Soft Computing (2023) 27:5003–5019
https://doi.org/10.1007/s00500-022-07582-y

APPLICATION OF SOFT COMPUTING

Intelligent attendance monitoring system with spatio-temporal human action recognition

Ming-Fong Tsai[1] · Min-Hao Li[1]

Accepted: 7 October 2022 / Published online: 28 October 2022

Abstract
This paper proposes an intelligent attendance monitoring system based on spatio-temporal human action recognition, which includes human skeleton gait recognition, multi-action body silhouette recognition and face recognition. Our system solves several problems, for example, when a mask is worn to conceal the face, which leads to a decrease in recognition accuracy performance, and when a 3D face mask is used to fake an identity. The skeleton gait feature of our intelligent attendance monitoring system uses a temporal weighted K-nearest neighbours algorithm to train the recognition model and carry out identification, while the multi-action body silhouette feature uses a multiple K-nearest neighbours algorithm to train the recognition model, identify the person and vote on the outcome. Using the proposed system, which integrates skeleton gait features, action silhouette features and face features, more effective recognition can be achieved. When the system encounters a situation with feature masking, such as when an individual is wearing a mask or has changed their clothes, or when the viewing angle is masked, it can continue to deliver good recognition ability through multi-angle skeleton synthesis gait recognition. Our experimental results show that the recognition accuracy of the system is 83.33% when a specific person wears a mask and passes through a monitored area. The intelligent attendance monitoring system uses a LINE messaging API as the access control notification function and provides a responsive web platform that allows managers to perform follow-up management and monitoring.

Keywords Intelligent attendance monitoring · Skeleton gait recognition · Silhouette recognition · Face recognition

1 Introduction

An access control system can efficiently manage the entries and exits of specific people within a given area, as it can identify a person in real time and record information in a cloud database system, which allows for the analysis of the huge amounts of data that can be collected in this way. This means that managers are able to fully understand the changes in the movements of specific people and track the movements of a specific person entering and exiting an area through the realisation of an intelligent attendance monitoring system for accurate personnel identification. Traditional access control systems use radio frequency identification to identify a specific person entering and exiting an area. In recent years, in order to effectively solve the problem of using radio frequency identification technology as a substitute for check-in, access control systems were designed that used deep learning face recognition technology to identify a specific person entering and exiting a given area. To construct and implement an access control system based on deep learning face recognition technology, the facial image features of the specific people entering and exiting the area must be obtained in advance as the basis for classification. Related studies have used the Faster R-CNN deep learning object detection model to train the face recognition model and identify people (Tsai and Li 2021; Ren et al. 2017). This approach uses a camera device to capture images of specific people entering the monitored area in real time, to enable a fast identification process. Not only can this effectively solve the problem of check-in as a substitute for a physical radio frequency identification card, but it can also prevent the risk of infection by a bacteria or virus due to contact with the radio

✉ Ming-Fong Tsai
mingfongtsai@gmail.com

[1] Department of Electronic Engineering, National United University, Miaoli, Taiwan

frequency identification card reader when entering and exiting an area. A related study proposed a multi-angle facial feature classification and recognition method to solve the problem of occlusion of the viewing angle arising from image capture by a single camera (Shakhnarovich et al. 2001). A multi-angle flat face image was projected onto a 3D head cylinder, and the rigid body motion theorem was then applied to handle the deformation problem of stitching together multiple multi-angle face images (Kong et al. 2005). A multi-angle flat face image stitching technique was used to improve the recognition accuracy of the access control system based on deep learning face recognition. However, due to the recent rapid developments in deep learning face recognition technology, this approach remains vulnerable to people who aim to fake identities by re-photographing facial images to pass through the access control system, even if subtle, multi-angle human facial features are used for feature classification and recognition. The use of a 3D face mask to fake an identity is a problem that has yet to be solved in the context of access control systems. At the same time, as the COVID-19 pandemic rages around the world, the wearing of masks has become one way to avoid the spread of infection. Access control systems based on deep learning face recognition technology often have reduced performance in terms of recognition accuracy due to the problem of masks covering faces (Cheema and Moon 2021).

With an access control system based on the above-mentioned deep learning face recognition technology, it is easy to use 3D printing technology to copy and simulate a face in order to bypass an identity authentication check. One study (Wang et al. 2004) has proposed that the way in which the human body moves and walks, otherwise known as gait, is a more difficult aspect of identity to replicate than fingerprints and iris recognition features and that capturing the gait of the human body can allow dynamic identification feature recognition to be performed in a long-distance and non-invasive way. A previous study (Wang et al. 2003) proposed an identification model training sample set for human gait identification, using gait silhouette map information. However, this information is susceptible to changes in the shape of the human body, which makes it impossible to recognise the identity features. For example, the gait silhouette map of a person wearing a heavy coat and carrying hand luggage will show a large difference from the normal silhouette. Hence, improving the gait silhouette map information can be treated as a dynamic identification feature recognition problem. In previous research, human skeleton key point detection and drawing technology has been used to obtain dynamic human posture information. In addition, in order to obtain dynamic identification feature information for human movement and walking gait recognition, one study (Yan et al. 2018) has proposed a Spatial Temporal Graph Convolutional Networks (ST-GCN) model that can be used to train and identify the human motion state recognition model. In this approach, OpenPose technology is used to process the continuous images of the states of human motion and to generate skeleton key point information for posture recognition. The ST-GCN model uses the key point information of the human skeleton to perform convolution calculations in the spatial and temporal dimensions. The changes in the skeleton key point information over consecutive frames in the space and time dimensions are used for deep learning of the state of motion of the human body. The Graph Convolutional Network (GCN) and Temporal Convolution Network (TCN) convolutional neural network architectures are integrated to form a deep learning recognition model training and identify with spatial and temporal. Since the ST-GCN model is manually adjusted to the topology of the human skeleton using a graph convolutional neural network, this reduces the recognition accuracy of the deep learning recognition model, and the Two-Stream Adaptive Graph Convolutional Networks (2S-AGCN) (Shi et al. 2019) model was therefore proposed to allow for deep learning of the state of motion of the human body. This model applies the concept of negative feedback to the deep learning process to achieve dynamic adjustment effects and uses J-stream to process the key points of the human skeleton and B-stream to process the connection information between them. Consider the vector change of the above-mentioned two-stream information as the training and identify of the human body motion state deep learning recognition model. In another paper, the ST-GCN system was introduced to perform GCN processing, which provides an embedding matrix to encode the connection relationships between the key points of the human skeleton and is suitable for all graph convolutional neural network layers. The Graph Convolutional Network Neural Architecture Search (GCN-NAS) (Peng et al. 2020) system was proposed in one study to dynamically adjust the embedding matrix, and this model uses multiple embedding matrices in each layer of the convolutional neural network to obtain the best training results. In order to reduce the number of calculations involved, the Neural Architecture Search (NAS) algorithm was developed with the aim of rapidly obtaining the best solution.

However, the use of only a single camera for continuous image shooting and framing will cause problems in terms of the viewing angle and body occlusion, which will reduce the accuracy of the dynamic posture information used in the detection and drawing of the key points of the human skeleton, thereby reducing the training and identification accuracy of the motion state recognition model. In this paper, we therefore propose an intelligent attendance monitoring system that can recognise both facial images

(static) and motion (dynamic) using spatial and temporal information. It uses multiple cameras for multi-directional continuous image capture to avoid problems with the viewing angle and body occlusion and relies on multi-directional facial images to perform face recognition to prevent people from using single face images to fake identities and to bypass the access control system. Multi-directional continuous images are used to detect the key points of the human skeleton, and this information is then fused into the multi-directional synthetic dynamic human posture information to train the recognition model and identify the state of motion of the human body. This approach prevents people from using 3D printing technology to fake an identity. At the same time, it is possible to obtain good recognition accuracy regardless of whether the people entering or exiting the monitored area are wearing masks. Moreover, when the silhouette changes, such as when the individual is wearing a jacket, this leads to abnormal features; however, the skeleton can be effectively identified without being affected.

