[1]Chen Li

[1]Yunqing Liu

[1]Junnian Wang

[1]Jianxin Li

[1]Chengtong Zhuang

Recognising pedestrian behaviour using a multi-channel spatiotemporal fusion network

Journal of Electrical Systems

Abstract: - To protect public safety, video abnormal behaviour detection is essential. The low accuracy of identifying abnormal behaviour in pedestrians is the focus of this study. To address this, an abnormal target identification method based on multi-feature fusion of trajectory skeleton is proposed. The process begins with defining the type of abnormal behaviour in accordance with the environmental requirements. Next, pedestrian identification is carried out in the designated area, the pedestrian is tracked to determine its movement trajectory, and the image coordinates of the corresponding relevant nodes are calculated by analysing the human posture. Lastly, the track and skeleton features are integrated to classify the normal and abnormal behaviours, and the identification of abnormal target behaviours is accomplished. Test results on the behavioural analysis database indicate that this algorithm's accuracy is 87.08%, which is higher than the single-feature identification method's detection efficiency of abnormal behaviours.

Keywords: anomalous behaviour; target detection; tracking; trajectory; skeleton

1 INTRODUCTION

The field of intelligent surveillance, traffic management, and security has placed increasing importance on the recognition of pedestrian behaviour on pavements due to the rapid advancement of information technology and the ongoing development of society [1]. Precise identification of pedestrian behaviour yields vital data for intelligent surveillance systems, augmenting public safety and optimising urban traffic flow management. However, because pedestrian behaviours are so diverse and complex—including changes in light, partial occlusion, and view angle—traditional behaviour recognition methods frequently encounter difficulties[2].

Particularly, as the number of cars and traffic accidents rises, there has been a great deal of concern about automotive assisted driving systems. Since pedestrians play a significant role in traffic, it is important to correctly recognise their behaviour ahead of vehicles and take appropriate early warning action to protect pedestrian safety and lower the frequency of traffic accidents[3]. Pedestrian behaviour recognition technology has emerged as a key component of automated and assisted driving systems, helping drivers recognise hazards and make decisions by analysing and

[1] School of Computer and Information Engineering, Luoyang Institute of Technology, Luoyang 471023, China

*Corresponding author: Liu Yunqing, lyunqing2023@163.com

judging the actions of pedestrians and implementing various countermeasures. This has a significant impact on improving road safety, traffic flow, and driving efficiency [4]. Therefore, in order to meet the technical requirements of vehicle assisted driving systems, it is imperative to establish a quick and efficient pedestrian behaviour recognition algorithm.

There are two primary categories of pedestrian behaviour recognition methods: deep learning-based recognition methods and traditional behaviour recognition methods [5-7]. Conventional techniques for recognising behaviour can be classified into two categories: those that rely on human geometric features and those that use spatiotemporal points of interest. However, these methods are limited to classifying actions in basic scenes. Pedestrian behaviour recognition encounters a number of difficulties, such as scene-specific lighting variations, partially occluded pedestrians, and shifting camera angles. These characteristics frequently make it challenging for conventional behaviour recognition techniques to produce the intended outcomes. Consequently, these issues must be addressed in this study in order to provide a more solid and trustworthy pedestrian behaviour recognition system.

Deep learning-based behaviour recognition techniques are now widely used in recognition. Convolutional neural networks (CNN) have been successfully applied to speech and image recognition, leading to the proposal of several studies on CNN-based techniques for pedestrian behaviour recognition in recent years. Better classification results are obtained in [8] when the CNN models AlexNet, GoogLeNet, VGG -Net, and ResNet are applied to the images and videos, respectively. However, the long and short-term memory network in the recurrent neural network is typically used to solve the correlation problem between consecutive frames in the video, as only taking into account the CNN will ignore the continuity problem of image frames. Accordingly, [9] proposes a sort of end-to-end fully-connected long and short-term memory (LSTM) network for the human behaviour recognition technology based on human behaviour; For example, [10] applies video recognition technology for the first time by combining a CNN with a long- and short-term memory network; [11–13] combines the two networks and uses the dual-stream network to extract features in the time and space dimensions, respectively, to achieve good recognition results. blended and used in the field of video recognition technology; [11–13] successfully recognised human behaviour by combining the two networks and using a dual-stream network to extract features in the temporal and spatial dimensions, respectively. Nevertheless, the majority of behaviour recognition techniques overlook the significance of pre-processing the video stream motion target, which can somewhat increase behaviour recognition accuracy.

The objective of this research is to enhance the precision and resilience of pedestrian behaviour identification through the implementation of trajectory skeleton and multi-feature fusion. Trajectory skeleton representation is becoming a useful tool for characterising pedestrian movement as deep learning and computer vision techniques advance, and multi-feature fusion aids in taking into account the information from various features in a comprehensive manner, which helps capture the characteristics of pedestrian behaviour more thoroughly.

This paper uses a multi-feature fusion method based on trajectory skeleton, which has better adaptability and generalisation ability than traditional methods. The objective of this study is to enhance the precision of pedestrian behaviour recognition and enhance its ability to adjust to intricate and dynamic real-world situations by merging the motion data of the trajectory skeleton with the spatiotemporal relationship of various features. It is anticipated that the research findings in the areas of traffic management and intelligent surveillance will benefit society.

2 RESEARCH METHODS

Figure 1 illustrates how anomalous behaviour is identified. Initially, the trajectory and skeletal features of pedestrians are obtained using a deep learning-based target detection and tracking algorithm. After analysing the pedestrian's posture, the skeletal and trajectory features of pedestrians are fused to classify the behaviours and ultimately identify the abnormal behaviours of pedestrians in the parking lot.

