Received February 12, 2018, accepted March 12, 2018, date of publication March 19, 2018, date of current version April 23, 2018.

Digital Object Identifier 10.1109/ACCESS.2018.2817253

Human Action Recognition by Learning Spatio-Temporal Features With Deep Neural Networks

LEI WANG[1,2], YANGYANG XU[1], JUN CHENG[1,2], HAIYING XIA[3], JIANQIN YIN[4], AND JIAJI WU[5], (Member, IEEE)

[1]Guangdong Provincial Key Laboratory of Robotics and Intelligent System, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518055, China
[2]The Chinese University of Hong Kong, Hong Kong
[3]Guangxi Normal University, Guilin 541000, China
[4]School of Automation, Beijing University of Posts and Telecommunications, Beijing 100876, China
[5]School of Electronic Engineering, Xidian University, Xi'an 710071, China

Corresponding author: Jun Cheng (jun.cheng@siat.ac.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant U1713213, Grant 61772508, Grant 61762014, Grant 61775175, and Grant 61673192, in part by the Guangdong Natural Science Funds under Grant 2014A030310252, in part by the CAS Key Technology Talent Program, in part by the Shenzhen Technology Project under Grant JCYJ20170413152535587, Grant JSGG20160331185256983, Grant JSGG20160229115709109, and Grant JCYJ 20170307164023599, in part by the Guangdong Technology Program under Grant 2016B010108010 and Grant 2016B010125003, in part by the Key Laboratory of Human–Machine Intelligence-Synergy Systems, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, under Grant 2014DP173025, in part by the Shenzhen Engineering Laboratory for 3D Content Generating Technologies under Grant [2017] 476, and in part by the CAS Key Technology Talent Program.

ABSTRACT Human action recognition is one of the fundamental challenges in robotics systems. In this paper, we propose one lightweight action recognition architecture based on deep neural networks just using RGB data. The proposed architecture consists of convolution neural network (CNN), long short-term memory (LSTM) units, and temporal-wise attention model. First, the CNN is used to extract spatial features to distinguish objects from the background with both local and semantic characteristics. Second, two kinds of LSTM networks are performed on the spatial feature maps of different CNN layers (pooling layer and fully-connected layer) to extract temporal motion features. Then, one temporal-wise attention model is designed after the LSTM to learn which parts in which frames are more important. Lastly, a joint optimization module is designed to explore intrinsic relations between two kinds of LSTM features. Experimental results demonstrate the efficiency of the proposed method.

INDEX TERMS Artificial intelligent, human action recognition, attention model, deep neural networks, robotic system.

I. INTRODUCTION

Human action recognition is one important task in robotics systems, especially for intelligent services. For example, in smart homes or smart factories, the robotics systems could assist human or collaborate with human, based on the recognition of action [1]. Combined with cyber-physical systems, action recognition can be used for other applications, such as health-care [2]. Also it can be applied in social activity analysis using cloud computing techniques [3]–[5]. However, with background clutter and occlusions in real world, human action recognition is still far from practical applications [6]–[8], especially in complex dynamic systems.

For video action recognition, previous approaches always take similar ideas with that of image recognition. But different from still images, human actions consist of ever-changing motions with different target objects, and different objects have various appearances in different scenes. So, it's indispensable to explore diverse spatio-temporal features for action recognition. To extract spatio-temporal features, Histogram of 3D Oriented Gradients (HOG3D) [9] and Histogram of Optical Flow (HOF) [10] have been proposed. The features will be further encoded or pooled in a hierarchic architecture, and input to Support Vector Machine (SVM) classifier. To make full use of motion information, one method based on dense-point trajectories has been proposed by computing optical flow of video frames [11]. The Motion Boundary Histograms (MBH) method has achieved good performance by extracting gradient features on horizontal and vertical components of optical flow separately [12].

In recent years, deep neural networks (DNNs) have obtained great achievement in many areas such as object detection, recognition, and image classification, due to its ability of automatically learning features from large datasets [13]–[16]. Spatial features of images can be extracted by convolution layers in Convolution Neural Network (CNN), which contains orientation-sensitive filters [17]. By extending the connectivity of the network in time dimension, CNN is also used to learn spatio-temporal features for large scale video classification [18]. As a typical recurrent neural network (RNN) architecture, Long Short Term Memory (LSTM) has the ability to preserve sequence information over time and capture long-term dependencies [19], so that it can extract temporal features. LSTM has been applied in many sequential modeling tasks such as machine translation, speech recognition, and visual descriptions [20]. With the aid of attention model, LSTM has achieved encouraging performance in machine translation [21] and image caption [22]. LSTM has potential ability in doing prediction tasks for videos, however, it does not take the spatial correlation into consideration. In some references, the original LSTM is referred to fully connected LSTM (FC-LSTM). Shi *et al.* [23] extended the FC-LSTM to convolution LSTM (ConvLSTM) to extract spatial and temporal information in a same LSTM unit. Modeling spatial and temporal features together will be beneficial for accurate recognition.

It is a natural way to detect the appearance of one object using CNN, and detect the motion using LSTM, according to human visual recognition mechanism. So, in this paper, we propose a new lightweight architecture for action recognition in videos based on DNN with only RGB data. Optical flow is not used, since its computation is too complicated for real-time applications. The proposed architecture consists of CNN, LSTM, attention model, and joint optimization. First, we extract two kinds of CNN features, *i.e.* spatial features and semantic features, produced by the convolution layer and fully connected layer, respectively. Correspondingly, for temporal feature extraction, two kinds of LSTM are built after convolution layer and fully connected layer of CNN, named as ConvLSTM and FC-LSTM, respectively. Two different attention models are designed for LSTM to provide insights into where the neural network is looking, find important parts of video, avoid the background noise's effect, and benefit the recognition. Each LSTM produces a vector to represent temporal feature of videos. There exist intrinsic relations between these two features, so we design a joint optimization module (JOM) to explore them.

The main contributions of this work can be summarized as follows. (1) We propose a feature extractor which consists of two kinds of LSTM after different layers of CNN to extract both spatial and semantic features in temporal domain. (2) We design a temporal-wise attention model after LSTM to learn temporal focus of actions. (3) We design a joint optimization module to train the network to be more robust.

II. RELATED WORKS

There exist a number of works for action recognition in videos, such as methods using hand-craft features (Harris3D, HOG3D, HOF, etc.) to generate spatio-temporal descriptors around the detected local interest points, and then using SVM for classification. Since the proposed method is based on deep neural network (DNN), in this section, we will only review related works based on DNN but not hand-craft-based methods.

