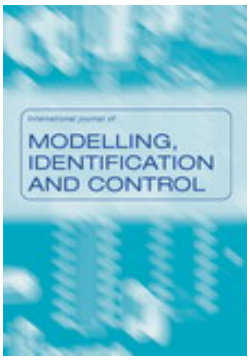


International Journal of Modelling, Identification and Control

ISSN online: 1746-6180 - ISSN print: 1746-6172
https://www.inderscience.com/ijmic

Multi-discriminant feature fall detection algorithm based on joints

Jimin Lai, Tonghui He

DOI: 10.1504/IJMIC.2023.10058394

Article History:

Received:	10 February 2023
Last revised:	25 May 2023
Accepted:	12 June 2023
Published online:	30 September 2024

Multi-discriminant feature fall detection algorithm based on joints

Jimin Lai and Tonghui He*

Hebei Software Institute,
Baoding, Hebei, 071000, China
Email: laijimin@126.com
Email: hetonghui2023@163.com
*Corresponding author

Abstract: Traditional fall detection algorithm is difficult to use accurately extract and recognise human posture features, and easily loses feature joints in the process of falling, resulting in low detection accuracy. Therefore, this paper proposes a multi-discriminant feature fall detection algorithm based on joints for nursing homes, medical rehabilitation centres and other places. Firstly, the initial features of human posture are obtained by the improved VGG-19 feature extraction model, and the initial positions of the joints are obtained and coded by adding a residual module. Secondly, the decoder network is used to complete deconvolution and upsampling operations to achieve greater fine-grained resolution. Finally, the image pose refinement module is designed to analyse the relationship between different adjacent feature nodes, so as to realise the accurate identification of the node position when the fall occurs. On this basis, the corresponding fall discriminant characteristics are proposed to achieve the detection of the elderly fall action. The results show that the proposed algorithm is more accurate and effective than other traditional algorithms on some datasets.

Keywords: fall detection; convolutional neural network; residual module.

Reference to this paper should be made as follows: Lai, J. and He, T. (2024) 'Multi-discriminant feature fall detection algorithm based on joints', *Int. J. Modelling, Identification and Control*, Vol. 45, No. 1, pp.11–18.

Biographical notes: Jimin Lai received his Master's in Software Engineering at the Beijing University of Technology. He is an Associate Professor at the Department of Computer Engineering, Hebei Software Institute. His research interests include issues related to the artificial intelligence. He is author of a great deal of research studies published at national and international journals, conference proceedings.

Tonghui He received his Master's in Information Science and Technology at the Hebei Agricultural University. He is an Associate Professor at the Department of Computer Engineering, Hebei Software Institute. His research interests include issues related to the computer vision. He is an author of a great deal of research studies published at national and international journals, and conference proceedings.

1 Introduction

By 2030, the proportion of people over 65 years will be expected to reach 25% in China, as the country faces the risk of an ageing population due to increasing life expectancy and declining fertility rates (Zhou et al., 2020). In the rapidly ageing society, the number of the semi-self-reliant elderly and the elderly living alone is increasing (Cao et al., 2017). How to ensure the health of the elderly is particularly important. Among many health problems, falling has the most serious impact on the health of the elderly, which can lead to different psychological and physical damage to the elderly, and may lead to death in severe cases. Therefore, different fall detection methods emerge at the historic moment. Human fall detection is realised on the basis of human posture recognition, which is a process of identifying human body parts in dynamic scenes (Wang et al., 2022). Now, posture recognition has been applied in motion recognition, monitoring scenes and other fields, and has broad market prospects (Yang et al., 2021). However, human posture recognition in colour images has been a complex problem in computer vision for a long time. The reasons include background scene changes, lighting, clothing colour, pose diversity, and occlusion, which make the accuracy of human pose recognition algorithm face great challenges (Li and Zhuang, 2021; Shi and Zen, 2021). Therefore, it has become a hot topic for scholars at home and abroad.

