World Journal of Information Technology
Print ISSN: 2959-9903
Online ISSN: 2959-9911
DOI: https://doi.org/10.61784/wjit3122

MULTI-MODAL LIGHTWEIGHT ELDERLY FALL RECOGNITION ALGORITHM FOR EMBEDDED EDGE HOME MONITORING DEVICES

XiaoYing Ran*, YouYan Dong, ZiHang Zhang
School of Automation and Electrical Engineering, Tianjin University of Technology and Education, Tianjin 300222, China.
Corresponding Author: XiaoYing Ran

Abstract: Machine vision-based elderly fall monitoring has become a mainstream solution for home elderly care, yet existing single-camera vision schemes suffer from low recognition accuracy for hunched and slow-moving seniors, frequent false alarms caused by illumination changes and furniture occlusion, and heavy computation load that fails real-time inference on low-cost embedded hardware such as Raspberry Pi. This paper develops a multimodal lightweight fall recognition framework based on MediaPipe Pose. A self-built dataset containing real home actions of elderly people is constructed, and transfer learning is adopted to fine-tune the pose model for special aged postures. Ultrasonic ranging, offline voice recognition and gas detection sensors are integrated to build weighted multi-modal joint discrimination logic. Dynamic frame extraction and dual-chip collaborative computing scheduling are designed to cut hardware computation overhead. The whole algorithm completes all inference locally without uploading video to cloud servers, effectively protecting household privacy. Experimental results verify that the proposed method suppresses false alarms significantly and supports stable real-time operation on low-cost edge terminals. It provides a feasible lightweight optimization scheme for privacy-sensitive home elderly monitoring systems.
Keywords: MediaPipe pose; Fall detection; Multi-modal data fusion; Edge computing; Lightweight pose estimation

1 INTRODUCTION

With the continuous deepening of global population aging, accidental falls have become a critical hidden danger threatening the life safety of empty-nest elderly residents. Vision-based fall monitoring technology has gradually become the mainstream solution for home elderly care, yet existing practical deployment schemes have obvious drawbacks. Wearable sensing devices suffer from low acceptance among the elderly, while cloud video monitoring faces hidden dangers of privacy leakage and service failure under offline conditions [1]. Most public human pose datasets are dominated by young and middle-aged samples; directly deploying models trained on such datasets to home monitoring equipment leads to poor recognition performance targeting aged groups [2]. Traditional feature extraction algorithms such as HOG+SVM rely on manual feature design and lack robustness under complex indoor lighting and occlusion interference, while complete networks like OpenPose contain massive parameters and cannot realize real-time reasoning on low-cost embedded boards [3]. Although MediaPipe Pose achieves a balance between detection speed and accuracy, it lacks targeted optimization for hunched, stooped and crutch-assisted postures of the elderly, and single visual judgment triggers frequent false alarms in daily household scenarios [4]. Therefore, developing a lightweight multi-modal fall detection algorithm supporting local edge computing is of great practical significance for promoting the large-scale application of home intelligent elderly monitoring equipment.

A large number of studies have explored lightweight visual detection and multi-sensor fusion monitoring methods in recent years. In terms of human target pre-screening, MobileNet SSD with depthwise separable convolution is widely adopted to filter irrelevant background areas and reduce invalid frame computing overhead, which effectively cuts down the idle power consumption of embedded devices [5]. For human skeleton extraction, MediaPipe Pose is extensively used to output 3D joint coordinates of the human body, but most pre-trained models ignore the unique body characteristics of the elderly and produce serious skeleton keypoint offset during low-stance movements [6]. On multi-modal collaborative discrimination, existing research mostly combines cameras with single inertial or ranging sensors to suppress false alarms, while few studies integrate ultrasonic ranging, offline voice recognition and gas detection sensors to build multi-dimensional cross-verification logic for comprehensive home risk assessment [7]. In terms of embedded deployment optimization, current lightweight strategies mainly focus on network layer pruning and quantization, without designing dual-chip collaborative task scheduling and dynamic frame extraction mechanisms to balance real-time performance and long-term standby power consumption [8]. In addition, most training datasets lack sufficient samples of elderly-specific postures, resulting in insufficient generalization ability of models in real home environments [9-10].

