A Skeleton-Based Deep Learning Framework for Gait Analysis and Fall Detection Using MediaPipe Pose

Pu-Sheng Tsai[1], Ter-Feng Wu[2]*

[1] Department of Electrical Engineering, Ming Chuan University, Taiwan
[2] Department of Electrical Engineering, National Ilan University, Taiwan
pusheng@mail.mcu.edu.tw, tfwu@niu.edu.tw

Abstract

This study proposes a skeleton-based human activity recognition framework that integrates MediaPipe Pose–based feature extraction with deep learning for intelligent healthcare applications, including gait analysis and fall detection. Skeletal time-series data from 33 keypoints are extracted from video streams and transformed into angle-based temporal features to characterize human motion patterns. Experimental results demonstrate that the proposed method effectively captures gait characteristics and achieves stable performance in both binary and multi-class fall detection tasks. In addition, an Internet of Things (IoT) architecture based on Raspberry Pi and ESP32 is developed to enable real-time fall detection and remote monitoring. The proposed system exhibits high computational efficiency and strong integration capability, making it suitable for smart home care and real-time safety monitoring applications.

Keywords: Skeleton-based human activity recognition, Convolutional Neural Network (CNN), Fall detection

1 Introduction

With the rapid advancement of artificial intelligence (AI) technologies, skeleton-based human motion analysis has emerged as a key research direction in intelligent perception systems. Compared with conventional vision-based approaches that rely on appearance cues, color information, or depth images, skeleton representations provide a highly structured and compact description of human motion while being inherently robust to illumination changes, background clutter, and environmental noise. These properties make skeleton-based methods particularly suitable for applications such as human activity recognition, gait-based identification, and health monitoring in real-world environments. Recent progress in pose estimation has enabled efficient extraction of human skeletal data from monocular RGB cameras. In particular, the MediaPipe Pose framework offers real-time detection of 33 human body keypoints using a lightweight pretrained model, allowing detailed motion analysis without specialized sensing hardware. Leveraging this capability, skeleton-based representations can serve as a unified intermediate feature space that supports multiple behavior analysis tasks within a single system architecture. In this study, gait-based identity recognition and fall detection are selected as two representative applications to demonstrate the versatility of the proposed skeleton analysis framework.

In terms of feature modeling, this study constructs motion representations based on skeletal keypoint coordinates and further derives joint angle sequences and their temporal variations as discriminative time-series features. Although joint angle dynamics are inherently one-dimensional temporal signals, the proposed approach organizes multiple joint-angle sequences into a structured two-dimensional representation by stacking temporal variations of selected joints in a unified and ordered manner. This image-like encoding enables explicit modeling of both inter-joint coordination and temporal motion evolution within a single representation. Through this temporal-to-spatial feature transformation, convolutional neural networks (CNNs) can effectively exploit their capability in capturing local spatial correlations, allowing the model to jointly learn cross-joint relationships and time-dependent motion patterns. Based on this representation, a lightweight deep learning–based classification architecture is designed to achieve a balance between recognition accuracy and real-time computational efficiency, enabling reliable performance even under limited training data and embedded computing constraints. Furthermore, to enhance practical applicability in real-world smart home care scenarios, an Internet of Things (IoT)–enabled feedback mechanism is integrated into the proposed system to support real-time fall event notification and remote monitoring. Overall, this study presents a skeleton-centric and task-general human behavior analysis framework that not only effectively supports gait-based identity recognition and fall detection but also demonstrates strong scalability and cross-task adaptability, serving as a foundational architecture for future multimodal human–machine interaction and behavior analysis research.

The main contributions of this study can be summarized as follows. First, a skeleton-based feature modeling strategy is developed to represent human motion using joint coordinates, angular dynamics, and temporal motion descriptors. Second, a unified and lightweight

*Corresponding Author: Ter-Feng Wu; Email: tfwu@niu.edu.tw
DOI: https://doi.org/10.70003/160792642026052703006

CNN-based classification framework is designed to accommodate multiple skeleton-driven recognition tasks through an efficient feature encoding scheme. Third, gait-based identity recognition and fall detection are investigated as representative application scenarios to validate the generality and effectiveness of the proposed framework. Finally, an IoT-enabled feedback mechanism is integrated to support real-time event reporting and remote monitoring, enabling practical deployment in intelligent living environments. Overall, this work highlights the potential of skeleton-based deep learning for intelligent living and safety monitoring applications and presents a scalable and task-adaptive framework that can serve as a foundation for future multimodal behavior analysis research.

2 Related Work

2.1 Skeleton-Based Human Motion Analysis

Early studies on skeleton-based human activity recognition primarily relied on wearable sensors or depth-sensing devices for data acquisition. With advances in computer vision, these approaches have gradually evolved toward image-based hu-man skeleton estimation, reflecting a clear shift in technological paradigms. Initial research commonly employed inertial sensors, accelerometers, or depth cameras such as Microsoft Kinect to capture three-dimensional joint positions, which were then used to infer human posture and motion characteristics. Representative work by Shotton et al. [1] proposed a pixel-wise body part classification approach, in which each pixel in a depth image is assigned to a specific body part, enabling rapid reconstruction of a three-dimensional skeletal structure from a single depth frame. This method demonstrated strong generalization capability and robustness to variations in pose, body shape, and clothing. Ye et al. [2] further improved 3D pose estimation accuracy by integrating an initial pose detection stage with a subsequent pose refinement mechanism. In addition, Li et al. [3] developed a non-contact real-time gesture recognition system based on Kinect, achieving satisfactory recognition performance under multi-gesture scenarios. Despite their effectiveness, depth-sensing devices are constrained by hard-ware cost, sensitivity to lighting conditions, and limited deployment flexibility, which restrict their applicability in practical and large-scale environments. As a result, recent research has increasingly shifted toward skeleton estimation and behavior analysis methods that rely solely on RGB images, aiming to enhance system scalability and feasibility for real-world intelligent applications.

In terms of skeleton feature modeling, early studies commonly combined traditional image processing techniques with statistical analysis methods. Human appearance or silhouette features were first extracted and then transformed using dimensionality reduction techniques such as Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) to reduce feature redundancy and improve classification performance. Huang et al. [4] utilized appearance-based human silhouettes extracted from images and employed a parametric canonical space for feature transformation and classification, demonstrating the feasibility of statistical feature modeling approaches in early vision-based gait analysis. In addition, several studies have explored vision-based fall detection approaches using human pose or structured human representations, such as depth images, human silhouettes, or structured posture features, in combination with machine learning or deep learning models for classification. Experimental results from these studies have demonstrated that vision-based structured human representations are capable of effectively capturing human motion variations and exhibit promising recognition performance and practical feasibility in fall detection tasks [5-6]. However, silhouette- and appearance-based representations are inherently sensitive to illumination variations, clothing changes, and occlusions, which often lead to incomplete semantic descriptions of human motion and limit their generalization capability in complex environments.

With the advancement of pose estimation techniques, research focus has gradually shifted toward skeleton-based models that emphasize joint topology and structural relationships. Skeleton representations describe human posture and motion using key joints and their connectivity, enabling more direct modeling of intrinsic joint dynamics while effectively suppressing background and appearance related interference. As a result, skeleton-based approaches have become a dominant paradigm in recent studies on human behavior analysis and gait recognition. Among commonly adopted skeleton extraction models, OpenPose and MediaPipe Pose are two representative frameworks. OpenPose [7] employs Part Affinity Fields (PAFs) to model the associations between body parts, allowing simultaneous keypoint localization and per-son-wise joint grouping in multi-person scenarios. It can estimate 18 two-dimensional joint positions from a single RGB image with high accuracy; however, its relatively high computational complexity limits real-time deployment on embedded or mobile platforms. In contrast, MediaPipe Pose is built upon the BlazePose architecture and emphasizes lightweight design and real-time inference capability. It outputs 33 human keypoints and incorporates structural inference mechanisms to recover missing joint information under partial occlusions, making it more suitable for cross-scene applications and mobile or edge-device deployment. In recent years, an increasing number of studies have applied pose estimation models to fall detection, gait analysis, and human action recognition, and further combined skeleton sequences with recurrent or convolutional temporal models for behavior modeling. Among these approaches, OpenPose and MediaPipe Pose have become widely adopted in vision-based skeleton analysis, as they can directly extract human keypoints from RGB images. For example, Salimi et al. [8] proposed a fall detection framework based on OpenPose and recurrent neural networks, in which temporal modeling of skeleton sequences was employed. Their experimental results demonstrated that the proposed method achieved a recognition accuracy of over 98% in practical fall detection tasks. Other studies have employed OpenPose-derived pose features in combination with one-

dimensional convolutional neural networks (1D-CNNs) to enable efficient and real-time fall classification [9]. These results collectively indicate that skeleton sequence modeling provides robust and reliable performance across multiple human behavior analysis tasks.

2.2 Gait Recognition Using Skeleton Analysis

In identity recognition applications, biometric modalities such as fingerprint, facial, and iris recognition have reached a high level of maturity and have been widely deployed in various authentication systems. In contrast, gait recognition has attracted increasing research attention due to its ability to perform identity verification at a distance, without physical contact, and with a relatively high resistance to spoofing. Unlike fingerprint-, face-, or iris-based approaches that require close-range acquisition, gait recognition enables identity verification under natural walking conditions, making it particularly suitable for intelligent surveillance and public-space monitoring scenarios. Early gait analysis methods primarily relied on human appearance or silhouette-based features, such as the Gait Energy Image (GEI). However, these approaches are highly sensitive to illumination variations, clothing changes, and background complexity, which limits their robustness across different environments. To improve cross-environment recognition stability, recent studies have gradually shifted toward skeleton-based gait analysis methods that emphasize structural motion representations. Wang et al. [10] constructed a multi-view gait database containing both three-dimensional skeletal data and two-dimensional silhouette information using Kinect V2, and proposed a matching-level fusion strategy to extract view-invariant static and dynamic features. Their method achieved recognition accuracies ranging from 90% to 94% under multi-view conditions, demonstrating the robustness of skeleton-based features against viewpoint variations.

With the rapid development of deep learning techniques, skeleton-based gait analysis has gradually shifted from static geometric feature representations toward temporal dynamic modeling. Numerous studies have demonstrated that recurrent neural networks are effective in capturing time-dependent variations in gait patterns. For instance, Zhang et al. [11] employed Long Short-Term Memory (LSTM) networks to analyze skeleton-based temporal features and achieved satisfactory recognition performance. Du et al. [12] proposed a hierarchical recurrent neural network (RNN) architecture to model temporal dependencies among skeletal joints. In addition, Wang [13] utilized Gated Recurrent Units (GRUs) to model continuous pose transitions in skeleton-based motion sequences, effectively improving temporal modeling capability and recognition stability in action recognition tasks. In parallel, Graph Convolutional Networks (GCNs) have been widely adopted for skeleton sequence analysis. These approaches represent the human skeleton as a spatio-temporal graph structure, enabling simultaneous modeling of spatial joint topology and temporal motion dynamics, and thus facilitating effective learning of complex human movement patterns. Representative studies include the Spatial–Temporal Graph Convolutional Network (ST-GCN) proposed by Yan et al. [14], as well as the adaptive graph convolution framework introduced by Shi et al. [15]. Both approaches have demonstrated strong generalization capability under cross-scene and cross-subject conditions, highlighting the effectiveness of skeleton-based GCN methods for gait and action analysis.

