Research on Fall Detection Based on Improved Human Posture Estimation Algorithm

ZHENG Yangjiaozi[1], ZHANG Shang[2]

(1. *China Three Gorges University*, *Computer and information institute*, *Yichang* 443002, *China*;
2. *Hubei Province Engineering Technology Research Center for Construction Quality Testing Equipments*, *China Three Gorges University*, *Hubei*, *Yichang*, 443002, *China*)

Abstract: According to recent research statistics, approximately 30% of people who experienced falls are over the age of 65. Therefore, it is meaningful research to detect it in time and take appropriate measures when falling behavior occurs. In this paper, a fall detection model based on improved human posture estimation algorithm is proposed. The improved human posture estimation algorithm is implemented on the basis of Openpose. An improved strategy based on depthwise separable convolution combined with HDC structure is proposed. The depthwise separable convolution is used to replace the convolution neural network structure, which makes the network lightweight and reduces the redundant layer in the network. At the same time, in order to ensure that the image features are not lost and ensure the accuracy of detecting human joint points, HDC structure is introduced. Experiments show that the improved algorithm with HDC structure has higher accuracy in joint point detection. Then, human posture estimation is applied to fall detection research, and fall event modeling is carried out through fall feature extraction. The designed convolution neural network model is used to classify and distinguish falls. The experimental results show that our method achieves 98.53%, 97.71% and 97.20% accuracy on three public fall detection data sets. Compared with the experimental results of other methods on the same data set, the model designed in this paper has a certain improvement in system accuracy. The sensitivity is also improved, which will reduce the error detection probability of the system. In addition, this paper also verifies the real-time performance of the model. Even if researchers are experimenting with low-level hardware, it can ensure a certain detection speed without too much delay.

Keywords: Fall Detection, Human Posture Estimation, Depthwise Separable Convolution, Convolutional Neural Networks, Feature Extraction

1 Introduction

According to statistics from the World Health Organization, “falling” is the second leading cause of death from all accidental or unintentional injuries in the world，with traffic accidents being the leading cause[1]. And according to research studies, approximately 30% of people who experienced falls are over the age of 65. Falls can cause great harm to the body of the elderly and bring them invisible pressure. This not only affects the daily lives of the elderly but also indirectly increases the pension burden of families and society. Therefore, for the elderly care monitoring problem, fall detection is a research content of great significance.

This research will minimize the physical injury caused by falls to the elderly and reduce their physical and mental suffering of the elderly. At the same time, it can also save social public medical resources and reduce the burden of family pension. In this paper, we propose a fall detection system based on human pose estimation and Convolutional Neural Networks. First, we preprocess the data and extract the position of human skeletal key-points from successive frames using the human pose estimation algorithm. The human body is extracted from the background by the algorithm while being tracked in real-time. Next, we extract the body descent rate and external profile features. After the primary judgment using the body descent rate, the secondary judgment is made by calculating the aspect ratio of the minimum external rectangular frame of the human body. After the secondary judgment, the fall event modeling is carried out to construct the fall features. Third, the designed Convolutional Neural Networks model is used to process and classify the above features. Finally, the classification results are applied to fall detection.

2 Related Work

2.1 Fall Detection Technology

In recent years, the trend of fall detection research has been increasing in recent years. Researchers have conducted more in-depth research on human behavior recognition and fall detection technology. From different implementation methods, fall detection technology can be roughly divided into the following three types, as shown in Fig.1: vision-based sensors, wearable device-based sensors, and ambient sensors[2].

2.1.1 Vision-based Sensors Fall Detection

Due to the widespread use of cameras in our lives, it has become a normal state to use cameras as a device for acquiring information in the field of computer vision research. The vision-based sensors mainly collect the data of human movement behavior through image or video capture devices installed in the user's living environment. And then the system uses image or video processing to perform human behavior recognition to determine if fall events have occurred. Chong et al[3]proposed a method that combines a background modeling approach of superpixel clustering with a background segmentation method based on the Horprasert algorithm. The author used two different methods, namely boundary box, and near elliptical motion quantization, to detect the drop of the foreground extracted by the method proposed in this paper. The results show that the method improves the processing speed and reduces the complexity of the original Horprasert segmentation.

The advantages of the vision-based sensors are easy to install the equipment and the user can visualize the video screen. However, at the same time, this method has some disadvantages, such as the target person may be obscured. To solve the shortcomings of conventional cameras, more and more researchers are choosing to use depth cameras. The most common method is to extract skeletal data from the human body. Bian et al[4]developed a two-tier system. First, the 3D coordinates of the body joints are removed from the segmented body parts by interpolation. Then, the position of the joint trajectory over time is extracted and the SVM classifier is used to detect the position of the joint.

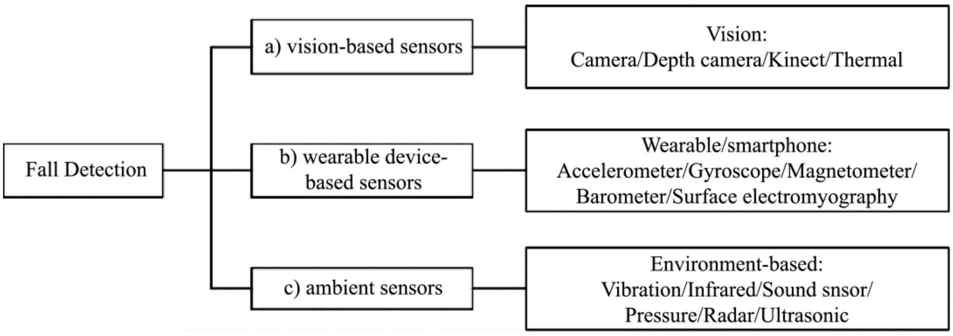

Fall Detection
a) vision-based sensors
Vision:
Camera/Depth camera/Kinect/Thermal
b) wearable device-based sensors
Wearable/smartphone:
Accelerometer/Gyroscope/Magnetometer/
Barometer/Surface electromyography
c) ambient sensors
Environment-based:
Vibration/Infrared/Sound snsor/
Pressure/Radar/Ultrasonic

Fig.1 Fall Detection Technology Classification

2.1.2 Wearable Device-based Sensors Fall Detection

Wearable device-based sensors for fall detection are developing rapidly and there are many kinds of devices available, the most common being accelerometers and gyroscopes. Instruments commonly used by researchers in fall detection studies include wearable watches, wearable belts, and wearable helmets. Rakhman et al[5] proposed a fall detection system based on a smartphone accelerometer and gyroscope. The authors used a three-axis accelerometer and gyroscope embedded in a smartphone to collect data on the human body's activity behavior and fall behavior. Next, the data characteristics features of different data are compared and analyzed with specific thresholds, and then a fall detection model is established. In addition, the fusion of wearable devices and other information technologies for multimodal information judgment is also a hot research topic. Yan Yujuan et al[6] designed a fall detection system that used an acceleration sensor, the ADXL330, combined with computable radio frequency identification tagging (CRFID) and pattern recognition technology. It used an acceleration sensor embedded in an ultra-high frequency CRFID to collect acceleration information of human movements to determine whether a fall occurred.

