Skeleton-Based Pedestrian Abnormal Behavior Detection with Spatio-Temporal Model in Public Places

Xiaojun Wu¹,*, Wuyao Zhao¹ and Sheng Yuan¹
¹School of Software Engineering, Xi’an Jiaotong University, Xi’an, 710049, China.
*Email:xiaojunw@xjtu.edu.cn

Abstract. In computer vision, there is growing interest in the recognition of pedestrian abnormal behaviors. The abnormal behavior of a person could be the sign of some dangerous activities. However, it’s still challenging to extract the discriminative spatial and temporal features effectively faced with video data. In this paper, we propose skeleton-based pedestrian abnormal behavior detection models. The base model is consisting of ResNet as a spatial feature extractor, LSTM as a global temporal feature extractor, and the ResNet network that use the dual-stream network to extract local temporal features. The proposed model is an improvement of all ResNet into Conv1x1_ResNet, and added a layer of Conv1x1_ResNet after dual-stream Conv1x1_ResNet to extract more accurate global space features. The proposed model achieved the highest accuracy of 89.29%, and the averaged get batch time is 0.3399 ms. The base model achieved 88.12% accuracy, and the averaged get batch time is 0.3174 ms less than the time taken by other models. Both models reach speed of 80 frames/sec. Compared with the models made in previous work, the base model has the shortest training time, and the proposed model provides the highest accuracy in the field of pedestrian detection.

1. Introduction
In terms of public safety, intelligent video surveillance has attracted more people's attention in recent years [1]. Pedestrian behavior detection plays an important role in various violations, such as pedestrians running, fighting, throwing punches and other abnormal events. Although abnormal behavior detection has been successfully applied to some places, it still faces a series of challenges for public places. First, it is difficult to define the boundary area of normal behavior in the video. The boundary between normal behavior and abnormal behavior is usually ambiguous. Second, the definition of the type of abnormal behavior is very subjective, and along with different scenarios. Third, the data used may contain noise, interfering with the results of the experiment. For the above problems, most of them should have specific application scenarios, and different methods and angles should be used to analyze different actions and appearances.

Recently, deep learning is now very popular and has been successfully applied to various machine learning tasks. By using the skeletal features [2], the display utilizes the general structure of the surveillance video, since the person and the objects attached to it form a static background. Some studies have shown that if a pre-training model[3] sees a particular object that will later be used in a transfer learning task, the performance of the target task will be better [4]. Pre-training and fine-tuning are very popular techniques in deep learning. In many cases, pre-trained ImageNet is selected to initialize the model. Studies have shown that in addition to adding LSTM after ResNet, and local temporal features extracted by optical flow are important [5]. The advantages of optical flow are its
It does not attempt to learn spatio-temporal features in a smaller period of time, but considers several methods to aggregate powerful CNN features over a longer period of time including videos of recurrent neural networks. Owing to the problem of vanishing and exploding gradients, standard recursive networks are difficult to learn on long sequences. However, LSTM [7] uses memory cell to better discover long-term temporal relationships.

The main contributions of this paper are summarized as follows:

1. Combining the KTH dataset and the Weizman dataset into one large dataset. Then using data preprocessing to depict the skeletal features of the pedestrian.
2. We proposed the base model and the proposed model. Since our dataset is small, it needs to use transfer learning to achieve high precision. Two models use Darknet19(on the ImageNet dataset) as a pre-trained model. Two models use a dual-stream network to extract local temporal features. The proposed model uses ResNet's improved Conv1x1_ResNet network to extract spatial features.
3. The experimental results of pedestrian behavior detection show that the proposed model achieves the highest accuracy, and the basic model converges fastest and has the shortest training time. Both models are comparable in real-time, speed, and accuracy with previous work.

2. Related research

The earliest research on human motion recognition can be traced back to the 1970s, when psychologists did the following experiments. People in dark environments installed light points on the joints, capturing the movements of these bright spots on the human body. Motion information is captured. Many research scholars have done a lot of detailed research in this field. For example, representing a video [8-9]. Some researchers analyzed the behavior of pedestrians and established global models to detect abnormal events, such as social power models [10] and interactive energy potential [11], but the above methods rely on complex manual functions and have great limitations.

Other studies have concentrated on the extraction and analysis of targeted and carefully selected descriptors, such as trajectories [12-13], histogram light flows [14-15], directional gradient histograms [16-17], and spatiotemporal directional energy [18]. In a particular use case, trajectory analysis is effective as a traffic monitoring or illegal exclusion zone. However, this approach requires the use of accurate pedestrian detection and tracking algorithms upstream, which is too complex and unreliable in scenarios where there are many people.