To fill the above research gaps, this paper proposes a fully locally operated multi-modal lightweight fall recognition algorithm oriented to embedded edge terminals, and its main marginal contributions are summarized as follows. First, a self-built home elderly action dataset covering various lighting and occlusion scenes is constructed, and transfer learning is applied to adjust pose discrimination thresholds adaptively for special aged postures such as hunchback and crutch walking. Second, a weighted multi-modal fusion discrimination framework integrating visual, ultrasonic, voice and gas sensing data is established to distinguish real falls from normal daily behaviors including bending over and squatting, which significantly reduces the false alarm rate of single vision systems. Third, a collaborative computing scheduling scheme based on Raspberry Pi CM4 and STM32 dual chips is designed, together with dynamic frame extraction and process priority control strategies to realize low-power real-time local reasoning without uploading any video data to cloud servers. The proposed algorithm provides a low-cost, privacy-preserving technical reference for edge intelligent home elderly fall monitoring systems.

2 THE OVERALL FOUR-LAYER EXECUTION FLOW OF THE ALGORITHM

The entire algorithm is developed using Python and the OpenCV library, optimized for the Raspberry Pi CM4 embedded Linux operating system, and employs a four-tier pipeline processing architecture. All data collected by cameras and sensors is timestamped to mitigate detection errors caused by asynchronous data acquisition from multiple sources. The algorithm's output—human body coordinates and fall risk scores—is transmitted via TCP packets over the local network to the underlying microcontroller, where it serves solely for local audiovisual alerts and continuous human target tracking.

2.1 Multi-Source Perception Collection Layer

The system synchronously collects real-time data from multiple peripherals, including 4K wide-angle camera video streams, ground elevation measurements from ultrasonic radar, gas concentration readings from MQ2 sensors, and wake -up signals along with keyword recognition results from the ASR PRO voice module. Given the varying data formats output by different sensors, the program first applies noise reduction and normalization processing to the raw values, then packages them uniformly before feeding them to the subsequent feature inference module. Detailed specifications of the relevant sensors are listed in Table 1:

Table 1 Sensors at Data Acquisition Layer

Sensor Equipment	Raw Output Data	Preprocessing Operations	Normalized Output Format
4K Camera	RGB Image	Size scaling, dynamic frame sampling, grayscale denoising	640 × 480 image matrix
Ultrasonic Radar	Voltage Value	Sliding mean filtering	Height of human body from ground
MQ2 Gas Sensor	Voltage Signal	Concentration calibration	Gas risk marker
Voice Module	Real-time audio stream	Environmental noise reduction, keyword matching	Floating-point confidence value of distress call

2.2 Feature Inference Layer

A two-stage filtering approach is employed to reduce computation on invalid images: the first stage utilizes the lightweight MobileNet SSD network to filter out irrelevant backgrounds such as tables, chairs, walls, and household appliances, thereby delineating the valid human region within the frame; the second stage employs the optimized MediaPipe Pose network to extract 33 key skeletal points of the human body, output pixel coordinates for each joint across consecutive frames, and simultaneously render the human skeleton for local debugging and visualization.

2.3 Fused Discrimination Layer

Specific posture thresholds are established based on the physiological characteristics of elderly individuals to preliminarily assess potential fall risks. Subsequently, a secondary weighted validation is performed using three types of sensor data—radar ranging, voice signals, and gas environment measurements—to differentiate between non-hazardous daily activities such as bending down to pick up objects, squatting, or sitting quietly for rest, ultimately generating a graded fall risk classification.

2.4 Simulation Output Layer

The program dynamically adjusts the image inference frame rate based on human presence in the frame, generating two key outputs: a fall detection alert to trigger local audio-visual alarms, and horizontal displacement coordinates of the human subject for continuous position tracking. Additionally, it optimizes CPU frequency and camera sampling intervals according to device status to reduce overall standby power consumption.

The hardware architecture employs a dual-chip architecture with specialized functions: the Raspberry Pi CM4 handles computationally intensive tasks such as image inference and speech analysis, while the STM32 microcontroller processes raw data from temperature/humidity sensors, radar, and gas sensors through simple preprocessing. The two chips exchange data periodically via TCP short messages, distributing computational load across both units to ensure smooth multi-task performance without latency.

3 DESIGN AND PARAMETER TUNING PROCESS FOR THE CORE VISUAL ALGORITHM

3.1 MobileNet SSD Pre-Processing Human Body Detection Network

To minimize unnecessary computations when no human subjects are present in the image, this study employs MobileNet SSD as the primary human detection network. Leveraging deeply separable convolutional architectures, it significantly reduces both network parameters and computational requirements. The multi-scale detection framework accommodates human targets at varying distances within home environments. Through extensive testing across multiple indoor scenarios, an optimal confidence threshold for human detection was established: candidate boxes below this threshold are discarded, and a non-maximum suppression algorithm is applied to eliminate redundant detections.
Experimental comparisons demonstrate that introducing a pre-screening module significantly reduces the CPU resource consumption during periods without human activity, effectively lowering long-term power usage.

