.* [19] used slow fusion model to fuse different images in video and constructed a CNN model of image sequence. Via this fusion method, the video sequence information can be effectively added to network, and the expression ability of behavior characteristics can be improved. But the input of the model is a single image selected from a video. Lately, many algorithms have been developed using optical flow as the motion information in videos. Simonyan *et al.* [22] introduced a two-stream network for action recognition, which takes a single RGB image (spatial information) and a stack of optical flow images (temporal information) as inputs. Feichtenhofer *et al.* [23] proposed a two-stream network with new fusion method. And each stream is still regular 2D CNN. [41] discovered one of the limiting factors about the optical flow. Optical flow is the apparent motion of intensity values, which can be produced by lighting changes without any actual motion. Therefore, the optical flow representation of real motion information has limitations. Shao *et al.* [42] proposed a spatiotemporal Laplacian pyramid coding method for video representation. Wang *et al.* [43] combined dense trajectory with CNN and proposed a method of using CNN to express trajectory features. Like [43], Lu *et al.* [44] also combined trajectory pooling method with 3D CNNs and introduced a multi-scale trajectory pooled 3D convolutional descriptor for action recognition. Wang *et al.* [45] introduced a multi-level video representation by stacking the activations of motion features, atoms and phrases. Zhao *et al.* [46] proposed an efficient pooling method called Line pooling, which pools stacked features along the timeline. Varol *et al.* [47] proposed a long-term 3D CNN to capture long-term temporal information. Uddin *et al.* [48] proposed a handcrafted feature descriptor, namely Weber's law-based Volume Local Gradient Ternary Pattern (WVLGTP) and a new convolutional network to extract deep spatial feature. Then, fusing the handcrafted spatiotemporal feature and deep spatial feature for action recognition.

## III. PROPOSED METHOD

There are usually two main steps for actions recognition, including action video representation extraction and classifier training. Spatiotemporal features are extremely important for the descriptor of behavior. Most of the current methods take a single image or a single clip as input of the model. These approaches are not able to fully extract spatiotemporal features from videos due to the lack of consideration of the diversity of scales in temporal domain.

In this work, we propose a novel action recognition approach. The proposed method employs a two-stream CNN Network structure to fuse short-term spatiotemporal features and long-term spatiotemporal features. Two-scale spatiotemporal features have been proven that can improve the results for action recognition, because the action in video has different lengths. First, we employ 3D convolutional network to extract long-term spatiotemporal features from clip selected from video. Then, we estimate the optical flow from two adjacent frames selected from clip and transform the optical flow. We employ 2D convolutional network to extract short-term spatiotemporal features from optical flow. Finally, we fuse long-term spatiotemporal features and short-term spatiotemporal features in fully-connected layer and feed the fusion features into an SVM [30] to achieve the final recognition predictions. In this section, we introduce the details of our proposed approach for human actions recognition.

### A. NETWORK STRUCTURE

Our proposed deep spatiotemporal features extraction model includes two subnetworks. The stacked RGB frames (capture long-term spatiotemporal features) and optical flow image (capture short-term spatiotemporal features) as input samples of the network. The combination of them has been proved to be effective.

In order to fully extract the spatial and temporal features of video data, a two-stream convolution network model is proposed. The two-stream convolutional model in this paper includes stacked static images express structure and optical flow images express structure, as shown in Figure 2. In the static images express structure, stacked RGB images are used as inputs, and C3D [29] network is used to extract long-term spatiotemporal features of videos. In the optical flow images express structure, three-channel optical flow graph is used as input, and VGG16 [50] network is used to extract short-term spatiotemporal features of videos.

A video V is partitioned into N clips and the clip contains 16 frames. V $= \{Ck\}_{k=1}^{N}$, where $C_k$ is the $k$-th clip. $C_k = \{I_t\}_1^m$, where $I_t$ is the $t$-th frame. The top convolution network structure operates on adjacent video clips, effectively capturing long-term spatiotemporal features from stacked still images for action recognition. The bottom convolutional network structure operates on optical flow images, effectively capturing short-term spatiotemporal features from optical flow images for action recognition. One video clip and one optical flow image are sampled from video V and they are used as the inputs of LT-Net stream and ST-Net stream, respectively. Each video clip and optical flow image are processed by the proposed two-stream CNN, then a per-clip action prediction is obtained. After processing all the clips, a series of clips results are obtained by the two-stream model. And then getting the prediction results of the whole

