A Survey of Computer Vision-Based Fall Detection and Technology Perspectives

Manling Yang[1], Xiaohu Li[2], Jiawei Liu[2], Shu Wang[3], and Li Liu[2](✉)

[1] Department of Microelectronics and Communication Engineering, Chongqing University, Chongqing, China
manling_yang@cqu.edu.cn

[2] School of Big Data and Software Engineering, Chongqing University, Chongqing, China
{xhlee,202124131055,dcsliuli}@cqu.edu.cn

[3] School of Materials and Energy, Southwest University, Chongqing, China
shuwang@swu.edu.cn

Abstract. With the increase in the number of elderly people living alone, the use of computer vision technology for real-time fall detection is of great importance. In this paper, we review fall detection based on computer vision from four perspectives: background significance, current research status, relevant influencing factors, and future research outlook. We summarized our approach by classifying the three types of input image data in fall detection systems: RGB (Red, Green, Blue), Depth, and IR (Infrared Radiation), outlining research in both target tracking and bone detection for basic image processing tasks, as well as methods for processing video data. We analyzed the possible effects of multiple factors on fall detection regarding camera selection, the individual object recognized, and the recognition environment, and collected the solutions. Based on the current problems and trends in vision-based fall detection, we present an outlook on future research and propose four new ideas including functional extensions using the easy fusion feature of Mask R-CNN (Mask Region with Convolutional Neural Network), the use of YOLO (You Only Look Once) family to improve the speed of target detection, using variants of LSTM (Long Short-Term Memory) such as GRU (Gate Recurrent Unit) to achieve more efficient detection, and using Transformer methods that have been migrated from natural language processing to computer vision for detection.

Keywords: Computer Vision · Deep Learning · Fall Detection · Neural Network · Video Surveillance System

1 Introduction

Falls are one of the main common risks faced by elderly and disabled people. A study conducted by the World Health Organization (WHO) in 2007 estimated that in people over 70 years of age, the probability of a fall event is as high as 42%, and 50% of them die unnaturally as a result of a fall [1]. Therefore, it is important to detect falls through technical means and provide timely assistance. Current fall detection methods can be

Y. Sun et al. (Eds.): ChineseCSCW 2022, CCIS 1682, pp. 595–609, 2023.
https://doi.org/10.1007/978-981-99-2385-4_45

divided into three main categories. The first category is based on wearable devices, including the monitored individual carrying sensors, and this type of system uses a wide range of technologies. The second category is based on systems with environmental arming, where sensors are placed around the monitored person, including infrared, RF sensors [2], etc. The third category is based on computer vision techniques, using cameras to capture video images for recognition, and classical common methods include Support Vector Machine (SVM) [3], Random Forest (RF) [4], Artificial Neural Networks, ANN) [5], etc. In the case of using wearable devices, elderly people need to wear the detection devices continuously, however, they may often forget to wear the devices or not wear them properly. Although the impact on the user's activity is small for environment-based sensors, the monitorable area still has limitations (Fig. 1).

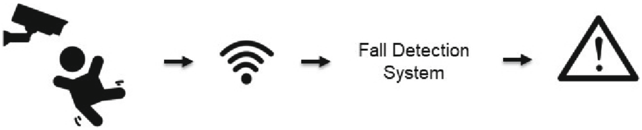

Fig. 1. General system architecture for image-based fall detection.

In summary, vision-based devices can replace the above sensors, provide a viable solution, and will be cheaper to implement. Cameras can also be installed in all rooms, with lower maintenance costs and more efficient replacement [6] (Table 1).

Table 1. Comparison of three classifications of current fall detection methods.

Compare angles	Wearable technology	Environmental placement	Based on Vision
Judgment signals	Signals from sensors	Sensor signals are placed in the environment	Image information
Hardware	Accelerometers, pressure sensors, inclinometers, microphones, etc	Pressure, acoustic, infrared, and RF sensors, etc	A variety of cameras, such as ordinary cameras, depth cameras, infrared cameras, etc
Advantage	Variety of detection information and analysis methods	Sensors don't affect the user being monitored	Less influenced by people, and visual information is more intuitive
Disadvantages	The monitored subject may refuse or forget to wear them	The equipment is influenced by the environment and sensor layout	Camera price is high, there is a risk of information leakage

2 A Review of Vision-Based Fall Detection-Related Research and Techniques

After researching related technologies, we classify the three commonly used input image data inputs in the field of computer vision into types RGB, Depth, and IR. The common datasets and fall detection models for each of these three types will be introduced in 2.1 with examples. For the basic image processing part, the more common and effective template matching algorithms and feature fusion techniques from the perspective of image processing techniques such as target tracking, motion prediction, and human skeleton detection are selected for illustration in 2.2. In 2.3, we focus on video data processing, with emphasis on image data processing with the addition of timeline, i.e., time-dependent sequential data, including two methods of deep bidirectional LSTM and video motion detection.

