Action Recognition with Motion Map 3D Network

Abstract

Recently, deep neural networks have demonstrated remarkable progresses for human action recognition in videos. However, most existing deep frameworks can not handle variable-length videos properly, which leads to the degradation in classification performance. In this paper, we propose a Motion Map 3D ConvNet(MM3D), which can represent the content of a video with arbitrary video length by a motion map. In our MM3D model, a novel generation network is proposed to learn a motion map to represent a video clip by iteratively integrating a current video frame into a previous motion map. A discrimination network is also introduced for classifying actions based on the learned motion map. Experiments on the UCF101 and the HMDB51 datasets prove the effectiveness of our method for human action recognition.

Keywords: Action recognition, video analysis, 3D-CNN, discriminative information

1. Introduction

Human action recognition aims to automatically classify the action in a video, and it is a fundamental topic in computer vision with many important applications such as video surveillance and video retrieval. As revealed by [1], the quality of action representations has an influence on the performance of action recognition, which means that learning a powerful and compact representation of an action is an important issue in action recognition.

In recent years, many approaches have been proposed to learn deep features of videos for action recognition. A slow fusion method [2] is presented to extend the connectivity of the network in temporal dimension to learn video features.

In [3], a two stream network is proposed to learn spatio-temporal features by using the optical flow and the original image at the same time. The C3D method [4] exploits 3-dimensional convolution kernels to directly extend the convolution operation of the image to the operation of the frame sequence. These methods can only learn the feature of the fixed-length video clip. Unfortunately, the lengths of videos are variable, and these existing works need resort to the pooling methods [4] or the feature aggregation methods [5][6] to generate a final representation of the entire action video.

In order to solve the problem of representing variable length videos, Bilen et al.[7] proposed dynamic images by using dynamic image network to represent action videos, which takes the order of video frames as the supervisory information without considering the category information of actions. The dynamic image is not able to capture the discriminative information of videos, resulting in the degradation of recognition accuracy.

Figure 1: The motion maps generated by the generation network represent the motions and static objects of video in various categories. The discriminative information integrated in the motion map can be used to classify the categories of videos.

In this paper, we propose a novel Motion Map 3D ConvNet (MM3D) to learn a motion map for representing an action video clip. By removing a large number of information redundancy of an action video, the motion map is a powerful, compact and discriminative representation of a video. As shown in Fig. 1, the motion maps learned by our MM3D model can capture distinguishable trajectories around the human body.

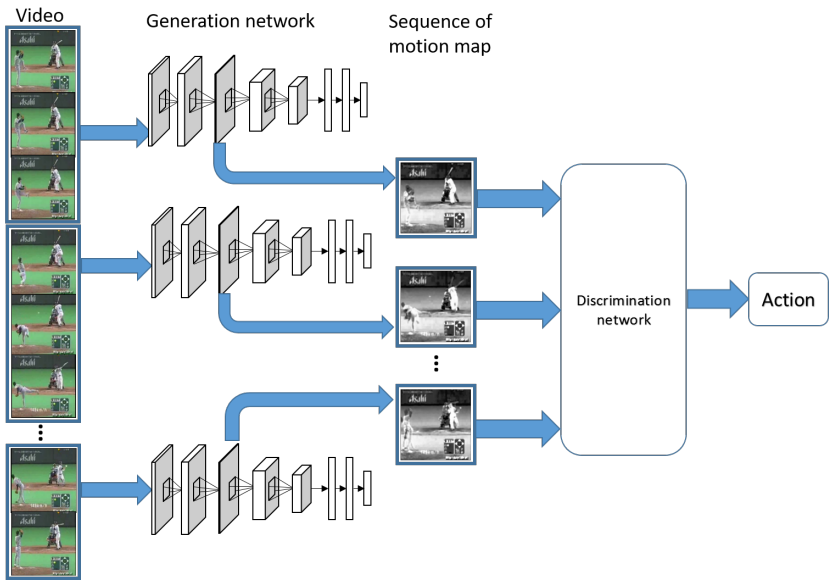

Figure 2: The framework of the MM3D

The proposed MM3D model consists of two networks: a generation network and a discrimination network. The framework of our MM3D is illustrated in Fig. 2. The generation network learns the motion maps of variable-length video clips, by integrating the temporal information into a map without losing the discriminative information of video clips. Specifically, it integrates the motion map of previous frames with the current frame to generate a new motion map. After the repetitive integration of the current frame, the final motion map is generated for the entire video clip by capturing the motion information. Besides, the action class labels are used as the supervisory information to train the generation network, so the learned motion map can also exploit the discriminative

information of action videos.