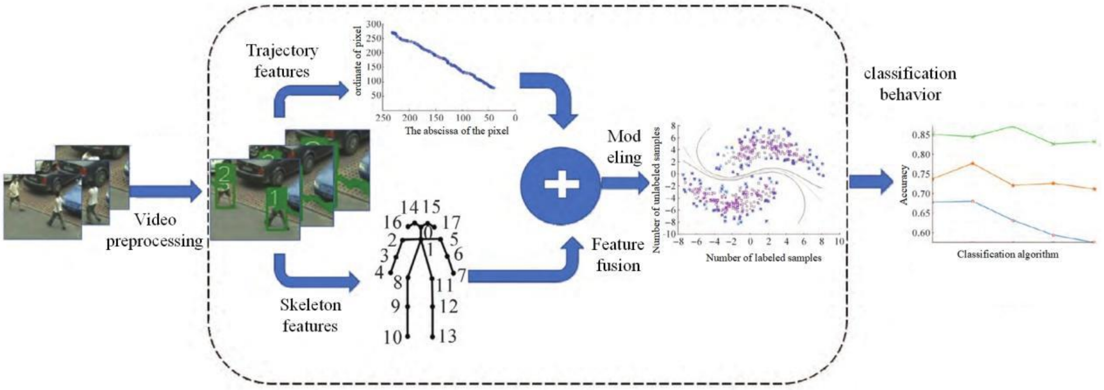


Figure 1 Abnormal behaviour identification process

2.1 Feature Extractions

Trajectory and skeleton features are chosen as the features for identifying abnormal behaviours because trajectory contains more comprehensive spatiotemporal information and has an inherent advantage in describing pedestrian targets, while skeleton features can better reflect the details of pedestrian movement and effectively express the state of pedestrian movement.

2.2.1 Feature extraction of trajectories

A trajectory is formed when a pedestrian moves along a set of center-of-mass points. To describe the pedestrian trajectory, extract each pedestrian target's position, immediate velocity, direction, and other relevant data, then form this data into a fixed-length feature vector [14]. The steps that are specific are as follows:

1) Modelling the trajectory:

$$f = \left[(x,y),(\mathrm{d}x,\mathrm{d}y),\left(\mathrm{d}^2x,\mathrm{d}^2y\right) \right] \quad (1)$$

Equation (1) denotes the positional features, directional features, and velocity features as $(x,y),(\mathrm{d}x,\mathrm{d}y)$, and $\left(\mathrm{d}^2x,\mathrm{d}^2y\right)$, respectively.

２）After extracting the position of each trajectory point, the difference data in the x and y directions contain the velocity and direction information of the target, so the difference information of the x and y axes $(\mathrm{d}x,\mathrm{d}y)$ is introduced, where

$$\begin{aligned} dx &= x_t - x_{t-1} \\ d_y &= y_t - y_{t-1} \end{aligned} \quad (2)$$

The definition of quadratic difference $(|\mathrm{d}^2x|,|\mathrm{d}^2y|)$ to represent the speed change characteristics, to provide data for the training process, and (3) The steering and speed of targets in different areas of the scene are often regular in order to detect the sudden acceleration of the traveller's abnormal behaviour [14].

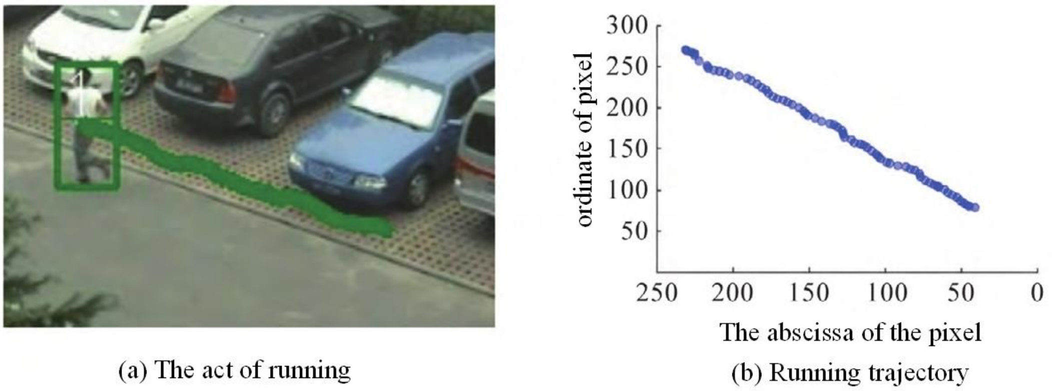
$$\begin{aligned}\mathrm{d}^2x&=x_t-2x_{t-1}+x_{t-2}\\ \mathrm{d}^2y&=y_t-2y_{t-1}+y_{t-2}\end{aligned}\quad(3)$$

4) The sample positional, directional, and velocity features of the trajectory are displayed in Table 1, which serves as an input sample for the machine learning classification algorithm that follows.

Table 1 Sample trajectories' positional, directional, and velocity characteristics

x	y	|dx|	|dy|	$\|d^2x\|$	$\|d^2y\|$
47.21	84.15	4.68	0.89	1.12	0.75
93.21	125.05	4.62	0.66	1.02	0.32
130.23	166.55	4.56	0.13	0.18	0.16
200.45	230.78	4.18	0.16	0.75	0.52
228.16	266.08	4.25	1.66	0.98	1.26

Combining the aforementioned characteristics, Figure 2 depicts the running trajectory and behaviour using pedestrian running as an example.



