Pattern Recognition Letters 118 (2019) 14–22

Contents lists available at ScienceDirect

Pattern Recognition Letters

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

A review of Convolutional-Neural-Network-based action recognition

Guangle Yao[a,b,c], Tao Lei[a,*], Jiandan Zhong[a,b,c]

[a] Institute of Optics and Electronics, Chinese Academy of Sciences, P.O. Box 350, Shuangliu, Chengdu 610209, China
[b] University of Electronic Science and Technology of China, No.4, Section 2, North Jianshe Road, Chengdu 610054, China
[c] University of Chinese Academy of Sciences, 19 A Yuquan Rd, Shijingshan District, Beijing 100039, China

ARTICLE INFO

Article history:
Available online 23 May 2018

Keywords:
Action recognition
Deep learning
Convolutional Neural Network
Action representation

ABSTRACT

Video action recognition is widely applied in video indexing, intelligent surveillance, multimedia understanding, and other fields. Recently, it was greatly improved by incorporating the learning of deep information using Convolutional Neural Network (CNN). This motivated us to review the notable CNN-based action recognition works. Because CNN is primarily designed to extract 2D spatial features from still image and videos are naturally viewed as 3D spatiotemporal signals, the core issue of extending the CNN from image to video is temporal information exploitation. We divide the solutions for exploiting temporal information exploration into three strategies: 1) 3D CNN; 2) taking the motion-related information as the CNN input; and 3) fusion. In this paper, we present a comprehensive review of the CNN-based action recognition methods according to these strategies. We also discuss the action recognition performance on recent large-scale benchmarks and the limitations and future research directions of CNN-based action recognition. This paper offers an objective and clear review of CNN-based action recognition and provides a guide for future research.

1. Introduction

Recognizing and understanding human actions and intentions is essential in robotics, human-computer interaction, intelligent surveillance, etc. and has thus been an important and popular research topic. In the past few decades, a large number of video action recognition methods have been proposed. In addition, various action recognition datasets and benchmarks for action recognition have been released. Since 2012, Convolutional Neural Network (CNN) [1] has been widely employed in the image domain, significantly improving the performance of image classification [2], object detection [3], scene classification [4], etc. The success of CNN in the image domain has inspired the research on CNN-based action video recognition. The methodologies and CNN architectures extended from image to video have undergone significant advancements. Therefore, CNN-based action recognition has become an active research field, with numerous CNN-based approaches, which have also achieved great success, recently emerging.

As illustrated in Fig. 1, video action recognition is divided into two kinds of tasks [5]: assigning a video to a set of predefined action classes and locating predefined action temporally in a video, which are referred to as classification and detection, respectively.

Compared with visible (gray or RGB) video, depth video is not sensitive to illumination and capable of capturing the geometric information of the object, while infrared thermal video is robust against the challenges of pool imaging light, illumination changes, etc. In addition, by deploying multi-view cameras, more information can be acquired from various views, resulting in more accurate action recognition. Therefore, action recognition in depth video, infrared video and multi-view video has gained much attention. Moreover, CNN has recently been applied to depth video action recognition [6], infrared video action recognition [7] and multi-view video action recognition [8].

Clearly, CNN-based action recognition can be extended from visible to depth and infrared video, from single-view to multi-view video. As the fundamental and essential task, classification, which has been studied extensively, widely and thoroughly, is embedded within the detection task. Therefore, this review focuses on the action classification task in visible single-view video.

In this paper, we provide a review of CNN-based action recognition, which is organized as follows. Section 2 introduces previous review works. In Section 3, an overview of action recognition is presented from two aspects: handcrafted and deep learning representation methods. Section 4 provides a comprehensive review of CNN-based action recognition according to the solutions for exploiting temporal information. Section 5 introduces the recently released large-scale action recognition datasets and discusses the performance of CNN-based action recognition on these

* Corresponding author.
E-mail address: taoleiyao@ioe.ac.cn (T. Lei).

https://doi.org/10.1016/j.patrec.2018.05.018
0167-8655/

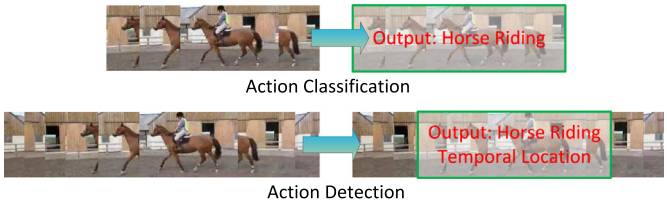

Fig. 1. Two kinds of tasks in action recognition: classification and detection.

benchmarks. Finally, the limitations and future research directions of CNN-based action recognition are concluded in Section 6.

2. Previous review works

Currently, numerous reviews and surveys on action recognition have been studied in the literature. The earlier works [9–12] obviously do not include a review of the recent action recognition approaches. Reviews [13,14] provided reviews of datasets and benchmarks designed for action recognition. In this section, we introduce the reviews and surveys related to action recognition published since 2014. Review [15] first identified the various challenges of action recognition. Then, it reviewed the action recognition approaches based on their capability to address these challenges and introduced the publicly available datasets with the identified challenges. By analyzing the feature representation and classifier, review [16] provided a review of action recognition and addressed some open problems of action recognition. Moreover, works [17,18] reviewed local representation methods for action recognition. Review [17] summarized related public datasets. Review [18] conducted an experimental comparison under unified settings on three datasets and thoughtfully discussed the findings from the experimental results and conclusions. Furthermore, works [19–21] reviewed action recognition, covering a wide range, from handcrafted to deep learning representation method. Specifically, review [19] explicitly discussed the superiorities and limitations of the existing methods; review [20] presented the dataset available and the application of action recognition and concluded with important discussions and research directions; and review [21] provided a high level analysis and discussed challenges and possible future directions. Unlike the abovementioned reviews, paper [5] described the THUMOS 2015 action recognition challenge in detail, including the benchmark, evaluation protocols for classification and detection, the results of submissions to this challenge, and the participating approaches. It also provided a study on how well the classification methods could be generalized for detection and proposed several directions and improvements for future THUMOS challenges.

CNN has attracted the most attention from the computer science community and CNN-based method has recently dominated the research on action recognition and has exhibited excellent performance. However, only a few reviews [19–21] regarded CNN-based method as a part of the deep learning representation action recognition. Therefore, this paper focuses on CNN-based action recognition.

3. Overview of action recognition

Typical action recognition is implemented in two steps: action representation and classification. Action representation, which is also referred to as feature extraction, is considered as the core of video action recognition. Effective action representation should be as follows: 1) Discriminative: The representations of actions from the same class carry similar information, while the representations of actions from different classes carry distinguishing information.

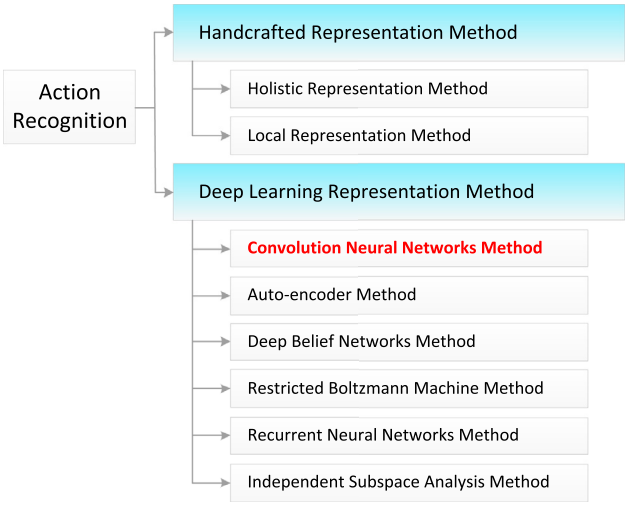

Fig. 2. The taxonomy of action recognition.

