Real Time Gesture (Fall) Recognition of Traffic Video Based on Multi-resolution Human Skeleton Analysis

Xinchen Xu[1], Xiaoqing Zeng[1], Yizeng Wang[2(✉)], and Qipeng Xiong[3]

[1] The Key Laboratory of Road and Traffic Engineering, Ministry of Education, Tongji University, No. 4800 Cao'an Road, Shanghai 201804, China
[2] Shanghai University, Shanghai, China
ll02790744@qq.com
[3] Shanghai FR Traffic Technology Limited Corporation, Shanghai, China

Abstract. The objective of this study was to detect abnormal behavior event (fall) by analyzing the existing monitoring video for ensuring the safety of rail transit platform passengers. Here, the key point coordinates and limb joints of human body are obtained based on PAFs (part affinity fields). Then feature information determined the fall is extracted based on the neck key point tracking algorithm. The feature for judging fall, namely the angle between leg and horizontal plane, is proposed in this study. The experimental data is based on the simulated fall videos took from Leqiao station of Soochow 1st metro line in Jiangsu province. Our results show that, the single picture and video are tested respectively, and it is found that the timing information is more fault-tolerant and more accurate in identifying falls, yet is more complex and more difficult to implement. What's more, because of the lack of sufficient fall videos, the results analysis based on the proposed algorithm remains much room for improvements. Also, more useful and efficient detection characteristics can be taken into account in the future.

Keywords: Gesture recognition · Rail transit · Video surveillance · Fall detection

1 Introduction

Falls are rare, while not timely detection of falls poses a great threat to personal safety and health. In public area, it may trigger panic, crowd gathering, fleeing stampede and other larger mass incidents. With the rapid development of computer vision, pose estimation is attracting more and more attention. Fall recognition, as a special case of human abnormal posture, has become a research hotspot.

Some methods used in fall recognition to date include two methods based on non-visual and visual. Non-visual detection methods mainly rely on wearable sensing device to detect acceleration signals when people fall. However, wearing those sensors on the body brings certain inconvenience to the action. This method is mostly used for the supervision of the elderly or the patient, and is not suitable in daily life.

X. Zeng et al. (Eds.): ITASC 2019, SIST 127, pp. 1–12, 2019.
https://doi.org/10.1007/978-981-13-7542-2_1

Now the mainstream approach is based on visual information such as video stream analysis based on depth camera, multi-camera and single fixed vision camera. With its depth information, the depth camera has more information and better effect than ordinary cameras, but its cost is higher and generality is not high. Multi-camera and depth cameras are very complex and costly, not suitable for those daily scenes. Monocular cameras are relatively simple and inexpensive to obtain. This article uses surveillance video from the single fixed-view camera. It can make full use of existing resources, and the results is expected to be applied to the reality.

Machine learning has been applied to more and more human pose recognition technologies. Tompson, Jain and others use the hybrid structure of depth convolution neural network and Markov random domain to recognize human pose [1]. Aslan, Sengur and others propose a fall detection system based on depth camera, where shape and support vector machine are used to distinguish falls from other daily behaviors [2]. The core technology in human behavior recognition is divided into three steps: human segmentation, feature extraction and representation, activity detection and classification. At present, the related methods of gesture recognition are mainly classified according to the models established. Non-model-based technology mainly extract features, and many body modeling are still 2D technology, which can obtain posture information more easily after labeling the human part [3].

PAFs (part affinity fields) [4] uses a non-parametric representation [4], learning to associate body parts with individuals in the image. It uses a bottom-up approach, which first regresses to the key points of everyone, and then divides the key points so that the key points can be assigned to everyone. The greedy parsing algorithm is sufficient to produce high-quality body posture analysis, even if the number of people in the image increases, the efficiency will remain efficient, which has a good application prospect for the scene of the mass transit platform. The real-time algorithm detects the two-dimensional pose of multiple people in the image, ensuring the timeliness of security monitoring and facilitating rapid rescue in emergency situations.

