Pattern Recognition Letters 107 (2018) 83–90

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

Going deeper with two-stream ConvNets for action recognition in video surveillance

Yamin Han[a], Peng Zhang[a,*], Tao Zhuo[b], Wei Huang[c], Yanning Zhang[a]

[a] School of Computer Science, Northwestern Polytechnical University, China PR
[b] Sensor-enhanced Social Media (SeSaMe) Centre, National University of Singapore, Singapore
[c] School of Information Engineering, Nanchang University, China

ARTICLE INFO

Article history:
Available online 12 August 2017

JEL classification:
41A05
41A10
65D05
65D17

Keywords:
Deeper
Two-stream
ConvNets
Action recognition
Video surveillance

ABSTRACT

Learning by deep convolutional networks have shown an outstanding effectiveness in a variety of vision based classification tasks, and for which, large datasets are the prerequisites to guarantee its high performance. But in many realistic circumstances, using a massive quantity of training samples to achieve more sophisticated analysis is hard to be fulfilled always, such as human action recognition in videos, and the resulting problem of data deficiency, especially for the labeled data, would critically limit the deeper model structure as a promising solution due to its high risk of overfitting. Additionally, in lacking of high modeling capacity constrained by of model depth, the high-level visual cues like object interaction, scene context and pose variations concurrent with human action also could become the extrinsic and intrinsic challenges for the traditional deep convolutional networks. For the limitations above, in this paper, we proposed a strategy of dataset remodeling by transferring parameters of ResNet-101 layers trained on the ImageNet dataset to initialize learning model and adopt an augmented data variation approach to overcome the overfitting challenge of sample deficiency. For model structure improvement, a novel deeper two-stream ConvNets has been designed for the learning of action complexity. With a dis-order strategy of training/testing video sets, the proposed model and learning strategy are able to collaboratively achieve a significant improvement of action recognition. Experiments on two challenging datasets UCF101 and KTH have verified a superior performance in comparison with other state-of-the-art methods.

1. Introduction

Owing to its potential applications in video surveillance [4], behavior understanding and video summarization, human action recognition in videos has drawn increasing attention in contemporary computer vision studies [1,3,9,16,20,23,27,31]. However, accurate action recognition is hard because the challenges in realistic scenarios may lead to the intra-class variations in the same action category, such as background clutter, scale change, dynamic viewpoint and abrupt motion in the datasets (e.g. HMDB51 [22] and UCF101 [29]). Meanwhile, the video inherent attributes, e.g. high dimension and low resolution, could further increase the difficulties of robust recognition [7,34]. Therefore, more abstractive representation [6,18] by imitating human understanding on videos has attracted up-to-date attentions in designing the classification model for action analysis.

In last decade, Convolutional Networks (ConvNets) [24] had obtained a series of breakthroughs in classification [21], detection [8], and recognition tasks [46]. The typical ConvNets based approaches including 3D ConvNets [16], Deep ConvNets [38] and two-stream ConvNets [27] had been applied to carry out more effective video-based action recognition. These approaches utilized the ConvNets trained on large-scale labeled datasets to learn a video representation from raw data and the two-stream ConvNets was the most competitive architecture among them. However, unlike being applied on image classification tasks [21], deep ConvNets failed to achieve a great improvement over traditional methods [35] for the reasons behind as: firstly, deep ConvNets based approaches require a large quantity of labeled samples for training, but unlike ImageNet dataset [5] containing thousands of data, the size of human action datasets are too smaller, which would lead to a high risk of overfitting when applying deeper ConvNets for training as on image classification. Secondly, the number of stacked layers in most

* Corresponding author.
E-mail addresses: yaminhan@mail.nwpu.edu.cn (Y. Han), zh0036ng@nwpu.edu.cn (P. Zhang).

http://dx.doi.org/10.1016/j.patrec.2017.08.015
0167-8655/© 2017 Published by Elsevier B.V.

current deep ConvNets is relatively sparse [27], e.g. the architecture of two-stream ConvNets only contains 5 convolutional layers and 3 fully-connected layers, it makes the ConvNets lack a high modeling capacity and be not able to handle a large categories of complex actions. Recent evidences suggested that the depth of networks is of crucial importance [28,30], which means that by exploiting very deep models, there might be a great performance enhancement benefiting from its high modeling capacity [8,25].

Motivated by analysis above, we proposed a novel deeper two-stream ConvNets for action recognition by adopting Residual Networks 101(ResNet-101) [11] as backbone. We removed layers after the pool5 layer of ResNet-101 and added an adaptation fully-connection layer, the output numbers of which were then related with the numbers of action classes of dataset. The proposed deeper model had a high modeling capacity which was able to deal with challenges of complex actions by building effective representations. For the overfitting problem, the parameters of ResNet-101 layers trained on the ImageNet were transferred to initialize the proposed model with an augmented data variation to increase the data diversity. By disordering the video sets listed in training/testing splits randomly to further enhance recognition accuracy, we can obtain a competitive performance on the UCF101 dataset [29] and KTH dataset [26] in comparison with the other state-of-the-art works.

The main contributions of this paper are summarized as below. Firstly, we designed a deeper two-stream ConvNets for action recognition by adapting recent very deep architectures-Residual Networks into video domain, which achieved comparable performance with the state-of-the-art two-stream approach [27]. Secondly, In order to solve the overfitting problems, we transferred parameters of layers trained on the ImageNet dataset to compute high-level representations for action recognition and proposed an augmented data variation strategy. Finally, we proposed a disordered strategy which has been utilized for video sets listed in training/testing splits to further enhance the final performance of evaluation.

2. Related work

In this section, we briefly introduced convolutional networks and deep learning based action recognition as the important preliminary knowledge to help understanding the technical descriptions presented in the proposed work.