(a) The act of running

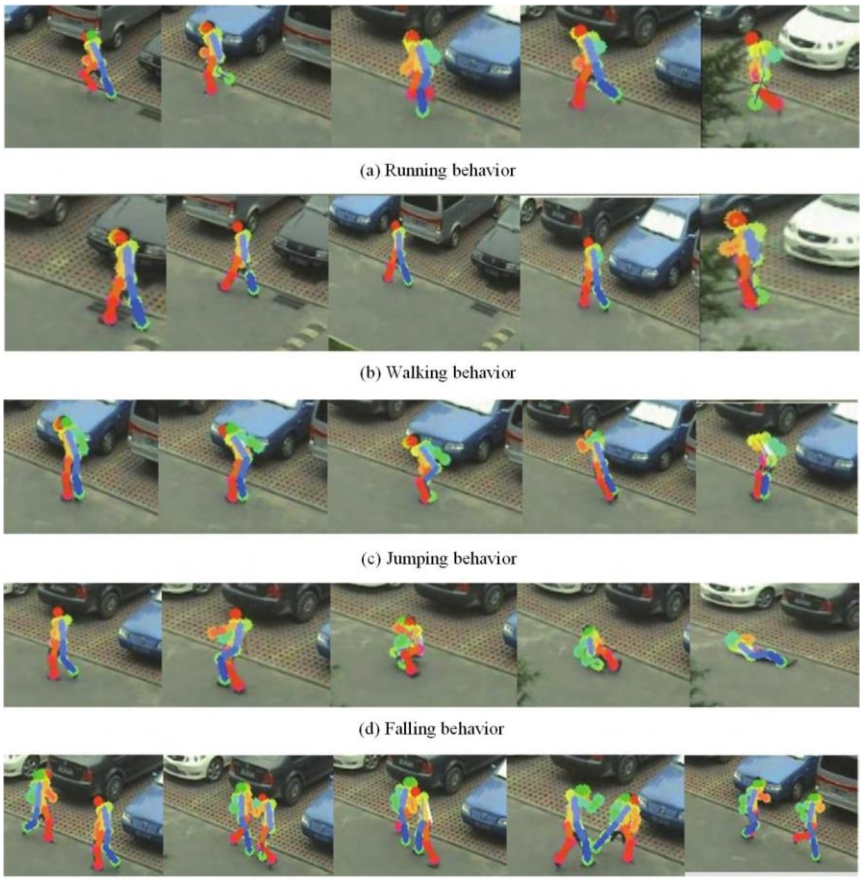
(b) Running trajectory

Figure 2 Running trajectory and behaviour of pedestrians

2.1.2 Skeleton feature extraction

Since trajectories alone cannot identify certain behaviours, like falling or jumping, the AlphaPose model [15–17] is utilised to extract information about the human skeleton and determine the coordinates of its important points.The AlphaPose model extracts the skeleton information in the following manner: First, on the basis of a single-person pose estimation structure, a symmetric spatial transformation network is presented. This network can extract the imprecise region frames in a high-quality human region; then employ data-driven to optimise the pose distance parameters and parametric pose non-maximal suppression to address the issues of detection frame positioning error and redundant detection; and lastly, use PGPG (pose-guided proposals generator) to improve the training data in order to extract the essential details for every target. to find the location of each target's skeleton key point.

The 5 behaviors—running, walking, jumping, falling, and fighting—are represented by skeleton maps in Figure 3. The AlphaPose model is used to extract the skeleton keypoints of these behaviours from the car park. Because the raw coordinates of these keypoints are unsuitable for direct use in gesture recognition, effective behavioural recognition features must be extracted from the raw features of the data.



(a) Running behavior

(b) Walking behavior

(c) Jumping behavior

(d) Falling behavior

(e) Fighting behavior

Figure 3 skeletal system representing the 5 actions of sprinting, walking, leaping, grappling, and fighting

Individual joints adjust in tandem with a person's posture, with structural vectors serving as a representation of each component's skeletal condition. Given that A (x_1, y_1) and B (x_2, y_2) stand for the left knee and left ankle coordinates, respectively, the left calf's skeletal posture, designated $V_{LKnee-LAnkle}$, is represented by vector $V_{AB} = (x_2 - x_1, y_2 - y_1)$. Four sets of structure vectors are constructed in order to describe the fall behaviour. The angle between the legs and the torso can be used to represent the fall behaviour. Additionally, the key point of the skeleton between the nose and the hip is selected to represent the torso part from the perspective of the surveillance video.

$$
\begin{aligned}
\mathbf{V}_{\text{Nose-LKnee}} &= \left(x_{\text{LKnee}} - x_{\text{Nose}}, y_{\text{LKnee}} - y_{\text{Nose}}\right) \\
\mathbf{V}_{\text{Nose-RKnee}} &= \left(x_{\text{RKnee}} - x_{\text{Nose}}, y_{\text{RKnee}} - y_{\text{Nose}}\right) \\
\mathbf{V}_{\text{LKnee-LAnkle}} &= \left(x_{\text{LAnkle}} - x_{\text{LKnee}}, y_{\text{LAnkle}} - y_{\text{LKnee}}\right) \\
\mathbf{V}_{\text{RKnee-RAnkle}} &= \left(x_{\text{RAnkle}} - x_{\text{RKnee}}, y_{\text{RAnkle}} - y_{\text{RKnee}}\right)
\end{aligned} \qquad (4)
$$

The structure vector from the nose to the left knee is represented by $\mathbf{V}_{\text{Nose-LKnee}}$ in Eq. (4), the left knee's coordinates are represented by $\left(x_{\text{LKnee}}, y_{\text{LKnee}}\right)$, the nose's coordinates are represented by $\left(x_{\text{Nose}}, y_{\text{Nose}}\right)$, and the structure vector from the nose to the right knee is represented by $\mathbf{V}_{\text{Nose-RKnee}}$;

The coordinates of the right knee are $\left(x_{\text{RKnee}}, y_{\text{RKnee}}\right)$, the left knee's coordinate is $\mathbf{V}_{\text{LKnee-LAnkle}}$, the structure vector from the left knee to the left ankle is $\left(x_{\text{LAnkle}}, y_{\text{LAnkle}}\right)$, the right knee's coordinate is $\mathbf{V}_{\text{RKnee-RAnkle}}$, and the right ankle's coordinate is $\left(x_{\text{RAnkle}}, y_{\text{RAnkle}}\right)$.