There are five main features of fall detection: aspect ratio (AR) [5], AR change, fall angle [6], center speed [6], and head speed [5]. These features all make full use of the posture difference between normal human and falling human. However, there are still few fall recognitions combined with the bone information of the human body. In order to identify the passenger's fall on the platform of rail transit, this paper uses daily surveillance video captured by the existing platform monitoring cameras. It is based on the current advanced bone extraction algorithm—PAFs (part affinity fields) [4], and analysis innovatively from the passenger's leg angle characteristics, through real-time calculation of the traveler's two thighs angle value and timing information.

2 Human Key Points Acquisition Based on Multi-resolution Part Affinity Fields Model

In this paper, the extraction of human key points and the connection of limbs and trunks are based on PAFs (part affinity fields) [4]. This method uses frame-by-frame processing to extract human skeleton without human detector in real time. Because of the large depth of rail transit platform, there may be a different number of available key point information in the surveillance video. Therefore, a pedestrian attitude database of rail transit is established from video images, and different number of key points are calibrated for different depth of field pedestrian instances. As shown in Fig. 1(a), (b), (c) correspond to the calibration of key points of human body in close, middle and long-range respectively, which can effectively distinguish the information of people in different depth of field. The original model is retrained using the database, and a new model with better recognition ability for rail transit scene is finally obtained. This paper uses this new model.

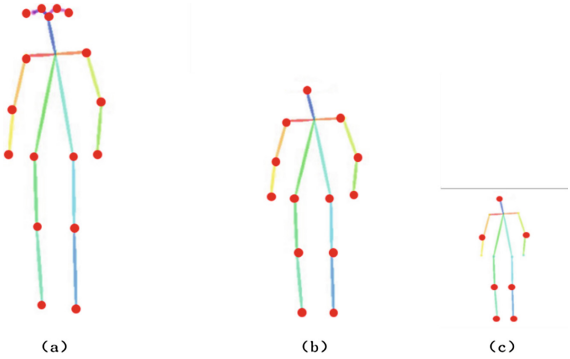

Fig. 1. Calibration of human key points in different depths of field

3 Fall Recognition Based on Multiresolution Part Affinity Fields Model

3.1 Leg–Horizontal Plane Angle Characteristics

Decision Rule. When a person falls, he will be tilted or even lie down. At this time, his legs will have a certain tilt angle, while the normal standing people's legs are almost upright, and for walking people, the angle will not be too large. In this paper, the angle of the two key points connecting the hip and knee, that is, the tilt angle of the thigh, is taken as the judgment condition of the fall, and the angles of the left and right

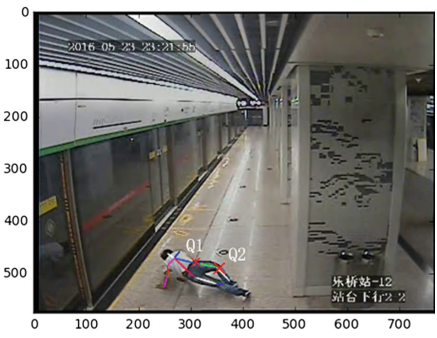

Fig. 2. Calculation points of leg angle

legs are calculated respectively. The formula for calculating the angle value of one leg is as follows:

$$\theta = \tan^{-1}[(y_{Q2} - y_{Q1}) \div (x_{Q2} - x_{Q1})] \tag{3.1}$$

x_{Q1}, y_{Q1}, x_{Q2}, y_{Q2} were calculated as the horizontal and vertical coordinates of the key points of the hip and knee on one side of the body. The specific instructions are shown in Fig. 2.

Table 1. Angle statistics of six figures

left and right leg angle / people / Figures	1	2	3	4	5	6	7	8
Fig.3	[93.37,91.74]	[-32.91,83.66]	[92.29,100.49]	[158.15,156.47]	[75.96,0.0]	[107.93,90.0]	158.58,92.49]	[87.61,*354.29*]
Fig.4	[91.51,93.27]	[98.53,101.89]	[64.23,85.76]	[95.19,95.19]	[70.71,59.93]	null	null	null
Fig.5	[88.49,93.27]	[97.60,99.78]	[98.13,92.73]	[90.0,103.67]	[87.14,98.53]	[95.19,100.78]	[85.24,-11.56]	[95.71,90.0]
Fig.6	[90.0,91.64]	[86.99,0.0]	[99.46,105.95]	[90.0,86.99]	[45.81,109.98]	[90.0,96.12]	null	null
Fig.7	[96.34,95.19]	[94.24,90.0]	[90.0,83.66]	[88.41,14.28]	[90.0,100.78]	[95.19,120.65]	null	null
Fig.8	[97.50,91.74]	[85.43,92.20]	[27.82,25.71]	107.35,112,75]	[92.73,100.78]	null	null	null