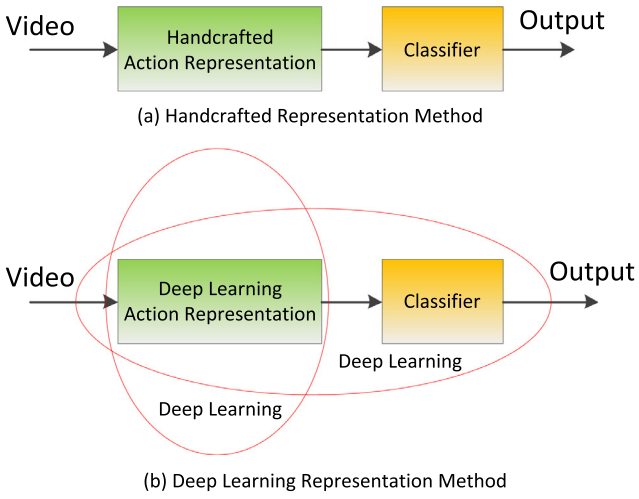

Fig. 3. Frameworks of handcrafted and deep learning representation action recognition.

2) Efficient: The action representation is easy to compute. 3) Low dimensional: This implies a low-cost of classification and feature saving.

The well-formed taxonomy of action recognition presented in Fig. 2 is based on the action representation. It divides action recognition into handcrafted representation method and deep learning representation method, which extract expert designed features and trainable features from video respectively. The frameworks of these two methods are presented in Fig. 3.

The research on action recognition begins in the 1980s with the study of the holistic representation method, which extracts global features, such as silhouette-based feature [22] and optical flow-based feature [23]. However, this method is limited to camera motion and requires pre-processing, such as background subtraction, foreground extraction, locating and tracking. Since 2003, the local representation method has been shown to be effective for action recognition, which extracts local features directly from the interest points in video and therefore avoids the pre-processing. Handcrafted representation method presented a milestone in action recognition. The relevant works on improved dense trajectories (iDT) [24,25], which encode extracted dense trajectories, trajectory-aligned HOG [26], HOF [27] and MBH [28] with the Fishier vector (FV) [29] or hybrid supervector (HSV) [25] deliver state-of-the-art

performances in terms of handcrafted representation action recognition and are often used as the baseline to evaluate new deep learning representation methods.

Different from the handcrafted representation method, in which the feature is designed manually, the deep learning representation method learns the trainable feature automatically from video. Fig. 2 presents the categories of the deep learning action recognition method based on the taxonomy of deep learning. The CNN-based method is the most studied method in various fields of computer vision, including action recognition, and has achieved great success. This review focuses on the CNN-based method, with further details presented in Section 4. In addition to CNN, other deep learning methods have also been applied to action recognition and have achieved good performances, although they are not very popular. Here, we give examples from each category. Taylor and Hinton [30] applied the restricted Boltzmann machine (RBM) [31] to capture the diverse motion-based features of action. Baccouche et al. [32] proposed an auto-encoder model that learns sparse over-completed spatiotemporal features automatically. Chen et al. [33] learned the invariant spatiotemporal features from video using a deep belief network (DBN) [34] model. To explore various convolutional temporal feature pooling architectures and model video as an ordered sequence of frames, Veeriah et al. [35] and Baccouche et al. [36] employed the recurrent convolutional network (RNN) known as long short-term memory [37] (LSTM) to learn the complex dynamics of action. In addition, Le et al. [38] and Pei et al. [39] presented an extension of the independent subspace analysis (ISA) network [40] algorithm to learn invariant spatiotemporal features from video action.

4. CNN-based action recognition

The Convolutional Neural Network (CNN) is a biologically inspired variant of multilayer perceptions that is a feed forward artificial neural network. It consists of an input layer, an output layer, and multiple hidden layers. The hidden layers are convolutional, pooling or fully connected. Convolutional layer applies a convolution operation and an additive bias to the input data and passes the result first through an activation function and then to the next layer. The convolution operation at position (x, y) in the jth feature map in the ith layer is formalized as Eq. (1).

$$v_{il}^{xy} = \varphi\left(b_{i,j} + \sum_{m}\sum_{p=0}^{P_i-1}\sum_{q=0}^{Q_i-1} w_{i,j,m}^{p,q} v_{(i-1),m}^{(x+p),(y+q)}\right) \tag{1}$$

where φ is a non-linear (e.g., Tanh, Sigmod or ReLU) activation function, w is the weight matrix, and P and Q are the height and width of the kernel, respectively. The pooling layer is a form of non-linear down-sampling. After the convolutional and pooling layers, the high-level reasoning in the CNN is carried out via fully connected layers, in which each neuron is connected to all activations in the previous layer.

In 1998, a CNN architecture named LeNet-5 [1] was designed to recognize digits in documents. However, the CNN developed slowly until its breakthrough in image classification [2] in 2012. In recent years, the CNN has experienced a significant advancement in image classification and object detection. Various CNN architectures, such as ZFNet [41], VGG [42], GoogLeNet [43], BN-Inception [44] and ResNets [45] have been designed, with the pre-trained models (weights) of these CNN architectures being obtained via pre-training on large-scale datasets. For a new small-scale dataset or a new modality, an additional training (transfer learning) is performed to fine-tune the pre-trained model of the network. Inspired by the success of CNN in the image domain, researchers have also employed CNN for video action recognition. An increasing number of CNN-based action recognition approaches have emerged with the demonstrated good performance.

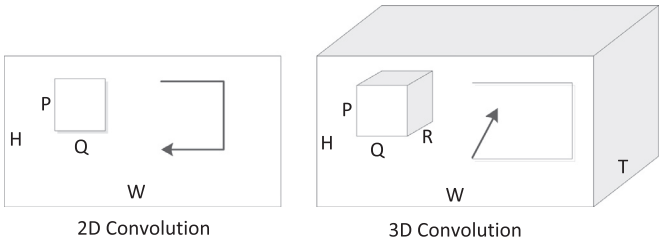

Fig. 4. Comparison of 2D convolution and 3D convolution.

As illustrated by Eq. (1), CNN is primarily applied in 2D space and has been proven to be powerful in terms of extracting spatial information from still images. Some action recognition works [46–50] adopted CNN to extract spatial information, which is considered as the auxiliary cue and combined with handcrafted feature iDT for the final action recognition. However, video is naturally viewed as 3D spatiotemporal signal, while the CNN was primarily designed to extract spatial features, and is thus not suitable for videos due to a lack of temporal information. Therefore, the core issue of extending the CNN from image to video is the exploitation of temporal information. We divide the solutions for exploiting of temporal information into three strategies: 1) 3D convolution; 2) taking motion-related information as the CNN input; and 3) fusion. In fact, these three strategies overlap with each other, e.g. a 3D CNN architecture with motion-related information as its input. In this section, we will review CNN-based action recognition according to these strategies.

