Multimed Tools Appl (2017) 76:15065–15081
DOI 10.1007/s11042-017-4514-3

Exploring hybrid spatio-temporal convolutional networks for human action recognition

Hao Wang[1] · Yanhua Yang[1] · Erkun Yang[1] · Cheng Deng[1,2]

Received: 27 November 2016 / Revised: 31 January 2017 / Accepted: 14 February 2017 /
Published online: 8 March 2017

Abstract Convolutional neural networks have achieved great success in many computer vision tasks. However, it is still challenging for action recognition in videos due to the intrinsically complicated space-time correlation and computational difficult of videos. Existing methods usually neglect the fusion of long term spatio-temporal information. In this paper, we propose a novel hybrid spatio-temporal convolutional network for action recognition. Specifically, we integrate three different type of streams into the network: (1) the image stream utilizes still images to learn the appearance information; (2) the optical stream captures the motion information from optical flow frames; (3) the dynamic image stream explores the appearance information and motion information simultaneously from generated dynamic images. Finally, a weighted fusion strategy at the softmax layer is utilized to make the class decision. With the help of these three streams, we can take full advantage of the spatio-temporal information of the videos. Extensive experiments on two popular human action recognition datasets demonstrate the superiority of our proposed method when compared with several state-of-the-art approaches.

Keywords Human action recognition · Convolutional network · Spatio-temporal information · Approximate rank pooling · Weighted fusion

✉ Cheng Deng
chdeng@mail.xidian.edu.cn

Hao Wang
hwang_3@stu.xidian.edu.cn

Yanhua Yang
yanghuahua.xd@gmail.com

Erkun Yang
ekyang@stu.xidian.edu.cn

[1] Department of Electronic and Engineering, Xidian University, Xi'an 710071, China

[2] The State Key Laboratory of Integrated Services Networks (ISN), Xidian University, Xi'an 710071, China

1 Introduction

With the tremendous growth of video capturing devices and storage space, video data have increased explosively. Meantime, human action recognition in videos has attracted much attention in the computer vision community due to the wide applications in video surveillance, multimedia analysis, human-computer interaction and healthcare. Although many efforts have been devoted into this domain, it is still challenging for action recognition because of the following two main reasons: (1) the low quality of video such as low resolution, camera motion and cluttered background; (2) the large intra-class variances due to the different motion speeds, intensity of illumination, and viewpoints. The crucial step for dealing with these obstacles is to design robust feature extraction method. As far as we known, there are mainly two categories of video feature representation in action recognition, hand-crafted features and deep-learned ones.

In the first category, the researchers use the hand-crafted local features including Cuboids [6], Space Time Interest Points [17], improved Dense Trajectories [31] and so on. The extraction of these features often consists of two steps: key points detection and feature extraction. However, the dimension of these local features is relative high and it thus increases the computational complexity. Besides, these hand-crafted local features may lack discriminative capacity for action recognition and are not optimal for visual representation.

In the second category, the researchers develop various deep neural network architectures to extract features, which have achieved great success in visual recognition recently. Based on the type of network architectures, we can further divide these works into three types. The first type of architecture uses 2D convolutional neural networks [2, 8, 13, 25, 32]. These architectures can utilize the power of pre-trained model on image recognition, but lose the capability for capturing spatio-temporal information simultaneously. The second type of architecture uses 3D convolutional neural networks [4, 11, 29]. They extend the 2D convolutional filters to 3D and apply them into the action tubes to capture spatio-temporal information simultaneously. Although this architecture exactly suits the video data structure for modeling spatio-temporal information, the initialization of parameters in network can not utilize the existing models which are pre-trained on large scale labelled image datasets. The third type of architecture is a hybrid [7, 38] of convolutional neural networks and recurrent neural networks since the recurrent neural networks can model the temporal information better. But the procedure for joint training is complex and the optimal solution is hard to be obtained.

In this paper, we proposed a novel hybrid spatio-temporal convolutional network by combining dynamic image stream with spatial and temporal streams to take full advantage of spatio-temporal information. In order to understand the information these three streams contain, we give an example of RGB image, optical flow image and dynamic image in Fig. 1. Our work is mainly inspired by Two-Stream networks [25], although the combination of spatial stream and temporal stream can fuse the appearance and motion information to obtain better performance, we find that it still has limit in temporal modeling because the spatial stream only trained on single still frames and the temporal stream which used optical flow loses much appearance details. To better incorporate the appearance information and the motion information, we introduce a novel dynamic image stream into the whole architecture. By using a ranking machine [2] to encode temporal evolution of frames in video, dynamic image can preserve details of objects as well as the motion information in a relative long time period simultaneously.

This paper has three main contributions:

(1) An improved dynamic image network is proposed and evaluated to show that dynamic images can capture spatio-temporal information simultaneously.
(2) We propose a novel hybrid spatio-temporal convolutional network by combining dynamic image stream with spatial stream and temporal stream to explore spatio-temporal information.
(3) Our approach obtains state-of-the-art performance on HMDB51 dataset (70.4 %) and comparable performance on UCF101 dataset (94.1 %).

The rest of this paper is organized as follows. In Section 2, some related networks are introduced and discussed. Section 3 provides the proposed approach, including network architectures, training and testing details. The experimental results are presented and discussed in Section 4. We draw the conclusions in Section 5 finally.

