A Multi-Modal Gait Analysis-Based Detection System of the Risk of Depression

Wei Shao, Zhiyang You, Lesheng Liang, Xiping Hu, Chengming Li, Wei Wang, and Bin Hu, *Senior Member, IEEE*

***Abstract*—Currently, depression has become a common mental disorder, especially among postgraduates. It is reported that postgraduates have a higher risk of depression than the general public, and they are more sensitive to contact with others. Thus, a non-contact and effective method for detecting people at risk of depression becomes an urgent demand. In order to make the recognition of depression more reliable and convenient, we propose a multi-modal gait analysis-based depression detection method that combines skeleton modality and silhouette modality. Firstly, we propose a skeleton feature set to describe depression and train a Long Short-Term Memory (LSTM) model to conduct sequence strategy. Secondly, we generate Gait Energy Image (GEI) as silhouette features from RGB videos, and design two Convolutional Neural Network (CNN) models with a new loss function to extract silhouette features from front and side perspectives. Then, we construct a multi-modal fusion model consisting of fusing silhouettes from the front and side views at the feature level and the classification results of different modalities at the decision level. The proposed multi-modal model achieved accuracy at 85.45% in the dataset consisting of 200 postgraduate students (including 86 depressive ones), 5.17% higher than the best single-mode model. The multi-modal method also shows improved generalization by reducing the gender differences. Furthermore, we design a vivid 3D visualization of the gait skeletons, and our results imply that gait is a potent biometric for depression detection.**

Manuscript received 30 July 2021; revised 29 September 2021; accepted 18 October 2021. Date of publication 26 October 2021; date of current version 5 October 2022. This work was supported in part by the National Key Research and Development Program of China under Grant 2019YFA0706200 and in part by the National Natural Science Foundation of China under Grants 61632014 and 61627808. *(Corresponding authors: Xiping Hu, Chengming Li, Wei Wang, and Bin Hu.) (Wei Shao and Zhiyang You contributed equally to this work).*

Wei Shao, Zhiyang You, and Lesheng Liang are with the Gansu Provincial Key Laboratory of Wearable Computing, School of information Science and Engineering, Lanzhou University, Gansu 730000, China (e-mail: shaow18@lzu.edu.cn; youzhy20@lzu.edu.cn; lianglsh19@lzu.edu.cn).

Xiping Hu is with the Gansu Provincial Key Laboratory of Wearable Computing, School of information Science and Engineering, Lanzhou University, Gansu 730000, China, and also with the School of Intelligent Systems Engineering, Sun Yat-sen University, Shenzhen 518107, China (e-mail: huxp@lzu.edu.cn).

Chengming Li and Wei Wang are with the School of Intelligent Systems Engineering, Sun Yat-sen University, Shenzhen 518107, China (e-mail: lichengming@mail.sysu.edu.cn; wangw328@mail.sysu.edu.cn).

Bin Hu is with the Gansu Provincial Key Laboratory of Wearable Computing, School of Information Science and Engineering, Lanzhou University, Lanzhou 730000, China, and also with the Institute of Engineering Medicine, Beijing Institute of Technology, Beijing 100000, China (e-mail: bh@lzu.edu.cn).

Digital Object Identifier 10.1109/JBHI.2021.3122299

***Index Terms*—Depression, gait, skeleton, multi-modal, fusion model.**

I. Introduction

DEPRESSION is a kind of mood disorder characterized by prolonged sadness [1]. Worse still, depression can lead to suicide. Moreover, the incidence of depression in postgraduate students worldwide is much higher than that of the general public [2], [3]. Postgraduate students tend to face more academic, affective, and financial challenges. It has been confirmed that depression affects suicidal ideation and suicide attempt in postgraduate students [4]. Clinical, depression is diagnosed by persistent depressive symptoms and assessed by scales, which depend on the General Practitioners (GPs) and cooperative patients. However, this conventional method is time-consuming and lacks objective metrics. Moreover, there is a serious shortage of professional psychiatrists required for the traditional diagnosis of depression worldwide, which makes it impossible for many patients to receive diagnosis and treatment in time. Furthermore, there existent an increasing trend within the population in the prevalence of depression [5]. Especially more mental health issues caused by the COVID-19 epidemic strengthen the pressure of the healthcare [6]. Therefore, it is essential to find a reliable, low-cost, rapid, and non-contact way to diagnose depression which can help large-scale screen of depression risk people. The assisted diagnosis of depression with the help of Artificial Intelligence (AI) technology is a solution to the shortage of healthcare due to its quick and objective judgment [7]. At present, some researchers have proposed some methods and models for recognizing depression, such as methods based on audio [8], video [9], text [10], [11], and physiological signals [12], [13]. These methods can assist medical staff in diagnosing depression to some extent, but these methods still have many drawbacks. For example, when collecting audio and video data, subjects generally require to communicate with AI assistants or experimenters in a restricted environment, which takes much time and cannot be used for large-scale depression screening. Gait, a traditional biometric, has shown manifold potentials such as early detection of neurodegenerative diseases (NDD) [14]–[16] and potentially dangerous situations recognition [17], due to the advancement of information acquisition and analysis. Besides, some neuroscience studies indicate a correlation between depression and gait [18]. Recent studies have proposed some depression recognition methods based on gait. Deligianni and

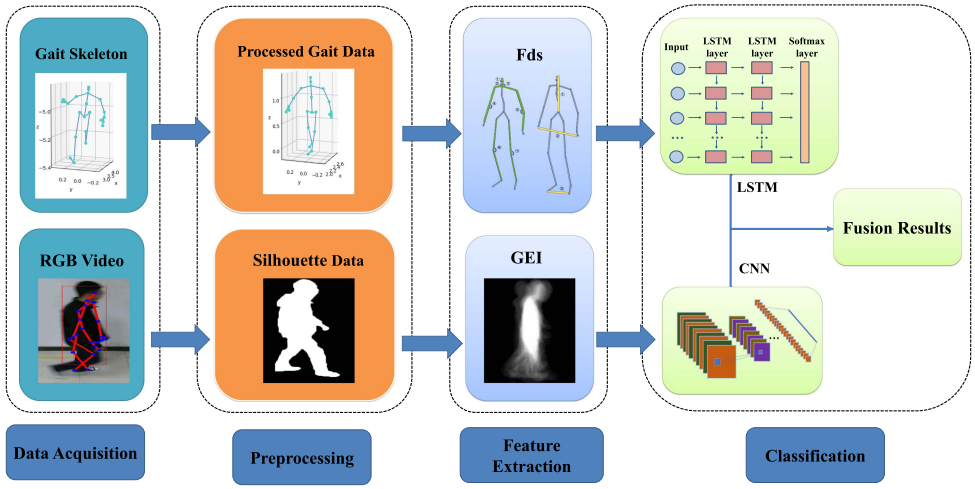

Fig. 1. The pipeline of our method. After acquiring the gait information (including skeleton and RGB videos), the step is to pre-process the data. For skeleton data, the pre-process steps include coordinate transformation and filtering, and extract corresponding features F_{ds}. For RGB videos, the pre-process step is mainly to extract silhouette and generate GEI as silhouette feature. Next, we use LSTM and CNN to classify depression with skeleton and silhouette features as input respectively. Finally, we fuse the classification results to obtain the final result.

