An improved two-stream CNN method for abnormal behavior detection

Luo Sha, Yuan Zhiwen*, Xu Kan, Zhang Jinli, Deng Honggang
State Grid Shanghai Municipal Electric Power Company, Electric Power Research Institute, Shanghai, China
yuanzhw@foxmail.com

Abstract—In recent years, abnormal behavior detection has become an active research field in computer vision and image processing. Several methods based on traditional two-stream CNN or 3D-CNN have been proposed and successfully applied in abnormal behavior detection. However, the abnormal behavior data is rarely less in the real scenario, in other words, the data is imbalanced, which may lead to overfitting problem and affect the final results. In this paper, an improved abnormal behavior detection method based on a two-stream CNN method was proposed to address the problem mentioned above. In the proposed method, DenseNet is adopted to extract both spatial and temporal features, and focal loss is employed to alleviate the influence of imbalanced data. The experiment results show that the proposed method provides a good performance in the real-world scenario.

1. INTRODUCTION
With the increment of surveillance camera numbers, it is impossible to detect abnormal behavior by humans. Video-based computer vision and image processing technology, such as abnormal behavior detection has become an active research field. By using the video-based image processing methods, the abnormal behaviors in the video clips can be detected quickly and automatically. Abnormal behavior detection methods can be roughly divided into traditional methods and deep learning methods.

Traditional abnormal behavior detection methods first extract some hand-crafted features, the behavior prediction results will be generated based on these features and trained models. Traditional abnormal behavior detection methods can only be used in the scenario with a simple environment and specific fewer behaviors. For example, Rajabi et al. used the filtering method and the gaussian mixture model to detect the moving target and human fall behavior. But it only works in a simple scenario. Chua et al. design and calculate three features to represent the human’s head, body, and leg, then detect human behaviors by calculating the similarity of each frame with the defined template. However, it will be difficult to calculate the similarity if the behaviors are complex and the background is complex. As we can see, it is difficult for the traditional abnormal behavior detection methods to deal with the high spatial complexity and large time difference in the video. Several studies have focused on deep learning-based abnormal behavior detection.

Deep learning-based abnormal behavior detection methods can be divided into two parts: 3D-CNN and two-stream CNN. Through 3D convolutional kernel, 3D-CNN can extract the motion information from the video in both spatial and temporal and achieve high-speed detection. For example, Shuiwang et al. developed a novel three-dimensional (3D) CNN model for action recognition. The model extracts features from spatial and temporal dimensions by performing 3D convolutions. It encodes...
motion information in multiple adjacent frames and generates a final feature representing human actions in real-world environments. Qiu et al. changed the inner connection of convolutional in ResNet, proposed a Pseudo-3D residual network (P3D) to detect human behaviors[6]. Besides 3D-CNN methods, Recent studies have proposed a two-stream CNN architecture that incorporates spatial and temporal networks to recognize action from videos[7,8]. Carreira J et al. introduced a new two-stream inflated 3D ConvNet (I3D) that is based on 2D ConvNet inflation, achieved good performance in two public datasets HMDB-1 and UCF-101[9]. Standard video actions are evaluated, and the results show that the architecture related above is comparable to the state-of-the-art methods.

Existing approaches indicate that the structural attribution of the deep learning model is effective in recognizing a specific object in an image or determining its spatial location as well as for behavior detection. However, in the real scenario, the abnormal behavior data is rarely less, which may lead to poor performance of the methods mentioned above, because of the overfitting problem caused by the imbalanced data. Thus, in this paper, an improved abnormal behavior detection method based on two-stream is proposed to address the problem in abnormal behavior detection. Firstly, spatial and temporal features are extracted from the original video clips by spatial stream network and temporal stream network. Then, the results are given using a softmax layer. Finally, the results from both two networks will be fused with different weights.

The organization of this paper is as follows. Section 2 gives the core idea of the proposed method. Section 3 describes the experimental results and discussions of the proposed method. Finally, the conclusion of this paper is presented in Section 4.

2. PROPOSED METHOD

2.1 Framework

To improve the performance of the abnormal behavior detection, an improved abnormal behavior detection based on the two-stream CNN method is proposed in this paper. The abnormal behavior detection framework has two parts: spatial stream network and temporal stream network. They all have the same structure with three dense blocks and two fully connected layers. The DenseNet is adopted to extract high-quality spatial features and temporal features simultaneously. The loss function of the softmax layer is modified by changing it to focal loss which can take the influence of imbalanced data into account. The prediction results of the two networks will be fused by considering the weight of the two networks at last. The structure of the proposed method is presented in Figure 1.