2 Related work

Researchers have devoted much efforts to design discriminative feature representations and effective classifiers in action recognition for decades. Many local image features have been generalized to videos such as 3D SIFT [21], extended SURF [37], 3D HOG [14]. These local features extracted around the detected interest points represent the 3D volumes. Recently the improved Dense Trajectories [31] has shown to be successful on a number of challenging datasets. Specifically, the information is encoded by HOG, HOF and MBH along with the trajectory to represent the action in video. However, these local features are designed specifically and hard to be generalized to other scenarios. Besides, they also lack enough high-level semantic information. To deal with these issues, Actons [48] and Action Banks [20] was proposed. In order to utilize the relations between action categories, Yang et al. [45] proposed a novel method based on multi-task learning framework with super-category. Alfaro et al. [1] proposed a novel scheme to quantify relative intra and inter-class similarities among local temporal patterns. For the action recognition on RGBD datasets, several methods have been proposed such as multi-modal multi-part learning framework [22], discriminative multi-instance multi-task learning framework (MIMTL) [43], bilinear heterogeneous information machine [15], and latent max-margin multi-task learning framework [44]. There are also some interesting works related to surveillance system, such as video structural description (VSD) [39–42] which represents and organizes the content in videos, and correspondence structure learning [24] for person re-identification and so on.

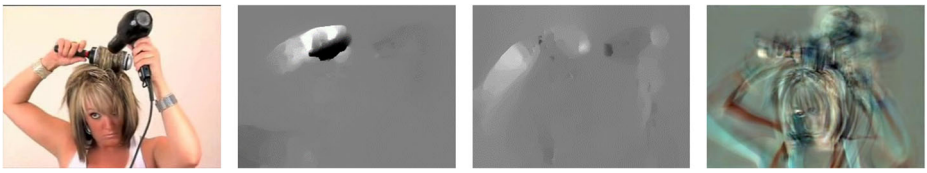

Fig. 1 Examples of RGB image, optical flow image and dynamic image. **a**: The RGB image contains scenes and objects information. **b**, **c**: The optical flow (x, y directions) reflect the motion information. **d**: The dynamic image contains appearance and motion information both

These hand-crafted methods are highly depended on expert knowledge to design and can not be trained in an efficient end-to-end way.

The great success of deep convolutional neural networks has attracted researchers to utilize deep features for action recognition. Ji et al. [11] extended 2D convolutional filters to 3D to apply them to the video tubes directly. The C3D model [29] used 3D convolutions and 3D pooling to explore spatio-temporal information simultaneously. These 3D models can not utilize the existing models which are pre-trained on large scale labelled image datasets. Some works use recurrent neural network to model the temporal evolution such as LRCN [7] and hybrid of CNN and LSTM framework [38]. But the complex joint training procedure and much parameters make it hard to obtain optimal soluiton. When the attention mechanism is introduced in action recognition, a simple soft attention mechanism [23], Video LSTM model [18] and two stream hierarchical attention model [36] are proposed. But these methods need specifically designed regularizer to guide the attention mechanism. In order to overcome the problem of limited temporal modeling, Feichtenhofer et al. [8] proposed several spatiotemporal fusion methods of video snippets. There are also excellent works [4, 47] for accelerating the speed of action recognition while preserving acceptable performance. Among these deep-learned methods, the most representative work is Two-Stream ConvNets [25] which uses two individually trained and complementary streams, i.e., spatial stream and temporal stream. This method firstly obtained comparable performance with hand-crafted features and we design our model based on it in this paper.

Recently, researchers have devoted much efforts to the temporal modeling due to the temporal evolution information is more discriminative for action recognition. The Temporal Segment Networks [34] segments the video into several clips and does sparse sampling in each clip to model the long term temporal information. Their experimental results show that the model can focus on useful information on the whole video. Bilen et al. [2] proposed a novel compact representation of videos: dynamic images. The dynamic images are generated by encoding the order of each frame in the video to capture the dynamics evolution.

Among these approaches, the Temporal Segment Networks [34] and Dynamic Image Networks [2] are most close to us. They both focus on modeling a long term temporal evolution and explore the appearance information and motion information to improve the performance. However, Temporal Segment Networks just divides the video into several clips and it is still hard to capture dynamic information between RGB frames in spatial stream. While dynamic images naturally capture a relative long term dynamics due to that it is generated by encoding a length of L (e.g. $L = 20$) consecutive frames and meanwhile it conveys complementary information both with still images and optical flow images. In this paper, we adopt dynamic image stream as the third stream and combine it with original two streams to take full advantage of spatio-temporal information.

3 Proposed approach

In this section, we describe the proposed hybrid spatio-temporal convolutional network for action recognition in details. Firstly, the overall frameworks is presented in Section 3.1. Then we describe the network architectures, training details in Section 3.2. Finally, the testing details is introduced in Section 3.3.

3.1 Hybrid spatio-temporal convolutional neural network

In this section, we propose a novel hybrid spatio-temporal convolutional network for action recognition and the framework is illustrated in Fig. 2. This framework contains

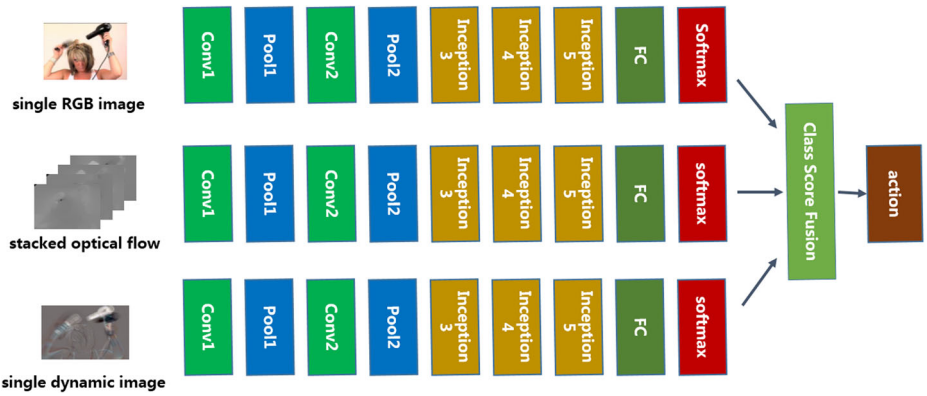

Fig. 2 Hybrid spatio-temporal convolutional networks: the network consists of three inputs, single RGB image, stacked optical flow and single dynamic image. The class scores is fused at the softmax layer to combine the motion and appearance information. Notice that each Inception block here has more than one Inception layers

three individual and complementary streams: spatial stream, motion stream and dynamic image stream. In the spatial stream, the RGB still images which contain the appearance information are processed. Similarly, the optical flow images which contain motion information are processed in the temporal stream. In the dynamic image stream, the dynamic images which take the correlation of space-time are processed. In the following we give a detailed description of each stream and summarize the advantages of our proposed scheme.