Existing human posture recognition algorithms include traditional detection methods and extraction of deep feature patterns. Most traditional methods use image segmentation and morphological analysis to locate human skeleton and node features in depth images. Such methods are subject to the influence of human flexible motion, and the model

scalability is poor. Yu et al. (2018) studied image segmentation of characters and their background, and took the extracted bending points of human skeleton as feature nodes. This method is simple to operate, but has low accuracy and is easily affected by characters occlusion, visual angle and other factors, so it cannot be widely applied. Li et al. (2015) designed a motion important feature extraction method based on frame spacing, which used the distance between quaternions to represent the difference of human posture, and constantly calculated and eliminated the frames whose difference was less than the threshold, finally, the posture characteristics of human body are obtained from the initial motion information. However, the extraction of human posture boundary information was not accurate. Grigorios et al. (2018) proposed a model that could model the visual appearance, modelling the structured limbs of the human body respectively, and expressing the kinematic dependence correlation of each part of the body through deformable configuration. However, the application scenarios of this model are very limited, which can only work in pure background, and the application effect is not ideal in most complex image scenes. Dantone et al. (2013) used spatial prediction technology to map input images directly to body joint coordinates through multi-stage sequential models. Although this method achieves good accuracy, it has low detection accuracy in occluded image scenes or complex body positions. Yang et al. (2016) proposed a new human posture recognition model, which used tree structure to construct key point locations of human body, and conducted end-to-end model training on tree structure and cyclic structure models. The pixel coordinates of key points in colour images were output, but the model parameters were too many, resulting in poor real-time performance. The reason is that most of the traditional methods used to distinguish falls are single and cannot accurately identify the complex human body structure.

At present, deep learning technology has made some achievements by using convolutional neural networks, which can automatically learn discriminant features from training sample to solve the problems of classification.

Kamel et al. (2020) proposed an improved stacked hourglass convolutional neural network model to predict human posture, and realised joint position correction of final posture after posture refinement by using pose correction network, so as to obtain accurate information of human posture. However, this method is only suitable for single-person posture detection, and its accuracy cannot be guaranteed when there are lots of people. Ke et al. (2022) proposed a unified three-stage network multi-person attitude recognition model, which extracted the human body boundary frame and key point information with the initial attitude, and effectively eliminated the error detection box through the filter module. Finally, the attitude refinement network was responsible for the regression of the complete attitude recognition results. Rana and Rawat (2021) proposed a simple and effective deep neural network, which adopted a top-down approach based on regional suggestion to detect action, and adopted non-maximum suppression processing to detect action recognition of actors in videos. Although it has a great improvement in performance method, it has limitations in the detection of dense scenes. Zhao et al. (2021) proposed a human body posture network structure, which can effectively ensure the spatial position relationship of human body in low-resolution situations and accurately extract semantic information to ensure the detection accuracy. However, the detection effect is poor in dim light scenes. Rogez et al. (2022) designed a multi-person action recognition model based on 2D-3D images. By using the convolutional neural network regression architecture of end-to-end positioning classification in different regions, 2D and 3D postures of figures in images can be detected to reconstruct the body postures. This kind of method can get the fall feature accurately, but it needs complex model, which leads to too long detection time and cannot meet the real-time detection requirements in multi-person scene.

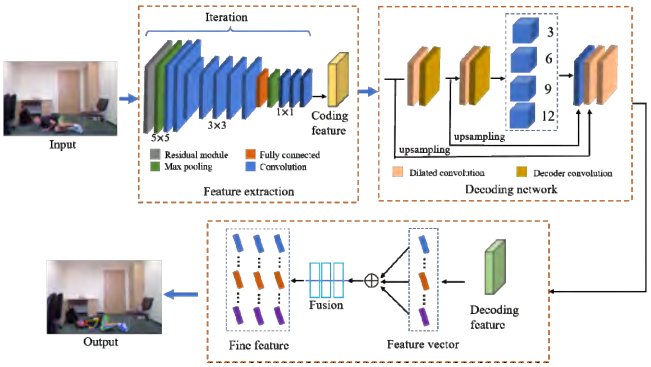
In order to solve the problems of low detection accuracy caused by inaccurate feature extraction, partial occlusion of human body, long time and other factors in the previous method, we propose a multi-discriminant feature detection algorithm for elderly falls based on joints. In order to make the algorithm more suitable for posture recognition of the elderly, higher requirements are put forward for the processing of human features. Therefore, multiple convolution kernels existing in the original VGG-19 feature extraction network are replaced by multiple small convolution kernels in this algorithm model, which effectively reduces the computation and improves the real-time detection capability. Moreover, residual module is added to enable the model to extract image edge information more accurately and obtain the exact location of the node. Secondly, coding and decoding networks are added to the model to obtain greater resolution, so that the human body features are more accurate and effective. Finally, a refined action recognition network was designed to predict the position of adjacent joints by using the unique graphic structure between skeletons, which effectively realises the accurate detection in the case of occlusion.