2.1.3 Ambient Sensors Fall Detection

Ambient-based sensors usually use sensors installed in the home environment to collect signals such as vibration signals, sound signals, and pressure signals to track the human body. Diego et al[7]proposed a detection method combining a class SVM (OCSVM) and a template matching classifier. The floor sound sensor captures the sound signal and then extracts the Mayer frequency cepstrum coefficient and Gaussian mean super vector (GMS). Users mark false-positive events and record them in the template GMS. The template matching classifier uses this template to make the final decision to distinguish human falls from non-falls. The ambient-based sensors allow direct access to target actions without violating the user's privacy. However, the detection environment can only be indoors, none of these systems can be used outdoors.

In addition to traditional acoustic and pressure sensors, many technologies such as ultrasound, infrared sensing, and radar detection are gradually being applied to systematic fall detection research. Zhang Dajun et al[8] proposed a method for fall detection using ultrasound at 20 kHz. The weightlessness caused by fall events increases the velocity of body movement, resulting in a Doppler effect that causes a shift in the reflected ultrasound frequency. Then, the system uses the offset data to determine whether the human has fallen. Sensors are sensitive to information such as sound and vibration, while their resistance to interference is weak. So, the most obvious disadvantage of this method is that it is easy to misjudge.

2.2 Skeleton Estimation

Generally speaking, human posture estimation algorithm is carried out by finding the position coordinates of human or object joint points. Take a person as an example, the joint points are elbow, knee, wrist and other joints. There are two types of pose estimation: multi pose and single pose. Single pose estimation is used to estimate the pose of a single object in a given scene, while multi pose estimation is used to detect the pose of multiple objects. Human posture estimation on popular MS COCO datasets can detect 17 different joint points (classes)[9]. Each joint point is annotated with three numbers (x, y, V), where x and Y mark the coordinates, and V indicates whether the joint node is visible

In this paper, we choose to apply human posture estimation to fall research. Through the analysis of human behavior using human posture estimation algorithm, we can distinguish between daily life behavior and fall behavior. The human posture estimation algorithm can extract the human joint point data under different postures. It can remove the background noise to the greatest extent, preserve the human structure, enhance the contrast between the human trunk and the background, and reduce the learning difficulty of the network. This can help us avoid the interference caused by human background to the greatest extent in fall detection research.

2.3 Depthwise Separable Convolutional

Convolutional Neural Networks (CNN) can be regarded as a variant of multilayer perceptron (MLP). Krizhevsky et al[10]applied CNN for the first time in the LSVRC-12 competition. By deepening the depth of the CNN model and combining relu and dropout technology, they achieved the best classification results at that time. The network structure has been named Alex Net. Simonyan et al[11]explored the importance of the 'depth' of CNN. Based on the existing network model structure, how to solve the network depth problem has become an important issue. To solve this problem, researchers proposed the inclusion of a convolutional layer with 3*3 convolutional kernels. And the experimental results showed that the performance of the model can be effectively improved when the number of weight layers reaches 16-19 layers, which is called VGG. Through continuous improvement and optimization, various convolutions such as Group convolution, Dilated convolution, and Depthwise separable convolution were created.

The convolution kernel can be regarded as a three-dimensional filter: channel dimension plus spatial dimension (width and height of the feature map). In regular convolution, the upper layer in the connection generally has multiple channels (we assume N channels here). Therefore, a filter must have N convolution kernels to correspond to it when performing convolution calculations. The essence of a filter completing one single convolution is that multiple convolution kernels are convolved with the feature maps of the corresponding channels of the previous layer. The convolution results are then summed to output a channel feature map for the next layer. In the next layer, if we need to get a feature map of multiple channels (we assume there are M channels), then we need to have M filters. The structure diagram of convolution layer of ordinary convolution is shown in Fig.2.

As shown in Fig.3, depthwise separable convolution decomposes the traditional convolution into a depthwise convolution and a 1×1 convolution (pointwise convolution). For the multi-channel feature maps from the previous layer, they are first all split into single-channel feature maps. The single-channel feature maps are convolved one by one for a single channel and then re-stacked together. Compared to conventional convolution, the biggest advantage of depthwise separable convolution is that it can greatly reduce the number of model parameters and calculations while maintaining a high level of accuracy.

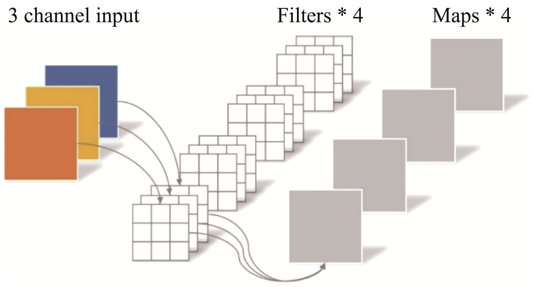


Fig.2 The Structure Diagram of Convolution Layer of Ordinary Convolution

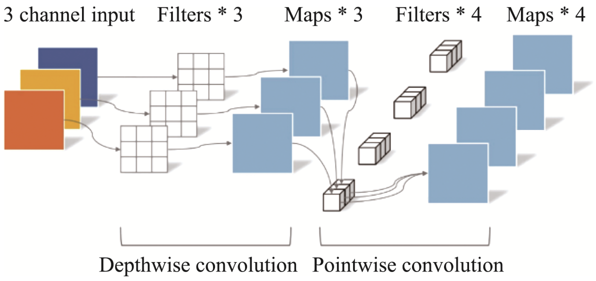


Fig.3 The Structure Diagram of Convolution Layer of Depthwise Separable Convolution

3 Methods

3.1 Openpose Algorithm

The openpose algorithm was first proposed in 2017. It is an open source library developed by the University of Nicky Mellon in the United States with Caffe as the framework on the basis of convolutional neural network and supervised learning[12]. Openpose algorithm takes as input a color image of size w*h and as output a two-dimensional skeleton image containing the location of key points for each person. The backbone network uses a VGG19 network model, where the input raw graph is initialized and fine-tuned by the first 10 layers to generate a set of feature maps F as input to the first stage. At this point, the algorithm enters the refinement stages. In the initial stage, there is the 3 * 3 convolution kernel in the convolutional layer, and in the refinement stage, it becomes the 7*7 convolution kernel. Starting from the second stage, the input to the stage t network consists of three components: S_{t-1}, L_{t-1}, F. S_{t-1}, L_{t-1} is the output of the stage t-1 network. In other words, the input to each stage of the network is the output of the previous stage of the network. We use a feedforward network to predict a two-dimensional confidence map S of body part locations and a set of vector fields L associated with body parts when using the algorithm to extract features. The degree of association between parts of the human body is obtained by encoding the association vector field. Finally, the confidence map and association field are resolved by the greedy algorithm, and the two-dimensional key points of everyone in the image are output.