Despite the good performance of the motion map on capturing the local temporal information of the video clips, a single motion map is not sufficient to capture the complex dynamics of an entire action. To learn the long term dynamics from the whole video and demonstrate the power of the motion maps on action recognition, a discrimination network is proposed. The architecture of this network is based on the 3D-CNN model [4] which has been shown powerful performance on action recognition. The input of the discrimination network is a sequence of the motion maps and the discriminative action feature based on the motion maps is extracted from the *pool*5 layer of the network.

The contributions of this work are two-fold. (1) We propose a new network to generate motion maps for action recognition in videos. The generated motion maps contain the temporal information and the discriminative information of the action video with an arbitrary video length. (2) We propose a discrimination network based on the motion maps to deal with the complex and long-term action video. The network can learn the discriminative features of the sequence of motion maps which benefits boosting the accuracy of action recognition.

This paper is an extended version of our prior conference publication [8]. The main differences are as follows. (1) This paper proposes a novel discrimination network which takes a sequence of motion maps as input. The discrimination network can learn the long term dynamics from the whole video and demonstrate the power of the motion maps on action recognition. (2) To show the effect of our discrimination network, an extended experiment is conducted for the comparison of the single image per video setting with the sequence of images per video setting. The observation of the results shows that the use of discrimination network can improve the accuracy of action recognition for our own motion map, up to 17.4% on the UCF101 and 18.7% on the HMDB51. Another extended experiment is provided to compare our method using discrimination network with the state-of-the-art methods. (3) This paper gives a more extensive overview and comparison of the related literature.

2. Related Work

Action recognition has been studied by the computer vision researchers for decades. To address this issue, various methods have been proposed, of which the majority is about action representations. These action representations can be briefly grouped into two categories: hand-crafted features and deep learning-based features.

Hand-crafted Feature: Since videos can be taken as a stream of video frames, many video representations are derived from the image domain. Laptev and Lindeberg [9] proposed space-time interest points (STIPs) by extending 2D Harris corner detector to 3D Harris corner detector. In the same way, SIFT and HOG are also extended into SIFT-3D [10] and HOG3D [11] for action recognition. The improved Dense Trajectories (iDT) [12], which is currently the state-of-the-art hand-crafted feature, can densely sample feature points from each frame at different scales and then use optical flows to track the feature points.

Motion is extremely important information of an action, and it is beneficial for action recognition. Ali et al [13] generated the video feature by calculating the optical flow of a sequence of frames. Kellokumpu et al [14] captured the motion information by the dynamic textures. Bilen et al [7] used the dynamic image network based on rank pooling [15] to generate dynamic images for the video. They learn the video representations end-to-end while being much efficient. Despite its good performance, the huge computational cost greatly limits the large-scale use of dense trajectory methods.

Deep Feature: Recently, deep features have also been investigated for action recognition with the availability of large amounts of training data. The methods of [16][17] have shown that learning visual representations with CNNs is superior to hand-crafted features for many recognition tasks in image domain. Extensions of CNN representations to action recognition in video have been proposed in several recent works. [2] investigates multiple approaches for fusing information over temporal dimension through the network. Multiple

stream based methods have been widely used in different kinds of contexts, and have achieved promising results in action recognition. Simonyan et al. [3] combined static frames and optical flow frames by using two-stream networks for action recognition. Fernando et al [18] used a ranking machine to learn ranking functions per video by temporally ordering the frames of the video. Ng et al [19] considered the video as an ordered sequence of still frames by modeling the temporal information using LSTM.

By extending 2D convolutional networks, the 3D Convolutional networks have been explored as natural methods for video modeling [4] [20] [21]. The 3D Convnet uses 3D convolutional kernels to extend the convolution operation on the spatio-temporal volumes, and can get hierarchical representations for the video volumes. Tran et al [4] trained the network on the Sports-1M dataset, showing the good performance of the 3D Convnet. Some other methods have improved the 3D Convnet by modifying the blocks of the network. Wang et al [22] proposed a new SMART block based on the 3D Convnet to model apperance and relation separately and explicitly with a two-branch architecture. Wang et al [23] inserted non-local blocks into C2D [4] and I3D [24] to turn them into the non-local 3D Convnets. However, the 3D Convnet is always fixed-length on the third dimension, which makes it difficult to learn complete temporal information. Besides, it is difficult to train a 3D convolutional network with a small scale dataset because of the large number of parameters.