4.1. 3D convolution

The straightforward solution to exploit spatiotemporal information is to perform 3D convolution on video which was validated in some pioneering CNN-based action recognition works [51–54] prior to 2012. 3D convolution is achieved by convolving a 3D kernel to a video clip. The operation at position (x, y, z) in the jth feature map in the ith layer is formalized as Eq. (2).

$$v_{il}^{xy} = \varphi\left(b_{i,j} + \sum_{m}\sum_{p=0}^{P_i-1}\sum_{q=0}^{Q_i-1}\sum_{r=0}^{R_i-1} w_{i,j,m}^{p,q,r} v_{(i-1),m}^{(x+p),(y+q),(z+r)}\right) \tag{2}$$

where φ is a non-linear (e.g., Tanh, Sigmod or ReLU) activation function, w is the 3D weight matrix, P, Q and R are the height, width and temporal length of the kernel, respectively. We illustrate the 2D convolution and 3D convolution in Fig. 4.

As one of the important pioneering works prior to 2012, Ji et al. [54] designed a 3D CNN architecture for action recognition, which consists of 1 hardwired layer, 3 convolutional layers, 2 subsampling layers and 1 fully connect layer. The hardwired layer generates the channels of gray, gradient and optical flow. Next, the convolution and subsampling are performed in each channel. The final action representation is then computed by combining the information from all channels. Recently, Ji et al. [55] improved their 3D CNN work [54] by regularizing the outputs with high-level features and combining the predictions of a variety of 3D CNN with different architectures. Meanwhile, Tran et al. [56] conducted a systematic study to find the best temporal kernel length for 3D CNN and developed a VGG-style 3D CNN architecture named C3D. As shown in Fig. 5, the C3D architecture consists of 8 convolutional layers with small $3 \times 3 \times 3$ convolutional kernels, five pooling layers and two fully connected layers. The extracted C3D features were demonstrated to be generic, efficient and compact. Furthermore, Tran et al. [57] conducted a 3D CNN search in a deep residual learning framework and developed a ResNet18-style 3D CNN

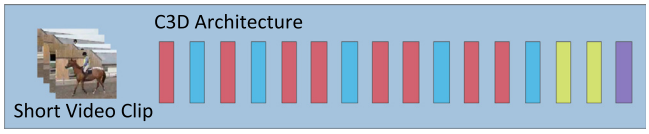

Fig. 5. C3D architecture. Red, green, blue, yellow and purple boxes indicate convolutional, normalization, pooling, fully-connected and softmax layers, respectively. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

architecture named Res3D, which outperforms C3D by a good margin in terms of recognition accuracy. In addition, Res3D is 2 times faster in run-time, 2 times smaller in mode size and more compact than C3D. By pre-training on the largest action recognition benchmark Sports-1M, the C3D and Res3D works both provided their pre-trained model, which can be used either as the initialization in transfer learning or as a fixed spatiotemporal feature extractor. Different from the abovementioned works [55–57], which learn spatiotemporal information from short video clips with 7, 16 and 8 frames, respectively, Varol et al. [58] developed a long-term temporal convolutions (LTC) architecture to represent action at full temporal scale by imposing 3D CNN on longer video clips, such as those with lengths of 60 or 100 frames, and demonstrated the importance of high-quality optical flow as the input of LTC.

Training 3D CNN is very expensive in terms of computation and memory. Creating or simulating 3D CNN using 2D CNN is an alternative implementation of 3D CNN. To reduce the number of network parameters and mitigate the compound difficulty of high kernel complexity and insufficient training video data, Sun et al. [59] proposed a factorized spatiotemporal convolutional network (FstCN), which factorizes the original 3D convolutional kernel learning as a sequential process of learning 2D spatial kernels in the lower layers, followed by learning 1D temporal kernels in the upper layers. They also proposed an effective training and inference strategy based on sampling multiple video clips from a given action video sequence for the issue of sequence alignment. To avoid training 3D CNN from scratch and instead use the knowledge learned by 2D CNN, Mansimov et al. [60] proposed several ways to initialize 3D convolutional weights using 2D convolutional weights, including averaging, scaling, zero weight and negative weight initializations. The experimental results showed that negative weight initialization yields the best performance among all these initializations. The 3D convolutional weight matrix $w^{p,\,q,\,r}$ with temporal dimension T created using 2D convolutional weight matrix $w^{p,\,q}$ by negative weight initialization is expressed as Eq. (3).

$$w^{p,q,r} = \alpha_t w^{p,q}, \text{where}\, \alpha_t = \begin{cases} \frac{2T-1}{T}, \text{if t} = 1 \\ -\frac{1}{T}, \text{otherwise} \end{cases} \tag{3}$$

By decoupling a $3 \times 3 \times 3$ 3D convolutional filter into a $1 \times 3 \times 3$ convolutional filter (equivalent to 2D CNN) and a $3 \times 1 \times 1$ convolutional filter (similar to 1D CNN), Qiu et al. [61] designed three variants of bottleneck building blocks in residual network, as depicted in Fig. 6, and proposed a Pseudo-3D Residual50 architecture named P3D ResNet, in which residual units are replaced with the variants of building blocks. To pursue structural diversity in the design of very deep networks [62], these three variants are mixed in interleaving order. Compared with the original 3D CNN, this Pseudo-3D CNN not only reduces the model size significantly but also enables the pre-training of 2D CNN on an image dataset.

4.2. Taking motion-related information as the CNN input

To exploit temporal information, some works have attempted to take motion-related information such as the optical flow and motion vector as the input of the CNN. The two-stream model

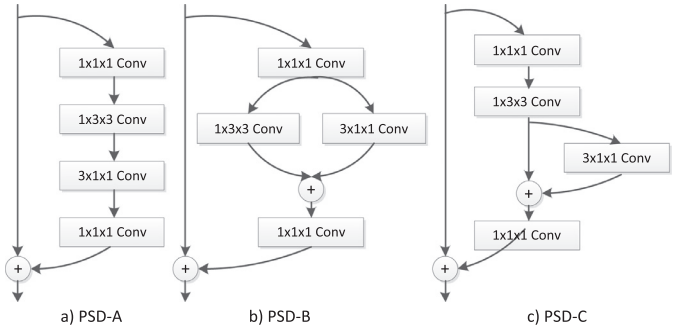

Fig. 6. Bottleneck building blocks of residual unit in pseudo-3D.