The proposed intelligent attendance monitoring system with spatio-temporal human action recognition is based on the continuous changes in the spatial and temporal features of the key points of the skeleton and integrates gait features with human body silhouette feature recognition technology. Multiple cameras are used for information fusion to ensure that the key points of the human skeleton do not cause masking errors due to the natural swing due to walking. Our approach is mainly based on the following two aspects for the extraction of the human body skeleton key point information feature:

- Against human body skeleton 17 key points for continuous time perform spatial relative change feature extraction. We treat each set of three key points of the skeleton as a unit to relative angle information formed plural continuous angle changes as specific action feature information. We also treat each pair of key points of the skeleton as a unit to relative distance information formed plural continuous distance changes as specific action feature information. The above-mentioned multiple continuous relative angle and distance information is used by the K-Nearest Neighbours (KNN) algorithm recognition model with temporal and spatial weight to training and identify.
- Against the above-mentioned plural continuous angle changes as the specific motion feature information, when the system determines that the key points of the human body skeleton are carrying out a specific motion, a Mask R-CNN is applied to capture multiple forms of continuous body silhouette information from the continuous motion. The system separately trains and identifies the KNN algorithm recognition model for the multiple continuous human silhouette information of the specific action. The spatio-temporal human action recognition system generates the final recognition result by voting, based on the above-mentioned multiple recognition model. Our approach extends previous related work in which face recognition was carried out by an attendance monitoring system based on a Faster R-CNN deep learning object detection model. Change to the continuous human body gait features recognition model of specific actions and the integration of multiple human body silhouette feature recognition model, voting on plural recognition models to obtain the final recognition results. This approach solves several problems such as the use of a 3D face mask to fake an identity and the reduction in the accuracy of the recognition model due to a mask covering the face. An access control notification function based on a LINE messaging API has also been added to our intelligent attendance monitoring system, with a responsive web page display.
- The following section presents background information and a review of related work. Section 3 introduces the proposed intelligent attendance monitoring system with face (static) and motion (dynamic) recognition based on spatial and temporal information, which integrates both gait and silhouette feature recognition technology. Section 4 describes the experimental method used in the paper and presents the experimental results and a performance analysis, while Sect. 5 contains the conclusion and suggestions for future work.

2 Background and related work

In this section, we describe the relevant technologies required for the proposed system. We discuss the Posenet architecture (Papandreou et al. 2018) and the Faster R-CNN (Ren et al. 2017) and Mask R-CNN (He et al. 2017) deep learning object detection models and review prior technologies that are similar to the proposed system. We also review the STV-GCN (Tsai and Chen 2021), GCN-NAS (Peng et al. 2020) and 2S-AGCN (Shi et al. 2019) deep learning motion detection systems, which are used in our performance analysis and experimental comparison.

2.1 Posenet skeleton key point detection

Our system applies Posenet (Papandreou et al. 2018) detection technology for the key points of the human skeleton, as proposed in a previous study. A total of 17 key points on the skeleton are used to carry out the human

action detection and action pose drawing functions. The key point detection method uses a GCN model for training and identification, and a CNN is used to capture the heatmap information for each skeleton key point in the image. The possible positions and a short-range offset feature are used to calculate the heatmap error, and Hough voting is then applied via an integrated voting function to obtain the lowest distortion of the skeleton key points stored in Hough arrays, based on the Hough score. The heatmap process divides the image into a 28×28 grid, calculates the probability of key points in each area and accurately corrects the coordinates through the use of a short-range offset. This approach is supplemented by the use of a mid-range offset feature to detect the links between the key points on the skeleton in order to reduce the distortion in the key points and improve the recognition accuracy, as shown in Fig. 1.

2.2 Faster R-CNN deep learning object detection model

The Faster R-CNN (Ren et al. 2017) deep learning object detection model includes Region Proposal Network (RPN) as the recognition box capture architecture, which generates detection box ranges of different proportions and sizes for different anchors and classifies the content of multiple detection boxes to obtain a detection box with high reliability, which is output as the recognition result. As shown in Fig. 2, RPN will be performed to generate region proposals after the image has been convolved. Based on the output region, the bounding box will be generated, and then, multiple identification boxes of different sizes will be generated in the middle of the box to obtain the correct identification target. Region of interest pooling is applied to the frame selection target to obtain the identification result. Faster R-CNN uses parallel processing via the Convolutional Neural Network (CNN) and RPN and combines them to form the object recognition model. For example, the access control system produces the recognition results based on face recognition and face box selection.

2.3 Mask R-CNN deep learning object detection model

Our scheme uses the Mask R-CNN (He et al. 2017) deep learning object detection model proposed in related work to identify the silhouette features of the human body. We use the semantic segmentation technology of the Mask R-CNN deep learning target detection model to obtain information on the silhouette. Our Mask R-CNN deep learning object detection model is based on a traditional Faster R-CNN model and combines the semantic segmentation algorithm with the FCN architecture. The first stage of the model uses a standard CNN to learn the image features; in the second stage, deconvolution feedback is applied to dynamically adjust the learning parameters, and the feature maps that have been classified are interpolated to achieve deconvolution learning. The final output of the system is a semantic segmentation map that is classified for each pixel. The use of semantic segmentation to give silhouette information allows us to obtain a silhouette map of the human body, as shown in Fig. 3. For the region of interest, a branch is generated for the segmentation mask. Each branch uses a small fully convolutional network to predict the segmentation mask in a pixel-to-pixel manner.