3. Method

In this section, we first introduce the concept of the motion map which is used to represent the video clips. Then, we present the architecture of the proposed network for learning motion map. Finally, we explain in detail the training and prediction procedures of our network.

3.1. The Motion Map

The motion map is an image used to represent a video clip. We denote a video clip as $V = \{f_i\}_{i=1}^{N}$, where f_i is the i-th frame in V, and N is the total

frame of the video clip. Then, define F_i as the motion map from f_1 to f_i, so the first motion map F_1 is actually the first frame f_1. In order to keep more important information, an iterative method is proposed to generate the motion map F_i by combining the motion map F_{i-1} with the video frame f_i through the generation network:

$$F_i = F_{i-1} \oplus f_i \tag{1}$$

After the iteration, the final motion map F_N of video V can be obtained. The discriminative information embodied in the single motion map F_N can be applied to the action recognition tasks.

Fig. 1 illustrats several examples of motion maps from various categories. Specifically, Fig.1(a) shows that the static objects such as windows and floors are presented as they are and the superposed silhouette incarnates the different location and posture of the man when he raises and lowers the body using the arms. Fig.1(d) shows that the woman's body and cello are barely moving, while the location of her head, arms, and fiddlestick are changing, hence the superposed shadows reflect the action of playing cello. Fig.1(l) shows that the arms and yo yo ball are the main features of this motion map while the rest of the motion map is diluted, which shows the relationship between the arms movements and playing yo yo ball. It proves that dynamic information is reflected by the superposition of different frames, when the static information in the video frame sequences can be retained. Thus, the motion map can effectively incarnate the discriminative information.

3.2. The Architecture of Network

3.2.1. The generation network

The architecture of the generation network is illustrated in Fig. 3. The network has two 3D convolutional layers with 64 and 1 feature maps, followed by two identical 2D convolutional layers with 128 feature maps. After that, there are 3 fully connected layers with the size of 2048, 2048 and the number of action classes in the network. Both 3D, 2D convolutional layers and fully

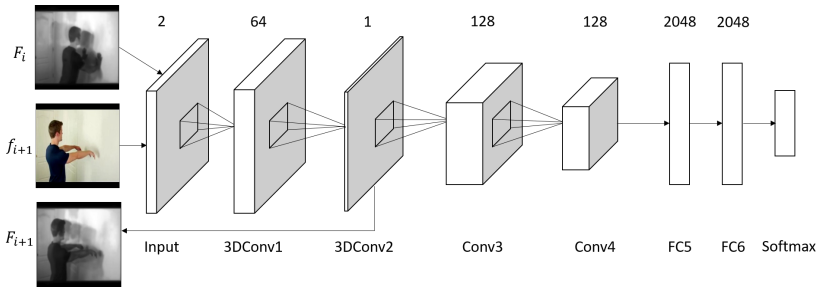

Figure 3: The architecture of the generation network.

connected layers are followed by a rectified linear unit. For the convolutional networks, we also use the max pooling layer behind the rectified linear unit. The kernel sizes of the 3D convolutional layers are $3 \times 3 \times 1$ and $3 \times 3 \times 2$, respectively. The kernel sizes of the following 3D max pooling layers are both $3 \times 3 \times 1$. The 2D convolutional filter is of size 9×9 and the 2D max pooling filter is of size 4×4. The softmax layer is used for image classification after the last fully connected layer.

The previous motion map and the current video frame are combined into a sequence as the input of our generation network, and the output layer is 3DConv2. The special structure of layer 3DConv2 can generate a motion map that integrates the information of input. During the training step, we consider the generation network as an action recognition network that exploits the information of both the previous motion map and the current video frame. The generation network can be trained in an end-to-end manner. After training with the labels of videos, the softmax layer output given by the network shows that the single motion map (3DConv2) provides discriminative information to distinguish actions. When the network is well trained, the effect of using a pair of images and using a single motion map is consistent.