A. 3D CONVOLUTION NETWORKS ON FRAMES

3D convolution networks [24], [25] have been employed on video frames to learn implicit motion features. The frames are from short video clips, the time of which is a few seconds. And the prediction results on clips are averaged at video level. The network performs just marginally better than single frame baseline [25], which indicates that the motion features have not been learned sufficiently.

B. CNN ON FRAMES AND OPTICAL FLOW

There are two pathways in human visual cortex, the ventral and dorsal streams, which performs object and motion recognition, respectively. According to this mechanism, a two-stream deep Convolution Network (ConvNet) is proposed for action recognition [26], which incorporates spatial and temporal networks. Spatial ConvNet works on a single frame, while temporal ConvNet works on multi-frame optical flow. Different fusion methods [27] of spatial and temporal convolution networks have been proposed to take advantage of the spatio-temporal information. But only up to 10 consecutive frames are used as a group for inference. As a result, only a small part of the full video's information is exploited. This will affect the recognition accuracy.

C. CNN/RNN ON FRAMES

A deep fusion framework of CNN and RNN is proposed in [28], and four fusion models are evaluated for recognizing human actions. In the first model, the last convolution layer of the VGG-16 network is connected with LSTM, followed by a soft-max layer. In the second model, the full connect layer output is fed to the LSTM, followed by a soft-max layer. In the third model, the outputs of LSTM in previous two models are merged and passed through a soft-max layer. In the fourth model, the outputs of convolution layer and full connected layer are fed into sequence-to-sequence LSTM, and then the results are fed into sequence-to-one LSTM. Experiments show that the fourth model has higher recognition ability.

D. CNN/RNN ON FRAMES AND OPTICAL FLOW

For accurate video classification or action recognition, it's important to learn a global description of the video's temporal evolution. Temporal feature pooling and LSTM are present

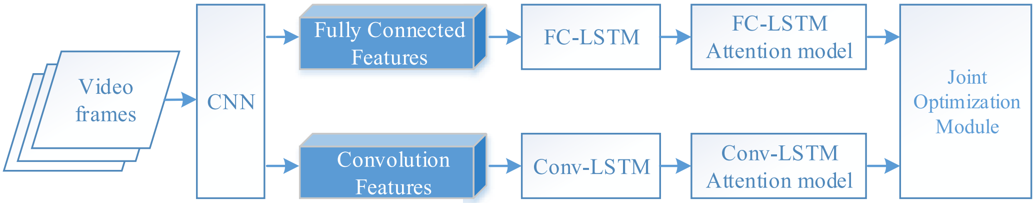


FIGURE 1. Flowgraph of the proposed method.

in [29] to utilize more frames (up to 120) for global information. By working on the last convolution layer across the video's frames, feature pooling model generates vector for video-level prediction. Employing LSTM on both frames and optical flow obtains good performance.

E. ATTENTION MODEL FOR VIDEO

According to the research in visual cognition, human always focus sequentially on different parts of the scene to extract relevant information, instead of on an entire scene at once. Therefore, employing attention mechanism will help to improve the performance in related learning tasks.

Recurrent soft attention model has been developed for action recognition [30]. LSTM is used to predict the probability of location and class label at next time step. And then, soft attention mechanism takes expectation over the feature slices at different regions to compute the expected value of the input at the next time-step. But since all the features are required to perform dynamic pooling, the method is computationally expensive.

A hierarchical attention network has been proposed in [31] for action recognition in video, which incorporates static spatial information, short-term motion information and long-term video temporal structures. First, two-stream ConvNets are used to extract appearance and motion features from frame images and corresponding optical flow images, respectively. Secondly, a hierarchical LSTM with two layers are used to model the video temporal structure. And then, attention weights are computed by using the appearance and motion features.

III. PROPOSED METHOD

In this section, we introduce the proposed method in detail.

The purpose of this work is to propose one DNN-based method for action recognition in video just using RGB data. The main idea is to use CNN to extract spatial features of each frame, use two kinds of LSTM with attention model to explore the temporal features between frames in video, and use joint optimization layer to fuse the two kinds of output temporal features to further extract relations. According to video labels, the entire network is trained for action recognition.

In CNN networks, the output results of the convolution layer and the fully connected layer are different. The convolution layer outputs spatial information, while the fully connected layer outputs semantic information. Both kinds of information are important for object recognition—spatial information (shapes, outlines, etc.) and semantic information (location invariance, rotation invariance, etc.). So both of them are used, and these two kinds of output results on video frames are processed with different LSTMs for temporal feature extraction. In the proposed method, convolutional LSTM (Conv-LSTM) and fully-connected LSTM (FC-LSTM) are performed on the output of convolution layer and fully connected layer of CNN, respectively. Attention model is relevant to human visual mechanism, since human always focus on moving objects instead of the whole picture or static background. Attention model adds a dimension of interpret-ability, and contributes to reduce the effect of background, so it will be benefit for recognition.

The framework of the proposed method is depicted in Figure 1. There are four primary components—spatial features extractor by CNN, Conv-LSTM with temporal-wise attention, FC-LSTM with temporal-wise attention, and joint optimization module.

A. SPATIAL FEATURES EXTRACTION BY CNN

Residual learning structure is adopted in CNN for spatial feature extraction [32]. The layers in this network structure are formulated to learning residual functions with reference to their input, as depicted in Fig. 2. The network parameters are derived by training on ImageNet dataset.

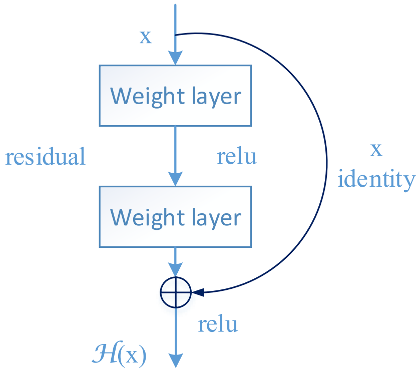


FIGURE 2. Residual mapping structure.