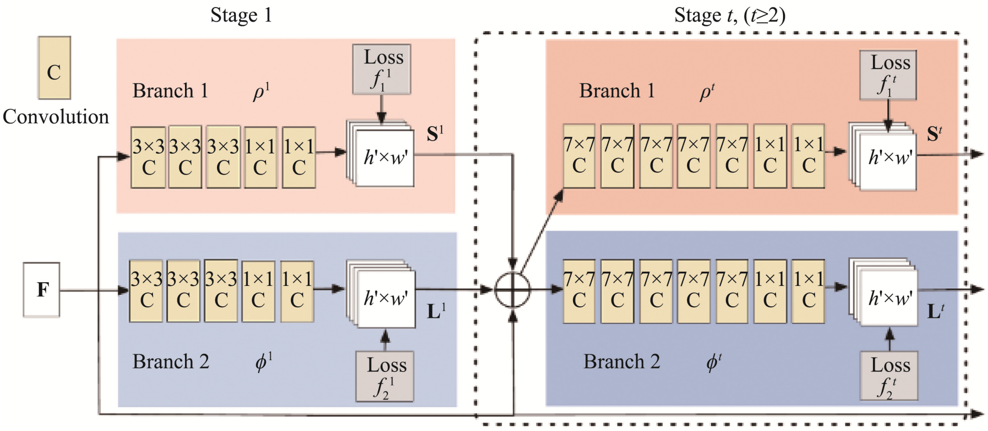
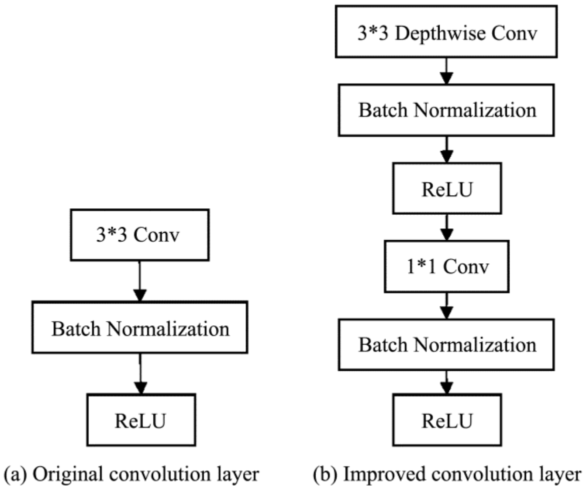


Fig.4 Network Structure of Openpose Algorithm[12]

3.2 Our Algorithm

To reduce the model parameters and increase the computational speed, we improve the Openpose algorithm from three perspectives. The first is for the backbone part of the network, the Openpose algorithm uses the VGG19 network. It uses the first four blocks of the VGG19 network and then adds two additional convolutional layers, later on, to complete the feature extraction before entering the initial stage. To reduce the model parameters and computational effort, we use depthwise separable convolution. The structure of the depthwise separable convolution is adapted to replace the VGG19 network. The comparison of convolution layer structure before and after improvement is shown in Fig.5.

The improved network structure is shown in Table 1. In the improved algorithm structure, the first two layers are traditional convolution layers, the size of convolution cores is 3*3, the number of convolution cores is 32, the step length is 1, and the filling coefficient is 1. The last ten layers establish a deep separable convolution neural network structure. The size of convolution kernel is 3*3, and the number increases in turn. The stride and filling coefficient are shown in the Table 1.



(a) Original convolution layer (b) Improved convolution layer

Fig.5 Comparison of Convolution Layer Structure before and after Improvement

Table 1 Improved Network Structure

Convolution Type	Convolution Kernel Size	Strides	Padding
Conv 1	3*3*32	1	1
Conv 2	3*3*64	1	1
Conv dw 1	3*3*64	1	0
Conv dw 2	3*3*128	2	0
Conv dw 3	3*3*128	1	0
Conv dw 4	3*3*256	2	0
Conv dw 5	3*3*256	1	0
Conv dw 6	3*3*256	1	0
Conv dw 7	3*3*512	1	2
Conv dw 8	3*3*512	1	0
Conv dw 9	3*3*512	1	0
Conv dw 10	3*3*512	1	0

In Table 2, it is shown the comparison of backbone network parameters and calculation before and after improvement. It is obvious from the data in the Table 2 that the number of parameters and calculation of the improved model are reduced a lot.

Table 2 Comparison of Parameters and Calculation of the Model before and after Improvement

Backbone Network	Parameters	Calculation
Original	$5.86*10^{6}$	$1.40*10^{10}$
Improved	$1.15*10^{6}$	$1.89*10^{9}$

Secondly, the Openpose algorithm used two branches to generate key-point heatmap and paf heatmap, respectively. The two branches have the same structure, the only difference is in the output phase, where the number of output results differs between the two branches. Therefore, we merge the two branches into one and add a 1*1 convolution to the output stage to split it into two branches. In addition, Openpose algorithm adopts a six stage multi-stage network. The prediction results of the previous stage are purified through the network of the later stage. However, in the actual training process of the model, we found that after stage 1, the performance of the network did not get much improvement, but the amount of computation significantly doubled. However, if only a single stage is used, although the training difficulty and calculation amount are greatly reduced, the effect is not ideal because the output error does not go through multi-layer back propagation. Therefore, this paper chooses to retain the first three stages and output loss in each stage to prevent the gradient from disappearing. The network structure is shown in Fig.6. The structure of Dilated MobileNet V2 in the Fig.6 will be further described below.

Third, in the refinement stage, the Openpose algorithm uses 7*7 convolution kernels. We replace it with a cascaded convolutional structure, using three 3*3 convolutional cascades instead of 7*7 convolution. In the current field of deep learning, more and more

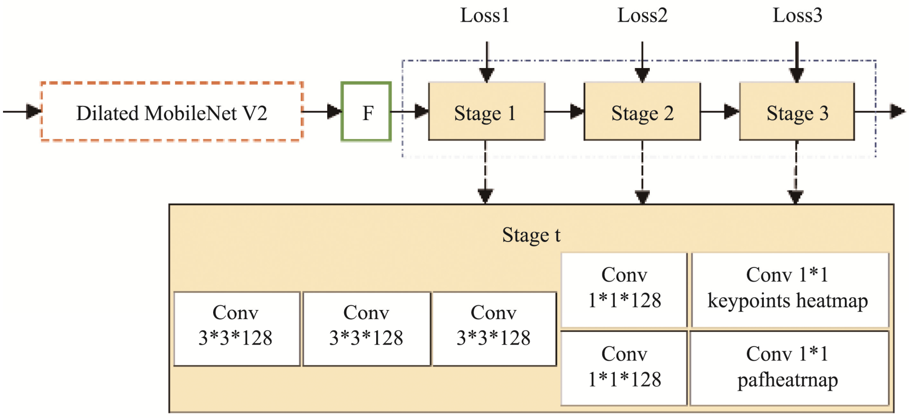


Fig.6 Network Structure of Our Algorithm

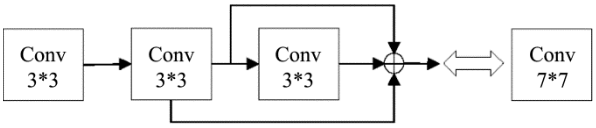


Fig.7 Three 3*3 Small Convolution Kernels Are Used to Replace 7*7 Convolution

researchers use multiple small convolution kernels instead of large convolution kernels. It can not only ensure the network performance, but also reduce the number of parameters and calculation. The repeated stacking of multiple small convolution cores makes multi-layer ReLU functions activated, which also increases the nonlinearity of convolution process and improves the ability of network nonlinear mapping.

The VGG19 network is replaced by adjusting the depthwise separable convolution structure, which greatly reduces the amount of model parameters and calculation. However, in the process of actual training model, we found that a few feature maps will lose some position information, resulting in the key nodes of human body cannot be well detected. Therefore, based on the improved algorithm, we introduce hybrid dilated convolution. Compared with dilated convolution, hybrid dilated convolution can increase the receptive domain and better ensure the continuity of the receptive domain. When it is applied to the existing human pose estimation algorithms for feature extraction, it can not only retain more feature information and ensure the robustness of the algorithm, but also not increase the computational power consumption.