Using the generation network to integrate two images into one, the motion map of a whole video can be iteratively generated. As shown in Fig. 4(a), the generation network accepts two images including a previous motion map and a current video frame, and produces a new motion map. According to Eq. 1,

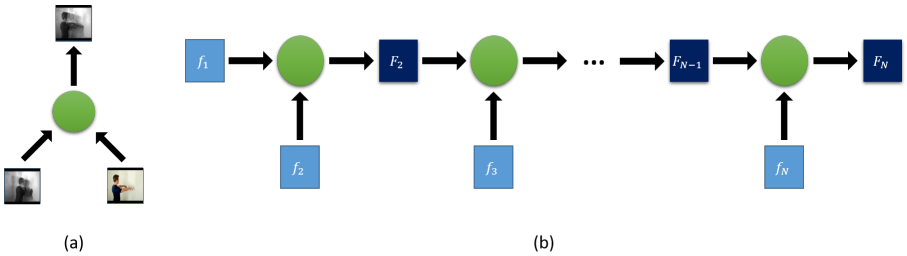

Figure 4: (a) The generation network takes two images as input, and generates a motion map by integrating the temporal information of input; (b) A motion map of a video is generated by using the generation network iteratively.

we use the generation network iteratively to get the final motion map, and the whole process is shown in Fig. 4(b).

3.2.2. The discrimination network

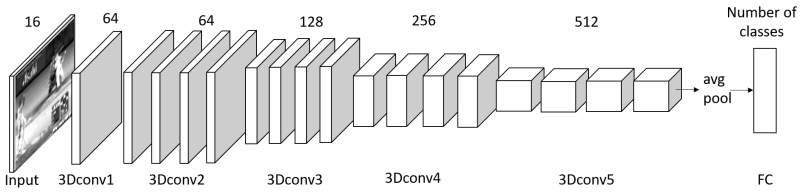

Figure 5: The architecture of the discrimination network.

Although the generation network can learn the motion maps over arbitrary length video sequences, a single motion map is not sufficient to capture the complex dynamics of the action video, when the video contains hundreds of video frames. In order to address the complex action videos and achieve the competitive performances on action recognition by the generated motion maps, we design the discrimination network. The architecture of the discrimination network is illustrated in Fig. 5, which is based on the 3D-CNN network [25] and the 3D residual architectures. A video is split into several video clips which are used to generate a sequence of motion maps. Each motion map contains both static and dynamic information of the video clip. The network takes a sequence of motion maps as input, and the features of *pool*5 are extracted for recognition.

The network is trained on the Sports-1M dataset as the initialization for our discrimination network and finetuned by the motion map.

3.3. Training

An iterative method is proposed to train the generation network. Before the training prrocedure, each video in the dataset is split into several video clips. K is the number of action classes in the dataset, and N_i represents the total number of F_i. For the first step, we construct the training dataset $C_1 = \{(f_1, f_2, k)_n\}_{n=1}^{N_2}$, where f_1 and f_2 are the first two frames of the video clip and the label $k \in \{1...K\}$ is the action class of the video clip. After training, the feature maps of the dataset can be extracted from the 3DConv2 layer, and transformed into the motion map F_2. We repeat this step, until the maximum number of video clip frames is reached. For the step s, $C_s = \{(F_s, f_{s+1}, k)_n\}_{n=1}^{N_{s+1}}$ is added into the training dataset, to train our model with the same method. Finally, the learned model we get can be used to generate the motion map of the video clips.

Algorithm 1 The training algorithm for the generation network

Input: The video clips set, V;
The video labels, K;
The maximum training iteration length, S;
The number of video clips, $N_1...N_S$;

Output: The model of the generation network, M;

1: Construct the initial training dataset $C_1 = \{(f_1, f_2, k)_n\}_{n=1}^{N_2}$
2: Use the training dataset to train the generation network;
3: **for** each $s \in 2, 3, ..., S$ **do**
4: Generate the motion map $\{(F_s)_n\}_{n=1}^{N_s}$ using the dataset of C_{s-1}
5: Add $C_s = \{(F_s, f_{s+1}, k)_n\}_{n=1}^{N_{s+1}}$ into the training dataset;
6: Use the training dataset to train the generation network;
7: **end for**
8: Return the model of the generation network;

For the discrimination network, each video is split into 16 video clips. Different video clips could have different numbers of frames. The motion map for each video clips can be generated using the generation network iteratively. For a video, 16 motion maps and one label is used to train our network. To achieve the competitive performance on the classification, we take the parameters of the C3D networks [4] training on the Sports-1M dataset as the initialization of our network, and finetune it using the motion maps of the video.