Spatial stream In this stream, the frames of the whole video have same label regardless that they are different from each other. We input the still images and obtain class scores at the softmax layer of this stream. As we can see, still RGB images contain static appearance information such as color, texture, particular scenes and objects. These information are strongly associated with the performed action. For example, the bow always exists in the archery action and horse always exists in the horse riding action. However, the limits of using still RGB images in action recognition are obvious. The cluttered background of video would decrease the performance and different action may have similar patterns in RGB images. For example, smiling and laughing, as well as walking and jogging. Only using static appearance information would result in confusion of which action is exactly performed.

Temporal stream This stream is intended to model the motion evolution of action. Here we use stacked optical flow fields to represent a motion pattern during a period of time and use them as the inputs of the stream. In this stream, various stacked optical flow in a video have same class label and they are calculated to obtain class scores at the softmax layer. Here we give a detailed description of optical flow. Assuming the intensity of light is basically consistent in corresponding region, optical flow is calculated via the relative movement between two consecutive frames. As an example in Fig. 3, we use $\mathbf{d}_t(x, y)$ to denote the displacement vector at the point (x, y) in the tth frame, which reflects the movement from current point to corresponding point in the following $(t + 1)$th frame. The $\mathbf{d}_t(x, y)$

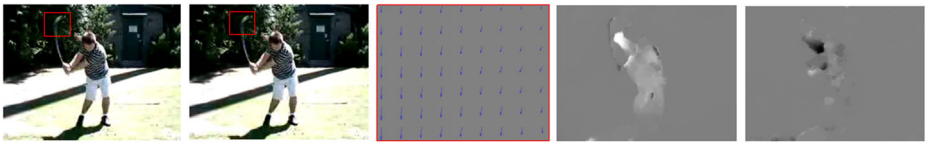

Fig. 3 Optical flow. **a**, **b**: Two consecutive frames with the area around a moving golf stick outlined with a red rectangle. **c**: A close-up of dense optical flow in the outlined area. **d**, **e**: The horizontal and vertical component of displacement vector fields

is composed of the horizontal vector $d_t^h(x, y)$ and vertical vector $d_t^v(x, y)$. Based on the consistency between two consecutive frames, we obtain the equation as bellow:

$$I(x, y, t) = I\left(x + d_t^h, y + d_t^v, t + 1\right), \tag{1}$$

where $I(x, y, t)$ represents the image function(i.e. gray value) of pixel at the location of (x, y) at time t. The linearized version of the equation by using the first-order Taylor approximation is illustrated as

$$\begin{aligned} I(x, y, t) &\approx I(x, y, t + 1) + \nabla I(x, y, t + 1)^T \mathbf{d}_t(x, y) \\ 0 &= \underbrace{I(x, y, t + 1) - I(x, y, t)}_{I_t(x,y,t+1)} + \nabla I(x, y, t + 1)^T \mathbf{d}_t(x, y). \end{aligned} \tag{2}$$

Then we obtain the optical flow constraint (OFC) equation as bellow:

$$OFC(d_t^h, d_t^v): \quad 0 = I_t + I_x d_t^h + I_y d_t^v, \tag{3}$$

where the partial derivatives of image function (i.e. gray value) are denoted as I_t, I_x, and I_y. Finally, we use this constraint and various methods to obtain the optical flow vector d_t^h and d_t^h. The methods and equations are complex so we do not illustrate more details in this paper due to we only use optical flow as one of the feature representations.

The experimental results show that the optical flow information is more discriminative than still RGB appearance information. However, it is ambiguous because of a single optical flow characterizes accurate motion information such as the moving violently block of the current frame. Besides, it also can be affected by subtle motion of camera.

Dynamic image stream This stream is an important component of our proposed work. We utilize the appearance and long term dynamics which are encoded in dynamic images to model the correlation of space and time. In this stream, the dynamic images are treated as RGB images for training and testing and calculated to obtain class scores. We give an example in Fig. 4 and clearly observe that the background is removed and the motion pattern is shown in dynamic image.

In this section, we firstly give the formulation used to generate dynamic images. Then we present the derivation of approximate rank pooling method due to its good balance between efficiency and accuracy. Finally, we describe the pipeline of the generation. It should be noticed that the process of generation is basically following the original work.

(a) (b) (c) (d)

Fig. 4 Generation of dynamic images. **a**, **b**, **c**: RGB frames during a period of time. **d**: The generated dynamic images. We can clearly observe that the background is removed and the motion pattern is shown in dynamic image

The core idea of dynamic image is encoding the order of frames into the video representations. The objective function for obtaining a optimal dynamic image d is presented bellow:

$$E(d) = \frac{\lambda}{2}\|d\|^2 + \frac{2}{T(T-1)} \times \sum_{q>t} \max\{0, 1 - S(q|d) + S(t|d)\}, \tag{4}$$

where the first term is usual quadratic regularizer and the second term is penalty for incorrectly ranking pairs. In here,

$$q > t \Rightarrow S(q|d) > S(t|d), \tag{5}$$

and

$$S(i|d) = \langle d, V_i \rangle, \quad i = q, t, \tag{6}$$

V_i is the representation at time i which includes the information happened in past, and q, t are time steps. Then we give the derivation of approximate rank pooling is as bellow:

$$\begin{aligned} \nabla E(\mathbf{0}) &\propto \sum_{q>t} \nabla \max\{0, 1 - S(q|d) + S(t|d)\}|_{d=\mathbf{0}} \\ &\propto \sum_{q>t} \nabla \max\{0, 1 - \langle d, V_q \rangle + \langle d, V_t \rangle\}|_{d=\mathbf{0}} \\ &= \sum_{q>t} \nabla \langle d, V_t - V_q \rangle = \sum_{q>t} V_t - V_q. \end{aligned} \tag{7}$$

