DOI: 10.1051/itmconf/20171201025

Design and Implementation of Behavior Recognition System Based on Convolutional Neural Network

Bo YU

School of Software Engineering, Beijing University of Technology, Beijing, China
e-mail: yuubbo@126.com

Abstract—We build a set of human behavior recognition system based on the convolution neural network constructed for the specific human behavior in public places. Firstly, video of human behavior data set will be segmented into images, then we process the images by the method of background subtraction to extract moving foreground characters of body. Secondly, the training data sets are trained into the designed convolution neural network, and the depth learning network is constructed by stochastic gradient descent. Finally, the various behaviors of samples are classified and identified with the obtained network model, and the recognition results are compared with the current mainstream methods. The result show that the convolution neural network can study human behavior model automatically and identify human's behaviors without any manually annotated trainings.

1 Introduction

In recent years, with the rapid development of electronic information science and technology and video signal collection technology, human behavior recognition based on computer vision [1] has attracted more and more attention based on. Behavior recognition in computer vision to achieve detection, tracking, analysis and recognition of human actions [2] in video image based on is the key technology in the field of intelligent control and pattern recognition. Intelligent monitoring is aimed at using the computer outside image information into a digital signal, and then through a series of calculation instead of the human brain processing and understanding of visual information, so as to realize the intelligent analysis and automatic recognition of monitoring picture.

At present, researchers at home and abroad for the analysis and recognition of human behavior in video surveillance has done a lot of work. The identification methods can be divided into two categories: Based on behavior recognition method based on model matching and behavior recognition method of state space. Behavior recognition method based on model matching (Template Matching [3,4]) refers to the first to establish a good image template sequence to represent the static target behavior of the human body, then it will match the target template video image sequence detection in the template and, if the match succeeds the behavior is the behavior characteristics, identification or template it is determined, this kind of behavior is not. Recognition method based on state space (State Space Approaches) [5,6] defines specific posture state, then the state is connected by the way of probability. The action sequence of test set can be seen as the traversal process of current static posture, and then through the joint probability calculation during operation in order to achieve the action of the action scores. Deep learning is a research direction in recent years very popular, convolutional neural network(CNN)[7] as the representative of deep learning network improves the traditional neural network recognition effect. And the convolutional network implements a recognition method of end to end, it does not need to manually design features, images can be directly used as the input data network, avoid data reconstruction and feature extraction of the complex process in the traditional recognition algorithm.

Deep learning has attracted many researchers to study it, and have in some fields of computer vision has been successful. Especially in recent years, deep learning algorithms have been successfully applied in various fields such as speech recognition, image recognition, and gradually extended to research on behavior recognition with time series. In this paper, a method based on convolution neural network is proposed. The unsupervised learning of the network is realized through the human behavior data set, and the deep learning network model is established. The experimental results show that the network can identify various behavior effectively, compared with other recognition rate has improved.

2 The Framework of Behavior Recognition System

Human behavior recognition is mainly divided into two processes: the identification and understanding of human behavior feature extraction and motion, as shown below. Feature extraction is to extract the key features of this information in the data in the video data or image data, the feature information is the key task of recognition, feature extraction in fact directly affects the final recognition result.

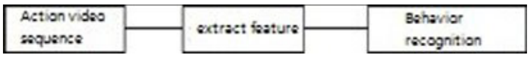


Figure 1 the main work of human behavior recognition

This algorithm is mainly composed of three parts, respectively, for video pretreatment, model training and behavior recognition part. In the video preprocessing part, firstly the original behavior of video preprocessing, using block updating background subtraction method to achieve target detection, two value image motion information is extracted, then the image input channel convolutional neural network, through the iterative training parameters of the network, to construct a model for convolution Behavior Recognition finally, you can use this network to identify human behavior, the flow chart of the system is shown in figure 2.

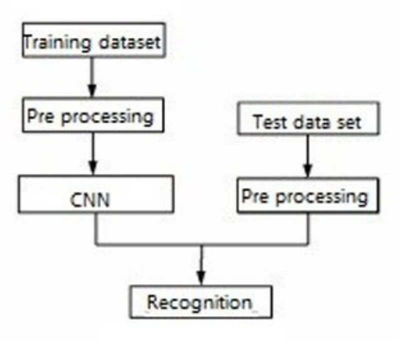


Fig. 2 flow chart of behavior recognition system

3 Behavioral Target Detection

Behavior of target detection is appropriate to process the image information of the human body contains behavior, remove the static background, detect the target motion information and behavior to carry, reliable data is provided for the subsequent recognition stage of the visual system. The behavior of target detection principle is as far as possible to retain the characteristic information of this behavior, while eliminating the redundancy information of the target, in the foundation of many application background detection is a very important step. This paper uses the Gauss mixture model based background subtraction in the background of the moving target detection. The principle is very simple, the basic process shown in figure 3.