In the experiments, the learning rate is 0.0003 at first, and reduced by 10 times every thirty thousand iterations with the momentum of 0.9 and the weight decay of 0.00005.

3.4. Prediction processing

During the prediction processing, a test video is split into 16 video clips. For each clip, a motion map can be generated step by step as the processing of training. The final model trained on the generation network is adapted for the prediction processing. One motion map with all of the discriminative information is available, when all the frames of the video clip is used. Then, the 16 motion maps are taken as the input of the discrimination network. Finally, we extract the feature of *pool*5 layer, and use a linear SVM to get the final recognition results.

4. Experiments

In this section, we validate the proposed network architecture on two standard action classification benchmarks, i.e., the UCF101 and HMDB51 datasets. Our method is firstly compared with two baseline methods, i.e., the single frame method and the dynamic image method. Then, we demonstrate the comparison results between our method and the state-of-the-art methods on the two datasets.

4.1. Dataset

4.1.1. UCF101

The UCF101 dataset [13] comprises of 101 action categories, over 13k clips and 27 hours of video data, and the mean video clip length is 7.21 seconds. The database consists of realistic user-uploaded videos containing camera motion and cluttered background, such as surfing and ice dancing. The dataset is trimmed, thus each frame of a video is related to its category.

4.1.2. HMDB51

The HMDB51 dataset [9] comprises of 51 action categories, which in total contains around 7,000 manually annotated clips extracted from a variety of sources ranging from digitized movies to YouTube videos. Compared with the UCF101 dataset, the HMDB51 dataset includes more complex backgrounds and more intra class differences.

4.2. Implementation Details

4.2.1. Video Frames

Since the generation network takes a sequence of still frames as input, the video needs to be converted into a sequence of video frames. When we convert the video into video frames, we find that different videos have different frame rates, which may have an adverse effect on the recognition. Therefore, the number of extracted video frames is based on the total time of the video rather than the number of video frames. In all the experiments, two frames are extracted per second, which makes it possible to capture the action changes with the generation network. For some short videos, some frames are filled for the video, thus each video has 16 frames at least.

4.2.2. Motion Map

As described in Section 3, the generation network is used to generate the motion map. The motion map cannot be directly obtained by the network. The feature map of 3DConv2 Layer can be extracted in the network. Then, we

propose to scale the value in feature map and perform histogram equalization to generate the motion map.

4.2.3. Data augmentation

The number of training data is very important for the generation network and the discrimination network. In order to get better performance, the mirroring and cropping methods are used to extend the original data. The difference between our method and the traditional methods is that a video clip instead of a video frame is the base unit for the expansion.

4.3. Generation network

In order to demonstrate the effectiveness of the motion map, we experiment with the single image per video setting, which means that an image or a map is used to represent a video. Firstly, the generation network generates a single motion map (SMM) for each video in each dataset. Then we use the motion maps to finetune the VGG network [26] for the action recognition task. The result is shown in Table 1 and Table 2.

Table 1: Accuracies (%) of single image per video setting on the UCF101 dataset

Approach	Split1	Split2	Split3	Average
Average Image	52.6	53.4	51.7	52.6
SDI [7]	57.2	58.7	57.7	57.9
SMM	**61.2**	**61.4**	**60.5**	**61.0**

Table 2: Accuracies (%) of single image per video setting on the HMDB51 dataset

Approach	Split1	Split2	Split3	Average
Average Image	34.1	33.8	33.9	33.9
SDI [7]	37.2	36.4	36.9	36.8
SMM	**39.3**	**37.1**	**37.0**	**37.8**

We use the motion maps generated by the generating network to compare with the Average Image and the SDI proposed in [7]. Average Image is the

average of all the video frames. The comparison results suggest that our motion map contains more discriminative information than the dynamic image and average image.

