Multimed Tools Appl
DOI 10.1007/s11042-016-3768-5

Stratified pooling based deep convolutional neural networks for human action recognition

Sheng Yu[1,2,3] **· Yun Cheng**[2] **· Songzhi Su**[1,3] **·**
Guorong Cai[4] **· Shaozi Li**[1,3]

Received: 9 December 2015 / Revised: 3 July 2016 / Accepted: 6 July 2016

Abstract Video based human action recognition is an active and challenging topic in computer vision. Over the last few years, deep convolutional neural networks (CNN) has become the most popular method and achieved the state-of-the-art performance on several datasets, such as HMDB-51 and UCF-101. Since each video has a various number of frame-level features, how to combine these features to acquire good video-level feature becomes a challenging task. Therefore, this paper proposed a novel action recognition method named stratified pooling, which is based on deep convolutional neural networks (SP-CNN). The process is mainly composed of five parts: (i) fine-tuning a pre-trained CNN on the target dataset, (ii) frame-level features extraction; (iii) the principal component analysis (PCA) method for feature dimensionality reduction; (iv) stratified pooling frame-level features to get video-level feature; and (v) SVM for multiclass classification. Finally, the experimental results conducted on HMDB-51 and UCF-101 datasets show that the proposed method outperforms the state-of-the-art.

Keywords Human action recognition · Convolutional neural networks (CNN) · Stratified pooling (SP) · Support vector machines (SVM)

✉ Shaozi Li
szlig@xmu.edu.cn

Sheng Yu
yushengxmu@stu.xmu.edu.cn

[1] Cognitive Science Department, Xiamen University, Xiamen 361005, China

[2] School of Information, Hunan University of Humanities, Science and Technology, Loudi 417000, China

[3] Fujian Key Laboratory of the Brain-like Intelligent Systems, Xiamen 361005, China

[4] Computer Engineering College, Jimei University, Xiamen 361005, China

Published online: 15 July 2016

1 Introduction

Human action recognition has becomes a hot research topic in computer vision. There is a growing number of applications related to human action especially in the area of video surveillance [12, 32, 35]. Meanwhile, it's still very challenging to identify human action recognition due to viewpoint changes, large intra-class variations, varying motion speed, partial occlusion, fast irregular motion, background clutter, etc. [46, 52]. Specifically, compared with object recognition in images, human action recognition is much more difficult because action always contains abundant spatial-temporal information.

Generally, the main pipeline of action recognition can be divided into three parts: feature extraction, feature coding to generate video-level feature descriptor and the descriptor classification. To improve the performance of recognition results, many research efforts have been concentrating on one of these processes, among which visual feature descriptors play the most important role in action recognition. Overall, there are mainly two types of video features for action recognition, including hand-crafted features such as histogram of oriented gradients (HOG) [24], histograms of optical flow (HOF) [24], motion boundary histogram (MBH) [44], The other one is deep-learned features such as stack convolutional independent subspace analysis [25], two-stream convolutional networks [41], 3D convolutional networks [16]. Although hand-crafted features achieved good results on human action recognition, it is worth noting that hand-crafted feature always fails in dealing with large intra-class variations and small inter-class variations.

Recently, deep convolutional neural networks (CNN) have been proposed for large-scale image processing [22, 39, 42]. Note that CNN have obtained state-of-the-art results on object detection, segmentation and recognition [10, 20, 22, 39]. Inspired by these impressive results, some studies attempt to employ CNN for action recognition, typically 3D convolutional neural networks [16], stacked convolutional independent subspace analysis [25], deep convolutional neural networks [20], two-stream convolutional networks [41], trajectory-pooled deep convolutional descriptors (TDD) [46] , etc.

However, there are several problems to directly apply existing deep CNN models to action recognition. Firstly, the structures of video-based CNN and image-based CNN are different, thus the video-based CNN weights must be trained on video dataset from scratch. Secondly, a CNN usually contains tens of millions of parameters, which rely on sufficient amount of training videos to prevent over-fitting. Thirdly, the network needs several weeks to train depending on the architecture. The work in [20] collected a new Sports-1M dataset, which is composed of 1 million videos and includes 498 classes of sports. In prior work [41], a fixed number of frames such as 25 frames are randomly selected and undergo cropping and flipping to reshape them into a compatible input format of an image based CNN. Because the different videos have different frame rate and duration, it is difficult to decide the number of video frames to be sampled. In [47], Wang et al. proposed that sampling frames from the given video may result in missing of some important information and suggested using all video frames as the network input. The algorithm reserves all video frames information, but greatly increased the burden of the subsequent steps.

To tackle the above problems, we first uniformly sample with stride of ten of the given video, and then using sampled frames to fine-tune AlexNet model [22] on the target dataset. Note that fine-tune existing CNN models not only effectively avoid over-fitting when dealing with small dataset, but also reduce the training time. Moreover, the sampling method preserves enough video information for action recognition in the case of the video frame rate less than 30 fps.

As for the performance evaluation, two-stream convolutional networks [41] and temporal pyramid pooling [47] match the best performance of the improved trajectories [45] of the hand-crafted feature, and TDD reach the state-of-the-art performance on HMDB-51 and UCF-101. TDD is based on convolutional activation features for human action recognition. But, convolutional layer activations belong to high-dimensional generic features, which are lacking discriminative capacity. At this point, a number of studies shown that the full-connected layer activations for image classification have much better performance [22, 42] than convolutional feature activations. The main reason is that the full-connected layer descriptor combines the benefits of strong semantics and low dimensionality. From the experimental results, we found that this virtue is also suitable for human action recognition. Therefore, in this paper, we select full-connected layer activations as frame-level features for action recognition.

