An elevator passenger behavior recognition method based on two-stream convolution neural network

Shengnan Lan¹, Saihua Jiang¹* and Gang Li²
1 Institute of Safety Science and Engineering, School of Mechanical and Automotive Engineering, South China University of Technology, Wushan Road 381, Guangzhou, 510641, P.R.China
2 Elevator IoT Lab, Guangzhou Academy of Special Equipment Inspection & Testing, Guangzhou, Guangdong, 510180, P. R. China
*Corresponding author’s e-mail: meshjiang@scut.edu.cn

Abstract. In order to reduce elevator accidents caused by passengers’ incorrect use, a method for elevator passenger dangerous behaviour recognition based on two-stream convolutional neural network is proposed after analysing existing elevator passenger behavior recognition methods. A two-stream convolution neural network model framework and its parameters for elevator passenger behavior recognition are also given. Herein, the appearance features of human behavior are extracted from space domain, and the motion features of human behavior are extracted from time domain. Finally, the softmax outputs of the two-stream are merged with the liner weighting to realize human behavior recognition. The model is trained and evaluated on the video dataset of elevator passenger behavior. The results show that this method can be used to identify the unsafe behavior of elevator passengers, and the average recognition rate is 96.87%.

1. Introduction
Elevators are the facilities most closely related to public daily life. On March 15, 2021, the State Administration of Market Regulation (SAMR) of China announced the national safety situation of special equipment in 2020. Data show that by the end of 2020, the number of elevators in China has reached 7.8655 million [1]. With the increase in the number of elevators, the relationship between elevator safety and public safety has been widely concerned. Especially when an elevator accident occurs in a crowded public place, it is easy to cause many casualties [2]. According to statistics [1], unsafe behavior of passengers is one of the key factors causing elevator accidents. For example, incorrect self-rescue behavior taken by passengers when they are trapped in the elevator[3]. These elevator accidents can be reduced by using video monitoring technology to monitor the passenger behavior in the elevator car. Thus, many cameras are installed in elevator cars. However, this traditional monitoring method cannot automatically identify the abnormal events in the elevator [4]. Therefore, how to automatically identify and alarm the unsafe behavior of passengers by elevator surveillance video has become a concern of researchers.

Several methods of abnormal behavior detection in elevator cars have been proposed in recent years. An abnormal behavior detection model based on corner kinetic energy was proposed in [5], it designed for robbery and violence behavior in the elevator car. According to the number of elevator passengers, Zhu et al.[6-7] propose three models and algorithms to detect the falling or crouching
behavior of one person, violent behavior between two-person, and group panic behavior, separately. A two-dimensional pose estimation method based on the Location-Focus Field (PAFs) was proposed in [8], which can detect the behavior of force the door open. However, these methods depend on artificial features, so it is difficult to automatically extract deep information from the image. And an algorithm in these methods can only detect one unsafe behavior.

Deep learning technology shows excellent performance of automatic feature extraction in image classification and detection [9-10]. Behavior recognition in video can be regarded as an image classification problem that changes with time. Therefore, deep learning methods in image recognition have been widely used in behavior recognition in video [11], among which the convolutional neural network (CNN) is the most widely used. Alexnet [12], GoogLeNet [13], VGGnet [14] and other classic CNN architectures have not only made breakthroughs in image processing, but also achieved remarkable results in video processing. CNN is applied to driver behavior recognition in [15], and realized the recognition of 10 types of driving behaviors, with an average recognition rate of 97.13%. However, deep learning technology has not been applied to elevator passenger behavior recognition.

Herein, an elevator passenger behavior recognition model based on two-stream CNN is proposed. And the model is trained and evaluated on the video dataset of elevator passenger behavior. This work is helpful for prevention of elevator accidents caused by passengers’ unsafe behavior and intelligent video monitoring of elevators.