To get spatial features, each video frame is fed into CNN, and two feature maps will be produced after the pooling layer and FC layer. Specifically, for the t^{th} frame, the last pooling layer outputs feature map f^t_{conv} and the fully connected layer outputs f^t_{fc}. The dimension of f^t_{conv} and f^t_{fc} is $K \times K \times C$ and D, respectively, where $K \times K$ is the shape of feature vector, C and D are filter numbers. At each time step, we can extract two feature vectors with dimension of $K \times K \times C$ and D. So, for the video with time length of T, feature maps can be present in the form of matrices as follows.

$$F_{conv} = [f^1_{conv}, \ldots, f^t_{conv}, \ldots, f^T_{conv}] \in \mathbb{R}^{K\times K\times C\times T} \quad (1)$$

$$F_{fc} = [f^1_{fc}, f^2_{fc}, \ldots, f^T_{fc}] \in \mathbb{R}^{D\times T} \quad (2)$$

B. FULLY-CONNECTED LSTM (FC-LSTM) WITH TEMPORAL-WISE ATTENTION

We designed LSTM with temporal-wise attention to explore the temporal features in video. LSTM is one kind of recurrent neural networks, which can preserve sequence information over time and capture long-term dependencies. One advantage of LSTM is that the gradient does not tend to vanish when trained with back propagation through time.

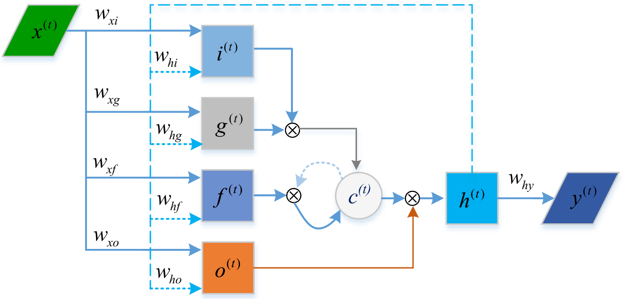


FIGURE 3. LSTM unit.

One unit of LSTM is depicted in Fig. 3. $x^{(t)}$, $c^{(t)}$, $h^{(t)}$ and $y^{(t)}$ stand for input vector, cell state, hidden state and output at the t^{th} state, respectively. The output $y^{(t)}$ depends on hidden state $h^{(t)}$, while $h^{(t)}$ depends on not only the cell state $c^{(t)}$ but also its previous state. Cell state $c^{(t)}$ is influenced by the input and memory information. $i^{(t)}$, $g^{(t)}$, $f^{(t)}$, and $o^{(t)}$ modulate the input and memory information, which stand for input gate, input modulation gate, forget gate, and output gate, respectively.

LSTM is implemented as follows:

$$i^{(t)} = \sigma_g(w_{xi}x^{(t)} + w_{hi}h^{(t-1)} + b_i) \quad (3)$$

$$f^{(t)} = \sigma_g(w_{xf}x^{(t)} + w_{hf}h^{(t-1)} + b_f) \quad (4)$$

$$o^{(t)} = \sigma_g(w_{xo}x^{(t)} + w_{ho}h^{(t-1)} + b_o) \quad (5)$$

$$g^{(t)} = \tanh(w_{xg}x^{(t)} + w_{hg}h^{(t-1)} + b_g) \quad (6)$$

$$c^{(t)} = f^{(t)} \odot c^{(t-1)} + i^{(t)} \odot g^{(t)} \quad (7)$$

$$h^{(t)} = o^{(t)} \odot \sigma_h(c^{(t)}) \quad (8)$$

where $h^{(t-1)}$ is the previous hidden state. w_x and w_h are weights of input vector and hidden state, respectively. b_i, b_f, b_o and b_g stand for the bias terms. $\sigma(\cdot)$ means a sigmoid function and $\odot$ indicated the Hadamard product. The cell state and output are computed step by step to capture long-term dependencies.

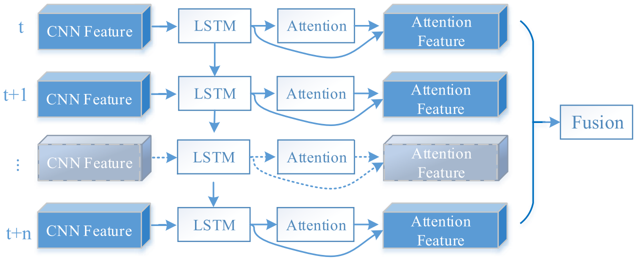


FIGURE 4. Unfolded LSTM with its attention model.

After LSTM, one temporal attention model is designed to decide which frames are important in videos for action recognition. The attention model is used to learn the representation of informative frames to produce a feature vector, which is computed as follows.

First, the output gate of LSTM $o^{(t)}$ pass through a fully connected layer and a tanh activation function to produce a mid-result $u^{(t)}$, which is computed as follows,

$$u^{(t)} = \tanh(W_u o^{(t)} + b_u) \quad (9)$$

where W_u and b_u are parameters in a fully connected layer, standing for weight and bias, respectively.

Secondly, we predict a SoftMax over L frames to produce a normalized importance weight $\alpha^{(t)}$. The focus softmax is defined as follows,

$$\alpha^{(t)} = p(F_t = t \mid u^{(t)}) = \frac{\exp({W_t}^T u^{(t)})}{\sum_{l=1}^{L} \exp({W_l}^T) u^{(t)}}, \qquad t \in 1 \ldots L, \quad (10)$$

where W_t^T stands for the weight mapping to the t^{th} element of the focus softmax, and L for the number of the frames. $\alpha^{(t)}$ is the probability with which the corresponding frame is thought to be important in this network, and it tells the network which time steps are needed to focus on.

Thirdly, a feature vector $\boldsymbol{s}$ is computed as the expected value of LSTM output feature at time-step t by taking expectation of all time steps' feature vector. And it is computed as follows,

$$s = E_{p(F_t = t \mid u^{(t)})}(o^{(t)}) = \sum_{t=1}^{L} \alpha^{(t)} o^{(t)} \quad (11)$$

Fig. 4 shows the unfolded LSTM and attention model. Different from previous attention models which mainly focus on regions in each frame, our attention model is built after LSTM to figure out which frame is important. The attention model is trained using back propagation to produce dynamical attention weight throughout the video sequence.