Motivated by the above analysis, this paper proposed a novel stratified pooling model which transforms frame-level features to video-level feature descriptor. The design of stratified pooling aims to allow an arbitrary number of the frame-level features to be aggregated into fixed-length video-level descriptor, maintain a low computational footprint and effectively prevent over-fitting. Firstly, we uniformly sampled video frames as the input of the CNN to fine-tune AlexNet model [22] on the target dataset. Secondly, the learned model is applied to extract activations from different layers of the CNN architecture, which including full-connected layers and convolutional layers. Thirdly, multiple frame-level activations are aggregated to form a final video-level feature descriptor. Finally, we use SVM with linear kernel [7] for action classification.

The main contributions in this study are summarized as follows: (1) The uniform sampling method preserves enough video information for action recognition. Furthermore, the network does not need huge amount of labeled video dataset during the training process. Then we obtain significant improvement on HMDB51 database, which only contains 6766 video clips. (2) We proposed a novel stratified pooling method that allows an arbitrary number of the frame-level features to be aggregated into fixed-length video-level feature descriptor. (3) We analyze the impact of activations of each layer in the CNN on action recognition, and our novel method has inspiring results on the HMDB-51 and UCF-101 datasets.

The rest of the paper is structured as follows. Section 2, the previous related work is introduced. Section 3 describes the proposed stratified pooling of deep convolutional neural networks activations for action recognition. In Section 4, we demonstrate the efficiency and practicality of the proposed method are conducted on two public datasets. Finally, Section 5 concludes the paper.

2 Related work

Hand-crafted features have obtained great success in recognition task. Typical examples of local feature descriptors include scale-invariant feature transform (SIFT) [31], spatio-temporal interest point (STIP) [23], histograms of optical flow (HOF) [24], histogram of oriented gradients (HOG) [24], histograms of oriented 3D spatio-temporal gradients (HOG3D) [21], speeded up robust features (SURF) [2] and motion boundary histogram (MBH) [44]. As for face recognition, Jian et al. [18] proposed an efficient method based on singular values and potential-field representation for face-image retrieval, in which the rotation-shift-scale invariant properties of the singular values are exploited to design a compact, global feature for face-image representation. In [19], Jian et al. proposed a novel singular value decomposition method for simultaneous hallucination and recognition of low-resolution faces, in which the singular values are first proved to be effective for representing face images.

As for action recognition, Wang et al. [44] proposed dense trajectory features (DTF) extended local features HOF, HOG, MBH and trajectory to align 3D volumes, which obtains much richer low level descriptors for representing the video. HOG descriptor depicts static appearance. HOF descriptor captures the local motion information. MBH descriptor captures the relative dynamic motion information [48]. So, DTF obtained a significant improvement on some challenging datasets which are UCF-101, HMDB-51, etc. Furthermore, in order to deal with camera motion, Wang et al. [45] proposed improved dense trajectory (IDT) which explicitly estimating camera motion to form a feature descriptor which has shown a good result on many action recognition datasets. However, local space-time features are sensitive to noise and then results in the instability of the recognition performance. In [13], the recognition performance significant improves which due to decomposing visual motion into residual and dominant motions. In [34], Peng proposed a new dense sampling strategy to reduce much valid trajectories while preserves the discriminative power. Chen et al. [4] proposed cluster trees model of improved trajectories not only obtain good recognition performance, but also reduce noisy clusters and alleviate intra-class variation. However, these local features were not robust to clutter motions, such as camera motions and background changes were accumulated.

Meanwhile, in order to transform local descriptors into video feature descriptor, there have been a number of feature coding algorithms proposed, such as sparse coding [5, 30], vector of local aggregated descriptors (VLAD) [14], naive bag of words (BoW) [8], fisher vector (FV) [37], improved fisher kernel [38], which achieves the successful of performance.

The hand-crafted features based action recognition achieve good performance, but these features are not optimized for visual representation and lack discriminative capacity when encounter background clutter, large intra-class variations videos for action recognition.

In recent years, deep learning models like deep Boltzmann machines (DBMs) [1, 28], deep belief networks (DBNs) [26, 27], stacked auto-encoders [9], Recurrent neural networks (RNNs) [6, 51] and CNN are used in computer vision applications. For human action recognition, RNNs and CNN have led to impressive performance. RNNs with Long Short-Term Memory (LSTM) units have been shown to perform well in the domain of image description [6, 49], image classification [33], video description [6, 50] and action recognition [6, 51]. Volodymyr et al. [33] proposed a visual attention mechanism based on RNNs model to select a sequence of locations for image classification task. Xu et al. [49] used attention mechanisms to generate image descriptions. Georgia et al. adapt R*CNN to use a primary region and a secondary region for image based action recognition [11, 33]. Inspired by those works on finding visual attention in images, Shikhar et al. [40] use RNNs-LSTM units to expand the visual attention on video based action recognition. The recognition performance is depends on the model pays attention to the action relevant region, but discriminative localization is a challenge problem for video. So, the recognition accuracy on HMDB-51 is only 41.3 %, which is lower than the state-of-the-art models about 30 %.

