Video Action Recognition based on Deeper Convolution Networks with Pair-Wise Frame Motion Concatenation

Yamin Han[1], Peng Zhang[1], Tao Zhuo[2], Wei Huang[3] and Yanning Zhang[1]

[1]School of Computer Science, Northwestern Polytechnical University, China
[2]Sensor-enhanced Social Media (SeSaMe) Centre, National University of Singapore, Singapore
[3]School of Information Engineering, Nanchang University, China

Abstract

Deep convolution networks based strategies have shown a remarkable performance in different recognition tasks. Unfortunately, in a variety of realistic scenarios, accurate and robust recognition is hard especially for the videos. Different challenges such as cluttered backgrounds or viewpoint change etc. may generate the problem like large intrinsic and extrinsic class variations. In addition, the problem of data deficiency could also make the designed model degrade during learning and update. Therefore, an effective way by incorporating the frame-wise motion into the learning model on-the-fly has become more and more attractive in contemporary video analysis studies.

To overcome those limitations, in this work, we proposed a deeper convolution networks based approach with pair-wise motion concatenation, which is named deep temporal convolutional networks. In this work, a temporal motion accumulation mechanism has been introduced as an effective data entry for the learning of convolution networks. Specifically, to handle the possible data deficiency, beneficial practices of transferring ResNet-101 weights and data variation augmentation are also utilized for the purpose of robust recognition. Experiments on challenging dataset UCF101 and ODAR dataset have verified a preferable performance when compared with other state-of-art works.

1. Introduction

Human action recognition in videos has drawn increasing attention from the research community [1, 5, 11, 13, 16, 20, 24], owing to its potential applications in many areas such as video surveillance, behavior analysis, and video content analysis. However, the problem of action recognition remains challenging when handling the realistic datasets (e.g. HMDB51 [15] and UCF101 [22]). The main difficulties are caused by large intra-class variations in the same action class, which may be caused by background clutter, scale, viewpoint change, and fast irregular motion. Meanwhile, the video inherent attributes, e.g. high dimension and low resolution, could further increase the difficulties of robust recognition. Thus, designing more abstractive representations by imitating human understanding on videos to deal with these challenges is of crucial importance.

In last decade, Convolutional Networks (ConvNets) [17] had led to a series of breakthroughs in image classification [14], object detection [4], scene recognition [34] and motion blur [7]. The typical ConvNets based algorithms including 3D ConvNets [11], Deep ConvNets [28], and two-stream ConvNets [20] had also been applied to the task of video-based action recognition. These methods utilized ConvNets trained on large-scale labeled datasets to automatically learn video representation from raw data and two-stream ConvNets was the most competitive architecture at present. However, unlike being applied on image classification tasks [14], deep ConvNets failed to achieve a great improvement over traditional methods [26]. In our views, this phenomenon may be caused by two major obstacles: firstly, deep ConvNets based methods required a large quantity of labeled samples for training. However, compared to the ImageNet dataset [3], the publicly available action recognition datasets are relatively small, which would lead to high risk of overfitting when applying deeper ConvNets for training as on image classification. Secondly, the number of stacked layers in most current deep ConvNets is relatively small [20], e.g. the architecture of two-stream ConvNets only contain 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 evidence [23, 21] revealed that the depth of networks is of crucial importance, which means that there

might be a great performance enhancement benefiting from its high modeling capacity by exploiting very deep models [18, 4].

Motivated by the above analysis. We proposed a novel deeper temporal ConvNets for action recognition, which adopts Residual Networks 101(ResNet-101) [8] as backbone. We removed layers after the pool5 layer of ResNet-101 and added an adaptation fully-connection layer, the output number of which are related with the numbers of action classes of dataset. The proposed deeper model had a high modeling capacity which was able to obtain effective representations to deal with challenges of complex actions. For the overfitting problem, the parameters of ResNet-101 layers trained on the ImageNet was transferred to initialize the proposed model and an augmented data variation to increase the data diversity. Besides, It can further enhance recognition accuracy by disordering the video sets listed in training/testing splits randomly. To verify the effectiveness of our method, A number of experiments were performed on UCF101 dataset [22] and ODAR2017 dataset and achieved a preferable performance.

The main contributions of this paper are summarized as below. Firstly, we designed a deeper temporal ConvNets for action recognition by adapting recent very deep architectures-Residual Networks into video domain, which achieved preferable performances on UCF101 dataset [22] and ODAR2017 dataset. 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 propose an augmented data variation strategy. Finally, a dis-ordered strategy which has been utilized for training/testing splits was proposed to further enhance the final performance of evaluation.

