Human Action Recognition based on Convolutional Neural Networks with a Convolutional Auto-Encoder

Chi Geng[1, a], JianXin Song[1, b]

[1]Nanjing University of Post and Telecommunications, Nanjing 210003, China.

[a]gengchi627@163.com, [b]songjx@njupt.edu.cn

Keywords: human action recognition, Convolutional Neural Networks, deep learning, pre-training.

Abstract. Human action recognition (HAR) research is hot in computer vision, but high precision recognition of human action in the complex background is still an open question. Most current methods build classifiers based on complex handcrafted features computed from the raw inputs, which are driven by tasks and uncertain. In this paper, type of deep model convolutional neural network (CNN) is proposed for HAR that can act directly on the raw inputs. In addition, an efficient pre-training strategy has been introduced to reduce the high computational cost of kernel training to enable improved real-world applications. The proposed approach has been tested on the KTH database and the achieved results compares favorably against state-of-the-art algorithms using hand-designed features.

Introduction

In the last decade, human action recognition (HAR) is becoming a more and more attractive research topic with several applications, such as video surveillance, virtual reality, intelligent human-computer interactions, etc. However, accurate recognition of actions is a highly challenging task due to cluttered backgrounds, occlusions, and viewpoint variations.

HAR consist of several stages, which describe the features that define activities or low level actions. A generic description of human action recognition from image sequence consist of two steps:1) extract complex handcrafted features from raw input video frames,2)build a classifier based on these features Some of the commonly used features for human action recognition are Histogram of Oriented Gradient (HOG)[1], Histogram of Optical Flow (HOF) , Motion Interchange Patters (MIP), Space-Time Interest Points (STIP), action bank features [2]and dense trajectories[3].However, these approaches are difficult and time consuming to extend these features to other systems. A large part of hand-design features are driven by task and different tasks may use completely different features. But in reality, it is hard to know what kind of feature is important to a specific task, so the feature selection is highly dependent on the specific problem. Especially for human action recognition, different kinds of sports show a very big difference in the appearance and motion model, it is hard to get the essential feature of action in the drastic change of environment .Therefore a generic feature extraction method is needed to be proposed to alleviate the need for hand-engineered features and reduces the calculation scale.

CNN[4] is a deep model that obtains complicated hierarchical features via convolutional operation alternating with sub-sampling operation on the raw input images. It is confirmed that CNN can gain more excellent performance in visual target recognition tasks through appropriate adjustment during the training. And CNN has invariance for a particular pose, illumination, and disorderly environmental change.

The first attempt for HAR using CNN was by[5] developing a novel 3D CNN model that extract features from both spatial and temporal dimensions by performing 3D convolutions, thereby capturing the motion information encoded in multiple adjacent frames. The developed model generated multiple channels of information from the input frames, and the final feature representation is obtained by combining information from all channels. In 2013, the same author improved the model by performing convolution and sub-sampling operations separately on gray-values of pixels, horizontal-gradient, vertical-gradient, horizontal optical flow and vertical optical flow channels extracted from adjacent input frames using hardwired layers. [6] proposed the used of multi-resolution CNN architecture and time information fusion for human action recognition on UCF-101 database using raw video as input. [7] proposed a deep convolutional network architecture for recognizing human actions in videos using action bank features of UCF50 database.[8] proposed a novel dynamic neural network model which can recognize dynamic visual image patterns of human actions based on learning. Convolutional neural network (CNN) and the multiple timescale recurrent neural networks (MTRNN) were introduced. [9] proposed a new method which combines part-based models and deep learning by training pose-normalized CNN.

Although CNN is a good option for HAR, this method still has a weakness that the kernels/weights employed in the convolution are trained by BP neural networks, which are very time consuming. In this paper, to solve this problem of HAR based on CNN, a convolutional auto-encoder (CAE) pre-training strategy has proposed. This method discovers good CNN initializations that avoid the numerous distinct local minima of highly non-convex objective functions arising in virtually all deep learning problems.

