Human Behavior Recognition Based On CNN

Qing YE¹, Jun DENG¹, Yong-mei ZHANG² and Wen-bo HUANG¹

¹College of Electronic Information Engineering, North China University of Technology, Beijing No.5 Jinyuanzhuang Road, Shijingshan District, Beijing, 100144, China

²College of Computer, North China University of Technology, Beijing No.5 Jinyuanzhuang Road, Shijingshan District, Beijing, 100144, China

Keywords: Convolution Neural Networks, Small sample, Single-person behaviors, Double-person interaction.

Abstract. The technology of human behavior recognition is a cutting-edge topic which covers many areas-including computer vision, pattern recognition and artificial intelligence, and it have been widely used in areas such as commercial products, medical testing and military fields. According to the principle of Convolution Neural Networks (CNN), this paper presents a CNN human behavior recognition method based on small sample, which can identify a variety of human behaviors and human interaction behaviors respectively. The experimental result shows that the method can accurately identify the behavior of single-person and double-person interaction, and the recognition rate is above 92%, which proves the effectiveness of the method.

Introduction

Convolution neural networks [1] have been developing in recent years, which is a type of recognition algorithm with high efficiency draws attention of all walks of life. Its origins date back to the middle of the last century. In the study of the cortex of a cat, two scholars found that its neurons was especially unique in the local sensitivity and direction choosing, and the network structure can simplify the feedback neural networks. At this point the convolution neural networks were proposed. Nowadays, convolution neural networks have attracted the attention of the scientific fields, especially in the area of pattern classification. Because of its simpleness and efficiency, its application area continues to expand.

The extraction and classification of image features have always been a basic and important research direction in the field of computer vision. Convolution neural networks provide an end-to-end learning model, in which parameters can be trained by traditional gradient descent methods. The trained convolution neural networks can learn the characteristics of the image and complete the extraction and classification of the image features. As an important research branch in the field of neural networks, the convolution neural network is characterized by that the characteristics of each layer are excited via the convolution kernel of the shared region by the local area of the upper layer. This feature makes the convolution neural networks more suitable for the learning and expression of image features than other neural networks.

Principle of CNN

The structure of CNN is different from network structures [2] with other performance, which is the transformation on the basis of the original theory. The calculation of the output value of this two uses the principle of forward propagation, while the negative propagation is used to adjust the weights and offsets. The difference between them is that the connection type between the neural units of the BP network [3] is full connection, while CNN uses the type of partial connection. In other words, the perceptual information of any neuron is derived from the upper part of the neuron, not all of them. And here are the three elements of CNN: perceived local area, weight sharing, sampling in the space or time.

The three characteristics of CNN make it very robust on the distortion of the input image data spatially and temporally. In this way, the convolution layer extract features, and then combine to
form a more abstract feature, and finally it forms a description of the picture object. Figure 1 is a typical example of a convolution neural network LeNet-5 [4].

![Figure 1. LeNet-5 structure.](image)

The structure consists of an input layer, a convolution layer, a down sampling layer (pool layer), a fully connected layer and an output layer. The input layer is a 32×32 pixel image (typically converted to grayscale image to be processed). For the input image, the greater the image is, the more information is available. And of course, this will increase the corresponding number of layers (without changing the convolution kernel and down sampling rate). Usually, in order to improve the convergence rate, the normalization of image processing is needed during network training.

According to the introduction of the CNN structure above, its working principle can be simply summed up by two parts:

**Define the network model.** The definition of a network model requires a specific definition of the problem. It requires to combine the specific application-related data needs or the own unique characteristics of the data, design and apply the role of various levels, as well as the functional characteristics of each set parameter (, etc.). At present, the model designs for CNN are plenty, such as the depth study of the model [5], CNN convolution step and the model design of relevant incentive functions [6].

