Fall Detection Algorithm for Elderly People Living Alone Based on ARKit and YOLOv5

Hui Jin
School of Computing and Engineering
University of Missouri-Kansas City
Kansas City, United States
jinhui13020609@163.com

Jayadithya Nalajala
School of Computing and Engineering
University of Missouri-Kansas City
Kansas City, United States
jnd9g@umkc.edu

Saniya Pandita
School of Computing and Engineering
University of Missouri-Kansas City
Kansas City, United States
spd6h@umkc.edu

Sai Jahnavi Devabhakthuni
School of Computing and Engineering
University of Missouri-Kansas City
Kansas City, United States
sdcvp@umkc.edu

***Abstract*— As the aging population continues to grow, the safety of elderly individuals living alone has become a pressing societal concern. The evolving social landscape presents new challenges for elderly care, particularly in addressing the high incidence of falls. These falls not only underscore the physiological vulnerability of the elderly but also highlight the deficiencies in current safety monitoring systems. Falls frequently result in severe injuries, reduced quality of life, and even life-threatening conditions. Therefore, developing a reliable and real-time fall detection technology is critical to ensuring the safety and well-being of this vulnerable demographic. This paper presents a fall detection algorithm tailored for elderly individuals living alone, leveraging the synergistic strengths of ARKit, YOLOv5, and Apple Vision Pro. The proposed system combines YOLOv5's rapid object detection, ARKit's precise spatial tracking, and Vision Pro's immersive environmental awareness to achieve accurate and low-latency fall detection. Furthermore, we introduce a novel enhancement by integrating YOLOv10's Top-1 bounding box selection strategy into the YOLOv5 pipeline, significantly improving detection precision and reducing false positives. This fusion of technologies enables swift identification of fall events, shortens emergency response times, and ultimately minimizes the physical and psychological consequences of such incidents.**

Keywords—elderly people, ARKit, YOLOv5, fall detection

I. INTRODUCTION

As the problem of aging in society continues to intensify, the life safety of elderly people living alone has gradually attracted widespread attention from the society. The new social environment has brought new challenges to the safety management of the elderly. In the current social context, the frequent occurrence of elderly falls not only reflects the physiological vulnerability of the elderly group, but also exposes the shortcomings of the existing safety protection system in dealing with the problem of elderly falls. Fall accidents often cause serious injuries to the elderly, affect their quality of life, and even threaten their life safety. Therefore, developing a technology that can accurately and quickly detect falls of elderly people living alone and respond in time is of great significance to improving the life safety of the elderly.

At present, there have been some studies on the problem of elderly fall detection[1]. Traditional fall monitoring methods are mainly based on sensors[2] or video surveillance technology[3], but there are certain limitations and problems. The sensor is fixedly installed in a specific position and cannot perceive the dynamics of the whole body, which is prone to blind spots; and video surveillance technology is limited by camera layout, lighting conditions and privacy protection, making it difficult to achieve real-time and accurate fall detection[4]. At the same time, traditional machine learning methods have the disadvantages of cumbersome detection steps, poor real-time performance, bloated model deployment, and lack of robustness in complex scenarios[5].

In order to solve the shortcomings of traditional methods, fall detection for elderly people living alone based on augmented reality and target detection algorithms has received widespread attention in recent years. AR technology can integrate virtual content with the real world, providing a more natural and intuitive way of interaction for elderly people living alone. The target detection algorithm can accurately detect fall behaviors in images or videos in real time, improving the accuracy and reliability of fall detection. The emergence of neural networks provides an effective way for feature extraction and utilization[6].

This paper mainly designs a fall detection algorithm for elderly people living alone based on ARKit and YOLOv5, aiming to provide an effective monitoring method for elderly people living alone and detect fall events in time. This will help shorten the rescue time and reduce the damage of fall events to the physical and mental health of elderly people living alone.

