IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 24 January 2024, accepted 8 February 2024, date of publication 12 February 2024, date of current version 16 February 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3365138

RESEARCH ARTICLE

Abnormal Activity Detection and Classification of Bus Passengers With In-Vehicle Image Sensing

HUEI-YUNG LIN[1,2], (Senior Member, IEEE), AND CHUN-HAN TSENG[2]
[1]Department of Computer Science and Information Engineering, National Taipei University of Technology, Taipei 106, Taiwan
[2]Department of Electrical Engineering, National Chung Cheng University, Chiayi 621, Taiwan

Corresponding author: Huei-Yung Lin (lin@ntut.edu.tw)

This work was supported in part by the Ministry of Science and Technology of Taiwan under Grant MOST 106-2221-E-194-004.

ABSTRACT As the self-driving technology is getting mature for public transportation applications, the safety concern of onboard passengers has become an important issue. It is essential to identify inappropriate or hazardous behaviors of passengers for the vehicles without human operators. In this work, we propose a technique to detect and classify the abnormal activities of passengers in a bus environment. Different from the existing human activity classification algorithms, our approach reduces the occlusion and increases the recognition rate by acquiring images from an overhead vision system. To overcome the increased complexity on feature extraction and classification, an action recognition network for top-view images are proposed by incorporating both spatial and temporal information. An image dataset, BUS-HAR, is generated for practical application scenarios with bus passengers. Experiments using real-world scene images have demonstrated the feasibility of our technique compared to existing approaches. The codes and image dataset are made available publicly at https://github.com/richardkuo1999/passenger-action-recognition.

INDEX TERMS Self-driving vehicle, action recognition, abnormal behavior detection, deep neural network, computer vision system.

I. INTRODUCTION

In the past few years, the development of self-driving technology has become increasingly mature, and advanced driver assistance systems (ADAS) are also adopted to practical uses. Many public transport vehicles are now equipped with ADAS to reduce driver fatigue and improve road safety. Meanwhile, autonomous driving systems have been gradually applied to mass transit vehicles to achieve automated public transportation. Currently, there are numerous government and industry collaborations on the testing programs of self-driving bus [1]. In this work we employ WinBus, a minibus jointly developed by Automotive Research and Testing Center of Taiwan and many companies, as our platform for experiments.[1]

In recent development, self-driving cars are able to achieve full autonomy in closed regions. Given a designated destination, they can plan the route and move accordingly based on sensing and perception of the surroundings. It is expected that no driver being onboard vehicles with fully automated future transportation. Since accidents occur and passengers may get injured due to unstable driving during the ride, it becomes an even more crucial issue if no human assistance in the vehicle. Thus, the passenger safety becomes an emerging problem for self-driving transport vehicles [2].

To address this issue, we propose a vision-based technique to recognize abnormal passenger behaviors in this paper. The passengers in a crowded space are first detected, followed by human activity classification using deep neural networks. We conduct the experiments in different scenarios with images collected using ceiling-mounted cameras. With the top-down viewpoint for image acquisition, occlusion among the passengers can be mitigated and the recognition performance is improved. In the proposed method, the objective is to identify improper or dangerous activities which need more attentions in unmanned operating environments. The abnormal behaviors include five types: falling, lying down, squatting, pulling the handrail, and waving.

The associate editor coordinating the review of this manuscript and approving it for publication was Junho Hong.

[1]https://www.artc.org.tw/en

To identify abnormal human behaviors, action recognition techniques are commonly used. They have been investigated by computer vision researchers for decades. Currently, some popular approaches include using human body skeleton, optical flow information, and RGB images. In our application scenario with limited bus interior space, a top-down view is adopted to cover a wider passenger region. It is not suitable to use body skeleton for action recognition since the existing approaches commonly adopt the images collected from side views for feature extraction. In this case, the recognition rate will be greatly affected due to the lower body being occluded by the upper part. For optical flow approaches, the mandatory preprocessing usually increases the computational complexity of the model and decreases the recognition speed. Hence, it is not possible to meet the real-time constraint of abnormal behavior detection for an immediate response.

There are generally two approaches for action recognition using RGB images, Long Short-Term Memory (LSTM) and Convolutional Neural Network (CNN). LSTM is an architecture improved from Recurrent Neural Network (RNN), while designed to address the issues of long-term memory in RNN. Although the sequence of images is taken as input for action recognition, this approach requires significant computational resources for training [3]. To address this problem, Lee et al. proposed a 3D CNN (3D Convolutional Neural Network) for activity classification by considering temporal information as an additional dimension to conventional 2D CNN [4]. A low-cost network model is utilized to extract continuous features from image sequence. In the previous works, 3D CNN often leads to model overfitting due to limited training data and the large number of parameters. Since the release of large action recognition datasets such as Kinetics-700 [5], sufficient data for model training have greatly improved the accuracy. Thus, it is advantageous to use 3D CNN as a framework for activity classification [6].

In this paper, we propose a 3D CNN-based network model for abnormal activity detection and classification of passengers in bus environments. Unlike most of the existing human activity recognition techniques, the images are captured from overhead cameras mounted on the ceiling to deal with occlusion. Consequently, the image feature extraction and network model training become more challenging. Moreover, with the downward looking viewpoint, the faces of passengers will not be seen from the images. It is able to provide a good privacy protection at the same time. Since there are no data available for abnormal behavior analysis in the bus environment, a new image dataset called BUS-HAR[2] is created for our training and testing. Our experiments are conducted in the real-world scenes, and the results demonstrate significant performance improvements compared to existing 3D CNN-based methods. The contributions of this work are as follows:

- Top-view images acquired in bus environments are used to perform passenger abnormal activity recognition.
- A 3D CNN-based network architecture is developed to achieve high accurate human behavior classification.
- The first abnormal activity recognition dataset are made available for transportation related research.