2.1 Example of Methods Related to Classification by Input Image Data Type

We first collected a summary of common datasets in current fall detection research, which is summarized in Table 2.

Table 2. Summary table of common falls datasets.

Dataset Name	Data Type	Features
Multi-camera fall dataset [7]	RGB	Multi-camera system acquisition contains fall simulations and normal daily activities in realistic situations
Le2i [8]	RGB	There is a realistic dataset of 191 videos containing fake falls and video frames without people
UR fall detection [9]	RGB	Contains a sequence of 30 falls and 40 activities of daily living, using a Kinect camera to record fall events
UP fall detection dataset [10]	RGB	17 healthy young adults performing 11 activities with data from wearable sensors, environmental sensors, and visual devices
SDU falls [11]	Depth	Kinect depth camera acquisition, consisting of six types of movements from ten subjects
TST Fall Detection [12]	Depth	Includes depth images and skeletal joint data collected using Microsoft Kinect v2
CMU Graphics Lab [13]	RGB and Depth	A total of 2605 sequences, divided into 6 categories and 23 subcategories

(*continued*)

Table 2. (*continued*)

Dataset Name	Data Type	Features
IASLAB-RGBD fall dataset [14]	RGB and Depth	Contains 15 different people, acquired in two different laboratory environments
Fall Detection Dataset [15]	RGB and Depth	Raw RGB and depth images were recorded using a single uncalibrated Kinect sensor, consisting of 8 different views in 5 different rooms with 5 different participants
CMD fall dataset [16]	RGB and Depth	Acquired by seven overlapping Kinect sensors and two wearable accelerometers, including 20 activities from 50 subjects and multimodal multi-view data

Deep Learning Fall Detection Based on RGB Images with Parameter Optimization. G. Anitha and S. Baghavathi Priya proposed a new image-based fall detection system that involves different operational stages including pre-processing of images, feature extraction, classification, and parameter optimization of the detection system [17] (Fig. 2).

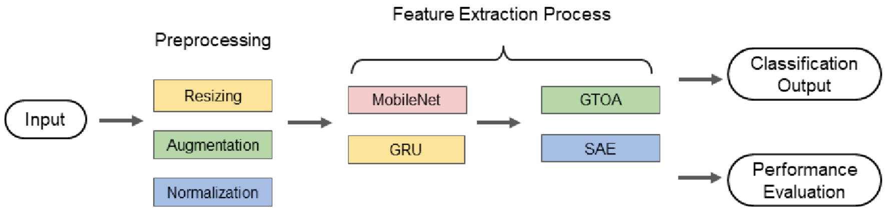

Fig. 2. The working process of the VEFED-DL model.

To improve the image quality and eliminate noise, the system will process the extracted frames at three levels including resizing, image enhancement, and min-max-based normalization [17].

IoT Fall Detection System Based on Deep Image HOG-SVM. The highly developed IoT technology and machine learning have enabled multimedia devices to be used in various environments where special people need to be protected. The Ritsumeikan University proposed a fall detection IoT system for the elderly based on HOG-SVM (Histogram of Oriented Gradients-Support Vector Machine) [18].

Non-invasive Multi-person Fall Detection Based on IR Images. Most of the studies on thermal vision-based fall detection methods, mainly focus on single-person occupancy scenarios, so they are not fully applicable to real life (Fig. 3).

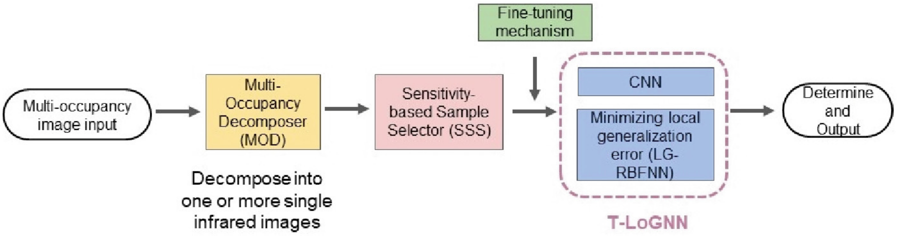
Fine-tuning mechanism
Multi-occupancy image input
Multi-Occupancy Decomposer (MOD)
Sensitivity-based Sample Selector (SSS)
CNN
Minimizing local generalization error (LG-RBFNN)
Determine and Output
Decompose into one or more single infrared images
T-LoGNN

Fig. 3. Multi-Person Fall Detection (MoT-LoGNN) Method Flow.

The Key Laboratory of Computational Intelligence and Cyberspace Information of the South China University of Technology proposed a non-invasive thermal vision-based fall detection method for multiple people, which consists of four components: T-LoGNN, fine-tuning mechanism, multi-occupancy decomposer (MOD), and sensitivity-based sample selector (SSS) [19].