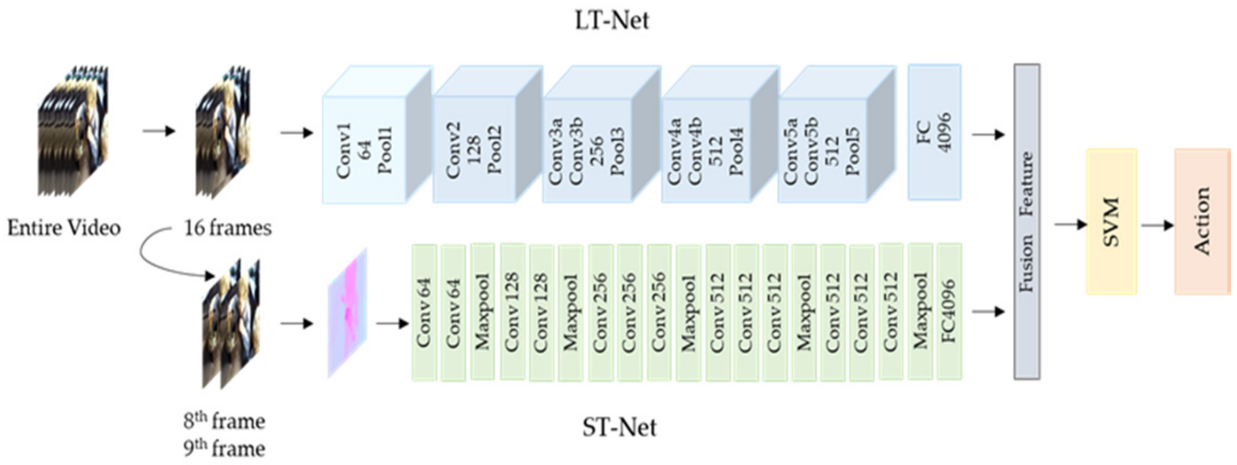


**FIGURE 2.** **The proposed two stream CNN architecture.**

video by counting the results of all cut clips. The architecture of the proposed CNN model presented in Figure 2. We will introduce the proposed action recognition model in detail below.

### B. LONG-TERM SPATIOTEMPORAL FEATURES EXTRACTION

In order to improve the performance of the spatiotemporal features, we propose long-term spatiotemporal features which represent both spatial and temporal information. Long-term spatiotemporal features in video play an important role for action recognition task. For example, run, walk, shoot ball and some other actions look like the same behavior in a short time. How to distinguish these behaviors, long-term spatiotemporal feature extraction is particularly important. For extracting long-term spatiotemporal features, LN-Net takes stacked RGB images as inputs. We employ 3D convolutions to process multiple consecutive pictures. In 3D convolutional network, 2D convolutions are converted to 3D by inflating filters from square to cubic. 3D convolutions are a natural generalization of 2D convolutions for video data which is 3D. Therefore, 3D convolutions have strong spatiotemporal feature representation capability. In this work, we choose C3D [29] as the backbone network, which can capture long-term spatiotemporal features from stacked RGB images. The input of model is the sequence that obtained from video clips, as shown in Figure 3. This stream features extraction network structure contains 8 convolution layers and 5 pooling layers. There is only one type of convolution filter, i.e. $3 \times 3 \times 3$ with stride $1 \times 1 \times 1$. Pooling layers are $2 \times 2 \times 2$ with stride $2 \times 2 \times 2$ except for the first pooling layer which has kernel size of $1 \times 2 \times 2$ due to preserving the temporal information in early phase. The number of filters per layer is shown in Figure 2. 16-frame clips with an overlap of 15 frames are selected from video and they are used as the inputs of LT-Net. For 3D convolutions network, there is an obvious challenge that it is difficult to train due to the number of model parameters.

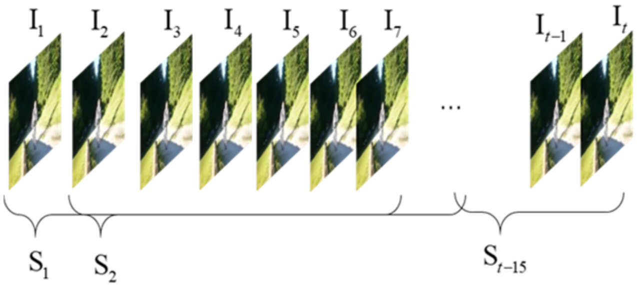


**FIGURE 3.** **The Input for Long-term Spatiotemporal Features Extraction Network.**

We can get more data from videos as training samples by setting adjacent clips with an overlap of 15 frames.

### C. SHORT -TERM SPATIOTEMPORAL FEATURES EXTRACTION

#### 1) OPTICAL FLOW EXTRACTION

video is usually expressed by continuous multi-frame images in temporal domain, so extracting temporal feature information is a very important step for video analysis. For action recognition, short-term temporal features have been demonstrated can be effective [22,23]. In this paper, the optical flow information is used as the input of ST-Net to capture short-term spatiotemporal features. Brox [49] optical flow is used to estimate the motion field of adjacent frames, and the estimated motion field is normalized. The two-channel optical flow field is transformed into a three-channel form, which is consistent with the dimension of RGB image. During the training period, we can initialize the net with a model pre-trained on RGB image dataset.

The two channels of optical flow graph are not color channels of RGB image, but two velocity vector channels. In order to transform optical flow information into three-channel optical flow graph $F$, $\theta \in [0,2\pi]$ and $M$ are introduced. The three channels of optical flow graph $F$ are represented as:

$$F_1 = sin\,\theta = \frac{u}{M}, \tag{1}$$

$$F_2 = cos\,\theta = \frac{v}{M}, \tag{2}$$

$$F_3 = M = \sqrt{u^2 + v^2}, \tag{3}$$

In Figure 4, we visualize the optical flow, the top line frames are the original images, which selected from image sequence. The second line images show the displacement of pixels in y direction and the third line images show the displacement of pixels in x direction. Three-channel optical flow images are shown in the bottom line.

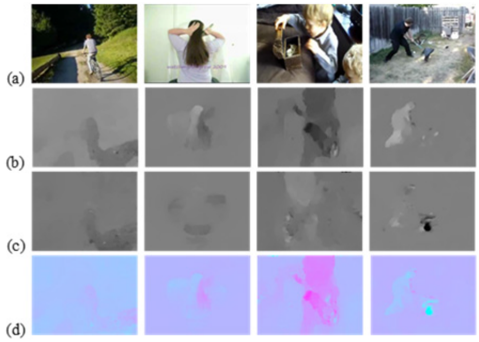


**FIGURE 4.** (a) Original images (b) the displacement of pixels in y direction (c) the displacement of pixels in x direction (d)three-channel optical flow images.