2 Algorithm framework

The overall block diagram of the algorithm is shown in Figure 1. First, the convolutional kernel 7×7 existing in the original VGG-19 network is replaced with multiple small convolution kernels, and a residual module is added to the network to improve the learning ability of edge features, meanwhile, the acquired human features are encoded and processed to obtain preliminary features of interest. Secondly, the features extracted by the encoder network are decoded into a larger fine-grained resolution through deconvolution and up-sampling operations. The features are up-sampled to convolution kernels of 3×3 and 6×6 to reduce parameters, and skip joins are adopted to retain the suppressed features. Finally, the decoded features are refined. At the same time, considering the significant graphic structure and adjacent relationship between

different key points, the implicit graph structure information is used for feature fusion processing and refined features, and the accurate human posture prediction output image is finally obtained.

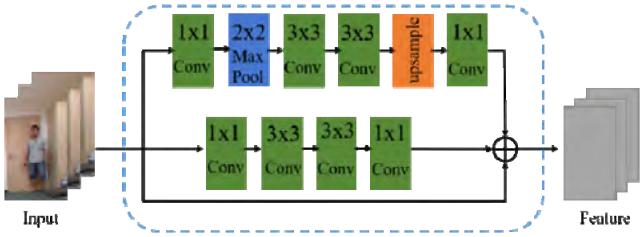
Figure 1 Overall flowchart of algorithm (see online version for colours)



2.1 Feature extraction

The VGG-19 network can be used for feature extraction of input images, but subsequent studies found that continuously deepening the number of layers of the network will greatly improve the performance of network detection, and has good scalability.

Figure 2 Sketch of residual module (see online version for colours)



Therefore, in order to increase the depth of the network and reduce network parameters, the convolutional kernel 7 × 7 is replaced with multiple and small convolutional kernels 1 × 1 and 3 × 3, and residual modules are added, as shown in Figure 2.

In Figure 2, the structure is a module with a deeper level. Input frames are processed by the convolution layer, maximum pooling layer and upper sampling layer respectively to enhance the refinement process. The output is a feature graph of 64 × 64 × 64, where 64 represents the height and width of the feature graph and 256 indicates the number of feature graph. Independent training is adopted in the training process, so that the module can obtain a unique loss function, and can independently learn the body posture characteristics. The training stages are 200, the iteration times are 1,000, the learning rate is 0.0025 to obtain a lower loss function value. When the atlas or video sequence is input into the model, the position confidence graph $W^*_{n,m}$ will be generated for the key point position of the person in each frame image, as the predicted position of the key point. The formula is as follows:

$$W^*_{n,m}(Q) = \exp\left(-\frac{\left\|Q - X_{n,m}\right\|_2^2}{\sigma^2_{n,m}}\right) \tag{1}$$

In the formula, m is the person's coded number, n is the key points' coded number, Q is any point in the figure, $X_{n,m}$ is the real location of key points, and $\sigma_{n,m}$ is the probability distribution of key points. Take the maximum value of the predicted position to get the unique key point position corresponding to the confidence graph. The calculation formula is as follows:

$$W^*_n(Q) = \max_m W^*_{n,m}(Q) \tag{2}$$

The coordinates of key points n obtained at this time can be used as the coordinates of points Q. The position coordinates of all key points are obtained, and the position of the human skeleton is predicted by the affinity vector field between the key points n_1 and the key points n_2, which can be expressed by calculating the integral value M. The formula is as follows:

$$M = \int_{t=0}^{t=1} k\left(Q(t)\right) \cdot \frac{n_2 - n_1}{\left\|n_2 - n_1\right\|_2} dt \tag{3}$$

$$Q(t) = (1-t)n_2 + tn_1, t \in (0,1) \tag{4}$$

In the formula, n_1 and n_2 is the adjacent key point, $\|n_2 - n_1\|$ is the corresponding bone length, $Q(t) \in [n_1, n_2]$. When any point Q is on the position of the skeleton, $k(Q(t)) = v$, v is the unit vector, otherwise, $k(Q(t)) = 0$. The larger the integral value is, the closer the key points are to the real bone position. Therefore, all accurate human bone positions can be obtained by selecting the maximum integral value.

