ISSN No:-2456-2165

A Computer Vision-Based Dual Network Approach for Indoor Fall Detection

Lu Zhang
Shandong University of Technology
Zibo, China

Chun Fang
Shandong University of Technology
Zibo, China

Ming Zhu*
Shandong University of Technology
Zibo, China

Abstract:- In order to strengthen the monitoring of the elderly and reduce the safety risks caused by falls, a video-based indoor fall detection algorithm using a dual network structure is proposed. Firstly, for the recorded video stream, we apply the fine-tuned YOLACT network to extract the contours of the human body, and then design a simple convolutional neural network to distinguish the categories of different family activities (including bending, standing, sitting and lying), and finally make a fall decision. When a lying position is detected on the floor region, it is considered as a fall. Experiments show that the proposed algorithm can successfully detect fall events in different indoor scenarios, and have a low false detection rate on the constructed data set.

Keywords:- *health care; YOLACT; convolutional neural network; gesture recognition; fall detection.*

I. INTRODUCTION

In recent years, with the development of living standards and medical technology, the aging of the population has become more and more serious, and the number of elderly people living alone has also increased, which has aroused widespread social concern. Accidental falls have become one of the main reasons that threaten the health of the elderly. The World Health Organization reported that 28% to 35% of people over 65 years old would fall each year, and this number would rise to 32% to 42% of people over 70 years old [1]. Once a fall occurs, it may cause soft tissue damage, fractures, head injury, etc., causing physical and psychological damage and increasing family burden [2]. In order to provide qualified medical and health services for the elderly living alone and reduce the negative impact of falls, it is essential to develop an effective fall detection system.

The current research work on fall detection algorithms is mainly divided into three types: wearable device-based methods [3], scene sensor-based methods [4], and computer vision-based methods. Wearable-based fall detection methods are to wear sensor devices including accelerometers, gyroscopes, etc. on the body, collect motion information, and detect human activities and states to determine whether they fall. The scene-based fall detection method is to deploy sensors, such as pressure sensors, sound sensors, etc., in the human activity area, and use such devices to obtain information about human activities, and then make fall judgments. Computer vision-based fall detection is to collect human activity videos through camera equipment, and use image processing technology to analyze and process them, and finally determine whether they fall. Compared with the first two methods, visual processing-based methods have their unique advantages. They are non-invasive (the elderly do not need to wear special equipment), will not affect human activities, and the cost of monitoring equipment is lower. The video has richer semantics, which is convenient for later review.

Processing methods based on computer vision are widely studied [5]. At present, many researchers use machine learning algorithms to perform fall detection. The focus of this work is mainly on extracting foreground moving targets in videos, extracting effective features of falls, and selecting classifiers with better performance. In terms of extracting human targets in videos, the main background subtraction algorithms, such as mixture of Gaussians (MoG) [6], Codebook [7], Vibe [8], etc. Fall-related features mainly include shape features and motion features. Rougier *et al.* [9] used a monocular camera to track the head movement of the active human body, and then extracted the head 3D motion trajectory curve, and then extracted the head speed feature on this basis. Judge the fall. Vaidehi *et al.* [10]extract static features such as human body aspect ratio and tilt angle to detect falls. C-Y Lin *et al.*[11] combined Motion History Image (MHI) to analyze falling behavior, and proposed two additional features, acceleration and angular acceleration, to comprehensively determine whether a fall event has occurred. Commonly used classifier models include Support Vector Model (SVM), Hidden Markov Model (HMM) and (convolutional neural networks) CNN. Mirmahboub *et al.* [12] use contour regions obtained from multiple consecutive frames extracted from continuous Gaussian average background differences as features, and input them into the SVM model to classify different types of normal activities. K. Tra *et al.* [13] used five features extracted from the ellipse model and input them to two HMM models to classify fall events and normal events. Adhikari *et al.* [14] preprocess the RGB-D image, combined with the CNN network model to determine whether the human body fall, the detection accuracy can reach 81%.

This paper proposes a new fall detection algorithm based on computer vision, which mainly uses a dual network structure. First, the YOLACT network is used to extract the contour of the foreground target, and then the features of the human posture are automatically extracted through the convolutional neural network structure to

complete the posture classification. When the detected human lying on the ground reaches a certain time threshold, a fall alarm signal is issued. Compared with other methods, the main contribution of this paper is to use the YOLACT network to replace the traditional background subtraction algorithm, which can improve the robustness of human body extraction in complex environments, and can achieve real-time extraction. In addition, the self-designed convolutional neural network can automatically extract the effective features of the human silhouettes, avoiding the complex feature extraction and data reconstruction process in traditional algorithms.