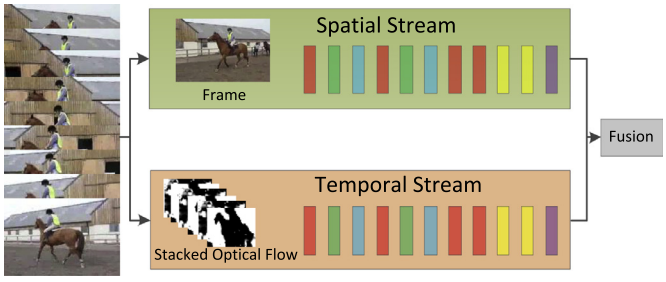

Fig. 7. Two-stream model for action recognition. Red, green, blue, yellow and purple boxes indicate convolutional, normalization, pooling, fully-connected and softmax layers, respectively. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

[63] shown in Fig. 7, which takes the optical flow as the input of the CNN to extract motion information, is a popular and important method for action recognition. According to the visual pathway hypothesis [64], in which the ventral stream plays the major role in the perceptual identification of objects and the dorsal stream mediates the required sensorimotor transformations for visually guided actions directed at such objects, Simonyan and Zisserman [63] proposed a two-stream model for action recognition, in which the spatial stream performs action recognition from still frames, while the temporal stream is trained to recognize action from the motion information. Each stream in this model is implemented with a CNN architecture, and the softmax scores of the two streams are combined via a late fusion (averaging fusion or SVM fusion) to obtain the final result. They adopted the CNN-M-2048 [65] architecture for each stream. Since the CNN for the spatial stream is essentially an image classification, they pre-trained the spatial CNN on the ImageNet dataset [66], a large-scale image dataset, and fine-tuned the spatial CNN on the evaluation datasets. Unlike the spatial CNN, which can be pre-trained on a large image classification dataset, the temporal CNN needs to be trained on evaluation datasets, which are rather small. To avoid over-fitting, multi-task learning was employed for temporal CNN training. However, the improvement offered by the two-stream model [63] in action recognition is not so evident. By comparing the two-stream model [63] with object recognition in still images, Wang et al. [67] argued that there are two reasons for this result. One is that the adopted CNN architecture is shallow. The other is that the scale of the training dataset is small. To address these issues, they adopted the GoogLeNet [43] and VGG-16 [42] deep CNN architectures to design a very deep two-stream model and proposed several good practices for training, including pre-training for both streams, smaller learning rates, more data augmentation and high dropout ratio. This very deep two-stream model [67] performed better than the original two-stream model [63]. To best take advantage of the spatiotemporal information in the two-stream model, Feichtenhofer et al. [68] studied a number of ways

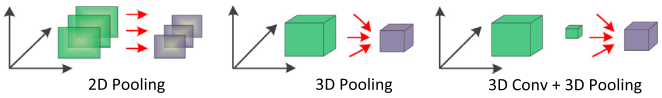

Fig. 8. Comparison of 3D Conv + 3D pooling with other kinds of pooling. 3D Conv + 3D pooling performs a convolution with a fusion kernel that spans the feature channels, space and time before 3D pooling.

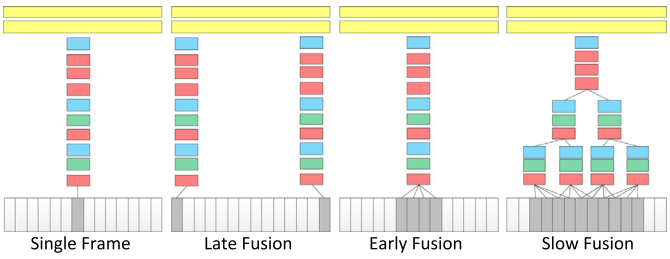

Fig. 9. Temporal information exploitation by fusing the spatial information. Red, green, blue and yellow boxes indicate convolutional, normalization, pooling and fully connected layers, respectively. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

to fuse two streams spatially and temporally and proposed an improved two-stream model with a novel convolutional fusion layer between two streams and a novel temporal fusion layer that incorporates 3D Conv + 3D Pooling. The difference between this pooling and other kinds of pooling is presented in Fig. 8. Due to the recent success of ResNets [45] for extremely deep network training, Feichtenhofer et al. [69] also introduced a spatiotemporal ResNets as a combination of ResNets and the two-stream model, which allows the hierarchical learning of spatiotemporal features by injecting residual connections between the spatial and temporal streams. Furthermore, [69] transferred both streams from the spatial to the spatiotemporal domain by transforming the dimensionality mapping filters of a pre-trained model into temporal convolutions, initialized as residual filters over time. In the two-stream model, the most computationally expensive step comes from the calculation of the optical flow. Zhang et al. [70] accelerated this model by replacing the optical flow with the motion vector, which can be obtained directly from compressed videos without extra calculation. However, the motion vector lacks fine structures, and contains noisy and inaccurate motion patterns. To resolve this problem, they proposed to transfer the knowledge learned from the optical flow CNN to the motion vector CNN.

4.3. Fusion

Fusion is also used to exploit temporal information. Various approaches have been proposed to fuse the information in the temporal domain or conduct pooling or aggregating the frame-level information and short-clip-level information into the video-level information. Karpathy et al. [71] investigated several connectivity patterns to fuse spatial information temporally: single-frame, early fusion, late fusion and slow fusion. All these kinds of fusion are shown in Fig. 9. Their experimental results showed that slow fusion performs better than early fusion and late fusion but achieves only a similar level of performance with respect to that of the single-frame model, which is a purely spatial network. These results indicate that these fusions in work [71] cannot capture sufficient motion information. In addition, Gao et al. [72] introduced a kind of feature alignment to generate a compact video representation that exploits temporal correlations in the spatial features, and Yu et al. [73] attempted to fuse the CNN frame-wise spatial feature to video-wise spatiotemporal feature via stratified pooling.

Moreover, fusion is a general concept in action recognition. It is used to not only exploit temporal information from spatial features but also exploit information by fusing, pooling or aggregating various kinds of extracted information, including spatial, temporal and spatiotemporal information for the final video representation. Examples include the aforementioned spatiotemporal fusion and temporal fusion in the improved two-stream model [68] and the residual connections between the spatial and temporal streams in the ResNet two-stream model [69]. Furthermore, Feichtenhofer et al. [74] combined the spatial and temporal streams in the ResNet two-stream model via motion gating and injected identity mapping kernels as temporal filters to learn long-term temporal information. To find a proper temporal feature pooling for CNN in action recognition, Ng et al. [75] investigated various ways of temporal pooling, which are performed on frame-wise spatial and temporal features extracted by the two-stream model, and experimentally selected temporal max-pooling as the main feature aggregation technique for action recognition. In contrast to temporal pooling, which produces order-invariant representation for action, they also proposed to represent an action as an ordered sequence of the frame by imposing LSTM on the spatial and temporal features, as shown in Fig. 10. Because LSTM uses memory cells to store, modify and access the internal state, it can discover long-term temporal information from the extracted information.

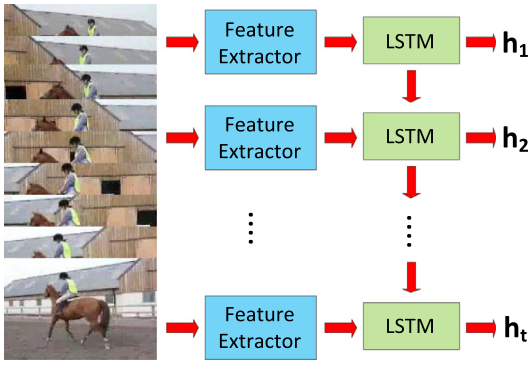

Fig. 10. Long-term temporal information exploited using LSTM.

Table 1
Recognition results on sports-1M. Hit@5 values indicate the fraction of test samples that contained at least one of the ground-truth labels in the top 5 predictions.

Method	Hit@5 (%)	CNN architecture	Input of the CNN
SlowFusion [71]	80.2	AlexNet	RGB
LSTM/ConvPooling [75]	90.4/90.8	GoogLeNet	RGB + OF
HumanSkeleton [87]	81.7	GoogLeNet	RGB
C3D [56]	85.2	C3D	RGB
Res3D [57]	87.8	Res3D	RGB
P3D-ResNet [61]	87.4	P3D-ResNet	RGB

In addition to [75], other works [76–78] exist that imposed LSTM on the spatial and temporal features obtained via the two-stream model to learn long-term temporal information. Meanwhile, Srivastava et al. [79] employed LSTM on the features obtained via the two-stream model in an encoder-decoder framework, and Yao et al. [80] proposed to use 3D CNN features and an LSTM decoder in an encoder-decoder framework.