More recently, some impressive results has been achieved using CNN for human action recognition in videos. Ji et al. [16] trained 3D convolutional operation to extract spatio-temporal features from raw sequence data for action recognition, which may appear of over-fitting phenomenon when without enough samples for training. The performance is worse than hand-crafted representation [45]. Karpathy et al. [20] trained CNN structures on the Sparts-1M dataset. The network consists of thousands of 3D convolutional filters which involves tens of millions of parameters, which is computationally intensive. In [41], Simonyan et al. proposed two-stream convolutional networks which incorporate temporal and spatial net. At testing time, the class scores of the video are gained by average pooling the scores of all selected frames. Wang et al. [46] proposed a trajectory pooled deep convolutional descriptor (TDD) to describe the video features. However, the convolutional layer activations are lack of semantics. The [47] proposed temporal pyramid pooling (TPP) that allows an arbitrary number of frames as the network input. The activations of full-connected layer 7 were encoded and pooled to a fixed-length feature vector. But, the encoding process leads to feature descriptor lose some semantics. Inspired by this, we directly stratified pooling full-connected layer activations as video-level feature vector to preserve more semantic and temporal information of video data. Meanwhile, we verify which layer activations of CNN have the best performance on action recognition.

3 The proposed method

In this section, we introduce our network architecture and the fine-tuning process of CNN on the target dataset, then describe how to use stratified pooling method to generate video-level feature descriptor. The pipeline of deep-learned feature for action recognition is illustrated in Fig. 1 which can be divided into three parts: video sampling, frame-level feature process and stratified pooling for action recognition. In the first part, we uniform sampling frame with constant stride ten of each given video as the input of the CNN. In the second part, we use fine-tuned CNN model to extract frame-level feature and select PCA to reduce feature dimensionality. In the last part, we use stratified pooling converts multiple frame-level to a video-level feature descriptor. The linear kernel SVM [7] has been employed for video classification.

3.1 Network configuration

In this paper, the network is based on AlexNet deep model [22] consisting of five convolutional layers, two full-connected layers and a softmax layer. In the rest of this paper, the notations follow in [3, 10], i.e. conv1 refers to the activation of the first convolutional layer, full6 refers to the activation of the full-connected layer 6. The detail configuration is shown in Table 1. The first line is the number of filters and specifies height and width of each

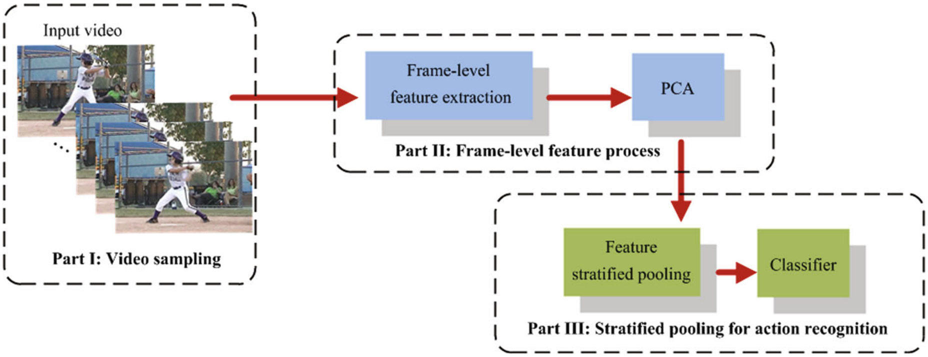

Input video
Part I: Video sampling
Frame-level feature extraction
PCA
Part II: Frame-level feature process
Feature stratified pooling
Classifier
Part III: Stratified pooling for action recognition

Fig. 1 The pipeline of the proposed network

Table 1 The configuration of convolutional deep neural network on HMDB-51 dataset

Conv1	Conv2	Conv3	Conv4	Conv5	Full6	Full7	Softmax
96 × 11 × 11	256 × 5 × 5	348 × 3 × 3	348 × 3 × 3	256 × 3 × 3	4096 × 1	4096 × 1	51 × 1
stride 4	stride 1	stride 1	stride 1	stride 1	dropout 0.5	dropout 0.5	–
pad -	pad 2	pad 1	pad 1	pad 1	–	–	–
LRN	LRN	–	–	–	–	–	–
pool 3 × 3	pool 3 × 3	–	–	pool 3 × 3	–	–	–
stride 2	stride 2	–	–	stride 2	–	–	–

filter. Stride indicates the intervals at which to apply the convolution kernel to the input. Pad indicates the number of pixels to add to each side of the input. Local response normalization abbreviate LRN and the parameters are $n = 5$, $\alpha = 0.0001$, $\beta = 0.75$, respectively.

3.2 Fine-tuning the CNN on the target dataset

Generally, a CNN configuration contains tens of millions of parameters. If the model is trained from scratch on video dataset, it is easy to appear of over-fitting phenomenon and it needs several weeks to train depending on the architecture. So our CNN parameters are initialized from a pre-trained AlexNet model and then fine-tuned using HMDB-51 or UCF-101 as the target dataset. The network parameters are trained the same as [22] by using the mini-batch stochastic gradient descent with momentum is 0.9 and weight decay 5×10^{-4}. The learning rate is set to 10^{-2}, and then decreased by a factor of 10 each after 14k iterations. At training time, the input of network is fixed to 224 × 224.