2.2 Basic Image Processing in Fall Detection

Target Tracking Based on Template Matching Algorithm. Motion target detection is the basis for advanced tasks such as motion target tracking and behavior recognition [20]. Intelligent video surveillance can automatically detect, identify and track targets in video scenes without human intervention [21] with the powerful computing power of computers combined with other technologies. The template matching algorithm uses the optimized absolute sum of differences (OSAD) to detect and recognize objects with high tracking accuracy, stable performance, and independence of illumination conditions [22]. Separating the target from the background, which is easy to target detection, and also the size of the tracking window can be adjusted according to the distance between the target and the camera [23]. Assuming that the average distance of the previous frame has been obtained and that the target will not go beyond that range in the next frame, all pixels in the tracking window whose depth is within that range can be selected and used for new distance calculation [23] (Fig. 4).

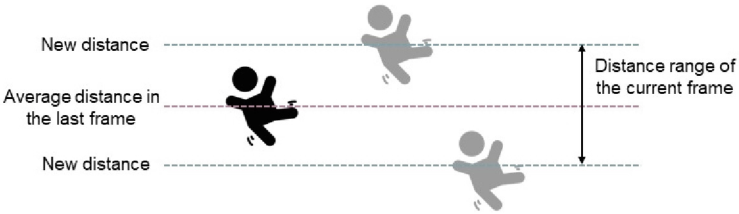
New distance
Average distance in the last frame
New distance
Distance range of the current frame

Fig. 4. Average distance measurement schematic

Feature Fusion-Based and Human Skeleton Detection Action Recognition. Most current action recognition methods can be divided into the following three categories:

depth sequence-based, human skeleton detection-based, and feature fusion based. The depth sequence-based methods use various deep learning models to process and judge the RGB-type image information, and their main advantage is the richness of appearance information. The human skeletal sequence-based approach uses the changes in human joint points between video frames to describe the action, including the relative position and appearance changes of the joint points [24]. (Fig. 5).

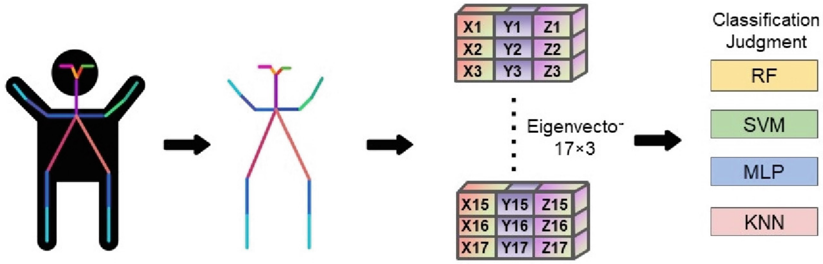

Fig. 5. Workflow of Skeletal Recognition.

The main problem with skeleton-based detection is that when occlusion occurs in the scene, the estimation of joint points can be lost, which affects the action recognition results. The fusion of skeletal features and depth information features can effectively overcome skeletal feature errors due to occlusion and perspective changes [25].

The main issue to consider when using feature fusion techniques is how multiple types of data can be fused to make them more effective in their ways. However, the problem exists that multimodal data fusion requires handling more data volume, higher feature dimensionality, and more complex action recognition computations [26].

2.3 Typical Video Data Processing Methods

Deep Bidirectional LSTM Based Video Sequence Action Recognition. Videos are sequential data in which the motion of the visual content is represented in a sequence of frames, and the sequence of frames helps to understand the context of the action. Long-term sequences are subject to forgetting the earlier inputs of the sequence, and such problems are known as gradient disappearance problems. LSTM can be used to solve such problems [27]. Sejong University proposed a novel approach to action recognition by processing video data using convolutional neural networks (CNN) and deep bi-directional LSTM (Densely-connected Bi-directional Long Short-Term Memory, DB-LSTM) networks [28]. In a bidirectional LSTM, the output at time t depends not only on the previous frame in the sequence but also on the upcoming later frames. In contrast, a bidirectional RNN (Recurrent Neural Network) is a stacking of two RNNs together, with one RNN proceeding forward and the other RNN proceeding backward.

DNN-Based Video Motion Monitoring. DNNs are suitable for dealing with problems related to time series, and videos are time-dependent, so video motion detection requires using the current frame, previous frame, and next frame of a given video. The Department of Computer Science at Auckland University of Technology proposed a deep learning-based model to implement video motion detection by combining CNN and RNN to build a DNN (Deep Neural Network) to accomplish video motion detection [29]. Integrating CNN and RNN can significantly reduce the size of video data and training time. However, the system only implements dynamic video detection, which cannot accomplish real-time object tracking and dynamic event recognition, and requires a large number of real videos for training and testing to produce more accurate results.

3 Other Influencing Factors Analysis

3.1 Camera Choice

RGB Cameras and RGB-D Cameras. Traditional RGB color cameras can only capture image data within the camera's field of view, record the size of the R, G, and B values of pixel points, and image as a 2-dimensional color image. The disadvantage is that the acquired image data information is extremely limited. For RGB-D depth camera or called a 3D camera, can obtain RGB images and depth images at the same time.