4.4. Summary

In this section, we reviewed the CNN-based action recognition works according to the solutions for exploiting temporal information. The 3D CNN and two-stream methods are important and essential for learning spatiotemporal information. Fusion in CNN-based action recognition is a more general concept that is used to exploit spatiotemporal information by fusing, pooling, or aggregating various kinds of extracted information.

Table 2
Recognition results on UCF101 and HMDB51 datasets. The mAP is the mean accuracy across three splits. The state-of-the-art CNN-based approaches are highlighted. In the brackets, the recognition result when only taking the RGB as the input is shown.

Method	UCF101 mAP(%)	HMDB51 mAP(%)	CNN Architecture	Input of the CNN
Handcrafted representation action recognition				
iDT + FV [24]	85.9	57.2	–	–
iDT + HSV [25]	87.9	61.1	–	–
CNN-based action recognition (2D CNN)				
SlowFusion [71]	65.4	–	AlexNet	RGB
TDD [88]	90.3	63.2	ZFNet	RGB + OF
LRCN [77]	82.3 (68.2)	–	ZFNet	RGB + OF (RGB)
Two-stream SVM [63]	88.0 (73.0)	59.4 (40.5)	CNN-M	RGB + OF (RGB)
Improved Two-stream [68]	92.5	65.4	VGG-16	RGB + OF
HRP [89]	91.4	66.9	VGG-16	RGB
Transformation [90]	92.4 (80.8)	62.0 (44.1)	VGG-16	RGB + OF (RGB)
Temporal Scale-Invariant [91]	93.7	69.5	VGG-16	RGB + OF
ActionVLAD [92]	92.7	66.9 (49.8)	VGG-16	RGB + OF (RGB)
AdaScan [93]	89.4 (78.6)	54.9 (41.1)	VGG-16	RGB + OF (RGB)
GRP [94]	91.9	65.4	VGG-16	RGB + OF
Multi-Resource [95]	89.1	54.9	VGG-19	RGB + OF
LSTM/ConvPooling [75]	88.6/88.2	–	GoogLeNet	RGB + OF
HumanSkeleton [87]	86.9	55.3	GoogLeNet	RGB
KeyVolume [96]	93.1	63.3	GoogLeNet	RGB + OF
TSN [97]	**94.2**	**69.4**	BN-Inception	RGB + OF
SSN [98]	**94.8**	**73.8**	BN-Inception	RGB + OF
Pyramid Two-stream [99]	**94.6**	**68.9**	BN-Inception	RGB + OF
Trajectory Pooling [100]	92.1	65.6	VGG-16, CNN-M	RGB + OF
Hybrid Deep Framework [101]	91.3	–	VGG-19, CNN-M	RGB + OF
Residual Two-stream [69]	**93.4 (82.3)**	**66.4 (43.4)**	ResNet-50	RGB + OF (RGB)
Multiplier Two-stream [74]	**94.2**	**68.9**	ResNet-50	RGB + OF
CNN-based action recognition (3D CNN)				
C3D [56]	85.2	–	C3D	RGB
VLAD3 [102]	90.5	–	C3D	RGB
Spatiotemporal LSTM [103]	85.4 (83.0)	55.2 (51.2)	C3D	RGB + OF
LTC [58]	91.7 (82.4)	64.8	LTC	RGB + OF (RGB)
Res3D [57]	85.8	54.9	Res3D	RGB
P3D-ResNet [61]	88.6	–	P3D-ResNet	RGB
CNN-based action recognition (2D CNN + 3D CNN)				
Multi-Granular [104]	90.8	63.6	VGG-19, C3D	RGB + OF
ST-VLMPF(DF) [105]	**93.6**	**69.5**	VGG-16, VGG-19, C3D	RGB + OF

Additionally, the spatiotemporal information extracted by CNN falls into one of four categories from the view of the temporal scale: (from small to large) apparent (spatial) information, motion information, short-term temporal information and long-term temporal information. Apparent, motion and short-term temporal information are modeled by CNN on the frame, CNN on the optical flow and 3DCNN on the short clip, respectively. The long-term deep temporal information is modeled in various ways. As aforementioned, it can be modeled by imposing LSTM on the extracted deep features [75–78], employing a long-term temporal 3D convolutions for the long clip [58], or injecting identity mapping kernels as temporal filters [74].

5. Performance

In this section, we introduce the recently released large-scale action recognition datasets. For comparison, we illustrate the action recognition performance by presenting the recognition accuracy on these datasets.

5.1. Recent large-scale benchmarks

With the improvement of action recognition, different kinds of benchmarks and datasets have been released. Hassner et al. [14] provided a review of action recognition datasets developed over the years: early datasets in the lab, e.g., Weizmann [81]; interim datasets in the TV, e.g., UCF Sports [82,83]; and recent datasets in the wild, e.g., UCF101 [84]. In this section, we introduce several recent 'wild' large-scale action recognition datasets.

HMDB51 [85]: The videos in HMDB51 were collected from various internet sources and digitized movies, and the human actions in this dataset are representative of daily actions. Some of the key challenges in this dataset are large variations in camera viewpoint and motion, cluttered background, and changes in the position, scale, and appearances of the actors. HMDB51 contains 51 distinct action categories, each containing at least 101 clips for a total of 6766 video clips. The action categories can be grouped into five types: 1) general facial actions; 2) facial actions with object manipulation; 3) general body movements; 4) body movements with object interaction; and 5) body movements for human interaction. For each action category, the video clips are split into training set with 70 clips and testing set with 30 clips, which fulfills the 70/30 balance, and the clips in the training and testing sets cannot come from the same video file. This dataset was annotated with action category labels, video quality and meta information, including visible body, lower body or full body, camera motion and the number of people involved in the action.

UCF101 [84]: UCF101, an extension of the UCF50 [86] dataset, was collected from YouTube and has 13,320 videos from 101 action categories. UCF101 offers the largest diversity in terms of actions and with the presence of large variations in camera motion, object appearance and pose, object scale, viewpoint, cluttered background, illumination conditions, etc. The videos in each action category are grouped into 25 groups, with each group consisting of 4–7 videos of an action. The action categories can be divided into five types: 1) human-object interaction, 2) body-motion only, 3) human-human interaction, 4) playing musical instruments, and 5) sports. Three training/testing splits are recommended; each test

split contains seven different groups, with the respective remaining 18 groups being used for training.

Sports-1M [71]: To obtain a sufficient amount of data to train CNN architectures in the video domain, the Sports-1M dataset, which consists of 1 million YouTube videos annotated with 487 classes, was developed. There are 1000–3000 videos per class and approximately 5% of the videos are annotated with more than one class. The dataset is split by assigning 70% of the videos to the training set, 10% to a validation set and 20% to a test set. As with the 2D CNN model pre-trained on the ImageNet [66], 3D CNN pre-trained models, such as the C3D and Res3D pre-trained models, are obtained by pre-training on the Sports-1M.