Six different pictures, each containing a fallen person, were skeletally extracted to calculate the angle of the person's left and right thighs. The green and white lines represented the person's left and right thighs respectively. As shown in Table 1, the left and right leg angles with corresponding serial numbers in each drawing are listed (the persons in each drawing are independently numbered), corresponding to Figs. 3, 4, 5, 6, 7 and 8 respectively, in which the red data is the angle values of the fallen persons' two legs.

Fig. 3.

Fig. 4.

Fig. 5.

Fig. 6.

Fig. 7.

Fig. 8.

It is found that the leg angle of normal erect pedestrians is about 90° and that of fallen pedestrians is quite different from 90°.

Result Evaluation. Using the abnormal angle of the leg for judging falling accords with the general law. It can accurately determine whether the person in the picture is in the abnormal attitude of falling or normal standing posture.

3.2 Falling Track Based on Continuous Frames

The original human skeleton extraction algorithm PAFs (part affinity fields) [4] uses frame-by-frame real-time processing for video without human tracking. Considering that the fall behavior includes three states: before fall, during fall, after fall, the corresponding state of leg angle will be from normal to abnormal, finally to normal, if such information can be obtained from the video frame, it can be more accurate to determine the presence of a fall in this surveillance video.

Fig. 9. Tracking through key coordinates of the neck in consecutive frames

As shown in Fig. 9, in the four consecutive frames, the necks of A and B are marked in red. It can be seen that the distance between the adjacent frames is very short. In the current video frame, if there is a neck coordinate in the neighborhood of the neck key point coordinate in the previous frame, the two will be judged to belong to the same person.

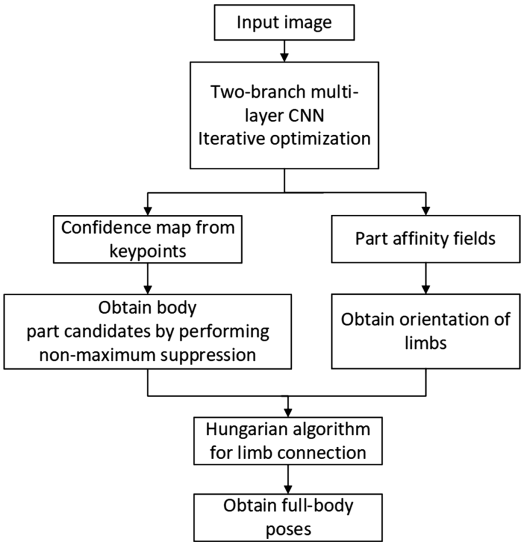

Fig. 10. Algorithm flow of PAFs [4]

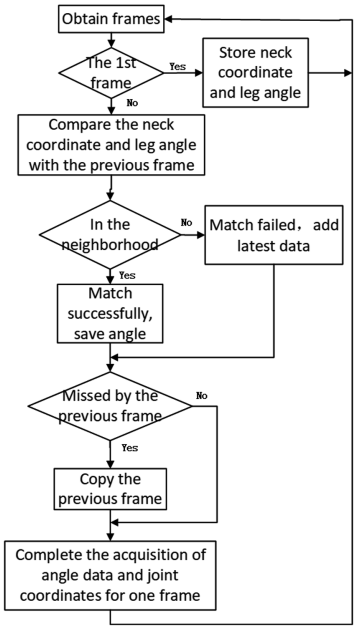

Fig. 11. Algorithm flowchart

Based on PAFs (part affinity fields) [4], this paper extracts the key points of human body and the information of limb connection. The algorithm flow of PAFs (part affinity fields) [4] is shown in Fig. 10. After that, the human is tracked by the position coordinates of the neck, and the fall behavior is monitored from the tilt angle of the thigh.