ConvNets based Image Classification. Deep learning techniques had shown its outstanding effectiveness in a variety of image based tasks [9,11,12,19,21,28,30,45,46]. Over the past few years, researchers had proposed many well-known network structures for image based tasks. For ImageNet Classification, Krizhevsky et al. [21] proposed an AlexNet model that contained eight learned layers. Zeiler and Fergus [45] introduced a novel visualization technique and found an eight layer ConvNets model named ClarifaiNet that outperform AlexNet. Simonyan and Zisserman [28] investigated the effect of the convolutional network depth on its accuracy in the large-scale image recognition setting and proposed a deeper model, VGGNet (up to 19 layers). Szegedy et al. [30] proposed GoogleNet, a 22 layers deep network. Besides, He et al. [11] proposed the deepest networks RestNets up to more than 1000 layers at present. Analyzing the evolution of from AlexNet to RestNets, we find that the depth of ConvNets is deeper and deeper. This evidence revealed that on the challenging imageNet dataset, methods that exploited deeper models lead greatly performance due to its high modeling capacity.

Deep Learning based Action Recognition. Since great advances had been verified in image recognition tasks based on convolutional networks, the ConvNets largely promoted the development of action recognition in videos [16,20,23,27,27,31,32,40,41]. To analyze the video understanding more effectively, Ji et al. [16] adapted 2D ConvNets to 3D for video-based action recognition on relatively small datasets. Similarly, Taylor et al. [31] used a 3D convolutional RBMS to learn spatio-temporal features in unsupervised. Karpathy et al. [20] evaluated several deep ConvNets on large-scale video classification using a large dataset, called Sports-1M. However, these models did not capture the motion information properly, compared to shallow hand-crafted representation [35], they only obtained a lower performance. To explicitly model the motion pattern, Simonyan and Zisserman [27] designed two-stream ConvNets composed of spatial and temporal net. The spatial net mainly captured appearance features utilizing video frames as inputs, while the temporal net learnt effective motion features by using optical flow fields between two consecutive frames. Based on this mechanism, the two-stream ConvNets could eventually reach a state-of-the-art performance, and followed it with multiple semantic channels, Wang et al. [40] proposed a two-stream semantic region based CNNs to accomplish more complex action recognition.

Unfortunately, most of those current deep models lacked high modeling capacity which was constrained by depth of their models. In this work, a new deeper two-stream ConvNets for action recognition was proposed by introducing the strategies of CNN weights transferring, data variation augmentation and dis-ordering.

3. Proposed work

In this section, we began the presentation of the proposed work deliberatively. First of all, an overview of the deeper architecture based on two-stream ConvNets will be introduced. After that, the details about getting rid of overfitting effect during training will be discussed. In the end, we give the descriptions of network testing and a dis-ordering strategy for final recognition performance improvement.

3.1. Deeper two-stream convolutional networks

The two-stream ConvNets [27] had shown its outstanding performance in the task of video action recognition. The original two-stream ConvNets contains two separate recognition streams composed with spatial and temporal stream, and then combined by late fusion. The spatial nets performs action recognition from still video frames, and the temporal net is trained to recognize action classes from motion information in the form of dense optical flow. But as most deep models, training a deeper structure with the two-stream ConvNets is not easy due to the problem of vanishing or exploding gradients [10]. The vanishing gradients is a well known nuisance in neural networks with many layers [2], as the gradient information is back-propagated, repeated multiplication or convolution with small weights cannot render the gradient information notably in the earlier layers. Other several approaches tried to reduce such an effect practically through careful initialization [10,42] or Batch Normalization [13], but the performance enhancement is still limited.

Instead of hoping each few stacked layers directly fit a desired underlying mapping, the Residual Networks (ResNets) [11] makes use of identity shortcut connections that enable flow of information across layers without attenuation, and allows extremely deep networks structure up to more than hundreds of layers, which is just needed by the designing of a deeper two-stream ConvNets. In ResNets, a building block defined as:

$$y = F(x, \{W_i\}) + x. \tag{1}$$

Here, x and y are the input and output vectors of the layers considered, W_i denotes the weights of i_{th} layer. The function $F(x, \{W_i\})$ represents the residual mapping to be learned, e.g. in the left of Fig. 2 that has two layers, $F = W_2 \times ReLU(W_1 x)$. The operation $F + x$ is performed by a shortcut connection and element-wise addition,

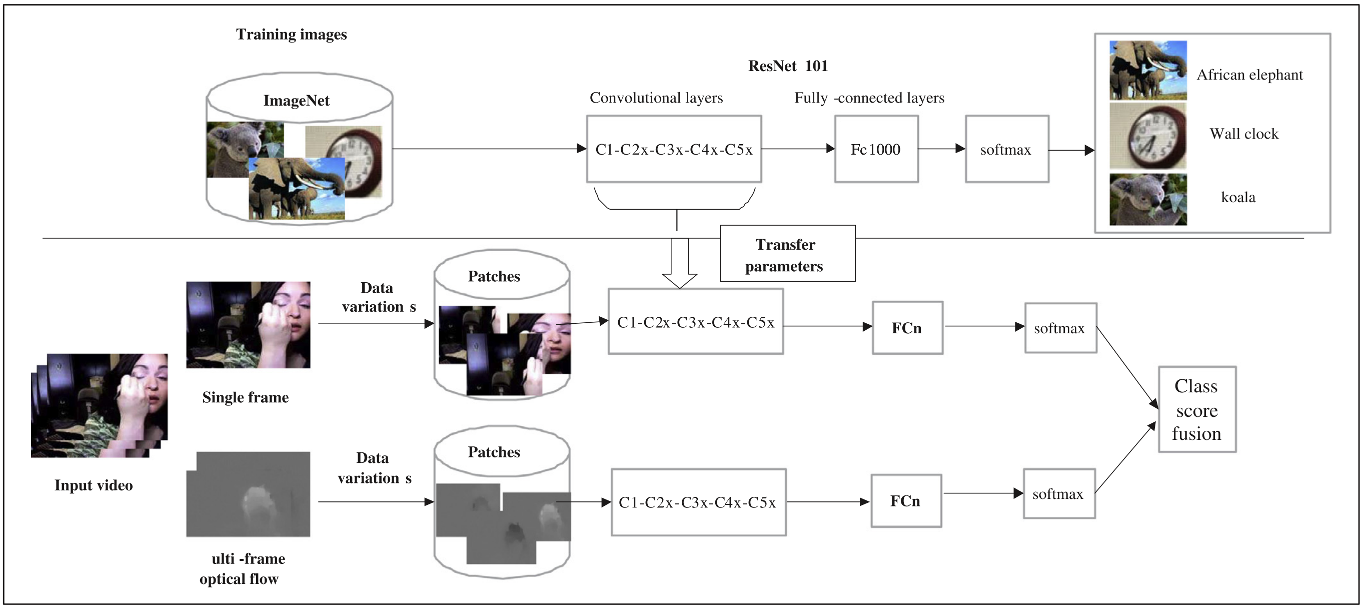


Fig. 1. Pipeline of our approach. We proposed a novel deeper two-stream ConvNets for the task of action recognition. First, we adapted the original ResNet101 to design of deeper two-stream ConvNets for action recognition in videos. Second, we proposed a strategy of dataset remodeling by transferring parameters of ResNet-101 layers trained on the ImageNet dataset to initialize learning model and adopted an augmented data variation strategy to overcome the overfitting challenge. At last, we predicted the scores of testing videos with our deeper two-stream ConvNets by late fusion.