5.2. Recognition results

Due to the large scale, the action recognition works evaluated on the Sports-1M dataset are few. We present the results of the CNN-based works on Sports-1M in Table 1. The HMDB51 and UCF101 datasets are the most popular benchmarks used to evaluate the recent action recognition methods. We provide a comprehensive list of the results of those CNN-based works that are influential or obtain remarkable results on these two datasets in Table 2. For each method, we specify the recognition accuracy, CNN architecture and the input of the CNN. To compare with state-of-the-art traditional handcrafted presentation method, we list the results of iDT-related action recognition [24,25]. The accuracies are reported directly from the original works. In addition, we also highlight state-of-the-art CNN-based approaches in red.

A quick look at the results in Table 2 reveals that action recognition accuracy achieved on UFC101 is over 94%. However, these same methods achieve an accuracy of only approximately 70% on HMDB51, which suffers from large viewpoint variation, cluttered background, and changes in the position, scale and appearances of actors. This reveals that the available action recognition methods are not able to overcome these challenges. On the other hand, we notice that the earlier CNN-based works [56,71] did not performed as well as the iDT-related methods. With the deeper CNN architecture and new technology adopted, the recently developed CNN-based methods resulted in a significant improvement, and outperformed the iDT-related methods.

5.3. Summary

From deep AlexNet and ZFNet to very deep VGG and GoogLeNet then to extremely deep ResNet, and from C3D and LTC to Res3D then to P3D-ResNet, the employed 2D and 3D CNN architectures for CNN-based action recognition became deeper. These results reveal that with this increase in depth of the CNN, the recognition performance was boosted. On the other hand, taking both the RGB and optical fields as the input of the CNN increases the reliability of recognition. By focusing only on the RGB input, we can determine that the most powerful architecture for extracting spatiotemporal information for action recognition is P3D-ResNet, which is reasonable. P3D-Res3D is an extremely deep 3D CNN architecture that outperforms the deep and very deep CNN architectures as well as the extremely deep 2D CNN architectures.

6. Conclusions

Recently, CNN-based action recognition, which has achieved remarkable performance and outperforms the handcrafted representation method on challenging datasets, has received the extensive attention from the computer science community. The strong capability to extract spatial information from 2D space has been proven in the image domain and video is naturally viewed as 3D spatiotemporal signal. The core issue of imposing CNN in the video domain is the exploiting of temporal information. In this paper, we summarized the solutions for exploiting temporal information and provided a comprehensive review of CNN-based action recognition according to these solutions. We also presented and compared the results of the CNN-based action recognition method on two challenging datasets, i.e., UCF101 and HMDB51.

To conclude this review, the limitations as well as future research directions of CNN-based action recognition are explored.

While CNN-based action recognition has recently achieved outstanding progress, the challenges of action recognition, such as viewpoint variation, occlusion, etc., have still not been overcome. Currently, due to the lack of a systemic and complete theory, it is difficult to explain the meaning of the deep features extracted by the CNN. Only some visualization techniques [41,106] give insight into the function of intermediate feature layers and the operation of the classifier. On the other hand, the videos in the existing datasets are not identified with the specific challenges of action recognition. Only the overall performance can be evaluated; the robustness of an action recognition method against a specific challenge still cannot be assessed. Essentially, it is difficult to understand the capability of CNN-based method to address specific challenges from theory or evaluation. Some negative natures of the CNN should also be addressed here. CNN is both computation and memory intensive. Hence, CNN is difficult to be deployed on embedded systems with limited hardware resources. In addition, the CNN-based method relies on large amount of data; yet, many realistic scenarios lack sufficient data for training, even though some large-scale datasets have been developed to make fine-tuning of the CNN architecture possible.

From the point view of improving action recognition, we anticipate several research directions for the CNN-based method in the future. 3D CNN are more suitable for spatiotemporal feature learning compared with 2D CNN. The emerging powerful 2D CNN architectures should be migrated to 3D CNN. CNN-based action recognition has benefited from the input of optical flow, a kind of popular motion-related information. However, the optical flow incurs a high computational cost. Taking new efficient motion-related information as the CNN input should be investigated. As a general concept, fusion will always be a keyword in the CNN-based action recognition, which is conducted to exploit spatiotemporal information by fusing, pooling, or aggregating various kinds of extracted information. The exploration of systemic theory is a big challenge. Much effort must be expended to fill the gap between performance and theory. In addition, efforts must be made to reduce the resource requirement of the CNN and to adapt the CNN-based method for small datasets. All these anticipated research directions converge toward the solutions for spatiotemporal information exploiting and the alleviation of the abovementioned limitations.

Conflicts of Interest

The authors declare no conflict of interest.

Acknowledgment

This work was supported by Youth Innovation Promotion Association, Chinese Academy of Sciences (grant no. 2016336).

References

[1] Y. Lecun, L. Bottou, Y. Bengio, P. Haffner, Gradient-based learning applied to document recognition, Proc. IEEE (1998) 2278–2324.
[2] A. Krizhevsky, I. Sutskever, G. Hinton, Imagenet classification with deep convolutional neural networks, in: Proceedings of the Annual Conference on Neural Information Processing Systems (NIPS), 2012, pp. 1097–1105.
[3] 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 (CVPR), 2014, pp. 580–587.