2. Related Work

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

Convolutional Networks for Image Classification. Deep learning techniques had shown its outstanding effectiveness in a variety of image based tasks [14, 33, 34, 5, 21, 23, 9, 8]. Over the past few years, researchers had proposed many well-known network structures for image based tasks. For ImageNet classification, Krizhevsky *et al*. [14] proposed an AlexNet model that contained eight learned layers. Zeiler *et al*. [33] introduced a novel visualization technique and found an eight layer ConvNet model named ClarifaiNet that outperformed AlexNet. Simonyan *et al*. [21] 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). In [23], Szegedy proposed GoogLeNet, a 22 layers deep network. Besides, He *et al*. proposed the deepest networks RestNets up to more than 1000 layers at present [8]. Analyzing the evolution of from AlexNet to RestNets, It was obviously found that the depth of ConvNets was deeper and deeper. These evidence revealed that on the challenging imageNet dataset, methods that exploited deeper models leading greatly performance due to its high modeling capacity.

Deep Learning based Action Recognition. Since great advances had been achieved in image recognition task based on Convolutional networks, which largely promoted the development of ConvNets based action recognition in video [20, 11, 24, 13, 25, 16, 20, 29]. In order to deal with video data, Ji *et al*. [11] adapted 2D ConvNets to 3D for video-based action recognition on relatively small datasets. Taylor *et al*. [24] used a 3D convolutional RBMS to learn spatio-temporal features in unsupervised. Karpathy *et al*. [13] 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 well, they achieved lower performance in comparison with shallow hand-crafted representation [26]. Recently, to explicitly model the motion pattern, Simonyan *et al*. [20] designed two-stream ConvNets composed of spatial and temporal net. The spatial net mainly captured appearance features utilizing video frames as inputs. Meanwhile, temporal net learnt effective motion features by using optical flow fields between several consecutive frames, and finally it matched the state-of-the-art performance. Besides, Wang *et al*. [29] proposed two-stream semantic region based CNNs with multiple semantic channels.

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

3. The Proposed Work

In this section, we introduced our proposed approach in detail. First of all, an overview of the new deeper temporal ConvNets will be introduced. After that, the details about getting rid of overfitting effect during training will be discussed. At last, we discuss detailed implementation considerations including network input configurations, dis-order scheme and network training or testing.

3.1. Deeper Temporal Convolutional Networks

Motion information encoded in multiple contiguous frames is of critical for action recognition. Indeed, motion information contains dynamics of human which can be used to recognize what the individuals are doing. Motion

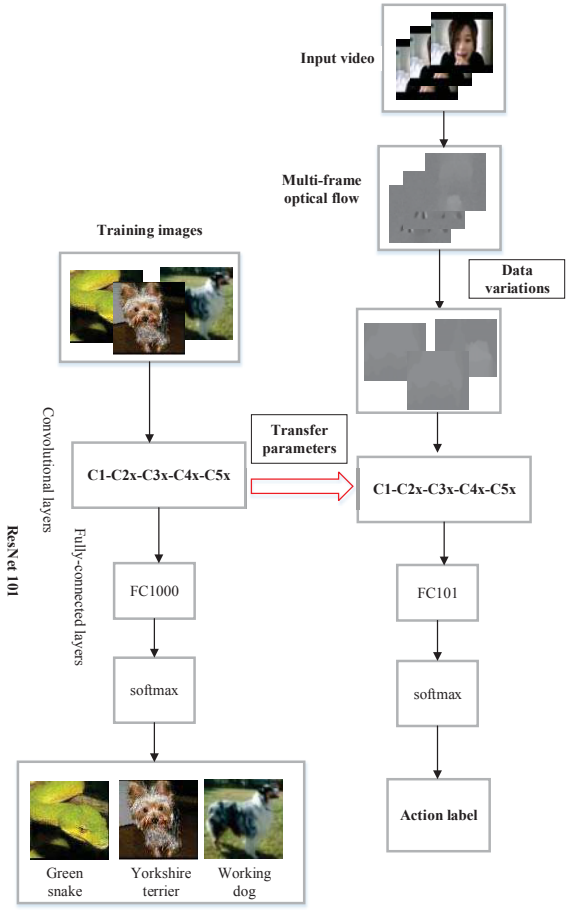

Figure 1. Pipeline of our approach. We incorporate the original ResNet101 into the framework of temporal ConvNets for action recognition in videos, and propose a new architecture, called deeper temporal ConvNets.

information based methods had shown its outstanding performance in the task of video action recognition [20]. The temporal ConvNetss is trained to recognize action classes from motion information in the form of dense optical flow. But as most deep models, training a deeper temporal ConvNets is not easy due to the problem of vanishing or exploding gradients [6]. 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 renders the gradient information ineffectively small in earlier layers. Several approaches tried to reduce such an effect practically through careful initialization [6], Batch Normalization [10], but the performance enhancement was still limited.