3.2 MediaPipe Key Point Extraction and Posture Adjustment for Elderly Users

Testing with the official pre-trained MediaPipe model revealed that its detection logic, designed for upright young adults, frequently suffers from keypoint displacement and motion misidentification when dealing with postures common among older adults—such as kyphosis, knee flexion, or using a cane. To address this, the system incorporated a pose correction module after extracting joint coordinates from 33 individuals, dynamically adjusting fall detection weights based on the vertical height difference between the shoulders and hips, thereby reducing the reliance on fixed parameters inherent in the original model and better accommodating the unique physical characteristics of elderly users.
The native MediaPipe Pose model incurs high per-frame inference costs, making it inadequate for indoor real-time monitoring applications. This paper proposes a dynamic frame extraction strategy: reducing image detection frequency when no human subjects are present; performing inter-frame inference during normal walking or sitting states of elderly individuals; and switching to real-time per-frame computation upon detecting significant torso tilt. Experimental results on embedded hardware demonstrate that the optimized approach significantly reduces per-frame inference overhead while balancing algorithmic real-time performance with hardware resource consumption. The key point visualization comparison is shown in Figure 1:

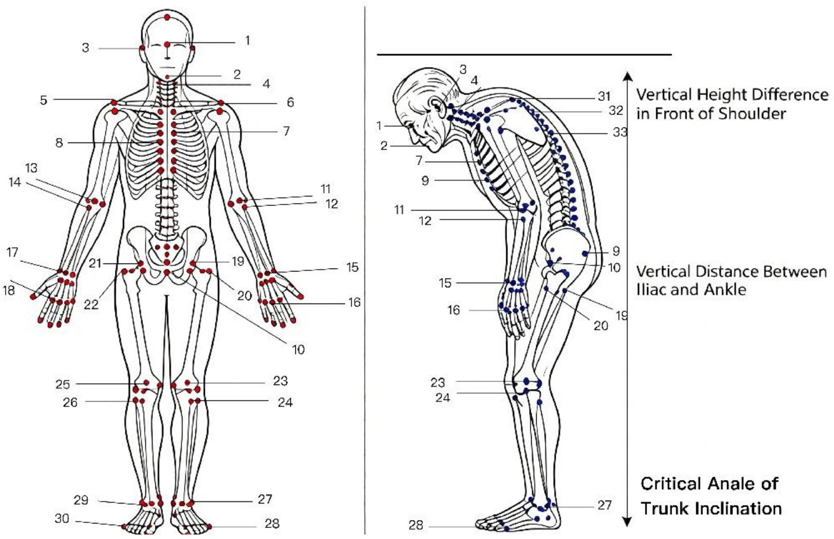

Figure 1 Media Pipe Pose: Visual Comparison of 33 Key Body Points

3.3 Self-Developed Home-Based Elderly Data Set and Parameter Tuning for Transfer Learning

Using the home simulation environment established by the laboratory, multiple sets of human motion sequence images were collected on-site, covering three high-frequency activity areas for elderly individuals—living room, kitchen, and bathroom—and encompassing four typical indoor lighting conditions: strong light, soft light, twilight backlight, and partial furniture obstruction. Manual annotations categorized movements into five types: normal walking, prolonged sitting with bending, squatting to pick up objects, side fall, and frontal fall. The dataset was specifically enriched with extensive real-world footage of elderly individuals aged 60 and above exhibiting kyphosis or using walking aids, addressing the insufficient representation of elderly posture samples in publicly available datasets.
During the model training phase, the core feature extraction network is frozen, with optimization focused solely on the backend behavior classification head. The loss function combines cross-entropy loss and temporal smoothing loss to mitigate false alarms caused by transient fluctuations in individual frames. The Adam optimizer is employed with a properly set initial learning rate, and an early stopping strategy is implemented after multiple iterations to prevent overfitting. Upon completion of training, redundant layers are removed to optimize the model for deployment on ARM-based embedded devices.

The suspected fall determination requires simultaneous fulfillment of multiple temporal feature conditions: a vertical height difference between the hip and ankle below a set threshold, a trunk tilt angle exceeding a critical value for a sustained period, and a body height above ground below a safety threshold. Upon meeting these criteria, the system proceeds to a multi-sensor secondary verification process to further filter out false triggers. To address the temporal variation patterns of human posture, a temporal similarity analysis approach is introduced to enhance classification accuracy.