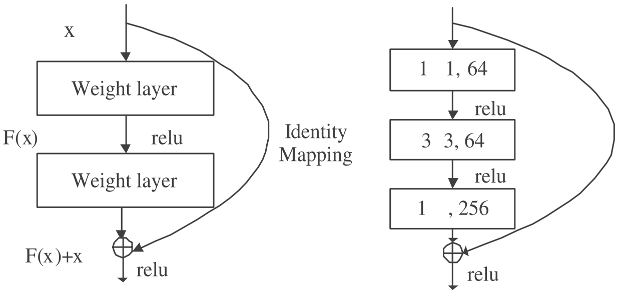


Fig. 2. A deeper residual function F.

which is performed channel by channel on two feature maps. In this process, the identity mapping shown in the left of Fig. 2 introduces neither extra parameters nor computation complexity, this is attractive in practice. The dimensions of x and F are required to be equal in Eq. (1). If not, a perform linear projection W_s can be performed to match the dimension as:

$$y = F(x, \{W_i\}) + W_s x. \tag{2}$$

In Eq. (2), the W_s is only used during matching dimensions. If dimensions of $F(x, \{W_i\})$ and x are identical, we can directly perform identity mapping using Eq. (1). Then, we involve a residual function F that has three layers with flexible form in the proposed work, which is shown in the right part of Fig. 2.

The overview of the proposed deeper two-stream ConvNets is shown in Fig. 1. In this structure, we constructed a deeper two-stream ConvNets based on a modified RestNets 101 by removing the layers after the pool5 layer of original ResNet-101 and adding an adaptation fully-connection layer, whose output numbers were related with the numbers of action classes of dataset. The networks architecture is the same for both spatial and temporal net except for the input data layer. Table 1 gives the details of the proposed ConvNets, where the spatial nets takes as input a square 224 × 224 pixel RGB image, and the input of temporal nets are volumes of stacking optical flow fields (224 × 224 × 2F, F is the number of stacking flows).

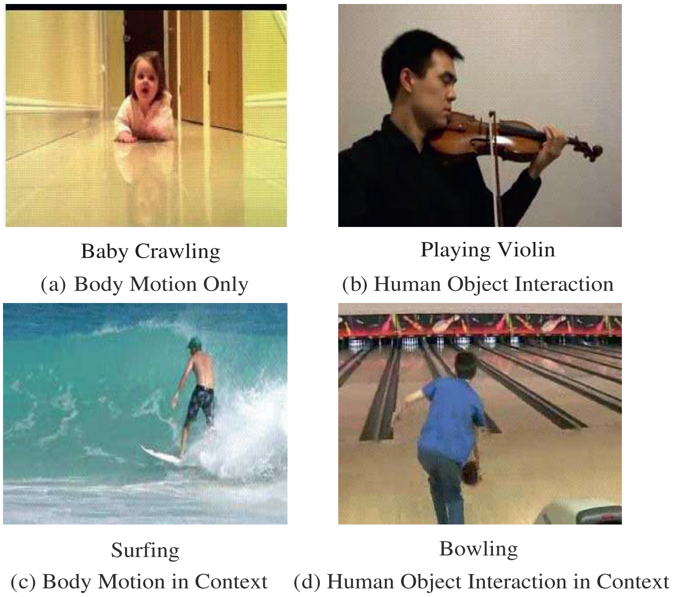
Table 1
ConvNets Architectures. The symbol of "[]" represents building blocks as shown in ResNet-101 [11]. e.g. In third rows, second column, 1 × 1 represents that the kernel size of convolutional layer is 1. The channels of convolutional layer is 64. In Conv2_x, there are 3 stacked building blocks. The final output size of Conv2_x is 56 × 56. Down sampling is performed by Conv3_x, Conv4_x, and Conv5_x. Our ConvNets architectures are similar with ResNet-101. We remove layers after the pool5 layer of original ResNet-101 and add an adaptation fully-connection layer FCn for the task of action recognition, where n is the numbers of action classes in dataset.

Layer	101-layer		Output size
Conv1	7 × 7 64		112 × 112
Conv2_x	1 × 1, 3 × 3, 1 × 1 64, 64, 256	× 3	56 × 56
Conv3_x	1 × 1, 3 × 3, 1 × 1 128, 128, 512	× 4	28 × 28
Conv4_x	1 × 1, 3 × 3, 1 × 1 256, 256, 1024	× 23	14 × 14
Conv5_x	1 × 1, 3 × 3, 1 × 1 512, 512, 2048	× 3	7 × 7
Pool5	7 × 7 2048		8 × 8
FCn	– n		–

3.2. Beneficial practices for network training

Since the performance of deeper neural networks need to be guaranteed by training on sufficient labeled data, existing datasets are small to train deeper two-stream ConvNets due to the overfitting problem. We propose several beneficial practices to make the training of deeper two-stream ConvNets stable and reduce the effect of overfitting.

ResNet Weights Transferring: current benchmark datasets for action recognition are mainly from daily life, the classes of which can be roughly grouped into four types as shown in Fig. 3: (1) **body motion only**, actions fully described by human movement like "Baby crawling"; (2) **human object interaction**, actions involving specific objects such as "Playing violin"; (3) **body motion in context**, body movement taking place in the specific environment like "Surfing"; (4) **human object interaction in context**, actions


Baby Crawling

(a) Body Motion Only

Playing Violin

(b) Human Object Interaction

Surfing

(c) Body Motion in Context

Bowling

(d) Human Object Interaction in Context

Fig. 3. Action classes of UCF101 dataset can be grouped into several types. The recognition of those actions is contributed by high-level visual cues, like human object interaction, scene context and pose variations.