![Framework of the proposed method](image)

**Figure 1. Framework of the proposed method**

2.2 DenseNet

Traditional CNN calculates the output of the \( p \)th layer by applying a series of nonlinear functions \( H \) to the previous layer’s output \( x_{i-1} \), which is given as:

\[
x_i = H_i(x_{i-1})
\]

Through consecutive convolution and pooling, the network can obtain final features in the last layer. However, some details of the original data tend to disappear in the last of the network, when the network goes deeper and deeper. To improve the information flow between layers, DenseNet provides a simple connection pattern, that is every layer of the network receives all the data in the previous
layers\textsuperscript{[10]}. DenseNet is a variation of standard CNN network which can strengthen the feature propagation, and substantially reduce the number of parameters. With the special architecture, high-quality features can be extracted, and the training efficiency can also be improved. Consequently, the \( l \)-th layer receives the feature-maps of all preceding layers, \( x_0, \ldots, x_l \), as input:

\[
x_l = H_l((x_0, \ldots, x_l))
\]  

In this manner, even the last layer can access the original information of the data.

In this paper, the DenseNet in the spatial and temporal stream network has three dense blocks, each of which has six dense layers. Between the dense blocks, there are transition down layers consisting of a \( 1 \times 1 \) convolution followed by a \( 2 \times 2 \) maxpooling operation. The structure of the used DenseNet is shown in Figure 2.

![Figure 2. Structure of DenseNet](image)

### 2.3 Spatial stream network

The spatial stream network is used to extract the spatial features from the images, which are randomly sampled from the video. In the feature extracting process, the background and the pose of the detection object are vitally important. To extract more robust spatial features, the dataset should be collected in a simple environment with a fixed background. The spatial features are extracted by multiple convolutional layers and pooling layers. The first classical convolutional neural network is LeNet, which contains 5 layers. Afterward, different convolutional neural networks have been proposed, such as AlexNet with 8 layers, VGGNet with 19 layers, GoogleNet with 22 layers, and DenseNet with up to 264 layers\textsuperscript{[11]}. Previous studies have proven that with the networks going deeper, the feature extraction ability becomes stronger. However, the deep network will bring the gradient vanishing problem, and the performance of the model will decrease rapidly. To address this problem, ResNet is proposed by adding another linear function besides the nonlinear function. Based on the ResNet, DenseNet not only adds more layers but also strengthen the feature propagation by adding more input to each layer. Thus, in this paper, DenseNet is adopted to extract the spatial features.

### 2.4 Temporal stream network

Besides the spatial information, temporal information in the video is also important for the detection of abnormal behavior. The optical flow is wildly used in the image processing field for it is simple and easy to be used for describing the motion information in the video. B.K.P Horn et al. first derived the formula for optical flow calculation\textsuperscript{[12]}. However, the value of optical flow will decrease to zero which cannot input to the network. To address the problem, a linear transformation is applied to transform the optical flow to grayscale images. In this paper, DeepFlow, one of the commonly used optical flow extraction methods, is adopted to extract the optical flow from the video clips. The original image and the extracted optical flow are given in Figure 3.
2.5 Focal loss Considering imbalanced data

As we described before, in the real scenario, the abnormal behavior data is rarely less, which may cause overfitting of the models. To address the problem above, an improved loss function named focal loss is used in this paper by considering the influence of the imbalanced data. The focal loss is base on the cross-entropy (CE) loss, which is calculating by:

$$CE(p, y) = \begin{cases} -\log(p) & \text{if } y = 1 \\ -\log(1 - p) & \text{otherwise} \end{cases}$$

In the above, \( y \) means the real label of the data, and \( p \in [0,1] \) is the probability for the class with the label \( y = 1 \).