The sample distribution table for data collection is shown in Table 2:

Table 2 Data Set Sample Distribution Table

Action Category	Total Samples	Collection Scenarios	Lighting Conditions	Proportion of Special Posture Samples
Walking	420	Living room, kitchen, corridor	Bright, soft, backlight	Hunchback, crutch-assisted walking
Squatting	280	Floors of living room and kitchen	Indoor natural light	Low knee-bending posture
Static Sitting	200	Sofas and chairs	Dim light, backlight	Static relaxed posture
Side Fall	150	Bathroom, corridor	Household occlusion, backlight	Torso with large tilt angle
Collapse to Ground	150	Living room floor	Indoor natural light	Body extremely close to ground

Relying solely on intermittent human detection cannot achieve continuous target tracking. This paper introduces the CSRT visual tracking algorithm to enhance the continuous monitoring mechanism. Upon initial detection of a human subject via MobileNet SSD, the system extracts texture and color features to generate a dedicated tracking template; subsequent image frames no longer require full-body detection. If the target is lost due to obstruction by furniture or walls, the system automatically restarts MobileNet SSD for reidentification, leveraging research principles from multi-target visual tracking to ensure uninterrupted home monitoring.

4 JUDGMENT LOGIC AND EMBEDDED COMMUNICATIONS

4.1 Design of Multimodal Fusion Discrimination Logic

Visual images alone are prone to false alarms due to interference from lighting or clothing. The system integrates data from three sensor types—range radar, voice recognition, and gas detection—to establish a dual-layer verification mechanism, employing a weighted voting approach for comprehensive risk assessment.

1) Radar distance verification: If a fall is visually detected but the radar measures the person's height above ground as greater than 50 cm, the alarm is immediately suppressed to distinguish between bending down to pick up items and actual falling; 2) Environmental risk linkage: When MQ2 detects excessive gas levels, an environmental alarm is triggered independently even if no fall has been identified; simultaneous occurrence of both fall and gas leak raises the alarm priority; 3) Voice-assisted judgment: The voice module increases the likelihood of visual recognition of a suspected fall when prolonged moaning or distress calls are detected; if no voice signal is received for more than 3 minutes after a fall, the highest-level emergency alarm is activated.

The weight distribution was finalized through multiple rounds of debugging: visual features account for 0.6, radar data for 0.2, while speech and gas detection each contribute 0.1; only when the composite score meets the threshold does a valid warning trigger, significantly reducing pure visual false alarm rates. The multimodal data weighting architecture is illustrated in Figure 2.

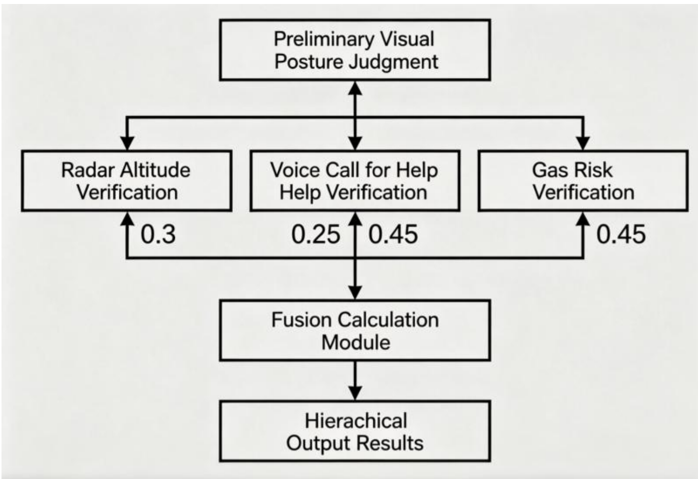

Figure 2 Multimodal Data Weighting Architecture

4.2 Dual-Chip TCP Data Interaction Solution

The hardware employs a collaborative architecture featuring a Raspberry Pi CM4 and an STM32 dual-controller unit: the Raspberry Pi handles AI image inference tasks, while the STM32 reads raw data from various sensors. Data exchange between the two chips is facilitated via the TCP protocol over a local network. The STM32 periodically uploads sensor floating-point data, with each message kept concise; the Raspberry Pi transmits human motion tracking coordinates and risk alert indicators. The communication module incorporates built-in heartbeat packets and automatic reconnection logic, enabling rapid recovery after interruptions to ensure uninterrupted algorithm execution.