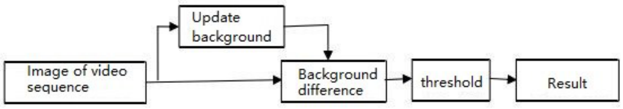


Figure 3 behavior target detection process

First, using the formula (2-1) f_{bk} image background image f_k and calculate the current frame difference, then the difference image D_k filtering processing, can extract moving objects.

$$\mathrm{D}_k(x,y) = |f_k(x,y) - f_{bk}(x,y)| \qquad (2\text{-}1)$$

Among them, $f_k(x,y)$ for the current frame image, $f_{bk}(x,y)$ as the background image, $D_k(x,y)$ for the difference image.

$$R_k(x,y) = \begin{cases} 1\ D_k(x,y) > T \\ 0\ D_k(x,y) \le T \end{cases} \qquad (2\text{-}2)$$

Among them, the threshold value of the image two is set to T, and the target area in the image is set to 255. The background subtraction method to the current frame and background image subtraction, the accuracy of motion target extraction is accurate or not will directly affect the final results. So if there is no background model update, it will lead to large error detection, such as light change accumulated to a certain time will be beyond the range of motion detection, the background part may be mistakenly classified as foreground moving objects.

When the background initialization is complete, need to adaptively update the parameters of the background image, to obtain an updated background estimation image for the $B_t = [\mu_t, \sigma_t^2]$ the use of sigma, type (2-3) update parameters.

$$\begin{cases} \mu_t = (1-\alpha)\mu_{t-1} + \alpha \times f_t \\ \alpha_t^2 = (1-\alpha)\sigma_{t-1}^2 + \alpha(f_t - \mu_t) \\ \alpha = K\frac{1}{\sqrt{2\pi\delta}}e^{-\frac{(\mu_{t-1}-f_t)^2}{2}} \end{cases} \qquad (2\text{-}3)$$

Alpha is the learning rate, indicating the background update rate, between 0 and 1. The following figure shows the KTH data set by the algorithm in different scenarios under the behavior of foreground extraction results.

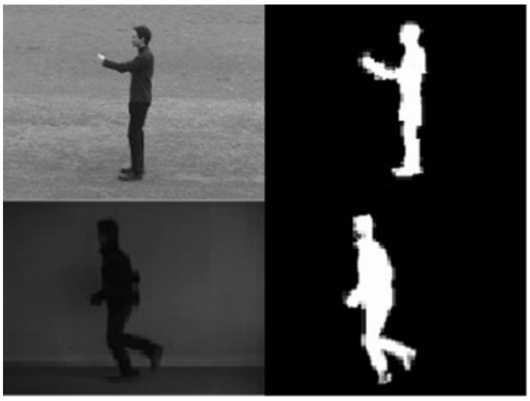


Fig. 4 foreground extraction

4 Convolutional Neural Network

For deep learning, suppose we have a training sample set (x^i, y^i), then the neural network can provide a kind of complex nonlinear model of $h_{W,b}(x)$, it has the parameters W, b as the weight parameters of the neural network, so as to fit the data in the dataset. The neural network consists of neurons, as shown in figure 5.

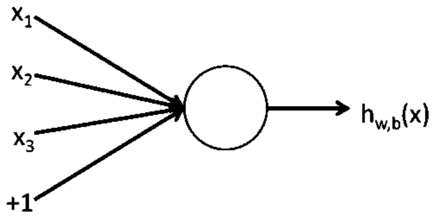


Fig. 5 neurons in neural networks

This neuron is a $\{x_1, x_2, x_3\}$ and intercept operation unit B as the input value, the output is

$$h_{W,b}(x) = f(W^T x) = f(\textstyle\sum_{i=1}^{3} W_i x_i + b) \qquad (3\text{-}1)$$

Among them, the function f is called the activation function. In this paper, we use the ReLU function as the activation function.

The neural network is many neurons together before a neuron output layer is after a neuron's input, as shown in figure 6.

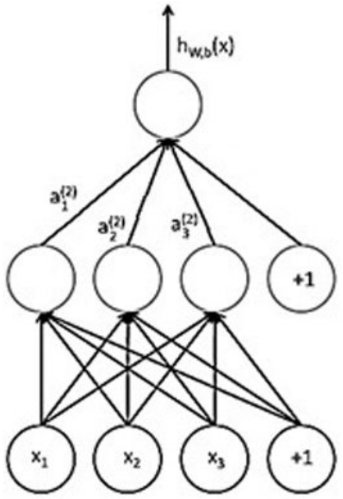


Fig. 6 neural network

The use of the circle to represent the nodes of the network, labeled "+1" node called the bias node. The left side of the neural network is the input layer, the right side of the output layer, the middle layer is called the hidden layer. It can be seen that there are 3 input units, 3 hidden units and 1 output units.