Automatic extraction of spatiotemporal features can use deep learning methods. The architecture[19] in includes a large number of convolutional layers, followed by a maximum pooling operation for extracting discriminative features and a convolution length memory for encoding frame-level changes. This feature describes the existence of the video Violent scene. The end-to-end spatiotemporal attention model [20] learns to selectively focus on the discriminative joints of each frame in the input frame and pay different levels of attention to the output of different frames. It does not try to learn spatiotemporal features over a small period of time, but considers several methods to extract features from CNN [21], which are capable of learning over a long time including videos of recurrent neural networks.

3. Our two models

So as to extract the features of the discriminative spatiotemporal domain automatically in the video, and the network structure proposed is based on transfer learning, ResNet, Conv1x1_ResNet, LSTM and dual-stream network. In this paper, the base model proposed the architecture is shown on the Figure 1(a), and the proposed model is shown on the Figure 1(b).
3.1. Extraction of spatial features

The lack of richness of large video dataset is a problem. Although the KTH dataset and the Weizman dataset provide 683 videos, most categories are about pedestrian behavior. Pre-training Darknet19 as a spatial feature extractor. Darknet19 is used to extract the characteristics of pedestrian space structure. The Darknet19 classification training uses SGD and momentum, the iterative update formula of momentum algorithm is equation 1.

\[
\begin{align*}
  w &= \beta v + (1 - \beta) dw \\
  w &= w - \alpha w
\end{align*}
\]

(1)

where \( dw \) is the calculated original gradient, and \( v \) is the gradient calculated by the exponential weighted average. The typical value of \( \beta \) is 0.9, and \( \alpha \) is the learning rate.

3.2. Extraction of temporal features

It is assumed that the effect of the optical flow is simulated by taking two frames in a video as input, and the Darknet19 processes two input frames. The effect of two parallel processed network analog optical flow methods is named 2POF (Two Parallel Optical Flow). If dual-stream network are ResNet, it is called 2POF_ResNet. The two frame outputs from the bottom of the pre-training model fed into a dual-stream network. In the base model, since the output from the Darknet19 is considered a low-level feature, 2POF_ResNet should learn the local motion features as well as the appearance invariance features by comparing the both frame feature maps. The two frame outputs from the top layer of the pre-training network are also merged and fed into 2POF_ResNet to compare the advanced functions of the two frames. The LeakyReLU excitation function is used in the convolutional network to extract nonlinear features. In the process of neural network training, because the learning ability is strong, and using dropout operation to prevent overfitting.

Using LSTM to extract global time domain features and long-term dynamic information, preventing overfitting using L2 regularization as shown in equation 2. The output of the LSTM cell use the fully connected dichotomy to output the detection results (abnormal and normal).

\[
y = f_{LSTM} + \alpha \| w \|_2^2
\]

(2)

Where output \( y \) refers to the sum of squares of each element in the weight vector \( w \) and square root, and a coefficient \( \alpha \) is added before the regularization term. \( f_{LSTM} \) denotes the LSTM cell function.
3.3. Description of input and output

The working steps of the base model and the proposed model are given in algorithm 1. UNROLLED_SIZE: the number of frames in the video. In all models, it split the video data of the KTH dataset and the Weizman dataset into 26 frames to generate 2570 videos. Pedestrian walking is normal behavior, and all others are abnormal behaviors, such as jogging, running, boxing, clapping, jumping, bending, one-legged jumping and waving. Among them, the normal data of pedestrians has 1289 videos, and the pedestrian abnormal behavior has 1281 videos. We randomly divided data into verification set, test set and training set according to the ratio of 1:4:5.

Algorithm 1 the base model and the proposed model

| Step 1: Data preprocessing |
|----------------------------|
| Use AlphaPose for data preprocessing to depict the skeletal features of the pedestrian. |

| Step 2: Pre-train Darknet19 |
|-----------------------------|
| Read sequence of frames in 4d tensor [BATCH_SIZE*UNROLLED_SIZE, w, h, c]. Apply pre-trained Darknet19 for frames to extract spatial features. |

| Step 3: Establish the deep neural network |
|------------------------------------------|
| if the base model |
| Apply 2POF_\text{Conv1x1}\_ResNet to capture local features. |
| end if |
| else if the proposed model |
| Apply 2POF_\text{Conv1x1}\_ResNet to capture local features, and then using a layer of \text{Conv1x1}_\text{ResNet} to extract more accurate global space features. |
| end else |