II. RELATED WORKS

In recent years, intelligent health monitoring systems based on deep learning have made significant progress, especially with the rapid development of augmented reality technology. Our work combines ARKit and YOLOv5 algorithms to design a fall detection algorithm for elderly people living alone. The system realizes 3D spatial modeling of the environment through ARKit, and combined with the YOLOv5 object detection algorithm, can analyze monitoring data in real time and accurately identify events where the elderly fall. Based on multi-sensor fusion technology, the system can continuously monitor the activities of elderly people living alone through efficient information transmission, improve the response speed and accuracy to abnormal behaviors, and effectively ensure the home safety of the elderly.

A. ARKit

ARKit is an augmented reality development kit launched by Apple, which has powerful image recognition and tracking capabilities. ARKit integrates device camera image information and device motion sensor (including LiDAR) information. ARKit integrates device motion tracking, camera

image acquisition, image visual processing, scene rendering and other technologies, and provides a simple and easy-to-use API (Application Programming Interface) to facilitate developers to develop AR applications. Developers no longer need to pay attention to the underlying technical implementation details, which greatly reduces the difficulty of AR application development. ARKit uses the image information (including information collected by LiDAR) collected by the monocular camera of mobile devices (including mobile phones and tablets) to realize advanced functions such as plane detection and recognition, scene geometry, ambient light estimation, ambient light reflection, image recognition, 3D object recognition, face detection, and human motion capture. On this basis, it can create a scene of virtual and real fusion. ARKit integrates AVFoundation, CoreMotion, and CoreML frameworks, and integrates and expands on this basis. Among them, AVFoundation is a framework for processing time-based multimedia data, CoreMotion is a framework for processing sensor data information such as accelerometers, gyroscopes, and LiDAR, and CoreML is a machine learning framework. ARKit combines video image information from AVFoundation with device motion sensor data from CoreMotion, and uses CoreML computer image processing and machine learning technology to provide developers with a stable three-dimensional digital environment.

In the future, ARKit technology has broad application prospects in the field of fall detection for elderly people living alone, which can further improve the quality of life and health management level of elderly people living alone.

B. *YOLOv5*

YOLOv5 (You Only Look Once)[7] is an object detection algorithm based on deep learning. The algorithm can achieve real-time object detection and positioning by converting the object detection task into a regression problem. Compared with its predecessor YOLOv4, YOLOv5 has achieved a better balance between accuracy and speed.

The core idea of the YOLOv5 algorithm is to divide the input image into several grids and predict the bounding box and category probability of the target on each grid.

The backbone network used by the YOLOv5 algorithm adopts the CSPDarknet53 architecture, which can extract high-dimensional feature representations from the input image. These feature maps contain rich target information, which is conducive to accurate detection of targets.

The YOLOv5 algorithm introduces a feature fusion mechanism, which can improve the detection ability of targets of different sizes by fusing feature maps of different scales. Specifically, the YOLOv5 algorithm fuses low-resolution feature maps with high-resolution feature maps, thereby retaining both global context information and detail information.

The YOLOv5 algorithm achieves efficient and accurate target detection by converting the target detection task into a regression problem and adopting key steps such as anchor box generation, feature extraction, feature fusion and target detection. The algorithm has the characteristics of high speed, high precision and flexibility, and has broad application prospects in fall detection of elderly people living alone.

III. SYSTEM ARCHITECTURE

The algorithm aims to design an efficient and reliable method for detecting falls of elderly people living alone. It combines the 3D spatial modeling of ARKit with the object detection technology of YOLOv5, and achieves accurate modeling of the monitoring scene through the depth perception and scene tracking capabilities provided by ARKit. The core detection algorithm uses YOLOv5, which can analyze the video stream in real time, detect the posture and behavior patterns of the elderly, and accurately identify fall events. By combining the spatial positioning of ARKit and the fast visual detection of YOLOv5, the algorithm can quickly make judgments when the elderly fall, providing support for further alarms and processing.

The algorithm consists of multiple modules. First, the input image or video is loaded through the data loader, and then enters the preprocessing module for preliminary processing. The processed data enters the feature extraction network (CNN)[8] the YOLO network to extract key features, and is classified and regressed through the structured output network. The loss function is used for feedback optimization to improve detection accuracy. Finally, the output is processed in combination with the spatial information of ARKit to generate the fall detection result. Fig. 1 presents the YOLOv5 network structure diagram, and Fig. 2 illustrates the ARKit-based YOLOv5 algorithm flowchart.