Based on the aforementioned research developments, skeleton-based gait analysis has evolved from early sensor-driven and statistical feature modeling approaches into a mature paradigm that leverages image-based skeleton estimation combined with deep learning–based temporal modeling. Following this research trend, the present study adopts MediaPipe Pose as a unified skeleton data source, enabling extraction of structured multi-joint coordinates of 33 human keypoints using only a standard RGB camera. By integrating limb-torso joint angles, relative joint displacements, and temporal sequence features, a convolutional neural network-based classifier is employed to construct a robust and accurate skeleton-based gait identity recognition module. This module serves as a representative validation case for the proposed framework under multi-task and cross-application scenarios.

2.3 Fall Detection Using Skeleton Analysis

Falls are among the most common and hazardous incidents affecting older adults and may result from factors such as hypotension, medication side effects, declining balance ability, and mobility deterioration. Consequently, the ability to accurately and promptly detect fall events in daily living environments while minimizing false alarms has long been a critical research topic in intelligent healthcare and human behavior monitoring. Early fall detection approaches primarily relied on static image analysis to determine whether a subject exhibited a fallen posture. However, due to the absence of temporal information, such methods are prone to misclassifying daily activities such as squatting, bending, or object picking as fall events, thereby limiting their practical applicability. To address this issue, Vallabh et al. [16] incorporated squatting motions into the classification model and conducted fall detection experiments using the MobiFall dataset, which includes both fall events and activities of daily living (ADL). By applying a feature selection strategy to reduce data dimensionality and comparing multiple classifiers, their results showed that the k-nearest neighbors (kNN) approach achieved the best performance, with an accuracy of 87.5%, demonstrating that appropriate feature modeling can effectively enhance fall detection performance. In addition to vision-based approaches, Nadee et al. [17] proposed a fall detection method based on ultrasonic sensors, in which sensor arrays were installed on ceilings and walls, and the acquired signals were transmitted to a computer via an Arduino platform using Wi-Fi communication. By analyzing differences between vertical and lateral signal patterns, the system was able to distinguish between standing, sitting, and falling behaviors, while also differentiating human motion from non-human objects. The reported fall detection accuracy reached 92%. Nevertheless, such

ultrasonic sensor–based approaches remain constrained by sensor deployment requirements and environmental conditions, which may limit their applicability in general residential or open-space environments.

With the emergence of deep learning techniques, fall detection research has gradually shifted from static posture judgment toward analysis approaches centered on dynamic posture and temporal information. Skeleton-based models, such as OpenPose and MediaPipe, have become an important foundation in fall detection research because they provide structured pose information in the form of human keypoints. Many studies further combine these skeletal representations with temporal models, including recurrent neural networks (RNNs), long short-term memory (LSTM) networks, or gated recurrent units (GRUs), to classify the time-series data formed by skeletal keypoints for fall event recognition. Several studies have employed OpenPose- or MediaPipe-based pose skeleton models to construct fall detection systems and have demonstrated their effectiveness in practical application scenarios [18-19]. Kuan et al. [20] employed OpenPose to extract 25 human joint keypoints and applied linear interpolation to compensate for missing joints, thereby constructing continuous skeleton sequences for fall recognition using RNN-, LSTM-, and GRU-based models. By utilizing multi-view and multi-environment video data and replacing raw image features with joint motion dynamics, their approach effectively reduced the impact of illumination variations and motion blur. Experimental results showed that the proposed LSTM-RNN and GRU models achieved an accuracy of 98.2% in action sequence recognition, representing an improvement of approximately 9.3% over baseline methods. These findings indicate that skeleton-based fall detection approaches, which are inherently less sensitive to background complexity, illumination changes, and clothing variations, have become a dominant paradigm in recent fall detection research. More recently, Google's MediaPipe Pose framework has attracted attention due to its lightweight design and real-time inference capability. MediaPipe Pose outputs 33 human keypoints and incorporates structural inference mechanisms to recover missing joint information under partial occlusions, thereby enhancing the stability of skeleton sequences. Bugarin et al. [21] utilized MediaPipe Pose to extract skeletal joint coordinates and trained classifiers for fall detection, achieving favorable recognition performance. Other studies have further deployed a MobileNetV2-based MediaPipe Pose model on smartphones, training the system using multiple RGB image sequences and integrating a random forest classifier to improve recognition performance on mobile devices. In addition to providing IoT-based notifications and real-time video feedback, experimental results demonstrated high detection accuracy and strong practical applicability in real-world scenarios.

Based on the above research developments, fall detection technologies have evolved from early static posture analysis relying on single images to dynamic decision-making approaches centered on skeleton sequences and deep learning models. Skeleton data can effectively capture critical motion characteristics such as center-of-mass descent, joint angle variations, and posture transitions, making them particularly suitable for recognizing fall events that exhibit pronounced temporal dynamics. In this study, MediaPipe Pose is employed to extract human skeletal keypoints, and a fall detection framework integrating a Convolutional Neural Network (CNN) with a temporal screening strategy is proposed to achieve both high sensitivity and high recognition accuracy. Furthermore, an Internet of Things (IoT) mechanism is integrated into an ESP32-based platform, enabling real-time alert transmission upon fall event detection and providing immediate feedback and remote safety monitoring. This integration enhances the practical applicability of the proposed system in intelligent healthcare scenarios. Overall, the proposed skeleton-based fall detection approach not only improves recognition performance but also demonstrates the potential of human-centered intelligent healthcare applications by emphasizing the practical value of technology in safety protection and social care. Building upon this research context, the present work further adopts MediaPipe Pose as a unified skeleton data source and constructs a deep learning framework applicable to multiple human behavior recognition tasks. Gait-based identity recognition and fall detection are subsequently employed as representative case studies to validate the generality and practical feasibility of the proposed framework.

Recent studies on human behavior analysis and intelligent systems can be broadly categorized into several directions. IoT-based and embedded system architectures have been widely developed for real-time monitoring and distributed applications, such as lightweight wireless platforms and fog computing frameworks (e.g., see [22-23]). Convolutional neural network (CNN)-based approaches have demonstrated strong capability in image processing and feature extraction tasks, particularly in learning spatial representations for applications such as contrast enhancement and fall detection (e.g., see [24-25]). In addition, vision-based recognition methods have been applied to various practical problems, including digit recognition and coordinate extraction from structured images (e.g., see [26-27]). Furthermore, several studies have explored the integration of computer vision with robotic and intelligent systems, such as robotic manipulation and UAV-based applications (e.g., see [28-29]). These approaches highlight the feasibility of combining perception, decision-making, and actuation in real-world environments.

Final, compared with existing skeleton-based approaches that directly utilize joint coordinates or temporal sequences, the proposed method transforms skeleton data into angle-based image representations. This representation effectively captures the relative motion patterns of body segments while reducing the dimensionality of the input data. Furthermore, by converting temporal features into image-based forms, the proposed approach enables the use of lightweight CNN

models instead of computationally expensive sequence models such as LSTM or GRU. As a result, the proposed framework achieves a balance between representation capability and computational efficiency, making it more suitable for real-time and embedded applications. In addition, unlike many existing studies that focus on a single task, the proposed framework demonstrates its generality by being applicable to both gait-based identity recognition and fall detection tasks.

3 Skeleton-Based Feature Representation

3.1 Overview of the Proposed Framework

This study proposes a unified skeleton-based deep learning analysis framework for two human behavior recognition tasks: gait-based identity recognition and fall detection. The overall system consists of five main modules, including data acquisition, skeleton extraction, feature construction, model training, and real-time inference. The complete research workflow of the proposed framework is illustrated in Figure 1. RGB video data are first captured and processed to extract human skeletons using MediaPipe/OpenPose. Temporal skeleton features are then constructed and encoded into image-like representations, which are fed into a shared CNN backbone to produce task-specific outputs for gait-based identity recognition and fall detection.

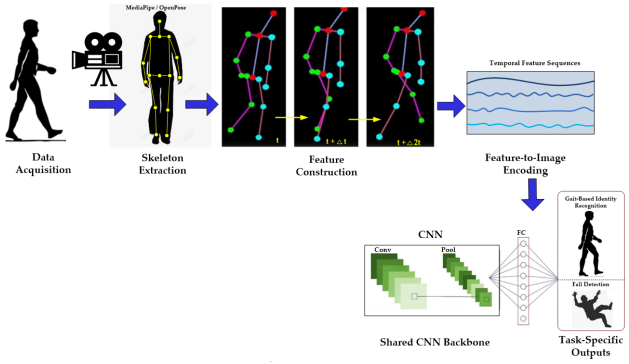

Figure 1. Overview of the proposed unified skeleton-based framework

In the data acquisition stage, RGB cameras are used to record subjects' walking behaviors and fall-related activities, which serve as the input data for subsequent skeleton analysis. In the skeleton extraction stage, the captured image sequences are processed using Google MediaPipe Pose to perform human pose estimation and automatically infer the coordinates of 33 human skeletal keypoints. To reduce the effects of variations in body height, camera distance, and image jitter, the extracted skeletal data are further normalized.

In the feature construction stage, task-specific skeleton feature representations are designed according to the characteristics of each application. For gait-based identity recognition, the feature design focuses on limb joint angles, relative joint displacements, and their temporal variations in order to capture stable gait cycle characteristics and discriminative motion patterns across individuals. For fall detection, the feature modeling emphasizes abrupt posture changes, angular discontinuities, and sudden displacement variations to highlight the distinctive temporal dynamics associated with fall events. In the model training stage, the constructed skeleton feature sequences are fed into convolutional deep learning models to train dedicated classifiers for gait identity recognition and fall detection, respectively. In the real-time inference stage, the system continuously acquires image data and generates updated skeleton sequences, which are subsequently input into the trained models for online recognition. The gait module is capable of identifying individuals passing through the camera's field of view in real time, while the fall detection module can promptly determine the occurrence of a fall upon detecting abnormal posture transitions.

Furthermore, an Internet of Things (IoT) alert mechanism is implemented through integration with an ESP32 microcontroller. When a fall event is detected by the system, the event is triggered at the edge device and transmitted via Wi-Fi through an IoT communication layer to a Raspberry Pi–based backend server and database for storage and management. A web-based frontend interface continuously retrieves the updated database content to display the latest fall status on mobile phones, tablets, or other smart devices, while simultaneously triggering warning indicators such as LEDs, buzzers, or emergency notifications to caregivers. Through this integrated architecture, the proposed system forms a practical, scalable, and real-time intelligent fall monitoring and notification platform, as illustrated in Figure 2.

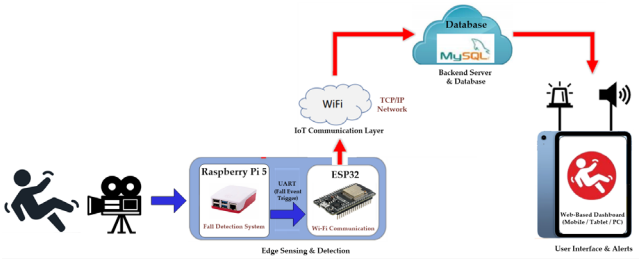

Figure 2. System architecture of the proposed fall detection and IoT-based alert platform

3.2 Data Acquisition

In this study, standard RGB cameras are employed as the primary data sources to collect image sequences of gait behaviors and fall-related activities, supporting subsequent skeleton extraction and the construction of behavior recognition models. To enhance data diversity and improve model generalization, multiple imaging devices are utilized during data collection, including built-in laptop cameras, smartphone cameras, and external Full HD cameras. Devices with different resolutions and installation positions are intentionally incorporated to reduce the model's dependency on specific hardware configurations and to improve cross-device applicability. Data acquisition is conducted across multiple indoor environments under varying lighting conditions, including

natural light, fluorescent illumination, and mixed lighting scenarios. As a result, the captured images exhibit variations in brightness distribution, shadows, reflections, and background complexity. This design prevents the model from overfitting to specific recording conditions and enhances its robustness across different deployment environments. All recorded image sequences are initially collected in an offline manner for quality inspection to ensure completeness and skeleton detectability. During system validation, real-time image streaming is employed to evaluate the model's responsiveness and robustness under practical application scenarios.