Depth Camera Comparison. Depth cameras can be divided into three types according to their principles: active projection structured light depth cameras, passive binocular depth cameras, and depth cameras based on Time of Flight (TOF) measurements (Table 3).

Table 3. Comparison of the main performance and parameters of three different principles of depth cameras.

Principle	Active projection structured light	Passive binocular	Based on the reflection time length
Measurement Accuracy	Decreases with increasing distance	Short distance range 0.01mm to 1cm	Stable at 1cm
Dark Environment	Applicable	Not applicable	Applicable
Main Advantages	Low power consumption, low cost, suitable for low light Suitable for use in low light conditions	Low hardware cost, can be used indoors and outdoors, strong and low light effect reduced	Longer measuring distance and maintain accuracy, can directly output 3D data of the measured object
Main Disadvantages	Poor accuracy at long distances, strong light easily interferes with the projected light	In strong/low light, the single environment has a big impact and can cause matching failure and complicated algorithm	Stable but not high accuracy, high time measurement requirements, basically cannot be used under outdoor bright light conditions
Applicable Scenarios	Smartphone front photography, face recognition, AR/VR, etc	Driverless, gesture recognition, depth detection, etc	Dynamic scenes unmanned, smartphone rear photography, etc

Table 4. Infrared imaging principle and infrared lens selection

Classification	Passive infrared imaging	Detect the infrared radiation of the target, since only the distribution of the object's temperature
	Active infrared imaging	Infrared lights produce infrared radiation to irradiate the object
Considerations	Infrared Sensing	Infrared light will cause a color CCD off-color impact, so ordinary color cameras will use a filter to filter out. Color infrared cameras require the use of a dual-peak single filter
	Lens selection	Improper matching of the sensor and lens will appear dark corners or lens angle waste. The lens angle also needs to match the angle of the selected infrared light emission
	Power supply selection	General power consumption needs to be greater than 20% of the overall system requirements. If below this value, the power supply will be fully loaded state

Infrared Camera Types. The biggest advantage of an infrared lens is to be able to apply it to the night vision scene. Infrared light selection is the choice of an infrared camera is a very important issue, you need to consider the camera, lens, power supply and other aspects of the comprehensive then choose (Table 4).

3.2 Individual Influence of the Identified Object

Basic User Parameters. In practical applications, the system may not be able to process the prepared training data as expected due to the differences in users and usage scenarios. In the system proposed by Ritsumeikan University Institute of Science and Technology for solving this problem [30], the method used is that the basic parameters of the user are calculated by the edge node and sent to the cloud, and then the cloud server calculates the best detection model and sends it back to the edge node.

Due to the small amount of data in the overall test condition, as much data as possible needs to be collected for the model to be trained in the follow-up. At the same time, a portion of experimenters with closer body sizes should be selected to reduce the influence of the model by individual differences such as height and weight.

3.3 Environmental effects during recognition

Light Condition Changes. Improving the accuracy of fall detection in complex environments such as changing room light conditions is an important issue in RGB image-based fall detection. The problem of the temporal evolution of visual data is solved by using dynamic images [31]. Sagar Chhetri and Abeer Alsadoon et al. from the School of Computer and Mathematics at Charles Sturt University proposed a mechanism that can improve the performance of image preprocessing by capturing each dynamic action in a video into a single image using a dynamic optical flow technique [32] (Table 5).

Table 5. Comparison of the prior art with the solution proposed by this system [32]

Comparative Aspects	Current Solutions	Proposed Solution
Method Name	TVL-1 optical flow algorithm	Enhanced dynamic optical flow algorithm
Accuracy	Improved accuracy of fall detection sensitivity under stable lighting conditions	Improves the accuracy of fall detection sensitivity under dynamic lighting conditions
Processing Time	Higher processing time and required processing power in the pre-processing stage	Reduces processing time in the image pre-processing stage using enhanced dynamic optical flow

It solves the process of processing video into image sequences, an operation that requires high processing power, and can reduce the processing power required for preprocessing. Encoding the temporal data of the optical flow video by rank pooling not only reduces the processing time but also improves the performance of the fall detection classifier under various lighting conditions.

Impact of Camera Height and Layout on Recognition Accuracy. Cameras placed in a lower position may have problems with occlusion and single monitoring view. Due to room differences, users will not set the cameras or sensors at the same height. Ritsumeikan University proposed the enhanced tracking algorithm and the noise-reducing Alex-Net (ETDA-Net) algorithm to improve the related performance [33]. In the later research, an algorithm of camera adaptive height calculation can be introduced to accurately measure the camera height and input to the model, to improve the model's adaptability to image input at different position heights. To adapt to single-camera equipment conditions, a late fusion technique is proposed that can improve the accuracy of existing fall detection systems [34] (Fig. 6).