Then (7) can be formulated into (8),

$$d^* \propto \sum_{q>t} V_q - V_t = \sum_{q>t} \left[\frac{1}{q} \sum_{i=1}^{q} \phi_i - \frac{1}{t} \sum_{j=1}^{t} \phi_j \right] = \sum_{t=1}^{T} \alpha_t \phi_t, \tag{8}$$

where ϕ is the feature vector extracted from one single frame and V is the time varying mean representations of frames. The final coefficient α_t is presented as bellow:

$$\alpha_t = 2(T - t + 1) - (T + 1)(H_T - H_{t-1}), \tag{9}$$

where $H_t = \sum_{i=1}^{t} 1/i$ and we set $H_0 = 0$.

For the generation process, we firstly choose T consecutive frames and perform a non-linear transformation (e.g. square root operation) to each frame. Then we use the coefficients calculated above to generate the initial dynamic images. Finally a minmax normalization for each color channel is performed and final dynamic image is merged from them.

The combination of spatial stream, temporal stream and dynamic image stream can model the whole action better by capturing appearance information and motion information. The dynamic image stream we added play an important role as richer feature representation

which modeling appearance and long term motion information simultaneously. It can alleviate the problems caused by spatial stream (e.g. the spatial stream is only trained on single still image) and temporal stream (e.g. the temporal stream only capture relative short term dynamics).

3.2 Network training

In this paper, we adopt the Inception Network [28] with Batch Normalization [10] as building block. The Inception Unit has three convolution subunits and one pooling subunit meanwhile the size of 5×5 filter is replaced with two 3×3 size filters. In here adding the Batch Normalization unit could accelerate the convergence speed. To further improve the capability of modeling temporal information, we adopt the temporal segment skills [34] to divide the whole input video into several clips and do spare sampling in each clip at training stage. The fusion of these three streams are performed at the softmax layer. In here we treat the generated dynamic images as RGB images for training and testing. To better improve the capability of generalization, the dropout layer is added and dropout ratio used is set to 0.8 for spatial stream network, 0.7 for temporal stream network and 0.8 for dynamic image stream network.

3.3 Network testing

The network inputs have three types: RGB images x_a which contain appearance information, stack of optical flow fields x_m which contain motion information, dynamic images x_d which contain appearance and motion information both. These three inputs go through the convolutional neural network to obtain the class scores of each input. For each training example $x = \{x_a, x_m, x_d\}$ with the label $k \in \{1, 2, ..., K\}$, we compute the class probability $p(k|x) = exp(z_k)/\sum_{i=1}^{K} exp(z_i)$. Here z_i are unnormalized log probabilities. Then based on these three class scores, a weighted fusion is performed

$$p(k|x) = w_a\, p(k|x_a) + w_m\, p(k|x_m) + w_d\, p(k|x_d) \tag{10}$$

to obtain the final class scores. The effect of fusing three stream is not like in the original two stream. Because the fusion of three stream can be seemed as three combination: combination of spatial stream and dynamic image stream, combination of temporal stream and dynamic image stream and combination of pure spatial stream and pure temporal stream. The fusion is illustrated in Fig. 5. In our view this fusion includes three weighted combination of two stream and the experimental results show the superiority of our approach.

4 Experiments

In this section, we firstly introduce the existing two large and popular action recognition datasets: UCF101 [27] dataset and HMDB51 [16] dataset in Section 4.1. The training and testing details for each stream is presented in Section 4.2. The effect of using deeper network is evaluated in Section 4.3. Then the combination of dynamic images with single RGB and stacked optical flow is evaluated in Section 4.4. The results show that the dynamic images not only capture the appearance information but also capture the motion information. In Section 4.5 we compare our proposed approach to state-of-the-arts to show the superiority of our approach and we also give some examples to show why our approach can improve the performance.

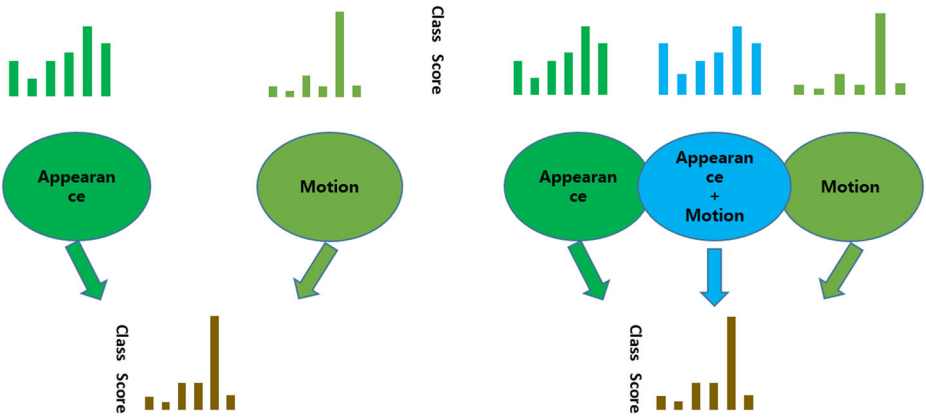

Fig. 5 Original two stream fusion v.s. Three-Stream fusion: compare with two stream fusion, our approach utilizes dynamic images to better explore spatio-temporal information

4.1 Datasets and evaluation protocol

In order to verify the effectiveness of our proposed method, extensive experiments are performed in the following challenging datasets. Both of them are widely used in action recognition. We follow the standard setup and report average accuracy over splits.