2. Principle

2.1. Convolutional neural network

Convolutional neural network (CNN) is mainly composed of convolution layers, pooling layers and dense layers. It can imitate the process of image processing and recognition in the visual cortex, so it has a wide range of applications in the field of image recognition. CNN automatically extracts image features through convolution and pooling operations, and combines feature extraction and classification output into a whole, so as to obtain higher recognition efficiency and better performance.

Figure 1 shows a simple convolutional neural network. It consists of two convolution layers, two pooling layers and two dense layers. Convolution layer and pooling layer are the key to automatic feature extraction.

2.1.1. Convolution layer. Convolution layer use several different convolution kernels to convolute the input of this convolution layer(Figure 2), then add bias, and then several feature maps are obtained by activation function. The calculation process of convolution layer can be expressed as follows:

\[ x'_j = f(\sum_{i \in M_j} x_{i-1} * W_{ij} + b'_j) \]  

(1)
Where $x_j^l$ is the $j$ feature map of $l$ layer, $W_{ij}^l$ is the convolution kernel between the $i$ feature map of $l-1$ layer and the $j$ feature map of $l$ layer, $b_j^l$ is the bias of $i$ feature map of $l$ layer, $M_j$ is the set of input feature map of $j$ feature map, * denotes convolution operation (as shown in Figure 2), $f(\cdot)$ is an activation function.

![Figure 2. Schematic diagram of convolution.](image)

2.1.2. Pooling layer. Pooling layer, also known as the lower sampling layer, is to sample the feature map of the upper layer to achieve dimensionality reduction, so as to reduce the calculation of model. Pooling operation is a special convolution operation. Common pooling methods are mean-pooling and max-pooling[16], as shown in Figure 3. Mean-pooling keeps the background information as much as possible by averaging the pixel values in the neighborhood. Max-pooling is to take the maximum value of pixels in the neighborhood, which can better retain the texture information of the image. What we need is the information of passenger behavior in elevator, not the background information of the car. Therefore, the pooling layer of the model constructed in this work is max-pooling.

![Figure 3. Schematic diagram of pooling. (a) Mean-pooling. (b) Max-pooling.](image)

2.1.3. Dense layer. Dense layer is also known as the fully connected layer. Its function is map the features extracted by the network to the label space, thus the output layer could output the recognition results conveniently. Dense layer transforms several two-dimensional feature maps into a one-dimensional vector, which reduces the influence of position of eigenvalue on classification result. To solve the multi-classification problem, dense layer is usually used in conjunction with softmax layer. Because the size of each element in the vector obtained by dense layer is unlimited, and softmax layer can normalize it to get the probability that the sample is identified as each category.
2.2. Two-stream convolutional neural network
In order to make better use of the information contained in temporal component of video, Simonyan et al.[16] proposed a two-stream convolutional neural network framework in 2014. Two-stream convolutional neural network consists of spatial stream CNN and temporal stream CNN, as shown in Figure 4. Spatial stream CNN extracts action appearance information in the spatial domain from a single frame of video to identify action. Temporal stream CNN obtains the motion information of human behavior by calculating the optical flow between continuous multi-frames of video, and then realizes behavior recognition. Finally, the predicted scores of each stream are combined by linear weighting.

![Figure 4. Two-stream convolutional neural network architecture.](image)

3. Model and dataset

3.1. Elevator passenger behavior recognition model
In order to make full use of the appearance information and motion information in the elevator monitoring video, a two-stream convolutional neural network model for elevator passenger behavior recognition is proposed. The architecture of this model is shown in Figure 5. It corresponds to original two-stream convolutional networks model architecture of [16].

