Research on Recognition Algorithm of Abnormal Behavior of Workers in Two-Stream Convolutional Network

Yinong Zhao, Ruoxi Li, Xuedong Zhang, Yuhan Cao and Xiaojing Chen

Computer Science and Software Engineering, University of Science and Technology, Anshan, Liaoning, China
Email: 784542390@qq.com

Abstract. In order to identify the abnormal behaviors of workers in the factory, this paper proposes an improved algorithm for identifying abnormal behaviors of workers in a two-stream convolutional network. The workers' body contour shape information extracted from the convolutional neural network is input into the LSTM network in order to extract timing information between frames. Secondary extraction of the dense optical flow image of the video image, sparse extraction of pixels with small changes in optical flow value in the dense optical flow image, and then put the new continuous optical flow image into the continuous optical flow image network. The two networks are fused after softmax classification to get the final recognition result. Experiments on the CAVIAR dataset, CASIA dataset, and self-built behavior dataset show that compared with other abnormal behavior detection methods and traditional two-stream convolution algorithms, the accuracy of the improved algorithm in this paper is improved by 1%-4%.

Keywords. Abnormal behavior recognition; two-stream convolutional network; LSTM network

1. Introduction

Accidental injuries usually occur in industrial production. At present, most video surveillance equipment can only store the occurrence of accidents, but cannot detect workers’ accidents in real time. Meanwhile, manual identification has the problems of high labor intensity, wrong detection and missed detection. Therefore, the real-time detection of abnormal behavior of workers that can cause accidents has attracted widespread attention [1].

Workers’ abnormal behavior recognition methods can be divided into traditional methods and deep learning classification and recognition methods. The traditional method is to use human-defined features to determine abnormal human behavior. Fujiyoshi [2] proposed to determine abnormal behavior by extracting human contour features. Wang [3] et al. fuses various human characteristics and use SVM for behavior detection. Bobick [4] proposed a view-based human motion representation and recognition method. Fu [5] proposed a new human motion recognition framework based on a mixed event probability sequence of hidden Markov models. The abnormal behavior recognition rate of these methods is particularly sensitive to custom features, which is the limitation of custom features. Due to the characteristics of high efficiency and high recognition rate of deep learning classification and recognition, it has gradually replaced traditional recognition methods, such as CNN neural network [6], AlexNet model [7], VGG model [8], InceptionNet model [9], TSN model [10], 3D-CNN model [11], CNN + LSTM model [12].
At present, there are some problems in both traditional methods and deep learning methods. First, the traditional method has high requirements for the background environment of video shooting and the hardware performance of the shooting equipment. Meanwhile, the feature extraction method based on artificial priors increases the labor cost. Second, at present, deep learning methods have achieved good results in the accuracy of identifying abnormal behavior in still images. Due to the temporal correlation between the behavior of the moving target in each frame of the video sequence, there is a problem that the convolutional neural network cannot extract in the time dimension. Third, for abnormal behavior data sets, there are no large data sets similar to UCFI at home and abroad, and more diverse and standardized data sets need to be established for different backgrounds.

Therefore, in view of the above problems, this paper proposes an improved two-stream convolutional network based on video images, and establishes a self-built data set to verify the method in this paper.

2. Improved Two-stream Convolutional Network Model
In the improved two-stream convolutional network, the video frame image network training part uses a 2D convolutional network to extract spatial features, and then inputs it to the LSTM network to obtain the time series features. Calculate the LSTM value at each video frame moment and average it. In the continuous optical flow image network part, framed the moving target after extracting the dense optical flow, and then the sparsely extracting dense optical flow of the contour of the moving target. Finally input to the 2D convolutional network for training. The improved network can not only extract more target motion information, but also improve the network’s ability to distinguish similar actions. The two parts of the network do linear weighted fusion by outputting the softmax function. As shown in figure 1.

![Figure 1. Improved two-stream convolutional neural network structure.](image)

2.1. Fusion of Video Frame Image Network and LSTM Neural Network
The network input value $X_t$, the last time LSTM output value $H_{t-1}$, the last time state unit $C_{t-1}$, the current time LSTM output value $H_t$ and the current time Unit status $C_t$. The network controls the content of the unit state $C_t$ through the forgetting gate $f_t$ and the inputting gate $i_t$, and uses the outputting gate $O_t$ to control the outputting quantity of $C_t$.