II. RELATED WORKS

In recent years, abnormal activity recognition is increasingly gaining attention, as an important technique to ensure environment safety. Some application scenarios include outdoor spaces such as train stations, airports, and subway stations, as well as indoor locations like public building, shopping malls and elderly care centers. In general, surveillance personnel are responsible to monitor the areas manually. As the number of surveillance regions grows, it would lead to the decreased concentration and increased fatigue among the security staffs. Since abnormal activities occur infrequently, it further makes continuous human surveillance challenging. As a result, the automated system capable of detecting abnormal activities is favorable for practical applications [7], [8].

Most current works directly define the kinds of actions as abnormal activities. Popoola et al. [9] pointed out that it was mandatory to consider what constituted abnormal activity in terms of actions. They treated abnormal activities as actions that stand out prominently and distinctly within normal cases. If a group of people running with only one person is walking, this person is different from the rest notably. This distinction will be used to define walking as potentially abnormal activity in the situation. By content analysis, it is then possible to determine if abnormal activities occur from image sequences.

Some abnormal activity recognition methods focus on the analysis of a group. In an early work, Mehran et al. employed Social Force Model to detect abnormal activities in a crowd [10], [11]. They utilized computer vision techniques to extract image features and incorporate optical flow information. The behavioral change over a period of time is detected by optical flow for assistance in identifying abnormal action. In addition to using optical flow, Basharat et al. [12] developed a model of normal movement trajectories based on learned patterns. Individual objects were tracked within surveillance footage to determine whether abnormal activities occur. Due to some unusual situations frequently arose on a train platform, such as people accidentally falling onto the tracks, Delgado et al. proposed a system that automatically detected if passengers jumped or fell from the platform [13]. In recent years, deep learning techniques has gained significant popularity, many researchers have applied learning-based methods to abnormal activity detection. Sun et al. employed an RNN network for image feature extraction and combine it with SVM (support vector machine) for abnormal behavior detection [14]. More recent investigation can be found in [15] for a comprehensive review.

[2]The image dataset is available publicly at https://github.com/richardkuo1999/Passenger-Action-Recognition.

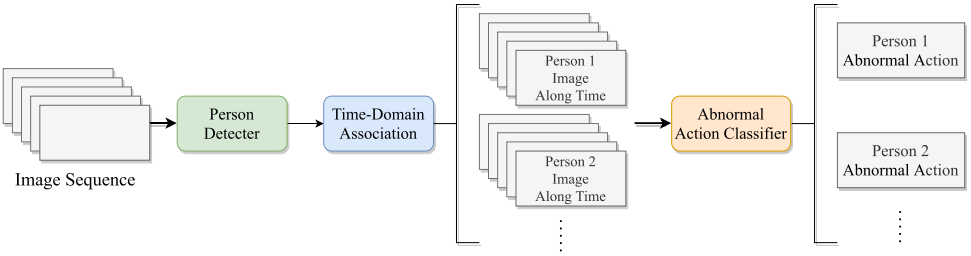


FIGURE 1. The system flowchart of our abnormal action recognition technique.

From the above literature survey, abnormal action recognition can be broadly categorized into two approaches: those focusing on individual objects as detection subjects and those treating crowds as collective entities for action classification. Nevertheless, the techniques based on object detection often encounter accuracy reduction in a crowded scenario. Objects can easily be occluded by other individuals or the trajectories of tracked objects might be lost. To cope with this difficulty, Zhou et al. proposed a CNN model that is capable of extracting spatial information from individual frames and capturing complex motion relations from consecutive frames [16]. The optical flow data were used to identify regions with abnormal movement in sequences of frames. It was processed by a spatiotemporal CNN to extract motion features, facilitating the detection and recognition of abnormal activities in crowds. In earlier research, action classification within vehicle interiors was most commonly applied to understand driver behaviors. It was important when drivers experienced fatigue or lacked concentration. In this regard, Yan et al. used CNNs to predict whether a driver's attention is focused [17]. They segmented the facial regions such as the eyes, mouth, and ears from the driver's images. Through CNN-based classification, it could identify whether the driver closed eyes or was using a mobile phone while driving.

While there are significant developments on driver behavior analysis technologies, relatively less research concerning passenger activity are investigated. Moreover, the availability of suitable public datasets are also limited. To mitigate this problem, Tu et al. proposed a technique to transform a large number of driver's images into the viewpoints for passengers through projection [18]. They utilized CNNs for passenger behavior recognition, distinguishing activities such as drinking water, talking on the phone, typing, resting. Most existing methods primarily rely on the individual images for activity recognition. Lacking temporal information for some actions makes classification more challenging. In [19], a sequence of images were used to recognize passenger activities inside a vehicle. A 3D CNN was employed to process cropped image frames for action recognition.

In the current literature, most driver or passenger behavior analysis focus on the recognition within small vehicles. There exist very limited research addressing activity recognition of passengers in public transportation systems. It presents many challenges of action classification in public transport vehicles due to the image occlusion posed by the horizontal viewpoint. The lack of suitable public datasets also lead many works to collect their own data. Velastin et al. employed histogram of oriented gradients (HOG) for passenger detection, followed by passenger action recognition using SVM [20]. Nevertheless, action recognition within transport vehicles requires to address the common passenger movements, such as walking, standing and sitting. Kao and Lin [21] proposed an improved 3D CNN-based model to recognize such passenger activities. The system aims to understand the passenger status such that comfort driving speed and turning movement can be adjusted accordingly.