For gait data acquisition, subjects enter the camera's field of view from outside the scene, walk continuously across the monitored area, and then exit the view, ensuring that complete gait cycles are fully captured, as illustrated in Figure 3. Fall-related data cover three primary states, including normal activities prior to a fall, posture transitions during the falling process, and static lying postures after a fall, as shown in Figure 4. All recordings ensure that the subjects' full bodies remain within the camera frame to prevent incomplete skeleton extraction, which could otherwise lead to missing keypoints or erroneous feature representations. Through the above data collection strategy involving multiple devices, multiple indoor environments, and diverse motion postures, a representative image dataset with sufficient variability is established. This dataset provides a reliable foundation for subsequent skeleton-based feature construction and deep learning model training. The dataset consists of video sequences captured by a monocular camera. Each video is converted into frames, from which skeletal keypoints are extracted using MediaPipe Pose. The analysis is therefore based on skeleton-derived representations rather than raw images. The dataset was divided into training and testing sets with a ratio of 80% to 20%. To avoid data leakage, the partition was performed on a per-subject basis, ensuring that samples from the same individual did not appear in both sets. During training, 20% of the training data were further used as a validation set to monitor the learning process and prevent overfitting. The data partition was performed prior to sliding window generation. The 60-frame sliding windows were then generated separately within the training and testing sets. Although overlapping windows may exist within the same sequence, no overlapping or highly correlated segments were shared between the training and testing sets.

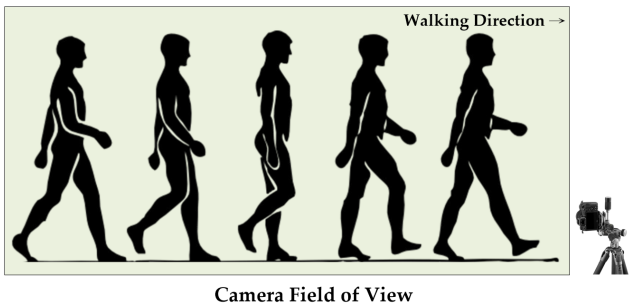

Figure 3. Gait data acquisition for identity recognition

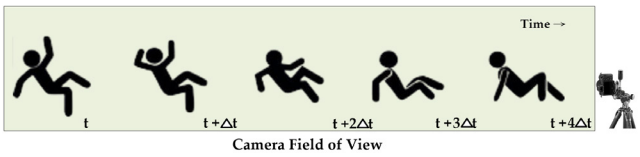

Figure 4. Fall data acquisition across temporal progression within the camera field of view

3.3 Skeleton Extraction

This study employs MediaPipe Pose to extract the three-dimensional coordinates (x_m, y_m, z_m) and the corresponding visibility score v_m for each of the 33 human skeletal keypoints (m = 1, …, 33). The resulting skeleton sequences generated for each video frame are further preprocessed to enhance data consistency and the reliability of subsequent feature construction. The definitions of the skeletal keypoints are illustrated in Figure 5.

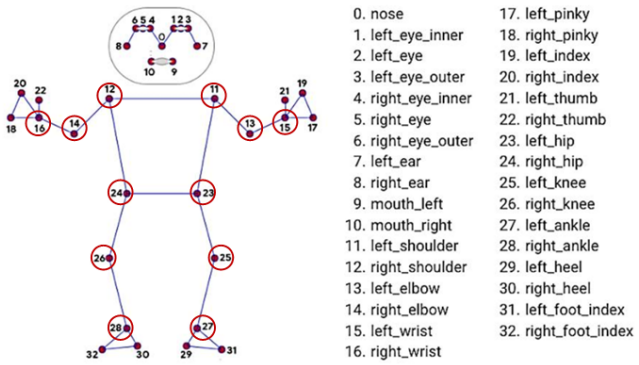

Figure 5. MediaPipe 33-point pose landmark configuration with selected joints highlighted

Within a Python environment, the recorded image sequences are read using OpenCV, and MediaPipe Pose is subsequently applied to perform frame-by-frame human pose estimation. The extracted skeletal joint coordinates are stored in a list structure to facilitate subsequent temporal feature analysis and deep learning model training. The overall preprocessing procedure consists of the following three main steps.

3.3.1 Coordinate Normalization

To reduce the scale variations caused by differences in subjects' height, body proportions, and camera distance, the hip center is defined as the coordinate origin, and the shoulder width is employed as a scaling factor for spatial normalization. Through this normalization process, all skeleton sequences are transformed into a unified relative coordinate system, thereby improving the feasibility of cross subject comparison. Since MediaPipe Pose does not provide a dedicated hip center joint, the midpoint between the left hip (joint 23) and the right hip (joint 24) is defined as the skeleton origin, which is computed as:

$$\text{hip\_center(hc)} = \frac{P_{23} + P_{24}}{2}, \tag{1}$$

Furthermore, to characterize the overall posture of the trunk, a shoulder center is defined as a reference point for

the upper body. Since MediaPipe Pose does not provide a single joint representing the shoulder center, the midpoint between the left shoulder (joint 11) and the right shoulder (joint 12) is adopted. The shoulder center is computed as follows:

$$\text{should\_center(sc)} = \frac{P_{11} + P_{12}}{2}, \tag{2}$$

The defined shoulder center is subsequently used for computing the trunk orientation angle, which describes the posture of the trunk relative to the vertical direction during walking. Scale normalization is performed using the shoulder width, defined as the Euclidean distance between the left shoulder (joint 11) and the right shoulder (joint 12):

$$d = \| P_{11} - P_{12} \|, \tag{3}$$

Accordingly, the normalized coordinate of the *i*-th joint is expressed as:

$$\overline{P}_i = \frac{P_i - \text{hip\_center}}{d}, \tag{4}$$

where P_i denotes the original three-dimensional coordinate of the *i*-th joint. This normalization strategy effectively mitigates scale variations arising from differences in body size, camera distance, and camera placement, thereby ensuring improved consistency and comparability among skeleton sequences across different subjects.

3.3.2 Temporal Smoothing

Since skeletal keypoint detection may be affected by illumination variations, partial occlusions, or rapid body movements, temporal jitter can occur in the extracted skeleton sequences. To mitigate such noise while preserving the primary motion patterns, a moving-average filter is applied to the normalized temporal skeletal coordinates, thereby improving the stability of subsequent temporal feature representations.

Let $\tilde{P}_i(t)$ denote the normalized coordinate of the *i*-th joint at frame *t*. The smoothed coordinate obtained using a moving-average filter with a window length of *k* is defined as:

$$\hat{P}_i(t) = \frac{1}{k} \sum_{j=0}^{k-1} \overline{P}_i(t-j), \tag{5}$$

where $\hat{P}_i(t)$ represents the smoothed joint coordinate and *k* denotes the length of the averaging window. This smoothing strategy effectively suppresses high-frequency fluctuations in the skeletal sequences while retaining the dominant temporal motion characteristics of human activities.

3.3.3 Joint Data Imputation (Occlusion Handling)

When certain joints are occluded or fail to be detected during pose estimation, missing joint data may occur, as indicated by a visibility score of $v = 0$. Such missing data lead to discontinuities in the skeleton sequences, which can adversely affect feature construction and the stability of model training. To preserve temporal continuity, a linear interpolation strategy based on neighboring frames is adopted to impute missing joint data.

Specifically, if the visibility of the *i*-th joint at frame tis $v(t) = 0$, and the smoothed joint coordinates at the nearest valid preceding frame t_1 and subsequent frame t_2 are given by $\hat{P}_i(t_1)$ and $\hat{P}_i(t_2)$, respectively, the imputed coordinate at frame tis computed using linear interpolation as follows:

$$\hat{P}_i(t) = \hat{P}_i(t_1) + \frac{t - t_1}{t_2 - t_1} \left[\hat{P}_i(t_2) - \hat{P}_i(t_1) \right], \tag{6}$$

This interpolation approach effectively mitigates feature discontinuities and temporal fluctuations caused by short-term occlusions or detection failures. After completing the aforementioned preprocessing steps, the resulting stable skeleton sequences are further utilized for task-specific feature extraction, including relative joint positions, joint angle information, and their temporal variations, to construct discriminative features for gait recognition and fall detection.

3.4 Gait Feature Construction

3.4.1 Gait Joint Angle Features

Since human gait exhibits pronounced periodic characteristics, this study selects 12 major joints that are highly relevant to gait motion from the 33 skeletal joints provided by MediaPipe Pose. These joints include the left and right shoulders (joints 11 and 12), elbows (joints 13 and 14), wrists (joints 15 and 16), hips (joints 23 and 24), knees (joints 25 and 26), and ankles (joints 27 and 28), as illustrated in Figure 5. This set of joints effectively describes limb swing behaviors and trunk motion trajectories during walking, and covers the most discriminative dynamic information within a gait cycle. In contrast to fall detection, which primarily focuses on abrupt changes in trunk position, gait-based identity recognition is more sensitive to subtle limb movements and the rhythmic patterns of whole-body motion that characterize individual walking styles. Accordingly, the aforementioned 12 joints are adopted as the core inputs for gait feature construction to ensure that subject-specific gait patterns can be sufficiently captured. Based on these selected joints, the temporal variations of joint angles relative to the vertical direction are utilized as the gait feature representation. Such angle-based features, expressed as structured temporal sequences, can simultaneously characterize relative limb postures and overall trunk motion trends within a unified representation, thereby providing discriminative information for subsequent classification models.

Considering that, during planar walking, the angles of body segments relative to the vertical direction can directly reflect gait variations, this study selects seven

vertical joint angles as the primary posture-related features. These angles include the left and right upper arms, left and right thighs, left and right lower legs, as well as the main trunk angle relative to the vertical direction, as illustrated in Figure 6. By converting the dominant postural variations of the limbs and trunk into angle-based temporal sequences, discriminative motion patterns within the gait cycle can be effectively captured. Although the complete details of human gait cannot be fully described using a limited number of angular features, under two-dimensional image conditions, the selected seven vertical angles sufficiently capture key information related to limb swing amplitudes and trunk stability, which are essential components of gait dynamics. Compared with approaches that rely on only a small number of angular features, the seven-angle representation adopted in this study provides a more comprehensive and representative description of human posture, thereby contributing to improved performance in gait-based identity recognition. To further clarify the angle definition, each angle is defined based on the relative orientation between connected body segments. These angles are computed from skeleton keypoints and are independent of any global coordinate system.

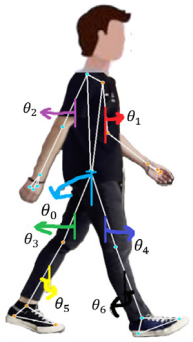

Figure 6. Definition of the seven body-segment angles used for feature extraction

The definitions of the selected body segment angles are summarized in Table 1.