The improved algorithm network structure in this paper has twelve layers, of which the first two layers are conventional convolution layers and the last ten layers are convolution layers based on depthwise separable convolution. Hybrid dilated convolution is introduced into the last three layers of the network, namely layers 10-12. The hybrid dilated convolution of sawtooth structure with dilation rate set to 1, 2 and 5 is added respectively. In this paper, the structure is named Dilated MobileNet V2 structure. The network with HDC structure is shown in Fig.8 below, which is called Dilated MobileNet V2.

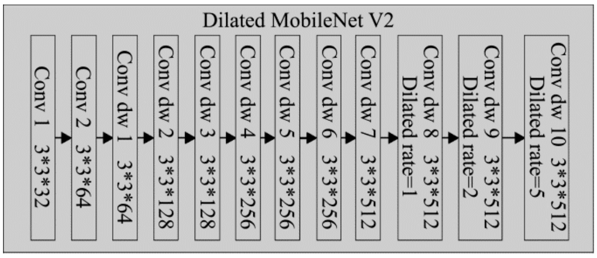


Fig.8 Structure of Dilated MobileNet V2

3.2 Analysis of Falls Constructing Model

Falls are momentary movements. There is a rapid fall process when a person falls. The center of the human body will drop rapidly from a higher position to the ground or near the ground. Changes in the human descent rate and the external contour are two intuitive features when a fall event is detected.

3.2.1 Velocity Analysis

The falling process of a human body can be divided into a weightless phase, an impact phase, and a resting phase. In Fig.9, (a) shows the coordinate system for acceleration and indicates the acceleration along the x, y and z axes respectively. (b) shows the coordinate system for the angular velocity and 3D angle of the human body and represents the angular velocity of the human body around the x, y, and z axes respectively. We use the change in the velocity of movement of the body's center point in the y-axis direction as a feature.

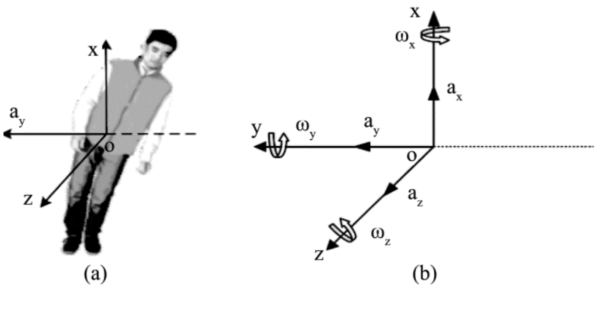


Fig.9 Movement Models of Human Activity.

We selected a total of six key node’s locations for reference: the head, the left and right shoulders, the body center position, the left and right knees. Sampling is done once at an interval of 12 frames. The average of the velocity of the six nodes moving in the y-axis direction between the two sampling frames is used as the body descent velocity.

As an example, in the standing stage, we take the diagonal intersection of the smallest rectangular frame outside of body as the position of the body's center mass. We assume that the coordinates are (x, y), where y is the height of the body central position in the y-axis direction. The rate of descent of the body centroid between two sampled frames is v_y.

$$v_y = \frac{(y_{i+f} - y_i)}{f} \tag{1}$$

y_{i+f} and y_f denote the vertical coordinates of the body center point in the two sampled frames, respectively. f is the time difference between the two sampled frames.

We assume that the coordinates of the 6 joint nodes are $(x_1, y_j) \sim (x_6, y_j)$. Average of the velocities of the 6 nodes moving in the y-axis direction is v.

$$v = \frac{\sum y_{j+f} - \sum y_j}{6f} \tag{2}$$

y_{j+f} and y_j denote the vertical height of the jth node between the two sampled frames, respectively. f is the time difference between the two sampled frames.

The method of our fall detection algorithm for fall speed is as follows:

(1). Falls are instantaneous action in which the body's rate of descent increases dramatically during a fall. Therefore, when we detect a small change in the rate of descent of human activity, we can conclude that it is not a fall. In other words, we determine that the body is non-falling when v is less than the threshold $v_{m.}$

(2). We find that the rate of descent of the human body may be more than the threshold v_m when squatting or jumping. However, falls occur only once and do not recur. Sometimes the human body will even syncope after a fall event occur when the body speed will remain at 0 for a certain period of time. (If there is no syncope occurred, the body speed may not be 0, but will very slightly over a certain period.) However, movements such as squatting and jumping do not occur as described above in most cases. The speed changes of the human body after a movement such as squatting or jumping will be greater than after falls. Here we set the speed change threshold to v_n. We carry out several sampling sessions and capture 24 frames per second and calculate the velocity of the human body in the y-axis direction every 12 frames. It takes about 5 seconds for each fall to occur. So, if we detect the behavior that v is more than threshold v_m within 5 seconds and in the following time, v is less than threshold v_n, then we can tentatively identify the behavior as a fall.

In this paper, we select 200 consecutive frames of video sequences with 5 different poses in the same video. Through experiments result as show in Fig.10, the speed differences of five different behavior states including falls are verified, and the corresponding speed curves are obtained. Through the analysis of velocity curve in Fig.10, we can see that the value of v_m is between 0.9-0.95m/s, and the value of v_n is between 0–0.05m/s.

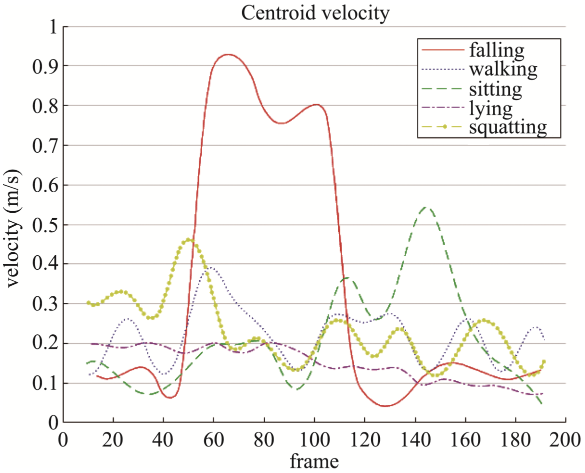


Fig.10 Centroid Velocity under Different Condition

3.2.2 External Contour Analysis

In a few cases, such as when the human body remains in a continuous squatting position after squatting. In such cases, a fall can easily be misjudged by the change in speed of descent of the body alone. So, then, we make a secondary judgment by the aspect ratio of the minimum external rectangular frame of the human body.

We complete the delineation of the human body with an external rectangular frame, as shown in Fig.11, which is the minimum external rectangular frame of the human body in the standing position.