II. METHODOLOGY

In this paper, the flow structure of the fall detection algorithm based on the dual network structure is shown in Fig 1. First, the YOLACT network is used for preprocessing to extract the human silhouettes from the original video stream, and then automatically extract the features of standing, sitting, bending and lying postures to construct a posture classification CNN. Finally, when lying on the floor area is detected the posture is regarded as a fall, and an alarm is issued when the time exceeds the threshold.

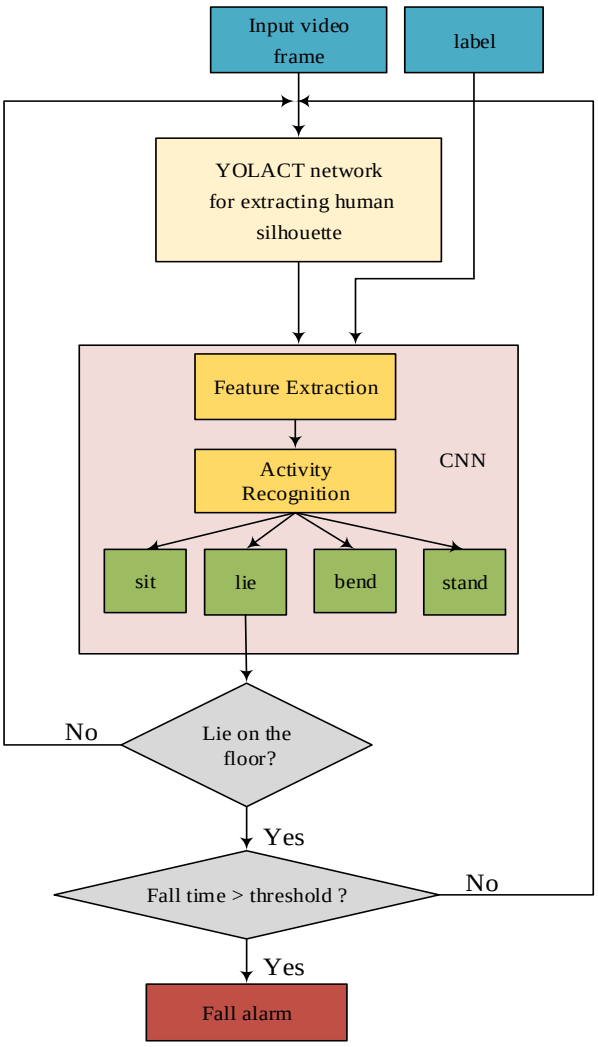


Fig 1:- Overall flow chart of fall detection

A. YOLACT network for human silhouette extraction

The YOLACT network [15] is a first-level instance segmentation model based on a first-level target detector. In this paper, the network structure is used to extract the contours of the human body in the video. Compared with the traditional background subtraction method, YOLACT can extract the human silhouettes in the real and complex background environment with excellent results. In addition, the extraction speed of the network is fast, and real-time segmentation can be realized. The overall network structure of YOLACT is shown in Fig 2.