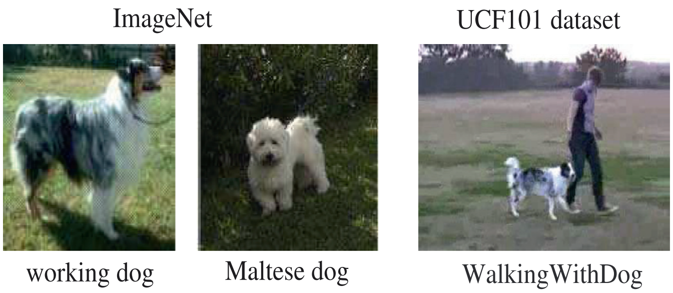


Fig. 4. Illustration of common statistics between the ImageNet and UCF101.

containing representative objects and occurring in certain context, such as “Bowling”. Since any given human action types need to be identified by the high-level visual cues like human object interaction, scene context and pose variations ([40]), and models trained on ImageNet dataset can be regarded as some mid-level understandings of object categories. Based on investigation, we found there exist ‘common statistics’ between the ImageNet dataset and UCF101 dataset, e.g. “Walking With Dog” in UCF101 dataset involves a “dog” class, while the ImageNet also contains many samples of dog such as “Maltese dog”, “working dog” and etc as observed in Fig. 4. This intrinsic connection inspired us to carry out a weights transferring from ResNet-101 layers trained on ImageNet to deeper two-stream ConvNets for model initialization.

During this process, the parameters of layers C1,⋯, C5x were initially trained on the ImageNet and transferred to our deeper spatial net directly in the next. However, the input of deeper temporal net were volumes of stacking optical flow fields (different from RGB images), which led to the channel number of first layer C1 in temporal net not be the same as ResNet-101 (20 vs. 3). For adjustments, we averaged the ResNet-101s filters of first layer C1 across its channels to initialize each channel of C1 layer filters in temporal net, and the parameters of the rest layers in ResNets 101 were then transferred to our temporal net for training continuation. It has proved that this strategy of weights transferring can effectively overcome the overfitting caused by sample deficiency.

Data Variation Augmentation Unlike image, video is a 3 dimensional data and has variable temporal duration. Thus, to utilize the ConvNets for video action recognition based on images, pre-processing is usually needed. In original two-stream ConvNets [27], videos were decomposed into frames according to time interval, and the motion information was modeled by extracting the optical flow fields between those frames. However, the data redundancy between consecutive frames would cause the deficiency of discriminative capacity for action recognition. Not as [21] only cropping the salient regions of the image center, in the training of proposed work, we introduced a scheme of data variation augmentation to increase the data diversity. With the fixed frame size of 256×340, each frame was cropped 4 corners and 1 center by randomly selecting from {256, 224, 192, 168} as the width and height, which was designed to take advantage of multi-scale representations. After the cropped regions being resized to 224×224 and flipped horizontally, there are 10 inputs (4 corners, 1 center, and their horizontal flipping) for the proposed model training. Such an augmentation scheme substantially increase the variations of inputs, which also help to get rid of the problem of overfitting.

3.3. Network testing

For the comparison with other state-of-the-art recognition approaches, the measurements in [27,38] had been employed in the proposed work. We sampled 25 frames or optical flow fields with equal temporal spacing in a given video, then followed by the data variation augmentation on each selected frame or optical flow field. As a sequence, there was 10 inputs for each frame in training deeper two-stream ConvNets and the video class score in separate nets was obtained by averaging all the scores of the sampled frames and their crops. By late fusion, the prediction scores of each stream nets were combined by using a weighted linear strategy. In testing process, it was found that randomly disordering the video sets listed in training/testing splits can achieve significantly improvement of recognition accuracy, even though the original video sets listed in training/testing splits were listed orderly according the categories. For each frame, we obtained 10 inputs for training deeper two-stream ConvNets. The class scores for the whole video in separate nets was obtained by averaging the all the scores of the sampled frames and their crops. Finally, prediction scores of each stream nets were combined by a weighted linear combination.

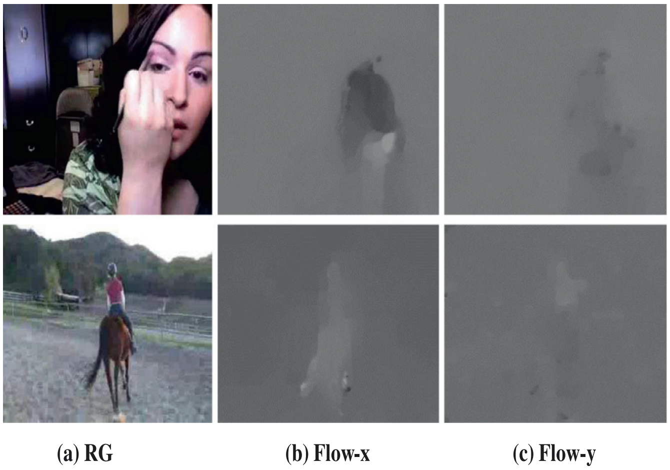
3.4. Implementation considerations

To implement the proposed recognition, some important issues need to be taken into account, and we discussed them as below.

Network Input: in deeper two-stream ConvNets, the spatial net utilized RGB images as the input. For temporal net, multi-frame (10 in our experiment) was taken as a unit to capture the motion information by using TVL1 optical flow [44] between each consecutive frame pair. To fed into temporal net, we discretized the values of optical flow fields into integers and set their range as the same as RGB images ([0,255]) with a linear transformation. Some frame samples and the corresponding optical flow fields were shown in Fig. 5. It can be found that a remarkable amount of horizontal movement or vertical movement were highlighted in the background. And then, the data variation augmentation was performed to generate more training samples.

Dis-ordering Strategy: to evaluate the dis-ordering strategy, we tested the spatial nets and temporal nets using three original splits and our dis-order splits respectively. As the results shown in Fig. 6, the proposed dis-ordering strategy can effectively improve the performance of both spatial net and temporal net. In the following experiments, the dis-ordered splits is put to use for overall evaluation.