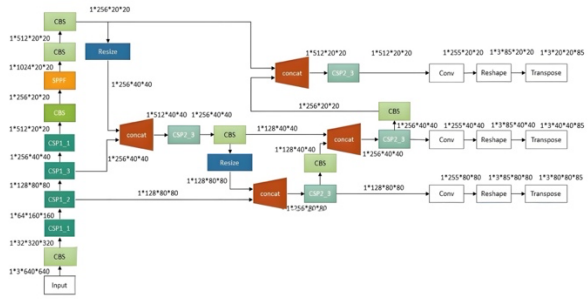

Fig. 1. YOLOv5 network structure diagram

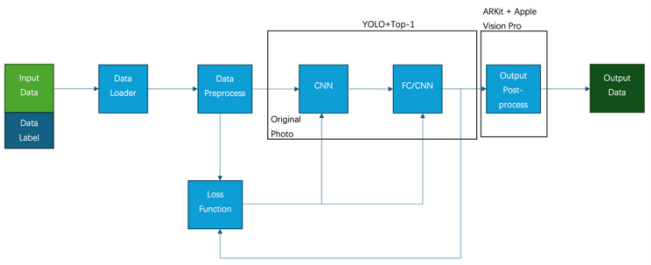

Fig. 2. ARKit-based YOLOv5 algorithm flow chart

IV. MODEL TRAINING

The model training process involves several key steps. First, the dataset is constructed. Next, model training begins, hyperparameters are set, and the training process is initiated. Finally, the model is evaluated to assess its performance. Each of these steps is critical to ensuring that the model achieves optimal accuracy and efficiency.

A. *Dataset Construction*

In the fall detection research based on ARKit and YOLO, building a high-quality and diverse fall image dataset is

crucial for accurate model training. This study collected 8713 fall images from the Internet, covering various types, to improve the generalization and practicality of the model in fall detection.

The dataset covers a variety of scenes from indoor to outdoor. It has a variety of fall types, designed to reflect the complexity of falls in real life. Fig. 3 shows an example from the dataset.

Fig. 3. Fall dataset

We used LabelImg for annotation in the YOLO format, manually drawing bounding boxes around flames and labeling them as class 0 ('falling'), generating corresponding .txt files.

B. *Training Process*

After data screening, cleaning and labeling, the fall detection dataset collected a total of 8713 images, with label 0 representing "falling" and label 1 representing "fall", with a total of 8713 data labeled. The dataset is divided into training set and validation set in a ratio of 8:2. The training set is used for model feature extraction, and the validation set is used to monitor the loss in model training, adjust epochs and hyperparameters.

The hyperparameters of model training are set as follows: the initial learning rate (Lr0) is 0.01, combined with the SGD optimizer to stabilize training; the final OneCycleLR learning rate (lrf) is a fraction of Lr0 to ensure smooth convergence of the model in the later stage. The momentum is set to 0.937 to accelerate training and reduce oscillation, and the weight decay is 0.0005 to prevent overfitting. The warm-up phase lasts for three epochs, and the training is stabilized by gradually increasing the learning rate. The object loss gain (obj) is set to 1.0 to enhance the detection of fall events and improve the recall rate. The IoU threshold (iou_t) is 0.20 to optimize the prediction of the bounding box. The horizontal flip probability (fliplr) is 0.5 to increase the diversity of the data and the generalization ability of the model. The mosaic enhancement probability is 1.0 to ensure that all images are enhanced.

The training process starts with configuring the hyperparameters. We used the YOLOv5 model (yolov5x) and trained it on a custom dataset. The hyperparameters are specified in the data/hyps/hyp.scratch-low.yaml file. The batch size of the training is 64 and the input image size is 640x640. The optimizer uses SGD and the training results are saved in the exp folder under the default directory.

During the forward propagation, the preprocessed image is input to the neural network. The network processes the input through layers such as convolution, activation, pooling, full connection, and batch normalization, and outputs the predicted bounding box, the size and category of the fall posture. The training goal is to minimize the difference between the predicted result and the actual label through loss functions including classification loss (cross entropy) and localization loss (IoU).