UCF101 The dataset annotated 13320 videos into 101 action categories which can be divided into sports video, human motion, human-object interaction and so on. The videos are collected from YouTube and in each category the videos are further divided into 25 groups. In each group the background or viewpoint is similar.

HMDB51 The dataset includes 6766 videos which are divided into 51 action categories. The videos are collected from movies, videos on Google, YouTube and so on. Due to different sources the background is cluttered and viewpoint is variable.

We follow the evaluation protocol provided by the organizers of datasets: each of them has three splits for training and testing, and the final accuracy is averaged across the splits. The details of datasets are presented in Table 1 and the mean accuracy precision is defined as bellow:

$$P = \sum_{i=1,2,\ldots,C} P_i / C \tag{11}$$

Where P_i is the accuracy of each category and C is the number of total categories.

4.2 Experiments settings

Training stage In spatial stream, the sizes of frames extracted from the video is various from video to video. So we follow the operation from the original two stream work and make the smallest side of frames equal 256. Then a 224×224 region is randomly cropped. Also we adopt the randomly horizontal flipping and scale jittering. At training stage, the learning rate starts from 0.001 which is reduced by a factor of 10 every 1000 iterations and stops at 2500 iterations. We use a pre-trained Inception network model on ImageNet [5] to initialize the parameters of spatial network. The batch size is 32 and split segment is 3.

Table 1 Details of datasets

	UCF101	HMDB51
Action category	101	51
Action type	5	5
Total video clips	13320	6766
Video clips for training	≈ 9600	3570
Video clips for testing	≈ 3700	1530

In temporal stream, we use a stack of 10 consecutive optical flow fields as input. The optical flow we use in this paper is TVL1 [46] optical flow, which could be computed in OpenCV with GPU. The optical flow fields are linearly re-scaled to a [0,255] range and stored as JPEG form to avoid to save them as float type, which would significantly save the space of storage. The learning rate starts from 0.005 which is reduced by a factor of 10 at 12k, 18k iterations and stops at 20k iterations. The batch size is 30 and we use a modified pre-trained Inception network model on ImageNet to initialize the parameters of temporal network. Specifically, the channels of first layer in pre-trained model is averaged and then copied N times. Here N denotes the number of channel in the first layer for temporal network.

In dynamic image stream, we generate the dynamic images by following the approximate rank pooling operations from the original dynamic image networks. The window size used for extracting dynamic images is 20 for UCF101 dataset and 10 for HMDB51 dataset. The stride is 1 for both datasets. The setting of window size and stride are adopted from the Discriminative Hierarchical Dynamic Image Networks [9]. For UCF101 dataset, the learning rate starts from 0.001 which is reduced by a factor of 10 every 3000 iterations and stops at 6500 iterations. For HMDB51 dataset we use the same learning rate, iterations and batch size as spatial stream.

Testing stage Notice that we treat dynamic images as RGB images for training and testing. At testing stage we extract 25 frames from one video with equal temporal interval and average the scores of all 25 samples to compute the final class scores in spatial stream and dynamic image stream. As for temporal stream, we also extract 25 samples but the difference is that each sample consists of 10 consecutive fields (x, y directions of 5 consecutive optical flow).

4.3 Exploration of deeper network for dynamic images

In the original Dynamic Image Networks [2], the researchers used CaffeNet [12] which only has five convolutional layers and three fully connected layers and did not use any data aug-

Table 2 Evaluation of using deeper network for dynamic images

Method	UCF101	HMDB51
Rank pooling [2]	72.2 %	40.9 %
Recursive rank pooling [9]	75.6 %	45.8 %
Hierarchical rank pooling [9]	78.8 %	47.5 %
Deeper rank pooling(our)	83.5 %	53.6 %

Table 3 Comparison between original dynamic image network and our improved version

	Original network [2]	Our network
Based network architecture	CaffeNet-based	Inception-based with batch normalization
Data augmentation technology	No	Yes(multi-scale crop, flip)
Long-term modeling technology	No	Temporal segment technology [34]

mentation skills, so the performance is relative low. In Discriminative Hierarchical Dynamic Image Networks [9], the researchers used VGG16 [26] network to extract frame features and then performed hierarchical rank pooling to obtain higher order dynamic images, however, the process of extracting dynamic images becomes much complex due to the hierarchical operation. We use the Inception with Batch Normalization as building block due to its good balance between accuracy and efficiency. We observe that the result for dataset UCF101 increases from 72.2 % to 83.3 % and for dataset HMDB51 increases from 40.9 % to 53.6 %. Compared with vanilla rank pooling and hierarchical rank pooling , the experimental results show that using deeper network with more data augmentation skills could improve the performance significantly and the results are illustrated in Table 2. Table 3 shows the most significant difference between original dynamic image network and our improved dynamic image network.

It should be noticed that the results are compared without combination of hand-crafted features.

4.4 Effectiveness analysis

In this section, we combine the dynamic image stream with appearance stream and temporal stream respectively and the results are illustrated in Table 4. When we remove each of the three streams, the performance decrease is presented in Table 5. The results show that the motion information is crucial for action recognition, and dynamic image stream we proposed is effective to improve the performance by capturing both appearance information and motion information.