The vector angle is chosen as a motion feature to represent the various motion states of the human body because when people move, the angle of their limbs will change significantly. The cosine values of the clip angles in each of the four structural vector groups mentioned above are computed independently and serve as characteristics of the fall behaviour [18]. The cosine values of the two sets of constructed vectors' angle cosines are described as follows.:

$$
\cos\theta_{\text{Nose-LAnkle}} = \frac{\mathbf{V}_{\text{Nose-LKnee}} \cdot \mathbf{V}_{\text{LKnee-LAnke}}}{\left|\mathbf{V}_{\text{Nose-LKnee}}\right|\left|\mathbf{V}_{\text{LKnee-LAnkle}}\right|} \qquad (5)
$$

$$
\cos\theta_{\text{Nose-RAnkle}} = \frac{\mathbf{V}_{\text{Nose-RKnee}} \cdot \mathbf{V}_{\text{RKnee-RAnkle}}}{\left|\mathbf{V}_{\text{Nose-RKnee}}\right|\left|\mathbf{V}_{\text{RKnee-RAnkle}}\right|} \qquad (6)
$$

According to Equations (5)–(6), the angle between the nose and the left ankle is $\theta_{\text{Nose-LAnkle}}$; the structure vectors from the nose to the left knee and from the left knee to the left ankle are $\mathbf{V}_{\text{Nose-LKnee}}$ and $\mathbf{V}_{\text{LKnee-LAnke}}$, respectively; the angle between the nose and the right ankle is $\theta_{\text{Nose-RAnkle}}$; the structure vectors from the nose to the right knee and from the right knee to the right ankle are $\mathbf{V}_{\text{Nose-RKnee}}$ and $\mathbf{V}_{\text{RKnee-RAnkle}}$, respectively.

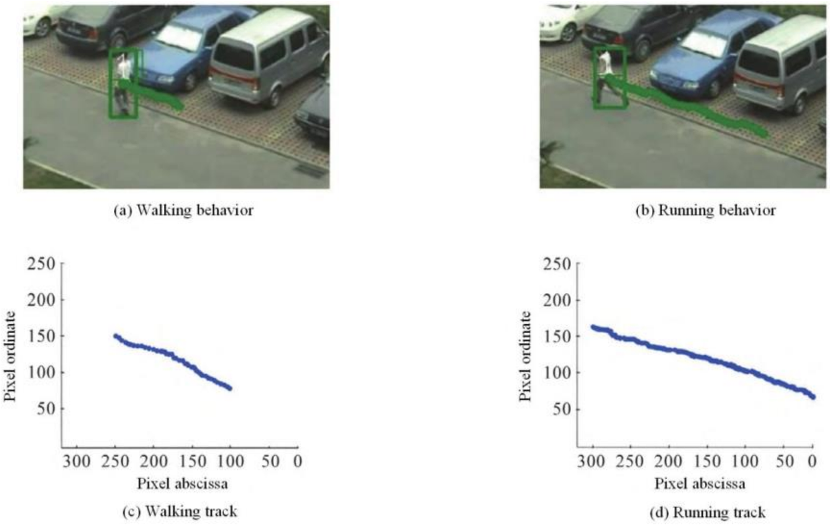
2.2 Feature fusion

The process of creating new, more effective features for classification by combining various extracted features is known as feature fusion. Feature fusion is more accurate than single feature classification because it can fully utilise the advantages of various features to form complementary features. Because the abnormal behaviours of pedestrians are so varied and complex, it is impossible for a single feature to accurately capture the type of behaviour. Therefore, fusion of skeleton and trajectory features will increase the classification accuracy of pedestrian behaviours.

2.2.1 Under-analysis of abnormal behavioural discrimination based on trajectory

After calculating Eqs. (1) to (6), some velocity and acceleration features are also obtained for each trajectory point of the travelling process during the monitoring process. These features are utilised as inputs to the classification algorithm. It is challenging to come up with a feature that can specifically represent a given activity because different activities, like walking and running, may have similar characteristics. Figure 4 compares the trajectories of walking and running. As the trajectory plots of walking and running behaviours (Figs. 4(c) & (d)) demonstrate, it is challenging to tell a behavior's trajectory whether it is walking or running.

An additional issue with relying solely on trajectories to identify behaviour is that the majority of real-world activities are composite actions rather than straightforward, repetitive ones. It is challenging to characterise composite activities solely on the basis of trajectories because most composite activities are connected by a number of straightforward actions. It is evident that additional features are required in addition to trajectory features in order to accurately depict human behaviour.


(a) Walking behavior

(b) Running behavior

(c) Walking track

(d) Running track

Figure 4 Comparison of walking and running trajectories

2.2.2 Under-analysis of abnormal behavioural discrimination based on skeletons

In parking lots, there are a variety of strange behaviours that occur, like chasing and falling. Some characteristics of these abnormal behaviours are shared by all of them. For instance, falling actually occurs as a walking-falling-state sequence, and it usually happens very quickly—sometimes in the span of a few seconds. Similarly, jumping involves the following steps: walking, jumping, walking, and jumping. The number of abnormal and normal behaviours is highly unevenly distributed because abnormal behaviour is abrupt, the fall or jump process lasts only a short while, and during the actual monitoring, the pedestrian may be walking or running normally prior to the fall or jump. The percentage of abnormal behaviours in some samples is shown in Table 2, from which it can be seen that the number of frames of abnormal behaviours does not account for a high percentage in the whole video. Three frames of each image of jumping behaviour and falling behaviour at different time periods are intercepted from the surveillance video, as shown in Fig. 5, from which it can be seen that the abnormal behaviour occurs only at a particular time period. In surveillance video, it is necessary to identify both normal and abnormal behaviours at the same time, and it is difficult to meet the realistic requirements by only taking the skeleton as the characteristic behaviour.