3.3 Video representation

In this section, we present a novel simple but effective video representation method, called stratified pooling based on deep convolutional neural networks (SP-CNN), which consists of two main parts: frame-level feature extraction and stratified pooling.

3.3.1 Frame-level feature extraction

Given the sampled video frame, we first crop ten 224 × 224 sub-images, which consist of four corner sub-image, one center sub-image and these five sub-images horizontal flipping. Secondly, we use the fine-tuned CNN model to extract some CNN layer feature activations respectively from these ten sub-images. Thirdly, we average pooling these activations as a frame-level feature descriptor. Lastly, PCA is used to reduce feature dimensionality and to de-correlate the feature descriptors. The architecture of frame-level feature extraction is illustrated in Fig. 2. Top-left: Different color dashed boxes represent extracted sub-images. To show it compactly, we only draw the top left corner and bottom right corner dashed box. Top-right: the left video frame undergoes horizontal flipping. The middle box represent the trained CNN model, which is used to extract sub-image activation.

3.3.2 Stratified pooling

It is important to aggregate an arbitrary number of frame-level features to fixed-length video-level feature. We propose a simple but effective three-level stratified pooling

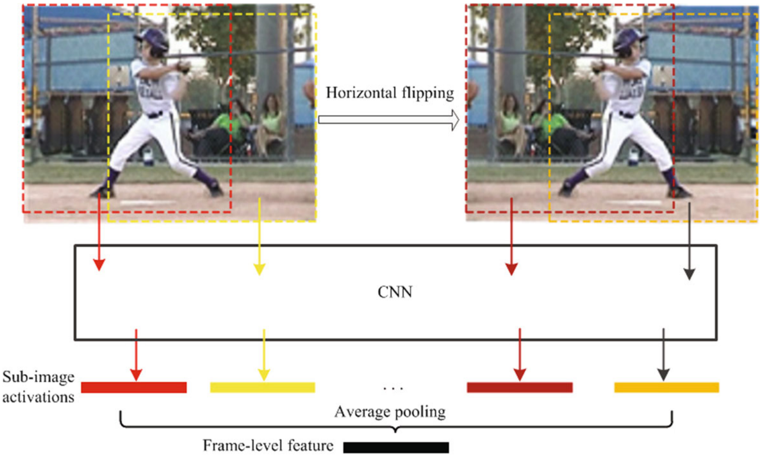


Fig. 2 Frame-level feature extraction

structure. Firstly, we uniform sampling video frames with stride ten from the given video. Then these selected frames are arranged in chronological order and denoted as $X = [x_1, x_2, \cdots, x_n]$, where x_1 is the first frames extracted of the given video and n is the number of frames. The method is illustrated in Fig. 3: (a) the first level covers all sampled frames and all frame-level features are pooled into the first level feature X_1. Such as X_1 represents an average pooling feature of the sample video volume.

$$X_1 = \frac{1}{N}\sum_{i=1}^{N} x_i \tag{1}$$

where x_i is the frame-level feature descriptor with d dimensionality, and N is the number of frames extract from the sub-volume. (b) In level 2, the sampled video volume is evenly divided into 2 sub-volumes. The sub-volume features are formed by pooling sub-volume's all frame-level features, respectively. And then the second level feature is obtained by concatenating the both sub-volume feature, $X_2 = [X_2^1, X_2^2]$. (c) In level 3, the sampled video volume is evenly divided into 4 sub-volumes, and the level 3 feature formed in the same way as level 2, $X_3 = [X_3^1, X_3^2, X_3^3, X_3^4]$. Lastly, these three level features are concatenated into a final video-level feature descriptor, $X_c = [X_1, X_2, X_3]$, with dimensionality $(1+2+4) \times d$.

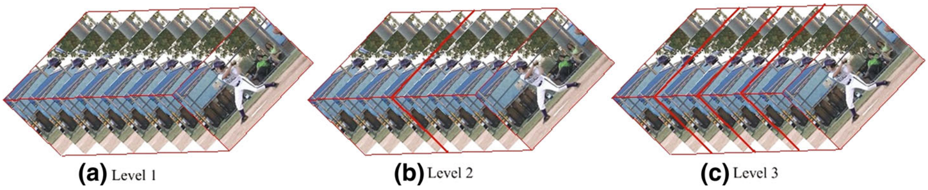


Fig. 3 Stratified pooling structure

4 Experiments

4.1 Datasets

The experiments are performed on HMDB-51 [15] and UCF-101 [43] which are considered as two most challenging video datasets for human action recognition. Some frames sampled from the UCF-101 and HMDB-51 are shown in Fig. 4.

HMDB-51 dataset includes 6766 video clips of 51 action classes, which are manually annotated clips selected from various sources such as YouTube, movies, etc. The dataset is divided into three splits for training and testing, with each split containing 3.7K training clips and 1.5K testing clips. Mean average precision (mAP) for action recognition is applied to evaluate the performance.

The UCF-101 dataset is composed of 13320 videos of 101 classes with each class having 100 clips at least. The same as HMDB-51, we use the standard splits for training and testing videos, and then mean average precision as the measurement of the final performance.

4.2 Experimental results

In the following experiments, the Caffe toolbox [17] is used for CNN implementation. Firstly, we investigate the impact of different network layers, feature dimensionality and pooling methods, and then evaluate the performance of our SP-CNN. Lastly, we compare our proposed SP-CNN with the state-of-the-art action recognition methods on HMDB-51 and UCF-101 datasets.