4.1convolution layer

The process of convolution layer processing is a layer of the image and the convolution kernel convolution, convolution processing is a basic operation of image processing. Related formulas are as follows:

$$y_{mn} = f\left(\sum_{j=0}^{J-1}\sum_{i=0}^{I-1} x_{m+i,n+j} w_{ij} + b\right),$$

$$(0 \leq m < M, 0 \leq n < N) \qquad (3\text{-}2)$$

Among them, X is a layer of two-dimensional data of the input image, w convolution kernel, B is a layer of bias, Y output, ReLU activation function. In the high dimensional input processing like images, each neuron is only partial region and input data connection, and use parameters in roll layer sharing is to control the amount of parameters.

4.2 sampling layer

The feature map sampling layer will be a layer of the sampling operation, reduce space dimension data layer, reduce the number of network parameters, making the computing resources cost less, and can effectively prevent over fitting. In this paper, the maximum sampling is used:

$$y_{mn} = \frac{1}{S_1 S_2}\left(\sum_{i=0}^{S_1-1}\sum_{j=0}^{S_2-1} x_{m*S_1+i,n*S_2+j}\right)$$

(3-3)

The sample size (S_1, S_2) is two-dimensional vector, x means input, y means output.

4.3 activation function

The input of each activation function is a number, and then a fixed mathematical operation is performed. The activation function used in the ReLU function in this paper, the function formula is f (x) =max (0, x), as shown in figure 7.

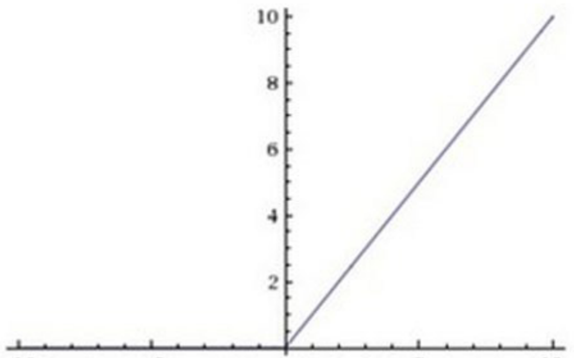


Fig.7 ReLU activation function

5 Experimental Process and Result Analysis

The selection of training samples from the KTH database, the database provides 6 kinds of behavior video, including normal walking, jogging, running, boxing, waving, clapping hands waving, each class has 25 people to complete the action, and each divided into four scenes, respectively in indoor and outdoor.

Convolutional network structure in this paper, a total of seven layers, including five volumes of sediments from three sampling layer, a connecting layer and an output layer. First, the input video data in accordance with the method of section second of the processed image, then a convolution of adjacent four successive images, the second layer is the largest under sampling layer, layer convolution kernel after deconvolution on a layer of feature map through the 3 volume layer finally fully connected layer processing using softmax regression model classification. Structure as shown in figure 8.

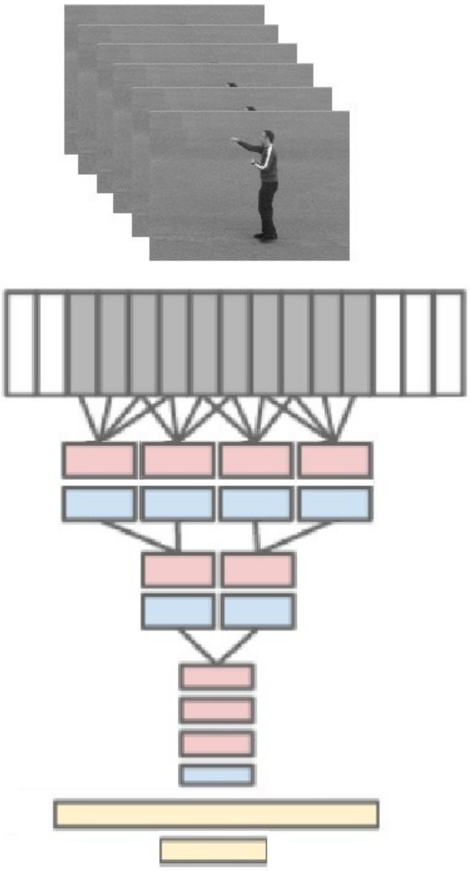
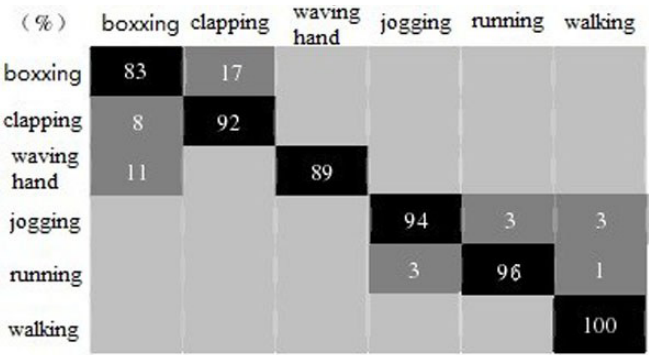