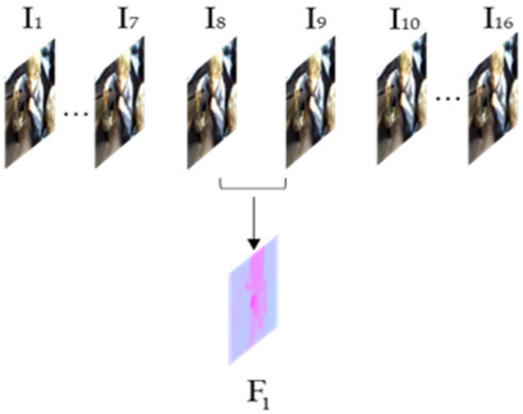


**FIGURE 5.** The Input for Short-term Spatiotemporal Features Extraction Network.

### 2) ST-NET STRUCTURE

To improve the performance of short-term temporal features, we introduce 2D convolutions which have strong spatial feature representation capability. Optical flow images as the input samples of 2D convolutions network. In this work, we employ VGG16[50] as the backbone network, which can capture short-term spatiotemporal features from optical flow information. The input of the model is a three-channel optical flow image that extracted from two adjacent frames which selected from clip, as shown in Figure 5. Short-term spatiotemporal features extraction network structure contains 13 convolution layers and 5 pooling layers. The convolution layers extract spatiotemporal feature maps and the pooling layers decrease the dimensionality of feature maps. Furthermore, the convolution layers followed by the batch normalization (BN) layers and Rectified Liner Unit (ReLU) which are benefit to decrease training time and to overcome overfitting. ST-Net takes single optical flow image as the input with size 256 × 256 × 3, which capture short-term spatiotemporal information from two adjacent frames. In this network, there is only one type of filter, i.e. 3 × 3. The number of filters per layer is shown in Figure 2. Normally, action videos contain different objects which perform different speed behaviors. The appearance information represented by spatial feature and the short-term temporal information represented by optical flow. There are differences in behavior on time scale. Different scale spatiotemporal features need to be considered in feature extraction. Therefore, short-term spatiotemporal information is also important for action recognition.

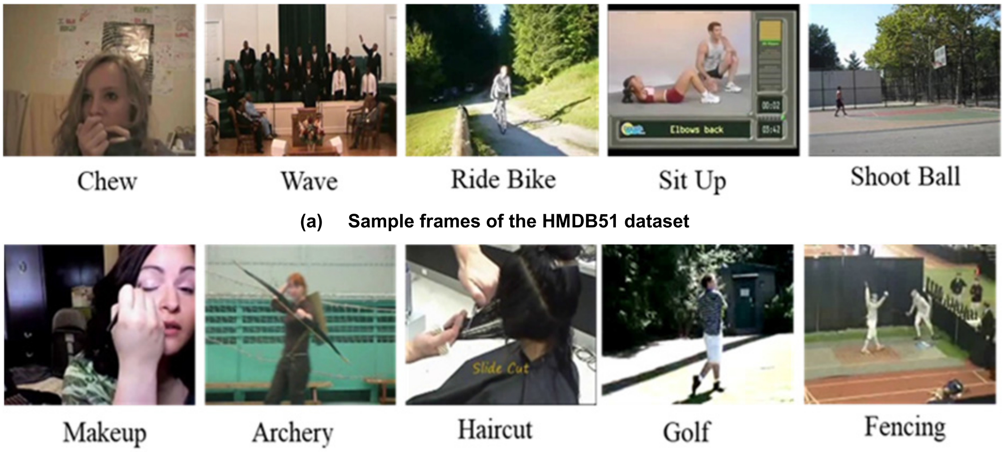
## D. NETWORKS TRAINING AND CLASSIFICATION

### 1) TRAINING METHOD

The proposed network employs a two-stream network, which is composed of two subnetworks. It is difficult to train the whole network from end to end. Since the limited number of samples and complex model, they are easy to cause overfitting problem when training network. In this paper, LT-Net and ST-Net are trained separately. Pre-training model and data augmentation strategies are employed to solve the overfitting problem. LT-Net and ST-Net are trained using public action recognition datasets with fine-tuning method. We initialize LT-Net with a model pre-trained on Sport-M [19] which is the largest video recognition benchmark. Since ST-Net takes three-channel optical flow image as input, the pre-training model on ImageNet [51] can be employed to initialize the ST-Net. During training period, every labeled video is partitioned into N clips to be train samples for LT-Net. Each clip contains 16 frames. We can obtain more data from the video as training samples by setting adjacent clips with an overlap of 15 frames. For each clip, fifteen optical flow images are calculated to be the train samples for ST-Net. LT-Net and ST-Net are both trained with cross entropy loss function. The loss function is the same as C3D net [29]. Training is done by the stochastic gradient descent (SGD) algorithm.

### 2) CLASSIFICATION USING SVM

After training, the two subnetworks can be used as feature extractors. To extract long-term spatiotemporal features, a video is split into 16-frame long clips with a 15-frame overlap between two consecutive clips. To extract short-term spatiotemporal features, two consecutive images are selected from the 16-frame long clips. After extracting spatiotemporal features, short-term spatiotemporal features and long-term



(a) Sample frames of the HMDB51 dataset

(b) Sample frames of the UCF101 dataset

**FIGURE 6.** Sample frames of datasets.

spatiotemporal features are concatenated and fed into SVM for classification. SVM is a common classifier for pattern recognition. The performance of SVM is completely dependent on support vector. The complexity of trained model is determined by the number of support vectors, not by the dimension of data. Therefore, SVM is not easy to over fitting.