Feldman *et al.* [18], [19] used the network control theory to analyze the relationship between gait and mood disorders, such as depression and anxiety. Moreover, they indicate that those suffering from depression tend to show a slumped posture with a slow gait. However, these researches only using 3D gait skeleton data [20]–[23], and the classification results depend on the data quality of the estimated skeleton locations by Kinect devices. Kinect devices have been widely used for the data acquisition of gait [24], yet, it has a ranging limit of 1.2-3.5 m distance, which cannot cover all data collection scenarios. Furthermore, when subjects walk from perspectives other than the front of the devices, the skeleton estimation will be inaccurate. On the contrary, gait recognition using silhouette data [25], [26] do well in many perspectives. Especially, gait energy image (GEI) reducing the amount of calculation becomes a widely used silhouette feature [27]. However, silhouette data extracted from 2D images, which neglect the distance information thus may have poor recognition performance.

In this paper, we propose a multi-modal depression recognition method by combining skeleton data and silhouette data of gait. The pipeline of our method is depicted in Fig. 1. First, we construct the data set consisting of 200 people (86 students at risk of depression, 114 healthy controls) including skeleton and silhouette data, and perform data pre-processing to improve data quality. Second, for skeleton data, the skeleton feature F_{ds} for depression recognition is proposed, which includes spatiotemporal features and kinematics features. For silhouette data, a Convolutional Neural Network (CNN) model with a new loss function is designed to extract silhouette features. Finally, we merge the silhouette features of the front and side views; then in the decision level, we fuse the classification results of skeleton data and silhouette data. Thus, a multi-modal fusion model is constructed for depression recognition based on gait analysis, which achieves an accuracy of 85.45%, and improves the classification performance compared with a single modality. The main contributions of this article are as follows:

1) We propose a gait analysis-based depression detection system with skeleton and silhouette modal, which achieves high accuracy with reducing gender differences. As far as we know, this is the first attempt to combine skeleton data and silhouette data for depression detection.
2) Considering the clinical symptoms of gait in patients with depression, we propose a skeleton feature set including spatiotemporal features and kinematics features to describe depression.
3) For silhouette images, two CNN models with the max merge strategy are proposed to extract silhouette features in both front and side perspectives. Furthermore, a new loss function is designed to accelerate convergence.
4) We demonstrate the 3D visualization of the gait skeleton based on Virtual reality (VR) technology, which provides the walking skeleton image from 360 viewing angles and shows the gait difference directly between the normal and people at risk of depression.

The remainder of this paper is organized as follows. Section II analyzes the correlation between gait and disease, especially depression at first. Then reviews researches using skeleton data or silhouette data. Section III describes the proposed system in detail. The experimental are presented in Section IV. Section V discusses the results. Section VI concludes the paper.

II. Related Works

A. Diseases Detection With Gait

Advances in gait data and image technology promote medicine. Videos were fed in graph convolution neural network perform well in detecting freezing of gait (FoG) to predict Parkinson's disease [15], the experiment was conducted in the

dataset of 45 patients with an AUC of 0.887. In [14], a new neural network architecture was designed for extracting temporal features and spatial features and fused the results at the decision level, which outperforms in classifying NDD. These findings are encouraging but do not fully explore except in NDD. Different from the cumbersome and costly acquisition of some physiological signals, gait, a novel biometric [28], arouses attention due to its advantages that it can be recognized from a distance for acquisition without the subject's cooperation and difficult to disguise. Accordingly, it is object way to detect depression by gait.

Lemake *et al.* [29] are one of the pioneers in the study of the correlation between depression and gait. Jonathan *et al.* [30] posed the association between depressive mood and activity in Parkinson's Disease. They calculated the spatiotemporal gait parameters of patients with major depression disorder and found that the patients show significant reductions in stride length, cycle time, and lower limb support. Moreover, substantial epidemiological studies [29]–[35] show that the gaits of depression patients are different from the normal's. For instance, the gait pattern in depression appears to a slower gait speed and shorter step length [32], [36]. Furthermore, in [37], the experimental results of the two-year follow-up modeled by logistic regression indicate that gait disturbance predicts incident depression. Murri *et al.* [38] deemed that depression in adults correlates with impairments of posture and gait while the elder's gait may depend on the interplay of diseases, cognitive impairment, and mood. They mainly used sensors to record gait data and calculated hand-crafted features [31], [35], [39] such as gait speed, stride length, stride duration to serve as temporo-spatial gait parameters. Although the hand-crafted method increases the processing procedure, it is conducive to enhancing interpretability.

B. Gait Analysis With the Single-Modal

Gait recognition has thrived on skeleton-based methods because the 3D representations are smaller data sizes with rich information. Notwithstanding, pose estimation algorithms such as Kinect are vulnerable for the occlusion of clothing. In [40], authors constructed frame-level features vectors using position vectors calculated from coordinates. Experimental results indicate the features are robust and effective in gait recognition. Considering that depressed individuals suffer prolonged sadness, we also reviewed the mood recognition by gait and pay attention to the sad class in this part. Wang *et al.* [41] used skeleton data estimated by Kinect devices to predict depression risk in the dataset of 43 scored-depressed and 52 non-depressed individuals. They combined the time domain information, frequency domain information, and spatial geometric features of gait information. The experimental results show that spatial features help a lot in the evaluation of depression. Bhattacharya *et al.* [18] proposed new features called Affective Features composed of movement features and posture features, calculated using 3D skeleton data extracted from RGB video. The hand-crafted features perform better than those only used features such as the angle, distance, frequency domain information, etc. What's more, they proposed the new Affective Features in [42] which serve as a constraint in training the autoencoder. Encouragingly, they achieved average precision of 89% in the sad class. Zhao *et al.* [21] used the gait features in the frequency domain to regress the PHQ-9 scores, and the correlation coefficient reached 0.51 by Gaussian processes. Chiu *et al.* [43] estimated the position of the human skeleton from side views by a deep learning method from the video and then calculated gait parameters as features. As a result, the best accuracy reached only 64% in classifying five emotion labels. However, these researches only using gait skeleton data so that the classification results depend on the quality of the estimated joints locations. To enhance the accuracy, we combine skeleton and silhouette data to recognize depression.

Silhouette data are considered a powerful tool to describe gait due to that the binary images characterize the position information with little redundancy details such as the dressing. Furthermore, it benefits to achieve the cross-view gait recognition tasks. GEI has been a widely used silhouette feature attribute to the simple processing and strong anti-noise ability. After extracting silhouettes data and centering the upper half part, the GEI calculates as follow:

$$G(x, y) = \frac{1}{N} \sum_{t=1}^{N} B_t(x, y) \tag{1}$$

where x and y are coordinates of the pixels in the images of the silhouettes at t-th frame, N denotes gait cycle length in frames.Benefit from the time normalization and space normalization based on the cycle length, GEI reduces calculation and memory without loss of performance and has strong anti-noise capability [27].