Through cyclic reading video stream, the above method can track the same person, and store the angle of each frame of each person, then estimate the attitude by calculating the average angle of each frame in real time for a period of time. If numerical abnormalities occur repeatedly, it is judged that a fall occurs in the surveillance video. The algorithm flowchart is shown in Fig. 11.

4 Experimental Results

4.1 Single Frame Fall Judgment

Recognition Criteria. Before recognizing a fall, we should analyze the difference between the body's thigh angle value in normal and fall. Figure 12 is a few pictures containing these two postures, and Fig. 13 is the leg angle data of all the video frames corresponding to the above postures. It can be found in the picture that the person is in a fall from more than 50 frames to more than 300 frames. And the leg angle from the scatter plot in Fig. 13 can be found to deviate from 90 while the value of the other frames is around 90.

Fig. 12. Falling keyframe

By analyzing the characteristics of angle values in different states, this paper determines the classification rule of behavior in a single picture as follows:

(a) normal

$$|\text{leg angle} - 90^{\circ}| < 30^{\circ};$$

(b) fall down

$$|\text{leg angle} - 90^{\circ}| >= 30^{\circ}.$$

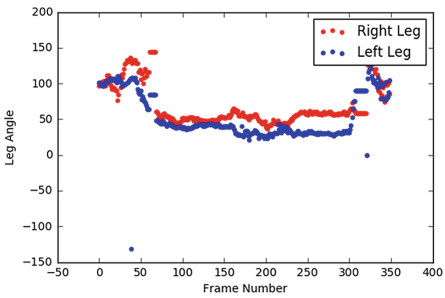

Fig. 13. Leg angle of all frames

Results. In this paper, 76 pictures were selected from fall videos, including 60 fall pictures and 16 normal posture pictures. Some pictures contain a single person and some have a number of people. On the premise of extracting human skeleton correctly, there are 60 fallen people and 117 normal people in these 76 pictures. Based on the above criteria, the final results are summarized as shown in Table 2.

Table 2. Gesture recognition results

Actual posture	Decision posture		
	Judged as fall	Judged to be normal	Accuracy
Falling	50	10	Fall detection 83.33%
Normal	7	110	Error alarm 5.98%

Result Analysis

(1) fall is judged to be normal

The main reasons for falling judged as normal posture are:

(a) the particularity of falling posture.

As shown in Fig. 14, the falling person's upper body bends, while the lower limbs are still upright. This paper judges the fall on the basis of legs bending, so it can be misjudged.

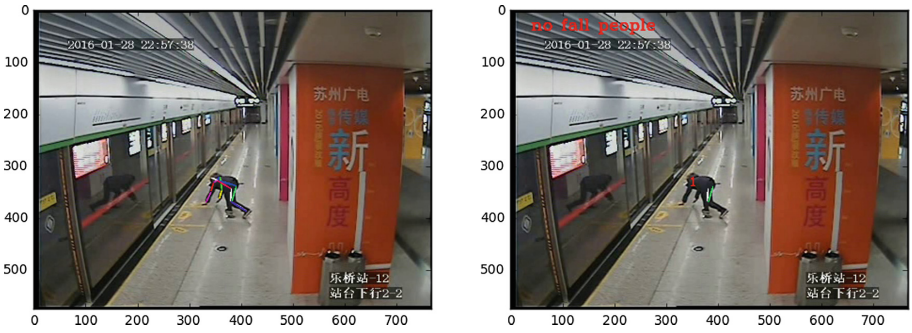

Fig. 14. Special fall posture

(b) the particularity of fall direction.

As shown in Fig. 15, the human body is in the same direction as the surveillance shooting, resulting in this method cannot distinguish the change of leg angle in this direction, although his two legs have changed from vertical to horizontal. So, it cannot identify the fall.