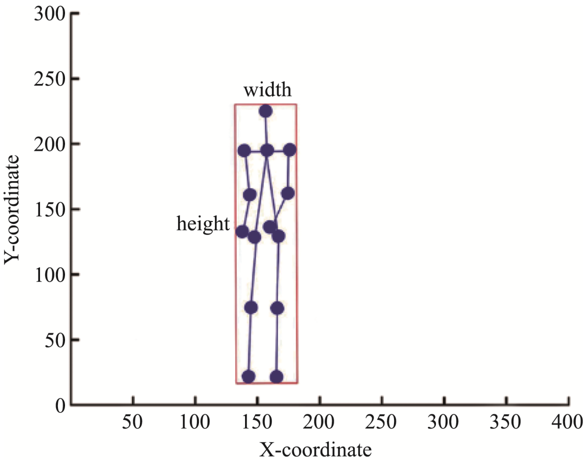
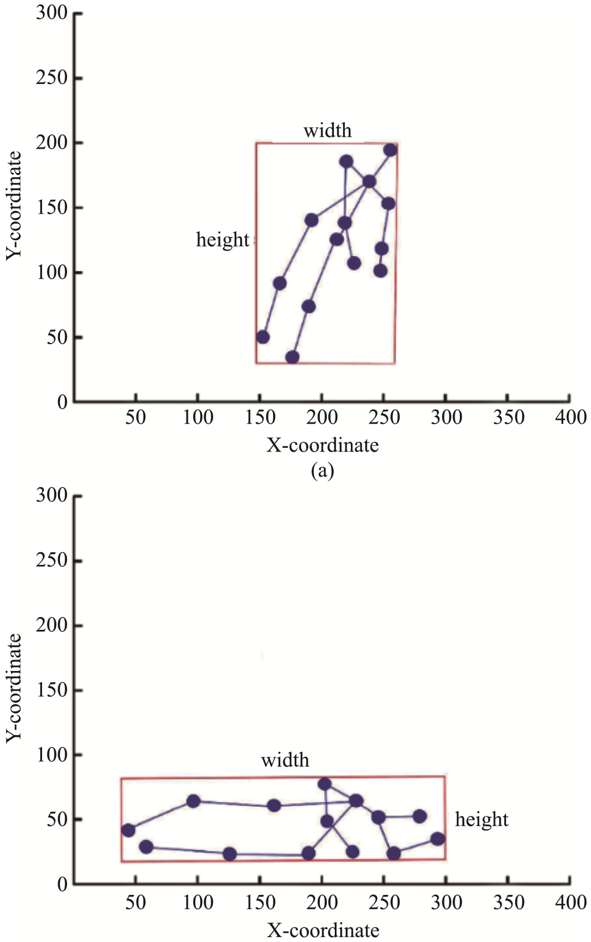


Fig.11 Minimum External Rectangular Frame of the Human Body in the Standing Position

In Fig.12, (a) shows the Minimum external rectangular frame of the human body during a fall, while (b) shows the Minimum external rectangular frame of the human body after a fall.

The most obvious change in the external profile of the human body during a fall can be seen in the ratio of the width to the height of the minimum external rectangular frame of the human body, as it is shown as Fig.13

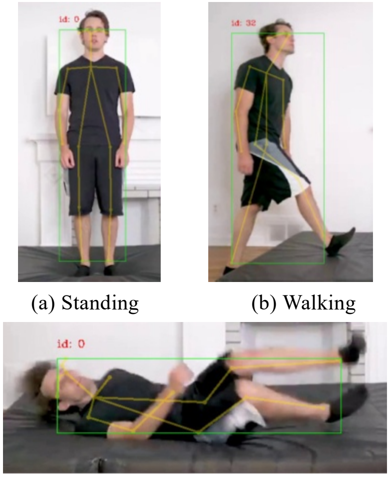
As Fig.13 shows that the experimental effect of the above video sequence on the improved model. It includes several different human states of standing, walking and falling. In the external rectangular frame above, we set the width to w, the height to h, and the aspect ratio $L = w/h$. When L is more than the threshold L_m, we identify the state as a fall state. We also select 200 consecutive frames of video sequences with 5 different poses in the same video. The experimental results are shown in Fig.14. Through curve analysis, we can see that the aspect ratio of Minimum external rectangular will change dramatically during the fall. This changeset is not found in other behaviors. Combined with the experimental results, we can determine that the value of L_m can be set to 2.5.



(a)

(b)

Fig.12 Minimum External Rectangular Frame of Human Body When a Fall Occurs, (a)shows Minimum External Rectangular Frame of the Human Body during Falling, (b)minimum External Rectangular Frame of the Human Body after Falling



(a) Standing

(b) Walking

(c) Falling

Fig.13 Experimental Results on the Improved Model

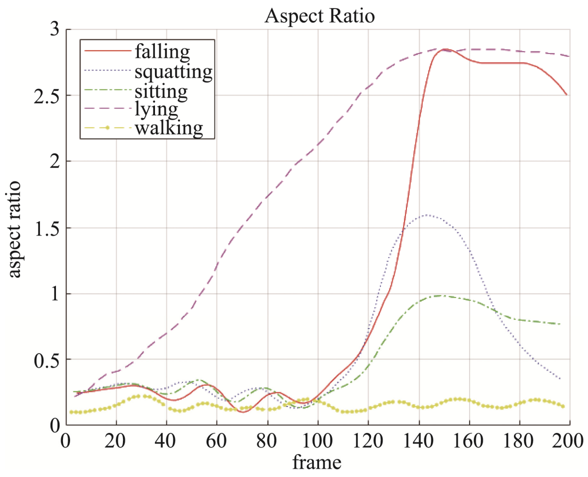


Fig.14 Aspect Ratio of Minimum External Rectangular under Different Posture

In Fig.14, it is worth noting the case of lying. When the human body is lying down, because the human body is gradually approaching the horizontal direction, the final action form is very similar to falling. Therefore, it can be seen from the Fig.14 that when the human body is lying down, the L value is almost the same as that in the case of falling. However, according to the analysis of fall above, fall is an instantaneous behavior with large speed change and short time. Therefore, we can see that the falling state takes less time when the width height ratio L of the external rectangle reaches the maximum value. In other words, the width height ratio curve of the external rectangle in the falling state changes greatly and steeply, and the width height ratio curve of the external rectangle in the lying state is relatively flat.

3.3 Analysis of Falls Constructing Model

Convolutional Neural Networks (CNN) is a kind of feedforward neural network with convolution calculation and depth structure. It is now widely used in many research fields, such as natural language processing and computer vision.

In simple terms, the process of fall detection is equivalent to a classification problem. A serious problem of fall detection using computer vision is that the spatial background information of human body will produce some interference. Although now in many kinds of studies, researchers have invented algorithms that can separate the human background to eliminate interference. However, since we preprocess the data with the estimation algorithm in advance and extract the human skeletal map in different poses, it can help us to avoid the interference caused by the human background to the greatest extent. Using the human key-points map as the input of convolution neural network can remove the background noise to the greatest extent, preserve the human structure, enhance the contrast between the human trunk and the background, and reduce the learning difficulty of the network.

In this paper, we choose Convolutional Neural Networks to design fall detection model. The structure of the model designed in this paper is shown in Fig.15 below.