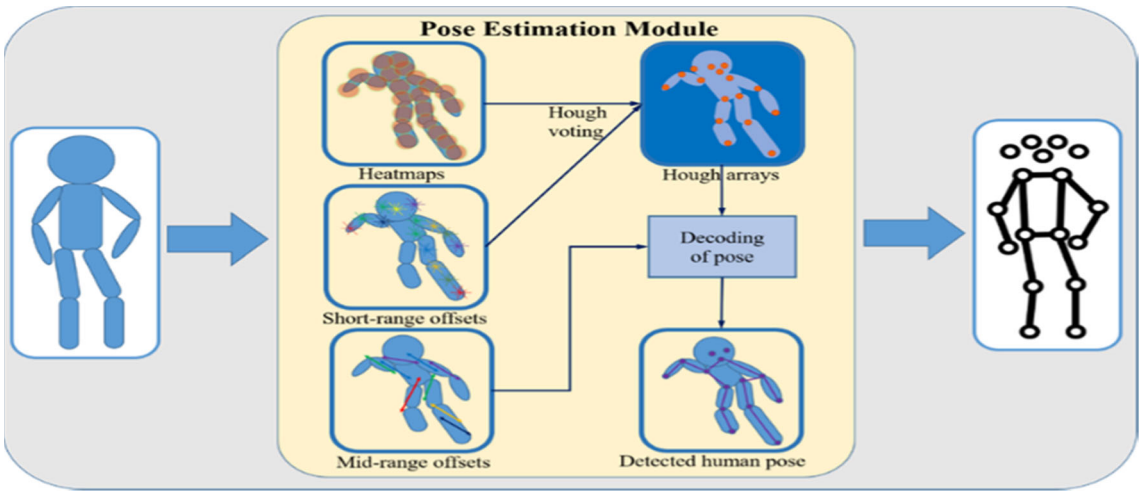


Fig. 1 Pose estimation based on detection of human skeleton key points

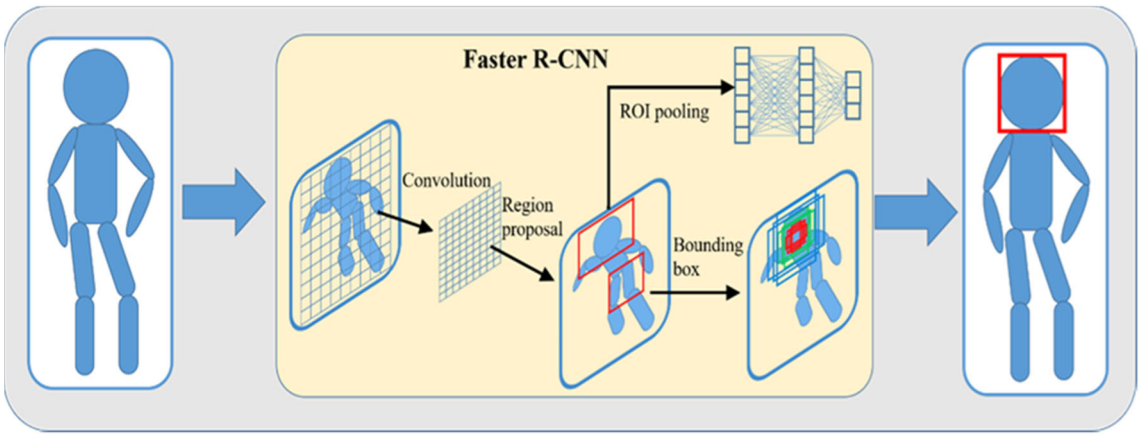


Fig. 2 Facial recognition architecture based on Faster R-CNN

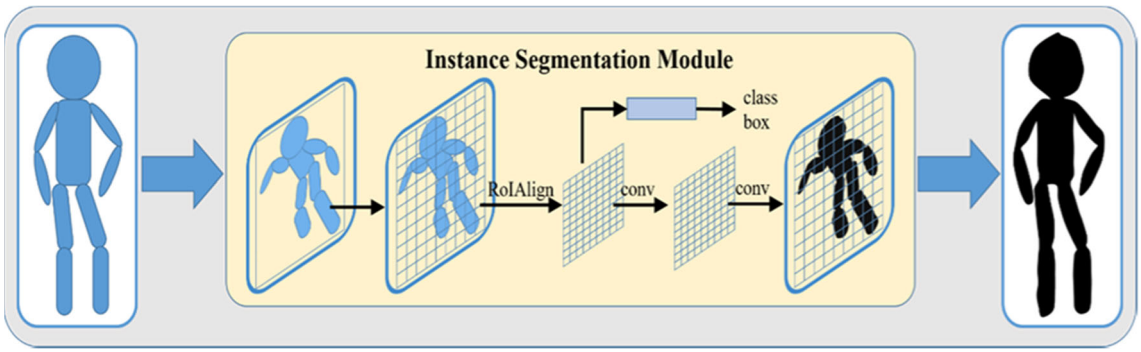


Fig. 3 Silhouette recognition architecture using Mask R-CNN

2.4 STV-GCN deep learning action detection system

The STV-GCN (Tsai and Chen 2021) deep learning action detection system was developed in a prior study to train and identify human motion recognition models in the form of non-traditional image files. The information used for training and identification by the deep learning motion detection system is the key point features of the human body skeleton. STV-GCN is a human motion recognition system that combines the KNN algorithm with an ST-GCN deep learning motion detection system. The GCN used in the ST-GCN model performs pattern recognition that is not limited to traditional two-dimensional graphs: it can be trained to recognise topological graphs or three-dimensional graphs composed of points and lines. The recognition model applies graph convolutional neural network learning to the spatial and temporal changes in the key points of the human skeleton. Its architecture relies on spatial GCN and TCN. The graph convolutional neural network uses nine alternately overlapping spatial and temporal feature extraction layers for computational learning, and a fully connected layer is finally applied to classify the action features of the key points of the human skeleton. In a prior study, human actions were classified based on different emotions and different speeds of motion, and good recognition accuracy was obtained. Our access control system involves the application of deep learning motion detection technology to the recognition of human gait. That is, the training of the recognition model and identification are carried out based on the spatial and temporal changes in the key points of the skeleton generated by walking and other types of movement. However, the recognition accuracy of the recognition model that the current deep learning motion detection system performs the same motion on different human bodies still needs to be strengthened.

2.5 GCN-NAS deep learning action detection system

In order to strengthen the recognition performance of the ST-GCN deep learning action detection model, a prior study added a NAS function in the model training stage, with the main aim of optimising the learning network structure. In the search space stage, reinforcement learning was used to find the model with the highest recognition accuracy. Traditional GCN convolution operations all use a

stable embedding matrix (a first-order Chebyshev polynomial) to control the correlation between the nodes of the topology graph. The developers of GCN-NAS used multi-order Chebyshev polynomials to generate multiple embedding matrices for convolution operations. The use of a neural architecture search function to find the best solution in the search space stage can avoid long learning and calculation times. It was shown in one study that GCN-NAS (Peng et al. 2020) still exceeded the ST-GCN deep learning motion detection system in terms of the training time of the model. This motion detection model can be applied to identify different types of motion and has achieved good recognition accuracy. However, the recognition accuracy of the recognition model that the current deep learning motion detection system performs the same motion on different human bodies still needs to be strengthened.

2.6 2S-AGCN deep learning action detection system

The 2S-AGCN (Shi et al. 2019) deep learning action detection system was developed with the aim of strengthening the GCN convolution operation weight of ST-GCN. In order to improve the traditional GCN convolution operation, which uses a stable embedding matrix to control the correlation between the nodes of the topology graph, a two-stream convolutional network architecture was developed. One of the streams used three embedding matrices as the weight change of the convolution operation: the first had the original form, the second was used to strengthen the learning of the correlations between nodes, and the last used a Gaussian embedding function to capture the relationships between nodes. These three different embedding matrices were used to perform an integrated calculation, and obtained the convolution operation weights more efficiently. The other stream focused on learning the connection relationships between nodes and use motion detection model training input the connections between the topological graph nodes to perform convolution operations, with the final output being the correlation features. The training time for 2S-AGCN in a motion detection model has been shown in prior work to be longer than for GCN-NAS. This motion detection model can be applied to identify different types of motion and obtains good recognition accuracy. However, the recognition accuracy of the recognition model of the current deep learning action detection system performing the same action on different human bodies still needs to be strengthened. The system proposed in this paper can effectively address the problems identified in the above-mentioned related work.

2.7 Human emotion recognition using ST-GCN

One study (Tsai and Chen 2022) in the literature proposed an emotional action recognition system in which ST-GCN was applied to the human skeleton to achieve recognition of different emotional actions. However, due to the loss of subtle features caused by the use of convolutional neural network technology in the training process, it is proposed to extract facial action features to further improve the recognition effect. The swing and degree of change in the characteristics of the human face are analysed. As shown in Fig. 4, we record the change in the up-down and left–right swings of the face, capture the continuous changes along the two axes of the face and use the K-nearest neighbours classifier for classification and identification, combining the two identification results to achieve better identification. However, for the recognition of human walking movements, the scheme in the literature cannot distinguish the subtle differences between different people with the same emotion, resulting in a low recognition ability for gait, and it cannot analyse facial changes when a person is wearing a mask.

2.8 Human action recognition system using skeleton point correction

When using a skeleton for human action recognition, a system is often limited by the image shooting angle and visual occlusion, which leads to misjudgement of the key points of human skeleton and affects the accuracy of action recognition. The study in Tsai and Huang (2022) proposed an action recognition system that included key point correction of the human skeleton, and the recognition accuracy was improved by corrections to the skeleton. A basic correction algorithm is used to correct points based on the symmetry of the human body, as shown on the left of Fig. 5, while an advanced correction algorithm is used to correct keypoints based on the range of the human shield map, as shown on the right of Fig. 5. In view of the problems with skeleton masking, in this paper, we use skeletons from different perspectives for synthesis and obtain the correct continuous changes in the skeleton positions by synthesising the skeleton. Compared with

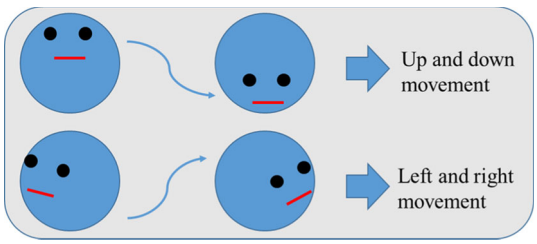


Fig. 4 Schematic diagram of face swing

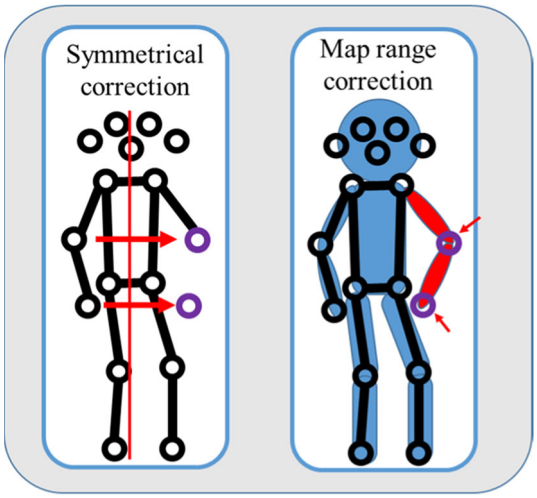


Fig. 5 Schematic diagram of skeleton point correction

prior schemes, we use skeleton features from different perspectives with higher correlation and achieve a better identification effect.

3 Using spatio-temporal human action recognition in an intelligent attendance monitoring system

We propose an intelligent attendance monitoring system that combines continuous human gait features with silhouette feature recognition technology to perform plural recognition model voting. We also use multiple cameras for multi-directional continuous image capture to avoid problems associated with the angle of view and body occlusion, to improve the overall recognition accuracy. In this section, we describe the system architecture, the data collection process, the training of the recognition model and the overall process of the proposed system.

3.1 System architecture

In view of the fact that training of our human movement recognition model and identification of a specific person is based on gait and silhouette features, the image frame of the human body movement process must reduce the problems related to viewing angle and body occlusion. Our system uses multiple cameras to capture image frames representing the motion of the human body. We use real-time images captured by multiple cameras to position the human body, to ensure that a specific person is located in the correct recognition area. The images captured by multiple cameras are also used to draw the key points of the human skeleton and to synthesise the gait or action based on these key points. We can use a dual camera as an example. This system captures dynamic images of the user while walking: the left camera captures dynamic images of the left side of the face and body, while the right camera captures images of the right side. The skeleton key points in the left body image are combined with those in the images of the right side, and the system identifies the key points of the synthesised skeleton to carry out motion recognition for a specific person. The motion recognition process can be divided into the following three main actions:

- The system performs specific identification of human body gait features, uses the Posenet human body skeleton key point detection model to obtain the skeleton key point information from each image and calculates the continuous angle and distance changes based on the key point information from the continuous image. This information is used as an identification feature for the time-sequenced gait of a specific person and to carry out classification training of the KNN recognition model and identification with weighted parameters.
- The system recognises a specific person based on the silhouette features of the human body and uses the continuous changes in the angles of the skeleton key points to determine the periodicity of the gait of the person. Multiple time points of the same angle are used as the basis of sampling for a time-sequenced gait. The Mask R-CNN deep learning object detection model is used to generate time-sequenced multiple silhouette feature information and then to perform classification training of the KNN recognition models and identification of the silhouette features. Finally, the identification results from the multiple KNN recognition models are submitted to a voting process to determine the prediction results for a specific person.
- The system recognises a specific person based on the symmetry features of the human face and carries out face recognition from the face images captured by the multiple cameras. It then confirms that the angle and position state of the face recognition result conform to the principle of the left–right symmetry characteristic of the human face. The Haar facial feature cascade classifier is applied to determine whether a human face exists, and this stage uses a Faster R-CNN deep learning object detection model as the network architecture. In future work, this module will be replaced by a new network architecture or deep learning object detection model with higher recognition accuracy and will be supplemented by the use of continuous human gait and silhouette feature recognition technology for specific actions to obtain the final recognition result.