Table 1. The description of seven body segment angles

Angle ID	Body segment	Joint pair (a→b)	Description
θ_1	Left upper arm	11 → 13	Shoulder to elbow (left)
θ_2	Right upper arm	12 → 14	Shoulder to elbow (right)
θ_3	Left thigh	23 → 25	Hip to knee (left)
θ_4	Right thigh	24 → 26	Hip to knee (right)
θ_5	Left shank	25 → 27	Knee to ankle (left)
θ_6	Right shank	26 → 28	Knee to ankle (right)
θ_7	Trunk	hc → sc	Main trunk orientation

3.4.2 Gait Joint Angle Features

In planar images, the orientation of a body segment relative to the vertical direction can be directly derived from the joint coordinates. For any two joint points $P_a(x_a, y_a)$ and $P_b(x_b, y_b)$, the corresponding limb segment vector is defined as:

$$\overrightarrow{P_aP_b}(t) = \left(x_b(t) - x_a(t), y_b(t) - y_a(t)\right), \tag{7}$$

Since all joint angles in this study are measured with respect to the vertical direction, the angular deviation of a limb segment formed by the proximal joint P_a and the distal joint P_b can be defined accordingly. Specifically, the angular deviation of the segment relative to the vertical direction at time t, denoted as $\theta_{a\to b}(t)$, is computed as the ratio between the horizontal and vertical components of the segment vector:

$$\theta_{a\to b}(t) = \tan^{-1}\left(\frac{x_b(t) - x_a(t)}{y_b(t) - y_a(t)}\right), \tag{8}$$

where $\theta_{a\to b}(t)$ represents the angular deviation of the limb segment from joint P_a to joint P_b relative to the vertical direction. This angular representation is used to characterize the posture variations of the upper limbs, lower limbs, and the main trunk during the gait process. Based on the skeletal joint coordinates provided by MediaPipe Pose, seven vertical joint angles are computed accordingly and represented as seven angle-based temporal sequences. These angle trajectories serve as the primary representation of gait motion in this study. Figure 7 illustrates the angle variation curves extracted from a complete walking sequence, along with the corresponding color-coded mapping for each joint angle.

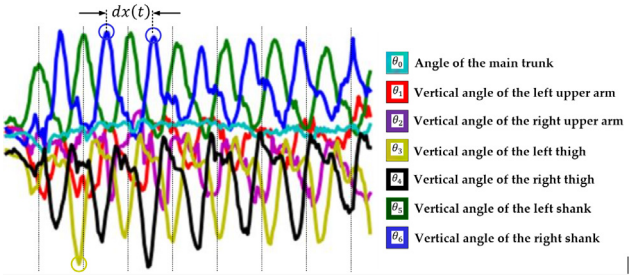

Figure 7. Temporal variations of the defined body-segment angles (θ_0–θ_6) during walking

3.4.3 Gait Cycle–Aligned Angle Representation

During gait locomotion, the horizontal distance between the two ankle joints exhibits a clear periodic pattern, characterized by a cyclic variation in which the distance gradually decreases from a maximum value and subsequently increases again. To quantify this behavior, the horizontal displacement between the left and right ankles is defined as

$$dx(t) = x_{27}(t) - x_{28}(t), \tag{9}$$

where $x_{27}(t)$ and $x_{28}(t)$ denote the horizontal coordinates of the left ankle (joint 27) and the right ankle (joint 28) at frame ttt, respectively.

As illustrated in Figure 8(a) $dx(t)$ represents the horizontal separation between the ankle joints with respect to a vertical reference. When $dx(t) > 0$ and reaches its maximum value, the left foot is located in front of the right foot, corresponding to a maximal stride condition. Conversely, when $dx(t) < 0$ and its absolute value reaches a maximum, the right foot leads while the left foot trails. The condition $dx(t) = 0$ indicates that the two ankles complete a forward–backward positional exchange in the horizontal direction. Furthermore, as shown in Figure 8(b), the temporal signal of $dx(t)$ exhibits a clear alternation between positive and negative values with distinct extrema. Although the leading foot alternates between the left and right sides across consecutive strides, this role switching does not affect gait cycle boundary detection, since the absolute value $|dx(t)|$ consistently forms prominent extrema at each maximal stride.

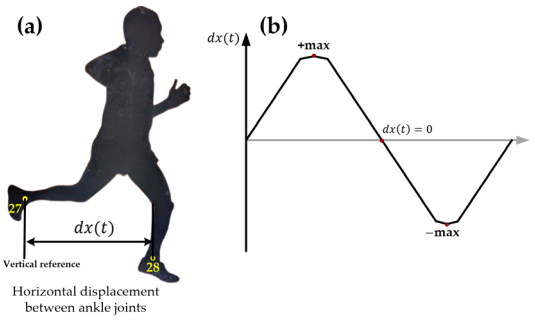

Figure 8. Horizontal ankle displacement $dx(t)$: (a) Definition between ankle joints; (b) Alternating extrema over gait cycles

Based on this observation, the extrema associated with sign changes in $dx(t)$ are employed as gait cycle segmentation points, and a complete gait cycle is defined as the interval between two successive maximal stride events. This segmentation strategy effectively removes redundant or incomplete gait segments and enables the subsequent segmentation and temporal alignment of joint angle sequences in both duration and phase. As a result, a cycle-consistent angle representation is obtained, improving the consistency of the extracted gait features and benefiting subsequent feature analysis and deep learning–based recognition. Figure 7 illustrates the long-term temporal behavior of the defined joint angle features across multiple consecutive gait cycles, where the detected cycle boundaries are indicated. After segmentation and normalization, the vertical joint angle trajectories corresponding to a single representative gait cycle are shown in Figure 9.

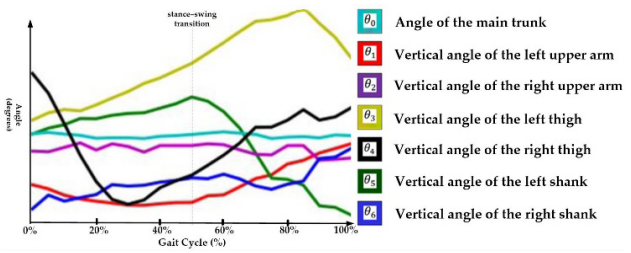

Figure 9. Normalized vertical joint angle trajectories over a single gait cycle (0–100%)

3.5 Fall Detection Feature Construction

3.5.1 Fall-Related Joint Angle Features

In the fall detection task, significant posture changes are primarily attributed to the global displacement and rotation of the trunk, whereas limb movements often exhibit high variability due to natural swinging motions or protective reactions during falling events. Such variability may adversely affect feature stability and classification reliability. Therefore, this study focuses exclusively on six joints provided by MediaPipe Pose that are most relevant to the trunk structure, including the left shoulder (joint 11), right shoulder (joint 12), left hip (joint 23), right hip (joint 24), left knee (joint 25), and right knee (joint 26), as illustrated in Figure 5. These joints are sufficient to characterize the principal dynamic posture changes of the trunk during a fall, while effectively reducing the interference caused by arm and forearm movements.

Since the six selected joints correspond to a total of twelve coordinate values (x, y), directly using the raw joint coordinates as features would result in a relatively high-dimensional representation and make posture variations less intuitive to interpret. To reduce feature dimensionality and enhance robustness, the left–right symmetric joint pairs at the shoulder, hip, and knee are further averaged to form three representative trunk nodes, namely the shoulder center, hip center, and knee center. Through this transformation, the feature dimensionality is effectively reduced from 12 to 6, while mitigating the influence of individual body shape variations (e.g., waist width) on the analysis results. The midpoint coordinates of each symmetric joint pair are computed according to Equations (10)–(12). Specifically, the shoulder center is obtained from joints 11 and 12, the hip center from joints 23 and 24, and the knee center from joints 25 and 26. For notational simplicity, the midpoints of these symmetric joint pairs are hereafter denoted as the shoulder center (x_{sc}, y_{sc}) the hip center (x_{hc}, y_{hc}), and the knee center (x_{kc}, y_{kc}) respectively, as defined in the corresponding formulations.

$$x_{sc} = \frac{x_{11} + x_{12}}{2}, y_{sc} = \frac{y_{11} + y_{12}}{2}, \quad \textbf{(10)}$$

$$x_{hc} = \frac{x_{23} + x_{24}}{2}, y_{hc} = \frac{y_{23} + y_{24}}{2}, \quad \textbf{(11)}$$

$$x_{kc} = \frac{x_{25} + x_{26}}{2}, y_{kc} = \frac{y_{25} + y_{26}}{2}, \quad \textbf{(12)}$$

By analyzing the temporal variations of these three representative trunk nodes, key dynamic characteristics of fall events, including trunk descent, inclination, and translational motion, can be effectively captured. These features are subsequently used as inputs to the deep learning–based classification model, thereby enhancing the stability and recognition performance of the proposed fall detection system.

3.5.2 Trunk Angle and Vertical Displacement Features

Building upon the skeleton representation in which

the hip center is defined as the coordinate origin, this study defines two trunk-related vectors: the vector from the hip center to the shoulder center and the vector from the hip center to the knee center, as illustrated in Figure 10(a). These vectors effectively characterize the principal temporal variations of trunk posture. During different activities, such as standing, sitting, and falling, variations in trunk flexion relative to the vertical direction, as well as overall body displacement, are reflected in changes in the direction and magnitude of these vectors. Since a fall event is inherently associated with pronounced downward displacement and inclination of the trunk, tracking only three representative joints, namely the shoulder, hip, and knee, is sufficient to capture the critical postural information required for fall detection. This design also mitigates the influence of large limb movements that may otherwise introduce instability and noise into the feature representation.

In the feature design stage, a structured representation is further constructed based on the three representative trunk joints described above. First, the vertical coordinate variation of the hip center is retained to reflect temporal changes in waist height. During a fall event, the hip height typically exhibits a pronounced decrease within a short time interval. In contrast, horizontal displacement of the hip center mainly reflects lateral position changes of the subject in the image plane and is less directly related to fall occurrences. Therefore, it is not included in the subsequent feature set. On this basis, the vector from the hip center to the shoulder center and the vector from the hip center to the knee center are defined as the upper trunk vector V_1 and the lower trunk vector V_2, respectively, as illustrated in the figure. Through this vector-based representation, the original 12-dimensional coordinate information derived from six joint points is transformed into five primary features, consisting of the planar components (x, y) of the two trunk vectors and the vertical height of the hip center. The vertical height of the hip center at time t, denoted as $y_{hc}(t)$, is computed as the average vertical coordinate of the left and right hip joints in the image coordinate system and is expressed in Equation (11).

To further characterize trunk bending and inclination during a fall event, two trunk vectors are introduced: the upper trunk vector $\mathbf{V}_1$ and the lower trunk vector $\mathbf{V}_2$. The directions of $\mathbf{V}_1$ and $\mathbf{V}_2$ are defined from the hip center toward the shoulder center and from the hip center toward the knee center, respectively, as illustrated in Figure 10(a). These two vectors effectively describe the overall temporal evolution of trunk posture. The angle between the two vectors can be computed using the vector dot product, and its mathematical formulation is given as follows:

$$\mathbf{V}_1 \cdot \mathbf{V}_2 = \| \mathbf{V}_1 \| \, \| \mathbf{V}_2 \| \cos\theta , \tag{13}$$

Here, θ denotes the angle between the upper and lower trunk vectors, and $\|\mathbf{V}_1\|$ and $\|\mathbf{V}_2\|$ represent the magnitudes of vectors $\mathbf{V}_1$ and $\mathbf{V}_2$, respectively. Let $\mathbf{V}_1 = (x_1, y_1)$ and $\mathbf{V}_2 = (x_2, y_2)$; the angle θ can then be expressed as

$$\theta = \cos^{-1}\left(\frac{x_1 x_2 + y_1 y_2}{\sqrt{x_1^2 + y_1^2}\sqrt{x_2^2 + y_2^2}} \right), \tag{14}$$

This angular feature effectively captures the degree of trunk bending and inclination occurring during a fall event and serves as an important geometric indicator of abrupt postural changes. When combined with the displacement features of the aforementioned trunk joints, the proposed vector and angle representations provide discriminative and robust fall-related features for subsequent deep learning–based classification models.