Table 2 Percentage of aberrant behaviour in a sample of choice

Video Category	Normal behavior/frame	Abnormal behavior/frame	Total number/frame	The proportion of abnormal behavior
Jumping Video 1	85	52	130	0.387
Jumping Video 2	145	70	220	0.352
Jumping Video 3	120	60	185	0.355
Jumping Video 4	105	55	160	0.362
Falling Video 1	195	85	284	0.356
Falling Video 2	182	92	278	0.362
Falling Video 3	130	75	210	0.324
Falling Video 4	265	114	385	0.421

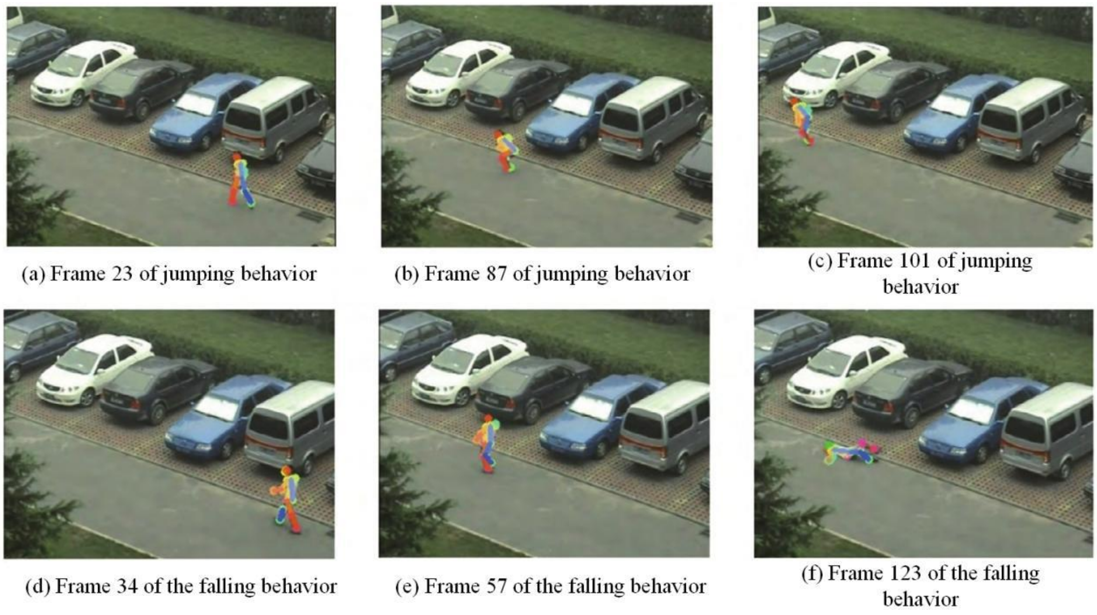

(a) Frame 23 of jumping behavior
(b) Frame 87 of jumping behavior
(c) Frame 101 of jumping behavior
(d) Frame 34 of the falling behavior
(e) Frame 57 of the falling behavior
(f) Frame 123 of the falling behavior

Figure 5 3 frames of each image of jumping behaviour and falling behaviour

The skeleton features are added as an extra technique to identify abnormal behaviours. While this study defines features like velocity and acceleration to form the training samples for potential abnormal behaviours, a variety of abnormal behaviours may occur in reality. Skeletal characteristics by themselves are also unable to satisfy the demands of reality, as abnormal behaviours like jumping happen abruptly and without continuity, and there is no discernible shift in a person's posture prior to or following the abnormal behaviour. Therefore, using a fusion model of skeleton features and pedestrian trajectory features is being considered to identify abnormal and normal behaviours in order to meet the realistic needs.

2.2.3 Feature Fusion Methods

The skeleton feature can only momentarily reflect the pedestrian's action, and because the video information is more complex, it is easier to use the trajectory as a feature. Consequently, when classifying the abnormal behaviour, feature fusion will produce a better classification effect. The fusion of trajectory features and skeleton features is used to identify the abnormal behaviours of pedestrians because, in contrast, the abnormal behaviours of pedestrians in car park surveillance videos are frequently composed of composite activities, and a single trajectory feature or skeleton feature cannot describe the whole process of the pedestrian's activities. The skeleton models of Eqs. (1) and (2) to (6) are fused to form a new trajectory skeleton model, as shown in Eq. (7), in order to address the aforementioned issues. An algorithm is proposed to fuse pedestrian trajectory and skeleton features:

$$f = \left[x, y, \mathrm{d}x, \mathrm{d}y, \mathrm{d}^2x, \mathrm{d}^2y, \cos\theta_{\text{Nose-LAnkle}}, \cos\theta_{\text{Nose-RAnkle}} \right] \quad (7)$$