Weighted Fusion: the fusion of spatial and temporal stream networks was carried out by a weighted average. With the investi-



(a) RG (b) Flow-x (c) Flow-y

Fig. 5. Examples of video frames and their corresponding optical flow fields.

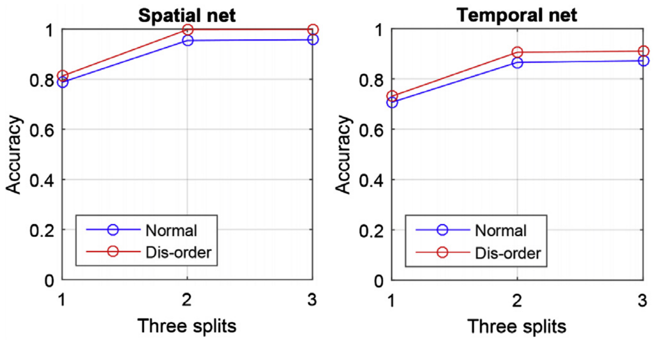


Fig. 6. Exploration of effectiveness of a dis-order strategy on UCF101 dataset. Left: We performed some experiments on spatial net using dis-ordering strategy or not. Right: Dis-ordering strategy or original splits were performed on temporal net. "Normal" meant we used original three training/testing video splits when testing. "Dis-order" meant that we evaluated our deeper two-stream ConvNets with a dis-order strategy of three training/testing video splits.

Table 2
Exploration of weighted fusion on UCF101 dataset. S denoted the spatial net. T denoted the temporal net. 2 × S+1 × T denoted setting weight of spatial net as 2 and that of temporal net as 1.

Weights	Accuracy
2 × S+1 × T	94.5%
1.5 × S+1 × T	94.8%
1.2 × S+1 × T	95.1%

gation of the dis-ordering strategy, we found the performance gap between spatial net and temporal net was much smaller than that was in the original two-stream ConvNets. In the proposed work, the accuracy of spatial net was higher than that of temporal net, for this reason, we designed a pairwise proportional scheme of weights to achieve an optimized combination of two streams. From the results shown in Table 2, three configurations was found to obtain a satisfactory fusion performance.

Network Training: Training the proposed deeper two-stream ConvNets is based on Caffe toolbox [11]. The mini-batch size of stochastic gradient descent (SGD) operation is set to 256, and the momentum is set to 0.9. For model initialization as discussed in the Section 3.2, the parameters of ResNet-101 layers trained on the large-scale ImageNet dataset is transferred, and fine tuned using a smaller learning rate. For both spatial net and temporal net, the learning rate is initialized to be 0.0001 and is decreased to its 0.0005 every 30, 000 iterations. The maximum iteration is set as 110, 000.

Table 3
Exploration of the effect of data augmentation on the performance for Deeper two-stream ConvNets on the UCF101 dataset (split 1).

Training setting	spatial ConvNets	temporal ConvNets
Baseline [27]	77.95%	67.91%
Data variation augmentation	81.26%	73.09%

4. Experiments

We conducted our experiments on UCF101 dataset [29] and KTH dataset [26]. The UCF101 dataset contains 13, 320 video clips totally which have been divided into 101 action classes and there are at least 100 video clips for each class. By following the procedure of THUMOS13 challenge [17], adopt the proposed dis-ordering strategy of three training/testing splits for evaluation. We used the dis-ordered UCF101 dataset for training deeper two-steam ConvNets and three dis-ordered testing splits to evaluation the performance of our proposed model. Finally, we reported the average accuracy across three splits.

KTH dataset [26] contains six types of human actions, which is performed by 25 actors in four different scenarios- outdoors, outdoors with scale variation, outdoors with different clothes and indoors. There are a total of 600 video sequences in the dataset. Following the evaluation in [26], the whole dataset were divided into a training set from 8 subjects (subjects 11, 12, 13, 14, 15, 16, 17 & 18), a validation set from 8 subjects (subjects 19, 20, 21, 23, 24, 25, 1 & 4) and a test set from 9 subjects (subjects 22, 2, 3, 5, 6, 7, 8, 9 & 10). Finally, we reported accuracy of test set.

4.1. Exploration study

In this section, we focus on the investigation the effect of data augmentation on the performance for network training described in Section 3.2. Specifically, we use proposed the deeper two-stream ConvNets and perform all experiments adopt different training settings on the split1 of UCF101 dataset.

We compare two training settings: (1) baseline setting in original two-stream ConvNets [27], where a fixed-size crop is randomly cropped and flipped from the whole frame. (2) data augmentation, which is described in Section 3.2. The results are shown in Table 3. We see that the performance of data variation augmentation scheme is much better than that of baseline setting (81.26% vs 77.95% in spatial ConvNets, 73.09% vs 67.91% in temporal ConvNets), which proves the good effect of carefully designed data augmentation for network training.

4.2. Recognition evaluation

The overall recognition evaluation was performed on the UCF101 dataset and KTH dataset with the proposed beneficial practices. Table 4 illustrates the accuracy difference of spatial net or temporal net on UCF101 dataset. We carried out recognition evaluation of the proposed deeper two-stream ConvNets on UCF101 dataset with different hand-designed features including HOG, HOF, MBH and etc. Table 5 shows that the convolutional descriptors of deeper spatial net or temporal net are much better than those hand-designed descriptors, which also indicates that the deep-learned features have stronger discriminative capability in recognition task.

Table 5 demonstrates the advancement of deeper structure compared with the original two-stream ConvNets [27] and TDDs [37], the deeper spatial net outperforms other spatial nets by a

Table 4
Performance of our deeper two-stream ConvNets on the UCF101. We conducted experiments over three dis-order splits on the UCF101 dataset. Deeper two stream results were combined by fusing class scores of spatial and temporal net with the strategy of setting weight of temporal nets as 1 and that of spatial net as 1.2. Then we averaged scores of spatial net over three splits. And the same average operating have been done for temporal net and deeper two stream.

Split	Split1	Split2	Split3	Avg
Spatial net	81.26%	99.81%	99.84%	93.64%
Temporal net	73.09%	90.60%	91.04%	84.91%
Deeper two stream	85.73%	99.76%	99.89%	95.13%

Table 5
We compared our proposed model with HOG descriptors [35], two-stream ConvNets [27] and TDD [37] on UCF101. Experiments verify the effectiveness of the our spatial net, temporal net or their combined model.