And SVM is suitable for individual training. In this paper, LT-Net, ST-Net and SVM are trained separately. Therefore, we utilize SVM with the RBF kernel function [30] to classify the actions via the feature vector. During training phase, model is trained with hinge loss function [30].

## IV. EXPERIMENT

To compare the performance of the proposed LSF CNN model with the current state-of-the-art methods, we first evaluated the proposed three-channel optical flow and the LSF CNN on two public datasets: UCF101[52] and HMDB51[53]. Then the role of the fusion of long-term and short-term spatiotemporal features was investigated by comparing the results with both LT-Net and ST-Net.

### A. DATASETS

HMDB51 dataset is consists of 6766 video clips with 51 classes of human actions. The spatial resolution is 320 × 240. All these videos are obtained from YouTube and digital movies. HMDB51 dataset is a very challenging dataset since most of the videos are taken by non-fixed cameras in real scenes. There are a lot of facial, limb and interactive actions.

The UCF101 dataset is set up by the University of Florida. It is one of the largest human action datasets, which consists of 13,320 video clips with 101 action classes. The videos are divided into 25 groups and each group contains at least 100 videos. Video lengths range from 29 frames to 1776 frames. The spatial resolution is 320 × 240. Video shooting scene is more complex since there are background disturbance, camera motion, scale and illumination changes.

Both datasets are divided into three splits, and the average accuracy of the three splits is used to evaluate the algorithm. Some examples from the datasets are presented in Figure 6.

### B. IMPLEMENTATION DETAILS

We used pytorch [54] to implement our algorithm. The platform used in experiment uses Intel Xeon(R) CPU, and Nvidia K80 GPU. We used MATLAB to complete optical flow estimate operation.

Firstly, we performed the pre-processing operations that include optical flow extraction and transformation. Then, we trained the two streams networks separately. For training LT-Net, the stochastic gradient descent (SGD) algorithm is employed with a batch size of 20. We set a learning rate 0.005 for long-term spatiotemporal features extraction. The input unit size is $112 \times 112 \times 3 \times 16$. For training ST-Net, training is done by SGD with a batch size of 32. We set a learning rate 0.001 for short-term spatiotemporal features extraction. As datasets for action recognition are significantly smaller than datasets for image classification, the chance of overfitting occurring in action recognition is higher. Therefore, data augmentation is crucial for the performance of our method. During training phase, we randomly extract four regions that are the center or corners of the image. After training, the two models can be used as descriptors extractor. During testing phase, a 16-frame long clip sampled and an optical flow image form an input unit. And the optical flow image is calculated from the middle two images of 16-frame long clip. For video prediction, we count clip predictions of all the clips which are sampled from the whole video. For both datasets, researchers provide three splits into training and testing data. The performance is measured by the mean accuracy of three splits.

### C. EXPERIMENTAL RESULTS AND ANALYSIS

We performed various experiments on the action datasets to investigate the effectiveness of the proposed method to recognize human actions. The public action datasets were divided into training set and test set. The researchers who built the public action datasets randomly selected 70% of the data as the training set and the remaining 30% as the test set. And random sampling three times forming three splits on each dataset. The performance is measured by the mean recognition accuracy across three splits. In the experiments, we report the average accuracy over the splits on both HMDB51 and UCF101.

**TABLE 1.** Recognition accuracy of HMDB51 %.

| Method | Split 1 | Split 2 | Split 3 | Average |
|---|---|---|---|---|
| Still image | 49.1 | 50.6 | 49.6 | 49.8 |
| Two-channel OP | 55.2 | 53.4 | 54.7 | 54.4 |
| Three-channel OP | 58.6 | 56.3 | 56.8 | 57.2 |

**TABLE 2.** Recognition accuracy of UCF101 %.

| Method | Split 1 | Split 2 | Split 3 | Average |
|---|---|---|---|---|
| Still image | 81.4 | 80.5 | 81.9 | 81.3 |
| Two-channel OP | 79.3 | 81.5 | 79.5 | 80.1 |
| Three-channel OP | 85.1 | 86.4 | 84.8 | 85.4 |

### 1) EFFECTIVENESS OF THREE-CHANNEL OPTICAL FLOW IMAGE

In this subsection, we explored the performance of different input samples for ST-Net. Different inputs (still image, two-channel optical flow image and three-channel optical flow image) of the single stream were empirically tested and the results are summarized in Table 1 and Table 2 in terms of classification accuracy. For fair comparison and easier network training, we have carried out several experiments under the VGG16 [50] net with three splits data on public action datasets. As shown in Table 1 and Table 2, the three-channel optical flow image is optimal for the action recognition model. The result shows that the three-channel optical flow is an efficient video representation, which is suitable to be the input sample of convolution network. The better performance is caused by the effective dynamic information that carried from optical flow. In addition, the proposed optical flow image has the same dimension as RGB image. In order to improve the effect of the net, numbers RGB images can be used to pretrain the model. The results prove that the new expression for optical flow is more suitable to be the input of convolution network for action recognition. In the following experiments, optical flow images used in this paper are all represented by the new expression.