In the improved network in this paper, the feature vector is extracted from the still video frame by 2D convolutional network. The each 2D convolutional network is used to extract the feature vector from the RGB pictures at each $t$ frame time. One outputting $R_t$ of the 2D-CNN layer corresponds to the LSTM at time $t$ enter $X_t$. The network convergence structure is shown in figure 2.
\[ W_f \] is the weight matrix of the forgetting gate, The matrix dimension is the dimension of the input plus the dimension of the hidden layer, then multiplied by the dimension of the unit state \([H_{t-1}, R_t]\) is the two vectors \(R_t\) and \(H_{t-1}\) connected into a longer vector. \(b_f\) is the bias term of the forgetting gate. \(\sigma\) is the sigmoid function. Equation for forgetting gate:

\[
f_f = \sigma(W_f [H_{t-1}, R_t] + b_f)
\]

\[ W_i \] is the weight matrix of the inputting gate, and \(b_i\) is the bias term of the inputting gate. Equation for inputting gate:

\[
i_i = \sigma(W_i [H_{t-1}, R_t] + b_i)
\]

The current inputting unit state \(C_t\) is calculated according to the previous outputting and the current inputting, \(b_i\) is the bias term for the unit state \(C_t\), and \(W_{ri}\) is the weight matrix of the unit state \(C_t\). The equation is shown in equation (3):

\[
C_t = \tanh(W_{ri} [H_{t-1}, R_t] + b_i)
\]

Combine \(C_{t-1}\) and \(C_t\) together by equation (4) to form a new unit state:

\[
C_t = f_f \cdot C_{t-1} + i_i \cdot C_t
\]

\(b_o\) is the bias term of the outputting gate. \(W_{ro}\) represents the weight matrix of the outputting gate. Equation of the influence of outputting gate on current output:

\[
o_o = \sigma(W_{ro} [H_{t-1}, R_t] + b_o)
\]

The final output of the network is determined by the output gate and the unit state:

\[
H_o = o_o \cdot \tanh(C_t)
\]

2.2. Secondary Optical Flow Extraction

For a video, it is assumed that the luminance value of a pixel \((x, y)\) in the image at time \(t_0\) is \(I(x, y, t_0)\). Let \(u(x, y)\) and \(v(x, y)\) be the optical flow components of the pixel in the \(x\) direction and the \(y\) direction, that is, the velocity vector of the pixel. When the changing image sequence is within a relatively small time \(d,\) the partial derivatives in the \(x\) and \(y\) directions are \(u = \frac{dx}{dt}\) and \(v = \frac{dy}{dt}\). Based on the assumption of constant brightness, the brightness value (gray value) of the same pixel between the frames doesn’t change, then we get:

\[
I(x, y, t_0) = I(x + dx, y + dy, t_0 + dt)
\]

According to the Taylor series expansion equation, the right side of equation (7) can be expanded to obtain equation (8):
\[ I(x + dx, y + dy, t_0 + dt) = I(x, y, t_0) + \]
\[ \frac{\partial I}{\partial x} dx + \frac{\partial I}{\partial y} dy + \frac{\partial I}{\partial t} dt + \varepsilon \]

\[ \varepsilon \] in equation (8) can be ignored. Based on the assumption of constant brightness, equation (8) can be converted into equation (9):
\[ I_x u + I_y v + I_t = 0 \] (9)

\( I_x, I_y, I_t \) is the partial derivative of the brightness value of the pixel \((x, y)\) along the direction \(x, y, t\), and their values can be directly calculated by solving the video image sequence. Through the optical flow constraint equation, the relationship between the spatial gradient value and the optical flow velocity value can be obtained. In order to calculate the optical flow values \(u(x, y)\) and \(v(x, y)\) of this pixel in two directions at the current moment, constraints introduced by global optical flow estimation method-smooth constraint term \(\phi\), minimized the term:
\[ \phi = \iint \left[ \left( \frac{\partial u}{\partial x} \right)^2 + \left( \frac{\partial u}{\partial y} \right)^2 + \left( \frac{\partial v}{\partial x} \right)^2 + \left( \frac{\partial v}{\partial y} \right)^2 \right] dxdy \] (10)

Let the data items in equation (9) be \(\phi'\) and minimize them to get:
\[ \phi' = \iint (I_x u + I_y v + I_t) dxdy \] (11)

The solution of the dense optical flow field can be transformed into:
\[ \psi = \min \iint \left[ \frac{\left( I_x u + I_y v + I_t \right)^2}{\beta^2 + I_x^2 + I_y^2} \right] dxdy \] (12)

The weighting coefficient \(\beta\) is 0.5, and the variational method is used to obtain the values of the dense optical flow vectors \(u(x, y)\) and \(v(x, y)\) by solving the above equation.
\[ u_{t+1} = u_t - I_x \frac{I_x u_t + I_y v_t + I_t}{\beta^2 + I_x^2 + I_y^2} \] (13)
\[ v_{t+1} = v_t - I_y \frac{I_x u_t + I_y v_t + I_t}{\beta^2 + I_x^2 + I_y^2} \] (14)