4.2.1 Dimensional reduction of Convolutional layer features

The original dimensionality of conv5 is $43264D$ which is much higher than that of full-connected layer feature activations, and the video feature dimensionality reaches $302848D$. It is essential to study the feature dimensionality impact on performance which is the key to achieve a better tradeoff between performance and storage costs. In this paper, we use PCA to reduce feature dimensionality and to de-correlate the feature descriptors. Figure 5 shows the results that are obtained by average pooling method on HMDB-51 dataset. The feature dimensionality changes from $2048D$ to $8192D$ and the results show that $6144D$ descriptors achieve the highest accuracy, which indicates that higher dimensionality does

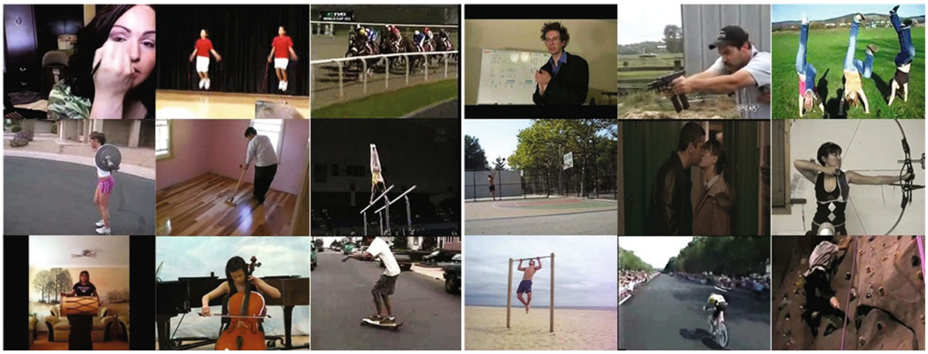


Fig. 4 Sample frames from UCF-101 and HMDB-51 datasets

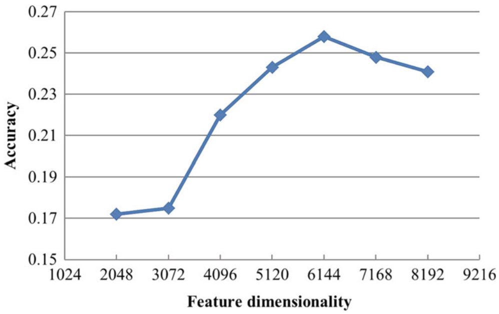


Fig. 5 Comparison the results with PCA on HMDB51 dataset (split 1)

not always lead to better performance. This is may be that the 6144D descriptors already carry enough information for recognition while high dimensions would result in instability problems adversely. Thus, we fix the dimensionality as 6144D for conv5 in the reminder of our experiments. For full-connected layer descriptors whose dimensionality is only 4096D. There is no necessity to reduce the dimensionality.

4.2.2 Comparison of fine-tuning methods

Transfer learning a new task from a pre-training model requires much less data and is faster than training from scratch. If using GPU fine-tuning, it is can learn an effective model in minutes or hours instead of days or weeks [17]. In this paper, we compare two fine-tuning methods: fine-tuning all layers which retraining all CNN parameters and fine-tuning the top 3 layers, i.e. retraining only the two full-connected layers and the last softmax layer. Figure 6 shows the results of the two different fine-tuning methods on the HMDB-51 dataset. In addition to the full6, the result of the other layers kept nearly unchanged. Possible reason may be that the convolutional features are generic features, such as edge features or color features which may be useful to most of recognition tasks. The full-connected feature contain more global semantic feature, which gradually more particular to the details of the classes contained in the original dataset. Therefore, fine-tuning convolutional parameters has no effect on classification performance. In contrast, it is necessary to fine-tune the full-connected parameters for recognition tasks. While, the difference between two methods in the accuracy of full7 is less than 0.5 %, the latter method requires less training time. Considering the trade-off between performance and fine-tuning time, we finally decide to fine-tune the top 3 layers.

4.2.3 Comparison of different layers

In deep CNN, different layers can carry different information of original images. The convolutional layers contain local features of the image like edges, texture, lines. While the full-connected layer usually denotes global strong semantic concepts. In this section, we study the influence of video-level features from different layers of CNN on HMDB-51 dataset. In term of convolutional layers, cross convolutional layer algorithm [29] is successful in image classification, and [46] proposed to use conv5 has good result for human action recognition. Therefore, we focus on comparing the influence of conv5 and full-connected layers activations on human action recognition. The results are reported in Fig. 6. As it can be seen from Fig. 6, convolutional layer only about 30 % mAP, indicating that convolutional features are inappropriate to be directly used for action recognition. Meanwhile, we notice

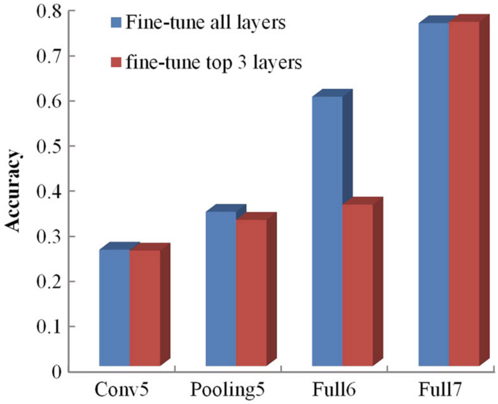


Fig. 6 Comparison of fine-tuning methods on HMDB51 dataset (split 1)

that using full7 can achieve the highest accuracy 76.3 % mAP, much better than the result of full6. The reason is that, the deeper full-connected layer has better discriminative power.