Instead of hoping each few stacked layers directly fit a desired underlying mapping, Residual networks (ResNets) [8] makes use of identity shortcut connections that enable flow of information across layers without attenuation. ResNets had achieved state-of-the-art performance in ImageNet classification task and allowed training of extremely deep networks up to more than hundreds of layers, which is just needed by the designing of a deeper temporal ConvNets. Formally, a building block in ResNets defined as: $y = G(x, \{W_i\}) + x$. Here, x and y are the input and output vectors of the layers considered. W_i represents the weights of i_{th} layer. The function $G(x, \{W_i\})$ represents the residual mapping to be learned, e.g. in the top of Figure 2 that has two layers, $G = W_2 \times ReLU(W_1 x)$. The operation $G + x$ is performed by a shortcut connection and element-wise addition. The element-wise additions performed on two feature maps, channel by channel. The identity mapping in the top of Figure 2 introduces neither extra parameters nor computation complexity, this is attractive in practice. The dimensions of x and G must be equal. If this is not the case, a linear projection W_s is performed to match the dimension: $y = G(x, \{W_i\}) + W_s x$. the W_s is only used when matching dimensions. The form of the residual function G is flexible. We involve the function G that has three layers as shown in the bottom of Figure 2.

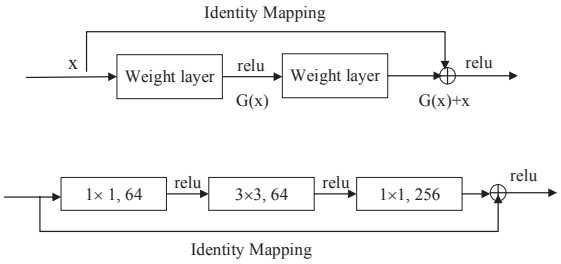

Figure 2. A deeper residual function F

The overview of the proposed deeper temporal ConvNets is shown in Figure 1. In this structure, we constructed a deeper temporal 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. Table 1 gives the details of the proposed ConvNets, where the input of deeper temporal ConvNets are volumes of stacking optical flow fields ($224 \times 224 \times 2F$, F is the number of stacking flows, here F=10). Some beneficial Practices such as Transferring ResNet-101 weights, Data Variation Augmentation are proposed for deeper temporal ConvNets training.

3.2. Beneficial Practices for Network Training

Deeper neural networks are more difficult to train on the condition of labeled data deficiency. Here we proposed

Layer	101-layer	Output size
Conv1	$\begin{bmatrix} 7\times 7 \\ 64 \end{bmatrix}$	112×112
Conv2_x	$\begin{bmatrix} 1\times 1, & 3\times 3, & 1\times 1 \\ 64, & 64, & 256 \end{bmatrix} \times 3$	56×56
Conv3_x	$\begin{bmatrix} 1\times 1, & 3\times 3, & 1\times 1 \\ 128, & 128, & 512 \end{bmatrix} \times 4$	28×28
Conv4_x	$\begin{bmatrix} 1\times 1, & 3\times 3, & 1\times 1 \\ 256, & 256, & 1024 \end{bmatrix} \times 23$	14×14
Conv5_x	$\begin{bmatrix} 1\times 1, & 3\times 3, & 1\times 1 \\ 512, & 512, & 2048 \end{bmatrix} \times 3$	7×7
Pool5	$\begin{bmatrix} 7\times 7 \\ 2048 \end{bmatrix}$	8×8
FC101	$\begin{bmatrix} - \\ 101 \end{bmatrix}$	-

Table 1. ConvNets Architectures. The symbol of '[]' represents building blocks as shown in ResNet-101 [8]. 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 ConvNet architectures are similar with ResNet-101. We remove layers after the pool5 layer of original ResNet-101 and add an adaptation fully-connection layer for the task of action recognition.

some beneficial practices for deeper temporal ConvNets training as follows.