After the loss is calculated, backpropagation calculates the gradient of each parameter and guides the model parameter adjustment to reduce the loss. This cycle of forward propagation, loss calculation, backpropagation, and optimization runs through the entire training process until the model is continuously optimized after each batch is completed.

C. *Model Evaluation*

Fig. 4 shows the model evaluation results. The bounding box loss of the training set and validation set gradually decreases, indicating that the accuracy of target localization is continuously improving. Object loss also decreased steadily, indicating that the model performed better at detecting the presence of fall events. The classification loss is close to zero, which is expected since the dataset only has two classes (standing and falling). As training progresses, both precision and recall improve significantly, with precision approaching 0.9 and recall reaching 0.9, indicating that the model's ability to correctly detect fall events is gradually increasing.

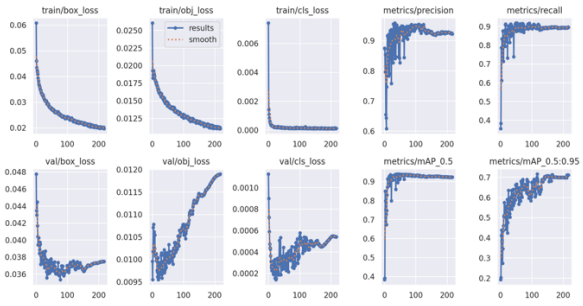

Fig. 4. Model Evaluation

The model achieved a high precision (P) of 0.9 and recall rate (R) of 0.9 on the custom fall detection data set, and was able to effectively detect most fall events with very few false positives or missed detections. This shows that the model has strong confidence in fall recognition and can minimize misses.

In terms of average accuracy (mAP), the model reached 0.9 on mAP50 and 0.7 on mAP50_95, indicating consistent performance under different IoU thresholds. A higher mAP50 indicates excellent detection accuracy under looser thresholds, while mAP50-95 reflects the robustness of the model under more stringent conditions, demonstrating the algorithm's ability to detect fall events at different angles and distances.

The algorithm based on ARKit and YOLOv5 performs well in real-time performance and can efficiently process video data to ensure timely detection of falls and rapid emergency response.

V. RESULTS

In this study, we designed an intelligent fall detection algorithm for elderly people living alone, combining the 3D spatial perception of ARKit and the real-time object detection capabilities of YOLOv5. The algorithm combines ARKit's precise mapping of the physical environment with YOLOv5's fast detection to ensure that fall events can be reliably identified, thereby improving safety and response efficiency.

We conducted two different tests to compare the algorithm. The first test evaluated the accuracy of fall detection without ARKit, while the second test evaluated the accuracy of fall detection with ARKit. As shown in Figure 5, the fall behavior can be detected without ARKit, and the model's confidence in this prediction is 0.83. As shown in Figure 6, there may be a problem of false positives, and the model may misjudge some non-fall postures (such as sitting) as falls. In Figure 7, the image of the detection results shows that the algorithm can effectively detect the fall event and annotate the event with a high confidence of 0.85. Using ARKit enhances the model's ability to detect falls in real-world environments, especially in understanding scene depth and spatial position. In Figure 8, a falls detection box is not assigned just because the person is only half.

Fig. 5. Fall detection results without ARKit

Fig. 6. Fall detection false alarm

Fig. 7. Fall detection results with ARKit

Fig. 8. Fall detection false alarm disappear

Furthermore, we integrated the YOLOv10 Top-1 bounding box selection strategy into the YOLOv5 framework to replace the traditional Non-Maximum Suppression (NMS) process. This optimization not only reduced false detection rates but also improved inference speed. Empirical tests revealed that the average detection time per frame was reduced by 0.1 milliseconds, enhancing the system's real-time performance without compromising accuracy. In Figure 9, it shows the results of using Top-1 and not using Top-1.