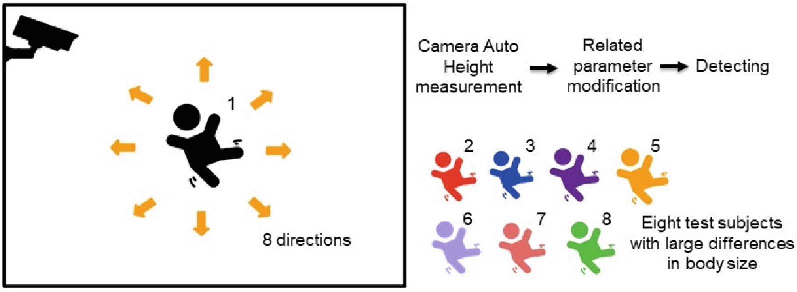
1
8 directions
Camera Auto Height measurement
Related parameter modification
Detecting
2
3
4
5
6
7
8
Eight test subjects with large differences in body size

Fig. 6. Experimental design diagram

4 Research Outlook

4.1 Image Data Type Selection

The image data type selection is considered from two perspectives: detection effect and practical application. For better detection effect, RGB type is a very classical data type in image processing, while Depth type can solve the problems of privacy and influence by ambient light that exist in RGB type. Meanwhile, the biggest difficulty of the IR type is that there are fewer public data sets. When conditions allow, we can choose to use RGB-D cameras to acquire both RGB and Depth image data inputs, RGB images for capturing color and appearance information, and Depth images for scenes with changing illumination conditions [35]. For a wider range of practical applications, if we choose to add a portable type of fall detection function module to the existing intelligent surveillance system, most of the image data of the current surveillance systems used in real life are RGB images with rich public data sets.

4.2 Target Detection Module

In the field of target detection, it can be divided into two mainstream types, one is the One-Stage algorithm of the Yolo series, SSD, etc., which directly predicts the class and location of different targets using only one CNN network. The other type is the R-CNN series of algorithms based on candidate regions (Region Proposal), which are Two-Stage and need to use Selective Search or CNN network (RPN) to generate candidate regions first, and then perform classification and regression. The advantage of the first type of method is faster detection, but the accuracy is relatively low, and the second type of method has relatively high accuracy but slow detection.

Mask R-CNN Combined with Human Bone Detection. CNN is a traditional target detection algorithm that has undergone continuous development and has undergone three progressive development stages, R-CNN, Fast R-CNN, and Faster R-CNN, with each structure contrasted as in Fig. 8.

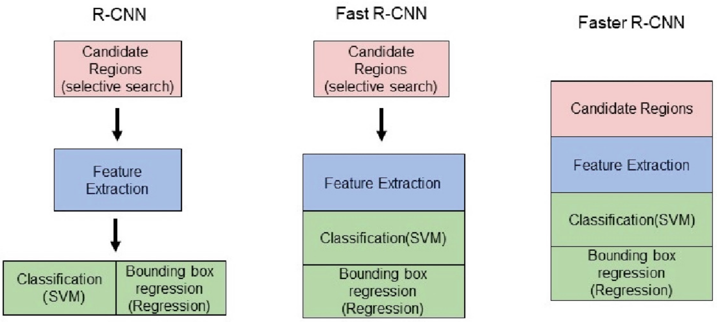
R-CNN
Candidate Regions (selective search)
Feature Extraction
Classification (SVM)
Bounding box regression (Regression)
Fast R-CNN
Candidate Regions (selective search)
Feature Extraction
Classification(SVM)
Bounding box regression (Regression)
Faster R-CNN
Candidate Regions
Feature Extraction
Classification(SVM)
Bounding box regression (Regression)

Fig. 7. Schematic diagram of R-CNN, Fast R-CNN, and Faster R-CNN frameworks

From the three layers at the beginning to the final unification into one layer, the parameters and operations were reduced and the detection speed was accelerated (Fig. 7).

The Mask R-CNN, which emerged subsequently, is a continuation of the development of the Faster R-CNN. In the paper that published Mask R-CNN, the authors then combined Mask R-CNN with KeyPoints Detection partly [36]. A similar idea was proposed by Sara MOBSITE et al. in 2018, where they used Mask R-CNN to output an image of a human silhouette. This can reduce unnecessary processing in the subsequent steps [37]. Linear interpolation is used in the OpenPose-based skeleton detection and LSTM/GRU model fall detection framework proposed by Chuan-Bi Lin et al. to compensate for the loss of joint points [38] to address the compensation of missing values. The problem that may arise when using image features captured by a normal RGB camera to acquire the skeleton is when the body is overlapped, occluded, and the body contours are not clear, all of which can cause losses and errors in generating the skeleton.

YOLO Series Selection. The Advantage of the YOLO Series Algorithm Over Two-Stage is Its Ability to Automatically Extract Features and Complete the Detection of Target Frames and End-To-End Prediction of Categories in One Go.