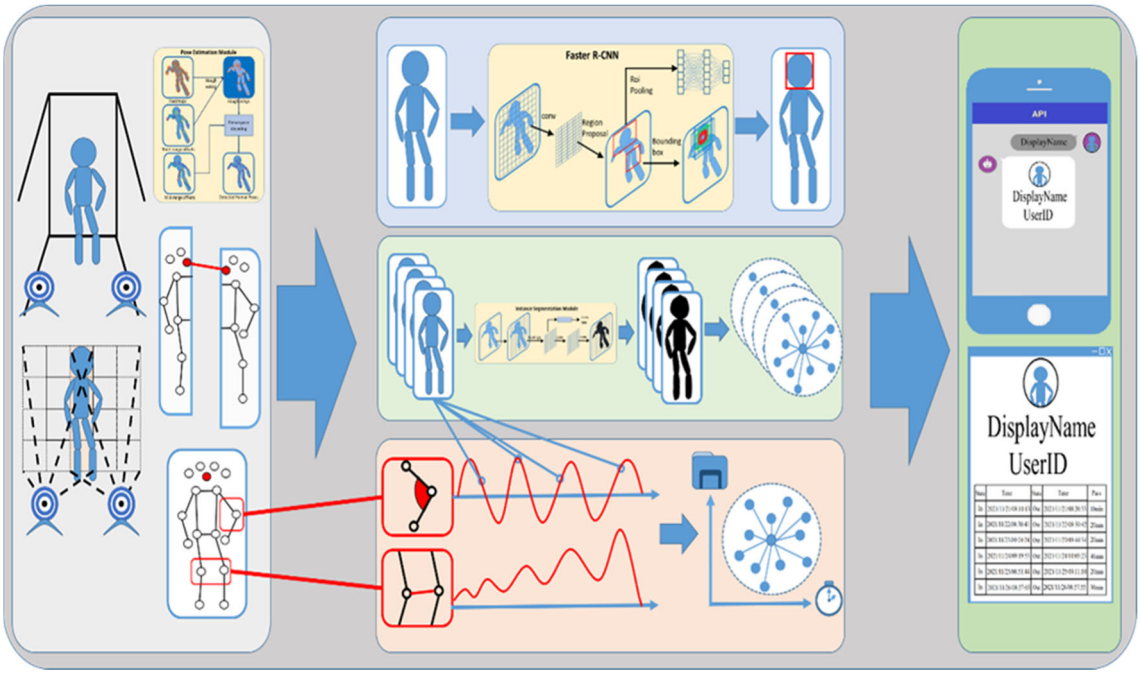


Fig. 6 Overview of the architecture of the proposed system

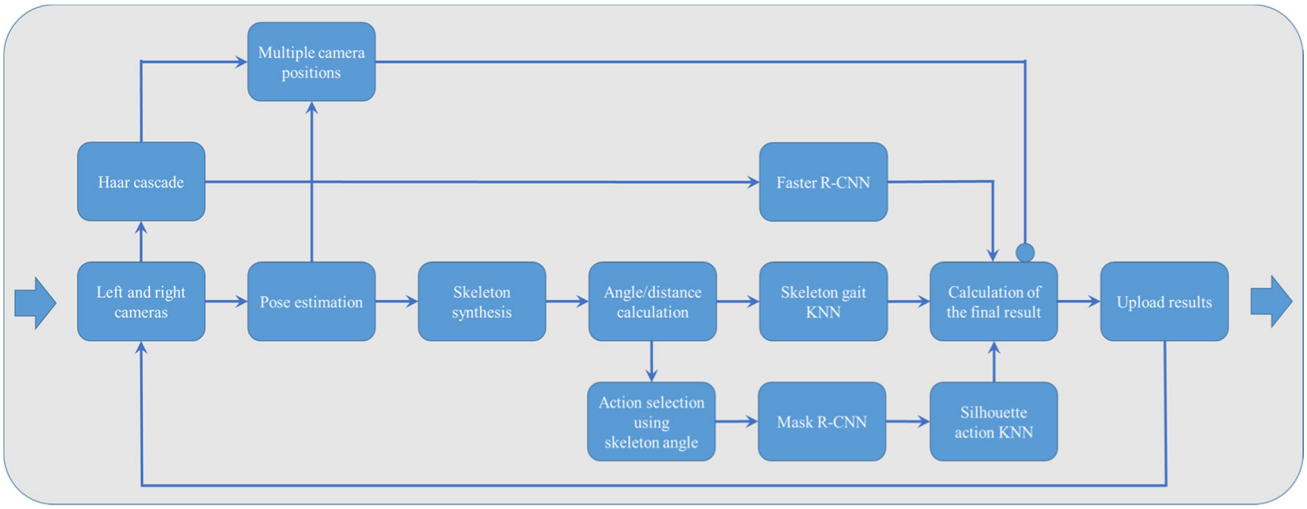


Fig. 7 Flowchart for the proposed system

The proposed intelligent attendance monitoring system uses the above-mentioned three types of recognition process to identify a specific person based on their actions, and will then allow them to enter the monitored area. The system can perform access control system management actions when a specific person has been successfully identified. A LINE messaging API is used to notify specific personnel that they have clocked in or out of work, and a responsive web platform can display historical information related to attendance or absence, as shown in Fig. 6.

A flowchart for the proposed intelligent attendance monitoring system is shown in Fig. 7. A flowchart is given and supplemented with pseudocode (number of lines). First, left and right cameras (1) capture left and right videos of people walking. Pose estimation (4) and a Haar cascade (5) are then used to identify and generate skeleton and face data representing human actions. We calculate the multiple camera positions (6) from the two pieces of data, measure the distance and map the pixels of the video in equal proportions, and judge the position based on the coordinate pixels of the data. We then use skeleton synthesis (16) to create the coordinates of the left and right skeletons. At the angle/distance calculation stage (22), we calculate the angle and distance of the skeleton by using three-point

coordinates to calculate the angle (23) and two-point coordinates to calculate the distance (32). We perform action selection using the skeleton angle (40) according to the skeleton angle data, mainly bend the left and right elbows by 150–170° and then input the selected action into a Mask R-CNN (55) to obtain the silhouette. We use the skeleton gait KNN (57) to identify the angle distance data and the silhouette action KNN (62) to identify silhouette data. Faster R-CNN (72) is applied to identify the face data, and we then calculate the final results (73) and upload them (74).

When the position of the person is outside the maximum range for image capture, no motion processing will be performed. When the position is between the maximum and minimum ranges for image capture, the system will recognise the skeleton key points and convert them into angle and distance data. The silhouette information is also recognised and processed based on the angle information. When the person's position is below the minimum range for image capture, calculation of the skeleton key points and silhouette recognition are stopped in order to perform face recognition processing. Based on the above-mentioned gait and silhouette features and face recognition technology, information fusion is performed to give the recognition result, as shown in Fig. 8.

3.2 Sensing and positioning using multiple cameras

Image capture with a single camera is likely to cause problems associated with the viewing angle and body occlusion, and we can illustrate this by taking as an example the Posenet skeleton key point detection model, which is used to obtain the key point information on the skeleton in each image. Continuous walking behaviour will cause the body to swing, with the front half of the body covering the back half, meaning that the key points of the skeleton cannot be identified and leading to a misjudgement.

The design of the system relies on two cameras to capture images from multiple angles. The left and right facial dynamic images captured by the two cameras are used for face recognition, to confirm that the angle and position of the face recognition results match the principle of left–right symmetry for a human face. The system applies the Posenet skeleton key point detection model to the continuous images captured by the two cameras to obtain the key point information of the human body skeleton. The information on the position of the face is used to synthesise the skeleton key points from the continuous images on both sides, to avoid problems arising from the viewing angle and body occlusion and the feature point information that strengthens the recognition of human action state, as shown in Fig. 9. The image captured by the left camera locates the position of the face in the image, such as the upper right quarter, while the image captured by the right camera locates the position of the face in the image, such as the upper left quarter. The position of the face on the actual level is determined by combining the positions of the left and right faces to locate the test subject.

3.3 Gait and contour characteristics

The Posenet skeleton key point detection model is used to obtain the key point information from each image, and the key points of the ears, eyes and nose are used as positioning reference points for the synthesis location of the key points of the continuous human skeleton. The key points of the left and right shoulders, left and right elbows, left and right wrists, left and right arms, left and right knees and left and right ankles are synthesised with the key points of the left and right skeleton to recognise an action by a specific person, as shown in Fig. 10.