C. CONVOLUTION LSTM WITH TEMPORAL-WISE ATTENTION

The spatial information is not encoded by the input-to-state or state-to-state transition in FC-LSTM. So we adopt a

Convolution LSTM network [23] to overcome this deficiency, with the state-to-state transitions in FC-LSTM to be replaced by convolution operations. ConvLSTM is computed as follow.

$$i^{(t)} = \sigma_g(w_{xi} * x^{(t)} + w_{hi} * h^{(t-1)} + b_i) \tag{12}$$
$$f^{(t)} = \sigma_g(w_{xf} * x^{(t)} + w_{hf} * h^{(t-1)} + b_f) \tag{13}$$
$$o^{(t)} = \sigma_g(w_{xo} * x^{(t)} + w_{ho} * h^{(t-1)} + b_o) \tag{14}$$
$$g^{(t)} = \tanh(w_{xg} * x^{(t)} + w_{hg} * h^{(t-1)} + b_g) \tag{15}$$
$$c^{(t)} = f^{(t)} \odot c^{(t-1)} + i^{(t)} \odot g^{(t)} \tag{16}$$
$$h^{(t)} = o^{(t)} \odot \sigma_h(c^{(t)}) \tag{17}$$

where $*$ and $\odot$ denote the convolution operator and Hadamard product, respectively. The input $x^{(t)}$, cell output $c^{(t)}$, hidden state $h^{(t)}$, and gates $i^{(t)}, f^{(t)}, o^{(t)}$ of ConvLSTM are 3D tensors (rows×columns×channels), while those in FC-LSTM are 1D tensors (channels only).

We designed one attention model after ConvLSTM. The output gate of ConvLSTM, $o^{(t)}$ is a 3D tensor, supposed to be $K \times K \times D$ without loss of generality. And our attention model is computed as follows.

$$u^{(t)} = Relu(W_k * o^{(t)} + b_k) \tag{18}$$
$$\alpha^{(t)} = p(F_t = t \mid u^{(t)}) = \frac{\exp(W_t{}^T u^{(t)})}{\sum_{l=1}^{L} \exp(W_l{}^T)u^{(t)}}, \quad t \in 1 \ldots L, \tag{19}$$
$$s^{(t)} = E_{p(F_t=t|u^{(t)})}(o^{(t)}) = \sum_{t=1}^{L} \alpha^{(t)} o^{(t)} \tag{20}$$

where *Relu*() stands for a rectified linear units activation function. Attention kernel weight W_k with size 1×1 is used to capture spatial information of feature vector. W_k and $o^{(t)}$ have the same channel number of D. Zero-padding is needed in convolution operator of attention map to make sure the attention map size not be reduced. W_t^T and b_k are parameters to learn, standing for weight and bias, respectively. Eq. (9) is different from Eq. (18) in that, the later one uses the convolution operator while the former uses the Hadamard product. Especially, our attention model in ConvLSTM not only focus on which frames are relevant to the video class, but also on the important regions in a frame. In other words, the attention model in ConvLSTM is both temporary and spatial.

D. JOINT OPTIMIZATION MODULE

We designed one joint optimization module for training ConvLSTM and FC-LSTM feature vectors together, in the purpose of fully exploring semantic, temporary and spatial features of video. Each feature vector, produced by these two LSTM networks with their attention model, contains discriminative characteristics, and there are intrinsic relations. General methods for action recognition always train two independent classifiers for each kind of feature vector, but this will miss the relations. So we jointly process two vectors by one classifier to explore their intrinsic relations. With twice the amount of vector data, the classifier will be trained more robustly for recognition.

We use cross-entropy as the cost function, and the object function is defined as follows,

$$\Phi = \arg\max \sum_{n=1}^{2} \sum_{i=1}^{C} y_{n,i} \log(\hat{y}_i) \tag{21}$$

where $y_{n,i}$ stands for the predicted label vector, $\log \hat{y}_i$ for the class probability of true label vector, C for the number of action categories, and n denotes the number of tasks.

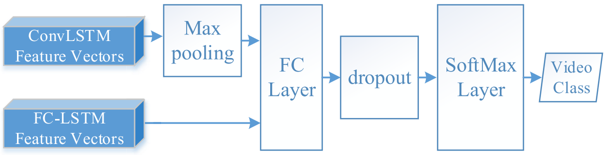


FIGURE 5. Joint optimization module.

The joint optimization module is depicted in Fig. 5. Before ConvLSTM feature passing through the joint optimization layer, max pooling is needed to sample the vector to the same size as FC-LSTM feature.

Our joint optimization module includes a fully connected (FC) layer and a SoftMax layer. Between FC and SoftMax layer, a dropout operation is used to prevent the model from over-fitting.

The JOM treats each feature vector as a separate classification task with the same label, and generates two class scores. During training, losses of two tasks are summed up as the final loss. In test process, we multiply two class scores produced by SoftMax function as the final score for the prediction.

IV. EXPERIMENT AND ANALYSIS

To evaluate the proposed method, experiments for video action recognition have been present in this section. First, we will give the descriptions of the datasets we have used. Following, experiment setting will be introduced, as well as the experimental results and analysis.

A. DATASETS

Three public datasets have been used in our experiment, i.e. UCF-11 [33], UCF Sports [34] and UCF-101 [35]. The action categories are listed in Table 1.

UCF-11 dataset contains 11 action categories with a total of 1600 videos, collected from YouTube. 955 and 645 videos are used for training and testing, respectively.

UCF Sports dataset contains 150 video sequence in 10 action categories, collected from television channels including BBC and ESPN. 75% of the videos in the dataset are used for training and 25% for testing.

UCF-101 dataset contains 13320 video clips in 101 categories, collected from YouTube. The dataset provides

TABLE 1. Information of Datasets used in our experiments.

Dataset	Video clips	Action Categories
UCF-11	1600	basketball shooting, biking/cycling, diving, golf swinging, horseback riding, soccer juggling, swinging, tennis swinging, trampoline jumping, volleyball spiking, and walking with a dog
UCF Sports	150	Diving, Golf Swing, Kicking, Lifting, Riding Horse, Running, Skate Boarding, Swing-Bench, Swing-Side, Walking
UCF-101	13320	Apply Eye Makeup, Apply Lipstick, Archery, Baby Crawling, Balance Beam, Band Marching, Baseball Pitch, Basketball Shooting, Basketball Dunk, Bench Press, Biking, Billiards Shot, Blow Dry Hair, Blowing Candles, Body Weight Squats, Bowling, Boxing Punching Bag, Boxing Speed Bag, Breaststroke, Brushing Teeth, Clean and Jerk, Cliff Diving, Cricket Bowling, Cricket Shot, Cutting In Kitchen, Diving, Drumming, Fencing, Field Hockey Penalty, Floor Gymnastics, Frisbee Catch, Front Crawl, Golf Swing, Haircut, Hammer Throw, Hammering, Handstand Pushups, Handstand Walking, Head Massage, High Jump, Horse Race, Horse Riding, Hula Hoop, Ice Dancing, Javelin Throw, Juggling Balls, Jump Rope, Jumping Jack, Kayaking, Knitting, Long Jump, Lunges, Military Parade, Mixing Batter, Mopping Floor, Nun chucks, Parallel Bars, Pizza Tossing, Playing Guitar, Playing Piano, Playing Tabla, Playing Violin, Playing Cello, Playing Daf, Playing Dhol, Playing Flute, Playing Sitar, Pole Vault, Pommel Horse, Pull Ups, Punch, Push Ups, Rafting, Rock Climbing Indoor, Rope Climbing, Rowing, Salsa Spins, Shaving Beard, Shotput, Skate Boarding, Skiing, Skijet, Sky Diving, Soccer Juggling, Soccer Penalty, Still Rings, Sumo Wrestling, Surfing, Swing, Table Tennis Shot, Tai Chi, Tennis Swing, Throw Discus, Trampoline Jumping, Typing, Uneven Bars, Volleyball Spiking, Walking with a dog, Wall Pushups, Writing On Board, Yo Yo.