Transferring ResNet-101 Weights. current benchmark datasets for action recognition are mainly derived from daily life, the classes of which can be roughly grouped into four types as shown in Figure 3: (1) **body motion only.** actions fully described by human movement like "Push up". (2) **human object interaction.** actions involving specific objects such as "Biking". (3) **body motion in context.** body movement taking place in the specific environment like "Skydiving". (4) **human object interaction in context.** actions containing representative objects and occurring in certain context, such as "Balance bean". Based on investigation, we found that any given human action types was identified by the high-level visual cues like human object interaction, scene context and pose variations [29]. Besides, as shown in Figure 4, there are "common statistics" between the ImageNet dataset and UCF101 dataset. e.g. "Walking With Dog" in UCF101 dataset involves the "dog" class, while the ImageNet also contains many samples of dog like "Maltese dog", "working dog". Models trained on ImageNet dataset can capture mid-level object parts information for action recognition. This intrinsic connection inspired us to transfer parameters of ResNet-101 layers trained on ImageNet to deeper two-stream ConvNets for model initialization to overcome the overfitting challenge of sample deficiency.

Figure 3. Action classes can be grouped into several types. The recognition of those actions is contributed by different semantics, like human object interaction, scene context and pose variations

Figure 4. Illustration of different dataset statistics between the ImageNet and UCF101.

During this process, due to the input of deeper temporal ConvNets are volumes of stacking optical flow fields that are different from RGB images, which led to the channel number of first layer C1 in temporal ConvNets are not same as that in ResNet-101 (20 vs. 3). For adjustments, we averaged the ResNet-101's filters of first layer C1 across its channel to initialize each channel of C1 layer filters in temporal ConvNets. Besides, the parameters of the rest layers C2x,$\cdots$,C5x in ResNets-101 were then transferred to our temporal ConvNets directly. The experiments proved that this strategy of weights transferring can effectively overcome the overfitting caused by sample deficiency.

Data Variation Augmentation. Unlike images, videos are 3D in nature and have variable temporal durations. In order to utilize the ConvNets for action recognition in videos, many methods process video data in advance. In two-stream ConvNets [20], videos are decomposed into many image frames according to the time interval and they extract optical flow fields between frames to model motion

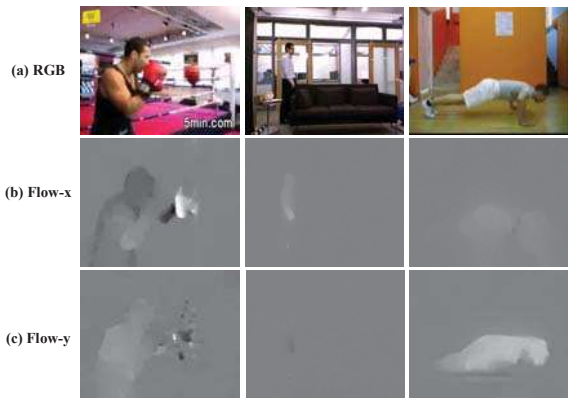

Figure 5. Examples of video frames and their corresponding optical flow fields.

information. However, the data redundancy between consecutive frames would cause the deficiency of discriminative capacity for action recognition. With the purpose of increasing the diversity of training data, we introduce an augmented data variation strategy. Firstly, with the fixed frame size of 256×340, each optical flow 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 advantages 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. Implementation Considerations

Network Input Configurations. Our deeper temporal ConvNets takes the multi-frame (10 in our experiment) optical flow fields which aim to capture the motion information as input. Here, the optical flow fields between two adjacent frames were extracted by adopting the TVL1 optical flow algorithm [32]. Then, in order to fed into temporal ConvNets, we discretize the values of optical flow fields into integers and set their range as [0, 255] by a linear transformation, which makes the range of optical flow fields to be just the same with RGB images. Some examples of video frames and their corresponding optical flow fields were shown in Figure 5. It can be seen that a remarkable amount of horizontal movement or vertical movement are highlighted in the background. Then, the data variation augmentation was performed to generate more training samples.

Dis-order Scheme. The original training/testing splits of UCF101 are listed orderly according the categories. In the process of training deeper model, we found that selecting labeled samples randomly to train deeper model each iteration can further improve the accuracy of training. In the same way, randomly disordering video sets listed in testing splits can also achieve significantly improvement of recognition accuracy when testing.