The system calculates the changes in the continuous angles and distances based on the synthesised skeleton key point information. The angle information refers to the angles formed between the key points of the skeleton and the joints. The change in the angle of the skeleton key points is used as the basis for calculating the continuous change in an action by a specific person. For example, the key points of the left and right shoulders, elbows and wrists of the human body can be used to generate eight sets of angle information. The distance information relates to the relative change in distance between the left and right joints of the skeleton. The changes in distance between the key points of the continuous skeleton are used to calculate the change in walking frequency of a specific person. For example, the key points of the left and right elbows and wrists of the human body can be used to generate four sets of distance information. The system is supplemented by the recognition of body silhouette features, in order to improve the accuracy of action recognition for a specific person. The system uses the continuous changes in angle as a basis for judging the periodic actions of the gait and takes the plural same angle time points for completing a time sequential gait action as the sampling basis. At the same time, the time-sequenced changes in the gait are used to classify training and identify using a multiple recognition model, and the movement far and near zooming of human action must be considered to normalise the images captured by the cameras. The system therefore uses the changes in the distances between the key points of the skeleton as the basis for normalisation of the image. A Mask R-CNN deep learning object detection model is used to generate time-sequenced multiple human silhouette feature information

```
1. Turn on left and right cameras
2. Loop until closed
3.     Collect left and right camera data, place in frameL and frameR
4.     Use Pose estimation for frameL and frameR, Generate data skeletonL and skeletonR
5.     Use Haar cascade for frameL and frameR, generate data face_data
6.     Calculate Multiple camera positions
7.         Let X,Y=MaxRange[X,Y]
8.         Let x,y=MinRange[X,Y]
9.         if(skeletonL,R==face_data[position] and skeletonL[nose][Y]==skeletonR[nose][Y]):
10.            L=((skeletonL[nose][X]-x)/(X-x))*Actual_distance_factor
11.            R=((skeletonR[nose][X]-x)/(X-x))*Actual_distance_factor
12.            Positioning_Data=(L+R)/2
13.    if Positioning_Data out of Max_Range:
14.        continue the loop
15.    else if Positioning_Data between Max_Range and Min_Range:
16.        Use skeletonL and skeletonR to run Skeleton Synthesis, Get Synthesis_Data
17.            Let Synthesis_Data[nose]=(0,0)
18.            For i in skeletonL[Left]:
19.                Synthesis_Data.append(skeletonL[nose][X]-i[X],skeletonL[nose][Y]-i[Y])
20.            For i in skeletonR[Right]:
21.                Synthesis_Data.append(skeletonR[nose][X]-i[X],skeletonR[nose][Y]-i[Y])
22.        Use Synthesis_Data for angle/distance calculation, Put the data in Angle_Distance_Data
23.            def Angle(A,B,C):
24.              AB=[A[0]-B[0],A[1]-B[1]]
25.              CB=[C[0]-B[0],C[1]-B[1]]
26.              AAB=(AB[0]**2+AB[1]**2)**0.5
27.              ACB=(CB[0]**2+CB[1]**2)**0.5
28.              DAC=AB[0]*CB[0]+AB[1]*CB[1]
29.              cos=DAC/(AAB*ACB)
30.              angle=(acos(cos)/pi)*180
31.              return angle
32.            def Distance(A,B):
33.              AB=[A[0]-B[0],A[1]-B[1]]
34.              distance=(AB[0]**2+AB[1]**2)**0.5
35.              return distance
36.            For i in range(7):
37.                Angle_Distance_Data.append(Angle(Synthesis_Data[i][0],Synthesis_Data[i][1],Synthesis_Data[i][2]))
38.            For i in range(3):
39.                Angle_Distance_Data.append(Distance(Synthesis_Data[i+8][0],Synthesis_Data[i+8][1]))
40.        Use Angle_Distance_Data for Action selection using skeleton angle calculation, Put the data in Data150L~170L, Data150R~170R
41.            if AngleDistanceData[Left Elbow][Angle]==150 degrees:
42.                Put frameL in Data150L
43.            if AngleDistanceData[Left Elbow][Angle]==160 degrees:
44.                Put frameL in Data160L
45.            if AngleDistanceData[Left Elbow][Angle]==170 degrees:
46.                Put frameL in Data170L
47.            if AngleDistanceData[Right Elbow][Angle]==150 degrees:
48.                Put frameR in Data150R
49.            if AngleDistanceData[Right Elbow][Angle]==160 degrees:
50.                Put frameR in Data160R
51.            if AngleDistanceData[Right Elbow][Angle]==170 degrees:
52.                Put frameR in Data170R
53.    else if Positioning_Data within Min_Range:
54.        Use Angle_Distance_Data for data preprocessing, Get Gait_Data
55.        Use Data150L, Data160L, Data170L, Data150R, Data160R and Data170R to run Mask R-CNN,
56.        Put the data in Multi_action_Silhouette_Data
57.        Use Skeleton gait KNN to identify Gait_Data, and output the result to Gait_Result
58.            for i in range(len(Gait_Data)):
59.                for t in range(i)
60.                    Temporal_weight_Gait_Data.append(Gait_Data[i]*gait_weight)
61.            Gait_result=KNN.predect(Temporal_weight_Gait_Data)
62.        Use Silhouette action KNN to identify Multi_action_Silhouette_Data, and output the result to
63.        Silhouette_Result
64.            scores=[0,0,0,0,0,0,0,0]
65.            scores[KNN150L.predect(Data150L)-1]=scores[KNN150L.predect(Data150L)-1]+silhouette_weight
66.            scores[KNN160L.predect(Data160L)-1]=scores[KNN160L.predect(Data160L)-1]+silhouette_weight
67.            scores[KNN170L.predect(Data170L)-1]=scores[KNN170L.predect(Data170L)-1]+silhouette_weight
68.            scores[KNN150R.predect(Data150R)-1]=scores[KNN150R.predect(Data150R)-1]+silhouette_weight
69.            scores[KNN160R.predect(Data160R)-1]=scores[KNN160R.predect(Data160R)-1]+silhouette_weight
70.            scores[KNN170R.predect(Data170R)-1]=scores[KNN170R.predect(Data170R)-1]+silhouette_weight
71.            Silhouette_Result=Max(scores)
72.        Use face_data to run Fast R-CNN, and output the result to Face_Result
73.        Calculate the final result of Gait_Result, Silhouette_Result, and Face_Result to get the Final_Result
74.    Upload results
75.        Use LINE message API to send the FinalResult to the user
76.        Upload FinalResult to the Web Server
77.        Clear all Data
78. Turn off Left and Right cameras
```

Fig. 8 Pseudocode for the proposed algorithm

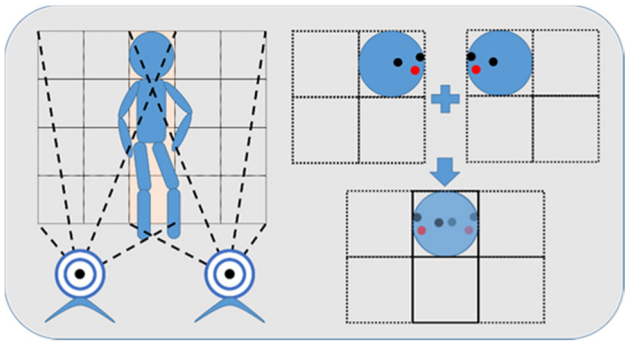


Fig. 9 Sensing and positioning using multiple cameras

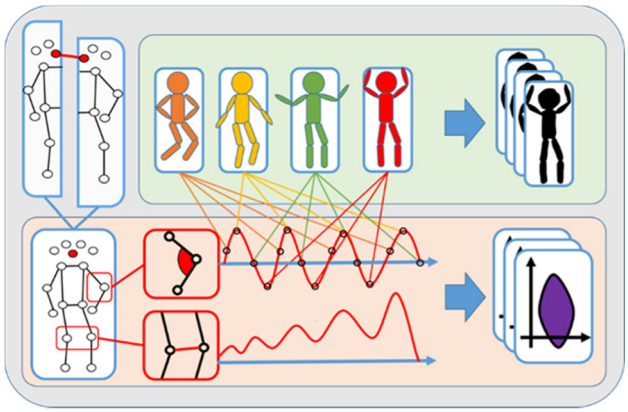


Fig. 10 Synthesis of the key points of the skeleton

and then to perform time-sequenced multiple KNN recognition model classification training and identify for multiple human silhouette. Finally, multiple KNN algorithms are used to identify the results of the recognition model and to vote on the final prediction result. The lower part of Fig. 10 shows how the timing changes of the joint angles are calculated, and how the action pictures are captured at different time points. The top part of Fig. 10 shows how a Mask R-CNN is used to generate the action silhouette data from the action pictures. The action silhouette data at similar time points are collected together and used for KNN image recognition. A KNN identification model is generated at each time point. The final continuous action silhouette identification result can be