2.2 Decoding network

The spatiotemporal features obtained by the network need to be decoded to the pixel level with a larger fine-grained resolution for subsequent human posture detection. The decoder network takes the resulting characteristic resolution [$T \times h \times w$] as input, and obtains the required resolution $\left[\frac{T}{2} \times \frac{h}{4} \times \frac{w}{4}\right]$ through two-layer deconvolution, two-layer expansion convolution, and two-stage up-sampling, T is the number of frames, h is the height, and w is the width. In the network, the features from the encoder network are input into each layer of deconvolution and expansion convolution by jumping connection, which is advantageous to decode the multi-scale context information around each pixel. The up-sampling method preserves the feature of the suppression, which is helpful to the feature decoding of the smaller object. Then, the pose map is constructed by guiding features, and the model of pose refinement is introduced, and the relationship between key points is fully considered to refine the features.

2.3 Motion feature recognition and refinement network

At present, human posture recognition has been extensively studied. However, most of these studies only focus on independent key point recognition and position, and lack of correlation between key points. Due to the unique physical properties of human body, the key points construct a remarkable graph structure, and there is a clear and accurate adjacency relationship between them. The location of the occluded key point can be well inferred by using the adjacency relation obtained between the nodes, so as to detect the human posture effectively. For example, if a key point is the right knee, its adjacent key points are the right ankle and right crotch point, then more supervised detection can be performed on these points rather than separate detection. Therefore, a figure posture refinement model is proposed to carry out feature refinement, and the graph structure is established for each key point, and the output is obtained by embedding supervision features. The formula is as follows:

$$g_n = \frac{1}{Z_n} \sum_{s_{n'} \in N(n)} \omega_{n'} T_{n'n} \left(f_{n'} \right) \tag{5}$$

$$\omega_{n'} = \begin{cases} h_{n'} \gamma \left(R_{n'} \right), & n' \neq n \\ 1, & n' = n \end{cases} \tag{6}$$

In the formula, $N(n)$ is the point set of the guide point and its neighbourhood, $T_{n'n}$ is the linear transformation between the guide points, $\gamma(\cdot)$ is the indicator function, Z_n for normalisation, $h_{n'}$ is the confidence score, $R_{n'}$ is the Boolean parameter filtering out the points with low connection quality and enhance the reliability of coding.

In order to better address the problem of human posture recognition, a weight is added to the connection relationship generated between the guide points to adjust the strength and reliability of the connection. In the meantime, the proposed algorithm can be trained better through the weight constraint. When $n' = n$, meet $\omega_{n'} = 1$, when $n' \neq n$, the guide point will make $\gamma(R_{n'}) = 0$, thus not affected by the strength of s_n itself and the connection reliability. The accurate key points that can represent the position of the body are obtained, and corresponding feature vectors are generated. The loss function is calculated by the fusion module, and the refined features of the prediction are output. After the fusion module calculates the loss function, the refined features of the forecast are output, which can be expressed as F_n and E_n for later attitude classification and regression. The corresponding loss function is calculated as follows:

$$L_n^S = \frac{1}{N_n} \sum_{s_n^i \in s_n} \partial_n^i L_s \left(F_n^i, 1 \right) \tag{7}$$

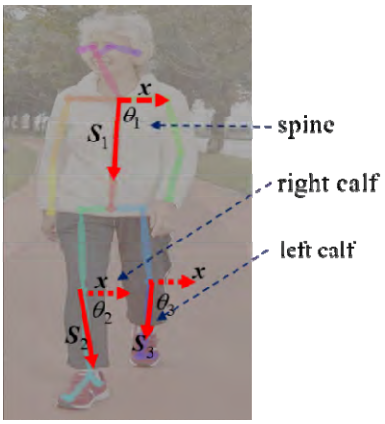
$$\partial_n^i = \exp\left(-\frac{\left(s_n^i - X_n \right)^2}{2\sigma^2} \right) \tag{8}$$

$$L_n^r = \frac{1}{N_n} \sum_{s_n^i \in s_n} L_s \left(E_n^i, X_n - s_n^i \right) \tag{9}$$

In the formula, L_n^s and L_n^r are softmax cross entropy loss and L_1 loss, respectively, and X_n is the true key point position.

Multiple vectors representing the position of the human torso are generated by processing the accurate position coordinates of the nodes to represent the pose information. By considering that the information carried by the trunk in the process of falling is richer than that carried by a single junction, vectors s_1, s_2 and s_3 representing the position and direction of the human spine and calves are obtained by coordinate transformation of the junction.