Network Training. We use the Caffe toolbox [8] for ConvNets implementation. The mini-batch size of stochastic gradient descent (SGD) operation is set to 256 and the momentum is set to 0.9. For UCF101 dataset, deeper temporal model is pretrained on ImageNet, except for the temporal ConvNets on ODAR dataset which was initialised from the deeper temporal ConvNets trained on UCF101 dataset, and fine tune the model parameters using a smaller learning rate. For both UCF101 dataset and ODAR2017 dataset, the learning rate of deeper temporal model 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.

Network Testing. For fair comparison with other state-of-the-art methods, we adopt the same testing strategies as shown in [20, 28]. We sample 25 optical flow frames with equal temporal spacing when given a video, then followed by the augmented data variation augmentation on each selected stacking optical flow fields. Therefore, there was 10 inputs for each selected stacking optical flow fields and the video class scores was obtained by averaging all the scores of the sampled stacking optical flow fields and their crops. In testing process, it was found that randomly disordering the video sets listed in testing splits can achieve significantly improvement of recognition accuracy, even though the original video sets listed in testing splits were listed orderly according the categories.

4. Experiments

In this section, we first introduce the datasets used for evaluation. Then, evaluations of deeper temporal ConvNets on two dataset are discussed in detail. In addition, we also do some additional experiments for deeper RGB ConvNets or combination of deeper RGB and temporal ConvNets. Finally, further insights into the learned deeper temporal ConvNets or RGB ConvNets are discussed.

4.1. Datasets

In order to evaluate our proposed method, we conduct experiments on UCF101 dataset [22] and ODAR 2017 dataset. UCF101 dataset is the largest available annotated video dataset, it 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. Following the evaluation scheme of The THUMOS13 challenge [12] and adopting the proposed dis-order scheme of three training/testing splits for evaluation, average accuracy across three splits are reported on UCF101 dataset.

Training setting	Deeper temporal ConvNets
Baseline [20]	73.24%
Data variation augmentation	76.19%

Table 2. Exploration of different training strategies for Deeper temporal ConvNets on the UCF101 dataset (split 2).

Split	Split1	Split2	Split3	Avg
Deeper temporal ConvNets	73.09%	76.19%	76.87%	75.38%
Deeper RGB ConvNets	81.26%	79.86%	81.25%	80.79%
Combined	85.73%	85.62%	85.77%	85.71%

Table 3. Performance of deeper temporal ConvNets, RGB ConvNets or combination of deeper RGB and temporal ConvNets on the UCF101.

Algorithm	UCF101
HOG [26, 27]	72.4%
iDT [26, 27]	84.7%
Spatial net [20]	73.0%
Temporal net [20]	83.7%
Two-stream ConvNets [20]	88.0%
Deeper temporal ConvNets	75.38%

Table 4. We compare our proposed model with HOG descriptors [26], two-stream ConvNets [20].

ODAR 2017 dataset consists of multiple publicly available datasets (e.g. IXMAS dataset [30], KTH [19], MSR [31] *etc*.) with carefully selected action classes that are common across these datasets. In this challenge, a protocol have been designed to ensure the training set, validation set, and test set contains action samples extracted from different domain. Overall, 12 common action classes and an additional class for all other action samples (i.e., "unknown" class) was selected.

4.2. Exploration Study

In this section, we focus on the investigation the data variation augmentation scheme for network training described in section 3.2. Specifically, we compare our training setting using data variation augmentation with Baseline setting in original two-stream ConvNets [20] where a fixed-size $224 \times 224 \times 2L(L = 10)$ is randomly cropped and flipped. The results are shown in Table 2. We see that the performance of training using data variation augmentation scheme is much better than that of Baseline setting, which implies carefully designed beneficial practices for network training is necessary to reduce the risk of overfitting.

4.3. Evaluation of Deeper Temporal ConvNets

The overall recognition evaluation of deeper temporal ConvNets was performed on the UCF101 dataset and ODAR2017 dataset with founded beneficial practices.

The results of deeper temporal ConvNets on the UCF101 dataset illustrated in Table 3. In Table 4, We first compare the performance of our deeper temporal ConvNets with hand-designed features like HOG descriptors, iDT descriptors. It shows that the convolutional descriptors of deeper temporal net are much better than HOG descriptors (75.38% vs 72.4%). However, there is a lower performance in comparison with iDT descriptors which was the most discriminative hand-designed features. To demonstrate the advancement of deeper structure compared with the original two-stream ConvNets [20], our deeper temporal ConvNets outperforms Spatial net [20] (75.38% vs 73.0%) but is weaker than Temporal net [20] and Two-stream ConvNets [20].