When we combine the spatial stream and dynamic image stream as two stream, we observe that the result increases both on UCF101 dataset (over spatial stream 3.0 % and over dynamic image stream 4.4 %) and on HMDB51 dataset (over spatial stream 5.9 % and over dynamic image stream 5.3 %). This proves that the dynamic images can capture motion information to improve the performance on spatial stream. When we combine the temporal stream with dynamic image stream, we observe the similar improvement on

Table 4 Evaluation of combination between three streams

Component	UCF101	HMDB51
RGB image	84.9 %	53.0 %
Optical flow	89.7 %	62.1 %
Dynamic image	83.5 %	53.6 %
RGB image + dynamic image	87.9 %	58.9 %
Optical flow + dynamic image	92.4 %	66.5 %
RGB image + optical flow	94.0 %	69.4 %
RGB image + optical flow + dynamic image	94.1 %	70.4 %

Table 5 Performance decrease of removing each of three streams

Removed component	UCF101	HMDB51
Without spatial stream	1.7 %↓	3.9 %↓
Without temporal stream	6.2 %↓	11.5 %↓
Without dynamic image stream	0.1 %↓	1.0 %↓

UCF101 dataset (2.7 % over temporal stream and 8.9 % over dynamic image stream) and on HMDB51 dataset (4.4 % over temporal stream and 12.9 % over the dynamic image stream). This shows that the dynamic image is complementary with optical flow. Notice that we use the weight 1 for spatial stream and 1 for dynamic image stream on UCF101 dataset when we combine these two streams. Similarly, we set the weight of spatial stream as 1 and set the weight of dynamic image stream as 1.2 on HMDB51 dataset. As for the combination of temporal stream and dynamic image stream, we use the equal weight on HMDB51 dataset and 1:0.7 on UCF101 dataset. The weight is 1:1.5 when we combine the spatial stream and temporal stream for both datasets. From the division of the weight and the improvement on each stream, we observe that dynamic image stream performs better on HMDB51 due to that the background of HMDB51 is more cluttered than UCF101.

4.5 Comparison with the state-of-the-arts

As we observed in Section 4.4, the dynamic images can not only model the motion information but also model the appearance information. And in original two stream, the spatial stream only trained on single frame and the length L of stacked optical flow is relative small (e.g. 10). These two shortcomings would degrade the performance for action recognition. So we combine the dynamic image with the original two stream to propose a novel hybrid spatio-temporal convolutional networks. The results presented in Table 6 shows that our approach outperforms the state-of-the-art method by 1.0 % on the HMDB51 dataset. The weight of spatial stream, temporal stream and dynamic image stream is set to 0.8, 1.0, and 0.1 on UCF101 dataset. While the weight of spatial stream, temporal stream and dynamic image stream is set to 1.1, 2.0, and 0.7 on HMDB51 dataset. In here, we compare our

Table 6 Comparison with the state-of-the-arts

Method	UCF101	HMDB51
DT+MVSV [3]	83.5 %	55.9 %
iDT+FV [31]	85.9 %	57.2 %
iDT+HSV [19]	87.9 %	61.1 %
MoFAP [33]	88.3 %	61.7 %
Two Stream [25]	88.0 %	59.4 %
TDD+FV [32]	90.3 %	63.2 %
LTC [30]	91.7 %	64.8 %
KVMF [49]	93.1 %	63.3 %
Transformation CNN [35]	92.4 %	63.4 %
Two Stream Fusion [8]	93.5 %	69.2 %
TSN [34]	94.2 %	69.4 %
Three Stream	94.1 %	70.4 %

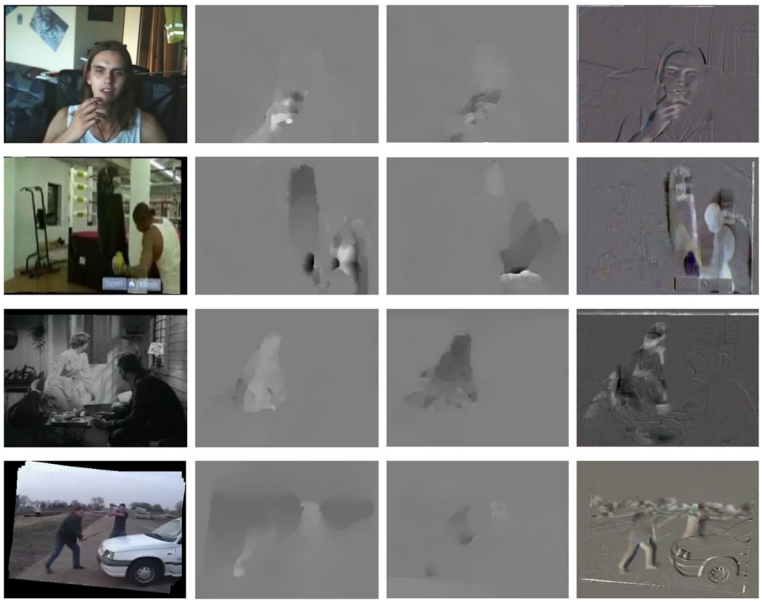

Fig. 6 In the *first row*, smoking is incorrectly classified as laughing in two stream while is correctly classified in our method. In the second row, punching is incorrectly classified as kicking in two stream while is correctly classified in our method. In the *third row*, standing is incorrectly classified as sitting in two stream while is correctly classified in our method. In the *fourth row*, hitting is incorrectly classified as fencing in two stream while is correctly classified in our method

approach with hand-crafted features based methods and deep features based methods both. Specifically, we choose improved Dense Trajectories (iDT) [31], MoFAP [33] method in hand-crafted features and Two-Stream ConvNets [25], Convolutional Fusion of Two Stream [8], Temporal Segment Networks (TSN) [34], trajectory-pooled deep convolutional descriptors (TDD) [32], long term convolution networks (LTC) [30], transformation CNN [35] and key volume mining framework (KVMF) [49] in deep-learned features.

Besides recognition accuracies, we want to attain further insight about why our approach can improve performance. From Fig. 6, we can observe that on spatial stream, the appearance information of smoking is similar to the appearance information of laughing. So the weight of appearance stream is relative high and it may result in this failure. However, the dynamic image stream can strength the motion information on spatial stream, so it is classified correctly.