Figure 3 Discriminant feature (see online version for colours)



In Figure 3, the red solid line is the human skeleton involved in the algorithm, and the dotted arrow represents the direction vector $x = (0, 1)$ of the X-axis, which is always parallel to the ground, so it can be used to represent the direction vector of the real ground. Then, the corresponding included angle between s_1, s_2, s_3 and x is:

$$\theta_i = \arccos \frac{[x, S_i]}{\|x\| \cdot \|S_i\|}, \; i = 1, 2, 3 \tag{10}$$

In the formula, θ_1 represents the angle between the spine and the ground, θ_2 represents the angle between the left calf and the ground, and θ_3 represents the angle between the right calf and the ground. Therefore, different thresholds can be set to accurately detect the fall action.

3 Experimental results and analysis

3.1 Experimental details

The hardware configuration in this paper is composed of Intel Core i9-10 900T 4.60 GHz processor, RTX 3080 GPU, 32 GB RAM, and the server bandwidth is 100 M. The software configuration consists of Ubuntu16.04LTS, Opencv3.4.10, Caffe9.0 and Openpose1.7.0. The number of iterations is set to 300, and the image input size is 512 × 512.

3.2 Experimental dataset

In the experiment, Penn Action dataset (Ji et al., 2022) and URFD dataset (Kottari et al., 2020) are selected for experimental analysis, and compared with the existing advanced methods. Among them, Penn Action is a dataset of movements with 2,326 video sequences, including 15 different movements such as jumping rope, bench press, baseball swing, pull-up and squats. Each frame of the video has a corresponding tag to combine the pixel coordinates of the joints (including head, neck, wrist, knee, elbow, etc.). The URFD public dataset includes 30 fall action sequences and 40 daily activity action sequences. The above dataset videos are all from real scenes, in which there are many uncertain factors such as noise, indoor objects interference, different movements, and different testers. Therefore, the fusion of all video sequence segments of the two datasets and random disruption as public datasets is still suitable for testing the performance of the algorithm. In addition, multiple cameras were installed in different rooms, and 10 experimenters were randomly selected to enter different rooms to perform actions (including 50 groups of squatting, 50 groups of walking, 50 groups of sitting and 50 groups of falling) as the video sequence of the self-built dataset. Each participant had different action speed, amplitude and direction each time, so as to ensure the validity of the dataset.

In order to make the proposed algorithm have more accurate recognition performance, many fall experiments are done to obtain the threshold of human trunk angle. During the experiment, ten participants stood on the same horizontal ground respectively, made the maximum left-leaning or right-leaning movement, and recorded the tilt angle at the moment when they lost their balance and fell, as shown in Table 1.

Table 1 Record of body inclination in fall

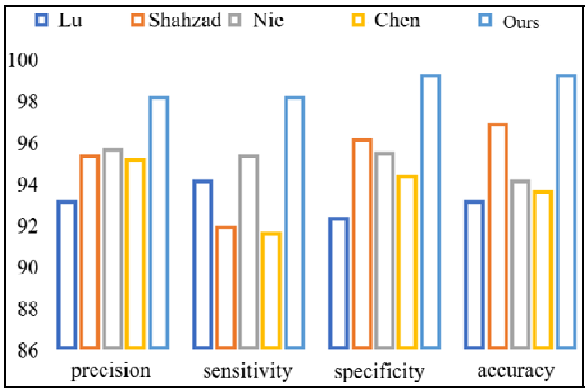
Tester	*Test 1*	*Test 2*	*Test 3*
1	72.3°	72.0°	71.0°
2	71.5°	73.0°	72.5°
3	71.6°	72.5°	71.5°
4	73.5°	74.0°	74.0°
5	71.0°	72.0°	73.0°
6	72.0°	74.0°	74.5°
7	73.0°	75.5°	71.0°
8	74.5°	74.0°	74.5°
9	74.0°	76.0°	73.0°
10	73.5°	76.0°	74.5°

As can be seen from Table 1, ten testers conducted three experiments, and the corresponding tilt angle was distributed between 70° and 80°. Moreover, the experimental results are more valid because the subjects have different ages, genders, heights and weights. The age of the experimenter in this study is younger than the elderly, and they have better balance ability. Therefore, the range of tilt angle threshold when falling is larger, which can be used as the threshold for the elderly to distinguish the tilt angle when falling.