TABLE 2. Accuracy on UCF-11. (a) Convolution with VGG-16 Net. (b) Convolution with ResNet-152.

(a)

Methods	without Attention	with Attention
FC-LSTM	94.57%	95.66%
ConvLSTM	96.59%	96.74%
Proposed Method	**97.52%**	**98.45%**

(b)

Methods	without Attention	with Attention
FC-LSTM	91.62%	91.78%
ConvLSTM	92.40%	98.14%
Proposed Method	**95.35%**	**98.76%**

three train-test splits, and we use the first split which has 9537 videos for training and 3783 videos for testing.

B. IMPLEMENTATION DETAILS

Video sequences have different number of frames. The most favorable time range of LSTM is a group of 40 frames [29], [30], which proves to be a good compromise between performance and complexity. In our experiments, 40 frames in each video are sampled equally to the recognition architecture. Data augmentation is also made by symmetric extension for training. Experiments are implemented based on Tensorflow and the code is available at https://github.com/Qingyang-Xu/DTA.

Adaptive Moment Estimation (Adam) is used for optimization of the proposed network, with the learn rate fixed to be 10^{-3} and the batch size fixed to be 60. In our FC-LSTM model, we use a two-layers LSTM structure, and the convolution kernel size of ConvLSTM is set to be 3×3 to capture fined feature information of videos.

C. RESULTS AND DISCUSSIONS

Firstly, the proposed method is evaluated on UCF-11, UCF Sports, and UCF-101 dataset to illustrate the efficiency of two kinds of LSTM and attention modal, and the results have been listed in Table 2, Table 3, and Table 4, respectively.

TABLE 3. Accuracy on UCF Sports. (a) Convolution with VGG-16 Net. (b) Convolution with ResNet-152.

(a)

Methods	without Attention	with Attention
FC-LSTM	72.97%	81.08%
ConvLSTM	78.38%	81.08%
Proposed Method	**83.78%**	**91.89%**

(b)

Methods	without Attention	with Attention
FC-LSTM	83.78%	83.78%
ConvLSTM	81.08%	86.48%
Proposed Method	**86.48%**	**91.89%**

TABLE 4. Accuracy on UCF-101.

Methods	without Attention	with Attention
FC-LSTM	77.42%	79.00%
ConvLSTM	73.53%	75.80%
Proposed Method	**82.62%**	**84.10%**

From the results, we can see that, the proposed method consistently performs better than methods using FC-LSTM and ConvLSTM separately. ConvLSTM performs better than FC-LSTM with an improvement of 1%∼5% percent, which proves the importance of the spatial features. And the proposed method performs better than ConvLSTM. This demonstrates that it's important and necessary to use both semantic and spatial information in temporal domain. In this way, more useful feature information can be explored for action recognition.

Experimental results also show that the attention module helps to improve the recognition accuracy. Especially the performance improvement is obvious on UCF Sports and UCF-101 dataset. The temporal attention module is very useful and performs much better than that without attention.

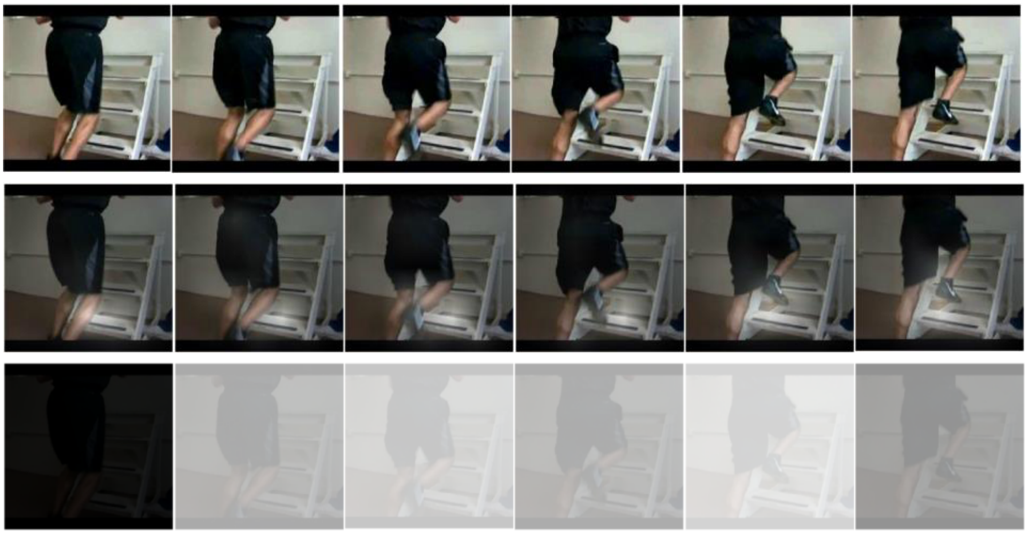


FIGURE 6. Visual attention of ConvLSTM in spatial and temporal. Top-row: Original video frames; Mid-row: ConvLSTM attention results in each frame; Bottom-row: ConvLSTM attention results on temporal frames.