III. APPROACH

The proposed abnormal action recognition technique mainly consists of two stages. The first part is the passenger detector, which aims to identify the passengers with abnormal actions. To achieve this, the Action Tube framework [22] is incorporated with YOLOv5 to derive the passenger bounding box for tracking. A sequence of bounding box across multiple frames is generated and fed into a 3D CNN for passenger abnormal behavior recognition in the second stage. Figure 1 depicts the system flowchart of the proposed method. It only requires the cameras mounted on the ceiling of the vehicle, and does not need additional hardware settings or dramatic infrastructure changes onboard the buses. As long as the camera viewpoints and image quality are satisfactory, the performance is based on the algorithm.

To classify actions with temporal information, 3D CNN is able to recognize anomaly such as falling [23]. In the recent studies of fall detection, many approaches adopt human joint positions to identify the occurrence of abnormal actions [24], [25]. However, due to the overhead viewpoint of the camera in the bus environments, it is not possible to use the human joint detection networks trained on MPII and COCO datasets (both containing side-view images) [26], [27]. The generated joint points do not fit well due to the difference in viewpoints. Thus, our abnormal action detection technique is carried out directly on image data, instead of preprocessing with human skeleton.

3D CNN is a network architecture that extends 2D convolution by adding a temporal dimension. We utilize time-series images as input, and apply 3D convolution to extract feature maps to encode the spatial and temporal information. Due to the extra dimension, the number of parameters of 3D CNN

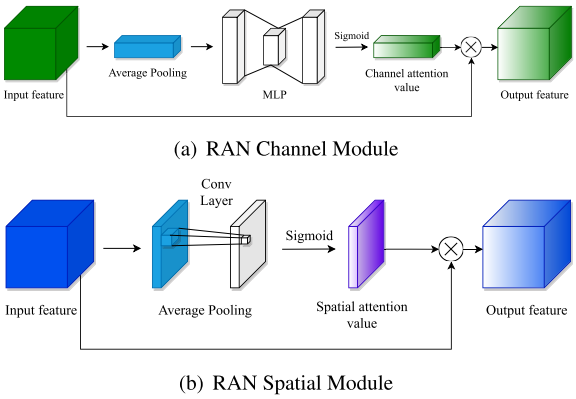


FIGURE 2. The attention modules adopted in this work.

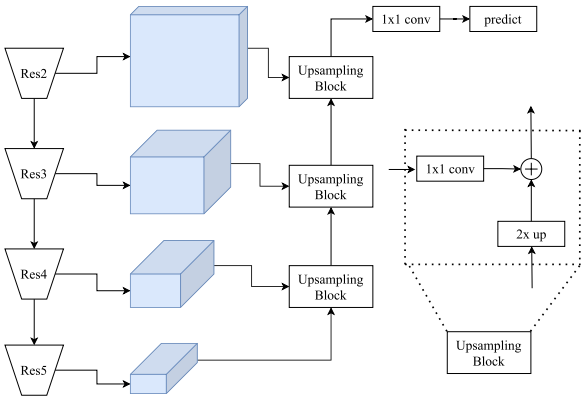


FIGURE 3. The FPN module used to merge the topmost feature map with other bottom-level feature maps.

is increased significantly. A commonly occurred problem is overfitting caused by the lack of sufficient training data. Thus, we utilize the Kinetics-400 dataset, with 240,000 training and 20,000 validation video clips, to mitigate the overfitting of 3D CNN [28]. Moreover, transfer learning is adopted to improve the performance by training on our BUS-HAR dataset for the bus application scenario. It is then followed by incorporating residual blocks to deal with gradient vanishing inherent from deep network structures.

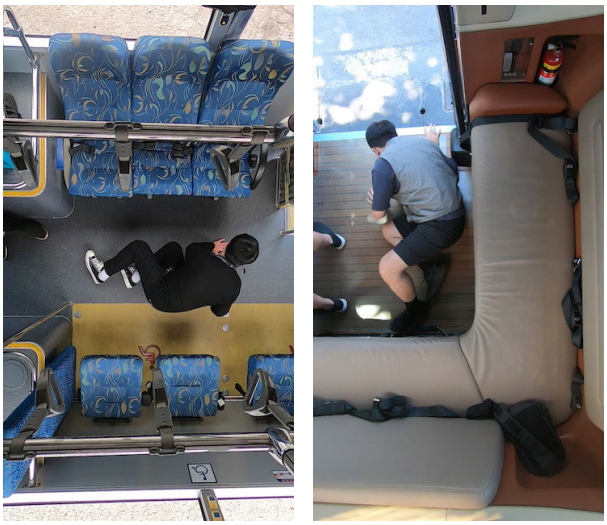
For classification of abnormal passenger activities onboard a bus, we consider 5 categories: attack, squatting, lying, fall, and handrail-pulling, An image sequence consists of twelve frames is utilized as input for 3D CNN. The backbone of the proposed network is based on the framework by Hara et al. [29], where 2D ResNet pre-trained on ImageNet is extended to 3D ResNet using Kinetics-400 for training. Transfer learning is then applied on the pre-trained model using UCF101 to improve its performance. Finally, the network is fine-tuned using the relatively small BUS-HAR dataset we collected.

3D CNN introduces an additional time dimension, which increases the model size significantly. As a result, the overfitting issue in 3D convolution makes the network more difficult to train. In this work, the 3D convolution is split into spatial and temporal components using separate kernels with sizes of $1 \times d \times d$ and $t \times 1 \times 1$ [30]. This approach makes deeper network models easier to derive the parameters. Furthermore, we integrate Residual Attention Network (RAN) and Feature Pyramid Network (FPN) into our network model to enhance the classification accuracy as follows.