3.5.3 Horizontal and Vertical Displacement Difference Features

When a subject is in an upright or supine posture, the angle θ between the upper and lower trunk vectors (Figure 10(a)) may both approach 180°, making it difficult to reliably distinguish between these two postural states using the angular feature alone. This limitation may lead to ambiguity in posture classification. To address this issue, the differences between the horizontal and vertical components of the upper and lower trunk vectors are introduced as auxiliary posture features. By comparing the relative variations of the two vectors along orthogonal directions, the proposed component difference features compensate for the insufficiency of a single angular descriptor and enhance the reliability of fall detection. Specifically, when the subject maintains an upright posture, the horizontal components of the upper and lower trunk vectors are similar, whereas their vertical components exhibit more pronounced differences. In contrast, when the subject is in a supine posture or after a fall event, the differences in the horizontal components of the two vectors increase significantly, while their vertical components tend to converge. This observation indicates that the vector component differences effectively capture the geometric transition of trunk posture from upright to lying, thereby providing discriminative auxiliary information for distinguishing different postural states.

To further illustrate the geometric interpretation of $\Delta x(t)$ and $\Delta y(t)$, Figure 10(b) schematically depicts the component relationships of the upper and lower trunk vectors in a two-dimensional planar coordinate system. In this study, the previously defined hip center is used as a common origin, from which the upper trunk vector and the lower trunk vector are constructed, respectively.

$$\begin{aligned} \mathbf{V}_1(t) &= \mathbf{P}_{sc}(t) - \mathbf{P}_{hc}(t) \\ &= (x_{sc}(t) - x_{hc}(t), y_{sc}(t) - y_{hc}(t)), \end{aligned} \tag{15}$$

$$\begin{aligned} \mathbf{V}_2(t) &= \mathbf{P}_{kc}(t) - \mathbf{P}_{hc}(t) \\ &= (x_{sc}(t) - x_{hc}(t), y_{sc}(t) - y_{hc}(t)), \end{aligned} \tag{16}$$

Here, $\mathbf{P}_{sc}$, $\mathbf{P}_{hc}$, and $\mathbf{P}_{kc}$ denote the coordinate positions of the shoulder center, hip center, and knee center, respectively. Under this formulation, the differences between the two vectors in the horizontal and vertical directions can be directly computed from their vector

components and are expressed as:

$$\Delta x(t) = V_{1x}(t) - V_{2x}(t) = x_{sc}(t) - x_{kc}(t), \quad (17)$$

$$\Delta y(t) = V_{1y}(t) - V_{2y}(t) = y_{sc}(t) - y_{kc}(t), \quad (18)$$

Here, $V_{1x}(t)$ and $V_{1y}(t)$, as well as $V_{2x}(t)$ and $V_{2y}(t)$, denote the horizontal and vertical components of the upper and lower trunk vectors, respectively. The horizontal component difference $\Delta x(t)$ characterizes the relative lateral displacement between the upper and lower trunk segments, whereas the vertical component difference $\Delta y(t)$ reflects their relative height variation along the vertical direction.

From a geometric perspective, when the subject maintains an upright posture, the horizontal components of the two trunk vectors tend to be similar, resulting in a small $\Delta x(t)$, while their vertical components differ significantly, leading to a large $\Delta y(t)$. In contrast, when the subject is in a supine posture or after a fall event, the horizontal component difference $\Delta x(t)$ increases markedly, whereas the vertical component difference $\Delta y(t)$ decreases, as the trunk orientation becomes more aligned with the horizontal plane. For intermediate postures, such as trunk bending, both $\Delta x(t)$ and $\Delta y(t)$ typically exhibit moderate values. These component difference features effectively complement the trunk angle descriptor by resolving posture ambiguity arising from angle similarity alone. When combined with the trunk angle feature and the vertical height of the hip center, the proposed representation provides a discriminative and robust set of fall-related posture features for subsequent classification.

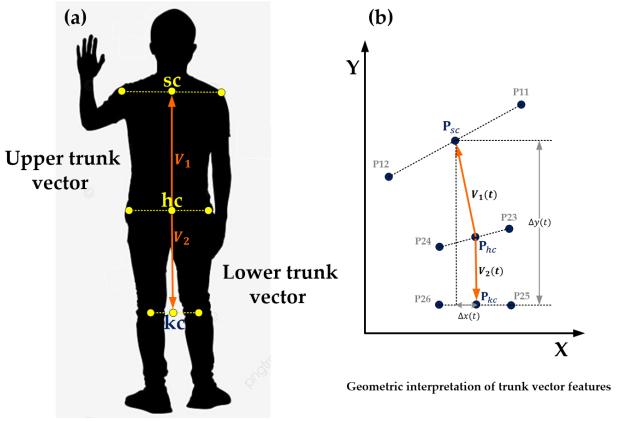

Figure 10. Trunk vector features: (a) Upper and lower trunk vectors; (b) Geometric representation in the X–Y plane

3.5.4 Temporal Feature Representation and Sliding Window Construction

Based on the above feature design, the original 12-dimensional skeleton coordinates derived from six joint points are first transformed into two representative trunk vectors and the vertical displacement of the hip center. Subsequently, the trunk angle θ and the component differences Δx and Δy are extracted. As a result, a four-dimensional fall posture feature vector, $\{\theta(t), \Delta x(t), \Delta y(t), y_{hc}(t)\}$, is constructed. This feature representation is physically interpretable and semantically explicit, and it serves as the input to the subsequent deep learning model to enhance the stability and recognition performance of fall detection.

All four features are time-varying sequences and must be considered along the temporal axis to fully characterize the dynamic nature of a fall event. In general, a fall occurs within approximately 2–3 seconds and is accompanied by pronounced postural and spatial changes. In this study, video data are captured at a frame rate of 30 fps, yielding 30 feature vectors per second. In practice, a sliding window with a length of 60 frames (approximately 2 seconds) is adopted to construct an action sequence sample. When 60 frames are accumulated, the first sample is formed. Thereafter, as each new frame is acquired, the oldest frame is discarded and the latest frame is appended, generating a new 60-frame feature sequence.

Through this sliding-window mechanism, the complete temporal evolution of a fall motion can be preserved under a fixed time scale, while the number of available training samples is significantly increased. In this study, the four features are visualized as temporal feature curves. Figure 11(a) illustrates the feature trajectories under normal standing or walking conditions, whereas Figure 11(b) depicts the case in which a fall event occurs within the sequence, where the four feature curves are defined using different colors. The pronounced variations observed in these feature curves provide important evidence for subsequent deep learning–based classifiers to identify fall events.

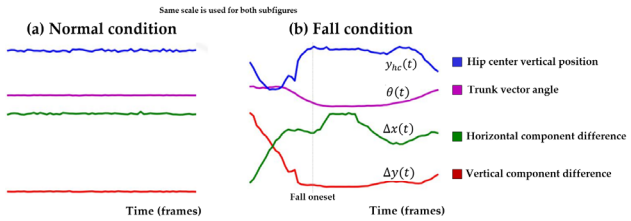

Figure 11. Feature trajectories under (a) Normal walking and (b) Fall conditions

4 Skeleton-Based Feature Representation

4.1 Skeleton-Based Temporal Encoding

Skeleton-based temporal features can generally be modeled using two different strategies. The first approach treats the multi-dimensional skeleton features at each frame (e.g., seven joint angle features) as a one-dimensional feature vector evolving over time and feeds the resulting $T\times7$ numerical sequence into sequential models such as 1D-CNNs, LSTMs, or GRUs. The second approach converts the multi-dimensional temporal signals into two-dimensional angle-plot images, enabling the use of 2D convolutional neural networks to capture local textures and shape patterns across different feature channels. Although the former approach is capable of handling pure temporal sequences, its input representation is limited to a $T\times7$

sequence and cannot explicitly encode phase differences, waveform shapes, or coordinated variations among different angle features. In contrast, the image-based representation preserves both the temporal dimension and the spatial relationships across features, resulting in richer texture and geometric patterns that are particularly well suited for high-level feature extraction using convolutional neural networks.

After comparing these two modeling strategies, this study adopts two-dimensional angle-plot images as the primary input representation. Compared with one-dimensional temporal sequences, the 2D images constructed from seven color-coded angle curves not only encapsulate the complete temporal evolution of gait or fall-related motions but also reveal inter-angle interactions. This representation allows CNNs to more effectively learn individualized gait characteristics and motion phase transitions. Experimental results further demonstrate that 2D-CNNs based on the proposed image-based encoding achieve superior performance in both gait identification and fall detection tasks. Accordingly, all subsequent model training and performance evaluations in this study are conducted using two-dimensional CNN architectures.

4.2 Gait Identification Model Based on Angle-Plot Images

To effectively exploit the inherent periodic characteristics of human gait and the coordination among multiple joints, this study proposes a gait identification approach based on angle-plot images. Unlike conventional methods that directly feed one-dimensional joint-angle time series into sequential models, the proposed method visualizes the temporal variations of seven joint angles ($\theta_0 \sim \theta_6$) and converts them into a two-dimensional image representation. In this manner, temporal information is re-embedded into a spatial structure, enabling effective feature extraction using convolutional neural networks (CNNs). Specifically, the selected seven joint angles correspond to the main trunk, the left and right upper limbs, and the left and right lower limbs. Their angle trajectories are plotted using fixed colors on a shared image plane, forming an angle-plot image that integrates dynamic information from multiple joints. Each image is generated from a sequence of 60 consecutive frames (approximately 2 seconds) of skeleton data, covering at least one complete gait cycle. Through this representation, not only are the periodic characteristics of gait motion preserved, but relative phase differences between the left and right limbs, joint-specific amplitude variations, and inter-joint coordination patterns are simultaneously captured. These characteristics are reflected in the shapes of the curves, the distribution of peaks, and the overlapping color structures, allowing gait differences to be clearly expressed in an image-based form.

For model design, a lightweight convolutional neural network is adopted as the classifier, treating the angle-plot image as a standard image input during training. The CNN extracts local texture and shape features through successive convolutional layers, reduces spatial dimensionality via pooling layers, and finally outputs subject identity predictions through fully connected layers, as illustrated in Figure 12. Since the angle-plot image jointly encodes temporal evolution and cross-angle spatial relationships, the CNN can automatically learn discriminative gait patterns without requiring additional temporal alignment or handcrafted feature design. Compared with traditional one-dimensional temporal models, the proposed angle-image representation more effectively preserves the structural information embedded in gait sequences, thereby enhancing the separability of gait patterns across different subjects. Subsequent experimental analysis will verify that the proposed angle-plot–based gait identification method improves classification accuracy, demonstrating the feasibility and effectiveness of transforming skeleton-based temporal features into two-dimensional images and applying CNN-based recognition.

4.3 Fall Detection Model Based on Vector-Angle Images

Fall events are characterized by distinct posture transitions, primarily driven by rapid trunk displacement, inclination, and a sudden decrease in body height. Therefore, compared with gait identification, which requires detailed limb dynamics, this study designs a simplified yet representative set of skeleton features specifically for fall detection. From the 33 joints provided by MediaPipe Pose, six joints that are most relevant to trunk motion are selected, namely the left and right shoulders (Joints 11 and 12), hips (Joints 23 and 24), and knees (Joints 25 and 26). These joints are further reduced to three midpoints: the shoulder center, hip center, and knee center. This simplification preserves the core postural information of the human body while substantially reducing the feature dimensionality, making the representation more suitable for real-time classification.