4.4. Discrimination network

In the previous experiment, to show the effect of the motion map, we compare our motion map with the dynamic image and the average video frame. The setting for the experiment is the single image per video. Next, we compare the single image per video setting with the sequence of images per video setting. For the single image per video setting, the steps are the same as the pervious experiment. For the sequence of images per video setting, a video is split into 16 video clips. We use different methods to integrate the video clip into a single image, therefore a sequence of images is used to represent the whole video. Then, the sequence of images is taken as the input to fine-tune our discrimination network for action recognition task. For the Sequence of Frames (Sof), we choose the first frame of each video clip. For the Sequence of Dynamic Images (SoDI), we get a dynamic image of each video clip using the method of SDI [7]. For the Sequence of Motion Maps (SoMM), the generation network is used to generate a motion map of the video clip. Table 3 reports the comparison on the UCF101 and HMDB51 datasets. The observation of the results shows that the use of discrimination network can improve the accuracy of action recognition for our own motion map, up to 17.4% on the UCF101 and 18.7% on the HMDB51. The other methods are also improved by the discrimination network according to the Table 1 and 2. The reason for the improvement is that a sequence of the images contains more discriminative information than a single image and our discrimination network can collect the discriminative information from the sequence of images. The comparison results also suggest that our motion map outperforms the other representation methods.

4.5. Comparison with the State-of-the-Art

Taking into account the complexity of the calculation and the accuracy of action recognition, we first combine our method with a hand-crafted feature

Table 3: Accuracy (%) comparison between the single motion map(SMM) and the sequence of the images (SoF,SoDI,SoMM) on the UCF101 dataset and the HMDB51 dataset

		Split1	Split2	Split3	Average
UCF101	SMM	61.2	61.4	60.5	61.0
	SoF	77.2	77.8	76.6	77.2
	SoDI	78.3	76.8	74.1	76.4
	SoMM	82.0	79.0	74.1	78.4
HMDB51	SMM	39.3	37.1	37.0	37.8
	SoF	55.1	54.2	53.7	54.3
	SoDI	53.8	54.3	53.1	53.7
	SoMM	59.7	54.7	55.0	56.5

called MIFS [27]. The combination method is the multiple kernels learning [28]. The results are shown in Table 4. Compared to the accuracy of MIFS, our method has achieved 8.6% improvement on the HMDB51 dataset and 2.8% improvement on the UCF101 data set.

Table 4: Accuracies (%) of the combination of our method and MIFS [27] feature

Approach	UCF101	HMDB51
MIFS [27]	89.1	65.1
Motion Map + MIFS [27]	91.9	73.7

We show a comparison of our method with the state-of-the-art method on the UCF101 and the HMDB51 datasets in Table 5. The table are divided into two parts. The methods of the first part only use the original RGB images, while the methods of the second part take both RGB images and the optical flow fields as the input of the network. Our proposed feature with the feature of MIFS achieve 91.9% on the UCF101 dataset and 73.7% on the HMDB51 dataset. Compared with the most methods, our method achieve the highest performance on the HMDB51 dataset, and is also competitive on the UCF101 dataset.

Table 5: Accuracy (%) comparison of our method with the state-of-the-art methods

Approach	UCF101	HMDB51
FV + IDT [29]	84.8	57.2
FSTCN [30]	88.1	59.1
2S-CNN + LSTM [19]	88.6	-
Dynamic Image Network [7]	89.1	65.2
C3D + IDT + SVM [4]	90.4	-
TDD + IDT [31]	91.5	65.9
Two-Stream Fusion + IDT [32]	93.5	69.2
TSN [33]	94.0	68.5
Our Method	91.9	73.7

5. Conclusion

In this paper, we have introduced the concept of a motion map. A motion map is a powerful representation of an arbitrary video which contains both the static and dynamic information. We also propose a Motion Map 3D ConvNet which can generate a motion map for a video clip and an iterative training method to integrate the discriminative information into a single motion map.

In future, we would like to extend our method on other tasks, such as temporal action localization, the action duration of which changes greatly, therefore precisely motion map learning can be useful for the localization.

References

[1] D. Weinland, R. Ronfard, E. Boyer, A survey of vision-based methods for action representation, segmentation and recognition, Elsevier Science Inc., 2011.

[2] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, L. Fei-Fei, Large-scale video classification with convolutional neural networks, in: Pro-

ceedings of the IEEE conference on Computer Vision and Pattern Recognition, 2014, pp. 1725–1732.