Attention modules are commonly used in neural networks to identify the importance of features through learning so the weights can be adjusted to improve classification results. The architecture Residual Attention Network was first proposed as a module for 3D CNN based on the concept of SENet [31]. As depicted in Figure 2, we adopt both the channel attention module and spatial attention module in sequence. The former emphasizes the channels of the feature maps, while the latter deals with the width and height of the feature maps. Although top-level feature maps consist of rich semantic features, they might not be able to identify the target location. To cope with this problem, we use Feature Pyramid Network to merge the topmost feature map with other bottom-level feature maps, as illustrated in Figure 3. Instead of predicting the feature maps after each iteration, the upsampling is carried out sequentially for feature fusion. This fusing approach can incorporates the information from different scales and enhance the classification performance.

The existing human action classification techniques generally input acquired images into the networks for recognition. In our application, it is also required to locate the position of passengers in the images when abnormal behavior occurs. To address this issue, we incorporate a passenger detector before action classification. The bounding box extracted from passenger tracking is fed into the action recognition network for classification. At this stage, the human detector does not only serves as a localization module but also allows the system to perform action recognition separately for individual passengers. The collected data images labeled with defined passenger abnormal actions are then used for classification.

Our proposed system is capable of tracking passengers to determine their relative positions in the images. It also stores a sequence of images for action classification. The frequency of 5 Hz is used to extract images from 30 FPS videos in the implementation. We use the duration of 12 frames as input to the classifier to identify fast abnormal actions. After the first frame is detected, a fixed identifier to each passenger is assigned and the bounding box is stored. In the subsequent frames, current and previous bounding boxes are compared using Intersection-Over-Union (IoU) metric. The Hungarian algorithm [32] is then utilized to match the most appropriate ROIs for passenger tracking. Comparing the IoU with visual tracking, the IoU method performs significantly fast as it only considers the size and position of the bounding boxes. In our application, passengers do not move rapidly inside a bus. The overhead camera setting reduces the possibility of passengers occluding each other. Consequently, mismatching or tracking loss do not seldom present in the image sequences.



(a) The image from a full-size bus. (b) The image from a minibus.

FIGURE 4. Typical passenger space images collected in our BUS-HAR dataset.

A. DATASETS

The passenger area inside a bus is narrow and crowded, so the images are acquired from a ceiling-mounted camera for activity recognition. Since it is very different from the current action classification research, own datasets are necessary for network training and testing. In our experiments, we consider two different application scenarios, a conventional full-size bus and a minibus. Figure 4 shows the images captured in the passenger spaces. To achieve a wide coverage of the interior, stereo cameras are used in both cases. The cameras are fixed at the heights of 2.26 and 1.90 meters from the ground for the bus and minibus, respectively. In our BUS-HAR dataset, the images are collected during daytime, and in bus and mini-bus environments. Since we utilize the downward facing cameras mounted on the ceiling to capture images from the top, it is not required to consider the demographic issue. Similar to most research in the activity recognition, the dataset is used to construct and evaluate the algorithm. It is then generalized to broader application scenarios during the development of specific system.

The most commonly used datasets for action recognition include UCF101 [33] and HMDB51 [34]. More recently, the Kinetics dataset [5] contains 700 categories for over 650,000 videos, with each category up to 700 videos. Most of the 3D CNN-based action recognition techniques rely on these three datasets for training and testing. In this work transfer learning is conducted on the Kinetics dataset. The pre-trained weights generated by this dataset containing a large amount of images are used to initialize our model.

The BOSS dataset [20] is the only dataset created with the viewpoint suitable for our application scenarios. It contains images acquired from a moving train using 10 cameras, with an image resolution of 720 × 576. Figure 5 shows two images obtained from top-view cameras. The dataset consists of nine categories: four personal actions and five interactive

FIGURE 5. Typical top-view images in the BOSS dataset.

actions, such as walking, sitting and standing, etc. We adopt 1,000 annotated images from the BOSS dataset to train our passenger detector. In addition, 5,000 images randomly selected from the people category of the COCO dataset are also included to augment the training samples [27].

IV. EXPERIMENTS

In the experiments, the feasibility of the proposed techniques is validated using images collected from a full-size bus (BRT) and a minibus (WinBus) for training and testing. It includes five abnormal actions in our experiment. They are considered according to the main safety issues during driving of vehicles. In general, the real-world variability such as the application scenarios with different lighting conditions and camera viewing directions can increase the system's applicability. It will be extended by including the more variabilities of samples in the dataset.

- Fall: Passenger falls in the cabin.
- Squatting: Passenger crouches in the cabin.
- Lying: Passenger lies on a seat in the cabin.
- Attack: Passenger holds a weapon and performs attacking actions
- Handrail-pulling: Passenger performs dangerous actions using the lever in the cabin.

The numbers of training and testing samples in the five action categories are 288/52, 320/22, 392/36, 256/18 and 412/38, respectively. Due to the difficulty on collecting the real-world images from different scenarios, it is not feasible to evaluate the techniques with the many scales of bus sizes. In this work, we consider two different bus sizes, BRT and WinBus. The crowd density is set as 1-3 passengers appeared in the camera viewpoint. The computational resources utilized in this work is a desktop with an Intel i7-8700 CPU and 16 GB RAM. The GPU for model training and testing is Nvidia GeForce RTX 2070 SUPER.

For abnormal action classification, we use a network based on 3D ResNet-50 with the temporal-spatial separable convolution (R2P1D) as our backbone model. This addresses the common overfitting issue observed in 3D ResNet-50. The use of separable convolution does not reduce the total number of parameters, but it enhances model optimization and mitigates overfitting. As illustrated in Figure 6, we modify 3D ResNet-50 with R2P1D by introducing the RAN (Residual Attention Network) modules (blue blocks) and FPN (Feature Pyramid Network) modules (green blocks). The RAN

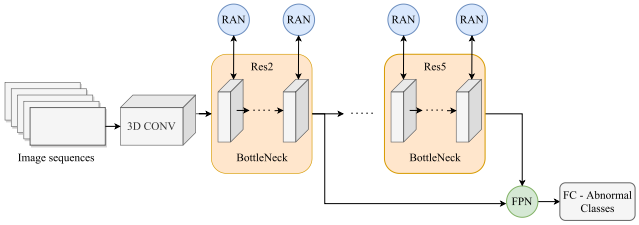


FIGURE 6. The proposed network architecture.It incorporates RAN and FPN modules for both channel and spatial aspects of feature maps after each bottleneck convolution.