To characterize posture changes during a fall, two adjacent vectors are constructed from the three midpoints, namely the shoulder-to-hip vector and the hip-to-knee vector, from which three categories of motion-sensitive features are extracted. The first category is the vector angle, which reflects posture transitions from standing (approximately 180°), to bending (less than 150°), and finally to a fallen state (approaching 180° again). The second category consists of vector component differences (Δx and Δy), which are introduced to resolve posture ambiguity that may arise when relying solely on angular information, such as the case where upright standing and supine postures both yield near-straight angles. The third category is the vertical height of the hip center (hip_y), which captures the sudden decrease in body height during a fall. Together, these four features ($\theta(t)$, $\Delta x(t)$, $\Delta y(t)$, $y_{hc}(t)$) form a time-varying feature set that effectively describes the fall process.

In this study, a temporal window of 60 frames (approximately 2 seconds) is used to construct an action sequence, which is continuously updated using a sliding-window strategy such that a new feature sequence is generated whenever a new frame is acquired. To facilitate effective learning by convolutional neural networks, the four-dimensional temporal signals are transformed into a two-dimensional curve image, referred to as the Vector–Angle Image. In this representation, the horizontal axis

corresponds to time, the vertical axis represents the feature values, and the four feature curves are distinguished using fixed colors. This image-based encoding preserves variations in velocity, orientation, and height throughout the fall process, while enabling the model to extract local shape patterns from image textures, such as steep slopes caused by rapid height drops, abrupt changes in vector differences, and sharp transitions in angular waveforms, all of which are indicative of fall events.

Based on the proposed image representation, a lightweight convolutional neural network (CNN) is employed as the classification model, as shown in Figure 12. The network consists of three convolutional blocks, each including a convolutional layer followed by a ReLU activation function and a 2×2 max-pooling layer. The numbers of filters in the three convolutional layers are m_1, m_2 and m_3, respectively, and increase progressively to enhance feature extraction capability. The input to the network is a grayscale feature image with a size of 432×288×1. After the final pooling layer, the extracted feature maps are flattened into a one-dimensional feature vector and fed into a fully connected layer with n neurons. Finally, a Softmax layer is used to produce the classification output, indicating either a fall or a non-fall event. During training, the model was optimized using the Adam optimizer with categorical cross-entropy loss. The batch size was set to 300, the number of epochs was set to 10, and 20% of the training data were used as a validation set. Classification accuracy was adopted as the evaluation metric.

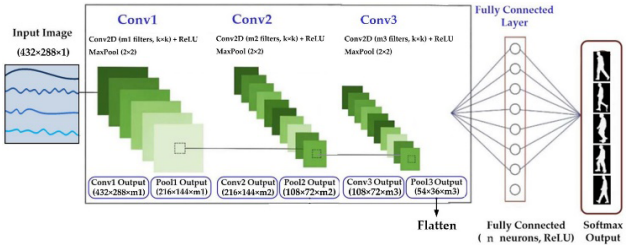

Figure 12. Architecture of the proposed CNN model

5 IoT-Enabled Fall Detection System

The proposed fall detection system comprises two core modules: a skeleton-based fall recognition model and an Internet of Things (IoT) communication framework. The recognition module acquires video streams from a camera and extracts temporal information from 33 human keypoints using MediaPipe Pose. These skeletal features are subsequently processed by a convolutional neural network (CNN) to classify the subject's activity state, including normal activities, stationary conditions, and fall events. Once a potential fall is detected, a trigger signal is immediately generated and transmitted through the built-in Wi-Fi module of an ESP32 device. The event data are sent via TCP/IP to a backend server deployed on a Raspberry Pi, as illustrated in Figure 13. The Raspberry Pi hosts an IoT platform built upon Apache, PHP, and MySQL. A PHP-based web service receives the uploaded event data and updates the MySQL database accordingly. In the proposed system, pose estimation, feature extraction, and classification are all performed on the edge device to ensure real-time processing and low latency, while only the classification results are transmitted to the IoT platform for notification and remote monitoring.

On the application side, a web-based monitoring interface implemented using HTML, CSS, and JavaScript dynamically retrieves the latest activity records through PHP scripts. Caregivers can access the server via its IP address using a smartphone, tablet, or personal computer to monitor the elderly user's activity status and fall history. Through this integrated architecture, AI-based fall recognition is seamlessly combined with IoT communication, enabling prompt event reporting and enhancing the responsiveness and reliability of home-based safety monitoring.

In the system architecture design, a conventional three-layer Internet of Things (IoT) framework is adopted to support the proposed fall detection application, as illustrated in Figure 13. The first layer corresponds to the sensing and feature extraction layer, which consists of a camera and a skeleton-based recognition model. Continuous image sequences are captured and processed to determine whether the subject is performing normal activities, remaining stationary, or experiencing a fall. Once a fall event is identified, a trigger signal is immediately generated. The second layer represents the network communication layer. An ESP32 module functions as an IoT node that transmits the trigger signal and event data to the backend server via Wi-Fi using TCP/IP protocols. The third layer is the application layer, implemented on a Raspberry Pi hosting an Apache, PHP, and MySQL server. Upon receiving event data from the ESP32, the backend server updates the MySQL database accordingly. PHP scripts periodically retrieve the latest activity records and dynamically generate a web-based monitoring interface for caregiver access. Caregivers can access the system via smartphones, tablets, or personal computers to monitor the elderly user's activity status and fall history. The overall system architecture establishes a complete workflow from sensing and detection to network transmission and web-based monitoring, as illustrated in Figure 14.

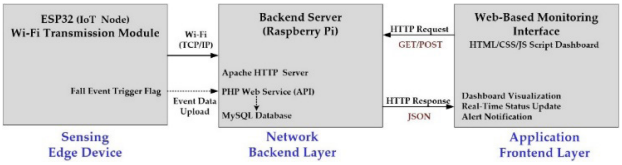

Figure 13. Three-layer IoT architecture for fall monitoring

Figure 14. IoT-based fall monitoring system deployment

The overall system latency is approximately 4 seconds. This includes both the temporal accumulation required by the 60-frame sliding window and the subsequent processing and communication stages. Specifically, approximately 2 seconds are required for the sliding window to capture sufficient temporal information for fall detection, followed by an additional delay of about 2 seconds for local processing, wireless transmission, database update, and web-page refresh. This end-to-end latency reflects the total time from the occurrence of a fall event to the appearance of the warning message on the user interface.

6 Experimental Results

6.1 Identity Recognition Performance Using Single-Step Angle Features

In the gait identity recognition experiment, a subject-independent evaluation protocol was adopted, where the training and testing sets consist of different individuals to ensure a fair assessment of generalization performance. In this study, 30 walking sequences were collected for each of six subjects, with all subjects walking from left to right. The system first extracted the leg opening angle based on skeleton keypoint detection, and a single gait step was defined as the temporal interval between two consecutive maxima of the leg opening angle. Each walking sequence was segmented into multiple step intervals, and the corresponding angle-plot images of individual steps were used as input features for the convolutional neural network (CNN) to analyze inter-subject differences in walking patterns. Figure 15 presents a comparison of classification performance under different convolution kernel sizes and CNN layer configurations. In this study, the CNN architecture is denoted as *a*–*b*–*c*–*d*, where *a*, *b*, and *c* represent the number of feature channels in the three convolutional layers, and *d* denotes the number of output classes in the final classification layer. For this identity recognition task, the number of output classes was set to match the number of subjects (six classes) as the standard configuration. In addition, CNN models with ten output classes were included as comparative baselines to evaluate the effect of output space redundancy on gait identity recognition performance.

Overall, varying the kernel size (3×3 vs. 5×5), channel width (16–32–32 vs. 32–64–64), and output dimension (6 vs. 10 classes) resulted in only marginal changes in recognition accuracy, with all tested models achieving performance within approximately 72% to 82%. Although configurations employing a 5×5 convolution kernel achieved slightly higher accuracy in some cases, no architecture consistently outperformed the others across all experimental settings. Similarly, increasing the channel width or expanding the output dimension from 6 to 10 classes did not produce a systematic or substantial improvement. The performance differences among all configurations were generally within a few percentage points, indicating that adjustments in kernel size or network capacity have limited influence under the proposed single-step angle feature representation.

This observation can be attributed to the inherent characteristics of single-step gait angle features, which primarily reflect lower-limb swing patterns within a single walking cycle and are highly sensitive to natural variations such as step length, walking speed, and individual gait habits. Consequently, single-step gait angle features represent a relatively challenging input for identity recognition. Compared with fall events, which involve rapid and pronounced posture transitions, gait identity recognition is intrinsically more difficult to achieve with high classification accuracy. The experimental results further indicate that when relying solely on single-step angle features, the average recognition accuracy remains around 78%, which is consistent with reported results in existing skeleton-based gait recognition studies. This outcome reflects the inherent performance ceiling of such features in identity recognition tasks.

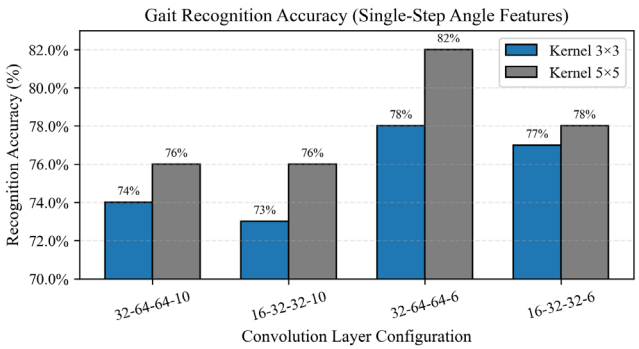

Figure 15. Gait recognition accuracy for single-step angle features

6.2 Gait Identification Performance Using Fixed Three-Step Angle Features

To further investigate the influence of extended temporal information on gait identification, an additional experiment was conducted using fixed three-step angle features collected from five subjects. The experimental protocol remained consistent with the previous setup, with all subjects walking from left to right. Each recording lasted approximately 6–8 s, yielding an average of 20 image frames per sequence. In this configuration, three consecutive complete gait steps were concatenated to form a single analysis sample. A gait segment was defined from the first occurrence of the maximum leg-opening angle to the completion of the third step. During data preprocessing, partial occlusion occurred in certain sequences, resulting in occasional missing skeletal keypoints. To maintain dimensional and temporal consistency, angle features associated with unstable joints were excluded, and only reliably detected joint angles were retained. The resulting multi-step angle-plot images were then used as input to the CNN models under different kernel sizes and channel configurations.

Figure 16 presents the identification performance under the fixed three-step setting. The results show that the recognition accuracy gradually increases as the convolution kernel size expands from 3×3 to 7×7. This

trend suggests that larger receptive fields are beneficial for capturing the more global temporal structure embedded in concatenated multi-step representations. Nevertheless, the overall recognition accuracy remains within a moderate range (approximately 60–68%), which is lower than that achieved under the single-step configuration. The reduced performance can be attributed to the absence of explicit temporal alignment across consecutive gait cycles. Although multi-step angle features theoretically contain richer temporal information, directly concatenating multiple gait cycles without phase normalization introduces inter-step misalignment and rhythmic distortion. Such inconsistencies weaken the structural coherence of the angle trajectories in the image space and reduce their discriminative capability. Furthermore, the multi-step configuration imposes stricter requirements on skeletal stability, as occlusion or keypoint jitter in any individual step propagates across the entire concatenated sequence. These findings indicate that increasing the number of gait steps does not automatically enhance identification accuracy. Instead, effective exploitation of multi-step gait information requires dedicated temporal alignment and normalization mechanisms to fully realize its potential advantages.