GEI captures gait features from side view well and subserves gait recognition from multi-view. Wu *et al.* [26] proposed an approach to multi-view human identification by deep CNN fed with GEI from view left to right, which achieved the best recognition rates on three challenging public datasets. Wang *et al.* [44] proposed a frame-by-frame GEI inspired by sliding window algorithm, and using Conv-LSTM model and outperforms others on the CASIA Dataset B for cross-view gait recognition. Furthermore, they achieved the average correct recognition rates at 95.9%. Inspired by GEI, Zhang *et al.* [45] presented a new gait representation called gait individual image, which removed redundant view-dependent parts from GEI and has good performance in cross-view identity recognition. However, there is no research to introduce silhouette data into depression recognition before. Therefore, we propose a method of fusing the two modalities, skeleton and silhouette, combining the advantages of the two modalities to detect depression.

III. Method

A. Data Acquisition

In this study, all participants were recruited from graduate students of Lanzhou University, including 86 ones at risk of depression and 114 controls. The protocol had obtained permission from the Institutional Review Board of Lanzhou University.

TABLE I
DEMOGRAPHIC AND STATISTICAL CHARACTERISTICS OF THE STUDY SAMPLE

Group	Male	Female	SDS	PHQ-9
Score-depressed	38	48	68.24	10.50
Health Control	54	60	34.52	2.43

* The table shows the mean of the SDS score, PHQ-9 score respectively, and the standard of division: SDS score $\geq$ 60 and PHQ-9 score $\geq$ 10.

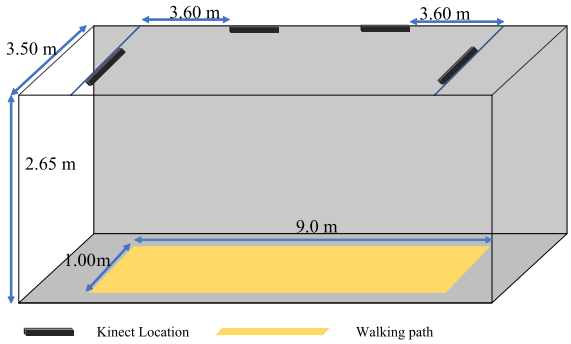

Fig. 2. The schematic diagram of data acquisition system, we deploy 4 Kinect cameras in a corridor, 2 cameras are located at both ends of the walking path, and other 2 cameras are located at the side of the path.

Each subject was asked to answer the Self-rating Depression Scale (SDS) and the Physical Health Questionnaire (PHQ-9) to assess their mental state and depression. Table I shows the demographic information and scales scores of the participants. The subjects are between 21 and 30 years old, and the number of women is slightly more than the number of men.

The Kinect cameras can not only capture real-time RGB and depth videos but also provide coordinates of joints from each frame of the videos. Our data acquisition system is similar to [16], 4 Kinect cameras are deployed above a 9-meters-long path, as shown in Fig. 2, two of them face each other. Furthermore, the tilt angle is set to 27° to have a larger acquisition perspective. Between the two Kinect devices, the other two are deployed on the side of the walkway. Thus, two devices collect data from a front perspective, and the other two collect data from a side perspective. Each subject was required to walk this path back and forth, which lasts one minute.

B. Pre-Processing

In this paper, both skeleton data and RGB image data are used for gait recognition. The skeleton data are the three-dimensional coordinates of 25 joints generated by Kinect devices. Due to the poor quality of the skeleton data generated in the side and back view, only the skeleton data under the front view were used. After discarding the incorrectly estimated coordinates or frames containing a malformed skeleton, we performed coordinates transformation on the data. As shown in Eq.2, we rotated the original coordinates to horizontal, $[x,' y,' z']$ and $[x, y, z]$ denote the coordinates after transformation and raw respectively, and the θ denotes the tilt angle. The reason for transformation is that the skeleton coordinates generated by Kinect with a tilt angle overlap easily, which harms recognition. After that, we applied Gaussian filtering to the coordinates on the x, y, and z-axis over time to smooth the data, whose coefficients are $h = \frac{1}{16}[1, 4, 6, 4, 1]$.

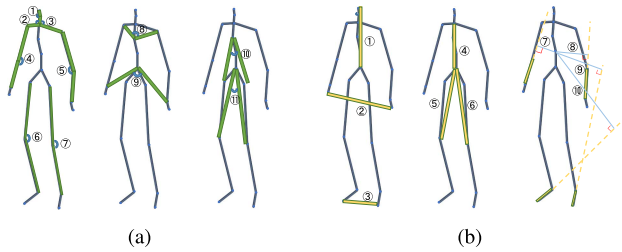

Fig. 3. The composition of spatiotemporal features. (a) Angle features include 11 angles between joints. (b) Distance features include 6 distances between joints and 4 distances from joints to skeletons.

$$\begin{pmatrix} x' \\ y' \\ z' \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos\theta & \sin\theta \\ 0 & -\sin\theta & \cos\theta \end{pmatrix} \times \begin{pmatrix} x \\ y \\ z \end{pmatrix}, \theta = 27°. \quad (2)$$

As for the RGB image data, we applied the Mask R-CNN [46] which output a binary mask for each region of interest to obtain the silhouette of subjects from the original data and transform it into a binary image. First, the video stream was divided into frames to obtain the picture that needs silhouette segmentation. Then we can get the required binary image containing silhouettes by feeding the obtained picture to the Mask R-CNN pre-training model. And we made the size of binary images consistent with the original one.

C. Feature Extraction

The skeleton feature is extracted to describe the movement of each joint in the entire gait cycle better. The previous research imply that there are significant differences between depressed individuals and normals in step length, arm swing and head movement [22], [32], [36]. Different from the Affective Features proposed in [42], we consider the head features on clinical manifestation and propose a novel skeleton feature, F_{ds}, for depression recognition. F_{ds} includes spatiotemporal features and kinematics features. The composition of two features are introduced separately below.

(1) Spatiotemporal features F_{st}. It includes the angle and distance between joints, as shown in Fig. 3. The angle features represent the angle between the joints in each frame. We consider not only the distance from the joint to the joint, but also the distance between the joint and the bone formed by two joints. The dimension of F_{st} is 21.

(2) Kinematics feature F_{dy}. It includes the speed, acceleration, and jerk of joints. We select the joints of the hands, ankles, and head as the observed joints. For these five joint points $p_j, j = 1, \ldots, 5$, we assumed that the coordinate vector of the j-th joint in the t-th frame is p_j^t, and we calculate the first-order partial derivative, the second-order partial derivative, and third-order partial derivative of p with respect to t respectively to obtain the speed, acceleration and jerk information. Finally, we combine

spatiotemporal features and kinematics features to get our final feature $Fds = Fst \cup Fdy$.

Regarding the research method in gait recognition, we use GEI as the silhouette feature for depression recognition. In this experiment, we calculated the GEI as Eq.1 serving as the silhouette features from different perspectives: front and side, to improve their classification performance.

D. Classification

1) Skeleton Modality: Inspired by the average speed and instant speed in physics, we propose two different strategies to construct classification models, named average strategy and sequence strategy.

(a) Average strategy: The averages of the skeleton features of a walking sequence are fed into the classification model. The dimension is the same as under a single frame. In this strategy, the features remain the same dimension as that of the single-frame but show the changes in a certain period of time. Since the experiment is a binary classification, we utilize three classic classification algorithms: K-Nearest Neighbors (K-NN) [47], Decision Tree (DT) [48], and Support Vector Machine (SVM) [49], which have been verified to have good performance.