![Figure 5. Architecture of elevator passenger behavior recognition model.](image)

The difference is that the original two-stream convolutional networks model has the same parameters on the two stream networks, while the parameters of our model is based on VGG-M-2048 model[17] and Flow_Net model[18], as shown in Table 1. In particular, activation function of the convolution layer in this model is sigmoid function, loss function is cross entropy loss function.
Table 1. Parameters of elevator passenger behavior recognition model.

| Layer | Spatial stream CNN | Temporal stream CNN |
|-------|--------------------|---------------------|
|       | Kernel | Stride | Pooling | Layer | Kernel | Stride | Pooling |
| Conv1 | 7×7×96 | 2×2    |         | Conv1 | 7×7×96 | 2×2    |         |
| Conv2 | 5×5×256 | 2×2   |         | Conv2 | 5×5×384 | 2×2   | 3×3     |
| Conv3 | 3×3×512 | 1     |         | Conv3 | 3×3×512 | 1     |         |
| Conv4 | 3×3×512 | 1     |         | Conv4 | 3×3×512 | 1     |         |
| Conv5 | 3×3×512 | 1     | 2×2     | Conv5 | 3×3×384 | 1     | 3×3     |
|       | Full6-4096 |       |         | Full6-4096 |       |         |
|       | Full7-2048 |       |         | Full7-2048 |       |         |
|       | Softmax   |       |         | Softmax   |       |         |

3.2. Unsafe behavior of elevator passengers

According to our statistics, 88.18% of elevator accidents in China from 2002 to 2019 are related to people's unsafe behavior, as shown in Figure 6. In addition, when the elevator fails, 78.92% of the accidents are related to people's unsafe behavior. Figure 7 shows that 61.64% of these accidents are related to elevator passengers’ unsafe behavior. Therefore, we summarize the unsafe behavior of elevator passengers involved in these accidents, as shown in Table 2.

![Figure 6: Causes of elevator accidents in China from 2002 to 2019.](image1)

![Figure 7: Distribution of unsafe behavior of different persons when elevator failure.](image2)
Table 2. Parameters of elevator passenger behavior recognition model.

| Item | Unsafe behavior | Scene |
|------|-----------------|-------|
| 1    | Kick the elevator door. | Use/Self-rescue stage |
| 2    | Force elevator door open. | Use/Self-rescue stage |
| 3    | Block the closure of elevator door with objects or limbs. | Use stage |
| 4    | Fell down in elevator car. | Use/Waiting for rescue stage |
| 5    | Fight in elevator car. | Use stage |

These unsafe behaviors in Table 2 can lead to elevator failure or accidents. In order to prevent the occurrence of elevator accidents caused by unsafe behavior of passengers, the model proposed in Section 3.1 is utilized to realize the automatic identification of passengers’ unsafe behavior in elevator car.

3.3. Experimental dataset

There is no standard elevator passenger behavior video dataset due to elevator monitoring video involving the privacy of passengers. Therefore, we recruited 10 volunteers to imitate various passenger behaviors in elevator, and collected these video for experiment. Figure 8 shows an example image of various behaviors in elevator.

Figure 8. Examples of various behaviors in elevator. (a) One person takes elevator normally. (b) Two people take elevator normally. (c) More than two people take elevator normally. (d) A passenger failing down. (e) Violence between two people. (f) Violence involving more than two people. (g) Force the door open. (h) Kick the elevator door. (i) Block the closing of elevator door.
3.4. Preprocessing of dataset

Firstly, video is divided into several single frame RGB images by extracting one frame per 10 frames, and 18000 single frame RGB images with a size of 1280 × 720 are obtained. However, an image that is too large is not conducive to training of network. In order to adapt to the input of network, we compress the image to 224 × 224.

The input of temporal stream CNN is optical flow image, so Lucas-Kanada algorithm[19] is used to calculate the optical flow. In this work, Lucas-Kanada algorithm is implemented by OpenCV. At the same time, each optical flow image has accumulated 10 frames of optical flow changes to keep the same amount of data with the spatial stream CNN.