3.3 Analysis of experiments on public datasets

In order to accurately evaluate the algorithm performance, this paper conducts comparative experiments with other four attitude recognition algorithms on the open dataset, and the detection results are shown in Figure 4.

Figure 4 Detection performance of different algorithms (see online version for colours)



The four methods are advanced human posture recognition algorithms, which increase the comparability of experiments. In order to comprehensively show the advantage of the algorithm and facilitate the comparison with the performance of other algorithms, four general evaluation indexes, including precision, sensitivity, specificity and accuracy are introduced in this paper. In the Figure 4, the proposed method has improved compared with other algorithms in various performance indexes, the accuracy can reach 99.0%, which proves that the algorithm still has good detection accuracy when detecting the complex environment, the diversity of personnel, and a large number of samples. In addition, the accuracy and sensitivity of the proposed algorithm reach 98.0%, and the highest results are 2.3% and 2.5% higher than other algorithm, which indicates that the algorithm is capable of detecting the marked error samples in the dataset.

In order to better highlight the advantages of the proposed algorithm, the number of parameters and time consuming of different algorithms in processing the same video were counted during the experiment, and the results are shown in Table 2. The proposed algorithm takes the second place in time, but it achieves the least result by calculating floating point numbers. Although Shahzad and Kim (2020) has achieved the minimum time, the number of parameters is large due to the existence of multiple convolution kernels to form the network and the need for discriminant actions based on threshold method. However, the methods proposed by Nie et al. (2019) and Chen et al. (2018) are of greater practical value than the proposed algorithm in terms of both time consuming and parameter quantity. Compared with Lu et al. (2019), the proposed algorithm adopts a lightweight network model with fewer

network parameters, which effectively reduces the complexity and ensures the real-time performance. The addition of residual modules enables the model to carry out more refined processing on the edge features of people, effectively improving the detection accuracy.

Table 2 Processing performance of different algorithms

Method	*FLOPs/G*	*Time/ms*
Lu	50.68	25
Shahzad	15.23	15
Nie	20.48	22
Chen	17.36	20
Ours	12.07	18

The detection effects of different human posture recognition algorithms are shown in Figure 5, where the first to fifth row pictures are Lu, Shahzad, Nie, Chen and Ours respectively, and the image types from the first to the fourth columns are normal actions, angle occlusion, body occlusion, border blur image. In Figure 5, the proposed paper can clearly recognise images with special conditions, while other algorithms can misrecognise or fail to recognise different human action conditions. Lu detects human posture only through the tilt angle of the body spine, and it can be seen that errors are prone to occur when detecting images such as the human spine being blocked or moving fast. Similarly, when Shahzad and Nie make use of background separation and human body boundary frame for initial extraction of human body pose, errors will occur due to the similar colour of the figure's dress to the background and the error of human body boundary point extraction. Chen judges the type of action by tracking the moving speed of some key points, resulting in low accuracy of detection results. In this paper, the algorithm obtains the edge information by adding the residual network module, and uses the refined network and the structural relationship of nodes to predict the blocked body parts, so as to effectively ensure the accuracy of recognition. In contrast, the traditional deep learning method only generates a single discriminant feature, such as the movement speed of key points and the proportion change of human body calibration frame, etc. which cannot accurately discriminate the pseudo-fall motion. Moreover, the traditional method cannot effectively extract human body features in the dimly lit workshop environment during image acquisition. The proposed algorithm fully considers the physical properties of the body in the event of a fall, and designs different discriminant features for the time domain and the space domain. Therefore, through the comparison of detection effects, it is proved that the proposed algorithm has good detection effectiveness.

3.4 Analysis of experiments on self-build datasets

The action sequences in the self-built dataset were collected in the real scene, and the randomness of camera angle and character posture in the picture increased the difficulty of detection to some extent. However, the detection accuracy of the algorithm in this paper was still up to 98.5%, which verified that the proposed algorithm has strong universality and can be used in many different detection scenarios. Other algorithms cannot accurately obtain the key point information of human body and the background separation is not accurate, which leads to the low recognition accuracy of different movements, as shown in Table 3.