4.3 Optimization of Process Scheduling in Embedded Systems

In Linux embedded systems, background processes are prioritized as follows: the image inference process is assigned medium priority, while log recording and background communication processes receive low priority. When no human presence is detected in the frame for an extended period, the CPU frequency is automatically reduced and camera sampling is temporarily disabled to minimize standby power consumption. A dedicated daemon thread has been added to ensure rapid automatic program restart upon abnormal termination of the image inference process, enabling round-the-clock device monitoring.

5 CONCLUSIONS

This paper designs a multimodal lightweight fall recognition algorithm suitable for low-cost embedded edge monitoring terminals for the elderly. The MobileNet SSD lightweight network is adopted to pre-filter human targets, and MediaPipe Pose is improved by transfer learning on a self-built elderly home action dataset to eliminate misjudgment caused by hunched, crutch-walking and other unique aged postures. Combined with the CSRT tracking algorithm, continuous target locking under furniture occlusion is realized. A weighted multi-modal joint discrimination mechanism integrating ultrasonic ranging, offline voice recognition and gas sensing data is constructed to distinguish real falls from daily bending and squatting behaviors, greatly lowering the false alarm rate of single visual detection. Meanwhile, a dual-chip collaborative scheduling scheme based on Raspberry Pi CM4 and STM32 is matched with dynamic frame extraction and embedded process priority optimization strategies to cut the long-term standby power consumption of equipment. All model reasoning and data processing tasks run locally on the terminal without uploading surveillance video to cloud servers, which effectively avoids household privacy leakage risks and meets the real-time monitoring demand of home pension scenarios. The overall system has low hardware cost and simple deployment conditions, possessing strong practical promotion value in community elderly care and family empty-nest elderly monitoring scenarios.

In future research, further lightweight compression of the pose estimation network can be carried out to reduce the memory occupation of embedded devices. Time-series neural networks will be introduced to mine long-term continuous motion characteristics of the elderly and improve the accuracy of early fall risk prediction. In addition, vital sign monitoring sensors such as heart rate and respiration can be added to expand multi-modal sensing dimensions and build a more comprehensive elderly safety early warning system. The model generalization ability will also be verified through more real home scene tests to adapt to complex lighting and occlusion environments in different families.

COMPETING INTERESTS

The authors have no relevant financial or non-financial interests to disclose.

FUNDING

The authors gratefully acknowledge the financial support from National Undergraduate Innovation and Entrepreneurship Training Program Project for 2025 (202510066040) funds.

REFERENCES

[1] Goyal N, Singh T, Goraya S M. Explainable Artificial Intelligence with Deep Convolutional Neural Networks for Real-Time Image-Based Yoga Posture Recognition in Fitness Training. New Generation Computing, 2025, 44(1): 5.
[2] Peng Wei, Deng Qiang, Li Ming, et al. Adaptive Multi-Lens Phase Modulation for Scale-Aware Privacy-Preserving Human Pose Recognition. IET Software, 2025, 2025(1): 7879383.
[3] Song Ze, Li Gang, Song Ming. Adaptive Attention Fusion Method for Fine-Grained Human Posture Recognition Based on Micro-Doppler. Journal of Physics: Conference Series, 2025, 3166(1): 012010.
[4] Junaid I M, Madarapu S, Ari S. Human gait recognition using attention based convolutional network with sequential learning. Signal, Image and Video Processing, 2024, 19(1): 157.
[5] Ali N, Alec J, T L S, et al. Real world validation of activity recognition algorithm and development of novel behavioral biomarkers of falls in aged control and movement disorder patients. Frontiers in Aging Neuroscience, 2023, 15: 1117802.

[6] Liu Ziyang, Li Wei, Xing Gang, et al. The Identification of Elderly People with High Fall Risk Using Machine Learning Algorithms. Healthcare, 2022, 11(1): 47.

[7] A A M A, M A H, Abdelghani D, et al. The Applications of Metaheuristics for Human Activity Recognition and Fall Detection Using Wearable Sensors: A Comprehensive Analysis. Biosensors, 2022, 12(10): 821.

[8] Amandine D, Titus B, JeanPierre B. Identifying Fall Risk Predictors by Monitoring Daily Activities at Home Using a Depth Sensor Coupled to Machine Learning Algorithms. Sensors, 2021, 21(6): 1957.

[9] Huang Kun, Yuan Qiongqian, Huang Zefan. A Two-Stage Fall Recognition Algorithm Based on Human Posture Features. Sensors (Basel, Switzerland), 2020, 20(23).

[10] Ye Cheng, Li Jie, Hao Sen, et al. Identification of elders at higher risk for fall with statewide electronic health records and a machine learning algorithm. International Journal of Medical Informatics, 2020, 137: 104105.