4.2.4 Comparison of pooling methods

As for aggregating frame-level features to get video-level features, we use two types of stratified pooling methods, including average pooling and max pooling two main types. We believe it is important to highlight the intuitive difference between max pooling and average pooling. In Fig. 7, we can find that max pooling outperforms average pooling in accuracy in conv5 layer and pool5 layer, but the opposite result occurs in the full-connected layer. Our study suggests that feature activations from convolutional layer are relatively sparse and actually strong activations in activated region, max pooling can effectively reserve those feature activations. The feature activations of full-connected layers are global feature. Compare with max pooling, average pooling has more resistance to noise and better performance.

4.2.5 Evaluation of SP-CNN

Stratified pooling aggregates an arbitrary number of the frame-level features to form the fixed-length global video-level feature descriptor, which is an essential component in our

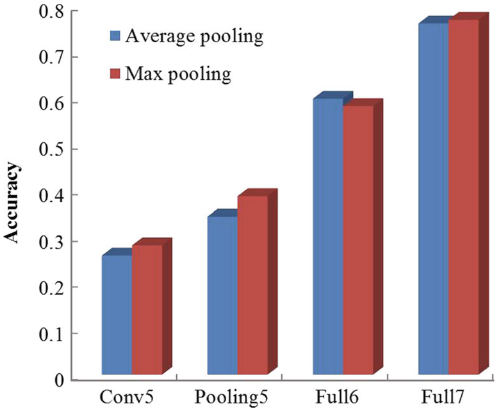


Fig. 7 Comparison of pooling methods on HMDB51 dataset (split 1)

Table 2 Different combinations of scale levels on HMDB-51 and UCF-101 (split 1)

Method	HMDB-51	UCF-101
Level1	75.2 %	88.1 %
Level2	76.4 %	89.1 %
Level3	76.1 %	89.0 %
Level1 + Level2	76.2 %	89.1 %
Level1 + Level3	76.3 %	89.4 %
Level2 + Level3	76.4 %	89.3 %
Level1 + Level2 + Level3 (SP-CNN)	76.5 %	89.6 %

method. In this section, we will evaluate our SP-CNN on different number of levels of stratified pooling.

The previous section has come to the conclusion that full7 reaches the highest accuracy. In this experiment, therefore, we use the full7 to investigate how many levels of stratified pooling affect recognition performance. The results are shown in Table 2. Firstly, we compare the accuracies resulting from three different numbers of levels, and find that level 2 achieves the highest accuracy 76.4 % mAP on HMDB-51 (split 1) and 89.1 % on UCF-101 (split 1). On the first split of two datasets, the accuracy of level 2 is 1 % over level 1, but roughly equivalent to level 3. This is shown that stratified pooling can retain some spatiotemporal information. Meanwhile, this is also illustrates that the higher level would not lead to better performance. And then we further compare performances of different combinations of descriptors generated from different numbers of levels on two datasets. From the results we can see that to concatenate all features of three scales of levels produces the highest recognition accuracy 76.5 % mAP and 89.6 % mAP, respectively. This is because that three scales of levels scheme carry the highest discrimination information for action recognition. Therefore, we concatenate features of all three scales of levels to form the final video-level feature descriptor.

4.2.6 *Comparison of the state-of-the-art*

Finally, we compare our best results with the state-of-the-art in Table 3. As can be seen from Table 3, in summary, deep-learned features reach higher accuracies compared to hand-crafted features. As for hand-crafted features, comparing HOF with MBH, we can see that MBH behaves better than HOF, resulting in about 3 % improvement but the feature dimensionality is doubled. Hybrid representation fusing multiple hand-crafted features and get 61.1 % mAP on HMDB-51 and 87.9 % mAP on UCF-101. However, along with the performance is improved. The feature dimensionality dramatically increase to 304524D. For deep-learned features, Wang et al. [46] combines TDD and IDT features which further improves the results reach the peak at 65.9 % mAP and 91.5 % mAP on HMDB-51 and UCF-101 respectively. Our proposed method with 74.7 % mAP on HMDB-51, which significantly outperforms the TDD and has only one-tenth of feature dimensionality with simple schemes. Our results achieve 91.6 % mAP on UCF-101, which is not obviously improved compared to the state-of-the-art method. The reason may be that there are some hard examples in the dataset for action recognition. It is difficult to keep improving the performance significantly when the accuracy has reached higher level.

Table 3 Comparison of our best results to the state-of-the-art (mean accuracy over three splits)

Feature type	Method	Feature dimensionality	HMDB-51	UCF-101
Hand-crafted	HOF [45]	27648	48.9 %	76.0 %
	MBH [45]	49152	52.1 %	80.8 %
	HOF+MBH [45]	76800	54.7 %	82.2 %
	IDT [45]	109056	57.2 %	84.7 %
	IDT+SFV [36]	204800	66.8 %	–
	Hybrid representation [35]	304524	61.1 %	87.9 %
Deep-learned	Two-stream ConvNets [41]	4096	59.4 %	88.0 %
	TDD [46]	32768	63.2 %	90.3 %
	TDD and IDT [46]	201729	65.9 %	91.5 %
	SP-CNN	28672	74.7 %	91.6 %

5 Conclusions

In this paper, we proposed a novel action recognition method called stratified pooling based on deep convolutional neural networks. Our method directly uses full-connected activations as frame-level features and get video-level feature after stratified pooling. The experimental results show that our method obtains the state-of-the-art performance on the two challenging datasets. In summary, our human action recognition method is simple but effective.