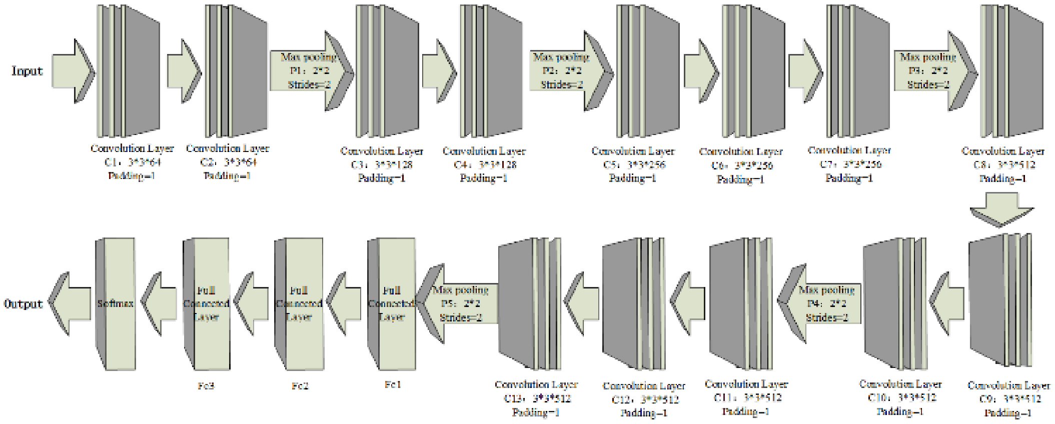

Input
Max pooling P1：2*2 Strides=2
Max pooling P2：2*2 Strides=2
Max pooling P3：2*2 Strides=2
Convolution Layer C1：3*3*64 Padding=1
Convolution Layer C2：3*3*64 Padding=1
Convolution Layer C3：3*3*128 Padding=1
Convolution Layer C4：3*3*128 Padding=1
Convolution Layer C5：3*3*256 Padding=1
Convolution Layer C6：3*3*256 Padding=1
Convolution Layer C7：3*3*256 Padding=1
Convolution Layer C8：3*3*512 Padding=1
Output
Softmax
Full Connected Layer
Full Connected Layer
Full Connected Layer
Max pooling P5：2*2 Strides=2
Max pooling P4：2*2 Strides=2
Fc3
Fc2
Fc1
Convolution Layer C13：3*3*512 Padding=1
Convolution Layer C12：3*3*512 Padding=1
Convolution Layer C11：3*3*512 Padding=1
Convolution Layer C10：3*3*512 Padding=1
Convolution Layer C9：3*3*512 Padding=1

Fig.15 Structure of the Designed Convolutional Neural Networks Model

The model is mainly composed of the convolution layer, the pool layers and the full connections. The size of each frame of the input layer is 224 * 224. In order to enable the action sequence to be used as the input of convolutional neural network, the input layer in the network is changed from receiving 3-channel RGB image to receiving 10 channel action sequence. As shown in Fig.15, in the model, C1-C13 are the convolutional layers. The size of the convolutional kernel is 3*3 and the parameter is set to 1. The number of convolution cores of C1-C2 is 64, that of C3-C4 is 128, that of C5-C7 is 256, and that of c8-c13 is 512. P1-P4 are all pooling layers. The main function of the pooling layer is down sampling, followed by compression features. It can effectively reduce the number of parameters and simplify the complexity of the network. It can also reduce the amount of calculation and expand the perceptual field of vision. The pooling operation used here is maximum pooling, with a step size of 2. This means that the maximum value of the image area is selected as the pooling value for that part. In terms of loss function, in this paper, the binary cross entropy loss function is used as the loss function of the model to express the difference between the predicted value and the true value. Compared with the ordinary cross entropy function, the binary cross entropy loss function can ensure the numerical stability in fall classification.

Deep learning models usually contain many parameters. When the amount of training data is not large enough, it will lead to overfitting. In general, one of the best ways to reduce overfitting is to increase the number of training samples. If there are enough training samples, even the most extensive networks are less likely to suffer from overfitting. Unfortunately, although more research results have been achieved in fall detection research, the number of fall datasets is not nearly large enough. To address this problem, it has been considered that expanding the training datasets is also a practical solution. The data expansion method used in this paper is image flipping.

4 Experiment

4.1 Experiment Dataset

In order to reflect the performance of the improved human posture estimation algorithm. We use COCO datasets[9] to train human pose estimation model. Each human body in the human joint point detection image in the coco data set is marked with the coordinate positions of 17 joint points. Based on the position of joint points marked in coco data set, this paper also adds an additional position of human joint points: two shoulder center points to complete human posture estimation.

In addition, in order to verify the effectiveness of the fall detection model based on pose estimation proposed in this paper, we use three datasets to train and test the model after completing design. These three datasets are publicly available, namely: UR Fall Detection Dataset (URFD)[13], Fall Detection Dataset (FDD)[14], and Multiple cameras fall dataset (Multicam) [15] . The FDD dataset contains 22636 images, of which 16794 images are available for training, 3299 images for validation and 2543 images for testing. The Multicam dataset has 192 videos of simulated falls and normal behavior taken from 8 different angles. In this paper, the Fall videos containing falling behavior are considered as positive samples, while the NotFall videos with other normal behavior are considered as negative samples

4.2 Evaluating Metric

In order to verify the effect of the improved human posture estimation algorithm, we choose to use the detection speed FPS and AP value (Average Precision) as performance indicators to evaluate the performance of the algorithm.

Average precision is an important index to measure the performance of the model in target detection. In this paper, a bottom-up method is used in the design of human pose estimation algorithm, that is, all human key-points on the image are detected first, and then these joint points are connected by cluster analysis. This method tests the ability of clustering analysis.

Object keyword similarity (OSK) is usually used to measure the similarity between true value and predicted value. The calculation formula is as follows:

$$OKS = \frac{\sum_i exp\{-d_p^2/2s_p^2\sigma_i^2\}\delta(v_{p^i}=1)}{\sum_i \delta(v_{p^i}=1)} \qquad (3)$$

d_p^2 represents the Euclidean distance between the true value and the predicted value of the joint node. p represents the ID of the human body in the truth value. p^i represents the ID of the joint node. $v_{p^i}=1$ indicates that the joint node is labeled and the label is visible. s_p represents the size of the area occupied by the human body calculated by the human body boundary box in the true value. δ_i represents the normalization factor of the ith joint point, which reflects the influence of the current bone point on the overall prediction. On the basis of OKS calculation results, the value of AP can be further calculated. The concept of AP is to calculate the ratio of the number of OKS greater than *t* to the total number of OKs under a given threshold *t*. When OKS is greater than *t*, the joint point is successfully detected; If it is less than *t*, the test fails. The calculation formula of AP is as follows:

$$AP@t = \frac{\sum_p \delta\,(OKS_p > t)}{\sum_p 1} \qquad (4)$$

Otherwise, sensitivity, specificity and accuracy as the judging criteria to verify the effectiveness of the fall detection model based on pose estimation proposed in this paper. We have also compared our method with some other fall detection methods. These three performance indicators are calculated using the following formulae.

$$Sensitivity = \frac{TP}{TP+TN} \qquad (5)$$

$$Specificity = \frac{TN}{TN+FP} \qquad (6)$$

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \qquad (7)$$

4.3 Results Analysis

4.3.1 Improved Human Posture Estimation Algorithm

We selected MSCOCO2014 dataset to train human posture estimation model. During training, the initial learning rate is set to 0.005, batch_size is set to 128; For every 500 epochs, the learning rate decreased by 0.6 times, and stopped after 3000 epochs.

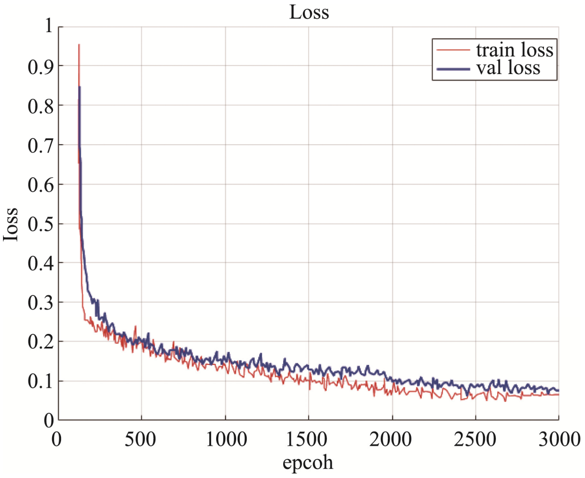


Fig.16 Loss Function Image

After the training, the generated weight model is saved and called. In this paper, we select MSCOCO2014 val dataset to test the running speed of the model. In this paper, the same video frame is used as the input are used for experimental comparison. Input a video sequence with a total of 5500 frames, and get the speed of model processing image through the ratio of video frames to program running time, that is, frame rate. According to the definition of frame rate, the higher the frame rate, the faster the detection speed. Through experiments, we get the ratio of video frames to program running time, that is, the result of frame rate, when the input is a video sequence with a total number of 5500 frames.