Algorithm	UCF101
HOG [35,36]	72.4%
HOF [35,36]	76.0%
MBH [35,36]	80.8%
HOF+MBH [35,36]	82.2%
iDT [35,36]	84.7%
Spatial net [27]	73.0%
Temporal net [27]	83.7%
Two-stream ConvNets [27]	88.0%
Spatial net conv4 and conv5 [37]	82.8%
Temporal net conv3 and conv4 [37]	82.2%
TDD [37]	90.3%
Deeper spatial net **(ours)**	93.6%
Deeper temporal net **(ours)**	84.9%
Deeper two-stream ConvNets **(ours)**	95.1%

large margin (93.6% vs 73.0% or 82.8%). Also for the temporal nets, the performance are better than others (84.9% vs 82.2% or 83.7%). Moreover, it is noticed that the deeper two-stream ConvNets with fusion scheme still outperform original two-stream ConvNets by around 7%. Those results, on the one hand, verify that the deeper two-stream structure is able to collaboratively achieve a significant improvement for action recognition by incorporating proposed beneficial practices. On the other hand, they further indicates that the weak performance of most current deep convolutional networks is caused by lacking of high modeling capacity constrained by depth of their models.

One thing that needs to be specified, the reason why temporal net with deeper ConvNets could not yield a same good performance as the deeper spatial nets (84.9% vs 93.6%) can be explained as follows. As discussed in Section 3.2, we used parameters of ResNet-101 layers trained on the ImageNet to initialize both deeper spatial and deeper temporal nets. However, the inputs of deeper temporal nets are the volumes of stacking optical flow fields, which are essentially different from the image inputs of ImageNet. Such a divergence caused the reused layers from ImageNet models could not work properly as a generic extractor of mid-level representations for temporal net. Besides, the optical flow fields extracted from videos are relatively insufficient for learning so many parameters as the initialization of temporal net.

At the end, we evaluated deeper temporal ConvNets on KTH dataset. The results was illustrated in Table 7. Combination is carried out by setting weight of deeper temporal ConvNets as 2 and that of deeper RGB ConvNets as 0.1. The best accuracy of 93.1% was achieved on KTH dataset. Besides, it was found that the deeper temporal net achieved much better performance than deeper spatial net (92.13% vs 61.11%). We concluded that this is due to two reasons. First, the KTH dataset are suffered from more poor image quality than UCF101, which was caused by darkness of light conditions and low resolution. Second, compared with UCF101 dataset, the actions of KTH datset are all performed by single person. As a consequence, the deeper spatial net extracted little high-level visual cues such as human object interaction, scene context etc. for action recognition.

Table 6
Comparison of deeper two-stream ConvNets to the state of the art on UCF101.

Method	Year	Accuracy
Two stream [27]	2014	88.0%
Two-stream+LSTM [43]	2015	88.6%
TDD+iDT [37]	2015	91.5%
Deep two-stream [38]	2015	91.4%
Two-stream SR-CNNS [40]	2016	92.6%
LTC [33]	2016	91.7%
KVMF [47]	2016	93.1%
TSN [39]	2016	94.2%
Ours	2016	**95.1%**

Table 7
Performance of our deeper two-stream ConvNets on the KTH dataset.

Method	Accuracy
Deeper spatial ConvNets	61.11%
Deeper temporal ConvNets	92.13%
Deeper two stream	93.1%
local SVM [26]	71.72%
convGRBM [31]	90.0%
3D ConvNets [16]	90.2%

4.3. Performance comparison

The comparison results between the proposed work against the other state-of-the-art on UCF101 dataset are summarized in Table 6. To be more specifically, in our experiments, we only compare with the deep learning based approaches. For the two-stream Convolutional networks [27] and its variants such as two-stream ConvNets with recurrent neural networks (two-stream+ LSTM) [43], Deep two-stream [38], TDD+iDT [37], and two-stream semantic region based CNNs (two-stream SR-CNNS) [40], the proposed work is better than all of these two-stream ConvNets based approaches. For other deep networks such as long term convolution networks (LTC) [33], key volume mining framework (KVMF) [47], and temporal segment networks (TSN) [39], the proposed deeper two-stream ConvNets is also the best among all. The superior performance on UCF101 dataset demonstrates the effectiveness of the deeper two-stream ConvNets and the beneficial practices using for training.

The comparison results between the proposed work against the other state-of-the-art on KTH dataset are summarized in Table 7. In Table 7, we compare the performance of our deeper two-stream ConvNets with hand-designed features like local features [14,15,26] and deep-learned models such as convGRBM [31] or 3D ConvNets [16]. It was shown that our proposed method had stronger discriminative capability in recognition task (93.1% vs 90.2% vs 90.0% vs 71.72%).

4.4. Visualization

To attain an insight into the learned deeper two-stream ConvNets, some video frames in UCF101 dataset belonging to different action classes, such as “Basketball”, “PommelHorse”, “ApplyEyeMakeup” and “PlayingGuitar”, are selected to feed into the deeper spatial net respectively. The feature maps of the last convolutional