5 Conclusion

In this paper, we firstly explore the deeper network for dynamic images and reveal that the dynamic images can not only capture appearance information but also motion information. Based on this, we proposed a novel hybrid spatio-temporal convolutional network by combining dynamic image stream with original two stream to explore spatio-temporal information. The fusion of three stream shows superiority compared with several state-of-the-art methods.

Acknowledgements The authors would like to thank the Editor-in-Chief, the handling associate editor and all anonymous reviewers for their considerations and suggestions. This work was supported by the National Natural Science Foundation of China (61572388).

References

1. Alfaro A, Mery D, Soto A (2016) Action recognition in video using sparse coding and relative features. In: Conference on computer vision and pattern recognition (CVPR), 2016, I.E. IEEE, pp 2688–2697
2. Bilen H, Fernando B, Gavves E, Vedaldi A, Gould S (2016) Dynamic image networks for action recognition. In: Conference on computer vision and pattern recognition (CVPR), 2016, I.E. IEEE, pp 3034–3042
3. Cai Z, Wang LM, Peng X, Qiao Y (2014) Multi-view super vector for action recognition. In: Conference on computer vision and pattern recognition (CVPR), 2014, I.E. IEEE, pp 596–603
4. Diba A, Pazandeh A, Gool LV (2016) Efficient two-stream motion and appearance 3d CNNs for video classfication. arXiv:1608.08851
5. Deng J, Dong W, Socher R, Li LJ, Li K, Li F-F (2009) Imagenet: a large-scale hierarchical image database. In: Conference on computer vision and pattern recognition (CVPR), 2009, I.E. IEEE, pp 248–255
6. Dollar P, Rabaud V, Cottrell G, Belongie S (2005) Behavior recognition via sparse spatio-temporal features. In: VS-PETS 2005
7. Donahue J, Hendricks LA, Rohrbach M, Venugopalan S, Guadarrama S, Saenko K, Darrel T (2015) Long-term recurrent convolutional networks for visual recognition and description. In: Conference on computer vision and pattern recognition (CVPR), 2015, I.E. IEEE, pp 2625–2634
8. Feichtenhofer C, Pinz A, Zisserman A (2016) Convolutional two-stream networks fusion for video action recognition. arXiv:1604.06573
9. Fernando B, Anderson P, Hutter M, Gounld S (2016) Discriminative hierarchical rank pooling for activity recognition. In: Conference on computer vision and pattern recognition (CVPR), 2016, I.E. IEEE, pp 1924–1932
10. Ioffe S, Szegedy C (2015) Batch normalization: accelerating deep network training by reducing internal covariate shift. In: International conference on machine learning (ICML), 2015, pp 448–456
11. Ji SW, Xu W, Yang M, Yu K (2013) 3D convolutional neural networks for human action recognition. IEEE Trans Pattern Anal Mach Intell (PAMI) 35(1):221–231
12. Jia YQ, Evan S, Jeff D, Sergey K, Jonathan L, Ross G, Sergio G, Trevor D (2014) Caffe: convolutional architecture for fast feature embedding. arXiv:1408.5093
13. Karpathy A, Toderici G, Shetty S, Leung T, Sukthankar R, Li Fei-Fei (2014) Large-scale video classification with convolutional neural networks. In: Conference on computer vision and pattern recognition (CVPR), 2014, I.E. IEEE, pp 1725–1732
14. Klaser A, Marszalek M, Schmid C (2008) A spatio-temporal descriptor based on 3D-gradients. In: 2008-19th British machine vision conference (BMVC), British machine vision association
15. Kong Y, Fu Y (2015) Bilinear heterogeneous information machine for rgbd action recognition. In: Conference on computer vision and pattern recognition (CVPR), 2015, I.E. IEEE, pp 1054–1062
16. Kuehne H, Jhuang H, Garrote E, Poggio T, Serre T (2011) HMDB: a large video database for human motion recognition. In: International conference on computer vision (ICCV), 2011, I.E. IEEE, pp 2556–2563
17. Laptev I (2005) On space-time interest points. Int J Comput Vis (IJCV) 64(2–3):107–123
18. Li ZY, Gavves E, Jain M, Snoek CGM (2016) VideoLSTM convolves, attends and flows for action recognition. arXiv:1607.01794
19. Peng X, Wang LM, Wang X, Qiao Y (2016) Bag of visual words and fusion methods for action recognition: comprehensive study and good practice. arXiv:14054506
20. Sadanand S, Corso JJ (2012) Action bank: a high-level representation of activity in video. In: Conference on computer vision and pattern recognition (CVPR), 2012, I.E. IEEE, pp 1234–1341
21. Scovanner P, Ali S, Mubarak Shah (2007) A 3-dimensional SIFT descriptor and its application to action recognition. In: ACM international conference on multimedia (ACM MM), pp 357–360
22. Shahroudy A, Ng TT, Yang Q, Wang G (2016) Multimodal multipart learning for action recognition in depth videos. IEEE Trans Pattern Anal Mach Intell (PAMI) 10:2123–2129
23. Sharma S, Kiros R, Salakhutdinov R (2015) Action recognition using visual attention. arXiv:1511.04119