Proposed Approach

The proposed approach consists of a feature extraction step using CNN and a pattern recognition step using a Support Vector Machine (SVM) classifier as shown in Fig. 1.

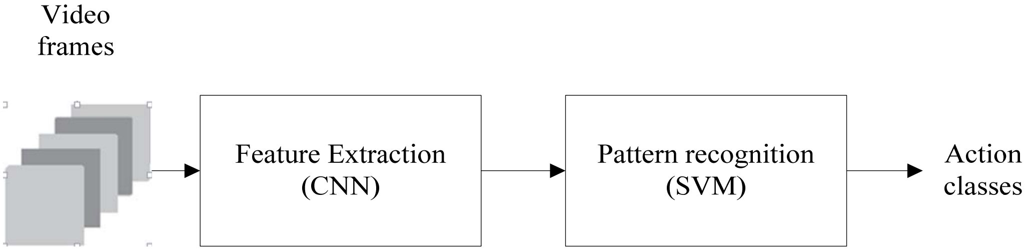

Fig. 1 Block diagram of the proposed approach

HAR with Deep CNN

In this section, we introduce using classical CNN into human action recognition in details. A CNN, consisting of multiple trainable stages stacked on top of each other, is employed to extract the features hierarchically, as in Fig. 2. Input is video frames size of 40 * 30 pixels that contains human actions. We will explain the flowchart of each layer of the network.

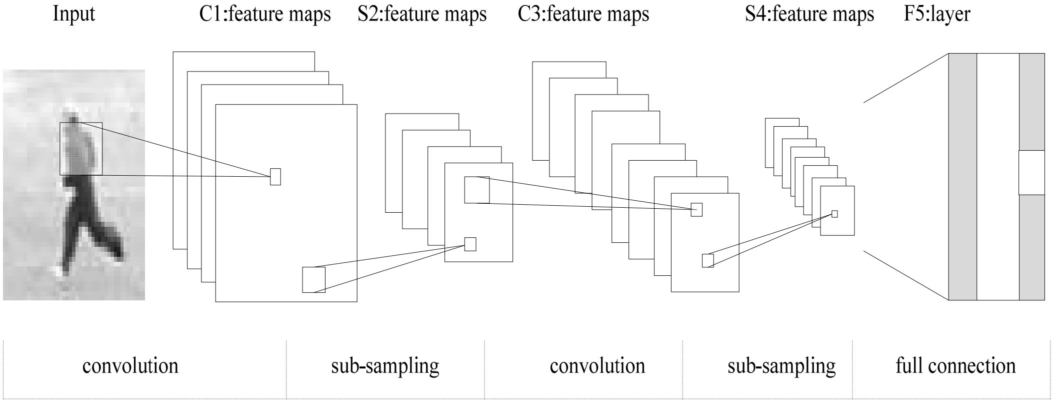

Fig. 2 Architecture of a convolutional neural network

Convoluting the input image with different trainable filters and additive biases, several feature maps can be generated in layer C1.Each feature map in S2 is obtained by a pooling operation that is performed on the corresponding feature maps in layer C1. The convolution and max-pooling procedures in layer C3 and layer S4 are the same as in layer C1 and layer S2.In the final recognition step, higher order features obtained after the final max-pooling layer S4 are eventually encoded into a 1-D vector, which is then categorized by a SVM classifier in the last layer of the CNN structure.

Generally, convolutional layer(C layer) is used for feature extraction, which input of each neuron is linked to the local receptive field of the previous layer, and extracting the local features. Once the local features are extracted, the location relationship between it and other features has been identified; Pooling layer(S layer) is considered to be a blur filter, playing the role of secondary feature extraction, that decrease spatial resolution between each hidden layer, and increase the number of plane in each layer, which can be used to extract multi-scale features from human action images.

A specific example of the convolution and max-pooling operation process is shown as Fig. 3.