Method	ODAR
Deeper temporal ConvNets	57.72%
Deeper RGB ConvNets	32.55%
Combined	34.83%

Table 5. Performance of deeper temporal ConvNets, deeper RGB ConvNets or combination of deeper temporal and RGB ConvNets on the ODAR2017 dataset.

One thing that needs to be specified, the reason why our Deeper temporal ConvNets could not yield a same good performance as the original Temporal net [20](75.38% vs 83.7%) can be explained as follows. As discussed in section 3.2, we use parameters of ResNet-101 layers trained on the ImageNet to initialize deeper temporal ConvNets. 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 ConvNets. 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 ODAR2017 dataset. The results was illustrated in Table 5. The best accuracy of 57.72% was achieved on ODAR2017 dataset. In Figure 6, we show the confusion matrix for ODAR2017 dataset classification using our deeper temporal ConvNets. Besides, the corresponding per-class recall are also visualised in Figure 7. It can be seen that the worst class recall corresponds to *unknow* class, which is mainly confused with *celltoear*, *drink* and *wave* class. We found that this is caused by the significant presence of same motion patterns (hand moving up and down) in both classes.

Limitation and Discussion. Due to the time limitations, we did not optimize the our model and achieve the higher performance on UCF101. However, our proposed ideas that deeper ConvNets obtained better performance are worth noting. Later, we will optimize the model by adjusting the parameters or some other measures. Believing that

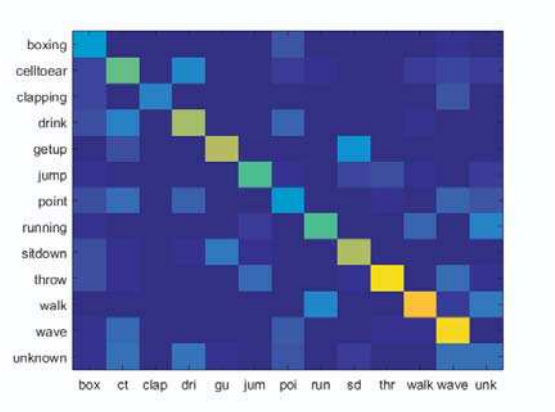

Figure 6. Confusion matrix of deeper temporal ConvNets on the ODAR dataset.

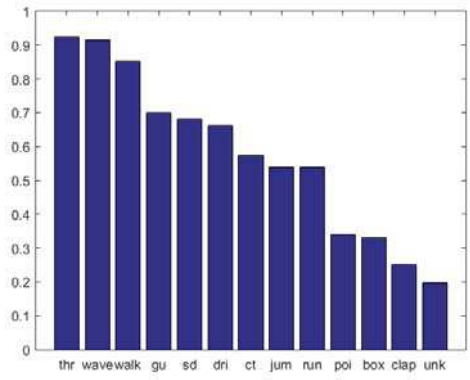

Figure 7. Per-class recall of deeper temporal ConvNets on the O-DAR dataset.

a promising performance should also be able to achieved on the dataset of UCF101 or other databases.

4.4. Additional Experiments

In this section, we also do some additional experiments for validating the deeper RGB ConvNets or combination of deeper RGB and temporal ConvNets. The deeper RGB ConvNets shares similar architectures with deeper temporal ConvNets except for the input data layer. The deeper RGB ConvNets takes as input a square 224×224 pixel RGB image and performs action recognition by exploiting appearance features. Besides, the training or testing of deeper RGB ConvNets are also similar with deeper temporal ConvNets. Combination of deeper RGB and temporal ConvNets means that softmax scores of each ConvNets are combined by late fusion.

The results of deeper RGB ConvNets or combination of deeper RGB and temporal ConvNets on the UCF101 dataset illustrated in Table 3. Based on investigation, it can be seen that the deeper RGB ConvNets achieved a superior performance than deeper temporal ConvNets. For combining of deeper RGB and temporal ConvNets, we take a weighted average of them. As shown in Table 3, the performance gap between deeper RGB ConvNets and deeper temporal ConvNets is much smaller. Based on this fact, we give more credits to the temporal ConvNets by setting its weight as 1 and that of RGB ConvNets as 1.2. The fusion of them can further boost the performance. This further improvement indicates that appearance features extracted from deeper RGB ConvNets are complementary to those motion features of deeper temporal ConvNets.