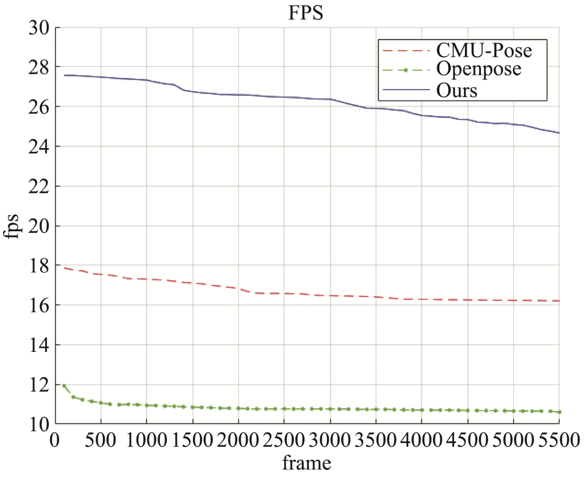


Fig.17 Experimental Results of Frame Rate of Different Models

It shows the experiment results of the same video sequence in different models in Table 3. According to the data in the Table 3, the frame rate of the improved algorithm model is increased by 57.03%.

Table 3 Comparison of Model Detection Speed before and after Improvement

Model	Total Input Frames	Average Frame Rate/fps
Openpose[12]	5500	16.71
Ours	5500	26.24

There is the threshold *t*, when OKs is greater than t, it indicates that the joint node has been successfully detected, and the value of AP can be further calculated. Table 4 shows the results of performance comparison experiments on mscoco2014 test dev dataset. In Table 4, AP represents the average accuracy. According to the target recognition evaluation standard of the International Conference on machine vision: AP^{50} represents the AP value obtained when the threshold *t* is 50%; AP^{75} represents the AP value obtained when the threshold *t* is 75%. In the COCO dataset, about 41% of the images are small-scale images less than 32 * 32. About 34% of the images are medium-sized images, and the image size is between 32 * 32-96 * 96. About 24% of the images are large-scale images larger than 96 * 96. Therefore, when the input image is a medium-scale image, the value of AP obtained by calculating all OKs thresholds is AP^M while AP^L is for the case that the input image is a large-scale image.

Table 4 Performance Comparison of the Improved Algorithm on MSCOCO2014 Test Dev Dataset

Model	AP	AP^{50}	AP^{75}	AP^M	AP^L
Openpose[12]	64.2	86.2	70.1	61.0	68.8
Ours	64.0	85.9	69.9	60.8	69.1

As can be seen from Table 4, compared with Openpose, the AP value decreased by 0.2, decreased by 0.31%, but the AP^L value increased by 0.3, increased by 0.44%. This shows that the performance of the improved model is better when the input image size is greater than 96 * 96. In addition, compared with the Openpose, the average detection speed of the improved model is increased by 57.03%, which can better meet the real-time requirements.

4.3.2 Fall Detection Model

The experimental results of the performance of the model designed according to the method proposed in this paper on the three datasets are shown in Fig.18. According to the final printed results, it is known that: the accuracy tested on the UR fall detection dataset is 98.53%, on the fall detection dataset is 97.71%, on the Multiple cameras falls dataset is 97.20%

The other results of the experiment, including specificity and sensitivity, are shown in Table 5 below.

To reflect the experimental results of the proposed method, we compare the experimental results of other methods and the proposed method in this paper on the same dataset, as shown in Table 6, Table 7 and Table 8.

Table 5 Experimental Results of the Proposed Method in This Paper on Different Datasets

Dataset	Sensitivity	Specificity	Accuracy
URFD	100.00%	98.44%	98.53%
FDD	99.83%	97.58%	97.71%
Multicam	99.80%	97.13%	97.20%

Table 6 Comparison of Our Experimental Results with Others on the URFD Dataset

Method	Sensitivity	Specificity	Accuracy
Ours	100.00	98.44	98.53
Khraief et al[17]	100.00	95.00	-
Harrou et al[18]	100.00	94.93	96.66

Table 7 Comparison of Our Experimental Results with Others on the FDD Dataset

Method	Sensitivity	Specificity	Accuracy
Ours	99.83	97.58	97.71
WANG et al[19]	97.78	97.37	97.33
Charfi et al[20]	98.00	99.60	-

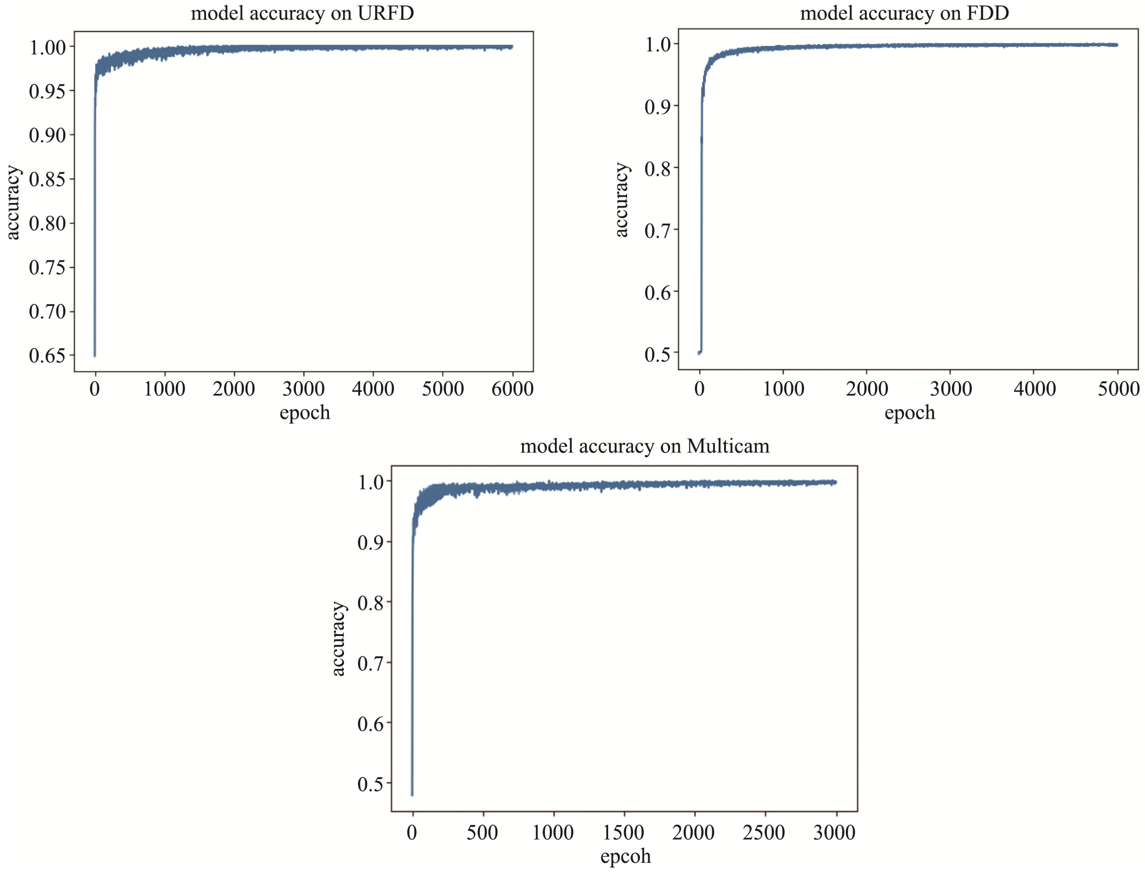


Fig.18 The Accuracy of the Proposed Method on the Three Datasets

Table 8 Comparison of Our Experimental Results with Others on the Multicam Dataset