[3] K. Simonyan, A. Zisserman, Two-stream convolutional networks for action recognition in videos, in: Advances in neural information processing systems, 2014, pp. 568–576.

[4] D. Tran, L. Bourdev, R. Fergus, L. Torresani, M. Paluri, Learning spatiotemporal features with 3d convolutional networks, in: Proceedings of the IEEE international conference on computer vision, 2015, pp. 4489–4497.

[5] H. Jégou, M. Douze, C. Schmid, P. Pérez, Aggregating local descriptors into a compact image representation, in: Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on, IEEE, 2010, pp. 3304–3311.

[6] F. Perronnin, C. Dance, Fisher kernels on visual vocabularies for image categorization, in: Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on, IEEE, 2007, pp. 1–8.

[7] H. Bilen, B. Fernando, E. Gavves, A. Vedaldi, S. Gould, Dynamic image networks for action recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 3034–3042.

[8] Y. Wennan, S. Yuchao, Y. Feiwu, W. Xinxiao, Representing discrimination of video by a motion map, The 2017 Pacific-Rim Conference on Multimedia.

[9] I. Laptev, On space-time interest points, International journal of computer vision 64 (2-3) (2005) 107–123.

[10] P. Scovanner, S. Ali, M. Shah, A 3-dimensional sift descriptor and its application to action recognition, in: Proceedings of the 15th ACM international conference on Multimedia, ACM, 2007, pp. 357–360.

[11] A. Klaser, M. Marszałek, C. Schmid, A spatio-temporal descriptor based on 3d-gradients, in: BMVC 2008-19th British Machine Vision Conference, British Machine Vision Association, 2008, pp. 275–1.

[12] H. Wang, C. Schmid, Action recognition with improved trajectories, in: Proceedings of the IEEE international conference on computer vision, 2013, pp. 3551–3558.

[13] S. Ali, M. Shah, Human action recognition in videos using kinematic features and multiple instance learning, IEEE transactions on pattern analysis and machine intelligence 32 (2) (2010) 288–303.

[14] V. Kellokumpu, G. Zhao, M. Pietikäinen, Human activity recognition using a dynamic texture based method., in: BMVC, Vol. 1, 2008, p. 2.

[15] B. Fernando, E. Gavves, J. Oramas, A. Ghodrati, T. Tuytelaars, Rank pooling for action recognition, IEEE transactions on pattern analysis and machine intelligence 39 (4) (2017) 773–787.

[16] 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.

[17] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, L. D. Jackel, Backpropagation applied to handwritten zip code recognition, Neural computation 1 (4) (1989) 541–551.

[18] B. Fernando, E. Gavves, J. M. Oramas, A. Ghodrati, T. Tuytelaars, Modeling video evolution for action recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 5378–5387.

[19] 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.

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

[21] 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.

[22] L. Wang, W. Li, W. Li, L. Van Gool, Appearance-and-relation networks for video classification, arXiv preprint arXiv:1711.09125.

[23] X. Wang, R. Girshick, A. Gupta, K. He, Non-local neural networks, arXiv preprint arXiv:1711.07971.

[24] J. Carreira, A. Zisserman, Quo vadis, action recognition? a new model and the kinetics dataset, arXiv preprint arXiv:1705.07750.

[25] D. Tran, J. Ray, Z. Shou, S.-F. Chang, M. Paluri, Convnet architecture search for spatiotemporal feature learning, arXiv preprint arXiv:1708.05038.

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

[27] Z. Lan, M. Lin, X. Li, A. G. Hauptmann, B. Raj, Beyond gaussian pyramid: Multi-skip feature stacking for action recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 204–212.

[28] M. Gönen, E. Alpaydın, Multiple kernel learning algorithms, Journal of machine learning research 12 (Jul) (2011) 2211–2268.

[29] X. Peng, C. Zou, Y. Qiao, Q. Peng, Action recognition with stacked fisher vectors, in: European Conference on Computer Vision, Springer, 2014, pp. 581–595.

[30] L. Sun, K. Jia, D.-Y. Yeung, B. E. Shi, Human action recognition using factorized spatio-temporal convolutional networks, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 4597–4605.

[31] 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.

[32] C. Feichtenhofer, A. Pinz, A. Zisserman, Convolutional two-stream network fusion for video action recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 1933–1941.

[33] 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.