**TABLE 3.** Recognition accuracy of the proposed network and other state-of-the-arts on the UCF101 dataset and HMDB51 dataset %.

| Method | HMDB51 | UCF101 |
|---|---|---|
| Handcrafted Methods | | |
| DT [31] | 55.9 | 83.5 |
| iDT [32] | 57.2 | 85.9 |
| MoFAP [45] | 61.7 | 88.3 |
| Deep learning Methods | | |
| Two-Stream [22] | 59.4 | 88.0 |
| Two-Stream-Fusion [23] | 62.1 | 90.8 |
| TDD+FV [43] | 63.2 | 90.3 |
| TC3D [44] | 64.5 | 90.1 |
| LTC [46] | 64.8 | 91.7 |
| Trajectory Pooling [47] | 65.6 | 93.7 |
| ST-Net | 57.2 | 85.4 |
| LT-Net | 61.3 | 89.6 |
| Proposed two-stream | 70.2 | 94.8 |

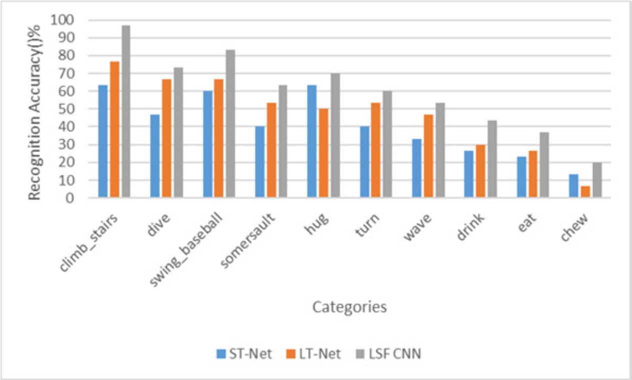


**FIGURE 7.** Accuracy comparison of some classes by ST-Net, LT-Net and LSF CNN on the HMDB51 action dataset.

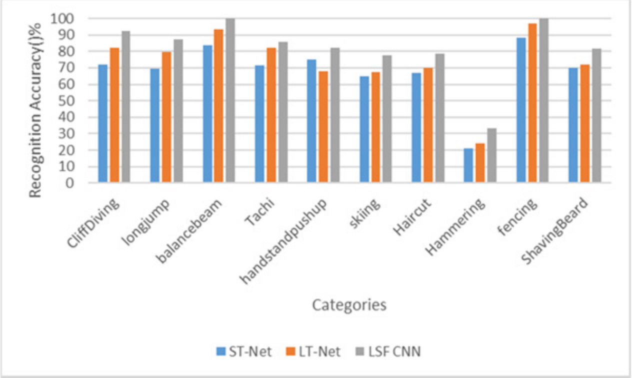
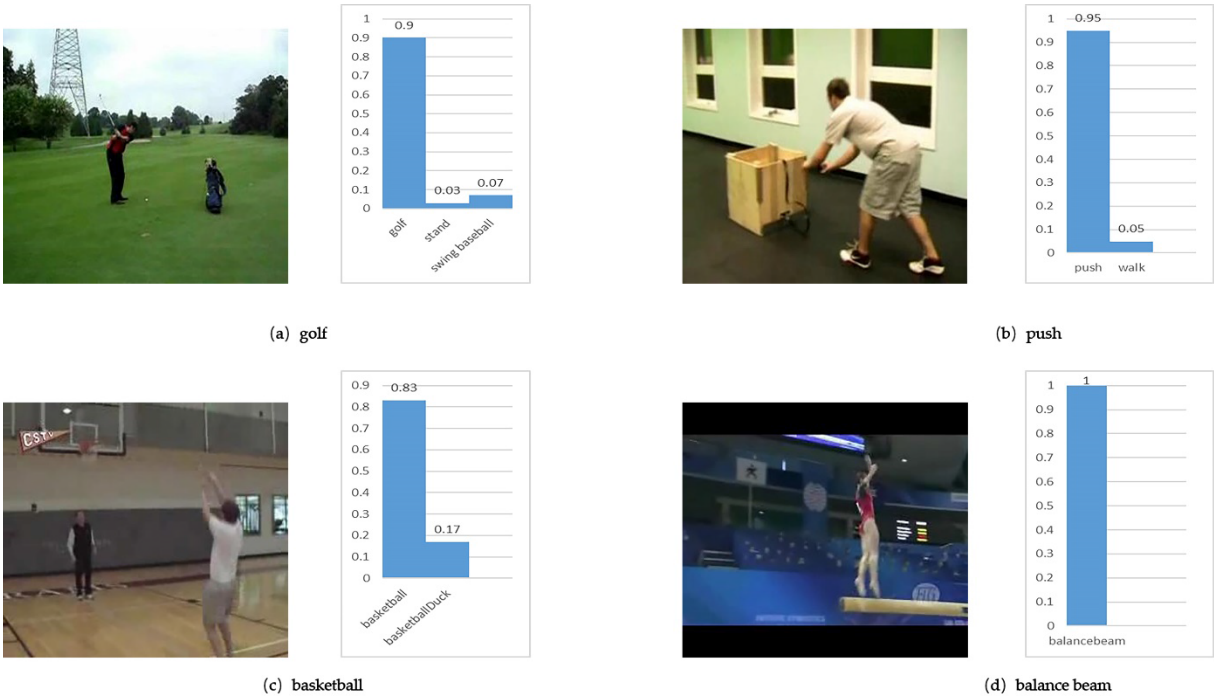


**FIGURE 8.** Accuracy comparison of some classes by ST-Net, LT-Net and LSF CNN on the UCF101 action dataset.

### 2) COMPARISON WITH THE STATE-OF-THE-ART METHODS

To fully evaluate the performance of the proposed method, we compared it with some existing state-of-the-art action recognition methods [22], [23], [31], [32], [43]–[47] and two subnetworks on HMDB51 and UCF101 datasets. Table 3 shows the comparison between the proposed network and other state-of-the-art methods.