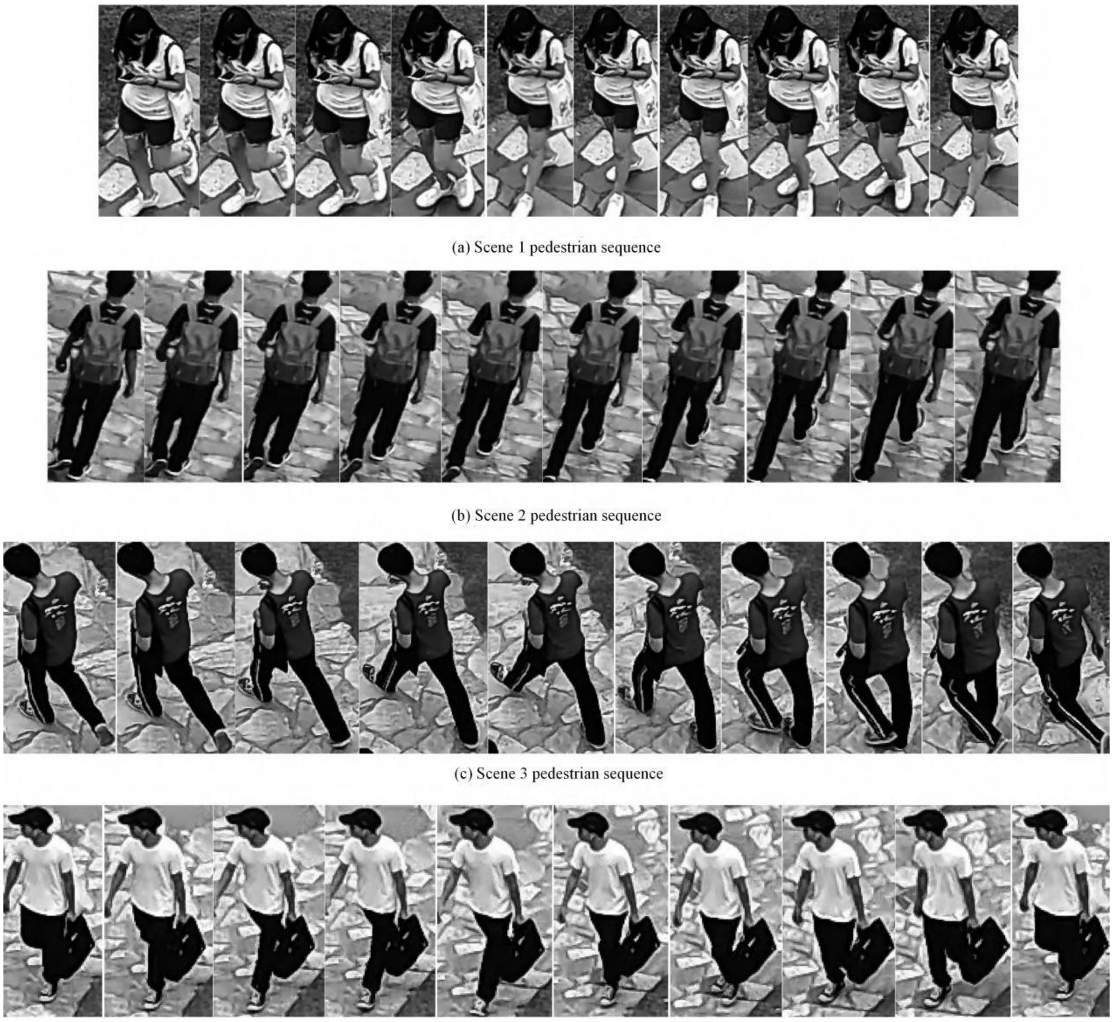
Table 3 displays a selection of the classification samples following feature fusion.

3 EXPERIMENTATION

3. 1 Introduction to the dataset

A sequence-image project set dataset of pedestrian attribute recognition is created in four distinct campus scenarios to confirm that the sequence-based pedestrian attribute recognition network performs better than the single-image-based pedestrian attribute recognition network.The dataset generated in this work is a segment of a variable-length pedestrian sequence that corresponds to a multi-classification label; that is, the chosen pedestrian attributes are maintained constant within a segment of a variable-length pedestrian sequence. Four distinct campus surveillance scenes provide the original data for the sequential pedestrian attribute recognition dataset. The surveillance cameras are positioned higher in each scene, so there is no overlap between them. There are a total of 653 pedestrian sequences of varying lengths in the four scenes, with an average length of 153 frames. There are 89 pedestrian sequences in Scene 1, 213 pedestrian sequences in Scene 2, 185 pedestrian sequences in Scene 3, and 166 pedestrian sequences in Scene 4.

The following 32 typical pedestrian attributes were chosen for annotation: hat, backpack, long hair, short hair, male, female, satchel, short sleeves on top, long sleeves on top, short dress on bottom, dress colour on top (black, blue, brown, green, grey, orange, pink, purple, red, white, yellow), and dress colour on bottom (black, blue, brown, green, grey, orange, pink, purple, red, white, yellow). This makes a total of 32 common pedestrian attributes. The distribution of the training and test sets was 6:4. As illustrated in Fig. 6, a shortened sequence of pedestrians from each of the four distinct campus scenarios is chosen as an example for illustration.



(a) Scene 1 pedestrian sequence

(b) Scene 2 pedestrian sequence

(c) Scene 3 pedestrian sequence

(d) Scene 4 pedestrian sequence

Figure. 6: Four distinct scenarios with examples of pedestrian sequences

3 2 Experimental setup

The dataset's organisation: The training set and the test set are the two sections that make up the dataset. The uniform sampling method is used to select the input images for the training set from the sequences, and the images sampled out at equal intervals serve as the network's data input. Using uniform sampling, 16 frames of images are chosen as network input for every pedestrian sequence of arbitrary length. Cyclic replication is used to satisfy the requirement for pedestrian sequences with fewer than 16 frames. A uniform sampling of 16 frames is also selected to be the network's input for the test set. During the testing phase, the entire sequence is divided into multiple uniformly sampled image sets, each of which is fed into the network for recognition in order to fully compare the performance of the pedestrian attribute recognition network under various feature fusion and temporal modelling approaches.