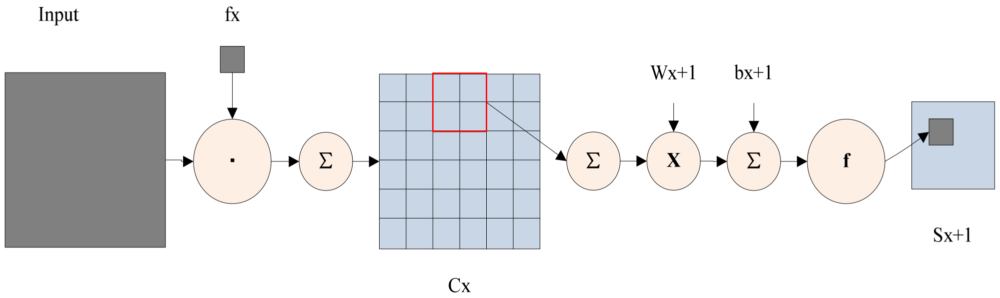

Fig. 3 convolution and max-pooling operation process

A convolution process including: use a trainable filter f_x to convolute the input image (the first stage is the input image, the remaining stages are the convolution feature maps), and then add a bias b_x, producing convolution layer C_x .where nonlinear mapping sigmoid(x) is defined as $f(x) = {}^{1}\!/_{(1+e^{-x})}$. Filter f_x is randomly initialized at first and is then trained with a well-known BP neural network. A max-pooling process includes: get the max pixel of every four pixels neighborhood l, then weighted by a scalar W_{x+1} and add a bias b_{x+1}, then through a sigmoid activation function, producing a feature map which size reducing four times.

CNN can automatically exact features from the raw image pixels, solving the blindness problem of task driven handcraft feature extraction method and improving the recognition accuracy[10]. The CNN structure combines three architectural ideas to ensure some degree of shift, scale, and distortion invariance, i.e., local receptive fields, shared weights, and pooling operations[11], which is particularly suitable for HAR in complicated real life.

Unsupervised Pre-training

As described in the previous section, filter f_x can be trained by a BP neural network; however, this can be very time consuming. We use a stacked Convolutional Auto-Encoders method for pre-training and obtain the first layer filters.

Considering the 2D image structure, the trend in vision and object recognition is to discover localized features that repeat themselves all over the input. Weights of CAEs are shared among all locations in the input, preserving spatial locality. The reconstruction is hence due to a linear combination of basic image patches based on the latent code. For a mono-channel input *x*, define the latent representation of the $l-th$ feature map to be

$$h^l = f\left(x \otimes W^l + b^l\right) \qquad (1)$$

Where $f(\cdot)$is an activation function, in this paper we use the sigmoid function:$f(z) = \frac{1}{1+e^{-z}}$. The bias is broadcasted to the whole map, and $\otimes$ denotes the 2D convolution. The reconstruction is obtained using

$$y = f\left(\sum_{l\in H} h^l \otimes \tilde{W}^l + c\right) \qquad (2)$$

Where again there is one bias *c* per input channels. *H* identifies the group of latent feature maps; $\tilde{W}$ identifies the flip operation over both dimensions of the weights.

The 2D convolution in equation (1) and (2) is determined by context. The cost function to minimize is the mean squared error (MSE):

$$E(\theta) = \frac{1}{2n}\sum_{i=1}^{n}(x_i - y_i)^2 \qquad (3)$$

Just as for standard networks the back propagation algorithm is applied to compute the gradient of the error function with respect to the parameters. This can be easily obtained by convolution operations using the following formula:

$$\frac{\partial E(\theta)}{\partial W^l} = x \otimes {\delta_h}^l + \tilde{h}^l \otimes \delta_y \qquad (4)$$

δ_h and δ_y are the deltas of the hidden states and the reconstruction, respectively. The weights are then updated using stochastic gradient descent.