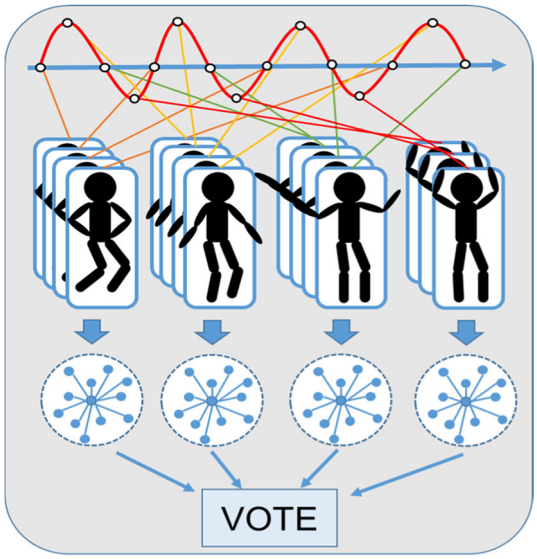


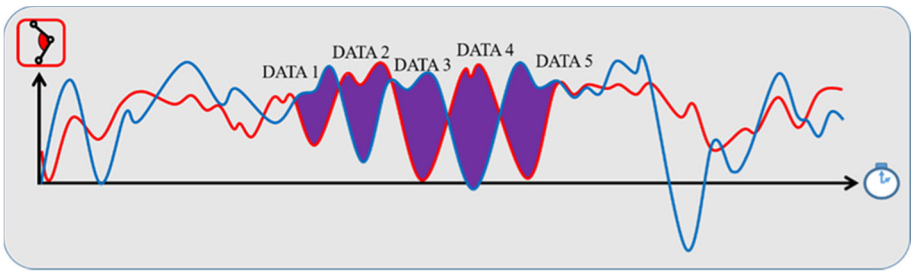
Fig. 12 Action silhouette training process

obtained by voting on the identification results from all the KNN identification models.

3.4 Data preprocessing

The system uses eight sets of angle information, based on the key points of the left and right shoulders, elbows and wrists of the human body, and four sets of distance information based on the key points of the left and right elbows and wrists, as data for preprocessing. Angle–time and distance–time relationship diagrams of the above information show that the human walking gait takes the form of a sine wave with a fixed cycle and maintains a certain degree of symmetry between the left and right sides. The system therefore uses a single period of this sine wave as the gait feature. When the key points of the right and left hands of the human body are at a particular angle, these become the starting point of a single cycle, and when they return to the same angle, this indicates the end of the cycle. As a piece of human gait feature data. In order to avoid misjudging the key points of the skeleton due to noise, Gaussian filtering is applied to remove the noise from the preprocessed data. For each human body, gait feature data to classification and judgment are performed to filter the

Fig. 11 Fixed cyclic sine wave



incomplete, off-peak and irregular feature data for a second time. The system finally obtains human gait feature data in the form of eight sets of continuous angle changes and four sets of continuous distance change information.

The above-mentioned gait feature data include continuous time changes that are equivalent to the walking speed, as shown in Fig. 11. The X axis shows the joint angle, the Y axis represents time, and the blue line indicates the continuous changes in the joint angle of the left skeleton, while the red line shows the continuous changes in the joint angle of the right skeleton. The joint angle gait data from the starting points of the left and right skeleton joint angles being the same to the next time the left and right skeleton joint angles are the same as the end point. The purple block shows the gait data after filtering. A complete walk contains four to six sets of gait data, and the figure shows the process for five sets of data.

3.5 Action silhouette and face recognition

Our system uses the KNN algorithm to train the network to identify the body motion of a specific person. The recognition model distributes the feature data on the feature plane, calculates the closest K data feature classification results of the feature data to be measured and uses a majority decision method to obtain the final classification. The recognition model described above is used to classify the gait and body silhouette features. The system includes angle and distance features in the training of the gait recognition model and adds time-sequenced weight changes. That is to say, weights are set in sequence for the time of human silhouette gait actions to strengthen feature learning and improve recognition accuracy. The system is designed to recognise the body silhouette features and uses multiple sequential KNN recognition models for classification training and identification. It uses a single cycle of a continuous silhouette gait action to sample multiple points at the same angle over time, and the KNN recognition model is applied for classification training and identification based on the silhouette features of each human body at the same angle. The system therefore generates a classifier based on multiple-KNN recognition model, and then adds the gait sequence to generate the weight settings, in order to strengthen the feature learning and improve the recognition accuracy. Finally, the recognition results of the multiple-KNN recognition model are subjected to a vote, to give the final prediction result in terms of identifying a specific person.

As illustrated in Fig. 12, the silhouette features of the test subject are extracted at different time points, and the silhouette features at the same feature time points are put together for KNN training and identification. At each time point, a set of KNN silhouette identification models is generated, and the identification results from all KNN identification models are voted on to give the continuous silhouette identification results for the complete walking gait. The system design combines the face recognition method of a traditional intelligent attendance monitoring system with Haar features, which are used in image processing and recognition technology, to extract non-specific

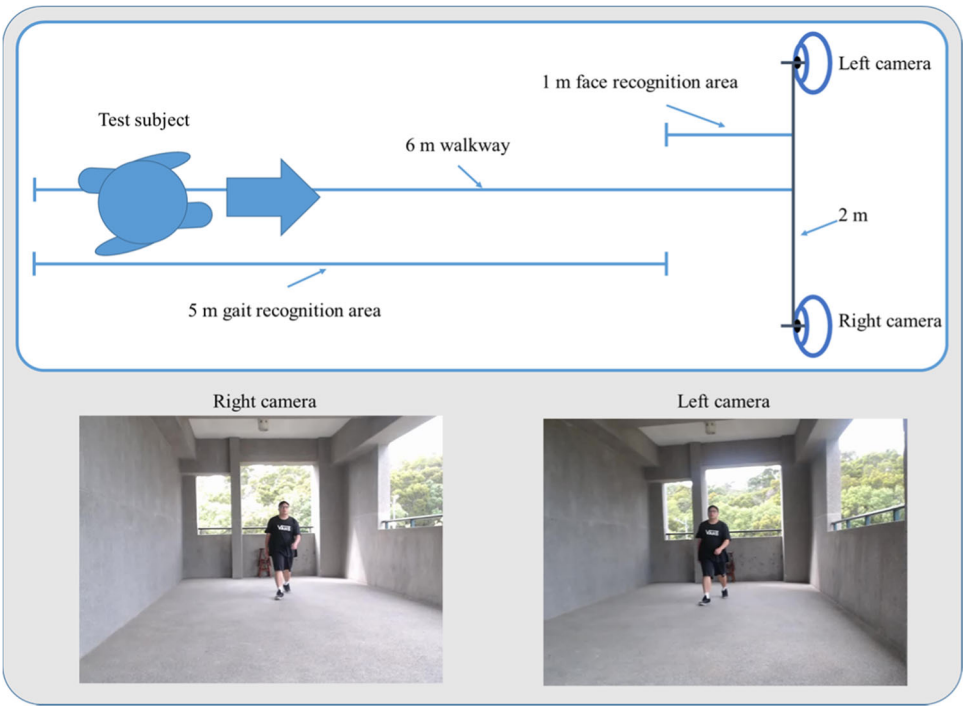


Fig. 13 Diagram showing the experimental setup and example images from the two cameras

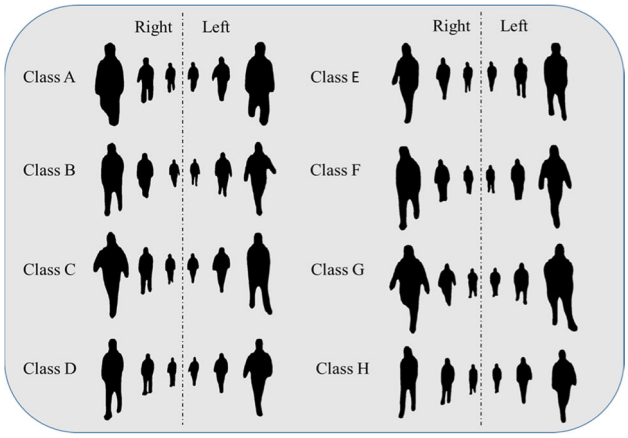


Fig. 14 Images from the dataset

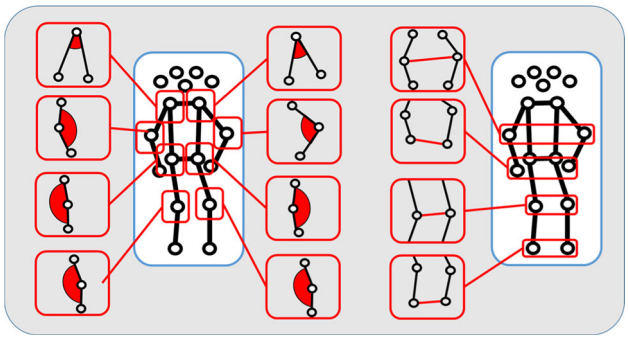


Fig. 15 Data collection for the gait recognition system

features of human faces, where the range is obtained using a cascade classifier. The system also uses the Faster R-CNN deep learning object detection model, as used in deep learning recognition, to perform specific face recognition for the above non-specific human face range. In future work, this part of the model will be replaced by a new network architecture or a deep learning object detection model with higher recognition accuracy. The system can therefore prevent problems such as the use of a 3D face mask to fake an identity and the reduction in accuracy due to a mask covering a face. The recognition system is supplemented by the use of continuous silhouette gait features for specific actions and the body silhouette features, meaning that the final results will have higher accuracy in terms of the recognition of a specific person.