Evaluation metrics: The multi-label classification task of sequence-based pedestrian attribute recognition uses average accuracy (mA), average check accuracy (m-prec), average check completeness (m-rec), and average F1 score (m-F1) as performance metrics to compare the effectiveness of various approaches in a thorough manner.

3. 3 Experiments with different input frame rates

In this paper, we adopt the uniform sampling method to feed the network with images sampled at equal intervals instead of the entire video sequence. Selecting fewer frames results in a larger sampling interval, larger image differences, and a comparatively smaller set of images that the network can refer to. Conversely, selecting more frames results in a smaller sampling interval and smaller image differences, but it also gives the network access to a larger image pool for learning. Table 3 shows that increasing the number of input frames from 4 to 16 frames enhances the network's overall performance when considering the four evaluation metrics taken together. At 16 input frames, the network performs at its best and achieves the highest ratings across all evaluation metrics.

Table 3 Experimental results comparing different input frame rates

Number of frames	mA(%)	m—prec(%)	m-rec(%)	m-F1(%)
4	81.25	83.85	77.28	80.15
8	81.22	86.12	77.35	81.45
12	81.63	85.44	78.05	81.56
16	81.86	86.65	78.25	82.25
20	80.96	87.65	76.32	81.45

When the number of input frames is increased to 20, all performance metrics decrease, with the exception of the average checking accuracy rate, which improves by 0. 98% when compared to the input of 16 frames. The average F1 score decreases by 0. 8%, the mA decreases by 0. 87%, and the average checking accuracy rate decreases by 2. 21%. To summarise, this paper proposes a multi-feature fusion pedestrian sequence image attribute recognition network with a temporal attention mechanism. The network accepts 16 input frames.

3.4 Trials utilising varying convolutional kernel depths in the temporal dimension

Compared to 2D convolution, 3D convolution has an extra depth channel in the temporal dimension.It can be understood that more frames are taken into consideration the larger the value assigned to the 3D convolution's time dimension. The null-temporal 3D convolutional attention factor-weighted feature aggregation branch conducts an experimental comparison of various values of the depth of the temporal dimension of a portion of the 3D convolutional layer. The values that were used in the experiments are 1, 3, and 5, respectively, and the results are shown in Table 4.

Table 4 Convolutional kernel experiment results at various time dimension depths

Depth of time dimension	mA(%)	m—prec(%)	m-rec(%)	m-F1(%)
Size=1	80.25	86.65	74.55	80.12
Size=3	86.65	85.65	75.28	82.12
Size=5	74.55	86.35	77.56	81.26

It is evident that the network performs worst when the depth of the time dimension is 1, as at this point the network does not take into account the relationships between neighbouring frames enough. The network performs worse when the depth of the time dimension is set to 5 than it does when it is set to 3. As compared to the optimal network performance, which occurs when the value is 3, the average F1 metrics and performance in mA are 0.36

and 0.41 percent, respectively. Furthermore, the network's overall performance remains unchanged despite the fact that there are more network parameters with a value of 5 than those with a value of 3, making training more challenging. In conclusion, the convolutional kernel's depth in the temporal dimension is set to three in this work.

3.5 Experiments with different feature aggregation and temporal modelling approaches

The frame-level feature extractors in the single-image-based pedestrian attribute recognition method and the sequence-based pedestrian attribute recognition method in the experiments are composed based on ResNet50 in order to fairly compare the two methods' pedestrian attribute recognition techniques.

Table 5 shows the performance of sequence-based pedestrian attribute recognition networks with multiple feature aggregation and temporal modelling approaches. As predicted, all the sequence-based pedestrian attribute recognition methods outperform the single-image-based pedestrian attribute recognition methods. The two temporal modelling approaches, RNN and LSTM, do not perform well in the pedestrian attribute recognition task. Although the overall performance is improved, it even decreases in the average detection rate, which may be due to the fact that there is no strong causality between multiple frames and the training of RNN and LSTM is difficult in the sequence-based pedestrian attribute recognition task.

Table 5 Experimental results of different feature aggregation and temporal modeling methods

Method	mA(%)	m—prec(%)	m-rec(%)	m-F1(%)
Based on a single image	77.85	80.25	70.42	75.52
RNN	79.05	77.85	73.25	75.66
LSTM	79.85	78.52	74.25	76.65
Average pooling	80.26	81.22	75.26	78.28
Maximum pooling	81.25	85.25	77.74	71.15
Reference [19] Method	81.25	85.58	77.45	81.26
Method of this article	81.26	86.65	78.28	82.35

Simple temporal maximum pooling and temporal average pooling do not require additional network parameters and can achieve better recognition results, but these two still have problems: when fusing multiple frame-level features into a sequence of features, they do not sufficiently take into account the connection between frames and do not determine the importance between multiple frame images, losing a large amount of valuable information. Both the temporal attention mechanism proposed in the literature [19] and the method proposed in this paper overcome the above shortcomings and achieve better pedestrian attribute recognition results than average pooling and maximum pooling. The last row in Table 5 shows the performance effect of the proposed method in this paper, and it can be seen that the multi-feature fusion pedestrian sequence image attribute recognition network combined with temporal attention mechanism proposed in this paper is the highest in each of the metrics, in which the metrics are elevated compared to the single-image based pedestrian attribute recognition method. In addition, when the method proposed in this paper is compared with the temporal attention mechanism proposed in the literature [20], it can be seen that the method proposed in this paper is higher in each of the evaluation indexes confirming the effectiveness of the framework and method proposed in this paper[21,22,23].