Different from the CE loss, focal loss (FL) introduces a modulating factor \((1 - p)^\gamma\) to the CE loss for addressing the influence of imbalanced data\(^{[13]}\). Which is given as:

$$FL(p_i) = -\alpha_i (1 - p_i)^\gamma \log(p_i)$$

In this formula, \( \alpha_i \) is the frequency of each label in the training set that can balances the importance of positive/negative examples, \( \gamma \) is the focusing parameter. When an instance is misclassified and \( p_i \) is small, the modulating factor is near 1 and the loss is unaffected. While \( p_i \to 1 \) the modulating factor goes to 0 and the loss for well-classified instances is down-weighted. Besides, the focusing parameter \( \gamma \) smoothly adjusts the rate at which easy instances are down-weighted. When \( \gamma = 0 \), \(-\alpha_i (1 - p_i)^\gamma\) is equal to 1, FL is equivalent to CE, and the effect of the modulating factor is likewise increased when \( \gamma \) is increased.

For better understanding, the more common format of the focal loss is given below. \( x \) means the input data, \( \text{class} \) means the label, \( j \) means the total number of the classes.

$$FL(x, \text{class}) = -\alpha_{\text{class}} \left( 1 - \frac{e^{\text{class}}}{\sum_j e^{\text{class}}} \right) \log \left( \frac{e^{\text{class}}}{\sum_j e^{\text{class}}} \right)$$

2.6 Results fusion

Both spatial stream network and temporal stream network will generate a behavior prediction results by the last softmax layer. To get the final result, a weighted sum is used to fusing the results from two networks, which is given as:

$$C_{\text{total}} = \arg \max(n_1 C_{\text{spatial}} + n_2 C_{\text{temporal}})$$

\( C_{\text{total}} \) means the final label of the instance, \( C_{\text{spatial}}, C_{\text{temporal}} \) means the output vector of two networks, \( n_1, n_2 \) means the weights of two network’s output. In this paper, the different weight combinations will be tested for getting the best performance of the abnormal behavior detection task. And the influence of different weight combinations will also be described.
3. EXPERIMENT

3.1 Experiment setup
A scheme capable of detecting and recognizing abnormal behavior was developed from ongoing behavior input form a video clip that consists of a standard RGB image frame sequence. The proposed method was trained and tested on a self-constructed dataset termed as electrical abnormal behavior dataset 5 (EABD5) collected from the power equipment laboratory. The dataset consists of 100 video clips in which each clip contains multiple moving human subjects with the follow five different behaviors: “Illegal Cutting”, “Walking”, “Carrying Equipment”, “Marking”, and “Running”. In this experiment, “Illegal Cutting” and “Running” are identified as abnormal behaviors with 5% of the total video clips, others are regarded as normal behaviors. The whole dataset was split into a training set with 80% and testing set with 20%. The resolution image frame corresponds to 480P (640×480), and the frame rate corresponds to 25 fps (frame per second).

The experiment was carried out on the ubuntu platform with Core I7-9700K CPU, NVIDIA GeForce GTX 1070 (8 GB). And the model is written by using a deep learning framework named PyTorch 1.4.0 based on python 3.6. The epoch of the experiment is 100 and the batch size is 5. To evaluate the proposed method, different abnormal behavior detection methods are selected as the compared methods, such as the Optical flow method and Dynamic particle field method as the traditional methods, 3D-CNN, traditional two-stream method, and two-stream method with LSTM as the deep learning method. Finally, Area Under Curve (AUC) and accuracy are selected to evaluate these methods.

3.2 Experiment results and discussion

1) Evaluation of different methods
The evaluation results of the different methods on the EABD5 dataset are list in Table 1. As we can see, the accuracy of the traditional methods is 89.36 and 90.59, which are less than the deep learning methods. It means the deep learning methods are better than the traditional methods. As for the deep learning methods, the accuracy of 3D-CNN is 91.63, because of the overfitting problem caused by a small training set. For the traditional two-stream method, the accuracy is 94.16. the reason is the network cannot extract high-quality features for abnormal behavior detection. The accuracy of two-stream + LSTM reaches 95.75, which is better than other methods expect the proposed method because the LSTM can extract and remember some information from the video clips in the temporal aspect. The proposed method gets the highest AUC among the methods, which shows the ability of the proposed method to address the influence of the imbalanced data.

| Methods                        | AUC   | Accuracy |
|--------------------------------|-------|----------|
| Optical flow method            | 75.43 | 89.36    |
| Dynamic particle field method  | 79.52 | 90.59    |
| 3D-CNN                         | 86.38 | 91.63    |
| Traditional two-stream method  | 88.57 | 94.16    |
| Two-stream + LSTM              | 90.42 | 95.75    |
| Proposed method                | **96.43** | **97.53** |