(b) Sequence strategy: Considering that walking is a dynamic process, the features from adjacent frames can also make up a sequence. Here we establish the classification model by the LSTM network [50] with the input of the skeleton feature of a gait cycle.

2) Silhouette Modality: We propose a 4-layer CNN architecture, similar with ConvNets. There are four convolutional layers in the entire network, and pooling layers are appended after the second and third convolutional layers. After that, we added a fully connected layer and a softmax layer for output. In detail, raw data are divided into different segments per the gait cycle, if two segments are from two adjacent time series of the same subject, the prediction shall be strikingly similar. Therefore, to decrease the difference between the prediction of two adjacent gait sequences in the time series, we design a new loss function:

$$\begin{aligned}\text{loss}\left(y^j, p_k\left(X_t^j\right)\right) = &\sum_{k=1}^{2} -\log\left(p_k\left(X_t^j\right)\right)\cdot\delta\left(l_k\right)\\ &+\ \alpha\cdot\sum_{k=1}^{2} -\log\left(p_k\left(X_t^j\right)\right)\\ &\cdot p_k\left(X_{t+1}^j\right), \qquad (3)\end{aligned}$$

$$\delta\left(l_k\right) = \begin{cases} 1 & l_k = y^j, \\ 0 & l_k \neq y^j. \end{cases} \qquad (4)$$

where l_k represents the predicted label, $p_k(X_t^j)$ and $p_k(X_{t+1}^j)$ represents the probability of two adjacent gait periods predicted as the k-th category, α is the coefficient of the penalty term. Besides, y^j represents the true label of the sample j, the δ function show as Eq.4. Moreover, the proposed loss function is to make the model more stable in the data of adjacent gait cycles.

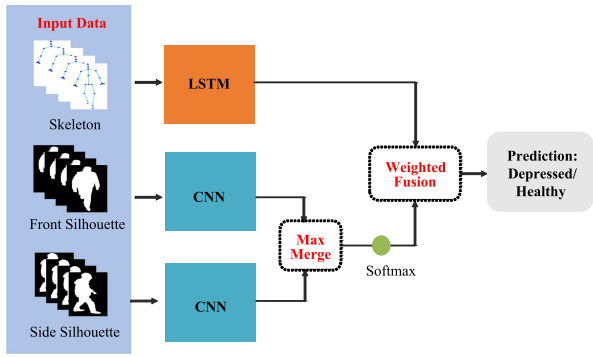

Fig. 4. Hybrid fusion model of skeleton and silhouette data. First, we regarded silhouette features of front and side perspectives as input, and at the feature-level, we fused the output of two CNNs and append a softmax layer. Then, at the decision-level, we fused the prediction results of skeleton and silhouette to obtain the final results.

3) Fusion Model: Since single-modal data failed to neglect the differences of individuals, the multi-modal fusion model can promote accuracy and generalization. Thus, we propose a hybrid fusion multi-modal classification model, including feature-level and decision-level fusion, as shown in Fig. 4. At the feature level, we merged the silhouette features of the front and side views via the CNN model. Then, the fused features are fed into a softmax layer to obtain the classification results.

In order to preserve the similarity of silhouette features from different views, we draw on the idea of the siamese network, [51] which shares weight parameters between two identical CNN networks. In the fusion method, we choose the maximum value fusion, which compares the value of each dimension and selects the maximum one as the value of this dimension. The decision-level fusion is the fusion of classification results of the skeleton and the silhouette after the feature-level fusion. And the strategy we used is weighted fusion. We define the recognition rate vectors P_i of two modalities as:

$$\boldsymbol{P}_i = \left(p_{i1}, p_{i2}\right)^{\mathrm{T}} \quad (i = 1, 2) \qquad (5)$$

where p_{i1} and p_{i2} denote the recognition rate of depression and health for the i-th modality respectively. Then we construct the weighting matrix:

$$\boldsymbol{W}_i = \begin{bmatrix} p_{i1} & 0 \\ 0 & p_{i2} \end{bmatrix} \quad (i = 1, 2) \qquad (6)$$

where $\boldsymbol{W}_i$ represents the weighting matrix of the i-th modality. Then the weighting matrix is multiplied with the recognition rate vector to get the result of linear weighted fusion C:

$$\begin{aligned} C &= \sum_{i=1}^{2} W_i P_i \\ &= \sum_{i=1}^{2} \begin{bmatrix} p_{i1} & 0 \\ 0 & p_{i2} \end{bmatrix} \begin{bmatrix} p_{i1} \\ p_{i2} \end{bmatrix} \\ &= \begin{bmatrix} \sum_{i=1}^{2} \left(p_{i1}\right)^2 \\ \sum_{i=1}^{2} \left(p_{i2}\right)^2 \end{bmatrix} \end{aligned} \qquad (7)$$

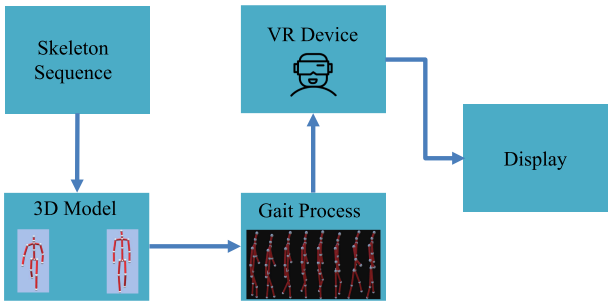

Fig. 5. The pipeline of skeleton visualization. First, we reconstructed 3D model of a person with skeleton sequence as input. Then we generated gait process of a gait cycle. Next with VR Device, we input our results to display the performance.

According to the maximum rule, the k-th state with the largest result is selected as the final recognition results:

$$\mathrm{MAX}_{j=1}^{2}\left\{\sum_{i=1}^{2}\left(p_{ij}\right)^{2}\right\}=\sum_{i=1}^{2}\left(p_{ik}\right)^{2} \tag{8}$$

E. Visualization

To visualize the skeleton sequence data and observe the dynamics, we build a VR application based on Unity to transform the data into the 3D model. In Fig. 5, at first, we built a 3D skeleton model via the three-dimensional coordinate. Since the six joints of the left-hand and right-hand parts and two of the feet affect the visualization performance, we removed these joints and used the remainder to construct spherical objects. Then we connected 17 joints with lines to construct a 3D human model. And then, we fed the model into a VR device. Besides showing the skeleton in the fixed position, we set up 9 cameras to observe the skeleton changes in the walking process from more perspectives. In which, 8 cameras will move with the human. They are located at the front, back, left, right, front right, front left, back right, and back left, respectively. We can switch between them through the buttons of the VR device to observe the model from different perspectives. The last camera can move freely, which means when we switch to this camera, we can set its position and angle to realize observation from any perspective.