Table 3 Accuracy of each algorithm detection

Method	*Squat*	*Walk*	*Sit*	*Fall*
Lu	92.0%	92.5%	95.0%	94.0%
Shahzad	93.0%	93.0%	94.0%	95.5%
Nie	92.5%	93.5%	93.5%	93.5%
Chen	91.0%	92.0%	92.5%	92.0%
Ours	95.0%	97.0%	98.5%	98.0%

The proposed algorithm can accurately detect human posture under different monitoring perspectives, and obtain better detection effects for different movements and different testers. The decoder network decodes the input image into pixel level with larger fine-grained resolution, which not only guarantees the accuracy of posture detection, but also improves the robustness of the algorithm as a whole. The detection effect on the self-built dataset is shown in Figure 6.

4 Conclusions

We propose a multi-feature fall algorithm based on joints in this paper. By improving the network model and introducing a residual module, we can refine the human edge posture features effectively. The decoding network is used to increase the fine-grained resolution of the input image. Meanwhile, the pose map refinement module is designed to ensure the recognition accuracy of the point when the figure is blocked. The three discriminant features proposed can accurately recognise the fall motion. Experimental results show that the accuracy of both the public dataset and the self-built dataset is significantly improved, which verifies that the proposed algorithm has better accuracy and universality than other pose recognition algorithms.

Although the proposed algorithm has good performance, it still has some shortcomings. The next stage of research will focus on improving the accuracy when the light is dim and the body is partially blocked. At the same time, it can judge the cause and severity of the fall according to the state information. Because the video of the old man falling down is imitated by the experiment, there are errors with the real situation. Therefore, more video data will be collected for detection in the future to improve the effectiveness of the algorithm.

Figure 5 Examples of detection effects of different algorithms (see online version for colours)

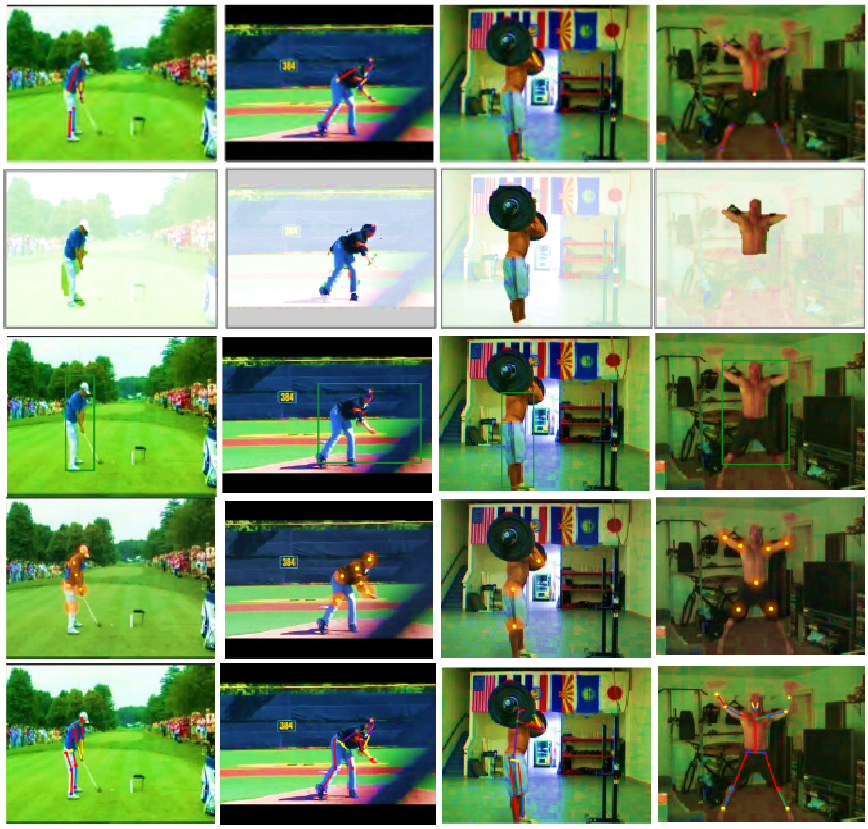


Figure 6 Example of detection effect (see online version for colours)

References

Cao, Z., Hidalgo, G., Simon, T., Wei, S.E. and Sheikh, Y. (2017) 'Realtime multi-person 2D pose estimation using part affinity fields', *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp.1302–1310.

Chen, Y.L., Wang, Z.C. and Peng, Y.X. (2018) 'Cascaded pyramid network for multi-person pose estimation', *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp.7103–7112.

Dantone, M., Gall, J. and Leistner, C. (2013) 'Human pose estimation using body parts dependent joint regressors', *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pp.3041–3048.