The dense optical flow extracts the optical flow values of all pixels in the entire picture, but in reality, we do not need to know the optical flow changes of all points, only the optical flow values of pixels in the active area of the moving target. In order to distinguish similar actions in the future, this paper extracts the optical flow in the moving area and superimposes it on the dense optical flow calculation to form a new optical flow image.
Sparse optical flow extraction based on local optical flow estimation. It is assumed that the optical flow value of each pixel in an area centered on \(p\) point is same, and different weights are given to different points in the area, so the calculation of optical flow is converted into calculate the error between the two frames that is the optical flow value:
\[ \lambda(D, A) = \sum_{i \in \Omega} W^2(x) \left[ I(x + dx, y + dy, t + dt) - I(x, y, t) \right]^2 \] (15)

The transformation matrix \( A \) can be expressed as equation (16), and use least squares method to solve \( D \) as equation (17):

\[ A^T W^2 A D = A^T W^2 b \] (16)
\[ D = (A^T W^2 A)^{-1} A^T W^2 b \] (17)

Expanding \( A^T W^2 A \) in equation (17) can get equation (18):

\[ A^T W^2 A = \left[ \sum W^2(x)I^2_i(x) \quad \sum W^2(x)I_i(x)I^2_j(x) \right] \] (18)

Because the sparse optical flow still satisfies the basic optical flow assumption, it can be obtained from equation (9):

\[ \begin{bmatrix} I_i \n I_j \end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix} = -I_i \] (19)

By bringing equation (19) into equation (18), the optical flow value in the sparse optical flow field can be obtained, that is, the optical flow values in the \( u \) and \( v \) directions.

\[ u = -\sum W^2(x)I^2_i(x)I_i - \sum W^2(x)I_i(x)I^2_j(x)I_j \] (20)
\[ v = -\sum W^2(x)I_j(x)I^2_i(x)I_i - \sum W^2(x)I_i(x)I^2_j(x)I_j \] (21)

2.3. Network Output

Video frame training network and continuous optical flow image training network contain the contour information and motion information of the moving target. According to the method in [12], this paper uses the video frame training network to output 1/3 weights after softmax output. The network weight accounts for 2/3, and the final network output is calculated as:

\[ R = 1/3R_1 + 2/3R_2 \] (22)

3. Experimental Results and Analysis

3.1. Data Set Selection and Experimental Environment

Based on the industrial background, we defined and subdivided the abnormal behaviors which usually occurred in factories. The main activities performed by workers in the industrial production process are four types of actions: equipment inspection, equipment maintenance, equipment inspection, and rest in place. The above actions are highly repetitive and intermittent. Therefore, for these features, this article filtered the two public data sets.

Select normal actions in the CAVIAR dataset, including walking, hovering, waving, and resting; abnormal movements, including falling, squatting. There are 6 types, 22 videos in total, video resolution is 384 x 288, frame rate is 25 fps.

Select normal actions in the CASIA dataset, including walking, running, bending over, jumping, and hovering. The abnormal actions include fainting, squatting. There are 7 types, 21 videos in total, video resolution is 320 x 240, frame rate is 25 fps.

To verify the accuracy and practicability of the method proposed in the paper, a self-built dataset was added to the experiment. In the laboratory and corridor, simulate the normal and abnormal movements of workers in the factory, and take a video to build a self-built data set. The normal
movements include walking, jumping, squatting (short time), hovering, two-person. Abnormal movements include squatting (long time), fainting. There are 7 types, 84 videos in total, video resolution is 1280×720, frame rate is 25 fps.

This paper is based on Python’s pytorch learning framework and performs experiments in the CPU and win7 system environments.

3.2. Network Parameter Setting
Because there are too many interference factors in the factory, the video data is preprocessed first, and then the video frames are input into the convolutional neural network for training. The initial learning rate is set to 0.001, SGD is used as the optimizer, the activation function is relu and the number of LSTM nodes is 64.

3.3. Extraction of Optical Flow Image and Optical Flow Change Rate
After extracting optical flow images by selecting different consecutive frames, the moving parts are marked with a frame, the moving parts in the frame are sparsely extracted, and then the processed optical flow map is sent to the network for training. Figures 3a-3c are comparison diagrams of continuous optical flow images of 2 frames, 5 frames, and 10 frames, respectively.

![Figure 3](image)

**Figure 3.** Comparison of optical flow of consecutive frames of 2 frames, 5 frames and 10 frames.

It can be seen from figure 3 that the optical flow image after sparse extraction clearly distinguishes between normal behavior and abnormal behavior, and when continuous video frame is selected as 5, it can better distinguish abnormal behavior from normal behavior.