| Step 4: Capture dynamic Information |
|-------------------------------------|
| Apply LSTM, the input shape is: [BATCH_SIZE, UNROLLED_SIZE, w, h, c]. The hidden layer of the LSTM is computed as follows: |
| $i_t = \sigma(w_{ii}x_t + w_{ih}h_{t-1} + w_{ic}c_{t-1} + b_i)$ |
| $f_t = \sigma(w_{ff}x_t + w_{fh}h_{t-1} + w_{fc}c_{t-1} + b_f)$ |
| $c_t = f_t c_{t-1} + i_t \tanh(w_{cx}x_t + w_{ch}h_{t-1} + b_c)$ |
| $o_t = \sigma(w_{oi}x_t + w_{oh}h_{t-1} + w_{oc}c_t + b_o)$ |
| $h_t = o_t \tanh(c_t)$ |
| where $\sigma$ is the logistic sigmoid function, and $i_t$, $f_t$, $o_t$, and $c_t$ are respectively the input gate, forget gate, output gate, and cell activation vectors at time $t$. |

| Step 5: Output the results |
|-----------------------------|
| After training the neural network, output the prediction. |

3.4. The difference between the proposed model and the base model

In the proposed model, improving the bottleneck base module in the ResNet network. A 1x1 convolution kernel is added to the left of the bottleneck. The improved ResNet network is called Conv1x1_ResNet. Conv1x1_ResNet introduces more non-linearities to deepen the network without increasing the receiving field, which increases the expressive power of the neural network. 2POF_\text{Conv1x1}_\text{ResNet} is used to extract local spatiotemporal features, and its output is connected to a layer of Conv1x1_ResNet to extract global spatial information. The proposed model not only improves the ResNet module, but also adds the depth of the network, which improves the accuracy and reduces the loss, while increasing the training time and calculation amount. The basic model not only accelerates the convergence speed and reduces the training time, but also reduces the model volume, and at the same time can accelerate the prediction time. The basic model has the fastest convergence and the shortest training time. The proposed model achieved the highest accuracy and the lowest loss.
4. Experimental results and discussion

Real-time video processing speed of two models on Nvidia RTX 2080 computer is 80fps. The experimental results of both models are shown in Figure 2. The first row to the third row are the abnormal behaviors of pedestrians, and the fourth row is the normal behavior of pedestrians. The red box around the video is an abnormal warning, and the green box is normal. The behavior of the pedestrian is normal before the start of each abnormal behavior, so the normal state is detected at the beginning. Except for the last two lines in Table 1, we have done a series of previous works. Since video accuracy is higher than frame accuracy in Table 1, it can be proved that the dual-stream network can accurately capture the relationship between multiple frames. We observe that the effect of the ResNet network is better than some CNN networks. The proposed model has a significant improvement in accuracy and loss compared to the base models and previous work models. The improved Conv1x1_ResNet has better effect and can extract the features of temporal and spatial.

![Figure 2. Pedestrian behavior detection results in public places.](image)

**Table 1. The results of comparison.**

| Method                        | Frame accuracy (%) | Video accuracy (%) | Loss   | Averaged getbatch time (ms) |
|-------------------------------|--------------------|--------------------|--------|----------------------------|
| Darknet19+LSTM                | 85.01              | 85.49              | 0.3509 | 5.4316                     |
| Darknet19+Conv+LSTM           | 86.33              | 86.47              | 0.3456 | 25.9196                    |
| Darknet19+2POF_CNN+LSTM       | 85.78              | 86.95              | 0.3236 | 0.3253                     |
| Darknet19+2POF_Conv1x1_Resnet+LSTM | 84.49             | 86.37              | 0.3251 | 0.3377                     |
| The base model                | 86.32              | 88.12              | 0.2980 | **0.3174**                 |
| The proposed model            | 86.82              | **89.29**          | 0.2869 | 0.3399                     |

As shown in Figure 3(a), the training set, validation set and test set have approximately the same distribution, and video accuracy rate gradually increases and stabilizes in the basic model. The generalization loss of the network is verified on the validation set. Until the network obtains a lower generalization loss on the validation set, the complete training process ends at a certain moment. The loss of the training set, test set, and verification set that calculated using the cross entropy loss function on the base model have been reduced and gradually stabilized as shown in Figure 3(b).
After training the proposed model, the video accuracy is shown in Figure 4(a). The loss is calculated using a cross-entropy function on the proposed model is shown in Figure 4(b). The distribution of the three curves in the accuracy and loss graphs in Figure 4 is more compact compared to Figure 3, indicating that the model trained by the proposed model has better fitting data and higher accuracy. On averaged getbatch time in table 1, we can see that the base model converges faster and the training time is shorter compared to the proposed model.