Fig. 9. Result for using Top-1 or not

VI. Conclusion

This paper focuses on the design and evaluation of a fall detection algorithm for elderly individuals living alone, leveraging ARKit, YOLOv5, and Apple Vision Pro. Through in-depth analysis of fall-related risks and practical challenges faced by this demographic, the proposed system demonstrates significant potential and advantages in real-time fall detection. By combining YOLOv5's object detection capabilities with ARKit's spatial perception and Apple Vision Pro's immersive and real-time environmental sensing, the algorithm achieves accurate and comprehensive identification of fall events.

Comparative experiments against conventional fall detection approaches confirm that the proposed method outperforms in both detection accuracy and real-time responsiveness. In particular, the integration of the YOLOv10 Top-1 bounding box selection strategy into the YOLOv5 detection pipeline has enhanced the detection precision while improving inference efficiency. This strategy replaces the traditional Non-Maximum Suppression (NMS) and reduces average detection latency by 0.1 milliseconds per frame, contributing to improved system performance in time-sensitive monitoring applications.

Despite these promising outcomes, several limitations remain. In real-world deployments, the algorithm's robustness and stability must be further strengthened to handle continuous, large-scale monitoring scenarios. Performance under varied lighting, occlusion, and posture conditions remains a challenge. The system's scalability and adaptability across diverse elderly populations and environmental settings also require further investigation.

Future work will focus on optimizing the model architecture and hyperparameters to improve detection efficiency and robustness. Efforts will also be directed toward enhancing adaptability to complex indoor scenes and postural variations, leveraging the spatial context awareness from ARKit and Vision Pro. Additionally, integrating the system with wearable devices or smart home infrastructure may further improve real-time feedback and user experience. Large-scale field testing and iterative user feedback collection will be essential for validating the algorithm's reliability and practical utility in daily living environments.

Acknowledgment

This research was conducted as part of the Computer Science CS 5588 Data Science Capstone course under the guidance of Yugyung Lee, Ph.D., Professor. Additionally, it

was supported by the technical resources by Sanda University and University of Missouri-Kansas City.

REFERENCES

[1] H. Le, M. Nguyen, W. Yan, and H. H. Nguyen, "Augmented Reality and Machine Learning Incorporation Using YOLOv3 and ARKit," Applied Sciences, vol. 11, no. 13, p. 6006, 2021.

[2] A. Gupta, R. Srivastava, H. Gupta, and B. Kumar, "IoT based fall detection monitoring and alarm system for elderly," in 2020 IEEE 7th Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering (UPCON), 2020, pp. 1-5.

[3] K. M. Sharook, A. Premkumar, R. Aishwaryaa, J. M. Amrutha, and L. R. Deepthi, "IFall Detection Using Transformer Model," Lecture Notes in Networks and Systems, pp. 58-66, 2022.

[4] Y. Yin, L. Lei, M. Liang, X. Li, Y. He, and L. Qin, "Research on Fall Detection Algorithm for the Elderly Living Alone Based on YOLO," 2021 IEEE International Conference on Emergency Science and Information Technology (ICESIT), vol. 12, no. 7, pp. 1515, Dec. 2021.

[5] A. Gupta, R. Srivastava, H. Gupta, and B. Kumar, "IoT based fall detection monitoring and alarm system for elderly," in 2020 IEEE 7th Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering (UPCON), 2020, pp. 1-5.

[6] G. Sun, S. Wang, and J. Xie, "An Image Object Detection Model Based on Mixed Attention Mechanism Optimized YOLOv5," Electronics, vol. 2023, pp. 403-408, 2023.

[7] Y. Sun, J. Song, Y. Li, Y. Li, S. Li, and Z. Duan, "IVP-YOLOv5: An intelligent vehicle-pedestrian detection method based on YOLOv5s," Connection Science, vol. 35, no. 1, p. 2168254, 2023.

[8] R. Y. Lee, H. Kamaludin, N. Z. M. Safar, N. Wahid, N. Abdullah, and D. Meidelfi, "Intelligence Eye for Blinds and Visually Impaired by Using Region-Based Convolutional Neural Network (R-CNN)," JOIV: International Journal on Informatics Visualization, vol. 5, no. 4, pp. 409-409, 2021.