Grigorios, C., Epameinondas, A. and Zafeiriou, S. (2018) 'IPST: incremental pictorial structures for model-free tracking of deformable objects', *IEEE Transactions on Image Processing*, Vol. 27, No. 7, pp.3529–3540.

Ji, B., Pan, Y. and Jin, X.G. (2022) 'Spatiotemporal neural network for video-based pose estimation', *Journal of Computer-Aided Design & Computer Graphics*, Vol. 34, No. 2, pp.189–197.

Kamel, A., Sheng, B. and Li, P. (2020) 'Hybrid refinement correction heatmaps for human pose estimation', *IEEE Transactions on Multimedia*, Vol. 23, No. 10, pp.70–74.

Ke, L.P., Ming-Ching, C. and Qi, H.G. (2022) 'Detposenet: improving multi-person pose estimation via coarse pose filtering', *IEEE Transactions on Image Processing*, Vol. 31, No. 3, pp.2782–2795.

Kottari, K.N., Delibasis, K. and Maglogiannis, G. (2020) 'Real-time fall detection using uncalibrated fisheye cameras', *IEEE Transactions on Cognitive and Developmental Systems*, Vol. 12, No. 3, pp.588–600.

Li, L. and Zhuang, Q.H. (2021). 'Prediction and simulation of human behavior continuity based on time domain segmentation', *Computer Simulation*, Vol. 38, No. 5, pp.339–343.

Li, S.Y., Jin, H. and Gan, L.Z. (2015) 'Extraction of motion key-frame based on inter-frame pitch', *Computer Engineering*, Vol. 41, No. 2, pp.242–247.

Lu, N., Wu, Y.D. and Li, F. (2019) 'Deep learning for fall detection: 3D-CNN combined with LSTM on video kinematic data', *IEEE Journal of Biomedical and Health Informatics*, Vol. 23, No. 1, pp.314–323.

Nie, X.C., Li, Y.C. and Luo, L.J. (2019) 'Dynamic kernel distillation for efficient pose estimation in videos', *Proceedings of the IEEE International Conference on Computer Vision*, pp.6941–6949.

Rana, J.A. and Rawat, Y.S. (2021) 'We don't need thousand proposals: single shot actor action detection in videos', *2021 IEEE Winter Conference on Applications of Computer Vision*, pp.2959–2968.

Rogez, G., Weinzaepfel, P. and Schmid, C. (2020) 'LCR-Net++: multi-person 2D and 3D pose detection in natural images', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 42, No. 5, pp.1146–1161.

Shahzad, A. and Kim, K. (2020) 'Falldroid: an automated smart-phone-based fall detection system using multiple kernel learning', *IEEE Transactions on Industrial Informatics*, Vol. 15, No. 1, pp.35–44.

Shi, Y.X. and Zen, Z.S. (2020) 'Temporal segment networks based on feature propagation for action recognition', *Journal of Computer-Aided Design & Computer Graphics*, Vol. 32, No. 4, pp.582–589.

Wang, Y., Xu, B. and Zuo, F. (2022) 'A video action recognition system based on HOF-CNN and HOG features', *Computer Simulation*, Vol. 39, No. 6, pp.179–182.

Yang, W., Ouyang, W.L. and Li, H.S. (2016) 'End-to-end learning of deformable mixture of parts and deep convolutional neural networks for human pose estimation', *2016 IEEE Conference on Computer Vision and Pattern Recognition*, pp.2432–2445.

Yang, Y. and Deva, R. (2013) 'Articulated human detection with flexible mixtures of parts', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 35, No. 12, pp.2878–2890.

Yu, T., Zheng, Z.R., Guo, K.W., Zhao, J.H., Dai, Q.H., Li, H., Pons-Moll, G. and Liu, Y.B. (2018) 'Double fusion: real-time capture of human performances with inner body shapes from a single depth sensor', *Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp.7287–7296.

Zhao, L., Wang, N.N. and Gong, C. (2021) 'Estimating human pose efficiently by parallel pyramid networks', Vol. 30, No. 5, pp.6785–6800.

Zhou, L.J., Li, W.Q., Philip, O. and Zhang, Z.Y. (2020) 'Jointly learning visual poses and pose lexicon for semantic action recognition', *IEEE Transactions on Circuits and Systems for Video Technology*, Vol. 30, No. 2, pp.457–467.