EXPERIMENT

Experimental Database

In our experiment, we test our method on the KTH database[12] as shown in Fig. 4. The video database containing six types of human actions (walking, jogging, running, boxing, hand waving and hand clapping) performed several times by 25 subjects in four different scenarios: outdoors s1, outdoors with scale variation s2, outdoors with different clothes s3 and indoors s4 as illustrated below. Currently the database contains 2391 sequences. All sequences were taken over homogeneous backgrounds with a static camera with 25fps frame rate. The sequences were down sampled to the spatial resolution of 160 * 120 pixels and have a length of four seconds in average.

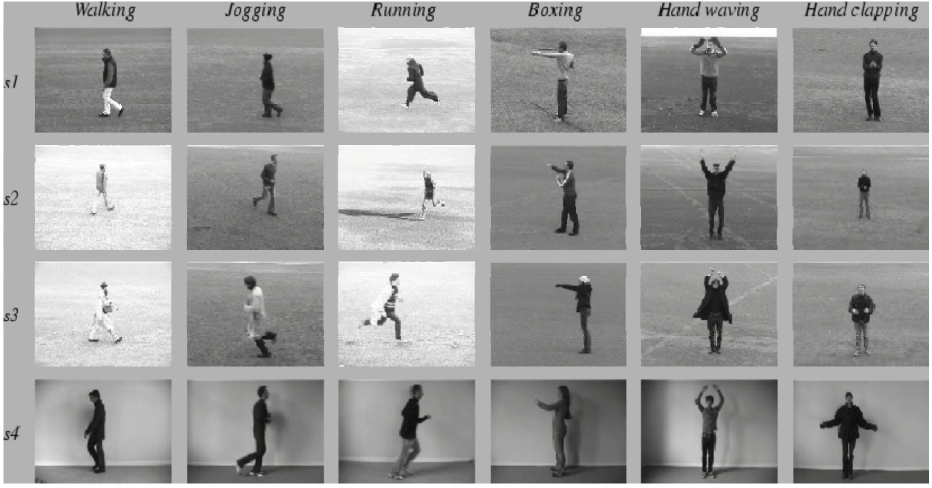

Fig. 4 KTH database

Experimental Set

The network input is a sequence of frames with size of 40 * 30 pixels by quadruple reducing the KTH frequencies as described in the previous section. In CNN-based HAR model, parameters of each layer in the CNN are presented in Table 1.

Table 1 specified parameters of the proposed CNN structure

Layer	Type	Feature maps	Kernel size
1	Convolution	6	5 * 5
2	Pooling	6	2 * 2
3	Convolution	12	5 * 5
4	Pooling	12	2 * 2

The network has 6 hidden layers: 1) convolutional layer with 6 5 * 5 filters per input channel; 2) max-pooling layer of 2 * 2; 3) convolutional layer with 16 5 * 5 filters per map; 4) max-pooling layer of 2 * 2; 5) a fully-connected layer of 120 hidden neurons. The SVM classifier in the last layer of the CNN is trained with the high order feature vectors of the training samples, and each query image can be categorized with its feature vector.

Experimental Results and Comparative Analysis

In the pre-training CNN-based HAR system, we train a CAE and use it to initialize a CNN with the same topology, to be fine-tuned for classification tasks. The classification accuracy for the six types of human actions from both the CNN and the pre-training CNN is summarized in Tables 2 and 3.

Table 2 classification accuracy across different human action recognition by the CNN

Action	Walking	Jogging	Running
Accuracy (%)	87.48	88.30	85.06
Action	Boxing	Hand waving	Hand clapping
Accuracy (%)	83.36	89.12	90.83
Average (%)	87.36		

Table 3 classification accuracy across different human action recognition by the pre-training CNN

Action	Walking	Jogging	Running
Accuracy (%)	92.20	93.53	90.04
Action	Boxing	Hand waving	Hand clapping
Accuracy (%)	89.27	94.41	95.50
Average (%)	92.49		

By comparing the results based on two different approaches in the experiment, the average recognition accuracy rate of the approach based on pre-training CNN is 5.13% higher than the results of the approach based on CNN. Because both the recognition approaches use the same classification method and run under the same system environment, so the major influence on the results should be the difference between the feature extraction methods.