TABLE 1. The evaluation results and the comparison with 3D ResNet-50 plus space-time split convolution model, 3D ResNet-50 (R2P1D).

	3D ResNet-50 (R2P1D)	Ours
Training dataset	BRT-train	BRT-train
Training data	1,668	1,668
Testing dataset	BRT-test	BRT-test
Testing data	166	166
Input size	224	224
Batch size	8	8
Pretrain weight	yes	yes
Epoch	100	100
Fall accuracy	90.4%	98.1%
Squatting accuracy	68.2%	81.8%
Lying accuracy	100%	100%
Pull rod accuracy	97.4%	100%
Attack accuracy	100%	100%
Overall accuracy	92.2%	97.0%

modules focus on both channel and spatial aspects of feature maps after each bottleneck convolution, and reveal prominent features. As for the FPN modules, they combine the feature maps from Res2, Res3, Res4, and Res5 convolutions through upsampling and 1×1 convolutions to fuse the outputs from different layers.

We use the weights pre-trained on Kinetics-700 and Moments in Time[3] datasets, with a total of 1,039 categories, and conduct transfer learning with our BUS-HAR dataset. During the network training, our input sample length is set to 12, and the original images are resized through scaling and filled to align with the shorter side of 224×224. The training parameters are given by weight decay of 0.0001, momentum of 0.9, and an initial learning rate of 0.1. We use the cosine decay strategy [35] for modifying the learning rate over time. To augment data for diversity and increase training samples, we apply horizontal flips and random rotations to the images. In the evaluation, we compare the proposed network model to the original 3D ResNet-50 plus space-time split convolution. The batch size and the number of epochs for network training are set as 8 and 100, respectively.

Our network model is built on the backbone of 3D ResNet-50 with space-time split convolution. In addition to incorporating RAN and FPN modules, we make adjustments to the learning rate change strategy, and modify the activation and loss functions. The commonly used cross entropy loss has the drawback due to imbalanced data for network training.

[3]http://moments.csail.mit.edu/

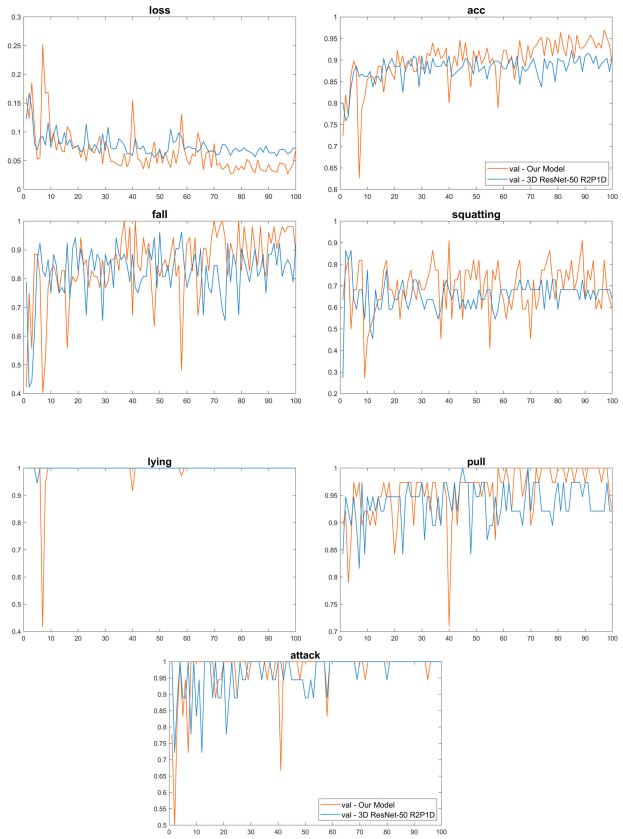


FIGURE 7. The test accuracy and loss curve of the two network models.

TABLE 2. The modifications of our network and training data for ablation study.

	Test accuracy
3D ResNet-50 (R2P1D)	86.1%
+ Leaky ReLU	86.7%
+ Focal Loss	89.8%
+ Add Training Data	92.2%
+ RAN Attention Module	95.2%
+ FPN Attention Module	97.0%

TABLE 3. The numbers of sample images for different action classes.

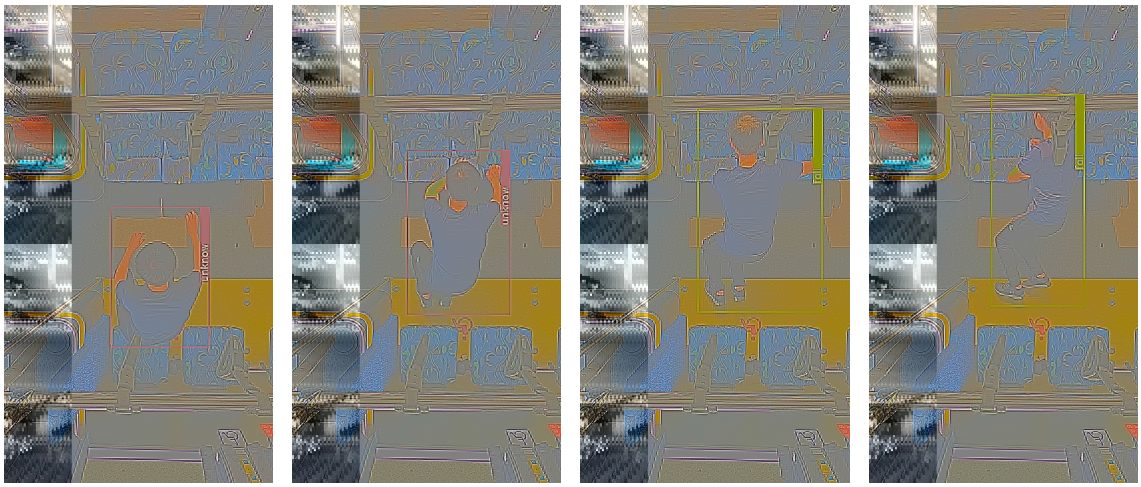
	fall	squatting	lying	attack
Training data	394	402	404	400
Testing data	114	106	104	120

TABLE 4. The testing accuracy of the WinBus dataset.