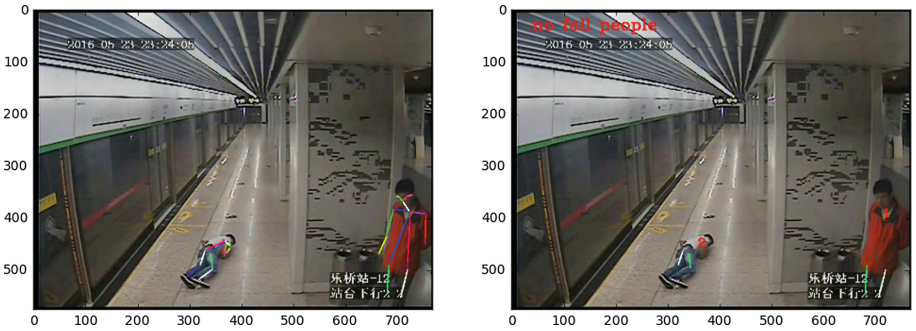

Fig. 15. Special fall direction

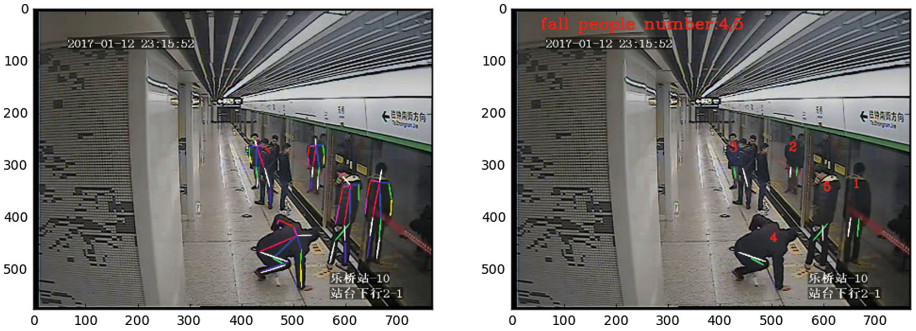

Fig. 16. Normal people misjudged

(2) normal posture is judged as fall

The main reasons for normal posture misjudged are:

(a) the particularity of normal standing posture

As shown in Fig. 16, the leg of standing person (marked as 5) is not completely upright, resulting in misjudgment. However, this attitude is very common in daily life, so error rate of relying only on a single picture for fall recognition is high. In contrast, determining from multiple pictures and multi-frame data is more reliable. Detailed results about real-time fall determination can be seen in the next section.

(b) imperfect posture recognition model

The attitude recognition algorithm is not perfect enough. As shown in Fig. 17, due to the depth of platform, the judgment of the leg angle of the person standing in the distance is wrong, which results in the algorithm judging the person wrongly as a fall.

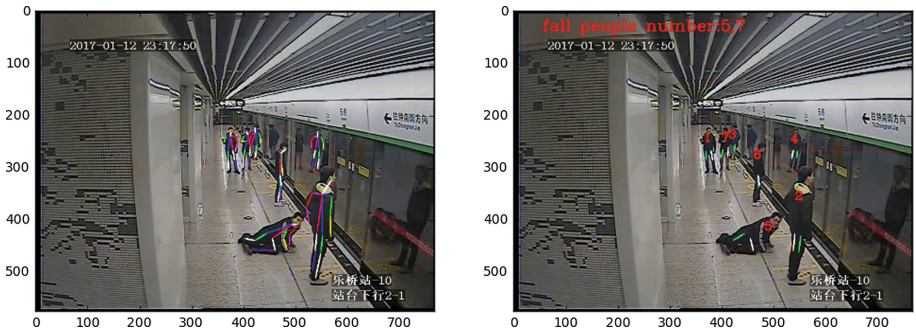

Fig. 17. Distant people misjudged

4.2 Real-Time Fall Judgment Through Continuous Frame

Judging a fall based on a single picture lacks fault tolerance, and some occasional special gestures can easily affect the final result. Therefore, this section analyzes the video frame by frame, and estimates the attitude by tracking the neck key points. In this paper, two videos were tested. The results show that the alarm can be sent out within two seconds of a fall, and the fall event can be accurately determined in the current video.