4 CONCLUSION

This study proposes a multi-feature fusion of trajectory skeleton-based pedestrian behaviour recognition method that integrates multi-feature information of both trajectory and skeleton to improve pedestrian behaviour recognition robustness and accuracy. Experimental results show that the method has better performance in dealing with composite behaviour and sudden abnormal behaviour. Future research can further optimise the feature extraction and fusion methods to adapt to more complex real-world scenarios, and explore more advanced deep learning models to improve recognition performance.

REFERENCES

[1] Qiu, S., Zhao, H., Jiang, N., Wang, Z., Liu, L., An, Y., ... & Fortino, G. (2022). Multi-sensor information fusion based on machine learning for real applications in human activity recognition: State-of-the-art and research challenges. *Information Fusion*, *80*, 241-265.

[2] Hou, R., Wang, Z., Ren, R., Cao, Y., & Wang, Z. (2023). Multi-channel network: Constructing efficient GCN baselines for skeleton-based action recognition. *Computers & Graphics*, *110*, 111-117.

[3] Qu, Y., Li, X., Qin, Z., & Lu, Q. (2022). Acoustic scene classification based on three-dimensional multi-channel feature-correlated deep learning networks. *Scientific Reports*, *12*(1), 13730.

[4] Li, Q., Gravina, R., Li, Y., Alsamhi, S. H., Sun, F., & Fortino, G. (2020). Multi-user activity recognition: Challenges and opportunities. *Information Fusion*, *63*, 121-135.

[5] Ounoughi, C., & Yahia, S. B. (2023). Data fusion for ITS: A systematic literature review. *Information Fusion*, *89*, 267-291.

[6] Hafeez, F., Sheikh, U. U., Alkhaldi, N., Al Garni, H. Z., Arfeen, Z. A., & Khalid, S. A. (2020). Insights and strategies for an autonomous vehicle with a sensor fusion innovation: A fictional outlook. *IEEE access*, *8*, 135162-135175.

[7] Wang, Y., Lu, T., Zhang, Y., Fang, W., Wu, Y., & Wang, Z. (2021). Cross-task feature alignment for seeing pedestrians in the dark. *Neurocomputing*, *462*, 282-293.

[8] Wang, Z., Zhang, W., & Zhou, H. (2019). Perception-guided multi-channel visual feature fusion for image retargeting. *Signal Processing: Image Communication*, *79*, 63-70.

[9] Zhang, P., & Zhang, J. (2022). Deep learning analysis based on multi-sensor fusion data for hemiplegia rehabilitation training system for stoke patients. *Robotica*, *40*(3), 780-797.

[10] Fu, X., Yu, G., & Liu, Z. (2021). Spatial–temporal convolutional model for urban crowd density prediction based on mobile-phone signaling data. *IEEE Transactions on Intelligent Transportation Systems*, *23*(9), 14661-14673.

[11] Noori, F. M., Riegler, M., Uddin, M. Z., & Torresen, J. (2020). Human activity recognition from multiple sensors data using multi-fusion representations and CNNs. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, *16*(2), 1-19.

[12] Xu, L., Yan, S., Chen, X., & Wang, P. (2019). Motion recognition algorithm based on deep edge-aware pyramid pooling network in human–computer interaction. *IEEE Access*, *7*, 163806-163813.

[13] Nazeer, S., Sultana, N., & Bonyah, E. Cycles and Paths Related Vertex-Equitable Graphs. Journal of Combinatorial Mathematics and Combinatorial Computing, 117, 15-24.

[14] Jingchun Zhou, Boshen Li, Dehuan Zhang, Jieyu Yuan, Weishi Zhang, Zhanchuan Cai. "UGIF-Net: An Efficient Fully Guided Information Flow Network for Underwater Image Enhancement," IEEE Transactions on Geoscience and Remote Sensing, vol. 61, pp. 1-17, 2023, Art no. 4206117, doi: 10.1109/TGRS.2023.3293912.

[15] Ali, J., Shan, G., Gul, N., & Roh, B. H. (2023). An Intelligent Blockchain-based Secure Link Failure Recovery Framework for Software-defined Internet-of-Things. Journal of Grid Computing, 21(4), 57.

[16] Wang, S., Cao, J., & Philip, S. Y. (2020). Deep learning for spatio-temporal data mining: A survey. *IEEE transactions on knowledge and data engineering*, *34*(8), 3681-3700.

[17] Liu, Y., Zhang, Q., & Chen, W. (2021). Massive-scale complicated human action recognition: Theory and applications. *Future Generation Computer Systems*, *125*, 806-811.

[18] Wang, J. T., Yan, G. L., Wang, H. Y., & Hua, J. (2018). Pedestrian recognition in multi-camera networks based on deep transfer learning and feature visualization. *Neurocomputing*, *316*, 166-177.

[19] Gao, S., Tan, A. H., & Setchi, R. (2019). Learning ADL daily routines with spatiotemporal neural networks. *IEEE Transactions on Knowledge and Data Engineering*, *33*(1), 143-153.

[20] Zhang, J., Yin, Z., Chen, P., & Nichele, S. (2020). Emotion recognition using multi-modal data and machine learning techniques: A tutorial and review. *Information Fusion*, *59*, 103-126.

[21] Chen, M., Banitaan, S., & Maleki, M. (2023). Enhancing Pedestrian Group Detection and Tracking Through Zone-Based Clustering. *IEEE Access*, *11*, 132162-132179.

[22] Berroukham, A., Housni, K., Lahraichi, M., & Boulfrifi, I. (2023). Deep learning-based methods for anomaly detection in video surveillance: a review. *Bulletin of Electrical Engineering and Informatics*, *12*(1), 314-327.

[23] Qin, L., & Kang, L. (2018). Application of video scene semantic recognition technology in smart video. *Tehnički vjesnik*, *25*(5), 1429-1436.