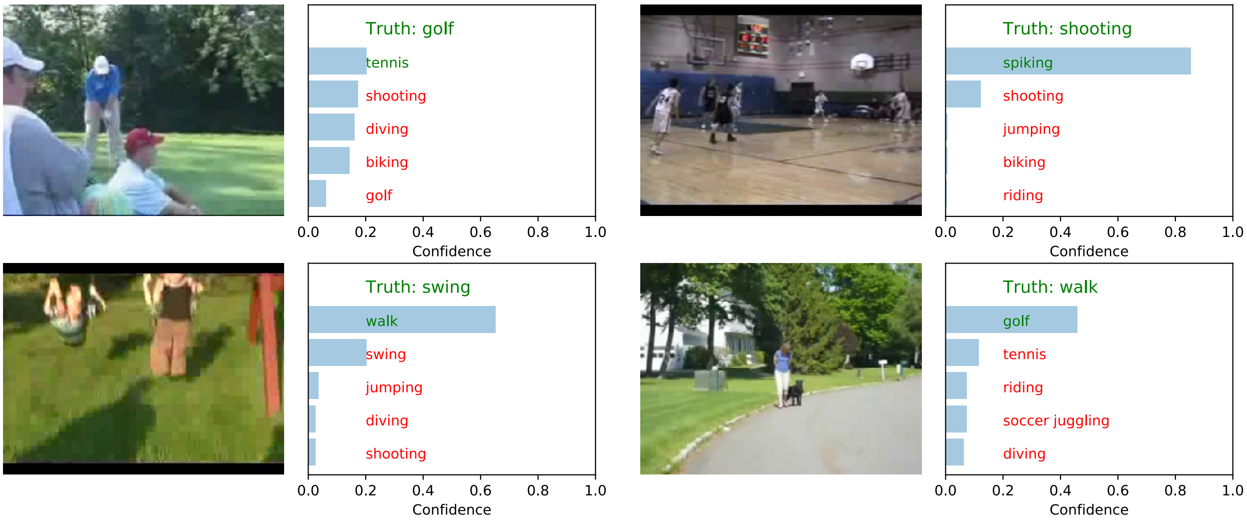


FIGURE 7. Examples of failed predictions using the proposed method.

TABLE 5. Comparison with other methods on UCF-11.

Methods	Accuracy
Incremental Activity (Hasan *et al.*, 2014) [36]	54.50%
Hybrid Features (Liu *et al.*, 2009) [33]	71.20%
Dense Trajectories (Wang *et al.*, 2011) [37]	84.20%
Soft attention (Sharma *et al.*, 2015) [30]	84.90%
Binary CNN-FLow (Ravanbakhsh *et al.*, 2015) [38]	84.30%
Two Stream LSTM (Gammulle *et al.*, 2017) [28]	94.60%
Proposed Method (Image Frames)	**98.76%**

TABLE 6. Comparison with other methods on UCF Sports.

Methods	Accuracy
Spatio-temporal Features (Wang *et al.*, 2009) [39]	85.60%
Hierarchical Feature (Le *et al.*, 2011) [40]	86.50%
Neighborhood Features (Kovashka *et al.*, 2010) [41]	87.20%
Dense Trajectories (Wang *et al.*, 2011) [37]	89.10%
Action Localization (Weinzaepfel *et al.*, 2015) [42]	90.50%
SGSH (Abdulmunem *et al.*, 2016) [43]	90.90%
Binary CNN-FLow (Ravanbakhsh *et al.*, 2015) [38]	**94.80%**
Proposed Method (Image Frames)	91.89%

For the convolution layers, we take both VGG-16 Net and ResNet-152 for testing, and ResNet-152 performs slightly better. This demonstrates that LSTM training is more important for videos. Due to the long training time, VGG-16 Net has not been used for UCF-101 dataset.

The performance of the proposed method has also been compared with other methods, including Incremental Activity [36], Hybrid Features [33], Spatio-temporal Features [39], Dense Trajectories [37], Two Stream LSTM [28], Soft attention [30], Two Stream ConvNet [26], 3D Fusion of

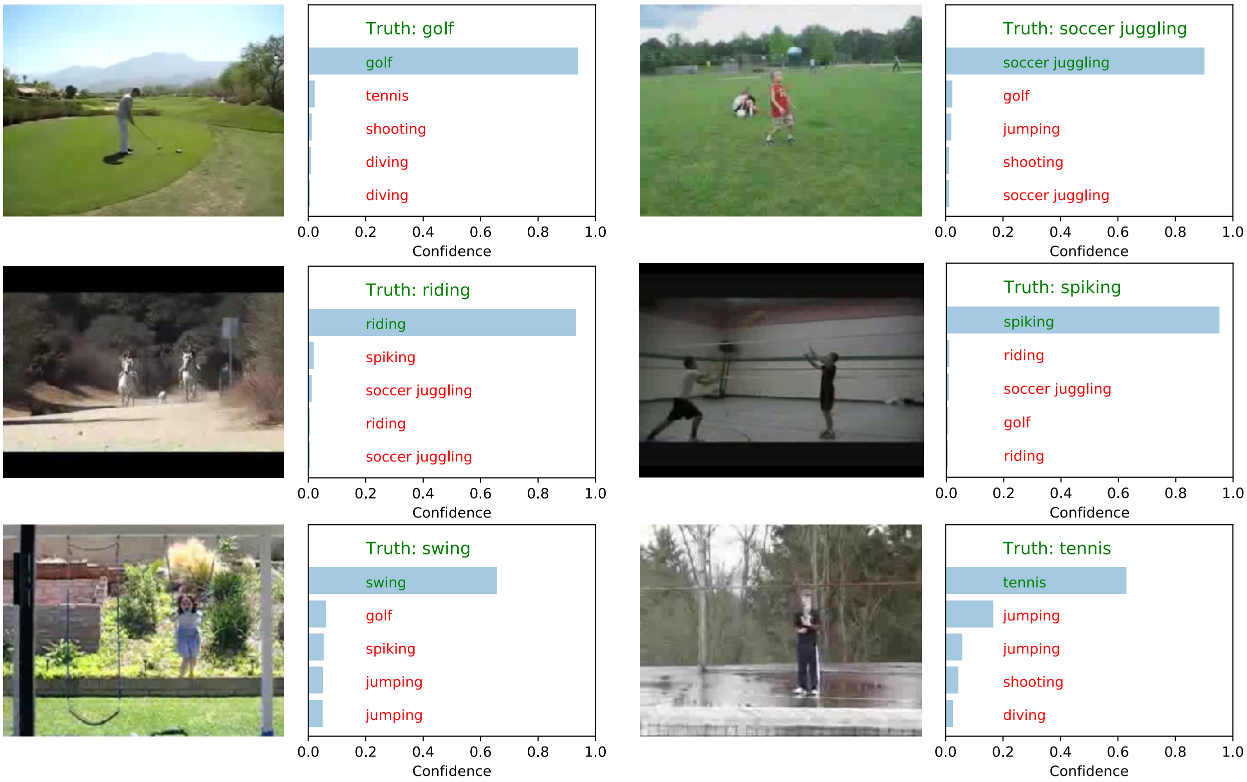


FIGURE 8. Examples of successful predictions using the proposed method.