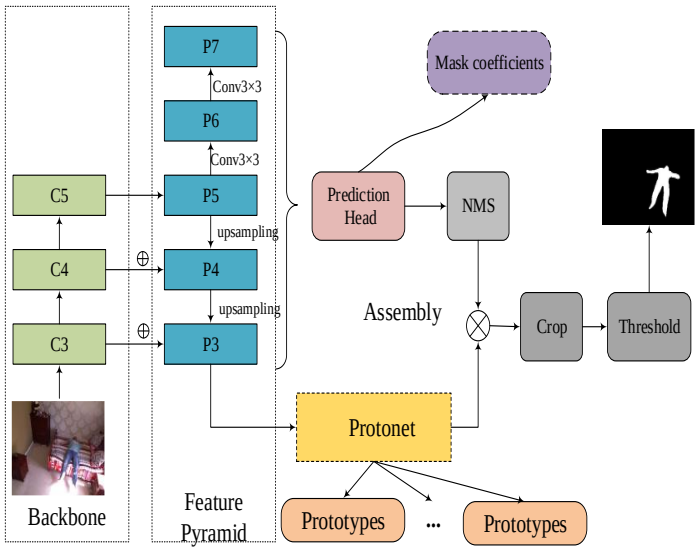


Fig 2:- YOLACT network

The goal of YOLACT is to add the mask branch to the existing one-stage target detection model. The general steps of detection are as follows: First, use a backbone network such as resnet101, resnet50 or VGG16 to extract multi-layer feature maps (as shown in Figure 2, from C3 to C5). Then generate P3, P4, P5 based on feature pyramid network (FPN), and generate P6 and P7 through P5. Next, the complex task of instance segmentation is decomposed into two simpler parallel tasks. The first branch uses the Protonet structure to predict different positions and foreground and background segmentation to generate multiple "prototype masks". The second one adds an additional head to the target detection branch to predict the vector used to encode the "mask coefficients" of each anchor represented by the instance in the prototype space. The two-branch prototype and mask coefficients can be calculated independently. Finally, for each instance after NMS, the mask segmentation result corresponding to each anchor is generated through linear combination. The specific principle details can be found in [15].

YOLACT training involves three parts: classification loss, bounding box loss, and mask loss. The classification loss and bounding box loss are the same as SSD [16], and the mask loss is the pixel-by-pixel binary cross entropy of the predicted mask and ground truth mask. Given a training data set containing images of human poses and related annotations (including human category annotations, human

ISSN No:-2456-2165

object bounding box annotations and human silhouette annotations), the YOLACT weight can be trained by minimizing the loss function. The trained YOLACT network can complete the extraction of human silhouette in video frames.

B. Convolutional neural network for human activity recognition

The image data used in this section is obtained through online resources and self-collecting methods. There are a total of 4,000 images including human standing (1000), sitting (1000), bending (1000), and lying (1000) in different indoor scenes. We divide the data set according to the ratio of training set: test set=8:2. After extracting the human body contour from the video record through the above YOLACT network, since the target contour only occupies a small part of the original image, the background part has no effect on the recognition of human activities. In order to eliminate image redundancy, reduce the amount of calculation, and increase the running speed, in advance, we extract the minimum bounding rectangle(MBR) of each contour, then uses the method of bilinear interpolation to resize the same size(64×64); finally each pixel of is normalized to [0,1].

In the field of computer vision research, deep learning techniques such as CNN [17] have been widely used in image classification. CNN can directly extract the information in the image layer by layer from a large amount of labeled data, extract the effective features of the image, and detect and classify the image. Compared with traditional manual features, CNN does not need to extract the contour features of the human body separately, and calculate information such as the position of the center of mass of the human body and the aspect ratio. The classic CNN network includes convolutional layer, pooling layer and fully connected layer.

- Convolutional layer. The input original image is subjected to a convolution operation with convolution kernels and an addable offset vector to obtain multiple mapping feature maps.
- Pooling layer. Usually after the convolutional layer, it is used for down-sampling to reduce the dimensionality of features. The two most traditional pooling methods are max-pooling and average pooling.
- Fully connected layer. After the original image is processed by multiple convolutional layers and pooling layers, the output feature image is flattened into a one-dimensional vector and used for classification. Other features can be added to this one-dimensional vector and used for classification. In CNN, one or more fully connected layers can be used for the final classification.

This paper designs a simple CNN network to complete the classification task of human posture, as shown in the structure Table 1.

Layers	Input size	Description	Output size
1	64*64*3	Convolutional layer, with 32, 3*3 convolution kernels	64*64*32
2	64*64*32	Pooling layer, with the 2*2 Maxpooling kernel	32*32*32
3	32*32*32	Convolutional layer, with 32, 3*3 convolution kernels	32*32*32
4	32*32*32	Pooling layer, with the 2*2 Maxpooling kernel	16*16*32
5	16*16*32	Convolutional layer, with 16, 3*3 convolution kernels	16*16*16
6	16*16*16	Pooling layer, with the 2*2 Maxpooling kernel	8*8*16
7	8*8*16	FC	1*1024
8	1*1024	FC	1*512
9	1*512	Softmax	1*4

Table 1:- Proposed the CNN structure

Taking the processed human contour image as the input of the convolutional layer, the CNN used is mainly composed of 3 convolutional layers, 3 pooling layers, 2 fully connected (FC) image feature layers and 1 fully connected classification feature layer. The Relu activation function is applied between the convolutional layer and the pooling layer to obtain fixed neuron output. The output after three convolutional layers is fed to two fully connected layers with 1024 and 512 neurons respectively. Finally, the probability of four representative activity categories is output through the Softmax function. In order to prevent over-fitting, a dropout layer is inserted between the last convolutional layer and the fully connected layer. The parameter of this layer is 0.4, that is, the neurons are turned off with a 40% probability to prevent over-fitting. In order to train the network, categorical_crossentropy is used as the loss function, and the root mean square propagation (RMSProp), which can automatically adjust the learning rate, is used as the optimization function to estimate the weight of the CNN, and the batch size is set to 32.