	Our Model
Training dataset	WinBus-train
Training data	1,600
Testing dataset	WinBus-test
Testing data	444
Input size	224
Batch size	8
Pretrain weight	yes
Epoch	60
Fall accuracy	100%
Squatting accuracy	97.2%
Lying accuracy	100%
Attack accuracy	100%
Overall accuracy	99.3%

In this work, we utilize the focal loss loss function

$$\mathrm{FL}\,(p) = -\alpha\,(1-p)^{\gamma}\log\,(p) \tag{1}$$



(a) (b) (c) (d)

FIGURE 8. The abnormal action recognition results on a BRT bus, with the confidence level set as 0.95.

TABLE 5. The testing accuracy using both BRT and WinBus data.

	Our Model
Training dataset	WinBus-train, BRT-train
Training data	2,856
Testing dataset	WinBus-test, BRT-test
Testing data	572
Input size	224
Batch size	8
Pretrain weight	yes
Epoch	60
Fall accuracy	95.2%
Squatting accuracy	94.5%
Lying accuracy	100%
Attack accuracy	100%
Overall accuracy	97.4%

TABLE 6. The number of parameters and accuracy of the network models tested on the UCF101 split 1 dataset.

	Parameters (Million)	Accuracy
3D ResNet-50	46.4	30
3D ResNet-50 [21]	61.5	35.3
3D ResNet-50 (R2P1D)	46.4	33.3
3D ResNet-101	85.5	25.8
Ours	66.4	36.3

TABLE 7. The results of the UR dataset with the models pretrained using the Kinetics dataset and our BUS-HAR dataset.

	Test1	Test2
Training dataset	UR dataset	UR dataset
Training data	62	62
Testing dataset	UR dataset	UR dataset
Testing data	15	15
Input size	224	224
Batch size	8	8
Pretrain weight	Kinetics 700 data	Our abnormal data
Overall accuracy	100%	100%

proposed by Lin et al. [36]. Eq. (1) is based on cross entropy loss, but includes one additional weighting factor. The focal loss function assigns higher loss values to challenging data points while assigning lower values to easier points. It is then possible to strengthen the learning process for hard samples.

Table 1 tabulates the evaluation results and the comparison with 3D ResNet-50 plus space-time split convolution model, 3D ResNet-50 (R2P1D). The test accuracy and loss trends of both network models are depicted in Figure 7. With the same training parameters, the results demonstrate that the proposed network model outperforms 3D ResNet-50 (R2P1D). Table 2 shows the modifications of our network architecture and the ablation study. By incorporating the FPN and RAN attention modules, utilizing leaky ReLU and focal loss in conjunction with augmented training samples, the accuracy has increases from the baseline R2P1D model's 86.1% to 97.0%.

To detect abnormal actions, we incorporate the classification network with passenger detection. By detecting passengers, tracking is initiated, and a sliding window of 12 frames is utilized to record extracted passenger bounding boxes for abnormal activity detection. Some action recognition results are shown in Figure 8. The threshold is set as 0.95 to enhance the confidence in action class prediction. The instances with low recognition confidence are labeled as 'unknown' in activity classification to mitigate false alarm. We also apply our method to the WinBus dataset, with the settings and training parameters remain the same as the previous BRT experiment. The numbers of sample images for different action classes are summarized in Table 3. Note that the WinBus dataset does not contain the 'handrail-pulling' category due to the limited passenger space. Table 4 shows the recognition for individual classes, with overall accuracy of 99.3%. Figure 9 shows the abnormal action results, with the confidence set as 0.95. Note that, since there are not previous works for abnormal activity recognition in the bus environment. It is not possible to utilize existing systems for direct comparisons. Nevertheless, we did conduct the experiments using different models, compare the performance, and provide ablation analysis.

To increase the generalization of our network models, the training samples are extended to include both the BRT and WinBus datasets. As illustrated in Table 5, there is a

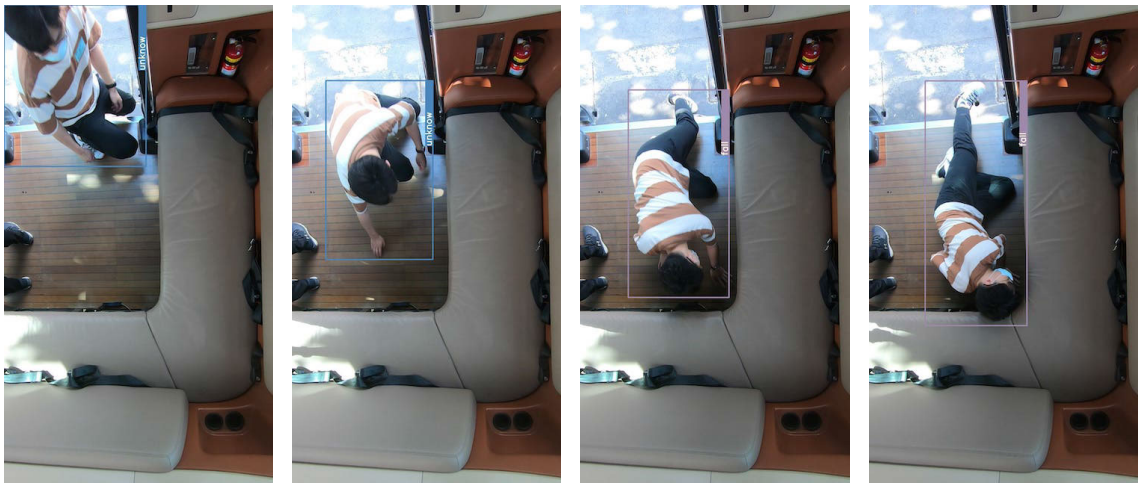


FIGURE 9. The abnormal action recognition results on a WinBus, with the confidence level set as 0.95.