![Figure 4](image)

**Figure 4.** Fusion structure.

An optical flow direction change rate map can be obtained from the optical flow field, and the optical flow change rate is the ratio of the number of pixels that change to the total number of pixels in the current frame image. The rate of change in optical flow in figures 4a and 4b shows the difference between normal and abnormal actions.
3.4. Experimental Results and Analysis of Different Data Sets

Based on the industrial background, this paper uses the filtered public data sets and self-built data for experimental comparison of different algorithms. This paper uses a supervised training method. The average recognition rate is the evaluation index of the model: the proportion of videos with the same motion category and real category predicted by the model.

First, the algorithm in this paper is compared with commonly used traditional methods, and obtained the average recognition rate of the corresponding algorithm.

From table 1, we can see that the recognition rate of the traditional method in the filtered data set and the self-built data set has decreased slightly, but the overall recognition rate remains at about 80%, and then the algorithm in this paper is compared with the deep learning method experimentally.

It can be seen from table 2 that although the improved network in this paper is based on a simple 2D convolutional network structure and doesn’t have a deeper network structure, for the workers in the factory often appear abnormal actions, this method has a better recognition rate.

| Method                  | CASIA dataset | CAVIAR dataset | Self-built dataset |
|-------------------------|---------------|----------------|--------------------|
| HOG+SVM                 | 79.3          | 78.4           | 81.7              |
| HOF+SVM                 | 78.0          | 82.1           | 81.2              |
| MBH+HOF                 | 80.7          | 81.7           | 83.3              |
| HOF+MBH+SVM             | 82.2          | 79.9           | 84.1              |

| Method                  | CASIA dataset | CAVIAR dataset | Self-built dataset |
|-------------------------|---------------|----------------|--------------------|
| CNN                     | 88.4          | 85.2           | 89.7              |
| AlexNet                 | 87.3          | 86.0           | 87.6              |
| VGGNet16                | 89.2          | 87.1           | 85.7              |
| LSTM                    | 88.1          | 86.1           | 88.4              |
| CNN+LSTM                | 89.8          | 89.5           | 88.3              |
| Traditional Two-stream  | 90.3          | 91.4           | 90.8              |
| Ours                    | 94.3          | 93.9           | 93.6              |

4. Conclusion

Based on the two-stream convolutional network, this paper analyzes the two problems in the original two-stream network where the video image frame network lack of temporally relevant information in extracting still video images and the continuous optical flow image network makes it difficult to distinguish similar behavior. Convolutional network algorithm, which combines the LSTM network with the video image frame network in traditional two-stream convolution to achieve the extraction of time series information in the video image, the secondary extraction of the dense optical flow image of the video image. Pixels with small changes in flow value are sparsely extracted, and a new continuous optical flow image is input to the continuous optical flow image network for training, so as to obtain workers’ motion information. Experiment on two filtered public data sets and self-built data sets, the average recognition rate has improved by 1%-4%.

References

[1] Xiang Y, et al. 2018 Review of human abnormal action recognition based on computer vision Infrared 39 (17) 1-6.
[2] Fujiyoshi H and Lipton A J 1998 Real-time human motion analysis by image skeletonization Fourth IEEE Workshop on Applications of Computer Vision (Breckenridge) pp 19-21.
[3] Wang D, et al. 2016 Abnormal behavior recognition of fall in surveillance video. *Electronic Design Engineering* **24** (22) 1126-1222.

[4] Bobick A F and Davis J W 2002 The recognition of human movement using temporal templates. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **23** (3) 257-267.

[5] Fu Y W and Yang S P 2015 Human action recognition by extracting motion trajectories. *International Conference on Digital Image Processing* pp 9631-96311.

[6] Lecun Y, Boser B, Denker J S, et al. 1989 Backpropagation applied to handwritten zip code recognition. *Neural Computation* **1** (4) 541-551.

[7] Krizhevsky A, Sutskever I and Hinton G 2012 Image net classification with deep convolutional neural networks. *NIPS* (Curran Associates Inc.) pp 1097-1105.

[8] Simonyan K and Zisserman A 2014 Very deep convolutional networks for large-scale image recognition. *Computer Science* **13** (2) 120-131.

[9] Szegedy C, Liu W, Jia Y, et al. 2015 Going Deeper with Convolutions. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* pp 1-9.

[10] Simonyan K and Zisserman A 2014 Two-stream convolutional networks for action recognition in videos. *Proc. of Advances in Neural Information Processing Systems* pp 568-576.

[11] Yuan Z 2017 A fall detection method based on two-stream convolutional neural network. *Journal of HeNan Normal University (Natural Science Edition)* **45** (3) 102-107.

[12] Si K, et al. 2018 Research on human fall behavior using CNN and LSTM-based hybrid model. *Application Research of Computers* **39** (12) 11-15.