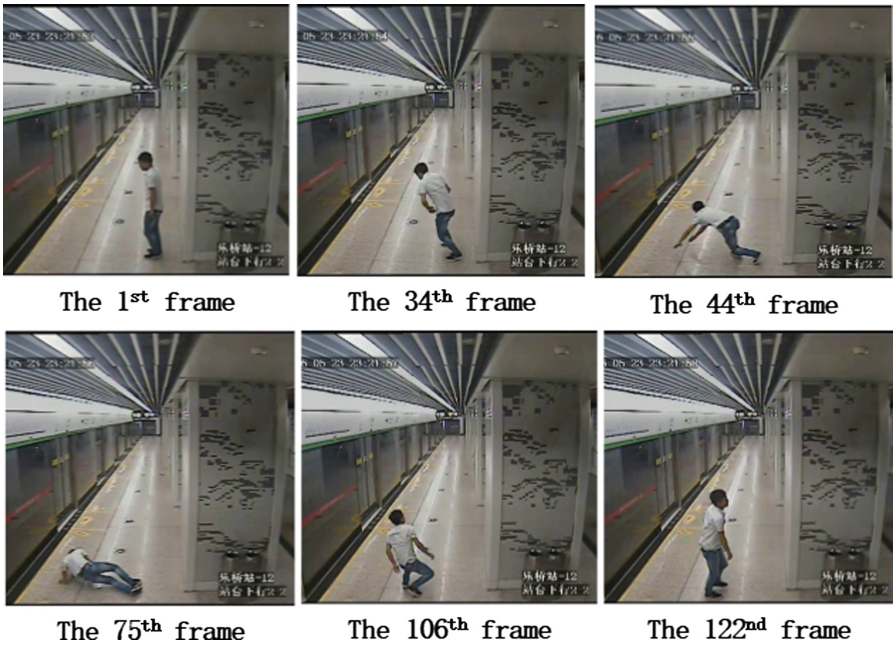

Fig. 18. Falling keyframes in the 1st video

Figure 18 is several key frames in the first video that containing falling. The corresponding leg angle values are shown in Fig. 19. It can be seen that the whole fall process is from about 40th frames to 150th frames.

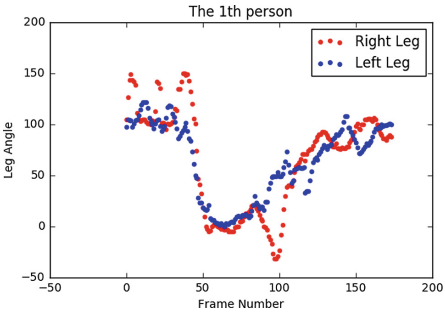

Fig. 19. Complete date

In this paper, the human leg data is judged in real time, and the alarm is sent in two seconds. As shown in Fig. 20(a), real-time decision data shows that an early warning is issued in frame 100, that is, a decision is made successfully by relying on the first 100 frames of data. Figure 20(b) is the 100th frame, and the person labeled 1 is decided as falling.