YOLO can achieve the best current performance while the inference speed is highly competitive. Most of the fall detection systems currently using the YOLO family of algorithms use the YOLO V5 algorithm. It has better convertibility and is easier to improve the detection accuracy, especially the YOLO V5S model with a smaller depth and width product factor, which helps to reduce the deployment cost [39]. Few attempts have been made to use the latest YOLO X. According to the data given in the paper, the detection speed can reach the millisecond level at the earliest [40], so it can achieve the requirement of target detection tracking for fall detection.

4.3 Falling Action Judgment

Improvements to the Problems of Current LSTM-Based Methods. In the case of detection based on the need to process video stream data, LSTM can handle the task of time series better than CNN as well as RNN. Also, LSTM solves the long-term dependency problem of RNN and alleviates the problem of gradient disappearance and gradient explosion caused by RNN backpropagation during training. But this is more time-consuming to train as the model structure of LSTM itself is relatively complex for the requirements of a fall detection system that needs fast detection.

If additional LSTM units need to be added to the network, the performance of the model can be improved, while the computational complexity increases accordingly. When increasing from using 24 to 512 units, the LSTM computation becomes approximately 10 times slower [41].

Advantages of GRU (LSTM variant) Implementation. Classification of fall events requires both temporal and spatial features to be considered. Recurrent neural networks can extract temporal features by remembering the necessary information from the past. However, the problem of gradient disappearance and gradient explosion may occur. Using Gated Recurrent Unit (GRU) network with its update and reset gates can solve this problem by being able to decide which information needs to be passed as output vectors. GRU is a variant of LSTM with a simple architecture, so GRU is faster than LSTM in terms of unit operations and detection speed.

Since LSTM is complex, many variants have been generated, and GRU is the most commonly used one among all LSTM variants. Other variants can be tried for comparison in the course of subsequent research.

4.4 Application of Transformer to Vision Domain

Although LSTM solves the problem of the limited memory length of RNN, it still cannot be parallelized. The data at the moment t_0 must be computed before the data at the moment $\mathrm{t}_0 + 1$ can be computed. Google proposed Transformer to replace the previous temporal memory network [42], and the memory time can be infinitely long if it is not affected by the hardware conditions. The parallelized implementation can greatly accelerate the training speed. The first proposed for the vision domain is the Vision Transformer (ViT), which applies the standard Transformer directly to images with as little modification as possible. When the model is pre-trained at a large enough scale and transferred to a classification task with fewer data points, the accuracy can be able to obtain significant improvement [43]. Swin Transformer [44] has made significant progress on a variety of image data processing tasks in the CV domain, mainly in three categories: image classification, target detection, and semantic segmentation.

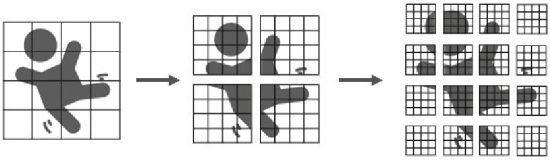

Fig. 8. Schematic diagram of Swin Transformer downsampling operation

The use of different downsampling rates in Backbone here helps to build target detection, instance segmentation, and other tasks on this basis. Secondly, unlike in ViT where Multi-Head Self-Attention (MSA) is performed directly on the whole feature map, Swin Transformer uses the concept of Windows Multi-Head Self-Attention (W-MSA) to divide the feature map into multiple disjoint regions, which in the paper It is called Window. Multi-Head Self-Attention is performed only within each window. This can reduce the computational effort when the shallow feature map is large (Fig. 9).

In early 2022 Microsoft Research Asia proposed Swin Transformer V2 [45] in response to three more major problems of the Swin Transformer for training and application to large visual models: instability of training, resolution gap between pre-training and fine-tuning and the large demand for labeled data. The corresponding solutions and improvements are proposed for each of the three problems mentioned above. It can be seen that the application of Transformer in the field of vision is developing rapidly and has great potential.

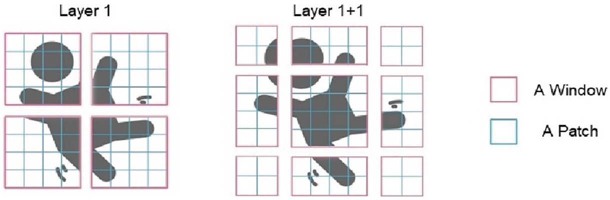

Fig. 9. Shifted Windows Multi-Head Self-Attention(SW-MSA) schematic

4.5 Future Research

Many insights about future research on optimization can be obtained from the updates of many previous methods. Firstly, inspired by the continuous updates of R-CNN, an important optimization idea is to improve the network's comprehensiveness and reduce the required parameters as much as possible. Secondly, based on the continuous updates of the YOLO series, it can be inspired that for the calculation of the loss values in the training part of the model, the equivalent conversion or simplification of the mathematical formula from the loss calculation is calculated.