IV. Experiments and Results

A. Implementation Details

K-fold cross-validation was applied to evaluate the performance of classification methods, here K=5. The dataset was divided into 5 subsets, in each trial, one of the subsets for testing and the remaining for training. The final performance is average of all trials. Other experimental setups are described in detail as following:

We applied three classification algorithms in average strategy. Regarding the parameters of K-NN, k is searched from 1 to 5, and the distance measurement method is Euclidean distance. Parameters of SVM are configured as the soft margin parameter $S=[10^{-3},10^{-2},\ldots,10^{3}]$, $\gamma=[10^{-5},10^{-4},\ldots,1000]$, and Gaussian kernel. In DT, the criteria for feature selection are information gain (Gain), information gain ratio (Gain Ratio) and Gini index (Gini), the parameter $max_{depth}=[3,4,\ldots,20]$. Furthermore, we applied the LSTM model for classification. As for parameter settings, the number of neurons of hidden layers is 128, dropout is set to 0.3, the activation function is softmax, and the loss function is cross-entropy. We use Adam optimization algorithm to train our model, where minibatch is 64 and epoch is 120.

As regards the silhouette data, we introduced a CNN model to extract the features for depression recognition. The model has 4 CNN layers, specifically, the numbers of convolution kernel are 32, 32, 64 and 128 respectively, the kernel size is all $3*3$, and the stride is 1. We append pooling layers after the second and third convolution layers. The pooling method is max-pooling with filters of size $2*2$ and the stride is 2. Finally, the number of neurons of the fully connected layer is 256.

TABLE II
Classification Results of Different Features and Classifiers Under Average Strategy

Feature	Classification	Accuracy	Sensitivity	Specificity	F1-score
F_{st}	K-NN	65.73%	79.21%	53.57%	68.67%
	DT	68.54%	73.27%	64.29%	68.84%
	SVM	68.08%	74.26%	62.50%	68.81%
F_{dy}	K-NN	70.42%	81.19%	60.71%	72.25%
	DT	70.89%	76.24%	66.07%	71.30%
	SVM	70.42%	78.22%	63.39%	71.49%
F_{ds}	K-NN	73.71%	**83.17%**	65.18%	75.00%
	DT	72.30%	72.28%	72.32%	71.22%
	SVM	**76.53%**	77.23%	75.89%	**75.73%**
Affective Features [42]	K-NN	71.36%	59.41%	**82.14%**	66.30%
	DT	73.71%	69.31%	77.68%	71.43%
	SVM	73.24%	67.33%	78.57%	70.47%

* The bold indicates the optimal result under the same index, the same as follow.

B. Results From Single-Modal Detection

Firstly, we have designed the experiment for comparison of different features extracted from skeleton data. Table II shows the classification performance of each classifier with different features by the average strategy. Initially, we have conduct an ablation study to compare the F_{st} calculated from 17 joints, F_{dy} calculated from 5 joints, and the fused feature F_{ds}. It can be seen from the table, the performance of F_{dy} is better than that of F_{st}. What's more, the fused feature, F_{ds}, improves each metric compared with single features. Among them, the accuracy of F_{ds} is up to 76.53% in the SVM classifier. The best performance of F_{st} is 79.21% while F_{dy} achieves the height sensitivity in K-NN classifiers, which is 83.17%. The highest f1-score in F_{ds} is 75.73%. In addition, we have compared the proposed F_{ds} with the Affective Features [42]. F_{ds} performs better in accuracy, sensitivity, and f1-score.

Then, we want to verify the efficacy of the proposed sequence strategy. By compared Table II and Table III , the same features taking sequence strategy outperform the average strategy. Furthermore, we have compared F_{ds} with other two features, that

TABLE III
CLASSIFICATION RESULTS OF DIFFERENT FEATURES UNDER SEQUENCE STRATEGY

Feature	Accuracy	Sensitivity	Specificity	F1-score
F_{ds}	**80.28%**	80.20%	**80.36%**	**79.41%**
Position Vectors [40]	74.18%	**89.11%**	60.71%	76.60%
Affective Features [42]	77.93%	85.15%	71.43%	78.54%

TABLE IV
CLASSIFICATION RESULTS BASED ON SILHOUETTE FROM DIFFERENT PERSPECTIVES

Perspectives	Accuracy	Sensitivity	Specificity	F1-score
Front	55.87%	53.47%	58.04%	53.47%
Side	**66.67%**	**64.52%**	**68.18%**	**61.54%**

TABLE V
CLASSIFICATION RESULTS UNDER DIFFERENT FUSION STRATEGIES

Feature fusion	Decision fusion	Accuracy	Sensitivity	Specificity	F1-score
Concatenation	Weighted Fusion	82.63%	87.13%	78.57%	82.63%
	Max Fusion	81.69%	84.16%	79.46%	81.34%
Max Merge	Weighted Fusion	**85.45%**	**91.09%**	80.36%	**85.58%**
	Max Fusion	82.16%	79.21%	**84.82%**	80.81%

is, the Position Vector [40] designed for frame-level matching and Affective Features [42]. As depicted in the Table III,the proposed features F_{ds} performs better than the Position Vector and the Affective Features in accuracy, specificity and f1-score. The specificity is significantly higher than the other two features with a value of 80.36%, which is higher 8.93% than that with Affective Features. Yet, the difference between the three f1-scores is unobvious. In terms of sensitivity, the Position Vectors and Affective Features perform better than F_{ds}, and the best results is 89.11%.

The Performances to detection using GEI from different views show in the Table IV. The classification results of the side views are better than the front views in all metrics, yet the best accuracy is only 66.67%

C. Results From Multi-Modal Detection

To confirm our hypothesis that the data of the skeleton and the silhouette are complementary to each other, we have trained the multi-modal model. According to the Table III, Table IV, and Table V, the multi-modal models are more accurate than the single-modal models, whatever taken any the strategies. What's more, Except for the fusion methods we proposed, we have compared our proposal with other common ones. At the feature-level fusion, concatenating features with identical dimensions has been applied in many neural networks [7], and we compared it with our methods, max merge. Regarding to decision-level fusion, we have compared our weighted fusion with the max fusion which chooses the class of the maximum probability as the prediction result. Consequently, the maximum merging strategy performs better than the concatenating one in the fusion strategy at the feature level. The accuracy of the concatenating one is 82.63%, while the accuracy of the maximum merging strategy reaches 85.45%. At the decision level, the weighted fusion strategy is superior to the maximum fusion one in the accuracy, sensitivity, and f1-score. In terms of sensitivity, the weighted fusion strategy is 11.88% higher than the maximum fusion one. It indicates that the performance of this strategy is more competent than the maximum fusion.

As we are aware, males and females tend to have dissimilar gaits. We wonder how genders impact the recognition rates under discrepant modalities or multi-modal, so we have compared the recognition results under different genders, as shown in Fig. 6. There are significant differences in the sensitivity, specificity, and f1-score of the two single-modal models. The sensitivity in females is higher than the males, with the highest difference of about 20%. By contrast, the specificity in males is higher than the females. In the fusion modality, the specificity in males is the highest, reaching 85.71%. Also, compared with single modalities, the difference in the multi-modal model of the accuracy and the f1-score resulting from different genders reduce significantly. Also, there are slight reductions in sensitivity and specificity results.