We report the performance of our method on KTH dataset in Table 4. In this table, we compare our test set accuracy against reported results in literature. We note that for this database, our method actions superior performance compared favorably against algorithms using hand-designed feature.

Table 4 average accuracy on KTH database

Algorithm	Accuracy (%)
Harris3D+ISA+HOFfrom[13]	92.10
Laptev et al.[14]	91.80
Wang et al.[15]	90.20
Proposed method	92.49

Summary

In this paper, we have investigated the ability for CNN to learn features from video frames. We proposed an efficient pre-training strategy to initialize a CNN with Trained CAES Weights. The results have demonstrated that our method is able to outperform the existing methods on a publicly available action database. However, our method is limited to handling 2D inputs. In future work, we shall continue investigating ways to deal with the 3D CNN model for action recognition. There are also other deep architectures, such as the deep belief networks[16], which achieve promising performance on action recognition tasks. It's also an interesting research direction.

Reference

[1] Huang Y, Yang H, Huang P. Action recognition using hog feature in different resolution video sequences[C]//Computer Distributed Control and Intelligent Environmental Monitoring (CDCIEM), 2012 International Conference on. IEEE, 2012: 85-88.

[2] Sadanand S, Corso J J. Action bank: A high-level representation of activity in video[C]//Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on. IEEE, 2012: 1234-1241.

[3] Wang H, Kläser A, Schmid C, et al. Dense trajectories and motion boundary descriptors for action recognition[J]. International journal of computer vision, 2013, 103(1): 60-79.

[4] LeCun Y, Bottou L, Bengio Y, et al. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE, 1998, 86(11): 2278-2324.

[5] Ji S, Xu W, Yang M, et al. 3D convolutional neural networks for human action recognition[J]. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 2013, 35(1): 221-231.

[6] Karpathy A, Toderici G, Shetty S, et al. Large-scale video classification with convolutional neural networks[C]//Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on. IEEE, 2014: 1725-1732.

[7] Ijjina E P, Mohan C K. Human Action Recognition Based on Recognition of Linear Patterns in Action Bank Features Using Convolutional Neural Networks[C]//Machine Learning and Applications (ICMLA), 2014 13th International Conference on. IEEE, 2014: 178-182.

[8] Jung M, Hwang J, Tani J. Multiple spatio-temporal scales neural network for contextual visual recognition of human actions[C]//Development and Learning and Epigenetic Robotics (ICDL-Epirob), 2014 Joint IEEE International Conferences on. IEEE, 2014: 235-241.

[9] Zhang N, Paluri M, Ranzato M A, et al. Panda: Pose aligned networks for deep attribute modeling[C]//Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on. IEEE, 2014: 1637-1644.

[10] Farabet C, Couprie C, Najman L, et al. Learning hierarchical features for scene labeling[J]. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 2013, 35(8): 1915-1929.

[11] Fan J, Xu W, Wu Y, et al. Human tracking using convolutional neural networks[J]. Neural Networks, IEEE Transactions on, 2010, 21(10): 1610-1623.

[12] Schuldt, Laptev and Caputo, Proc. ICPR'04, Cambridge, UK

[13] Niebles J C, Wang H, Fei-Fei L. Unsupervised learning of human action categories using spatial-temporal words[J]. International journal of computer vision, 2008, 79(3): 299-318.

[14] Wang H, Kläser A, Schmid C, et al. Action recognition by dense trajectories[C]//Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on. IEEE, 2011: 3169-3176.

[15] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld.Learning realistic human actions from movies. In CVPR,2008. 3362, 3366

[16] Ali K H, Wang T. Learning features for action recognition and identity with deep belief networks[C]//Audio, Language and Image Processing (ICALIP), 2014 International Conference on. IEEE, 2014: 129-132.