From the experiment results, we can see that the proposed LSF CNN shows superior accuracy over these handcrafted methods, such as DT [31], iDT [32], MoFAP [45]. Because handcrafted methods depend on tracking densely sampled points from moving objects and designing handcrafted feature descriptors. Generally, it is difficult to track



(a) golf

(b) push

(c) basketball

(d) balance beam

**FIGURE 9.** Some sample results of action recognition.

the foreground target due to many factors, e.g., the motion of cameras, shadows, complex background, etc. Furthermore, it's hard to design an excellent handcrafted feature descriptor to represent video data. Similarly, the proposed LSF CNN shows competitive performance with two-stream CNN [22], two-stream-fusion CNN [23], TDD-FV [43], TC3D [44], LTC [46], Trajectory Pooling [47]. These methods achieved discriminative representation by considering deep spatiotemporal features with CNN. However, these approaches are solely based on a single image or a single scale clip that sampled from video data. The proposed LSF CNN combined long-term spatiotemporal features and short-term spatiotemporal features, which improved the discriminative power. In this paper, we employ SVM which is a simple liner classifier and achieve better performance. Experimental results demonstrate that our proposed long-short-term spatiotemporal features have good generalization ability and distinguishability. In addition, in order to show more intuitively the performance obtained by our algorithm, two subnetworks recognition results are show in the following experiment.

### 3) COMPARE WITH TWO SINGLE STREAMS

In this subsection, we explored the performance of two single streams to show the role of two-scale features for action recognition. In order to verify the effectiveness of the network framework, the recognition results of two branches and the complete model in the database are tested respectively. Firstly, we measured the performance of ST-Net, which the input is three-channel optical flow image. The network model was pre-training on ImageNet followed by fine-tuning on action dataset. Then, we turned to LT-Net, which the input is clip. Training LT-Net on the public action datasets is challenging, due to the large model parameters and small size training set. Like ST-Net, the network model was pre-training on Sport-M dataset and fine-tuning on action dataset. Then the two single subnetworks were used as feature extractors. We trained and tested long-term spatiotemporal feature and short-term spatiotemporal feature using the linear SVM and reported the accuracy of some categories on the HMDB51 and UCF101.

In the experiment, chew, clap, hug and some other short-term actions categories that can be recognized via single image or two adjacent images have better performance with ST-Net than LT-Net. On the contrary, dive, flic-flac, pullup and some other longer-term actions categories that be recognized need image sequence have better performance with LT-Net. Figure 7 and figure 8 show the accuracy comparison of some categories that obtain definite improvement of recognition accuracy on the HMDB51 action dataset and UCF101 action dataset. From the figures, we can see that our proposed LSF CNN shows the superior performance over ST-Net and LT-Net. The better performance is caused by the fusion of different scale features. On the HMDB51 dataset, 'climb stairs' obtains the highest growth in recognition accuracy. The accuracy increases by 33.4%. On the UCF101 dataset, 'cliff diving' obtains the highest growth in recognition accuracy. The accuracy increases by 20.5%. The HMDB51 actions are more similar the actions in real scene, so the action on the HMDB51 is harder to recognize. However, our proposed LSF CNN has a higher growth on the HMDB51. The results of the public action datasets demonstrate that effective combination of long-term spatiotemporal features and short-term spatiotemporal features is conducive to the expression of video information.

To summarize, the above experimental results on the public action datasets show that the proposed LSF CNN obtains performance gains on HMDB51 dataset and UCF101 dataset, due to the effective combination of long-term spatiotemporal features and short-term spatiotemporal features. In order to show more intuitively the performance obtained by our algorithm, the recognition results of some categories are show in Figure 9.

In addition, we observe that the proposed LSF CNN achieves low recognition accuracy on some action categories, such as 'chew', 'eat', and 'drink' actions in the HMDB51. Because there are significant similarities between these actions and small range of these actions. On the public datasets, there are some videos contain complex background. For these videos, our proposed LSF CNN also has limitations for extracting information. Optical flow is sensitive to lighting changes. It can be produced by background motion without any actual motion. In other words, sometimes the optical flow is not the real action information.

## V. CONCLUSIONS

In this paper, we combined both optical flow and deep spatiotemporal features for human action recognition. In order to obtain deep spatiotemporal features from videos we proposed a novel two-stream Convolutional Networks structure, which extracts effective long-short-term spatiotemporal features from videos. The experimental results show that the proposed model has a good performance for the action recognition task, benefited from its ability to learn different scale spatiotemporal features effectively.

The proposed model is also an effective method to capture motion information for other video recognition task, such as people counting and abnormal event detection, where the temporal features may be can improve the recognition results. For potential future work, we are planning to further extend the fusion method for encoding spatiotemporal features effectively and efficiently. Furthermore, we are also planning to explore learning a CNN to predict optical flow, which also benefit for human action recognition.

## REFERENCES

[1] D. Ayers and M. Shah, "Monitoring human behavior from video taken in an office environment," *Image Vis. Comput.*, vol. 19, no. 12, pp. 833–846, Oct. 2001.

[2] M. T. López, A. Fernández-Caballero, M. A. Fernández, J. Mira, and A. E. Delgado, "Visual surveillance by dynamic visual attention method," *Pattern Recognit.*, vol. 39, no. 11, pp. 2194–2211, Nov. 2006.