[4] C. Farabet, C. Couprie, Y. LeCun, Learning hierarchical features for scene labeling, IEEE Trans. Pattern Anal. Mach. Intell. 35 (8) (2013) 1915–1929.
[5] H. Idrees, A. Zamir, Y. Jiang, A. Gorban, I. Laptev, R. Sukthankar, M. Shah, The THUMOS challenge on action recognition for videos "in the wild", Comput. Vision Image Understanding 155 (2017) 1–23.
[6] R. Yang, R. Yang, DMM-pyramid based deep architectures for action recognition with depth cameras, in: Proceedings of the Asian Conference on Computer Vision (ACCV), 2014, pp. 37–49.
[7] C. Gao, Y. Du, J. Liu, J. Lv, L. Yang, D. Meng, A. Hauptmann, InfAR dataset: infrared action recognition at different times, Neurocomputing 212 (2016) 36–47.
[8] R. Kavi, V. Kulathumani, F. Rohit, V. Kecojevic, Multiview fusion for activity recognition using deep neural networks, J. Electron. Imaging 25 (4) (2016).
[9] V. Krüger, D. Kragic, A. Ude, C. Geib, The meaning of action A review on action recognition and mapping, Adv. Rob. 21 (13) (2007) 1473–1501.
[10] R. Poppe, A survey on vision-based human action recognition, Image Vision Comput. 28 (6) (2010) 976–990.
[11] D. Weinland, R. Ronfard, E. Boyer, A survey of vision-based methods for action representation, segmentation and recognition, Comput. Vision Image Understanding 115 (2) (2011) 224–241.
[12] P. Turaga, R. Chellappa, V. Subrahmanian, O. Udrea, Machine recognition of human activities: a survey, IEEE Trans. Circuits Syst. Video Technol. 18 (11) (2008) 1473–1488.
[13] J. Chaquet, E. Carmona, A. Fernández-Caballero, A survey of video datasets for human action and activity recognition, Comput. Vision Image Understanding 117 (6) (2013) 633–659.
[14] T. Hassner, A critical review of action recognition benchmarks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 245–250.
[15] M. Ramanathan, W. Yau, E. Teoh, Human action recognition with video data: research and evaluation challenges, IEEE Trans. Hum.-Mach. Syst. 44 (5) (2014) 650–663.
[16] S. Kang, R. Wildes, Review of action recognition and detection methods. arXiv preprint, arXiv: 1610.06906. 2016.
[17] D. Dawn, S. Shaikh, A comprehensive survey of human action recognition with spatio-temporal interest point (STIP) detector, Vis. Comput. 32 (3) (2016) 289–306.
[18] X. Zhen, L. Shao, Action recognition via spatio-temporal local features: a comprehensive study, Image Vision Comput. 50 (2016) 1–13.
[19] F. Zhu, L. Shao, J. Xie, Y. Fang, From handcrafted to learned representations for human action recognition, Image Vision Comput. 55 (2016) 42–52.
[20] A. Sargano, P. Angelov, Z. Habib, A comprehensive review on handcrafted and learning-Based action representation approaches for human activity recognition, Appl. Sci. 7 (110) (2017).
[21] S. Herath, M. Harandi, F. Porikli, Going deeper into action recognition: a survey, Image Vision Comput. 60 (2017) 4–21.
[22] Z. Lin, Z. Jiang, L.S. Davis, Recognizing actions by shape-motion prototype trees, in: Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2009, pp. 444–451.
[23] A. Efros, A. Berg, G. Mori, J. Malik, Recognizing action at a distance, in: Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2003, pp. 726–733.
[24] H. Wang, C. Schmid, Action recognition with improved trajectories, in: Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 3551–3558.
[25] X. Peng, L. Wang, X. Wang, Y. Qiao, Bag of visual words and fusion methods for action recognition, Comput. Vision Image Understanding 150 (2016) 109–125.
[26] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2005, pp. 886–893.
[27] I. Laptev, M. Marszalek, C. Schmid, B. Rozenfeld, Learning realistic human actions from movies, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2008, pp. 1–8.
[28] N. Dalal, B. Triggs, C. Schmid, Human detection using oriented histograms of flow and appearance, in: Proceedings of the European Conference on Computer Vision (ECCV), 2006, pp. 428–441.
[29] F. Perronnin, C. Dance, Fisher kernels on visual vocabularies for image categorization, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2007, pp. 1–8.
[30] W. Taylor, G. Hinton, Factored conditional restricted Boltzmann machines for modeling motion style, in: Proceedings of the International Conference on Machine Learning (ICML), 2009, pp. 1025–1032.
[31] P. Smolensky, Information Processing in Dynamical Systems: Foundations of Harmony Theory, in: Parallel Distributed Processing: Explorations in the Microstructure of Cognition, MIT Press, Cambridge, MA, USA, 1986, pp. 194–281.
[32] M. Baccouche, F. Mamalet, C. Wolf, C. Garcia, A. Baskurt, Spatio-temporal convolutional sparse auto-encoder for sequence classification, in: Proceedings of the British Machine Vision Conference (BMVC), 2012.
[33] B. Chen, J. Ting, B. Marlin, N. Freitas, Deep learning of invariant spatio-temporal features from video, in: Proceedings of the Annual Conference on Neural Information Processing Systems (NIPS), 2010.
[34] G. Hinton, S. Osindero, Y. Teh, A fast learning algorithm for deep belief nets, Neural Comput. 18 (7) (2006) 1527–1554.
[35] V. Veeriah, N. Zhuang, G. Qi, Differential recurrent neural networks for action recognition, in: Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 4041–4049.
[36] M. Baccouche, F. Mamalet, C. Wolf, C. Garcia, A. Baskurt, Action classification in soccer videos with long short-term memory recurrent neural networks, in: Proceedings of the International Conference on Artificial Neural Networks, 2010, pp. 154–159.
[37] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Comput. 9 (8) (1997) 1735–1780.
[38] Q. Le, W. Zou, S. Yeung, A. Ng, Learning hierarchy invariant spatio-temporal features for action recognition with independent subspace analysis, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, (CVPR), 2011, pp. 3361–3368.
[39] L. Pei, M. Ye, X. Zhao, Y. Dou, J. Bo, Action recognition by learning temporal slowness invariant features, Visual Comput. Int. J. Comput. Graph. 32 (11) (2015) 1395–1404.
[40] A. Hyvärinen, J. Hurri, P. Hoyer, Natural Image Statistics, Springer, Heidelberg, 2009.
[41] M. Zeiler, R. Fergus, Visualizing and understanding convolutional networks, in: Proceedings of the European Conference on Computer Vision (ECCV), 2014, pp. 818–833.
[42] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, in: Proceedings of the International Conference on Learning Representation (ICLR), 2014.
[43] 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 (CVPR), 2015, pp. 1–9.
[44] S. Ioffe, C. Szegedy, Batch normalization: Accelerating deep network training by reducing internal covariate shift, in: Proceedings of the International Conference on Machine Learning (ICML), 2015, pp. 448–456.
[45] K. He, X. Zhang, S. Ren, Jian Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778.
[46] Jain M, Jan. Gemert, C. Snoek, University of Amsterdam at THUMOS challenge 2014, THUMOS'14 Action Recognition Challenge, 2014.
[47] D. Oneata, J. Verbeek, C. Schmid, The LEAR submission at THUMOS 2014, THUMOS'14 Action Recognition Challenge, 2014.
[48] L. Wang, Y. Qiao, X. Tang, Action recognition and detection by combining motion and appearance features, THUMOS'14 Action Recognition Challenge, 2014.
[49] S. Karaman, L. Seidenari, A. Bimbo, Fast saliency based pooling of Fisher encoded dense trajectories, THUMOS'14 Action Recognition Challenge, 2014.
[50] M. Jain, J. Gemert, C. Snoek, What do 15,000 object categories tell us about classifying and localizing actions? in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 46–55.
[51] H. Kim, J. Lee, H. Yang, Human action recognition using a modified convolutional neural network, in: Proceedings of the International Symposium on Advances in Neural Networks, 2007, pp. 715–723.
[52] M. Yang, S. Ji, W. Xu, J. Wang, Detecting human actions in surveillance videos, in: Proceedings of the TREC Video Retrieval Evaluation Workshop, 2009.
[53] M. Baccouche, F. Mamalet, C. Wolf, C. Carcia, A. Baskurt, Sequential deep learning for human action recognition, in: Proceedings of the International Conference on Human Behavior Unterstanding, 2011, pp. 29–39.
[54] S. Ji, W. Xu, M. Yang, Y. Kai, 3D convolutional neural networks for human action recognition, in: Proceedings of the International Conference on Machine Learning (ICML), 2010, pp. 495–502.
[55] 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.
[56] 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 (ICCV), 2015, pp. 4489–4497.
[57] D. Tran, J. Ray, Z. Shou, S. Chang, M. Paluri, ConvNet architecture search for spatiotemporal feature learning, arXiv:1708.05038. 2017.
[58] G. Varol, I. Laptev, C. Schmid, Long-term temporal convolutions for action recognition, IEEE Trans. Pattern Anal. Mach. Intell. (2017).
[59] L. Sun, K. Jia, D. Yeung, B. Shi, Human action recognition using factorized spatio-temporal convolutional networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 4597–4605.
[60] E. Mansimov, N. Srivastava, R. Salakhutdinov, Initialization strategies of spatio-temporal convolutional neural networks, arXiv preprint, arXiv: 1503.07274. 2015.
[61] Z. Qiu, T. Yao, T. Mei, Learning spatio-temporal representation with pseudo-3D residual networks, in: Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 5534–5542.
[62] X. Zhang, Z. Li, C.C. Loy, D. Lin, Polynet: a pursuit of structural diversity in very deep networks, arXiv:1611.05725. 2016.
[63] K. Simonyan, A. Zisserman, Two-stream convolutional networks for action recognition in videos, in: Proceedings of the Advances in Neural Information Processing Systems (NIPS), 2014, pp. 568–576.
[64] M. Goodale, A. Milner, Separate visual pathways for perception and action, Trends Neurosci. 15 (1) (1992) 20–25.
[65] K. Chatfield, K. Simonyan, A. Vedaldi, A. Zisserman, Return of the devil in the details: delving deep into convolutional nets, in: Proceedings of the British Machine Vision Conference (BMVC), 2014.
[66] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang,