Besides, the results on the ODAR dataset were illustrated in Table 5. Combination is carried out by setting weight of deeper temporal ConvNets as 2 and that of deeper RGB ConvNets as 1. It can be found that the deeper RGB ConvNets on ODAR dataset did not yield good performance as on the UCF101 dataset. As shown in Table 5, the performance of deeper RGB ConvNets was below that of deeper temporal ConvNets by a large margin. Besides, the deeper RGB ConvNets on ODAR dataset did not act as complementary feature to further boost the performance as deeper RGB ConvNets on the UCF101 dataset did. We concluded that this is due to three reasons. First, the ODAR dataset are suffered from more poor image quality than UCF101, which was caused by darkness of light conditions and low resolution. Second, higher inter-class similarities in ODAR dataset which caused by significant presence of same human or scene in different classes. Finally, compared with UCF101 dataset, the action classes of ODAR are rather simple actions such as “walk”, “running”, and “wave”, which involves only body movements. As a consequence, the deeper RGB ConvNets extracted little high-level visual cues for action recognition.

4.5. Visualisation

To attain an insight into the learned deeper RGB or temporal ConvNets, some frame images or their corresponding optical flow fields belonging to different action classes such as “PommelHorse”,“Basketball” in UCF101 dataset, and “getup”, “drink” in ODAR dataset are selected to feed into deeper RGB net or deeper temporal ConvNets respectively. The feature maps of last convolutional layer that corresponding to deeper RGB or temporal ConvNets are then visualized in Figure 8. From the visualized examples, we see that our feature maps of last convolutional are relatively sparse and exhibit a high correlation with the action areas, which indicates that our deeper RGB or temporal ConvNets has high modeling capacity to obtain discriminative features for action recognition in videos. Besides, it can be seen that the feature maps of ODAR dataset are largely more sparse than those of UCF101, it verified that the the deeper RGB or temporal ConvNets of ODAR dataset extracted little high-level visual cues for action recognition as discussed in

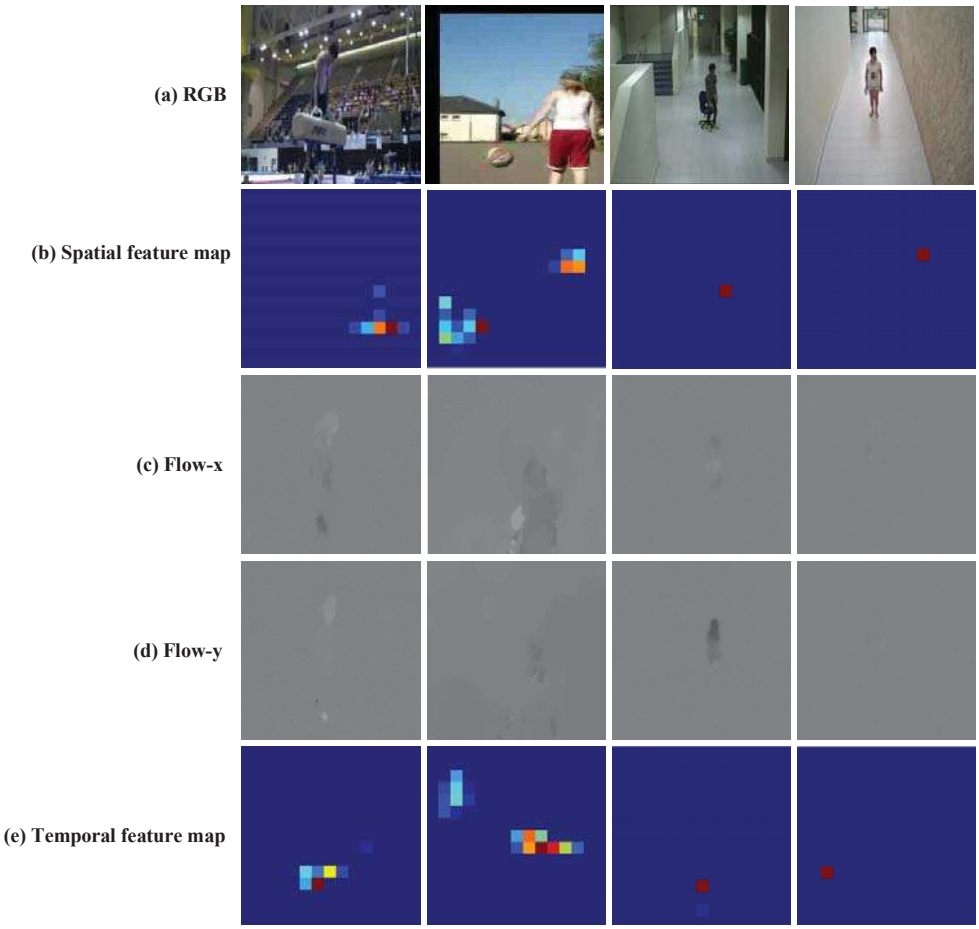

Figure 8. Visualisation of learnt convolutional filters. We select some video frames, optical flow fields, and visualize their corresponding feature maps of deeper RGB ConvNets and deeper temporal ConvNets.