[3] E. A. Suma, D. M. Krum, B. Lange, S. Koenig, A. Rizzo, and M. Bolas, "Adapting user interfaces for gestural interaction with the flexible action and articulated skeleton toolkit," *Comput. Graph.*, vol. 37, no. 3, pp. 193–201, May 2013.

[4] H. Kjellström, J. Romero, and D. Kragić, "Visual object-action recognition: Inferring object affordances from human demonstration," *Comput. Vis. Image Understand.*, vol. 115, no. 1, pp. 81–90, Jan. 2011.

[5] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie, "Behavior recognition via sparse spatio-temporal features," in *Proc. IEEE Int. Workshop Vis. Surveill. Perform. Eval. Tracking Surveill.*, Oct. 2006, pp. 65–72.

[6] A. Klaeser, M. Marszalek, and C. Schmid, "A spatio-temporal descriptor based on 3D-gradients," in *Proc. Brit. Mach. Vis. Conf.*, 2008, pp. 1–10.

[7] I. Laptev and T. Lindeberg, "On space-time interest points," *Int. J. Comput. Vis.*, vol. 64, nos. 2–3, pp. 107–123, 2005.

[8] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Anchorage, AK, USA, Jun. 2008, pp. 24–26.

[9] G. Willems, T. Tuytelaars, and L. V. Gool, "An efficient dense and scale-invariant spatio-temporal interest point detector," in *Proc. Eur. Conf. Comput. Vis.*, 2008, pp. 650–663.

[10] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.

[11] J. Wang, Y. Yang, J. Mao, Z. Huang, C. Huang, and W. Xu, "CNN-RNN: A unified framework for multi-label image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2285–2294.

[12] H. Lee and H. Kwon, "Going deeper with contextual CNN for hyperspectral image classification," *IEEE Trans. Image Process.*, vol. 26, no. 10, pp. 4843–4855, Oct. 2017.

[13] Y. Wang, L. Zhu, Z. Yu, and B. Guo, "An adaptive track segmentation algorithm for a railway intrusion detection system," *Sensors*, vol. 19, no. 11, p. 2594, 2594.

[14] R. He, X. Wu, Z. Sun, and T. Tan, "Wasserstein CNN: Learning invariant features for NIR-VIS face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 7, pp. 1761–1773, Jul. 2019.

[15] S. Banerjee and S. Das, "Mutual variation of information on transfer-CNN for face recognition with degraded probe samples," *Neurocomputing*, vol. 310, pp. 299–315, Oct. 2018.

[16] Y. Wang, Z. Yu, and L. Zhu, "Foreground detection with deeply learned multi-scale spatial-temporal features," *Sensors*, vol. 18, no. 12, p. 4269, 4269.

[17] N. Mahmoudi, S. M. Ahadi, and M. Rahmati, "Multi-target tracking using CNN-based features: CNNMTT," *Multimedia Tools Appl.*, vol. 78, no. 6, pp. 7077–7096, Mar. 2019.

[18] L. Leal-Taixé, C. Canton-Ferrer, and K. Schindler, "Learning by tracking: Siamese CNN for robust target association," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2016, pp. 33–40.

[19] 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.

[20] J. Yue-Hei Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, and G. Toderici, "Beyond short snippets: Deep networks for video classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 4694–4702.

[21] S. Ji, W. Xu, M. Yang, and K. Yu, "3D convolutional neural networks for human action recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 221–231, Jan. 2013.

[22] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2014, pp. 568–576.

[23] C. Feichtenhofer, A. Pinz, and A. Zisserman, "Convolutional two-stream network fusion for video action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1933–1941.

[24] C. Jin, S. Li, T. D. Do, and H. Kim, "Real-time human action recognition using CNN over temporal images for static video surveillance cameras," in *Advances in Multimedia Information Processing—PCM*. Springer, 2015, pp. 330–339.

[25] Z. Tu, J. Cao, Y. Li, and B. Li, "MSR-CNN: Applying motion salient region based descriptors for action recognition," in *Proc. 23rd Int. Conf. Pattern Recognit. (ICPR)*, Dec. 2016, pp. 3524–3529.

[26] B. Li, Y. Dai, X. Cheng, H. Chen, Y. Lin, and M. He, "Skeleton based action recognition using translation-scale invariant image mapping and multi-scale deep CNN," in *Proc. IEEE Int. Conf. Multimedia Expo Workshops (ICMEW)*, Jul. 2017, pp. 601–604.

[27] C. Lea, A. Reiter, R. Vidal, and G. D. Hager, "Segmental spatiotemporal CNNs for fine-grained action segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 36–52.

[28] M. Ravanbakhsh *et al.*, "Action recognition with image based CNN features," in *Proc. CVPR*, 2015.

[29] 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. (ICCV)*, Dec. 2015, pp. 4489–4497.

[30] C. J. C. Burges, "A tutorial on support vector machines for pattern recognition," *Data Mining Knowl. Discovery*, vol. 2, no. 2, pp. 121–167, 1998.

[31] H. Wang, A. Klaser, C. Schmid, and C.-L. Liu, "Action recognition by dense trajectories," in *Proc. CVPR*, Jun. 2011, pp. 3169–3176.

[32] H. Wang and C. Schmid, "Action recognition with improved trajectories," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 3551–3558.

[33] X. Peng, L. Wang, X. Wang, and Y. Qiao, "Bag of visual words and fusion methods for action recognition: Comprehensive study and good practice," *Comput. Vis. Image Understand.*, vol. 150, pp. 109–125, Sep. 2016.

[34] H. Bilen, B. Fernando, E. Gavves, A. Vedaldi, and S. Gould, "Dynamic image networks for action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 3034–3042.