Table 1 Accuracy of the proposed gait and action silhouette recognition system

Accuracy of all data						
Distance data		Angle data		Combined data		Final accuracy
Skeleton recognition						
73.42%		79.83%		83.33%		83.33%
Left hand 150° data	Left hand 160° data	Left hand 170° data	Right hand 150° data	Right hand 160° data	Right hand 170° data	Final accuracy
Silhouette recognition						
61.79%	63.98%	65.98%	63.50%	63.79%	67.65%	72.38%

3.6 Intelligent attendance monitoring system

This paper proposes an intelligent attendance monitoring system, based on temporal and spatial face (static) and motion (dynamic) recognition. It uses the LINE messaging API as the communication medium between system administrators and employees. When the system successfully recognises a specific person entering or exiting the monitored area, it will use the human–computer interaction API to notify the person of the check-in, via a confirmation message for commuting. At the same time, the user can also communicate with the management system via the human–computer interaction API and can carry out actions such as adding and deleting users, uploading images for face recognition training and viewing attendance management information. The intelligent attendance monitoring system is built using an Apache Web Server and a MySQL database and provides a background management web interface that allows administrators to control related information such as the user LINE display name, API transmission ID, user avatar, images for face recognition training and work records.

4 Experimental results

In this chapter, we describe the experimental environment and parameter settings, and analyse the experimental performance in comparison with three alternative deep learning motion detection systems: ST-GCN, GCN-NAS and 2S-AGCN.

4.1 Experimental environment and parameter settings

We use a dual camera system based on Logitech C925e webcams as the image sensing devices, and the system is implemented in Python development software in the Anaconda environment. We use TensorFlow, an open-source software library developed by Google, as a deep learning runtime package. The Posenet skeleton key point detection architecture, the Faster R-CNN deep learning object detection model and the Mask R-CNN deep learning object detection model all use the TensorFlow package. The KNN algorithm recognition model is based on the Scikit-learn

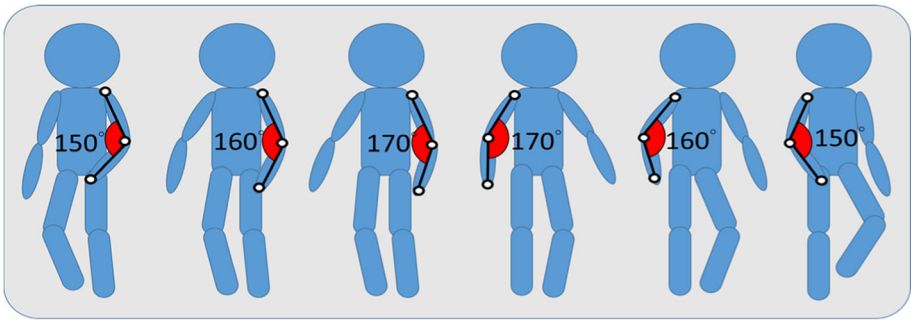


Fig. 16 Data collection for the action silhouette recognition system

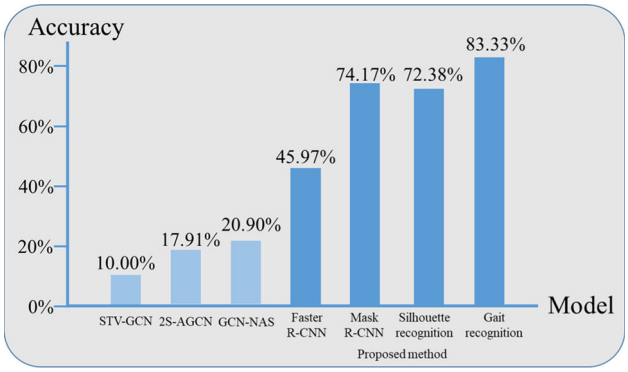


Fig. 17 Identification accuracy for specific people under various conditions

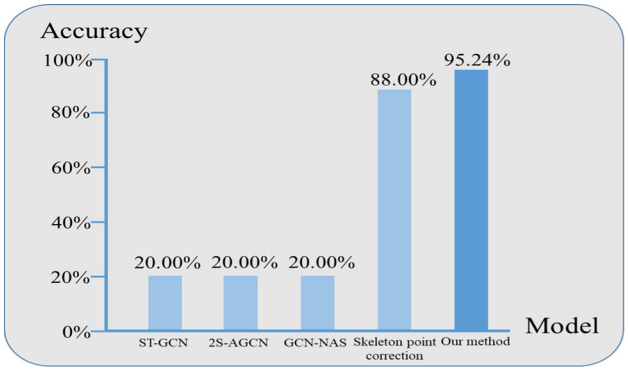


Fig. 18 Action recognition accuracy of our model and alternative schemes based on skeleton key point information

open-source machine learning software library, which is used as a training package, and the access control human–computer interaction API relies on the messaging API officially provided by the LINE developers to communicate with users. The intelligent attendance monitoring system platform uses an Apache Web Server and a MySQL database system.

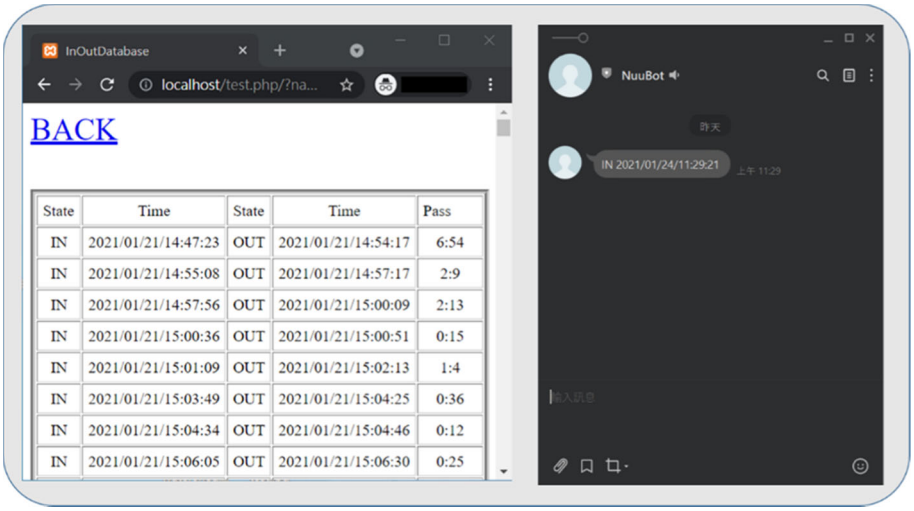
As shown in Fig. 13, the spatiotemporal human action recognition training system used two cameras that were set up on a six-metre-long walkway. The test subjects consisted of seven people, who walked from a distance of six metres from the dual camera setup to one metre away. This gave a total walking distance of five metres for each test subject, which was used as a training sample. Each subject walked this distance 350 times, giving 1292 gait features and 14,536 body silhouette data points. Figure 14 shows walking data for a total of eight people from the left and right cameras. The left and right silhouettes of testers who are 6 m, 3.5 m and 1 m away from the camera are used as a demonstration.

4.2 Performance results

In this section, we analyse the recognition accuracy of our gait feature recognition system. This experiment used a total of 1292 gait features generated from seven people. The gait features included the angle and distance information, as shown in Fig. 15. The angle data for the skeleton were based on eight sets of key points: the left and right shoulders, the left and right elbows, the left and right hips, and the left and right knees. The distance data for the skeleton were based on four sets of key points: the left and right elbows, the left and right wrists, the left and right knees, and the left and right ankles. The angles and distances between the pairs of key points of the skeleton were used as the feature recognition data, and the KNN algorithm was used for training of the recognition model and identification based on the spatio-temporal weight characteristics.

As shown in Table 1, the experimental results show that the recognition accuracy is 73.42% when the distances between the key points of the skeleton are used as features, and the recognition accuracy is 79.83% when the angles between the key points of the skeleton are used as features. The overall recognition accuracy of the system when both the angles and the distances are used as recognition features is 83.33%. Finally, for recognition situations where the subject is wearing a mask and a jacket, our skeleton synthesis gait recognition system has a recognition accuracy of 83.33%.