A. Karpathy, A. Khosla, M. Bernstein, A. Berg, F. Li, ImageNet large scale visual recognition challenge, Int. J. Comput. Vision 115 (3) (2015) 211–252.
[67] L. Wang, Y. Xiong, Z. Wang and Y. Qiao, Towards good practices for very deep two-stream convnets, arXiv:1507.02159. 2015.
[68] 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 (CVPR), 2016, pp. 1933–1941.
[69] C. Feichtenhofer, A. Pinz, R. Wildes, Spatiotemporal residual networks for video action recognition, in: Proceedings of the Conference on Neural Information Processing Systems (NIPS), 2016.
[70] B. Zhang, L. Wang, Z. Wang, Y. Qiao, H. Wang, Real-time action recognition with enhanced motion vector CNNs, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2718–2726.
[71] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, F. Li, Large-scale video classification with convolutional neural networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 1725–1732.
[72] Z. Gao, G. Hua, D. Zhang, N. Jojic, L. Wang, ER3: A unified framework for event retrieval, recognition and recounting, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.
[73] S. Yu, Y. Cheng, S. Su, Stratified pooling based deep convolutional neural networks for human action recognition, Multimedia Tools Appl. 76 (11) (2017) 1–16.
[74] C. Feichtenhofer, A. Pinz, R. Wildes, Spatiotemporal multiplier networks for video action recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.
[75] J. 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 (CVPR), 2015, pp. 4694–4702.
[76] J. Donahue, L. Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, T. Darrell, K. Saenko, Long-term recurrent convolutional networks for visual recognition and description, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 2625–2634.
[77] J. Donahue, L. Hendricks, M. Rohrbach, S. Venugopalan, S. Guadarrama, K. Saenko, T. Darrell, Long-term recurrent convolutional networks for visual recognition and description, IEEE Trans. Pattern Anal. Mach. Intell. 39 (4) (2017) 677–691.
[78] H. Gammulle, S. Denman, S. Sridharan, C. Fookes, Two stream LSTM: a deep fusion framework for human action recognition, in: Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV), 2017, pp. 177–186.
[79] N. Srivastava, E. Mansimov, R. Salakhutdinov, Unsupervised learning of video representations using LSTMs, in: Proceedings of the International Conference on Machine Learning (ICML), 2015.
[80] L. Yao, A. Torabi, K. Cho, N. Ballas, C. Pal, H. Larochelle, A. Courville, Describing videos by exploiting temporal structure, in: Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 4507–4515.
[81] M. Blank, L. Gorelick, E. Shechtman, M. Irani, R. Basri, Actions as space-time shapes, in: Proceedings of the IEEE International Conference on Computer Vision Pattern Recognition (CVPR), 2005, pp. 1395–1402.
[82] M. Rodriguez, J. Ahmed, M. Shah, MACH Action, a spatio-temporal maximum average correlation height filter for action recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2008.
[83] K. Soomro, A. Zamir, Action Recognition in Realistic Sports Videos, Computer Vision in Sports, Springer International Publishing, 2014.
[84] K. Soomro, A. Zamir, M. Shah, UCF101: A Dataset of 101 Human Actions Classes from Videos in the Wild, University of Central Florida, 2012 Technical Report.
[85] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, T. Serre, HMDB: a large video database for human motion recognition, in: Proceedings of the IEEE International Conference on Computer Vision (ICCV),, 2011.
[86] K. Reddy, M. Shah, Recognizing 50 human action categories of web videos, Mach. Vis. Appl. 24 (5) (2013) 971–981.
[87] B. Mahasseni, S. Todorovic, Regularizing long short term memory with 3D human-skeleton sequences for action recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 3054–3062.
[88] Wang L, Y. Qiao, X. Tang, Action recognition with trajectory-pooled deep-convolutional descriptors, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 4305–4314.
[89] B. Fernando, P. Anderson, M. Hutter, S. Gould, Discriminative hierarchical rank pooling for activity recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 1924–1932.
[90] X. Wang, A. Farhadi, A. Gupta, Actions ~ transformations, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2658–2667.
[91] H. Chen, J. Chen, R. Hu, C. Chen, Z. Wang, Action recognition with temporal scale-Invariant deep learning framework, China Commun. 14 (2) (2017) 163–172.
[92] R. Girdhar, D. Ramanan, A. Gupta, J. Sivic, B. Russell, ActionVLAD: learning spatio-temporal aggregation for action classification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.
[93] A. Kar, N. Rai, K. Sikka, G. Sharma, AdaScan adaptive scan pooling in deep convolutional neural networks for human action recognition in Videos, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.
[94] A. Cherian, B. Fernando, M. Harandi, S. Gould, Generalized rank pooling for activity recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.
[95] E. Park, X. Han, T. Berg, A. Berg, Combining multiple sources of knowledge in deep CNNs for action recognition, in: Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV), 2016, pp. 177–186.
[96] 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 (CVPR), 2016, pp. 1991–1999.
[97] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, L. Gool, Temporal segment networks: towards good practices for deep action recognition, in: Proceedings of the European Conference on Computer Vision (ECCV), 2016, pp. 20–36.
[98] Q. Chen, Y, Zhang, Sequential segment networks for action recognition, IEEE Signal Process. Lett. 24 (5) (2017) 712–716.
[99] Y. Wang, M. Long, J. Wang, P. Yu, Spatiotemporal pyramid network for video action recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.
[100] S. Zhao, Y. Liu, Y. Han, R. Hong, Pooling the convolutional layers in deep convnets for video action recognition, IEEE Trans. Circuits Syst. Video Technol. (2017).
[101] Z. Wu, X. Wang, Y. Jiang, H. Ye, X. Xue, Modeling spatial-temporal clues in a hybrid deep learning framework for video classification, in: Proceedings of the ACM Multimedia Conference (ACM MM), 2018.
[102] Y. Li, Li W, V. Mahadevan, N. Vasconcelos, VLAD3: encoding dynamics of deep features for action recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 1951–1960.
[103] Y. Ye, Y. Tian, Embedding sequential information into spatiotemporal features for action recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 1110–1118.
[104] Q. Li, Z. Qiu, T. Yao, T. Mei, Y. Rui, J. Luo, Action recognition by learning deep multi-granular spatio-temporal video representation, in: Proceedings of the ACM International Conference on Multimedia Retrieval, 2016, pp. 159–166.
[105] I. Duta, B. Ionescu, K. Aizawa, N. Sebe, Spatio-temporal vector of locally max pooled features for action recognition in videos, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.
[106] K. Simonyan, A. Vedaldi, A. Zisserman, Deep inside convolutional networks: visualising image classification models and saliency maps. arXiv:1312.6034. 2013.