4. Results and discussion

Elevator passenger behavior recognition model is trained and tested on the dataset of elevator passenger behavior by Python 3.8 in Anaconda 3. In this work, 80% of RGB images and optical flow images are used for model training and validation, and the remaining 20% are used for testing. During the model training, 20% of the data is used as the validation set to estimate whether the model is over fitting. The learning rate is 0.8, batch size is 32, and epoch is 100. The convergence curve of loss function is shown in Figure 9, and the accuracy curve is shown in Figure 10.

![Figure 9. Loss function convergence curve.](image)

![Figure 10. Accuracy curve.](image)

Figure 9 shows that the convergence curve of the loss function decreases rapidly at the beginning of training, which indicates that the learning rate we set is appropriate, and it is in the process of gradient descent. With the training, the convergence curve of loss function tends to be stable (the loss function gradually approaches 0), which indicates that the model converges stably in the process of training.

Figure 10 shows that with the increase of epoch, the accuracy curve of the model on training set and validation set gradually increases and tends to be stable (the accuracy rate gradually approaches 1), indicating that the generalization ability of our model is good.

When the epoch is lower than 40, the value of loss function in training set is greater than that in validation set(Figure 9), and the accuracy of training set is less than that of validation set(Figure 10). It shows that model is under fitting and needs to increase the training epoch. With the increase of epoch, the loss value decreases, and the accuracy increases. It shows the model fits well, and there is no over fitting.

We tested the model on the test set, and confusion matrix of test results is shown in Figure 11. The results show that the elevator passenger behavior recognition model can effectively identify the behavior of elevator passenger, and accuracy is 97.22%. According to the confusion matrix, we can calculate precision, recall, and F1 value of each behavior class, and macro precision, macro recall, and macro F1 value for all samples on test set, as shown in Table 3. The results show that the average recognition rate of the model is 96.87%.
Figure 11. Elevator passenger behavior recognition results (confusion Matrix).

Table 3. Elevator passenger behavior recognition results.

| Class | Precision % | Recall % | F1 Score % |
|-------|-------------|----------|------------|
| 1     | 93.27       | 98.74    | 95.93      |
| 2     | 98.15       | 98.15    | 98.15      |
| 3     | 99.15       | 98.32    | 98.73      |
| 4     | 95.83       | 96.28    | 96.05      |
| 5     | 95.85       | 98.3     | 97.06      |
| 6     | 97.67       | 98.13    | 97.90      |
| 7     | 99.61       | 94.8     | 97.15      |
| 8     | 97.14       | 95.44    | 96.28      |
| 9     | 100.00      | 93.67    | 96.73      |
| Macro average | 97.41 | 96.87 | 97.14 |

Table 3 shows that recognition rate of Class-1 (one person takes elevator normally) is highest, reaching 98.74%, while Class-9 (block the closing of elevator door) has lowest recognition rate with 93.67%. Class-9 is mainly incorrectly identified for Class-1 and Class-8 (kick the elevator door), and the error recognition rate is 3.62% and 1.81%, respectively (shown in Figure 11). After checking the
video dataset, we believe that model recognition errors have two reasons. On the one hand, in some videos, block the closing of elevator door cannot be captured by camera due to the position of the passengers standing in the car, so the model cannot detect this behavior. On the other hand, some passengers block the closing of elevator door with their feet, which is similar to kicking the elevator door and is misjudged by the model.

For the first situation, there are two solutions. The first method is to avoid the emergence of monitoring blind spots by using two cameras, but this method will increase the cost of monitoring. The second method is to adjust the installation position of the camera to the middle of the top of the car. For the second situation, the frequency of kicking the elevator door is significantly higher than the frequency of blocking the closing of the elevator door with your foot. Therefore, these two types of behavior can be effectively distinguished by adjusting the number of video frames for calculating optical flow.

5. Conclusions
Elevator accident is closely related to the unsafe behavior of passengers when the elevator fails. In this work, we summarize the unsafe behavior of elevator passengers, collect the elevator passenger behavior video dataset, and propose an elevator passenger behavior recognition model based on two-stream convolutional neural networks. The model is trained and evaluated on the collected elevator passenger behavior video dataset. The results show that the model can effectively identify unsafe behaviors of elevator passengers, with average recognition rate of 96.87%. Accuracy of the model has achieved 97.22% on the test set.