24. Shen Y, Lin WY, Yan JC, Xu ML, Wu JX, Wang JD (2015) Person re-identification with correspondence structure learning. In: International conference on computer vision (ICCV), 2015, I.E. IEEE, pp 3200–3208
25. Simonyan K, Zisserman A (2014) Two-stream convolutional networks for action recognition in videos. In: Annual conference on neural information processing systems (NIPS), pp 568–576
26. Simonyan K, Zisserman A (2015) Very deep convolutional networks for large-scale image recognition. In: International conference on learning representations (ICLR), pp 1–14
27. Soomro K, Zamir AR, Shah M (2012) UCF101: a dataset of 101 human actions classes from videos in the wild. arXiv:1212.0402
28. Szegedy C, Vanhoucke V, Ioffe S, Shlens J (2016) Rethinking the inception architecture for computer vision. In: Conference on computer vision and pattern recognition (CVPR), 2016, I.E. IEEE, pp 2818–2826
29. Tran D, Bourdev L, Fergus R, Torresani L, Manohar Paluri (2015) Learning spatiotemporal features with 3d convolutional networks. In: International conference on computer vision (ICCV), 2015, I.E. IEEE, pp 4489–4497
30. Varol G, Laptev I, Schmid C (2016) Long-term temporal convolutions for action recognition. arXiv:1604.04994
31. Wang H, Schmid C (2013) Action recognition with improved trajectories. In: International conference on computer vision (ICCV), 2013, I.E. IEEE, pp 3551–3558
32. Wang LM, Qiao Y, XO T (2015) Action recognition with trajectory-pooled deep-convolutional descriptors. In: Conference on computer vision and pattern recognition (CVPR), 2015, I.E. IEEE, pp 4305–4314
33. Wang LM, Qiao Y, Tang XO (2016) MoFAP: a multi-level representation for action recognition. Int J Comput Vis (IJCV) 119(3):254–271
34. Wang LM, Xiong YJ, Wang Z, Qiao Y, Lin DH, Tang XO, Gool LV (2016) Temproal segment networks: towards good practices for deep action recognition. arXiv:1608.00859
35. Wang XL, Farhadi A, Gupta A (2016) Action ~ transformation. In: Conference on computer vision and pattern recognition (CVPR), 2016, I.E. IEEE, pp 2658–2667
36. Wang YL, Wang SH, Tang JL, O'Hare N, Chang Y, Li BX (2016) Hierarchical attention network for action recognition in videos. arXiv:1607.0641
37. Willems G, Tuytelaars T, Gool LV (2008) An efficient dense and scale-invariant spatio-temporal interest point detector. In: Proceedings of the european conference on computer vision (ECCV), pp 650–663
38. Wu ZX, Wang X, Jiang YG, Ye H, Xue XY (2015) Modeling spatial-temporal clues in a hybrid deep learning framework for video classification. In: ACM international conference on multimedia (ACM MM), pp 461–470
39. Xu Z, Hu CP, Mei L (2016) Video structured description technology based intelligence analysis of surveillance videos for public security applications. Multimedia Tools and Applications (MTAP) 75(19):12155–12172
40. Xu Z, Liu YH, Mei L, Hu CP, Chen L (2015) Semantic based representing and organizing surveillance big data using video structural description technology. J Syst Softw 102:217–225
41. Xu Z, Mei L, Hu CP, Liu YH (2016) The big data analytics and applications of the surveillance system using video structured description technology. Clust Comput 19(3):1283–1292
42. Xu Z, Mei L, Liu YH, Hu CP, Chen L (2016) Semantic enhanced cloud environment for surveillance data management using video structural description. Computing 98(1–2):35–54
43. Yang YH, Deng C, Gao SQ, Liu W, Tao DP, Gao XB (2016) Discriminative multi-instance multi-task learning for 3d action recognition. IEEE Trans Multimedia (TMM). doi:10.1109/TMM.2016.2626959
44. Yang YH, Deng C, Tao DP, Zhang ST, Liu W, Gao XB (2016) Latent max-margin multitask learning with skelets for 3d action recognition. IEEE Transactions on Cybernetics (TCYB) 99:1–10
45. Yang YH, Liu RS, Deng C, Gao XB (2016) Multi-task human action recognition via exploring super-category. Signal Process (SP) 124:36–44
46. Zach C, Pock T, Bischof H (2007) A duality based approach for realtime TV-l1 optical flow. In: 29th DAGM symposium on pattern recognition, pp 214–223
47. Zhang BW, Wang LM, Wang Z, Qiao Y, Wang HL (2016) Real-time action recognition with enhanced motion vector CNNs. In: Conference on computer vision and pattern recognition (CVPR), 2016, I.E. IEEE, pp 2718–2726
48. Zhu J, Wang BY, Yang XK, Zhang WJ, Tu ZW (2013) Action recognition with actons. In: International conference oncomputer vision (ICCV), 2013, I.E. IEEE, pp 3559–3566
49. Zhu WJ, Hu J, Sun G, Cao XD, Qiao Y (2016) A key volume mining deep framework for action recognition. In: Conference on computer vision and pattern recognition (CVPR), 2016, I.E. IEEE, pp 1991–1999

Hao Wang received the B.E degree in Electronic and Information Engineering from Hangzhou Dianzi University, China, in 2015. He is currently pursuing his Ph.D. degree at School of Electronic Engineering, Xidian University. His main research interests include action recognition and pose estimation.

Yanhua Yang received the B.E degree in Electronic and Information Engineering and M.S. degree in Signal and Information Processing from Xidian University, China, in 2004 and 2007, respectively. She is currently pursuing her Ph.D. degree at School of Electronic Engineering, Xidian University. Her main research interests include complex action recognition and event detection.

Erkun Yang received the B.E degree in Electronic and Information Engineering from Xidian University, China, in 2013. He is currently pursuing his Ph.D. degree at School of Electronic Engineering, Xidian University. His main research interests include computer vision and machine learning.

Cheng Deng (S'09) received the B.E., M.S., and Ph.D. degrees in signal and information processing from Xidian University, Xi'an, China. He is currently a Full Professor with the School of Electronic Engineering at Xidian University. His research interests include computer vision, multimedia processing and analysis, and information hiding. He is the author and coauthor of more than 50 scientific articles at top venues, including IEEE TNNLS, TMM, TCYB, TSMC, TIP, ICCV, CVPR, IJCAI, and AAAI.