Fig. 8 convolution neural network structure

The video is divided into training set and test set for each video, video of the people will continue to repeat, the training set is used to construct the network, and then classify the test set. As can be seen from the following figure, in all kinds of behavior, the fist and applause in the test of the most confusing. At the same time, because between walking and running and jogging only in the video frequency and motion amplitude difference, so improve the recognition difficulty, for this system, the recognition rate is still relatively good.



（%）	boxxing	clapping	waving hand	jogging	running	walking
boxxing	83	17				
clapping	8	92				
waving hand	11		89			
jogging				94	3	3
running				3	96	1
walking						100

Fig.9 the recognition rate of the algorithm on the kth dataset

6 Copyright Forms and Reprint Orders

You must submit the IEEE Electronic Copyright Form (ECF) per Step 7 of the AT author kit's web page. THIS FORM MUST BE SUBMITTED IN ORDER TO PUBLISH YOUR PAPER.

Acknowledgment

This paper proposes a behavior recognition method based on convolutional neural network, learning the characteristics of human behavior through convolution unsupervised neural networks, then experiments are carried out by the neural network model of KTH human data set, experimental results show that the structure can do better on the human body for classification.

References

[1] Wang H, Klaser A, Schmid C, Liu C L. Action recognition by dense trajectories. Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Providence, RI: IEEE, 2011. 3169¡3176

[2] Soomro K, Zamir A. Action recognition in realistic sports videos. Computer Vision in Sports. 2014.

[3] Niu F, Adbel-Mottaleb M, HMM-based Segmentation and Recognition of Human Activities from Video Sequences[C], IEEE International Conference on Multimedia and Expo, Amsterdam, Netherlands, 2005

[4] Zhou Hanning, Kimber D.Unusual Event Detection Via Multi-Camera Video Mining[C]. 18th International Conference on Pattern Recognition, Hong Kong, China, 2006

[5] Brand M, Oliver N and Pentland A. Coupled hidden Markov models for complex action recognition. In: Proc IEEE Conference Computer Vision and Pattern Recognition, Puerto Rico, 1997, 994-999.

[6] Chen Yufeng, Liang Guoyuan, Lee K. Abnormal Behavior Detection by Multi-SVM based Beyesian Betwork[C].International Conference on Information Acquistion. Guangzhou, China, 2010.

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

[8] Wang H, Schmid C. Dense trajectories and motion boundary descriptors for action recognition[J]. International Journal of Computer Vision, 2013, 103(1): 60-79

[9] Zhang Fan, Gao li, Lu Haixian. Star Skeleton for Human Behavior Recognition[C].IEEE International Conference on Audio, Language and Image Processing, Shanghai, China, 2012.

[10] H. Wang and C. Schmid,Action recognition with improved trajectories, in ICCV, 2013.

[11] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, HMDB:a large video database for human motion recognition, in ICCV, 2011.

[12] I. S. Jacobs and C. P. Bean, "Fine particles, thin films and exchange anisotropy," in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.

[13] M. Baccouche, F. Mamalet, C.Wolf, C. Garcia, and A. Baskurt, Action classification in soccer videos with long short-term memory. recurrent neural networks, in International Conference on Artificial Neural Networks (ICANN), 2010.

[14] A. Farhadi, M. Hejrati, M. Sadeghi, P. Young, C. Rashtchian, J. Hockenmaier, and D. Forsyth, Every picture tells a story: Generating sentences from images, in ECCV, 2010.

[15] G. Kulkarni, V. Premraj, S. Dhar, S. Li, Y. Choi, A. C. Berg, and T. L.Berg, Baby talk: Understanding and generating simple image descriptions, in CVPR, 2011.

[16] Y. Yang, C. L. Teo, H. Daum´e III, and Y. Aloimonos,Corpusguided sentence generation of natural images, in EMNLP, 2011.

[17] M. Mitchell, X. Han, J. Dodge, A. Mensch, A. Goyal, A. Berg,K. Yamaguchi, T. Berg, K. Stratos, and H. Daum´e III, Midge: Generating image descriptions from computer vision detections, 2012