5. Conclusion

This work shows that using transfer learning, ResNet, Conv1x1_ResNet, LSTM and dual-stream network are the best way to achieve accurate, reliable, and fast models in pedestrian detection with limited datasets and computational resources. The base model and the proposed model can capture discriminative features in spatial configuration and temporal dynamics. In the future, we will improve the model by analyzing error samples and consider more contextual information, such as interactions, to help with identify behaviors. Furthermore, our work is to extend single pedestrian detection to multi-pedestrian behavior recognition, and we will do some special work to explore and create well-balanced new models with different video sources for pedestrian detection and to provide more categories of resources to detect pedestrian behavior itself.

References

[1] G. Tripathi, K. Singh,K. D. Vishwakarma, et al., “Convolutional neural networks for crowd behaviour analysis: a survey,” 2019, The Visual Computer, pp. 753-776.

[2] X. Nie, J. Feng, Y. Zuo, et al., "Human Pose Estimation with Parsing Induced Learner," 2018, Computer vision and pattern recognition, pp. 2100-2108.
[3] J. Redmon, and A. Farhadi, "YOLO9000: Better, Faster, Stronger," 2017, IEEE Conference on Computer Vision and Pattern Recognition, pp. 6517-6525.
[4] A. Karpathy, G. Toderici, S. Shetty, et al., "Large-Scale Video Classification with Convolutional Neural Networks," 2014, IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1725-1732.
[5] H. Y. Ng, M. Hausknecht, S. Vijayanarasimhan, et al., "Beyond Short Snippets: Deep Networks for Video Classification," 2015, Computer Vision and Pattern Recognition, pp. 4694-4702.
[6] L. Sevilla-Lara, Y. Liao, F. Guney, et al., "On the Integration of Optical Flow and Action Recognition," 2018, German Conference on Pattern Recognition, pp. 281-297.
[7] S. Hochreiter, and J. Schmidhuber, "Long short-term memory," 1997, Neural Computing, pp. 1735–1780.
[8] Y. Zhang, L. Qin, H. Yao, P. Xu, et al., "Beyond particle flow: Bag of Trajectory Graphs for dense crowd event recognition," 2014, IEEE International Conference on Image Processing, pp. 3572-3576.
[9] T. Wang, M. Qiao, Y. Deng, et al., "Abnormal event detection based on analysis of movement information of video sequence," 2018, Optik, pp. 50-60.
[10] R. Mehran, A. Oyama, and M. Shah, "Abnormal crowd behavior detection using social force model," 2009, IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), Miami, Florida, USA, pp. 935-942.
[11] X. Cui, Q. Liu, M. Gao, et al., "Abnormal detection using interaction energy potentials," 2011, Computer Vision and Pattern Recognition, pp. 3161-3167.
[12] S. Cosar, G. Donatiello, V. Bogorny, et al., "Toward abnormal trajectory and event detection in video surveillance," 2017, IEEE Transactions on Circuits and Systems for Video Technology, vol. 27, no. 3, pp. 683–695.
[13] T. V. Duong, H. H. Bui, D. Q. Phung, et al., "Activity recognition and abnormality detection with the switching hidden semi-markov model," 2005, IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), vol. 1, pp. 838–845.
[14] T. Wang, and H. Snoussi, "Detection of abnormal visual events via global optical flow orientation histogram," 2014, IEEE Transactions on Information Forensics and Security, vol. 9, no. 6, pp. 988–998.
[15] T. Wang, M. Qiao, A. Zhu, et al., "Abnormal event detection via covariance matrix for optical flow based feature," 2018, Multimedia Tools and Applications, vol. 77, no. 13, pp. 17375–17395.
[16] B. Zhao, L. Fei-Fei, and E. P. Xing, "Online detection of unusual events in videos via dynamic sparse coding," 2011, IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3313–3320.
[17] M. Javan Roshtkhari, and M. D. Levine, "Online dominant and anomalous behavior detection in videos," 2013, Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2611–2618.
[18] A. Zaharescu, and R. Wildes, "Anomalous behaviour detection using spatiotemporal oriented energies, subset inclusion histogram comparison and event-driven processing," 2010, European Conference on Computer Vision, pp. 563–576.
[19] S. Sudhakaran, and O. Lanz, "Learning to Detect Violent Videos using Convolutional Long Short-Term Memory," 2017, 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), pp. 1-6.
[20] S. Song, C. Lan, J. Xing, et al., "An End-to-End Spatio-Temporal Attention Model for Human Action Recognition from Skeleton Data," 2016, National Conference on Artificial Intelligence, pp. 4263-4270.
[21] S. Bouindour, R. Hu, H. Snoussi, et al., "Enhanced Convolutional Neural Network for Abnormal Event Detection in Video Streams," 2019, International Conference on Artificial Intelligence, pp. 172-178.