2) Evaluation of different combinations of weights
Both spatial stream network and temporal stream network can give the results of behavior detection. For getting better performance, the results will be fused with different weights. The fused results are listed in Table 2. As shown in Table 2, when the weight of the temporal stream network increasing, the accuracy is also increasing. In other words, the temporal stream network contributes more benefits to the final results. When the weight of the temporal stream network reaches 5, the model gets the best performance, which is 97.53 shown in the table.
TABLE II. Auc and accuracy of different combinations of weights

| $n_1 : n_2$ | AUC   | Accuracy |
|------------|-------|----------|
| 1:1        | 92.34 | 95.35    |
| 1:2        | 94.67 | 96.46    |
| 1:5        | 96.43 | 97.53    |
| 5:1        | 94.63 | 95.63    |
| 2:1        | 91.71 | 94.52    |

3) Evaluation of the training error

Finally, the training error of the proposed is analyzed by plotting the training error of each epoch. As described before, the experiment has a total of 100 epochs. As is shown in Figure 4, the error becomes decrease quickly when the epoch number reaches 10, which means it is the efficiency of the proposed method to meet the requirement of abnormal behavior detection task.

![Figure 4 Training error](image)

4. CONCLUSION

In this paper, an improved abnormal behavior detection method is proposed to detect abnormal behaviors. Firstly, DenseNet is adopted to extract high-quality spatial and temporal features. Then a modified loss function of the last softmax layer is adopted to eliminate the influence of imbalanced data. Finally, the results will be fused with different combinations of weights. According to the final experiment results, the proposed abnormal behavior detection method is superiority and possibility than the traditional methods and other deep learning-based methods.

ACKNOWLEDGMENT

This work was financially supported by the State Grid Shanghai Municipal Electric Power Company Research and Development Project (52094020001A).

REFERENCES

[1] Popoola O P, Wang K. Video-based abnormal human behavior recognition—A review[J]. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 2012, 42(6): 865-878.
[2] Rajabi H, Nahvi M. An intelligent video surveillance system for fall and anesthesia detection for elderly and patients[C]//2015 2nd International Conference on Pattern Recognition and Image Analysis (IPRIA). IEEE, 2015: 1-6.
[3] Chua J L, Chang Y C, Lim W K. A simple vision-based fall detection technique for indoor video surveillance[J]. Signal, Image, and Video Processing, 2015, 9(3): 623-633.
[4] Dong B, Wang X. Comparison deep learning method to traditional methods using for network intrusion detection[C]//2016 8th IEEE International Conference on Communication Software and Networks (ICCSN). IEEE, 2016: 581-585.
[5] Ji S, Xu W, Yang M, et al. 3D convolutional neural networks for human action recognition[J]. IEEE transactions on pattern analysis and machine intelligence, 2012, 35(1): 221-231.

[6] Qiu Z, Yao T, Mei T. Learning Spatio-temporal representation with pseudo-3d residual networks[C]//proceedings of the IEEE International Conference on Computer Vision. 2017: 5533-5541.

[7] Simonyan K, Zisserman A. Two-stream convolutional networks for action recognition in videos[C]//Advances in neural information processing systems. 2014: 568-576.

[8] Han Y, Zhang P, Zhuo T, et al. Going deeper with two-stream ConvNets for action recognition in video surveillance[J]. Pattern Recognition Letters, 2018, 107(7): 83-90.

[9] Carreira J, Zisserman A. Quo vadis, action recognition? a new model and the kinetics dataset[C]//proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 6299-6308.

[10] Huang G, Liu Z, Van Der Maaten L, et al. Densely connected convolutional networks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 4700-4708.

[11] Yoo H J. Deep convolution neural networks in computer vision: a review[J]. IEIE Transactions on Smart Processing & Computing, 2015, 4(1): 35-43.

[12] Horn B K P, Schunck B G. Determining optical flow[C]//Techniques and Applications of Image Understanding. International Society for Optics and Photonics, 1981, 281: 319-331.

[13] Lin T Y, Goyal P, Girshick R, et al. Focal loss for dense object detection[C]//Proceedings of the IEEE international conference on computer vision. 2017: 2980-2988.