[35] S. Dong, D. Hu, R. Li, and M. Ge, "Human action recognition based on foreground trajectory and motion difference descriptors," *Appl. Sci.*, vol. 9, no. 10, p. 2126, 2126.

[36] M. Oquab, L. Bottou, I. Laptev, and J. Sivic, "Learning and transferring mid-level image representations using convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1717–1724.

[37] E.-S.-A. El-Dahshan, H. M. Mohsen, K. Revett, and A.-B.-M. Salem, "Computer-aided diagnosis of human brain tumor through MRI: A survey and a new algorithm," *Expert Syst. Appl.*, vol. 41, no. 11, pp. 5526–5545, Sep. 2014.

[38] L. Chang, X. M. Deng, M. Q. Zhou, Z. K. Wu, Y. Yuan, and S. Yang, "Convolutional neural networks in image understanding," *Acta Autom. Sinica*, vol. 42, no. 9, pp. 1300–1312, 2016.

[39] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Semantic image segmentation with deep convolutional nets and fully connected CRFs," *Comput. Sci.*, vol. 4, pp. 357–361, Dec. 2014.

[40] G. Cheron, I. Laptev, and C. Schmid, "P-CNN: Pose-based CNN features for action recognition," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 3218–3226.

[41] Sevilla-Lara, Laura, "On the integration of optical flow and action recognition," in *Proc. German Conf. Pattern Recognit.*, 2017, pp. 281–297.

[42] L. Shao, X. Zhen, D. Tao, and X. Li, "Spatio-temporal Laplacian pyramid coding for action recognition," *IEEE Trans. Cybern.*, vol. 44, no. 6, pp. 817–827, Jun. 2014.

[43] L. Wang, Y. Qiao, and X. Tang, "Action recognition with trajectory-pooled deep-convolutional descriptors," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 4305–4314.

[44] X. Lu, H. Yao, S. Zhao, X. Sun, and S. Zhang, "Action recognition with multi-scale trajectory-pooled 3D convolutional descriptors," *Multimedia Tools Appl.*, vol. 78, no. 1, pp. 507–523, Jan. 2019.

[45] L. Wang, Y. Qiao, and X. Tang, "MoFAP: A multi-level representation for action recognition," *Int. J. Comput. Vis.*, vol. 119, no. 3, pp. 254–271, Sep. 2016.

[46] S. Zhao, Y. Liu, Y. Han, R. Hong, Q. Hu, and Q. Tian, "Pooling the convolutional layers in deep ConvNets for video action recognition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 8, pp. 1839–1849, Aug. 2018.

[47] G. Varol, I. Laptev, and C. Schmid, "Long-term temporal convolutions for action recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 6, pp. 1510–1517, Jun. 2018.

[48] M. Uddin and Y.-K. Lee, "Feature fusion of deep spatial features and handcrafted spatiotemporal features for human action recognition," *Sensors*, vol. 19, no. 7, p. 1599, 2019.

[49] T. Brox, A. Bruhn, N. Papenberg, and J. Weickert, "High accuracy optical flow estimation based on a theory for warping," in *Proc. ECCV*, 2004, vol. 3024, no. 10, pp. 25–36.

[50] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. ICLR*, 2014.

[51] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Miami, FL, USA, Jun. 2009, pp. 20–25, pp. 248–255.

[52] K. Soomro, A. Roshan Zamir, and M. Shah, "UCF101: A dataset of 101 human actions classes from videos in the wild," 2012, *arXiv:1212.0402*. [Online]. Available: http://arxiv.org/abs/1212.0402

[53] H. Kuehne *et al.*, "HMDB51: A large video database for human motion recognition," in *Proc. ICCV*, 2013, pp. 2556–2563.

[54] A. Paszke, S. Gross, S. Chintala, and G. Chanan. *Pytorch*. Accessed: 2019. [Online]. Available: https://github.com/pytorch/pytorch

**YANQIN WAN** was born in Nanchang, China. She received the B.S. degree in measurement and control technology and instruments from Beijing Jiaotong University, Beijing, China, in 2014, where she is currently pursuing the Ph.D. degree with the Key Laboratory of Vehicle Advanced Manufacturing, Measuring and Control Technology. Her research interests include image recognition and machine learning.

**ZUJUN YU** was born in Dangyang, China. He received the B.S. and M.S. degrees in thermal power machinery and devices and the Ph.D. degree in vehicle operation engineering from Beijing Jiaotong University, Beijing, China, in 1991, 1994, and 2009, respectively. He is currently a Full Professor with the School of Mechanical, Electronic and Control Engineering, Beijing Jiaotong University, where he is also the Vice-Principal. His research interests include embedded system and intelligent instruments, detection and control of vehicles and infrastructure, and state detection and fault diagnosis of electromechanical systems.

**YAO WANG** received the B.S., M.S., and Ph.D. degrees from Beijing Jiaotong University, Beijing, China. He is currently a Lecturer with the School of Mechanical, Electronic and Control Engineering, Beijing Jiaotong University. His research interests include embedded system and intelligent instruments, detection and control of vehicles and infrastructure, and state detection and fault diagnosis of electromechanical systems.

**XINGXIN LI** was born in Tianshui, China. She received the B.S. degree in measurement and control technology and instruments from Beijing Jiaotong University, Beijing, China, in 2014, where she is currently pursuing the Ph.D. degree with the Key Laboratory of Vehicle Advanced Manufacturing, Measuring and Control Technology. Her research interests include image recognition and machine learning.