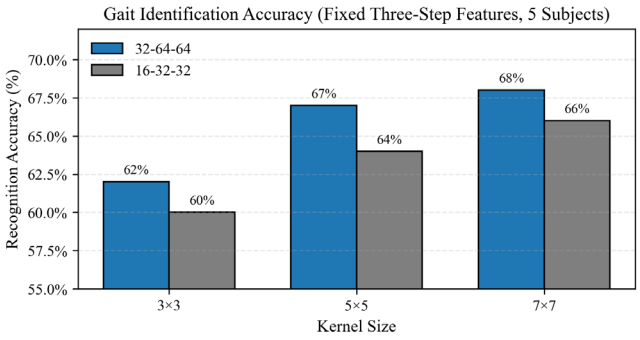

Figure 16. Gait recognition accuracy for three-step angle features

Taken together, the single-step and fixed multi-step experiments reveal that structural consistency plays a more critical role than architectural scaling in skeleton-based gait identification. Although multi-step angle features theoretically provide richer temporal cues, their effectiveness is constrained by phase misalignment and segmentation variability when temporal normalization is absent. In contrast, the single-step representation preserves intrinsic gait periodicity and therefore achieves more stable performance. These observations suggest that future improvements in gait identification should prioritize phase-consistent encoding and temporal alignment mechanisms rather than merely increasing receptive field size or network depth.

6.3 Fall Detection Performance Evaluation

In the fall detection experiments, this study focuses on short-term variations in human posture as the primary cue for distinguishing fall events from daily activities. Based on the previously described feature design, the original 12-dimensional skeleton coordinates derived from six selected joints are transformed into a four-dimensional posture feature vector $\{\theta(t), \Delta x(t), \Delta y(t), y_{hc}(t)\}$, which consists of the trunk vector angle θ, the horizontal and vertical component differences Δx and Δy, and the vertical height of the hip center y_{hc}. This feature set effectively captures the trunk inclination, displacement, and height variation that characterize fall events, while preserving clear physical interpretability. All four features are time-varying signals and must be analyzed along the temporal axis to fully describe the dynamic nature of a fall. In general, a fall event is completed within approximately 2–3 s and is accompanied by pronounced posture changes and spatial displacement. In this study, image sequences are captured at a frame rate of 30 fps, generating 30 feature vectors per second. For practical implementation, a sliding window of 60 frames (approximately 2 s) is adopted to form a single motion sequence sample. Once 60 consecutive frames are accumulated, the first sample is generated; subsequently, each newly acquired frame is appended to the sequence while the oldest frame is discarded, continuously producing updated feature sequences. This sliding-window strategy allows the system to preserve the complete temporal evolution of fall events within a fixed time scale while enabling continuous and real-time fall detection.

Unlike gait identification, fall detection targets non-periodic events characterized by abrupt and pronounced posture transitions. The primary focus of fall detection therefore lies in capturing rapid changes in human skeletal configuration over consecutive image frames, rather than subtle periodic variations associated with walking cycles. Accordingly, the fall detection experiments in this study do not involve step segmentation or gait alignment. Instead, continuously extracted feature sequences formed by a sliding window are directly adopted as the basic analysis units. This design preserves the complete temporal evolution of fall motions while effectively increasing the number of available training samples. Moreover, it more closely reflects the operational characteristics of real-time fall monitoring systems in practical home-care environments. Based on this formulation, the proposed fall detection method is systematically evaluated under three classification settings, namely binary classification, seven-class classification, and five-class classification, in order to assess the recognition capability and robustness of the constructed vector–angle image representation across different application scenarios.

6.3.1 Binary Fall Detection Performance

First, we conducted a binary classification experiment to distinguish fall events from non-fall activities. The training dataset contains 591 fall samples and 4819 non-fall samples, resulting in a highly imbalanced class distribution. This imbalance is consistent with real-world home monitoring scenarios, where falls occur far less frequently than routine daily activities. Despite this challenging condition, the proposed CNN model achieves an average accuracy of 98%, as shown in Figure 17.

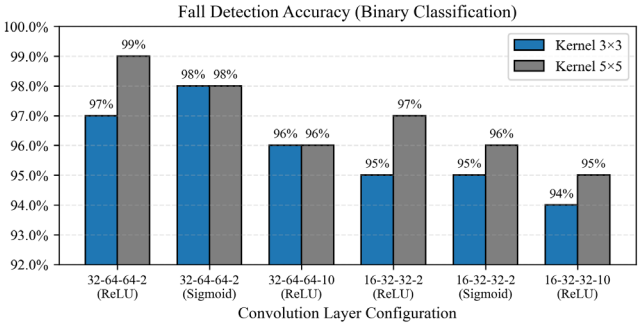

Figure 17. Binary fall detection accuracy for vector–angle features

This result suggests that fall events exhibit distinctive and consistent posture variations, such as rapid trunk descent, pronounced body inclination, and abrupt reductions in body height. These characteristics can be effectively captured by the proposed vector–angle image representation, enabling reliable discrimination between fall events and non-fall activities even under severely imbalanced training conditions.

6.3.2 Seven-Class Fall Detection Performance

Building upon the strong performance observed in the binary classification task, this study further extends the fall detection problem to a seven-class classification setting in order to evaluate the model's ability to simultaneously recognize daily activity states and their corresponding fall events. The seven classes include standing (1155 samples), standing-to-fall (266 samples), walking (968 samples), walking-to-fall (197 samples), sitting (1216 samples), sitting-to-fall (173 samples), and lying (1456 samples). This formulation requires the model to handle a broader range of daily postures as well as their associated fall transitions, thereby increasing both the classification complexity and the practical relevance of the fall detection task. Among these classes, standing, walking, sitting, and lying represent non-fall states, whereas standing-to-fall, walking-to-fall, and sitting-to-fall correspond to fall-related events. Despite the increased number of classes and the more complex data distribution, the proposed model achieves an average classification accuracy of approximately 96%, as shown in Figure 18.

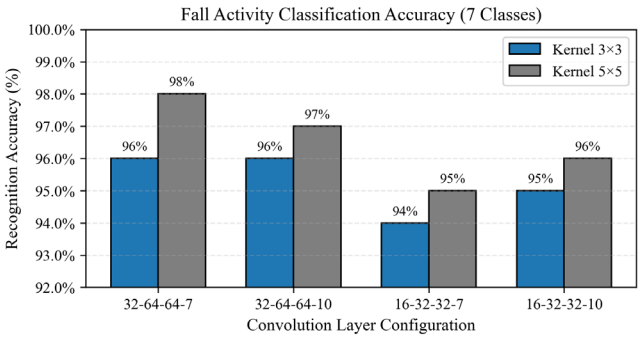

Figure 18. Seven-class fall classification accuracy for vector–angle features

The experimental results demonstrate that the proposed approach is capable of not only detecting fall events but also accurately distinguishing normal activity states from their corresponding fall transitions across different activity scenarios. This confirms that the vector–angle image representation exhibits strong robustness and discriminative capability in multi-class fall detection tasks.

6.3.3 Five-Class Fall Detection Performance

In the seven-class setting, the number of samples for each fall-related category is relatively limited, which may lead to class imbalance during training. From a practical deployment perspective, fall events occurring under different activity contexts can be treated as the same critical event in terms of alert generation and notification. Accordingly, this study further aggregates all fall-related samples into a single "fall" category, resulting in a five-class classification setting consisting of standing, walking, sitting, fall, and lying. Under this configuration, the training dataset includes 1155 samples for standing, 968 for walking, 1216 for sitting, 636 for fall, and 1456 for lying. The experimental results show that the proposed model achieves an average classification accuracy of approximately 96% in the five-class setting, which is the highest among the three evaluated configurations, as shown in Figure 19.

This result indicates that consolidating fall events into a single class effectively alleviates the class imbalance issue and allows the model to focus on the core posture transition characteristics shared by different types of falls. Moreover, the five-class setting better aligns with the requirements of practical fall detection systems, facilitating timely and reliable triggering of alert mechanisms in real-world monitoring applications.

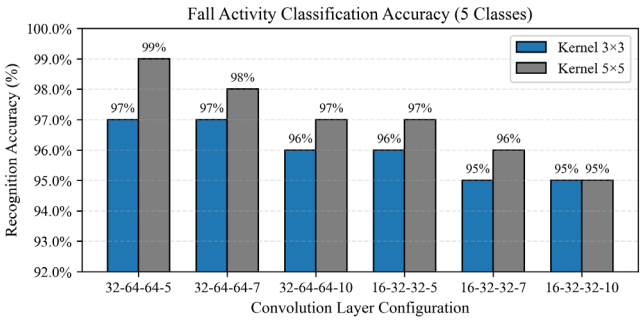

Figure 19. Five-class fall classification accuracy for vector–angle features

Overall, the experimental results presented in this chapter demonstrate that the proposed skeleton-based deep learning framework exhibits distinct performance characteristics in gait identification and fall detection tasks. In the gait identification task, when only a limited number of gait cycles represented by angle-plot images are used as input, the recognition performance is noticeably influenced by the inherent variability of natural human motion. In contrast, for the fall detection task, fall events are characterized by abrupt and consistent posture transitions, enabling the proposed vector–angle features to achieve stable and high recognition accuracy across different classification settings. These results indicate that the proposed feature representation offers superior robustness and applicability for fall-related event recognition.

7 Conclusions and Future Work

This study proposes a skeleton-based deep learning framework and conducts systematic experimental evaluations on two tasks: gait identification and fall detection. The experimental results indicate that, in the gait identification task, when angle-plot images derived from a single gait cycle or a fixed number of gait cycles are used as input features, the overall recognition accuracy ranges from approximately 72% to 82%. Moreover, variations in convolutional kernel sizes and convolutional layer configurations have a relatively limited impact on classification performance. These findings reflect the inherently high variability of human gait, in which angle-based features are easily influenced by individual walking habits, speed, stride length, and the stability of skeleton extraction, thereby imposing an upper bound on achievable identification performance. In contrast, for the fall detection task, the proposed vector–angle and component-difference features effectively capture the abrupt posture transitions and rapid height reductions that occur during fall events. Experimental results demonstrate that the proposed method consistently achieves recognition accuracies exceeding 96% under binary, five-class, and seven-class classification settings. In addition, the performance differences among various CNN architectures are relatively small, indicating that the proposed feature representation exhibits strong stability and discriminative capability. These results further confirm that, compared with the subtle and highly individualized variations in gait patterns, fall events present more consistent and distinguishable motion characteristics in the skeleton feature space.

Overall, the experimental results of this study indicate that although the proposed angle-based image representation can be applied to different human action recognition tasks, its performance is highly dependent on the intrinsic characteristics of the target action. For fine-grained and highly variable behaviors such as gait identification, relying solely on angle features extracted from a limited number of gait cycles inevitably imposes inherent performance limitations. In contrast, for fall detection, which is characterized by clear posture transitions and pronounced spatial displacement, the proposed skeleton-based features combined with a lightweight CNN architecture demonstrate strong recognition capability, highlighting their suitability and effectiveness for fall detection applications. The proposed fall detection approach consistently achieves recognition accuracies exceeding 90% under binary, five-class, and seven-class classification settings, indicating robust performance and high stability. Since fall events are typically accompanied by abrupt and well-defined posture changes, their corresponding features exhibit high separability in the image feature space, enabling the proposed vector–angle image representation to effectively support real-time recognition requirements. Owing to these properties, the proposed fall detection model can serve as a reliable event decision module in practical systems and is well suited for integration into IoT-based real-time notification frameworks. By combining edge-side real-time inference with network communication modules, fall events can be promptly transmitted to caregivers, thereby enhancing the timeliness and practical value of home-care safety monitoring systems.