D. Visualization

In Fig. 7, we have displayed the skeleton images and skeleton features between healthy control and people at risk of depression. The skeleton image demonstrates that people at risk of depression are prone to reduce their arms swing and stride. From the histograms in the figure, we can see the difference more clearly. The histograms show distributions of the first angle, the fourth angle, and the eighth angle in Fig. 3(a), respectively. It can be seen from the first two histograms that the distributions of people at risk of depression are more concentrated, indicating that compared with normal people, their head movement and arm swing are reduced. The peaks of people at risk of depression appear at a greater angle from the third histogram, which implies that the shoulders of depression patients are in a more relaxed state.

V. DISCUSSIONS

Competent skeleton features are proposed, especially with sequence strategy. Interestingly, the F_{dy} only contains kinematics parameters from five joints outstrips in the sensitivity the F_{st} and Affective Features calculated from 21 joints. In [20], rigid-body representations with 12 joints selected by mechanical energy achieved the best accuracy. Moreover, the five joints calculating F_{ds} show great mechanical energy relatively. By comparing the results of single features and the fused feature, we verified the effectiveness of the proposed feature from the perspective of an ablation study. But the fused features F_{ds} calculated from 17 joints perform well than the others, which implies that spatial information is also important [41]. Furthermore, the F_{ds} outperforms others in depression detection both in average and sequence strategy, which may be due to considering more clinical manifestation. Compared with silhouette data, skeleton data can describe dynamic changes during walking more accurately.

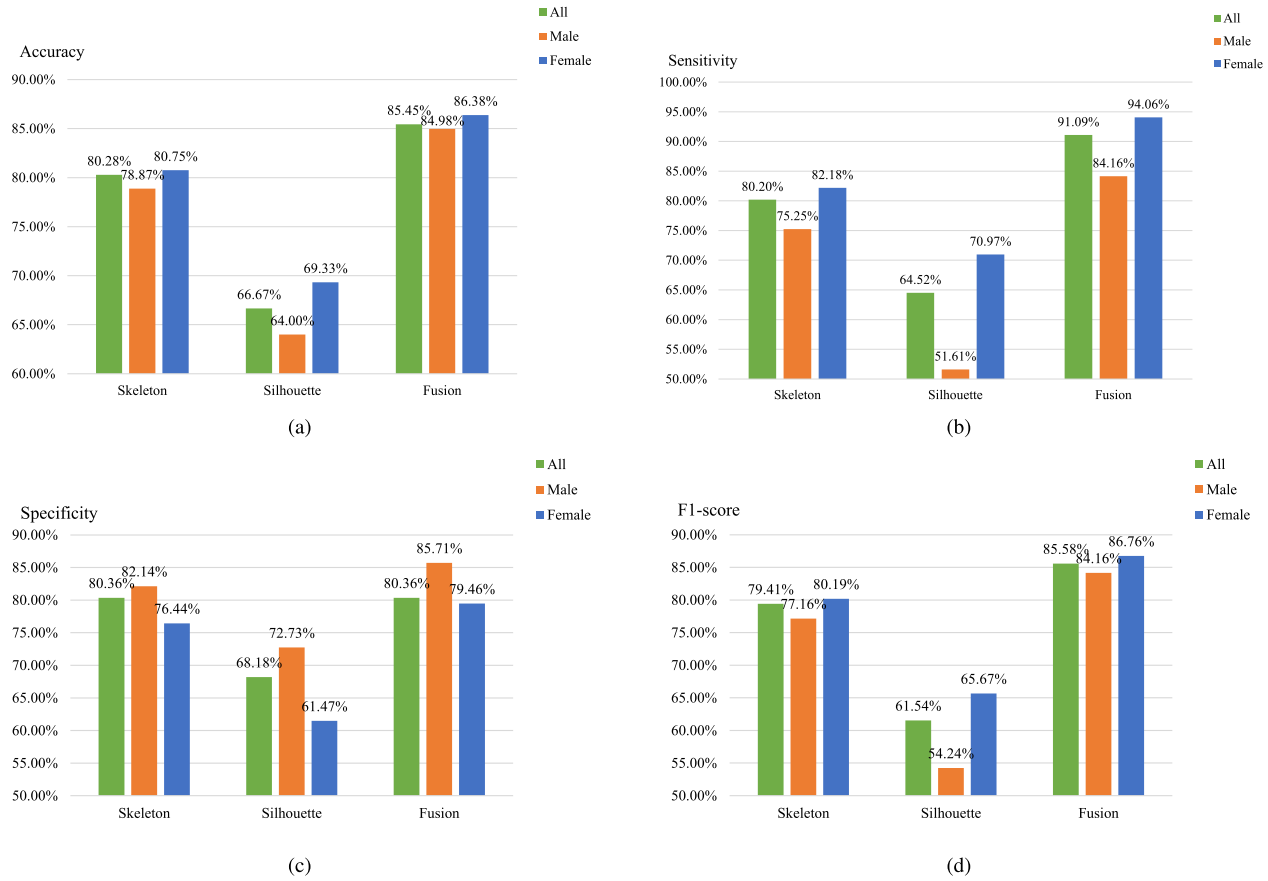

Fig. 6. The performance metrics under the gender difference, subgraph (a), (b), (c), (d) represents the accuracy, f1-score, sensitivity and specificity performance under the gender difference, respectively.

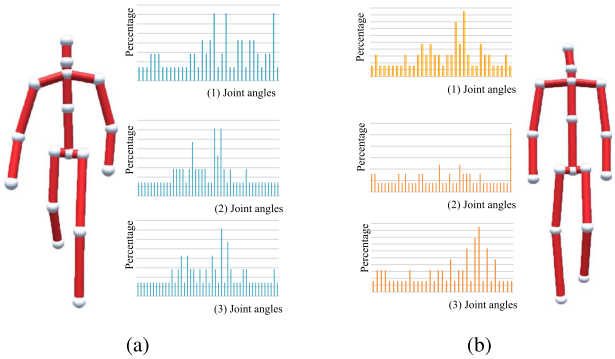

Fig. 7. VR Visualization of skeleton: (a) controls (b) people at risk of depression. The histogram from top to bottom represents the frequency distributions of three features respectively: (1) Joint angles of the head, neck, and spineshoulder. (2) Joint angles of the shoulder, elbow, and wrist in right. (3) Joint angles of the spinemid and shoulders.

Table II and Table III show that the classification results based on silhouette data still are not comparable to that of the skeleton data, even with the silhouette from side views which perform best in silhouette data. Then, we have first endeavored to use silhouette data to detect depression, although the results are discouraging in single-modal detection while contributes to the multi-modal. It can be seen that the classification results of the side views are better than the front views which is consistent with the results of many studies [26], [52] that the silhouette data in front views fail to reflect walking state well. In contrast, our skeleton data suit for front views. The two modalities may complement each other in perspectives. We obtained better classification performance using skeleton and silhouette fusion, which agrees that the skeleton and silhouette information are related and complementary. Comparing the fusion strategies, the maximum merging strategy and weighted fusion are intelligible and powerful.