TABLE 7. Comparison with other methods on UCF-101.

Methods	Accuracy
Single Frame CNN Model (Images Frames) (Simonyan *et al.*, 2014) [26]	73.00%
Single Frame CNN Model (Optical Flow) (Simonyan *et al.*, 2014) [26]	73.90%
Two stream ConvNet (Optical Flow + Image Frames, SVM Fusion) (Simonyan *et al.*, 2014) [26]	87.00%
3D Fusion of Two-Stream ConvNet (Optical Flow + Image Frames) (Feichtenhofer *et al.*, 2016) [27]	90.40%
Two-Stream I3D (RGB) (Carreir *et al.*, 2017) [44]	84.50%
Two-Stream I3D (Optical Flow) (Carreir *et al.*, 2017) [44]	90.60%
Two-Stream I3D (RGB + Optical Flow) (Carreir *et al.*, 2017) [44]	**93.40%**
Proposed Method (Image Frames)	84.10%

Two-Stream ConvNet [27] and Two-Stream I3D [44]. From Table 5 and 6, we can see that, the proposed method achieved better performance on UCF-11 and UCF Sports than the others methods. From Table 7, we can see that for UCF-101 dataset, methods based on two-stream networks perform better when using both RGB and optical flow data, while the proposed method just used RGB data. One problem of computing optical flow is the extra burden for both software and hardware, which makes it difficult in real-time applications. The proposed method performs better than Single Frame CNN Model [26] and competitively with Two-Stream I3D [44] when using RGB data only.

Visualization of the learned attention region is provided in Fig. 6. Original video frames are given in Fig. 6(a). Corresponding attention region in each frame is shown in Fig. 6(b). Attention weighted frames in temporal domain are shown in Fig. 6(c). From Fig. 6(b), we can see that, the regions around legs in a frame are brighter, which means more important. In Fig. 6(c), several frames with bigger range movement are brighter than others, meaning that they are paid more attention than other frames. The temporal-wise attention model encourages the model to focus on specific parts of specific frames during training, which will be more discriminative for classification.

Fig. 7 and Fig. 8 give failed and successful predictions for some test sequences, respectively. From Fig. 7, we can see these actions are very similar with the wrong classes in terms of both the appearance and motion patterns.

Since only RGB data is used, the computational complexity of the proposed method is low, and the test time for one video sequence is about 0.17 seconds.

V. CONCLUSION

In this paper, we propose a new lightweight architecture for video action recognition, which consists of CNN, LSTM and attention model. This architecture is designed according to human visual mechanism in the purpose of obtaining strong representational power for prediction just using RGB data, without optical flow data which needs additional computations. We use a convolution model to extract two kinds of features (semantic, spatial) for each frame, and followed by two kinds of LSTMs (FC-LSTM, ConvLSTM) with their temporal-wise attention model. We also designed a JOM to optimize the classifier and explore the intrinsic relations between feature vectors. The proposed recognition architecture has been tested on three widely used dataset, and achieved state-of-the-art performance compared with existing methods. In the future work, we will try to modify this network in pooling or fusion layers.

ACKNOWLEDGMENT

The authors would like to thank the editor and reviewers for their work on this manuscript.

REFERENCES

[1] R. Poppe, "A survey on vision-based human action recognition," *Image Vis. Comput.*, vol. 28, no. 6, pp. 976–990, 2010.

[2] Y. Guo, X. Hu, B. Hu, J. Cheng, M. Zhou, and R. Y. K. Kwok, "Mobile cyber physical systems: Current challenges and future networking applications," *IEEE Access*, vol. 6, pp. 12360–12368, 2018, doi: 10.1109/ACCESS.2017.2782881.

[3] Z. Ning *et al.*, "A cooperative quality-aware service access system for social Internet of vehicles," *IEEE Internet Things J.*, to be published, doi: 10.1109/JIOT.2017.2764259.

[4] X. Hu, T. H. S. Chu, H. C. B. Chan, and V. C. M. Leung, "Vita: A crowdsensing-oriented mobile cyber-physical system," *IEEE Trans. Emerg. Topics Comput.*, vol. 1, no. 1, pp. 148–165, Jun. 2013.

[5] X. Hu, T. H. S. Chu, V. C. M. Leung, E. C.-H. Ngai, P. Kruchten, and H. C. B. Chan, "A survey on mobile social networks: Applications, platforms, system architectures, and future research directions," *IEEE Commun. Surveys Tuts.*, vol. 17, no. 3, pp. 1557–1581, 3rd Quart., 2015.

[6] X. Ji, J. Cheng, D. Tao, X. Wu, and W. Feng, "The spatial Laplacian and temporal energy pyramid representation for human action recognition using depth sequences," *Knowl.-Based Syst.*, vol. 122, pp. 64–74, Apr. 2017.

[7] X. Ji, J. Cheng, W. Feng, and D. Tao, "Skeleton embedded motion body partition for human action recognition using depth sequences," *Signal Process.*, vol. 143, pp. 56–68, Feb. 2018.

[8] A.-A. Liu, N. Xu, W.-Z. Nie, Y.-T. Su, Y. Wong, and M. Kankanhalli, "Benchmarking a multimodal and multiview and interactive dataset for human action recognition," *IEEE Trans. Cybern.*, vol. 47, no. 7, pp. 1781–1794, Jul. 2017.

[9] A. Klaser, M. Marszałek, and C. Schmid, "A spatio-temporal descriptor based on 3D-gradients," in *Proc. 19th Brit. Mach. Vis. Conf. (BMVC)*, 2008, pp. 99.1-99.10.

[10] R. V. H. M. Colque, C. Caetano, M. T. L. de Andrade, and W. R. Schwartz, "Histograms of optical flow orientation and magnitude and entropy to detect anomalous events in videos," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 27, no. 3, pp. 673–682, Mar. 2017.

[11] H. Wang and C. Schmid, "Action recognition with improved trajectories," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Sydney, NSW, Australia, Dec. 2013, pp. 3551–3558.

[12] N. Dalal, B. Triggs, and C. Schmid, "Human detection using oriented histograms of flow and appearance," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2006, pp. 428–441.

[13] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, 2006.

[14] C. Farabet, C. Couprie, L. Najman, and Y. LeCun, "Learning hierarchical features for scene labeling," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 8, pp. 1915–1929, Aug. 2013.

[15] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, May 2015.

[16] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun. (2013). "OverFeat: Integrated recognition, localization and detection using convolutional networks." [Online]. Available: https://arxiv.org/abs/1312.6229

[17] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, vol. 25. Dec. 2012, pp. 1097–1105.