Conflicts of interest
There are no conflicts to declare.

Acknowledgements
This work is supported by the National Key Research and Development Program of China (NO.2018YFC0809005).

References
[1] State Administration for Market Regulation of China, (2021). Announcement on the Safety Status of Special Equipment in China 2020. http://gkml.samr.gov.cn/nsjg/tzsbj/202103/t20210315_326902.html
[2] Wang, W. (2015) Research on the status and problems of elevator public safety. Brand and Standardization, (10): 87–88.
[3] Lei, Y.C., Xu, Z.M., Nie, W.X. Reason for elevator accident and the preventive and rescue measures. China Elevator, 30(06): 46–47.
[4] Lan, S.N., Gao, Y.P., Jiang, S.H. (2021) Computer vision for system protection of elevators. Journal of Physics: Conference Series, 1848(1): 012156 (15pp).
[5] Jin, H.Y., Xiong, Q.Y., Shi, X., et al. (2012) Abnormal behavior detection in elevator car based on corner kinetic energy model. Application Research of Computers, 29(02): 775–778.
[6] Zhu, Y.J. (2017) Abnormal behavior detection of passengers in elevator. University of Science and Technology of China. Heifei, China.
[7] Zhu, Y.J., Wang, Z.F. (2016) Real-Time Abnormal Behavior Detection in Elevator. In: 4th Chinese Conference on Intelligent Visual Surveillance. Beijing, China. pp. 154–161.
[8] Ma, Z.W. (2018) Research on detection of abnormal behavior of passengers in elevator cars based on video. Southeast University. Nanjing, China.
[9] Xu, M.X. (2019) An overview of image recognition technology based on deep learning. Computer products and circulation, (01): 213.
[10] Zhong, P., Gong, Z.Q., Li, S.T., et al. (2017) Learning to Diversify Deep Belief Networks for Hyperspectral Image Classification. IEEE Transactions on Geoscience and Remote Sensing,
[11] Zhou, F., Xue, Y.Y., Li, Q.M. (2015) A Review of Video Monitor and Encoding Technology. Software, 36(04): 84-92.

[12] Krizhevsky, A., Sutskever, I., Hinton, G.E. (2017) ImageNet classification with deep convolutional neural networks. Communications of the ACM, 60(6): 84-90.

[13] Szegedy, C., Liu, W., Jia, Y.Q., et al. (2015) Going deeper with convolutions. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Boston, MA, USA. pp. 1-9.

[14] Simonyan, K., Zisserman, A. (2015) Very Deep Convolutional Networks for Large-Scale Image Recognition. In: The International Conference on Learning Representations (ICLR). San Diego, USA. (14pp).

[15] Xu, D., Dai, Y., Ji, J.H. (2019) Research on driver behavior recognition method based on convolutional neural network. China Safety Science Journal, 29(10):12-17.

[16] Simonyan, K., Zisserman, A. (2014) Two-Stream Convolutional Networks for Action Recognition in Videos. Advances in neural information processing systems, arXiv:1406.2199v2. (11pp).

[17] Chatfield, K., Simonyan, K., Vedaldi, A., and Zisserman, A. (2014) Return of the devil in the details: Delving deep into convolutional nets. In Proc. BMVC., arXiv:1405.3531v4. (11pp).

[18] Gkioxari, G., Malik, J. (2014) Finding Action Tubes. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 759-768.

[19] Lucas, Bruce D., and Kanade, T. (1981) An iterative image registration technique with an application to stereo vision. In: Proceedings of the 7th International Joint Conference on Artificial Intelligence (IJCAI). Morgan Kaufmann Publishers Inc., San Francisco, CA, USA. pp. 674–679.