Acknowledgments This work is supported by the Nature Science Foundation of China (No.61202143, No. 61572409, No.61571188), the Natural Science Foundation of Fujian Province (No.2013J05100), the Research Foundation of Education Bureau of Hunan Province(No.15C0726).

References

1. Aarts E, Korst J (1988) Simulated annealing and boltzmann machines
2. Bay H, Tuytelaars T, Van Gool L (2006) Surf: speeded up robust features. In: Computer vision–ECCV 2006. Springer, pp 404–417
3. Chatfield K, Simonyan K, Vedaldi A, Zisserman A (2014) Return of the devil in the details: Delving deep into convolutional nets. arXiv:1405.3531
4. Chen QQ, Zhang YJ (2015) Cluster trees of improved trajectories for action recognition. Neurocomputing
5. Coates A, Ng AY (2011) The importance of encoding versus training with sparse coding and vector quantization. In: Proceedings of the 28th international conference on machine learning (ICML-11), pp 921–928
6. Donahue J, Anne Hendricks L, Guadarrama S, Rohrbach M, Venugopalan S, Saenko K, Darrell T (2015) Long-term recurrent convolutional networks for visual recognition and description. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2625–2634
7. Fan RE, Chang KW, Hsieh CJ, Wang XR, Lin CJ (2008) Liblinear: a library for large linear classification. J Mach Learn Res 9:1871–1874
8. Fei-Fei L, Perona P (2005) A bayesian hierarchical model for learning natural scene categories. In: IEEE computer society conference on computer vision and pattern recognition, 2005. CVPR 2005, vol 2. IEEE, pp 524–531

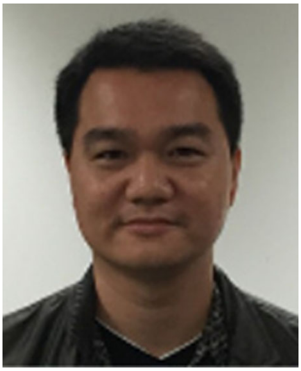
9. Gehring J, Miao Y, Metze F, Waibel A (2013) Extracting deep bottleneck features using stacked autoencoders. In: IEEE international conference on acoustics, speech and signal processing (ICASSP), 2013. IEEE, pp 3377–3381
10. Girshick R, Donahue J, Darrell T, Malik J (2014) Rich feature hierarchies for accurate object detection and semantic segmentation. In: IEEE conference on computer vision and pattern recognition (CVPR), 2014. IEEE, pp 580–587
11. Gkioxari G, Girshick R, Malik J (2015) Contextual action recognition with r* cnn. In: Proceedings of the IEEE international conference on computer vision, pp 1080–1088
12. Iosifidis A, Tefas A, Pitas I (2014) Class-specific reference discriminant analysis with application in human behavior analysis
13. Jain M, Jégou H., Bouthemy P (2013) Better exploiting motion for better action recognition. In: IEEE conference on computer vision and pattern recognition (CVPR), 2013. IEEE, pp 2555–2562
14. Jégou H., Perronnin F, Douze M, Sanchez J, Perez P, Schmid C (2012) Aggregating local image descriptors into compact codes. IEEE Trans Pattern Anal Mach Intell 34(9):1704–1716
15. Jhuang H, Garrote H, Poggio E, Serre T, Hmdb T (2011) A large video database for human motion recognition. In: Proceedings of IEEE international conference on computer vision
16. Ji S, Xu W, Yang M, Yu K (2013) 3d convolutional neural networks for human action recognition. IEEE Trans Pattern Anal Mach Intell 35(1):221–231
17. Jia Y, Shelhamer E, Donahue J, Karayev S, Long J, Girshick R, Guadarrama S, Darrell T (2014) Caffe: convolutional architecture for fast feature embedding. In: Proceedings of the ACM international conference on multimedia. ACM, pp 675–678
18. Jian M, Lam KM (2014) Face-image retrieval based on singular values and potential-field representation. Signal Process 100:9–15
19. Jian M, Lam KM (2015) Simultaneous hallucination and recognition of low-resolution faces based on singular value decomposition. IEEE Trans Circuits Syst Video Technol 25(11):1761–1772
20. Karpathy A, Toderici G, Shetty S, Leung T, Sukthankar R, Fei-Fei L (2014) Large-scale video classification with convolutional neural networks. In: IEEE conference on computer vision and pattern recognition (CVPR), 2014. IEEE, pp 1725–1732
21. Klaser A, Marszałek M., Schmid C (2008) A spatio-temporal descriptor based on 3d-gradients. In: BMVC 2008-19Th british machine vision conference. British Machine Vision Association, pp 275–271
22. Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems, pp 1097–1105
23. Laptev I (2005) On space-time interest points. Int J Comput Vis 64(2-3):107–123
24. Laptev I, Marszałek M., Schmid C, Rozenfeld B (2008) Learning realistic human actions from movies. In: IEEE conference on computer vision and pattern recognition, 2008. CVPR 2008. IEEE, pp 1–8
25. Le QV, Zou WY, Yeung SY, Ng AY (2011) Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis. In: IEEE conference on computer vision and pattern recognition (CVPR), 2011. IEEE, pp 3361–3368
26. Le Roux N, Bengio Y (2008) Representational power of restricted boltzmann machines and deep belief networks. Neural Comput 20(6):1631–1649
27. Lee H, Grosse R, Ranganath R, Ng AY (2009) Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations. In: Proceedings of the 26th annual international conference on machine learning. ACM, pp 609–616
28. Leng B, Zhang X, Yao M, Xiong Z (2015) A 3d model recognition mechanism based on deep boltzmann machines. Neurocomputing 151:593–602
29. Liu L, Shen C, Hengel AVD (2014) The treasure beneath convolutional layers: Cross-convolutional-layer pooling for image classification. arXiv:1411.7466
30. Liu R, Chen Y, Zhu X, Hou K (2015) Image classification using label constrained sparse coding. Multimedia Tools and Applications:1–15
31. Lowe DG (2004) Distinctive image features from scale-invariant keypoints. Int J Comput Vis 60(2):91–110
32. Luo J, Wang W, Qi H (2014) Spatio-temporal feature extraction and representation for rgb-d human action recognition. Pattern Recogn Lett 50:139–148
33. Mnih V, Heess N, Graves A et al. (2014) Recurrent models of visual attention. In: Advances in neural information processing systems, pp 2204–2212
34. Peng X, Qiao Y, Peng Q, Qi X (2013) Exploring motion boundary based sampling and spatial-temporal context descriptors for action recognition. In: British machine vision conference (BMVC)
35. Peng X, Wang L, Wang X, Qiao Y (2014) Bag of visual words and fusion methods for action recognition: comprehensive study and good practice. arXiv:1405.4506