Many current types of research in the field of vision not only improve and evolve the classical methods but also gradually produce completely new structural models different from the traditional methods. For example, the traditional LSTM method mentioned earlier has a large number of variants as well as the emergence of a completely new Transformer. Another example is the direct prediction for human pose judgment by the classical regression-based approach. Then the heat map-based approach gradually emerged, where the detection is performed by predicting the fraction of each key point appearing at each position. More and more methods have appeared in the field, but the goal has always been to improve detection accuracy and speed. At the same time, more researchers are focusing on improving the generality of the methods for more usage scenarios, and the three points mentioned above can be used as a guide for future research purposes.

Acknowledgment. This work was supported by grants from the National Natural Science Foundation of China (grant no. 61977012), the China Scholarship Council (grant no. 201906995003), the Central Universities in China (grant no. 2021CDJYGRH011), and the Key Research Programme of Chongqing Science \& Technology Commission (grant no. Cstc2019jscx-fxydX0054).

References

1. WHO global report on falls prevention in older age (2007)
2. Gutierrez, J., Rodriguez, V., Martin, S.: Comprehensive review of vision-based fall detection systems. Sens.-Basel **21**(3) (2021)
3. Chen, Z.J., Wang, Y.: Infrared-ultrasonic sensor fusion for support vector machine-based fall detection. J. Intell. Mater. Syst. Struct. **29**(9), 2027–2039 (2018)
4. Msaad, S., Cormier, G., Carrault, G.: Detecting falls and estimation of daily habits with depth images using machine learning algorithms. In: 42nd Annual International Conference of the IEEE-Engineering-in-Medicine-and-Biology-Society (EMBC), Montreal, Canada, pp. 2163–2166. IEEE (2020)

5. Yodpijit, N., Sittiwanchai, T., Jongprasithporn, M.: The development of artificial neural networks (ANN) for falls detection. In: 2017 3rd International Conference on Control, Automation and Robotics (ICCAR), pp. 547–50. IEEE (2017)
6. Ramanujam, E., Padmavathi, S.: A vision-based posture monitoring system for the elderly using intelligent fall detection technique. In: Mahmood, Z. (eds.) Guide to Ambient Intelligence in the IoT Environment. Computer Communications and Networks. Springer, Cham, pp. 249–69 (2019). https://doi.org/10.1007/978-3-030-04173-1_11
7. Auvinet ER, C., Meunier, J., St-Arnaud, A., Rousseau, J.: Multiple cameras fall data set. Technical Report Number 1350. University of Montreal: Montreal, QC, Canada (2011)
8. Charfi, I.: Optimized spatio-temporal descriptors for real-time fall detection: comparison of support vector machine and Adaboost-based classification. J. Electr. Imag. **22**(4) (2013)
9. Kepski, M., Kwolek, B.: Embedded system for fall detection using body-worn accelerometer and depth sensor. In: International Workshop Intelligent Data, pp. 755–759 (2015)
10. Martinez-Villasenor, L., Ponce, H., Brieva, J., Moya-Albor, E., Nunez-Martinez, J., Penafort-Asturiano, C.: UP-fall detection dataset: a multimodal approach. Sensors (Basel) **19**(9) (2019)
11. Ma, X.: Depth-based human fall detection via shape features and improved extreme learning machine. IEEE J. Biomed. Health Inf. **18**(6) (2014)
12. Cippitelli, E., Gambi, E., Gasparrini, S., Spinsante, S.: TST Fall detection dataset v2. IEEE Dataport (2016)
13. CMU Graphics Lab—Motion Capture Library (2021). http://mocap.cs.cmu.edu/
14. Munaro, M.: A feature-based approach to people re-identification using skeleton keypoints (2014)
15. Adhikari, K.: Activity recognition for indoor fall detection using convolutional neural network (2017)
16. Tran, T.H.: A multi-modal multi-view dataset for human fall analysis and preliminary investigation on modality (2018)
17. Anitha, G., Priya, S.B.: Vision based real time monitoring system for elderly fall event detection using deep learning. Comput. Syst. Sci. Eng. **42**(1), 87–103 (2022)
18. Kong, X.B., Meng, Z.L., Nojiri, N., Iwahori, Y., Meng, L., Tomiyama, H.: A HOG-SVM based fall detection IoT system for elderly persons using deep sensor. Procedia Comput. Sci. **147**, 276–282 (2019)
19. Zhong, C.N., Ng, W.W.Y., Zhang, S., Nugent, C.D., Shewell, C., Medina-Quero, J.: Multi-occupancy fall detection using non-invasive thermal vision sensor. IEEE Sens. J. **21**(4), 5377–5388 (2021)
20. Feng, Z., Zhu, X., Xu, L., Liu, Y.: Research on human target detection and tracking based on artificial intelligence vision. In: 2021 IEEE Asia-Pacific Conference on Image Processing, Electronics and Computers (IPEC), pp. 1051–1054 (2021)
21. Velasquez, J., Piech, K., Lehnhoff, S., Fischer, L., Garske, S.: Incremental development of a co-simulation setup for testing a generation unit controller for reactive power provision. Comput. Sci. Res. Dev. **32**(1–2), 3–12 (2016). https://doi.org/10.1007/s00450-016-0319-2
22. Satish, B., Jayakrishnan, P.: Hardware implementation of template matching algorithm and its performance evaluation. In: 2017 International Conference on Microelectronic Devices, Circuits and Systems (ICMDCS) (2017)
23. He, S.S., Liang, A., Lin, L., Song, T.: A continuously adaptive template matching algorithm for human tracking. In: 2017 First IEEE International Conference on Robotic Computing (IRC), pp. 303–309 (2017)
24. Ramirez, H., Velastin, S.A., Meza, I., Fabregas, E., Makris, D., Farias, G.: Fall detection and activity recognition using human skeleton features. IEEE Access **9**, 33532–33542 (2021)
25. Chaaraoui, A.A., Padilla-Lopez, J.R., Florez-Revuelta, F.: Fusion of skeletal and silhouette-based features for human action recognition with RGB-D devices. In: 2013 IEEE International Conference on Computer Vision Workshops (ICCVW), pp. 91–97 (2013)