By the complementarity of skeleton and silhouette, the model effectively reduces the impact of the genders, also improves the generalization ability. From Fig. 6, there are significant differences in the classification results of single-modal depression recognition under different genders. The fusion model proposed in this paper reduces the classification difference under different genders. There are few studies on the classification tasks using the fusion of skeleton and silhouette modalities, but in the field of 3D portrait model generation [53], researchers found that the skeleton can help to locate the spatial position of joints of the human body, and the silhouette can describe the detailed information of the human body shape so that the constructed 3D model is more vivid. It agrees that there is complementary information between the skeleton and the silhouette modality

and also verifies that the experimental results of this paper are valid.

The visualization of the skeleton information shows a skeleton sequence of a gait cycle and skeleton features. It provides convenience for understanding the gait abnormalities of patients with depression. In this way, on the one hand, to improve the interpretability of the model by comparing the skeleton information of people at risk of depression and healthy people. For instance, Fig. 7 illustrates the reduced arms swing and stride of people at risk of depression [35]. On the other hand, to facilitate the communication between doctors and patients through visualization to provide more reasonable explanations.

VI. Conclusion

In this article, we aimed at designing a non-contact depression recognition method to help GPs classify postgraduate students at risk of depression more effective and convenient. Our research proffered a new pipeline to analyze the gait data, combing depth and RGB information. The experimental results imply that the proposed method leveraging the complementarity gained a higher accuracy with reducing the gender differences. Concretely, for skeleton data, we proposed an effective feature set, called F_{ds}. As for RGB images, we extracted human silhouette and generated GEI. Then we proposed a hybrid fusion model with skeleton and silhouette to detect depression. To confirm our proposal, we constructed the skeleton and RGB dataset consisting of 200 participants. Furthermore, our best accuracy is 85.45%. Last but not least, to help GPs understand the classification results, we proposed the gait information visualization system that enables GPs to observe the walking sequence and compare with normal walking.

In the future, in order to obtain more reliable depression recognition result, we plan to collect gait data from more perspectives to analyze the mental state. Furthermore, considering the influence of demographic information, such as age and gender will be considered in the subsequent study to improve the classification. In conclusion, our research implies that silhouette data help the gait analysis. What's more, we hope the multi-modal depression recognition can help the depressed patients get an early intervention and treatment.

References

[1] "World Health Organization," Accessed: Aug. 11, 2020. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/depression

[2] K. M. Cooper, L. E. Gin, M. E. Barnes, and S. E. Brownell, "An exploratory study of students with depression in undergraduate research experiences," *CBE-Life Sci. Educ.*, vol. 19, no. 2, 2020.

[3] L. Guo *et al.*, "Prevalence and changes in depressive symptoms among postgraduate students: A systematic review and meta-analysis from1980 to 2020," *Stress Health*, Mar. 26, 2021.

[4] T. M. Evans, L. Bira, J. B. Gastelum, L. T. Weiss, and N. L. Vanderford, "Evidence for a mental health crisis in graduate education," *Nat. Biotechnol.*, vol. 36, no. 3, pp. 282–284, 2018.

[5] D. Moreno-Agostino, Y.-T. Wu, C. Daskalopoulou, M. Hasan, M. Huisman, and M. Prina, "Global trends in the prevalence and incidence of depression: A systematic review and meta-analysis," *J. Affect. Disord.*, vol. 281, pp. 235–243, 2020.

[6] K. Usher, J. Durkin, and N. Bhullar, "The COVID-19 pandemic and mental health impacts," *Int. J. Ment. Health Nurs.*, vol. 29, no. 3, pp. 315–318, 2020.

[7] M. Woźniak, J. Siłka, and M. Wieczorek, "Deep neural network correlation learning mechanism for CT brain tumor detection," *Neural Comput. Appl.*, pp. 1–16, 2021. [Online]. Available: https://doi.org/10.1007/s00521-021-05841-x

[8] X. Zhang, J. Shen, Z. U. Din, J. Liu, G. Wang, and B. Hu, "Multimodal depression detection: Fusion of electroencephalography and paralinguistic behaviors using a novel strategy for classifier ensemble," *IEEE J. Biomed. Health Informat.*, vol. 23, no. 6, pp. 2265–2275, Nov. 2019.

[9] S. Harati, A. Crowell, Y. Huang, H. Mayberg, and S. Nemati, "Classifying depression severity in recovery from major depressive disorder via dynamic facial features," *IEEE J. Biomed. Health Informat.*, vol. 24, no. 3, pp. 815–824, 2020.

[10] H. Jelodar, Y. Wang, R. Orji, and S. Huang, "Deep sentiment classification and topic discovery on novel coronavirus or COVID-19 online discussions: NLP using LSTM recurrent neural network approach," *IEEE J. Biomed. Health Informat.*, vol. 24, no. 10, pp. 2733–2742, Oct. 2020.

[11] B. Saha, T. Nguyen, D. Phung, and S. Venkatesh, "A framework for classifying online mental health-related communities with an interest in depression," *IEEE J. Biomed. Health Informat.*, vol. 20, no. 4, pp. 1008–1015, Jul. 2016.

[12] J. Shen *et al.*, "An optimal channel selection for EEG-based depression detection via kernel-target alignment," *IEEE J. Biomed. Health Informat.*, vol. 25, no. 7, pp. 2545–2556, Jul. 2021.

[13] H. Cai *et al.*, "A pervasive approach to eeg-based depression detection," *Complexity*, vol. 2018, 2018. [Online]. Available: https://doi.org/10.1155/2018/5238028

[14] A. Zhao *et al.*, "Multimodal gait recognition for neurodegenerative diseases," *IEEE Trans. Cybern.*, to be publish, doi: 10.1109/TCYB.2021.3056104.

[15] K. Hu *et al.*, "Vision-based freezing of gait detection with anatomic directed graph representation," *IEEE J. Biomed. Health Informat.*, vol. 24, no. 4, pp. 1215–1225, Apr. 2020.

[16] Z. You *et al.*, "Alzheimer's disease classification with a cascade neural network," *Front. Public Health*, vol. 8, 2020, Art. no. 665.

[17] M. Woźniak, M. Wieczorek, J. Siłka, and D. Połap, "Body pose prediction based on motion sensor data and recurrent neural network," *IEEE Trans. Ind. Informat.*, vol. 17, no. 3, pp. 2101–2111, Mar. 2021.

[18] F. Deligianni, Y. Guo, and G.-Z. Yang, "From emotions to mood disorders: A survey on gait analysis methodology," *IEEE J. Biomed. health Informat.*, vol. 23, no. 6, pp. 2302–2316, Nov. 2019.

[19] R. Feldman, S. Schreiber, C. Pick, and E. Been, "Gait, balance and posture in major mental illnesses: Depression, anxiety and schizophrenia," *Austin Med. Sci*, vol. 5, no. 1, 2020, Art. no. 1039.

[20] H. Lu, W. Shao, E. Ngai, X. Hu, and B. Hu, "A new skeletal representation based on gait for depression detection," in *Proc. IEEE Int. Conf. E-Health Netw., Appl. Serv.*, 2021, pp. 1–6.

[21] N. Zhao *et al.*, "See your mental state from your walk: Recognizing anxiety and depression through Kinect-recorded gait data," *PLoS One*, vol. 14, no. 5, 2019, Art. no. e0216591.

[22] J. Fang *et al.*, "Depression prevalence in postgraduate students and its association with gait abnormality," *IEEE Access*, vol. 7, pp. 174425–174437, 2019.