[18] 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. (CVPR)*, Columbus, OH, USA, Jun. 2014, pp. 1725–1732.

[19] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[20] J. Donahue *et al.*, "Long-term recurrent convolutional networks for visual recognition and description," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 4, pp. 677–691, Apr. 2017.

[21] M.-T. Luong, H. Pham, and C. D. Manning. (2015). "Effective approaches to attention-based neural machine translation." [Online]. Available: https://arxiv.org/abs/1508.04025

[22] K. Xu *et al.*, "Show, attend and tell: Neural image caption generation with visual attention," in *Proc. Int. Conf. Mach. Learn. (ICML)*, Jun. 2015, pp. 2048–2057.

[23] X. Shi *et al.*, "Convolutional LSTM network: A machine learning approach for precipitation nowcasting," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2015, pp. 802–810.

[24] 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.

[25] 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.

[26] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, vol. 1. 2014, pp. 568–576.

[27] 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)*, Las Vegas, NV, USA, Apr. 2016, pp. 1933–1941.

[28] H. Gammulle, S. Denman, S. Sridharan, and C. Fookes, "Two stream LSTM: A deep fusion framework for human action recognition," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2017, pp. 177–186.

[29] J. Y. 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.

[30] S. Sharma, R. Kiros, and R. Salakhutdinov. (2015). "Action recognition using visual attention." [Online]. Available: https://arxiv.org/abs/1511.04119

[31] Y. Wang, S. Wang, J. Tang, N. O'Hare, Y. Chang, and B. Li, "Hierarchical attention network for action recognition in videos," *CoRR*, 2016. [Online]. Available: https://arxiv.org/abs/1607.06416

[32] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 770–778.

[33] J. Liu, J. Luo, and M. Shah, "Recognizing realistic actions from videos 'in the wild,'" in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2009, pp. 1996–2003.

[34] K. Soomro and A. R. Zamir, "Action recognition in realistic sports videos," in *Computer Vision in Sports* (Advances in Computer Vision and Pattern Recognition). Cham, Switzerland: Springer, 2014, pp. 181–208.

[35] K. Soomro, A. R. Zamir, and M. Shah, "UCF101: A dataset of 101 human action classes from videos in the wild," Center Res. Comput. Vis., Univ. Central Florida, Orlando, FL, USA, Tech. Rep. CRCV-TR-12-01, Nov. 2012. [Online]. Available: http://crcv.ucf.edu/data/UCF101.php

[36] M. Hasan and A. K. Roy-Chowdhury, "Incremental activity modeling and recognition in streaming videos," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Columbus, OH, USA, Jun. 2014, pp. 796–803.

[37] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu, "Action recognition by dense trajectories," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Providence, RI, USA, Jun. 2011, pp. 3169–3176.

[38] M. Ravanbakhsh, H. Mousavi, M. Rastegari, V. Murino, and L. S. Davis. (Dec. 2015). "Action recognition with image based CNN features." [Online]. Available: https://arxiv.org/abs/1512.03980

[39] H. Wang, M. M. Ullah, A. Klaser, I. Laptev, and C. Schmid, "Evaluation of local spatio-temporal features for action recognition," in *Proc. Brit. Mach. Vis. Conf. (BMVC)*, 2009, pp. 124.1–124.11.

[40] Q. V. Le, W. Y. Zou, S. Y. Yeung, and A. Y. Ng, "Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Providence, RI, USA, Jun. 2011, pp. 3361–3368.

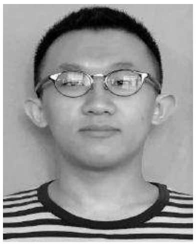
[41] A. Kovashka and K. Grauman, "Learning a hierarchy of discriminative space-time neighborhood features for human action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, San Francisco, CA, USA, Jun. 2010, pp. 2046–2053.

[42] P. Weinzaepfel, Z. Harchaoui, and C. Schmid, "Learning to track for spatio-temporal action localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Santiago, Chile, Dec. 2015, pp. 3164–3172.

[43] A. Abdulmunem, Y.-K. Lai, and X. Sun, "Saliency guided local and global descriptors for effective action recognition," *Comput. Vis. Media*, vol. 2, no. 1, pp. 97–106, Mar. 2016.

[44] J. Carreira and A. Zisserman. (2017). "Quo vadis, action recognition? A new model and the Kinetics dataset." [Online]. Available: https://arxiv.org/abs/1705.07750

LEI WANG received the Ph.D. degree in electrical engineering from Xidian University, China, in 2010. He was with Huawei Technologies Company Ltd., from 2011 to 2012, and the University of Jinan from 2012 to 2016. From 2014 to 2015, he was with the Department of Embedded Systems Engineering, Incheon National University, as a Post-Doctoral Fellow. He is currently an Associate Professor with the Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences. His research interests include image transforms, video coding, machine learning, and video analysis.



YANGYANG XU received the B.E. degree from the Yantai University, China, in 2015. He is currently pursuing the M.E. degree with Guangxi Normal University. Since 2017, he has been with the Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, as an intern. His research interests include deep learning and video action recognition.

JUN CHENG received the B.E. and M.E. degrees from the University of Science and Technology of China, Hefei, China, in 1999 and 2002, respectively, and the Ph.D. degree from The Chinese University of Hong Kong, Hong Kong, in 2006. He is currently with the Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China, as a Professor, and the Director of the Laboratory for Human Machine Control. His current research interests include computer vision, robotics, and machine intelligence and control.

HAIYING XIA received the M.E. and Ph.D. degrees from the Department of Electronic and Information Engineering, Huazhong University of Science and Technology, Wuhan, China, in 2007 and 2011, respectively. She is currently an Associate Professor with Guangxi Normal University. Her current research includes pattern recognition, medical image analysis, and neural networks.

JIANQIN YIN received the Ph.D. degree in control science and engineering from Shandong University in 2013. She is currently an Associate Professor with the School of Automation, Beijing University of Posts and Telecommunications, Beijing, China. Her current research interests include machine learning, video analysis, and action recognition.

JIAJI WU (M'07) received the B.S. degree in electrical engineering from Xidian University, Xi'an, China, in 1996, the M.S. degree from the National Time Service Center, Chinese Academy of Sciences, in 2002, and the Ph.D. degree in electrical engineering from Xidian University in 2005. He is currently a Professor with Xidian University. His current research interests include still image coding, hyperspectral/multispectral image compression, communication, and high-performance computing.