notable accuracy increase for the 'squatting' class. It suggests future explorations on refining approaches to handle distinct classes of image data during model training and testing. There are current five abnormal activities considered for training and testing. For the activities not defined explicitly for training, it is not feasible to assess their recognition rates. Nevertheless, more categories of actions will be included for the extensive evaluation.

In addition to performing experiments on our private data, we have also incorporated two public datasets, UCF101 split 1 and UR, for training and testing. The evaluation contains 3D ResNet-50, 3D ResNet-50 (R2P1D), 3D ResNet-101, 3D ResNet-50a [21], and our network model on the datasets. For fair comparison, all models are trained from scratch without using pre-trained weights. Table 6 tabulates the number of parameters and recognition accuracy of the baseline models and our method. The results show that our technique achieves the highest accuracy, with a modest increase in the model size compared to 3D ResNet-50. To evaluate the capability of our model on handling broader abnormal actions, the UR dataset is employed. The dataset contains images of falls and daily activities involving abnormal behaviors. Two distinct training sessions are conducted, both utilizing pre-training weights. In Table 7, Test 1 and Test 2 represent the models pre-trained on the Kinetics dataset and our BUS-HAR dataset, respectively. The results have demonstrated perfect detection for abnormal actions.

V. CONCLUSION

This work presents a novel 3D convolutional neural network designed for detecting and categorizing abnormal actions of passengers in a bus environment. Different from the existing action recognition algorithms, our proposed technique leverages sequences of images obtained from an overhead camera system. This represents an intricate task given the challenges caused by occlusion in the limited passenger space and the lack of publicly available datasets. Our proposed approach performs passenger detection and tracking, followed by human action recognition to identify abnormal activities. A new dataset, BUS-HAR, is created with the images collected from the passenger spaces of both the bus and minibus for training and testing. In the real experiments, the performance evaluation has demonstrated significant improvements compared to existing methods. Continue with this pioneer research related to the abnormal activity recognition of passengers onboard a bus, the future work will focus on enlarging the action classes for identification. The system deployment on resource-constrained edge devices will be considered in the investigation.

REFERENCES

[1] R. Buehler and V. Tech, "Can public transportation compete with automated and connected cars?" *J. Public Transp.*, vol. 21, no. 1, pp. 7–18, Jan. 2018.
[2] R. Bridgelall, "Using artificial intelligence to derive a public transit risk index," *J. Public Transp.*, vol. 24, 2022, Art. no. 100009.
[3] J.-Y. He, X. Wu, Z.-Q. Cheng, Z. Yuan, and Y.-G. Jiang, "DB-LSTM: Densely-connected bi-directional LSTM for human action recognition," *Neurocomputing*, vol. 444, pp. 319–331, Jul. 2021.
[4] H. Lee, Y.-S. Kim, M. Kim, and Y. Lee, "Low-cost network scheduling of 3D-CNN processing for embedded action recognition," *IEEE Access*, vol. 9, pp. 83901–83912, 2021.
[5] J. Carreira and A. Zisserman, "Quo vadis, action recognition? A new model and the kinetics dataset," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2017, pp. 6299–6308.
[6] C.-H. Tseng and H.-Y. Lin, "A vision-based system for abnormal behavior detection and recognition of bus passengers," in *Proc. IEEE 25th Int. Conf. Intell. Transp. Syst. (ITSC)*, Oct. 2022, pp. 2134–2139.
[7] E. Ramanujam, T. Perumal, and S. Padmavathi, "Human activity recognition with smartphone and wearable sensors using deep learning techniques: A review," *IEEE Sensors J.*, vol. 21, no. 12, pp. 13029–13040, Jun. 2021.
[8] M. Ehatisham-Ul-Haq, A. Javed, M. A. Azam, H. M. A. Malik, A. Irtaza, I. H. Lee, and M. T. Mahmood, "Robust human activity recognition using multimodal feature-level fusion," *IEEE Access*, vol. 7, pp. 60736–60751, 2019.
[9] O. P. Popoola and K. Wang, "Video-based abnormal human behavior recognition—A review," *IEEE Trans. Syst., Man, Cybern., C, Appl. Rev.*, vol. 42, no. 6, pp. 865–878, Nov. 2012.