Section 4.4.

5. Conclusions and Future work

In this paper, we proposed a novel deeper temporal ConvNets that learning of high-level motion information 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 dis-ordering strategy among video sets listed in testing splits. Besides, the additional experiments are carried out to explore how well deeper RGB ConvNets or combination of deeper RGB and temporal ConvNets performed on UCF101 and ODAR dataset. The empirical experiments have shows that the proposed deeper temporal ConvNets achieved a superior performance.

In experimental evaluation, we also found that deeper temporal ConvNets did not yield good performance as the deeper RGB ConvNets 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 future studies.

6. 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, and Key NSF grant S2017QNZDB0041 approved by the Jiangxi Provincial Department of Science and Technology, China, and the National Research Foundation, Prime Ministers Office, Singapore under its International Research Centre in Singapore Funding Initiative.

References

[1] J. K. Aggarwal and M. S. Ryoo. Human activity analysis: A review. *ACM Computing Surveys (CSUR)*, 43(3):16, 2011.

[2] Y. Bengio, P. Y. Simard, and P. Frasconi. Learning long-term dependencies with gradient descent is difficult. *IEEE Transactions on Neural Networks*, 5(2):157–166, 1994.

[3] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 248–255. IEEE, 2009.

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

[5] G. Gkioxari, R. Girshick, and J. Malik. Contextual action recognition with r* cnn. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1080–1088, 2015.

[6] X. Glorot and Y. Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Aistats*, volume 9, pages 249–256, 2010.

[7] D. Gong, J. Yang, L. Liu, Y. Zhang, I. Reid, C. Shen, A. v. d. Hengel, and Q. Shi. From motion blur to motion flow: a deep learning solution for removing heterogeneous motion blur. *arXiv preprint arXiv:1612.02583*, 2016.

[8] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *arXiv preprint arXiv:1512.03385*, 2015.

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

[10] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.

[11] S. Ji, W. Xu, M. Yang, and K. Yu. 3d convolutional neural networks for human action recognition. *IEEE transactions on pattern analysis and machine intelligence*, 35(1):221–231, 2013.

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

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

[14] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

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

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

[17] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[18] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.

[19] C. Schuldt, I. Laptev, and B. Caputo. Recognizing human actions: A local svm approach. In *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, volume 3, pages 32–36. IEEE, 2004.

[20] K. Simonyan and A. Zisserman. Two-stream convolutional networks for action recognition in videos. In *Advances in Neural Information Processing Systems*, pages 568–576, 2014.

[21] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[22] K. Soomro, A. R. Zamir, and M. Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.

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

[24] G. W. Taylor, R. Fergus, Y. LeCun, and C. Bregler. Convolutional learning of spatio-temporal features. In *European conference on computer vision*, pages 140–153. Springer, 2010.

[25] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri. Learning spatiotemporal features with 3d convolutional networks. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 4489–4497. IEEE, 2015.

[26] H. Wang and C. Schmid. Action recognition with improved trajectories. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3551–3558, 2013.

[27] H. Wang and C. Schmid. Lear-inria submission for the thumos workshop. In *ICCV workshop on action recognition with a large number of classes*, volume 2, page 8, 2013.

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

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

[30] D. Weinland, R. Ronfard, and E. Boyer. Free viewpoint action recognition using motion history volumes. *Computer vision and image understanding*, 104(2):249–257, 2006.

[31] J. Yuan, Z. Liu, and Y. Wu. Discriminative subvolume search for efficient action detection. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 2442–2449. IEEE, 2009.

[32] C. Zach, T. Pock, and H. Bischof. A duality based approach for realtime tv-l 1 optical flow. In *Joint Pattern Recognition Symposium*, pages 214–223. Springer, 2007.

[33] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In *European Conference on Computer Vision*, pages 818–833. Springer, 2014.

[34] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva. Learning deep features for scene recognition using places database. In *Advances in neural information processing systems*, pages 487–495, 2014.