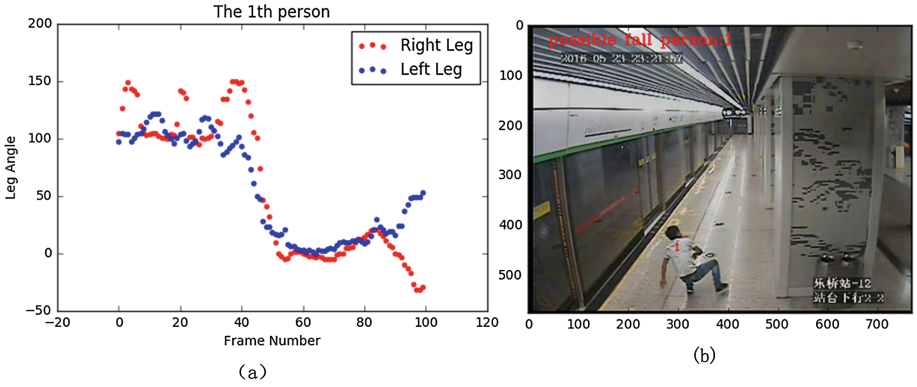

Fig. 20. Real-time decision data

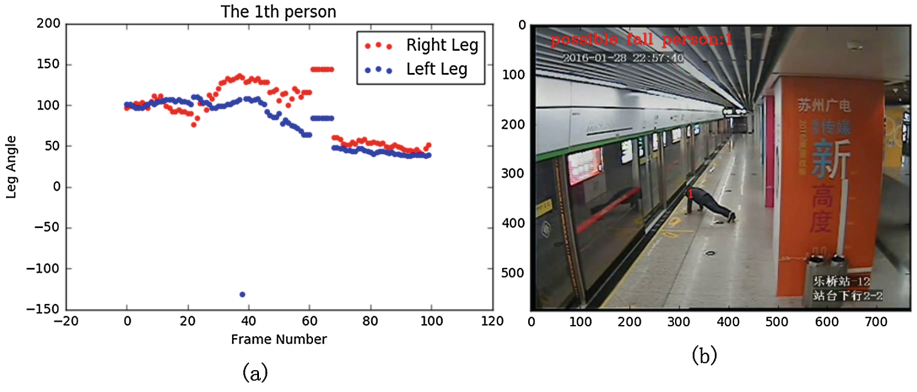

Fig. 21. Real-time decision data

The key frames in the second fall video are shown in Fig. 12. Figure 13 is the angle data read frame by frame from the whole video. As can be seen from the scatter plot, the people in the video is in the fall posture from about the 50th frame to the 300th frame. Figure 21(a) is the data obtained before a fall event is determined. It can be seen that the abnormal fall behavior in the video can be identified based on the first 100 frames. Figure 21(b) is the video frame when the alarm is issued at the corresponding time. The figure shows that the person labeled 1 has fallen.

5 Summary

In this paper, the threshold method is used to locate the neck key point of the same person, so as to achieve the tracking of the characters in the continuous video frames. Making full use of each frame of the video information, the real-time analysis can alarm after two seconds of fall. The bending of the human thigh, that is, the angle between the leg and the horizontal plane, is taken as the feature for the fall judgment, and the threshold value is determined by experience. Simplicity and certain accuracy are guaranteed at the same time. However, due to the limitation of the original attitude recognition algorithm in resolving the occlusion of personnel, the occasional occurrence of falls, and the limitation of the characteristics of falls judgment, there are inevitable errors in the process of image processing, resulting in the inconsistency of the number of people in the front and rear frames, the errors of key point coordinates and wrong limb connection. From the research point of view, the fall data used for training is insufficient. Different scenes, different shooting angles, static and dynamic camera data are lacking, so it is not adequate enough to verify this algorithm's accuracy. There remains many improvements in testing effects and optimizing the recognition algorithm.

References

1. Tompson, J.J., Jain, A., LeCun, Y., et al.: Joint training of a convolutional network and a graphical model for human pose estimation. In: Advances in Neural Information Processing Systems, pp. 1799–1807 (2014)
2. Aslan, M., Sengur, A., Xiao, Y., et al.: Shape feature encoding via fisher vector for efficient fall detection in depth-videos. Appl. Soft Comput. **37**, 1023–1028 (2015)
3. Ke, S.R., Thuc, H.L.U., Lee, Y.J., et al.: A review on video-based human activity recognition. Computers **2**(2), 88–131 (2013)
4. Cao, Z., Simon, T., Wei, S.E., et al.: Realtime multi-person 2D pose estimation using part affinity fields. In: CVPR 2017, vol. 1, no. 2, p. 7 (2017)
5. Foroughi, H., Pourreza, H.R.: Intelligent video surveillance for monitoring fall detection of elderly in home environments. In: 11th International Conference on Computer and Information Technology (2008)
6. Rougier, C., Meunier, J., St-Arnaud, A., et al.: Fall detection from human shape and motion history using video surveillance. In: 2007 21st International Conference on Advanced Information Networking and Applications Workshops, AINAW 2007, vol. 2, pp. 875–880. IEEE (2007)