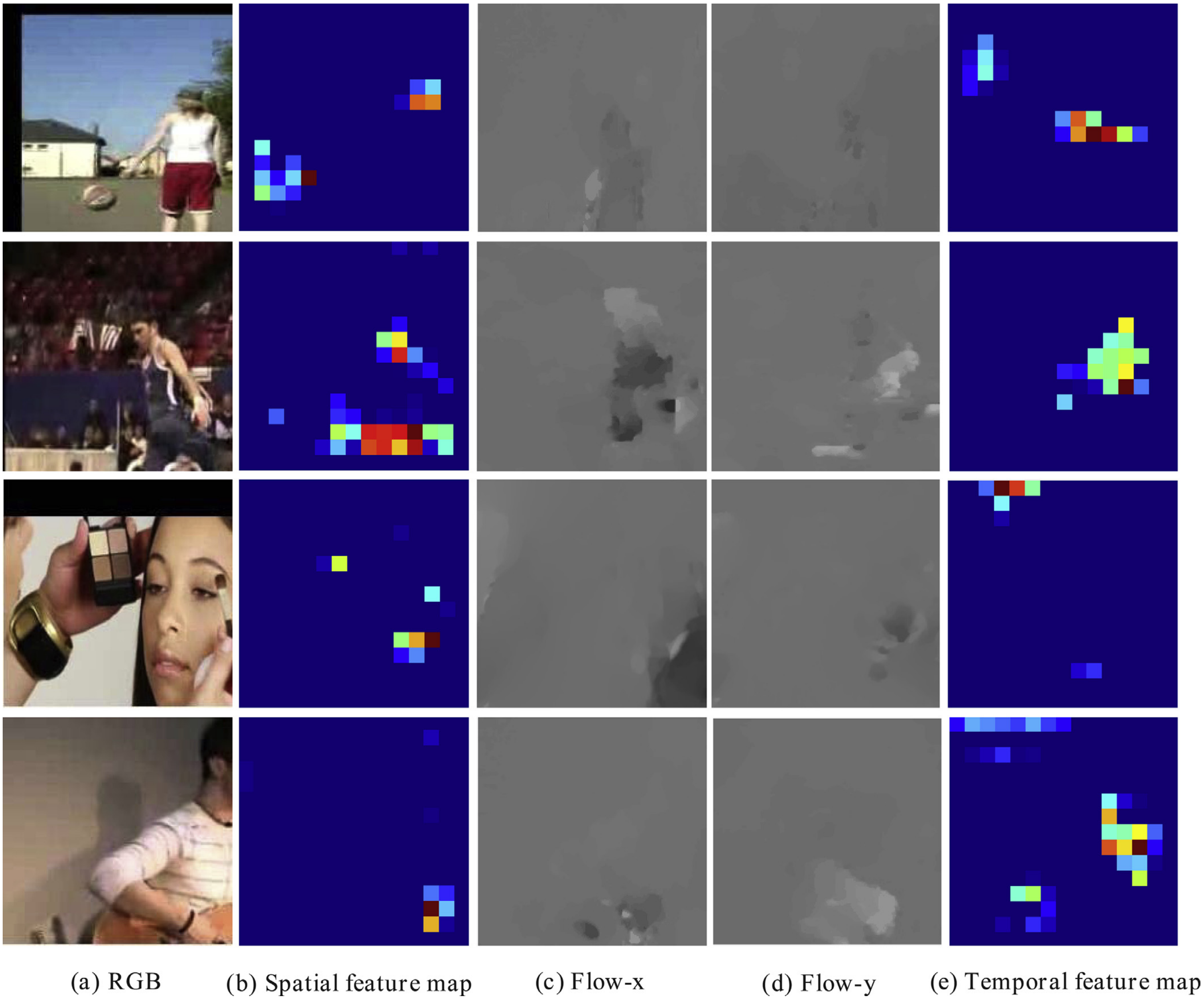


Fig. 7. Visualization of learnt convolutional filters. We select some video frames, optical flow fields, and visualize their corresponding feature maps of deeper spatial net and deeper temporal net.

layer are then visualized followed by pooling layer, whose outputs are fed into a fully-connected layer to obtain the final class scores. For the temporal net, the feature maps of same layer as spatial net are visualized using stacked optical flow fields corresponding to frame images of spatial net. The results of visualization are shown in Fig. 7. From the visualized samples, it can be seen that the feature maps of last convolutional are relatively sparse and exhibit a high correlation with the action areas. It indicates that the proposed deeper two-stream model has higher modeling capability to generate more discriminative features for action recognition in videos.

5. Conclusions and future work

In this paper, we proposed a novel deeper structure of two-stream ConvNets by learning high-level representations for action recognition in videos. To guarantee the learning performance, some beneficial practices are also introduced to overcome the overfitting challenge of sample deficiency. In testing phase, a significant improvement of action recognition has been achieved with a disordering strategy among video sets listed in training/testing splits. The empirical experiments have shown that the proposed deeper two-stream ConvNets can outperform the other state-of-the-art at recognition accuracy of 95.1% on the UCF101 dataset and 93.1% on the KTH dataset.

In experimental evaluation, we found that the temporal net with deeper ConvNets did not yield good performance as the spatial nets on UCF101 dataset. One promising way for this limitation is to capture the motion information with deeper temporal structure, which motivated us to adopt the deeper recurrent neural networks to model long term motion dynamics in our future studies.

Acknowledgments

This research is supported by National Natural Science Foundation of China 61571362 & 61601505 & 61403182 & 61363046, and by the National High-Tech Research and Development Program of China (863 Program) with No. 2015AA016402, as well as the key youth grant S2017QNZDB0041 approved by Natural Science Foundation of Jiangxi Province and the National Research Foundation, Prime Ministers Office, Singapore under its International Research Centre in Singapore Funding Initiative.

References

[1] J.K. Aggarwal, M.S. Ryoo, Human activity analysis: a review, ACM Comput. Surv. (CSUR) 43 (3) (2011) 16.

[2] Y. Bengio, P.Y. Simard, P. Frasconi, Learning long-term dependencies with gradient descent is difficult, IEEE Trans. Neural Netw. 5 (2) (1994) 157–166.

[3] C. Ding, D. Tao, Multi-task pose-invariant face recognition, IEEE Trans. Image Process. (T-IP) (2015).

[4] C. Ding, J. Choi, D. Tao, L. Davis, Multi-directional multi-level dual-cross patterns for robust face recognition, IEEE Trans. Pattern Anal. Mach. Intell. (T–PAMI) (2016).

[5] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: a large-scale hierarchical image database, in: Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on, IEEE, 2009, pp. 248–255.

[6] C. Ding, D. Tao, Robust face recognition via multimodal deep face representation for multimedia applications, IEEE Trans. Multimed. (T-MM) (2015).

[7] C. Ding, D. Tao, Pose-invariant face recognition with homography-based normalization, Pattern Recognit. (2017).

[8] R. Girshick, J. Donahue, T. Darrell, J. Malik, Rich feature hierarchies for accurate object detection and semantic segmentation, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2014, pp. 580–587.

[9] G. Gkioxari, R. Girshick, J. Malik, Contextual action recognition with r* cnn, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 1080–1088.