**The principles of network training.** CNN is often trained by two methods - the residual transmission and the set of each parameter. However, the convergence of training is often hindered by problems such as over-fitting, gradient elapse and explosion, and then the convergence is weakened. In recent years, convolution neural network is still moving forward, and it is also optimized and expanding in its scale, so it needs its algorithm to put forward higher theoretical requirements.

**Human Body Behavior Recognition Based on CNN**

CNN’s development speed is due to its advantages in the field of identification, and therefore it evolved a lot of CNN changes in the form, but its essence is still two parts -convolution and down sampling. In this paper, we use the convolution neural network based on improvements for human body behavior identification. In the feature extraction layer, it is not only in accordance with the original CNN feature extraction to complete, but the combination of human behavior edge and corner characteristics, so that the feature information is more comprehensive, the range is wider, which is more conducive to training and learning. At the same time, the feature mapping layer still uses the similar network structure of LetNet-5, and optimizes its network mapping relationship, improves the input combination on the basis of figure(1), which makes the network richer and has more features, and it is also conducive to expanding the network size and improving the recognition rate of human behaviors. In this paper, CNN training methods use a small-sample statistical learning, and in order to improve the convergence of the network, the design of a gradual attenuation of the learning rate is prepared for the use of relatively small learning rate. For how to improve the learning speed, the second derivative is used.

First, the optimal learning rate \( \eta \) is acquired. Assuming that the objective function can be approximated by the quadratic function, it can be launched by Taylor expansion:
In formula (1), . Take the derivative on both sides of formula (1):

\[
\frac{dE(W)}{dW} = \frac{dE(W_{c})}{dW} + (W - W_{c}) \frac{d^{2}E(W_{c})}{dW^{2}} + ...
\]  

(1)

In formula(1), \( \frac{dE(W_{c})}{dW} = \frac{dE}{dW} |_{W = W_{c}} \). Take the derivative on both sides of formula(1):

\[
\frac{dE(W)}{dW} = \frac{dE(W_{c})}{dW} + (W - W_{c}) \frac{d^{2}E(W_{c})}{dW^{2}}
\]  

(2)

Set \( W = W_{mn} \), so that \( \frac{dE}{dW} |_{W = W_{mn}} = 0 \).

\[
W_{mn} = W_{c} - (\frac{d^{2}E(W_{c})}{dW^{2}})^{-1} \frac{dE(W_{c})}{dW}
\]  

(3)

So the optimal learning rate is as shown in equation (4):

\[
\eta_{opt} = (\frac{d^{2}E(W_{c})}{dW^{2}})^{-1}
\]  

(4)

The above is only one-dimensional situation. In the multi-dimensional update formula, it is very difficult to get the optimal learning rate. Here, \( \frac{d^{2}E(W_{c})}{dW^{2}} = H \), then \( H \) will become the Hessian matrix.

For CNN’s multi-layer network, assume one of the layers has \( N_{i} \) inputs and \( N_{o} \) outputs, and \( O = F(W, X) \), \( \partial^{2}E / \partial O^{2} \) is a \( N_{o} \times N_{o} \) matrix, then:

\[
\frac{\partial^{2}E}{\partial W} = \frac{\partial O^{T}}{\partial W} \frac{\partial^{2}E}{\partial O \partial O} + \frac{\partial E}{\partial O} \frac{\partial^{2}E}{\partial O^{2}}
\]  

(5)

Under normal circumstances, the second item in the right side of formula (5) will be ignored. If the non-diagonal elements in \( \partial^{2}E / \partial O^{2} \) are set to zero, then:

\[
\frac{\partial^{2}E}{\partial W} = \sum \frac{\partial^{2}E}{\partial O_{i} \partial O_{j}} \frac{\partial O_{i}}{\partial W_{j}}
\]  

(6)