26. Zhang, H.B., Zhang, Y.X., Zhong, B., Lei, Q., Yang, L., Du, J.X., et al.: A comprehensive survey of vision-based human action recognition methods. Sensors (Basel) **19**(5) (2019)
27. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. **9**(8), 1735–1780 (1997)
28. Ullah, A., Ahmad, J., Muhammad, K., Sajjad, M., Baik, S.W.: Action recognition in video sequences using deep bi-directional LSTM with CNN features. IEEE Access **6**, 1155–1166 (2018)
29. Luo, H., Liao, J., Yan, X., Liu, L.: Oversampling by a constraint-based causal network in medical imbalanced data classification. In: 2021 IEEE International Conference on Multimedia and Expo (ICME), pp. 1–6 (2021)
30. Chen, Y., Kong, X., Meng, L., Tomiyama, H.: An edge computing based fall detection system for elderly persons. Procedia Comput. Sci. **174**, 9–14 (2020)
31. Fan, Y.X., Levine, M.D., Wen, G.J., Qiu, S.H.: A deep neural network for real-time detection of falling humans in naturally occurring scenes. Neurocomputing **260**, 43–58 (2017)
32. Chhetri, S., Alsadoon, A., Al-Dala'in, T., Prasad, P.W.C., Rashid, T.A., Maag, A.: Deep learning for vision-based fall detection system: enhanced optical dynamic flow. Comput. Intell. **37**(1), 578–595 (2020)
33. Kong, X., Chen, L., Wang, Z., Chen, Y., Meng, L., Tomiyama, H.: Robust self-adaptation fall-detection system based on camera height. Sensors (Basel) **19**(17) (2019)
34. Baldewijns, G., Debard, G., Mertes, G., Croonenborghs, T., Vanrumste, B.: Improving the accuracy of existing camera based fall detection algorithms through late fusion. In: P Annual International IEEE EMBS, pp. 2667–2671 (2017)
35. Khraief, C., Benzarti, F., Amiri, H.: Elderly fall detection based on multi-stream deep convolutional networks. Multimed. Tools Appl. **79**(27–28), 19537–19560 (2020). https://doi.org/10.1007/s11042-020-08812-x
36. He, K., Gkioxari, G., Doll´ar, P., Girshick, R.: Mask R-CNN. In: Facebook AI Research (FAIR) (2018)
37. Mobsite, S., Alaoui, N., Boulmalf, M.: A framework for elders fall detection using deep learning. IEEE (2020)
38. Lin, C.-B., Dong, Z., Kuan, W.-K., Huang, Y.-F.: A framework for fall detection based on OpenPose skeleton and LSTM/GRU models. Appl. Sci. **11**(1) (2020)
39. Yin, Y., Lei, L., Liang, M., Li, X., He, Y., Qin, L.: Research on fall detection algorithm for the elderly living alone based on YOLO. In: 2021 IEEE International Conference on Emergency Science and Information Technology (ICESIT), pp. 403–408 (2021)
40. Ge, Z., Liu, S., Wang, F., Li, Z., Sun, J.: YOLOX: exceeding YOLO Series in 2021 (2021)
41. Lu, N., Wu, Y., Feng, L., Song, J.: Deep learning for fall detection: three-dimensional CNN combined with LSTM on video kinematic data. IEEE J. Biomed. Health Inform. **23**(1), 314–323 (2019)
42. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., et al.: Attention is all you need. Comput. Lang. (2017)
43. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., et al.: An image is worth 16x16 Words: transformers for image recognition at scale. Google Res. Brain Team (2021)
44. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., et al.: Swin transformer: hierarchical vision transformer using shifted windows. In: Microsoft Research Asia (2021)
45. Liu, Z., Hu, H., Lin, Y., Yao, Z., Xie, Z., Wei, Y., et al.: Swin transformer V2: scaling up capacity and resolution. In: Microsoft Research Asia (2022)