[10] X. Glorot, Y. Bengio, Understanding the difficulty of training deep feedforward neural networks., in: Aistats, 9, 2010, pp. 249–256.

[11] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, arXiv preprint arXiv:1512.03385 (2015).

[12] K. He, X. Zhang, S. Ren, J. Sun, Delving deep into rectifiers: Surpassing human-level performance on imagenet classification, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 1026–1034.

[13] S. Ioffe, C. Szegedy, Batch normalization: accelerating deep network training by reducing internal covariate shift, arXiv preprint arXiv:1502.03167(2015).

[14] J. Yu, Y. Rui, B. Chen, Exploiting click constraints and multiview features for image reranking, IEEE Trans Multimed. 16 (1) (2014) 159–168.

[15] J. Yu, Y. Rui, D. Tao, Click prediction for web image reranking using multimodal sparse coding, IEEE Trans. Image Process. 23 (5) (2014) 2019–2032.

[16] S. Ji, W. Xu, M. Yang, K. Yu, 3D convolutional neural networks for human action recognition, IEEE Trans. Pattern Anal. Mach. Intell. 35 (1) (2013) 221–231.

[17] Y. Jiang, J. Liu, A.R. Zamir, G. Toderici, I. Laptev, M. Shah, R. Sukthankar, Thumos challenge: action recognition with a large number of classes, ECCV Workshop, 2014.

[18] J. Yu, Z. Kuang, B. Zhang, D. Lin, J. Fan, Image privacy protection by identifying sensitive objects via deep multi-task learning, IEEE Trans. Inf. Forensics Secur. 12 (5) (2017) 1005–1016.

[19] J. Yu, X. Yang, F. Gao, D. Tao, Deep multimodal distance metric learning using click constraints for image ranking, IEEE Trans. Cybern. (2016). doi: 10.1109/TCYB.2016.2591583.

[20] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, L. Fei-Fei, Large-scale video classification with convolutional neural networks, in: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, 2014, pp. 1725–1732.

[21] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, in: Advances in neural information processing systems, 2012, pp. 1097–1105.

[22] H. Kuehne, H. Jhuang, R. Stiefelhagen, T. Serre, Hmdb51: a large video database for human motion recognition, in: High Performance Computing in Science and Engineering 12, Springer, 2013, pp. 571–582.

[23] Q.V. Le, W.Y. Zou, S.Y. Yeung, A.Y. Ng, Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis, in: Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on, IEEE, 2011, pp. 3361–3368.

[24] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, Gradient-based learning applied to document recognition, Proc. IEEE 86 (11) (1998) 2278–2324.

[25] S. Ren, K. He, R. Girshick, J. Sun, Faster r-cnn: towards real-time object detection with region proposal networks, in: Advances in neural information processing systems, 2015, pp. 91–99.

[26] C. Schuldt, I. Laptev, B. Caputo, Recognizing human actions: a local svm approach, in: Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on, 3, IEEE, 2004, pp. 32–36.

[27] K. Simonyan, A. Zisserman, Two-stream convolutional networks for action recognition in videos, in: Advances in Neural Information Processing Systems, 2014, pp. 568–576.

[28] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, arXiv preprintarXiv:1409.1556 (2014).

[29] K. Soomro, A.R. Zamir, M. Shah, Ucf101: a dataset of 101 human actions classes from videos in the wild, arXiv preprintarXiv:1212.0402(2012).

[30] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 1–9.

[31] G.W. Taylor, R. Fergus, Y. LeCun, C. Bregler, Convolutional learning of spatio-temporal features, in: European conference on computer vision, Springer, 2010, pp. 140–153.

[32] D. Tran, L. Bourdev, R. Fergus, L. Torresani, M. Paluri, Learning spatiotemporal features with 3d convolutional networks, arXiv preprint arXiv:1412.0767(2014).

[33] G. Varol, I. Laptev, C. Schmid, Long-term temporal convolutions for action recognition, arXiv preprint arXiv:1604.04494(2016).

[34] W. Liu, Z. Zha, Y. Wang, K. Lu, D. Tao, p-laplacian regularized sparse coding for human activity recognition, IEEE Trans. Indust. Electr. 63 (8) (2016) 5120–5129.

[35] H. Wang, C. Schmid, Action recognition with improved trajectories, in: Proceedings of the IEEE International Conference on Computer Vision, 2013, pp. 3551–3558.

[36] H. Wang, C. Schmid, Lear-inria submission for the thumos workshop, in: ICCV workshop on action recognition with a large number of classes, 2, 2013, p. 8.

[37] L. Wang, Y. Qiao, X. Tang, Action recognition with trajectory-pooled deep-convolutional descriptors, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 4305–4314.

[38] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, Towards good practices for very deep two-stream convnets, arXiv preprint (2015). arXiv:1507.02159.

[39] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, L. Van Gool, Temporal segment networks: towards good practices for deep action recognition, in: European Conference on Computer Vision, Springer, 2016, pp. 20–36.

[40] Y. Wang, J. Song, L. Wang, L. Van Gool, O. Hilliges, Two-stream sr-cnns for action recognition in videos, BMVC, 2016.

[41] W. Liu, H. Liu, D. Tao, Y. Wang, K. Lu, Multiview hessian regularized logistic regression for action recognition, Signal Process. 110 (2015) 101–107.

[42] W. Liu, H. Zhang, D. Tao, Y. Wang, K. Lu, Large-scale paralleled sparse principal component analysis, Multimed. Tools Appl. 75 (3) (2016) 1481–1493.

[43] J. Yue-Hei Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, G. Toderici, Beyond short snippets: Deep networks for video classification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 4694–4702.

[44] C. Zach, T. Pock, H. Bischof, A duality based approach for realtime tv-l 1 optical flow, in: Joint Pattern Recognition Symposium, Springer, 2007, pp. 214–223.

[45] M.D. Zeiler, R. Fergus, Visualizing and understanding convolutional networks, in: European Conference on Computer Vision, Springer, 2014, pp. 818–833.

[46] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, A. Oliva, Learning deep features for scene recognition using places database, in: Advances in neural information processing systems, 2014, pp. 487–495.

[47] W. Zhu, J. Hu, G. Sun, X. Cao, Y. Qiao, A key volume mining deep framework for action recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 1991–1999.