We now analyse the recognition accuracy of the action feature recognition system. The experiment used a total of 14,536 human silhouette feature data points drawn from seven people, and the recognition results from the multiple-



State	Time	State	Time	Pass
IN	2021/01/21/14:47:23	OUT	2021/01/21/14:54:17	6:54
IN	2021/01/21/14:55:08	OUT	2021/01/21/14:57:17	2:9
IN	2021/01/21/14:57:56	OUT	2021/01/21/15:00:09	2:13
IN	2021/01/21/15:00:36	OUT	2021/01/21/15:00:51	0:15
IN	2021/01/21/15:01:09	OUT	2021/01/21/15:02:13	1:4
IN	2021/01/21/15:03:49	OUT	2021/01/21/15:04:25	0:36
IN	2021/01/21/15:04:34	OUT	2021/01/21/15:04:46	0:12
IN	2021/01/21/15:06:05	OUT	2021/01/21/15:06:30	0:25

Fig. 19 Access control notification in our intelligent attendance monitoring system

KNN recognition model were subjected to a voting process to generate the final prediction results in terms of identifying specific people. The action feature information was based on the angles between the key points of the skeleton as the basis for classification, as shown in Fig. 16. Six actions were identified: the left hand bending forward by 150, 160 and 170°, and the right hand bending forward by 150, 160 and 170°. The systems are, respectively, multiple KNN recognition model is trained based on the above six actions. As shown in Table 1, the experimental results show that the values of the recognition accuracy for the multiple-KNN recognition model were 61.79, 63.98, 65.98, 63.50, 63.79 and 67.65%, respectively, for these six actions. An additional 20 untrained human gait sample data points were used to vote to generate the final results of the multiple-KNN recognition model in terms of predicting specific persons. A total of 1469 silhouette feature data points were generated through information processing, and based on these, the system was able to predict and recognise specific individuals with a recognition accuracy of 72.38% when each one was wearing a mask and jacket.

We now analyse the performance of our model in comparison to similar action recognition schemes based on the key point information of the human skeleton. The state-of-the-art STV-GCN, GCN-NAS and 2S-AGCN systems use continuous key point information of the human skeleton as training data for the action recognition model. In this experiment, we used gait data samples from 20 people wearing masks and jackets to train our action recognition model based on the Posenet skeleton key point detection framework and trained the action recognition model for 5000 epochs. We obtained values for the recognition accuracy of 10% for STV-GCN, 17.91% for 2S-AGCN and 20.9% for GCN-NAS as shown in Fig. 17. Our approach uses a Faster R-CNN deep learning object detection model as the basic face recognition function, for which the recognition accuracy will be reduced when a mask is worn over the face. The face recognition part of our intelligent attendance monitoring system has a recognition accuracy of 45.97% and an average confidence level of only 0.0441 when a mask is worn. When the Mask R-CNN deep learning object detection model is used in the face recognition function, its recognition accuracy when a specific person is wearing a mask is 74.17%. Based on these experimental results, it can be seen that the Faster R-CNN and Mask R-CNN deep learning target detection models cannot deal with a situation in which a person is wearing a mask, as this leads to a decrease in the recognition accuracy. In our system, the silhouette feature recognition system has a recognition accuracy of only 72.38% when the subject is wearing a jacket. Therefore, the continuous gait features of specific actions and a body silhouette feature recognition model are integrated to create a multiple recognition model, which uses voting to generate the final recognition result. Our intelligent attendance monitoring system is able to carry out gait feature recognition when a specific person is wearing a mask and jacket, and its recognition accuracy is 83.33%. The main reason for this is that the STV-GCN, GCN-NAS and 2S-AGCN deep learning recognition model systems have a high level of recognition accuracy for different human actions, but do not give good recognition accuracy for the same actions carried out by different people. Through the use of multi-view features, our scheme not only avoids the problem of occlusion caused by actions, but also has the effect of feature amplification, and thus a higher recognition accuracy.

We now compare the performance of our model to similar action recognition schemes based on the key point information of the human skeleton. As shown in Fig. 18, the related work provides a database of action recognition, which includes the action data on the human body during a

golf swing. Using the feature extraction technology presented in this article, a total of 135 pieces of motion data were captured. The proposed KNN human skeleton gait recognition system with temporal weights was used for training, and an accuracy rate of 93% was obtained. The limitations on the features used in the alternative scheme results in an accuracy of only 88%, which is about 5% lower than the skeleton merge feature zoom proposed in this paper. Traditional skeleton recognition technologies such as ST-GCN, GCN-NAS and 2S-AGCN have a recognition accuracy of only 20% and cannot handle the problem of skeleton occlusion at all, resulting in very low recognition performance. Our system offers access control based on the identification results of the specific person in the monitored area, and uses the LINE messaging API to send attendance-related information messages to specific personnel, as shown in Fig. 19. At the same time, the system saves attendance-related information to a cloud database and the responsive web platform to allow managers to track the footprints of specific personnel and monitor changes in personnel flow.

5 Conclusions

Our intelligent attendance monitoring system can efficiently manage the attendance, absence and time records of specific personnel, thus allowing managers to understand the changes in the flow of people and the footprints of specific personnel entering and exiting a given area. This makes it possible to conduct data analysis and manage the flow of people within the scope of safety considerations for various attendance and absence records. Existing attendance monitoring systems based on face recognition technology cannot handle the problem that arises when masks are worn over the face, as this leads to a reduction in the performance accuracy; in addition, the use of a 3D face mask to fake an identity poses a problem. At the same time, the silhouette features cannot be effectively identified in the case where the subject has changed their clothes.

In this paper, we therefore propose an intelligent attendance monitoring system with spatio-temporal human action recognition, which combines the use of skeleton gait features, body action silhouette features and facial feature recognition technology. Faced with feature masking, such as wearing a mask, changes in clothes and viewing angle masking, good recognition ability can also be obtained through multi-angle skeleton synthesis gait recognition. Our experimental results show that the proposed intelligent attendance monitoring system has an accuracy of 93.33% in terms of identification when a mask is worn by a specific person in the monitored area. In future work, we will take into account the success of most recognition technologies and the failure of a single recognition technology to automatically collect samples and will retrain and identify the automatic recognition model to improve the recognition accuracy of the overall intelligent attendance monitoring system. At the same time, research on multi-target recognition is carried out, aiming at the skeleton formed by multiple people walking with different perspectives at the same time. In future work, when the same person is distinguished, the skeleton will be merged, which is expected to solve the problem of angle offset caused by different walking positions.

Funding This research was funded by National United University, Taiwan.

Data availability Enquiries about data availability should be directed to the authors.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

Ethical approval This article does not contain any studies with human participants or animals performed by any of the authors.

References

Cheema U, Moon S (2021) Sejong face database: a multi-modal disguise face database. Comput vis Image Underst 208–209:1–9

He K, Gkioxari G, Dollar P, Girshick R (2017) Mask R-CNN. In: IEEE international conference on computer vision, pp 2980–2988

Kong S, Heo J, Abidi B, Paik J, Abidi M (2005) Recent advances in visual and infrared face recognition—a review. Comput vis Image Underst 97(1):103–135

Papandreou G, Zhu T, Chen L, Gidaris S, Tompson J, Murphy K (2018) PersonLab: person pose estimation and instance segmentation with a bottom-up, part-based, geometric embedding model. In: European conference on computer vision, pp 282–299

Peng W, Hong X, Chen H, Zhao G (2020) Learning graph convolutional network for skeleton-based human action recognition by neural searching. Proc AAAI Con Artif Intell 34(03):2669–2676

Ren S, He K, Girshick R, Sun J (2017) Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Trans Pattern Anal Mach Intell 39(6):1137–1149

Shakhnarovich G, Lee L, Darrell T (2001) Integrated face and gait recognition from multiple views. In: IEEE computer society conference on computer vision and pattern recognition, pp 439–446

Shi L, Zhang Y, Cheng J, Lu H (2019) Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In: IEEE conference on computer vision and pattern recognition, pp 12018–12027

Tsai M, Li M (2021) Attendance monitoring system based on artificial intelligence facial recognition technology. In: ieee international conference on consumer electronics-Taiwan, pp 1–2

Tsai M, Chen C (2021) Spatial temporal variation graph convolutional networks (STV-GCN) for skeleton based emotional action recognition. IEEE Access 9:13870–13877

Tsai M, Chen C (2022) Enhancing the accuracy of a human emotion recognition method using spatial temporal graph convolutional networks. Multimed Tools Appl. https://doi.org/10.1007/s11042-022-13653-x

Tsai M, Huang S (2022) Enhancing accuracy of human action recognition system using skeleton point correction method. Multimed Tools Appl 81:7439–7459

Wang L, Tan T, Ning H, Hu W (2003) Silhouette analysis-based gait recognition for human identification. IEEE Trans Pattern Anal Mach Intell 25(2):1505–1518

Wang L, Ning H, Tan T, Hu W (2004) Fusion of static and dynamic body biometrics for gait recognition. IEEE Trans Circuits Syst Video Technol 14(2):149–158

Yan S, Xiong Y, Lin D (2018) Spatial temporal graph convolutional networks for skeleton-based action recognition. Assoc Adv Artif Intell 32:7444–7452