Finally, the trained CNN is applied to classify different types of poses. For ground areas, mark them manually. When the human body detects a lying position and is on the ground area, it is regarded as a fall.

III. EXPERIMENTAL RESULTS AND ANALYSIS

This section verifies the performance of the fall detection algorithm proposed in this paper in a real home environment. Invite 10 volunteers to simulate walking, sitting up, bending, lying up, and falling. A total of 320 video clips were recorded, including 90 falling clips (including back fall, back fall, and side fall).

A. YOLACT network for human extraction results

The experiment in this section uses the human body categories in the MS COCO dataset [18] to train the YOLACT network in advance to obtain a model that can recognize the human body. Experiments show that the model is not ideal for segmentation of falling postures. Therefore, we collected a total of 1,800 pictures including standing, sitting, bending and falling postures by myself. According to the experimental requirements, the pictures were labeled with the contour of the human body and the category person using the labelme tool, and converted into '.Json' format data with mask annotations Set, fine-tuned through migration learning to obtain a new human segmentation model. Experiments have proved that the network model can extract human contours in different scenarios. Some test results of YOLACT extracted from human body are shown in Fig 3.

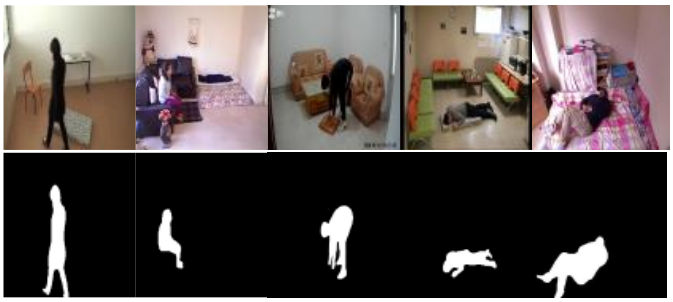


Fig 3:- Extraction results of human poses in different scenarios

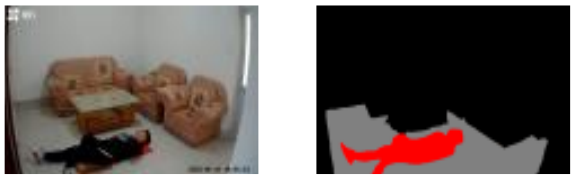
B. CNN activity recognition and fall detection

In order to prove the accuracy of the designed CNN network structure for human behavior classification, this paper has done a comparative study with the traditional support vector machine. Among them, for a fair comparison, all algorithms are tuned separately. The optimal parameters are selected for the grid search of SVM after 5-fold cross-validation on the training data, and finally the accuracy is compared on the same test set. The results are shown in the Table II. It can be seen that the designed CNN network has achieved a high accuracy rate in the posture classification of sitting, standing, bending and lying, which is better than the traditional SVM. Experiment shows that the designed CNN is feasible.

method	Stand/%	Sit/%	Bend/%	Lie/%
SVM	96.83	95.24	93.97	96.51
CNN	97.78	95.56	94.60	97.46

Table 2:- Classification accuracies comparisons by different classifies

As mentioned earlier, when CNN detects a lying posture on the floor region, it will detect a fall. As shown in the Fig 4, (a) is the original scene, (b) the red area is the human body area, and the gray area is the marked ground area.



(a)original scene (b) marked scene

Fig 4:- Original scene and marked floor region

The experiment in this paper verifies the feasibility of the proposed method. The classification results of fall activities and non-fall activities are shown in TABLE III. It can be observed that out of 90 falling videos of different types, only 3 were undetected, and in the other 230 normal activity videos, 4 videos were misdetected, and the overall classification accuracy rate reached 97.81%. By analyzing the misdetected video, it may be the wrong classification of similar postures due to the camera angle of view. Follow-up research can optimize the behaviors with low recognition rate according to the recognition characteristics of the current scheme to improve the overall recognition rate.