Based on the proposed skeleton-based deep learning framework for fall detection, several research directions remain worthy of further extension and enhancement, as described below. (1) At the feature modeling level, this study has demonstrated that low-dimensional skeleton features constructed from vector–angle relationships and component differences can effectively capture the abrupt posture transitions associated with fall events. However, the current feature design primarily focuses on the geometric relationships of the human trunk. Future work may further incorporate additional skeleton-based information related to balance control, such as joint velocities, angular velocities, or acceleration features, to enhance the model's capability in predicting imminent falls or atypical fall patterns. Such extensions would enable the system to evolve from post-event recognition toward early warning and pre-fall detection. (2) With respect to temporal modeling and sequence representation, this study adopts a fixed-length sliding window combined with image-based encoding to balance real-time performance and recognition accuracy. Future research may explore variable-length sequence modeling or the integration of attention mechanisms, allowing the model to automatically focus on the most discriminative temporal segments during fall events. This could reduce the influence of redundant information on classification decisions and improve adaptability to falls with different speeds and motion characteristics. (3) In terms of system deployment, this study has completed an integrated validation of skeleton extraction, deep learning inference, and IoT communication; however, the current implementation primarily targets scenarios involving a single camera and a single subject. Future extensions may include multi-camera or multi-view architectures to mitigate the impact of occlusion on skeleton extraction stability and to enhance system reliability in real-world home environments. In addition, for scenarios involving multiple users, integrating multi-person skeleton tracking and identity separation mechanisms would allow fall events to be accurately associated with specific individuals, thereby improving system applicability in long-term care facilities or public spaces. (4) From a smart healthcare application perspective, future work may integrate the proposed fall detection system with other physiological or environmental sensing data, such as wearable devices, physiological signals, or ambient home sensors, to establish a more comprehensive multimodal care platform. Through cross-sensor data fusion and joint analysis, such a platform could further improve the reliability of fall detection, support long-term behavioral pattern analysis, and facilitate health risk assessment, thereby providing a more forward-looking and practical solution for intelligent home-care systems.

Finally, it should be noted that the dataset used in this study is self-collected and limited in size, which may affect

the generalizability of the proposed method. Therefore, this study is mainly intended as a proof-of-concept to demonstrate the feasibility of the proposed skeleton-based framework. Future work will expand the dataset and validate the proposed approach on larger and publicly available datasets to further improve its robustness and practical applicability.

References

[1] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, A. Blake, Real-Time Human Pose Recognition in Parts from Single Depth Images, *IEEE Conference on Computer Vision and Pattern Recognition*, Colorado Springs, CO, USA, 2011, pp. 1297–1304. https://doi.org/10.1109/CVPR.2011.5995316

[2] M. Ye, X. Wang, R. Yang, L. Ren, M. Pollefeys, Accurate 3D Pose Estimation from a Single Depth Image, *IEEE International Conference on Computer Vision*, Barcelona, Spain, 2011, pp. 731–738. https://doi.org/10.1109/ICCV.2011.6126310

[3] Y. Li, Hand Gesture Recognition Using Kinect, *IEEE International Conference on Computer Science and Automation Engineering*, Beijing, China, 2012, pp. 196–199. https://doi.org/10.1109/ICSESS.2012.6269439

[4] P. S. Huang, C. J. Harris, M. S. Nixon, Recognising Humans by Gait via Parametric Canonical Space, *Artificial Intelligence in Engineering*, Vol. 13, No. 4, pp. 359–366, October, 1999. https://doi.org/10.1016/S0954-1810(99)00008-4

[5] A. Núñez-Marcos, G. Azkune, I. Arganda-Carreras, Vision-Based Fall Detection with Convolutional Neural Networks, *Wireless Communications and Mobile Computing*, Vol. 2017, No. 1, pp. 1–16, December, 2017. https://doi.org/10.1155/2017/9474806

[6] B. Kwolek, M. Kepski, Improving Fall Detection by the Use of Depth Sensor and Accelerometer, *Neurocomputing*, Vol. 168, pp. 637–645, November, 2015. https://doi.org/10.1016/j.neucom.2015.05.061

[7] Z. Cao, G. Hidalgo, T. Simon, S. E. Wei, Y. Sheikh, OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 43, No. 1, pp. 172–186, January, 2021. https://doi.org/10.1109/TPAMI.2019.2929257

[8] M. Salimi, J. J. M. Machado, J. M. R. S. Tavares, Using Deep Neural Networks for Human Fall Detection Based on Pose Estimation, *Sensors*, Vol. 22, No. 12, Article No. 4544, June, 2022. https://doi.org/10.3390/s22124544

[9] M. Dutt, A. Gupta, M. Goodwin, C. W. Omlin, An Interpretable Modular Deep Learning Framework for Video-Based Fall Detection, *Applied Sciences*, Vol. 14, No. 11, Article No. 4722, June, 2024. https://doi.org/10.3390/app14114722

[10] Y. Wang, J. Sun, J. Li, D. Zhao, Gait Recognition Based on 3D Skeleton Joints Captured by Kinect, *IEEE International Conference on Image Processing*, Phoenix, AZ, USA, 2016, pp. 3151–3155. https://doi.org/10.1109/ICIP.2016.7532940

[11] P. Zhang, C. Lan, J. Xing, W. Zeng, J. Xue, N. Zheng, View Adaptive Recurrent Neural Networks for High Performance Human Action Recognition from Skeleton Data, *arXiv preprint*, arXiv:1703.08274, pp. 1-10, August, 2017. https://arxiv.org/abs/1703.08274

[12] Y. Du, W. Wang, L. Wang, Hierarchical Recurrent Neural Network for Skeleton Based Action Recognition, *IEEE Conference on Computer Vision and Pattern Recognition*, Boston, MA, USA, 2015, pp. 1110–1118. https://doi.org/10.1109/CVPR.2015.7298714

[13] H. Wang, L. Wang, Modeling Temporal Dynamics and Spatial Configurations of Actions Using Two-Stream Recurrent Neural Networks, *IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, HI, USA, 2017, pp. 3633–3642. https://doi.org/10.1109/CVPR.2017.387

[14] S. Yan, Y. Xiong, D. Lin, Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition, *AAAI Conference on Artificial Intelligence*, New Orleans, LA, USA, 2018, pp. 7444–7452.

[15] L. Shi, Y. Zhang, J. Cheng, H. Lu, Two-Stream Adaptive Graph Convolutional Networks for Skeleton-Based Action Recognition, *IEEE Conference on Computer Vision and Pattern Recognition*, Long Beach, CA, USA, 2019, pp. 12018–12027. https://doi.org/10.1109/CVPR.2019.01230

[16] P. Vallabh, R. Malekian, N. Ye, D. C. Bogatinoska, Fall Detection Using Machine Learning Algorithms, International Conference on Software, *Telecommunications and Computer Networks (SoftCOM)*, Split, Croatia, 2016, pp. 1–9. https://doi.org/10.1109/SOFTCOM.2016.7772142

[17] C. Nadee, K. Chamnongthai, Ultrasonic Array Sensors for Monitoring of Human Fall Detection, I*nternational Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON)*, Hua Hin, Thailand, 2015, pp. 1–4. https://doi.org/10.1109/ECTICon.2015.7207097

[18] W. Chen, Z. Jiang, H. Guo, X. Ni, Fall Detection Based on Key Points of Human-Skeleton Using OpenPose, *Symmetry*, Vol. 12, No. 5, Article No. 744, May, 2020. https://doi.org/10.3390/sym12050744

[19] D. Soman, R. P. Singh, N. Prithika, M. Siri, S. Kumar, A Novel Fall Detection System Using MediaPipe, *International Conference on Circuits, Control, Communication and Computing (I4C)*, Bangalore, India, 2022, pp. 336–340. https://doi.org/10.1109/I4C57141.2022.10057642

[20] C. B. Lin, Z. Dong, W. K. Kuan, Y. F. Huang, A Framework for Fall Detection Based on OpenPose Skeleton and LSTM/GRU Models, *Applied Sciences*, Vol. 11, No. 1, Article No. 329, January, 2021. https://doi.org/10.3390/app11010329

[21] C. A. Q. Bugarin, J. M. M. Lopez, S. G. M. Pineda, M. F. C. Sambrano, P. J. M. Loresco, Machine Vision-Based Fall Detection System Using MediaPipe Pose with IoT Monitoring and Alarm, *IEEE Region 10 Humanitarian Technology Conference (R10-HTC)*, Hyderabad, India, 2022, pp. 269–274. https://doi.org/10.1109/R10-HTC54060.2022.9929527

[22] S. S.-D. Xu, C.-H. Chen, T.-C. Chang, Design of oneM2M-based fog computing architecture, *IEEE Internet of Things Journal*, Vol. 6, No. 6, pp. 9464–9474, December, 2019. https://doi.org/10.1109/JIOT.2019.2929118

[23] S. S.-D. Xu, I. G. D. Pancev, M. Y. Baihaqi, Design and implementation of a 6LoWPAN-based lightweight wireless

embedded Internet platform for IoT applications, *Journal of Information Technology*, Vol. 24, No. 2, pp. 323–332, March, 2023.
https://doi.org/10.53106/160792642023032402011

[24] A. Alarifi, A. Alwadain, Killer heuristic optimized convolution neural network-based fall detection with wearable IoT sensor devices, *Measurement*, Vol. 167, Article No. 108258, January, 2021.
https://doi.org/10.1016/j.measurement.2020.108258

[25] S. H. Yusuf, S. S.-D. Xu, G. N. Wedajew, C.-H. Chang, Convolutional neural network-based single-image contrast enhancement using multi-exposure training, *Digital Signal Processing*, Vol. 163, Article No. 105221, August, 2025.
https://doi.org/10.1016/j.dsp.2025.105221

[26] P.-S. Tsai, T.-F. Wu, W.-H. Chen, Generalized vision-based coordinate extraction framework for EDA layout reports and PCB optical positioning, *Processes*, Vol. 14, No. 2, Article No. 342, January, 2026.
https://doi.org/10.3390/pr14020342

[27] P.-S. Tsai, T.-F. Wu, J.-Y. Chen, J.-F. Huang, Integrating of Image Processing and Number Recognition in Sudoku Puzzle Cards Digitation, *Journal of Internet Technology*, Vol. 23, No. 7, pp. 1573–1584, December, 2022.
https://doi.org/10.53106/160792642022122307012

[28] P.-S. Tsai, T.-F. Wu, C.-T. Liao, Development of a robotic manipulator for piano performance via numbered musical notation recognition, *Machines*, Vol. 13, No. 12, Article No. 1121, December, 2025.
https://doi.org/10.3390/machines13121121

[29] P.-S. Tsai, T.-F. Wu, Y.-C. Wang, Automatic quadrotor dispatch missions based on air-writing gesture recognition, *Processes*, Vol. 13, No. 12, Article No. 3984, December, 2025.
https://doi.org/10.3390/pr13123984

Biographies

Pu-Sheng Tsai was born in Taiwan, R.O.C., in 1962. He received the M.S. degree in automatic control from the Feng Chia University, Taichung, Taiwan, R.O.C., in 1985 and the Ph.D. degree in electrical engineering from the National Taiwan University, in 1998. He is currently an Associate Professor in the Department of Electrical Engineering at Ming Chuan University, Taoyuan, Taiwan. His main research interests are artificial intelligence, virtual reality somatosensory interaction, internet of things, Embedded Micro-controller application and design.

Ter-Feng Wu was born in Taiwan in 1962. He received the B.S. degree in Department of Industrial Education from National Taiwan Normal University, Taipei, Taiwan, in 1986. He received the M.S. degree in Department of Control Engineering from National Chiao Tung University, Hsinchu, Taiwan, in 1993. He received the Ph.D. degree in Department Electrical Engineering from National Taiwan University, Taipei, Taiwan, in 2006. He is currently a Distinguished Professor and also serves as the Dean of College of Electrical Engineering & Computer Science, National Ilan University, Yilan, Taiwan. His research interests include intelligent control, neural network, fuzzy CMAC, unmanned aerial vehicles (UAVs), green energy and mobile robot, etc.