Method	Sensitivity	Specificity	Accuracy
Ours	99.80	97.13	97.20
Liu et al[21]	98.00	97.50	-
Rougier et al[22]	95.40	95.80	-

From the comparison of Table 6, Table 7 and Table 8, it can be seen that the method proposed in this paper improves the progress while also improving the sensitivity of the system to a certain extent and reducing false detections. In addition, in the actual detection process, this experiment is currently in a low hardware configuration. Nevertheless, the detection speed of the method proposed in this paper for video is around 23–26 fps, up to 28 fps, with low latency, which can basically meet the requirements of real-time detection. This can also provide some help for experiments without higher hardware equipment environment

5 Conclusion

In this paper, we propose a fall detection method based on human pose estimation. First, we combine the depthwise separable convolution and a pyramidal model based on the dilated convolution. A human pose estimation algorithm is obtained by improving on the structure of the original Openpose algorithm. The human posture estimation algorithm is used to track the human body in the video and extract the key-points of the human body under different behaviors and apply the results of human pose estimation to the research of fall detection. Then feature extraction is carried out on the extracted data. The falling speed of human body and the minimum external rectangular frame of the human body are used as the features for multiple judgments, so as to establish the fall event model and construct the fall features. Next, it is fed into the designed Convolutional Neural Networks model for classification. And the final classification results are applied to fall detection. Based on the experimental results, it is demonstrated that the fall detection method proposed in this paper achieves better detection results on all three publicly available fall detection datasets.On the premise of maintaining the original detection accuracy, the average detection speed of the improved human posture estimation algorithm is improved by 9.7fps. The experimental results on several public fall data sets also prove that the application of human posture estimation results to fall detection has better recognition effect on fall behavior.

References

[1] World Health Organization. Fact sheet on fall injuries in 2018 https://www.who.int/zh/news-room/fact-sheets/detail/falls. URL.

[2] Yu X. Approaches and principles of fall detection for elderly and patient[C]//e-healthNetworking, Applications and Services, 2008. Health Com 2008.10th International Conference on. IEEE, 2008:42-47.

[3] Chong C J, Tan W H, Chang Y C, et al. Visual based fall detection with reduced complexity horprasert segmentation using super-pixel,2015 IEEE 12th International Conference on Networking, Sensing and Control. IEEE, 2015: 462-467.

[4] Z. Bian, J. Hou, L. Chau, and N. Magnenat-Thalmann, “Fall detection based on body part tracking using a depth camera,” IEEE J. Biomed. Health Information., vol. 19, no. 2, pp. 430–439, Mar. 2015.

[5] Rakhman AZ, Nugroho LE, Widyawan, et al. Fall detection system using accelerometer and gyroscope based on smartphone// International Conference on Information Technology, Computer and Electrical Engineering. IEEE 2014:99-104.

[6] Yan YuJuan, Li Hua, Zhao JuMin, et al. Fall Detection System Based on CRFID and Pattern Recognition. Computer Engineering, 2019(6):297-302.

[7] Diego D, Daniele F, Emanuele P, et al. A Combined One-Class SVM and Template-Matching Approach for User-Aided Human Fall Detection by Means of Floor Acoustic Features. Computational Intelligence and Neuroscience, 2017, 2017:1-13.

[8] Zhang Dajun, Lan Henrong, Wu Youlong. Bathroom fall detection based on ultrasonic Doppler effect. Journal of Shanghai Normal University (Natural Science Edition),2018, v.47(02):92-96.

[9] Lin T Y, Maire M, Belongie S, et al. Microsoft coco: Com-
mon objects in context[C]//European conference on com puter vision. Springer, Cham, 2014: 740-755.

[10] Alex Krizhevsky; Ilya Sutskever; Geoffrey E. Hinton. ImageNet Classification with Deep Convolutional Neural Networks.

[11] Simonyan K, Zisserman A. Very Deep Convolutional Networks for Large-Scale Image Recognition. Computer Science, 2014.

[12] Zhe C, Simon T, Wei S E, et al. Realtime Multi-person 2D Pose Estimation Using Part Affinity Fields. IEEE, 2017.

[13] Bogdan Kwolek, Michal Kepski, Human fall detection on embedded platform using depth maps and wireless accelerometer, Computer Methods and Programs in Biomedicine, Volume 117, Issue 3, December 2014, Pages 489-501, ISSN 0169-2607.

[14] Adhikari, Kripesh, Hamid Bouchachia, and Hammadi Nait-Charif. “Activity recognition for indoor fall detection using convolutional neural network.” *Machine Vision Applications (MVA), 2017 Fifteenth IAPR International Conference on.* IEEE, 2017.

[15] E. Auvinet, C. Rougier, J. Meunier, A. St-Arnaud, J. Rousseau, “Multiple cameras fall dataset”, Technical report 1350, DIRO - Université de Montréal, July 2010.

[16] Z. Cao, G. Hidalgo, T. Simon, et al. OpenPose: Realtime multi-person 2d pose estimation using part affinity fields[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), 2019.

[17] Khraief C, Benzarti F, Amiri H. Elderly fall detection based on multi-stream deep convolutional networks. Multimedia Tools and Applications, 2020, 79(6).

[18] Harrou, Zerrouki F, Sun N, et al. Statistical control chart and neural network classification for improving human fall detection// 2016 8th International Conference on

Modelling, Identification and Control (ICMIC). IEEE, 2017.

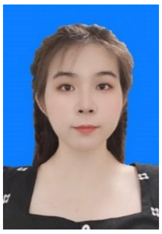
[19] Wang B H, Yu J, Wang K, et al. Fall Detection Based on Dual-Channel Feature Integration. IEEE Access, 2020, PP (99):1-1.

[20] Charfi I, Miteran J, Dubois J, et al. Definition and Performance Evaluation of A Robust SVM Based Fall Detection Solution// Eighth International Conference on Signal Image Technology & Internet Based Systems. IEEE, 2012.

[21] Liu J, Xia Y, Tang Z. Privacy-preserving video fall detection using visual shielding information. Visual Computer, 2020(1).

[22] Rougier C, Meunier J, St-Arnaud A, et al. Robust Video Surveillance for Fall Detection Based on Human Shape Deformation. IEEE Transactions on Circuits & Systems for Video Technology, 2011, 21(5):611-622.

Author Biographies



ZHENG Yangjiaozi is now a M.Sc. candidate of Three Gorges University. Her main research interest includes Internet of Things and deep learning.

E-mail: eden33s@163.com

ZHANG Shang received Ph.D. degree from China University of Geosciences (Beijing) in 2014. He is now an associate professor and master supervisor in Three Gorges University. His main research interests include Internet of Things, computer application technology, etc.

E-mail: wetoo@163.com