36. Peng X, Zou C, Qiao Y, Peng Q (2014) Action recognition with stacked fisher vectors. In: Computer vision–ECCV 2014. Springer, pp 581–595
37. Perronnin F, Dance C (2007) Fisher kernels on visual vocabularies for image categorization. In: IEEE conference on computer vision and pattern recognition, 2007. CVPR'07. IEEE, pp 1–8
38. Perronnin F, Sánchez J., Mensink T (2010) Improving the fisher kernel for large-scale image classification. In: Computer vision–ECCV 2010. Springer, pp 143–156
39. Sermanet P, Eigen D, Zhang X, Mathieu M, Fergus R, LeCun Y (2013) Overfeat: integrated recognition, localization and detection using convolutional networks. arXiv:1312.6229
40. Sharma S, Kiros R, Salakhutdinov R (2015) Action recognition using visual attention. arXiv:1511.04119
41. Simonyan K, Zisserman A (2014) Two-stream convolutional networks for action recognition in videos. In: Advances in neural information processing systems, pp 568–576
42. Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556
43. Soomro K, Zamir AR, Shah M (2012) Ucf101: A dataset of 101 human actions classes from videos in the wild. arXiv:1212.0402
44. Wang H, Kläser A., Schmid C, Liu CL (2013) Dense trajectories and motion boundary descriptors for action recognition. Int J Comput Vis 103(1):60–79
45. Wang H, Schmid C (2013) Action recognition with improved trajectories. In: IEEE international conference on computer vision (ICCV), 2013. IEEE, pp 3551–3558
46. Wang L, Qiao Y, Tang X (2015) Action recognition with trajectory-pooled deep-convolutional descriptors. arXiv:1505.04868
47. Wang P, Cao Y, Shen C, Liu L, Shen HT (2015) Temporal pyramid pooling based convolutional neural networks for action recognition. arXiv:1503.01224
48. Xu H, Tian Q, Wang Z, Wu J (2015) A survey on aggregating methods for action recognition with dense trajectories. Multimedia Tools and Applications:1–17
49. Xu K, Ba J, Kiros R, Courville A, Salakhutdinov R, Zemel R, Bengio Y (2015) Show, attend and tell: neural image caption generation with visual attention. arXiv:1502.03044
50. Yao L, Torabi A, Cho K, Ballas N, Pal C, Larochelle H, Courville A (2015) Describing videos by exploiting temporal structure. In: Proceedings of the IEEE international conference on computer vision, pp 4507–4515
51. Yue-Hei Ng J, Hausknecht M, Vijayanarasimhan S, Vinyals O, Monga R, Toderici G (2015) Beyond short snippets: deep networks for video classification. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4694–4702
52. Zhou Y, Ni B, Hong R, Wang M, Tian Q (2015) Interaction part mining: a mid-level approach for fine-grained action recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3323–3331

Sheng Yu is currently pursuing the Ph.D degree at the Department of Cognitive Science, School of Information Science and Engineering, Xiamen University, China. His research interests include computer vision and pattern recognition.

Yun Cheng is current a professor at the School of Information, Hunan University of Humanities, Science and Technology, China. His research interest is image processing and computer vision.

Songzhi Su is current an assistant professor at the Department of Cognitive Science, School of Information Science and Engineering, Xiamen University, China. His research interest is computer vision.



Guorong Cai is current an associate professor at the School of Computer Engineering, Jimei University, China. His research interest covers image processing and computational intelligence.

Shaozi Li is current a professor at the Department of Cognitive Science, School of Information Science and Engineering, Xiamen University, China. His research interest is multimedia computing and computer vision.