Action	Correct number	Wrong number	Accuracy/%
Walk	60	0	100
Sit-Stand-Sit	58	2	96.67
Bend	59	1	98.33
Fall	87	3	96.67
Lie	64	1	98.46
Total	313	7	97.81

Table 3:- Human action recognition results

IV. CONCLUSION

This paper proposes a computer vision-based fall detection algorithm using dual network structure processing. The YOLACT network is used to extract the contour area of the human body, then input to the trained CNN network after preprocessing to complete the activity classification and make a fall alarm decision. Experiments show that the algorithm has strong adaptability to the scene, the accuracy of activity classification is high, and the false detection rate of falls is low. In the future work, we will study the feature fusion of multiple features to obtain a more accurate and robust fall detection system.

REFERENCES

[1]. H. Liu and Y. Guo, "A vision-based fall detection algorithm of human in indoor environment," Second International Conference on Photonics and Optical Engineering. Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, 2017.

[2]. A. Iazzi, M. Rziza and R.O.H. Thami, "Fall detection based on posture analysis and support vector machine", 2018 4th International Conference on Advanced Technologies for Signal and Image Processing (ATSIP). IEEE, 2018.

[3]. M. Saleh, N. Georgi, M. Abbas and R.L.B. Jeannès, "A Highly Reliable Wrist-Worn Acceleration-Based Fall Detector", 2019 27th European Signal Processing Conference (EUSIPCO). 2019.
[4]. D. Droghini, E. Principi, S. Squartini, P. Olivetti and F. Piazz, "Human Fall Detection by Using an Innovative Floor Acoustic Sensor", Multidisciplinary Approaches to Neural Computing. 2018.
[5]. N. Lapierre, N. Neubauer, A. Miguel-Cruz, A. Rios Rincon, L. Liu and J. Rousseau, "The state of knowledge on technologies and their use for fall detection: A scoping review", International Journal of Medical Informatics, 2018, 111(MAR.):58-71.
[6]. I. Martins, P. Carvalho, L. Corte-Real and J.L. Alba-Castro, "BMOG: boosted Gaussian Mixture Model with controlled complexity for background subtraction", Pattern Analysis and Applications, 2018, 21(3):1-14.
[7]. M. Yu , Y. Yu , A. Rhuma and S.M.R. Naqvi, "An Online One Class Support Vector Machine-Based Person-Specific Fall Detection System for Monitoring an Elderly Individual in a Room Environment", IEEE Journal of Biomedical & Health Informatics, 2013, 17(6):1002-1014.
[8]. J. Gao and H. Zhu, "Moving object detection for video surveillance based on improved ViBe", Control & Decision Conference. IEEE, 2016:6259-6263.
[9]. C. Rougier and J. Meunier, "3D Head Trajectory using a Single Camera", International Journal of Future Generation Communication and Networking, invited paper for the special issue on Image and Signal Processing, 2010,3(4): 43-54.
[10]. V. Vaidehi, K. Ganapathy, K. Mohan, A. Aldrin and K. Nirmal, "Video based automatic fall detection in indoor environment", International Conference on Recent Trends in Information Technology. IEEE, 2011.
[11]. C.Y. Lin , S.M. Wang, J.W. Hong and L.W. Kang, "Vision-based fall detection through shape features", IEEE Second International Conference on Multimedia Big Data. IEEE, 2016.
[12]. B. Mirmahboub, S. Samavi, N. Karimi and S. Shirani, "Automatic monocular system for human fall detection based on variations in silhouette area", IEEE Trans Biomed Eng, 2013, 60(2):427-436.
[13]. K. Tra and T.V. Pham, "Human fall detection based on adaptive background mixture model and HMM", 2013 International Conference on Advanced Technologies for Communications, Ho Chi Minh City, Vietnam, 2013:95-100.
[14]. K. Adhikari , H. Bouchachia and H. Nait-Charif, "Activity recognition for indoor fall detection using convolutional neural network", 2017 Fifteenth IAPR International Conference on Machine Vision Applications (MVA). IEEE, 2017.
[15]. D. Bolya, C. Zhou, F. Xiao and Y.J. Lee, "YOLACT: Real-time instance segmentation", IEEE CVF International Conference on Computer Vision, 2019:9157-9166.
[16]. W. Liu, D. uelov and D. Erhan et al. "SSD: Single shot multibox detector". 2016.
[17]. A. Dhillon and G.K. Verma, "Convolutional neural network: a review of models, methodologies and applications to object detection", Progress in Artificial Intelligence, 2020:1-28.
[18]. J. Chen and X. Ran, "Deep Learning With Edge Computing: A Review", Proceedings of the IEEE, 2019, PP(99):1-20.