In the experiment, the number of design parameters is \( m \). The first step is to get the first-order partial derivative \( J \) of these parameters. The second step is to derive the second derivative of all the parameters according to the formula (2), and obtain the second-order partial derivative \( H \) and transform it into a diagonal matrix ( \( m \times m \) ). Finally, after a number of iterations, we get an average matrix \( \bar{H} \). From the Newton method mentioned above, we can see that it aims to obtain the exact inverse of the hessian matrix. Therefore, in this paper, in order to be more efficient, we adopt a simple and pragmatic approach to approximate the inverse matrix \( H^{-1} \) of the required matrix hessian.

\[
\Delta W = H^{-1}J = (I + \rho*H) \Delta \eta
\]  

(7)

In formula (7), \( H \) represents the diagonal matrix ( \( m \times m \) ), \( \rho \) represents the unit diagonal matrix weight (generally set to 0.2). In the experiment, the learning rate is expressed by \( \eta \).

**Experimental Results and Analysis**

In this paper, 13 kinds of human behavior in the laboratory environment are collected, which is completed by many people with different gender and of different body sizes. Among them, these human behaviors mainly include hands lifting, bending, walking, squatting, one hand waving, sideways stretching, stoop, lying, making phone calls, two-person communicating, two-person shaking hands, two-person bowing and two saluting, which are expressed as symbols B1, B2, B3, B4, B5, B6, B7, B8, B9, B10, B11, B12, B13 respectively. At the same time, in order to ensure the objectivity and universality of the experiment, a number of the men and women of different height take part in the experiment to collect data information.

---

In this paper, we use the small sample CNN to train 13 kinds of human behavior data for data training of different characteristics, for each action there are 100 behavior samples selected to train, and then 60 behavior images are selected as identification samples, and conduct multiple experiments. Figure 2 is the result image of the behavior identification processing on the Microsoft Visual Studio platform, we can see from the picture that it can successfully identify the type of human behavior from all types of behaviors at the end.

![Figure 2. Behavior identification results.](image)

(a) single-person behavior identification  
(b) two-person interaction identification

Table 1 shows the identification results of various types of human behavior using method in this paper. It is clear from Table 1 that the recognition rate of the 13 different behaviors obtained by this experiment has reached a high demand, which is due to that in the actual collection of behavioral data, each of the different collected personnel are in strict accordance with the requirements, careful implement of each of the criteria of the identified behavior, so that these different acts form a more obvious distinction.

<table>
<thead>
<tr>
<th>Behavior type</th>
<th>Recognition rate</th>
<th>Behavior type</th>
<th>Recognition rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>B1</td>
<td>98.8%</td>
<td>B2</td>
<td>94.2%</td>
</tr>
<tr>
<td>B3</td>
<td>97.1%</td>
<td>B4</td>
<td>93.5%</td>
</tr>
<tr>
<td>B5</td>
<td>97.9%</td>
<td>B6</td>
<td>96.3%</td>
</tr>
<tr>
<td>B7</td>
<td>94.8%</td>
<td>B8</td>
<td>98.2%</td>
</tr>
<tr>
<td>B9</td>
<td>99.1%</td>
<td>B10</td>
<td>92.7%</td>
</tr>
<tr>
<td>B11</td>
<td>94.6%</td>
<td>B12</td>
<td>95.2%</td>
</tr>
<tr>
<td>B13</td>
<td>93.8%</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Conclusion**

In this paper, CNN behavior recognition method based on small sample is proposed, and 13 different human behaviors are being identified and analyzed. We first introduce the structure and principle of CNN algorithm, and then combine the edge and corner features to improve the algorithm through the analysis and judgment of the original characteristics. Finally using the experimental data to prove the effectiveness and reliability of this method, that is ultimately conducive to efficient training, thereby improving the behavior recognition rate.

**Acknowledgement**

This paper is supported by Excellent Talent Training Project of Beijing (2013D005002000002), National Natural Science Foundation of China (Grant No.61371143), Natural Science Foundation of Beijing (Grant No.4132026), National Science and Technology Support Program of China (2012BAH04F00), Technology Project of Beijing Municipal Education Commission (KM201210009007), Excellent Young Teachers Training Project by North China University of Technology.
References