[10] D. Helbing and P. Molnár, "Social force model for pedestrian dynamics," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 51, no. 5, p. 4282, 1995.
[11] R. Mehran, A. Oyama, and M. Shah, "Abnormal crowd behavior detection using social force model," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 935–942.
[12] A. Basharat, A. Gritai, and M. Shah, "Learning object motion patterns for anomaly detection and improved object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2008, pp. 1–8.
[13] B. Delgado, K. Tahboub, and E. J. Delp, "Automatic detection of abnormal human events on train platforms," in *Proc. IEEE Nat. Aerosp. Electron. Conf.*, Jun. 2014, pp. 169–173.
[14] X. Sun, S. Zhu, S. Wu, and X. Jing, "Weak supervised learning based abnormal behavior detection," in *Proc. 24th Int. Conf. Pattern Recognit. (ICPR)*, Aug. 2018, pp. 1580–1585.
[15] R. Nayak, U. C. Pati, and S. K. Das, "A comprehensive review on deep learning-based methods for video anomaly detection," *Image Vis. Comput.*, vol. 106, Feb. 2021, Art. no. 104078.
[16] S. Zhou, W. Shen, D. Zeng, M. Fang, Y. Wei, and Z. Zhang, "Spatial–temporal convolutional neural networks for anomaly detection and localization in crowded scenes," *Signal Process., Image Commun.*, vol. 47, pp. 358–368, Sep. 2016.
[17] C. Yan, H. Jiang, B. Zhang, and F. Coenen, "Recognizing driver inattention by convolutional neural networks," in *Proc. 8th Int. Congr. Image Signal Process. (CISP)*, Oct. 2015, pp. 680–685.
[18] I. Tu, A. Bhalerao, N. Griffiths, M. Delgado, T. Popham, and A. Mouzakitis, "Deep passenger state monitoring using viewpoint warping," in *Image Analysis and Processing—ICIAP 2017*. Cham, Switzerland: Springer, 2017, pp. 137–148.
[19] I. Tu, A. Bhalerao, N. Griffiths, M. M. Delgado, A. Thomason, T. Popham, and A. Mouzakitis, "Dual viewpoint passenger state classification using 3D CNNs," in *Proc. IEEE Intell. Vehicles Symp. (IV)*, Jun. 2018, pp. 2163–2169.
[20] S. A. Velastin and D. A. Gómez-Lira, "People detection and pose classification inside a moving train using computer vision," in *Proc. Int. Vis. Inform. Conf.* Cham, Switzerland: Springer, 2017, pp. 319–330.
[21] S.-F. Kao and H.-Y. Lin, "Passenger detection, counting, and action recognition for self-driving public transport vehicles," in *Proc. IEEE Intell. Vehicles Symp. (IV)*, Jul. 2021, pp. 572–577.
[22] S. Saha, G. Singh, M. Sapienza, P. Torr, and F. Cuzzolin, "Deep learning for detecting multiple space-time action tubes in videos," in *Proc. Brit. Mach. Vis. Conf.*, 2016.
[23] Md. M. Islam, O. Tayan, M. R. Islam, M. S. Islam, S. Nooruddin, M. N. Kabir, and M. R. Islam, "Deep learning based systems developed for fall detection: A review," *IEEE Access*, vol. 8, pp. 166117–166137, 2020.
[24] N. Noor and I. K. Park, "A lightweight skeleton-based 3D-CNN for real-time fall detection and action recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2023, pp. 2179–2188.
[25] L. Yao, W. Yang, and W. Huang, "A fall detection method based on a joint motion map using double convolutional neural networks," *Multimedia Tools Appl.*, vol. 81, no. 4, pp. 4551–4568, Feb. 2022.
[26] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele, "2D human pose estimation: New benchmark and state of the art analysis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 3686–3693.
[27] T. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2014, pp. 740–755.
[28] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, M. Suleyman, and A. Zisserman, "The kinetics human action video dataset," 2017, *arXiv:1705.06950*.
[29] K. Hara, H. Kataoka, and Y. Satoh, "Can spatiotemporal 3D CNNs retrace the history of 2D CNNs and ImageNet?" in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6546–6555.
[30] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri, "A closer look at spatiotemporal convolutions for action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6450–6459.
[31] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.
[32] M. Ullah, A. K. Mohammed, F. A. Cheikh, and Z. Wang, "A hierarchical feature model for multi-target tracking," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2017, pp. 2612–2616.
[33] K. Soomro, A. R. Zamir, and M. Shah, "UCF101: A dataset of 101 human actions classes from videos in the wild," 2012, *arXiv:1212.0402*.
[34] D. S. Wishart et al., "HMDB: The human metabolome database," *Nucleic Acids Res.*, vol. 35, pp. D521–D526, Jan. 2007.
[35] T. He, Z. Zhang, H. Zhang, Z. Zhang, J. Xie, and M. Li, "Bag of tricks for image classification with convolutional neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 558–567.
[36] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, Jun. 2017, pp. 2980–2988.

HUEI-YUNG LIN (Senior Member, IEEE) received the Ph.D. degree in electrical and computer engineering from The State University of New York at Stony Brook, USA. He is currently a Professor with the Department of Computer Science and Information Engineering, National Taipei University of Technology. He also has a joint appointment with the Department of Electrical Engineering, National Chung Cheng University. In 2002, he joined National Chung Cheng University, Taiwan, as an Assistant Professor and he was promoted to a Full Professor, in 2013. He was the Director of Research Liaison Division, from 2009 to 2013, and the Director of the Academic Development Division, from 2012 to 2014, with Office of Research and Development, National Chung Cheng University. He is the author of over 180 international conferences and journal articles, and five book chapters. He holds 11 U.S. and nine Taiwan invention patents. He also serves as the organizing committee member and a program committee member of over 50 international conferences. He is a recipient of the Excellent Research Award from National Chung Cheng University, the Outstanding Academic-Industry Cooperation Award from Taiwan Association of System and Science and Engineering (TASSE), and the Outstanding Robotics Engineer Award from Robotics Society of Taiwan (RST). His research interests include machine learning, computer vision, robotics, and mechatronics. He is a fellow of IET and RST, and a Senior Member of Optica.

CHUN-HAN TSENG received the B.S. degree in electrical engineering from the National Taipei University of Technology, in 2019, and the M.S. degree in electrical engineering from National Chung Cheng University, Taiwan, in 2021. He is currently with Himax Technologies Inc., Tainan, as a Computer Engineer. His research interests include autonomous mobile robot, computer vision, machine learning, image processing, and computer graphics.