[23] B. Sun, Z. Zhang, X. Liu, B. Hu, and T. Zhu, "Self-esteem recognition based on gait pattern using Kinect," *Gait Posture*, vol. 58, pp. 428–432, 2017.

[24] F. Gholami, D. A. Trojan, J. Kövecses, W. M. Haddad, and B. Gholami, "A microsoft Kinect-based point-of-care gait assessment framework for multiple sclerosis patients," *IEEE J. Biomed. Health Informat.*, vol. 21, no. 5, pp. 1376–1385, Sep. 2017.

[25] H. Chao, Y. He, J. Zhang, and J. Feng, "Gaitset: Regarding gait as a set for cross-view gait recognition," *Proc. AAAI Conf. Artif. Intell.*, vol. 33, no. 1, pp. 8126–8133, 2019.

[26] Z. Wu, Y. Huang, L. Wang, X. Wang, and T. Tan, "A comprehensive study on cross-view gait based human identification with deep CNNs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 2, pp. 209–226, Feb. 2017.

[27] J. Han and B. Bhanu, "Individual recognition using gait energy image," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 2, pp. 316–322, Feb. 2006.

[28] S. Chen, J. Lach, B. Lo, and G.-Z. Yang, "Toward pervasive gait analysis with wearable sensors: A systematic review," *IEEE J. Biomed. Health Informat.*, vol. 20, no. 6, pp. 1521–1537, Nov. 2016.

[29] M. R. Lemke, T. Wendorff, B. Mieth, K. Buhl, and M. Linnemann, "Spatiotemporal gait patterns during over ground locomotion in major depression compared with healthy controls," *J. Psychiatr. Res.*, vol. 34, no. 4/5, pp. 277–283, 2000.

[30] J. A. Stamford, P. N. Schmidt, and K. E. Friedl, "What engineering technology could do for quality of life in Parkinson's disease: A review of current needs and opportunities," *IEEE J. Biomed. Health Informat.*, vol. 19, no. 6, pp. 1862–1872, Nov. 2015.

[31] N. T. Dragašević-Mišković *et al.*, "Impact of depression on gait variability in Parkinson's disease," *Clin. Neurol. Neurosurgery*, vol. 200, 2021, Art. no. 106324.

[32] R. Briggs *et al.*, "Do differences in spatiotemporal gait parameters predict the risk of developing depression in later life?," *J. Amer. Geriatrics Soc.*, vol. 67, no. 5, pp. 1050–1056, 2019.

[33] F. R. Marino *et al.*, "Gait speed and mood, cognition, and quality of life in older adults with atrial fibrillation," *J. Amer. Heart Assoc.*, vol. 8, no. 22, pp. 1–8, 2019.

[34] F. Pieruccini-Faria, S. W. Muir-Hunter, and M. Montero-Odasso, "Do depressive symptoms affect balance in older adults with mild cognitive impairment? results from the 'gait and brain study'," *Exp. Gerontol.*, vol. 108, pp. 106–111, 2018.

[35] J. Michalak, N. F. Troje, J. Fischer, P. Vollmar, T. Heidenreich, and D. Schulte, "Embodiment of sadness and depression-gait patterns associated with dysphoric mood," *Psychosomatic Med.*, vol. 71, no. 5, pp. 580–587, 2009.

[36] J. B. Sanders, M. A. Bremmer, H. C. Comijs, D. J. Deeg, and A. T. Beekman, "Gait speed and the natural course of depressive symptoms in late life; an independent association with chronicity?," *J. Amer. Med. Directors Assoc.*, vol. 17, no. 4, pp. 331–335, 2016.

[37] R. Briggs, D. Carey, R. A. Kenny, and S. P. Kennelly, "What is the longitudinal relationship between gait abnormalities and depression in a cohort of community-dwelling older people? data from the Irish Longitudinal Study on Ageing (TILDA)," *Amer. J. Geriatr. Psychiatry*, vol. 26, no. 1, pp. 75–86, 2018.

[38] M. B. Murri *et al.*, "Instrumental assessment of balance and gait in depression: A systematic review," *Psychiatry Res.*, vol. 284, 2020, Art. no. 112687.

[39] S. Radovanović, M. Jovičić, N. P. Marić, and V. Kostić, "Gait characteristics in patients with major depression performing cognitive and motor tasks while walking," *Psychiatry Res.*, vol. 217, no. 1/2, pp. 39–46, 2014.

[40] S. Choi, J. Kim, W. Kim, and C. Kim, "Skeleton-based gait recognition via robust frame-level matching," *IEEE Trans. Inf. Forensics Secur.*, vol. 14, no. 10, pp. 2577–2592, Oct. 2019.

[41] T. Wang *et al.*, "A gait assessment framework for depression detection using Kinect sensors," *IEEE Sensors J.*, vol. 21, no. 3, pp. 3260–3270, Feb. 2021.

[42] U. Bhattacharya *et al.*, "Take an emotion walk: Perceiving emotions from gaits using hierarchical attention pooling and affective mapping," in *Proc. Eur. Conf. Comput. Vis.*, Springer, 2020, pp. 145–163.

[43] M. Chiu, J. Shu, and P. Hui, "Emotion recognition through gait on mobile devices," in *Proc. IEEE Int. Conf. Pervasive Comput. Commun. Workshops*, 2018, pp. 800–805.

[44] X. Wang and W. Q. Yan, "Human gait recognition based on frame-by-frame gait energy images and convolutional long short-term memory," *Int. J. Neural Syst.*, vol. 30, no. 1, 2020, Art. no. 1950027.

[45] Z. Zhang, J. Chen, Q. Wu, and L. Shao, "GII representation-based cross-view gait recognition by discriminative projection with list-wise constraints," *IEEE Trans. Cybern.*, vol. 48, no. 10, pp. 2935–2947, Oct. 2018.

[46] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2961–2969.

[47] T. Cover and P. Hart, "Nearest neighbor pattern classification," *IEEE Trans. Inf. Theory*, vol. 13, no. 1, pp. 21–27, Jan. 1967.

[48] S. R. Safavian and D. Landgrebe, "A survey of decision tree classifier methodology," *IEEE Trans. Syst., Man, Cybern.*, vol. 21, no. 3, pp. 660–674, May/Jun. 1991.

[49] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, pp. 1–27, 2011. [Online]. Available: https://doi.org/10.1145/1961189.1961199

[50] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[51] J. Bromley *et al.*, "Signature verification using a 'siamese' time delay neural network," *Int. J. Pattern Recognit. Artif. Intell.*, vol. 7, no. 4, pp. 669–688, 1993.

[52] B. Li, C. Zhu, S. Li, and T. Zhu, "Identifying emotions from non-contact gaits information based on microsoft Kinects," *IEEE Trans. Affective Comput.*, vol. 9, no. 4, pp. 585–591, Oct.–Dec. 2018.

[53] E. Dibra, H. Jain, C. Öztireli, R. Ziegler, and M. Gross, "HS-Nnets: Estimating human body shape from silhouettes with convolutional neural networks," in *Proc. 4th Int. Conf. 3D Vis.*, 2016, pp. 108–117.