Hindawi
Complexity
Volume 2021, Article ID 1589716, 15 pages
https://doi.org/10.1155/2021/1589716

Research Article

GaitVision: Real-Time Extraction of Gait Parameters Using Residual Attention Network

Mohammad Farukh Hashmi,[1] **B. Kiran Kumar Ashish**,[2] **Prabhu Chaitanya**,[3] **Avinash Keskar**,[4] **Sinan Q. Salih**,[5,6,7] **and Neeraj Dhanraj Bokde**[8]

[1]*Department of Electronics and Communication Engineering, National Institute of Technology, Warangal, India*
[2]*Computer Vision, Barcelona, Spain*
[3]*University of North Texas, Denton, TX, USA*
[4]*Department of Electronics and Communication Engineering, Visvesvaraya National Institute of Technology, Nagpur, India*
[5]*Institute of Research and Development, Duy Tan University, Da Nang 550000, Vietnam*
[6]*Computer Science Department, Dijlah University College, Al-Dora, Baghdad, Iraq*
[7]*Artificial Intelligence Research Unit (AIRU), Dijlah University College, Al-Dora, Baghdad, Iraq*
[8]*Department of Mechanical and Production Engineering-Renewable Energy and Thermodynamics, Aarhus University, Aarhus 8000, Denmark*

Correspondence should be addressed to Sinan Q. Salih; sinanq.salih@duytan.edu.vn

Received 22 April 2021; Accepted 28 October 2021; Published 28 November 2021

Academic Editor: Salvatore Cuomo

Gait walking patterns are one of the key research topics in natural biometrics. The temporal information of the unique gait sequence of a person is preserved and used as a powerful data for access. Often there is a dive into the flexibility of gait sequence due to unstructured and unnecessary sequences that tail off the necessary sequence constraints. The authors in this work present a novel perspective, which extracts useful gait parameters regarded as independent frames and patterns. These patterns and parameters mark as unique signature for each subject in access authentication. This information extracted learns to identify the patterns associated to form a unique gait signature for each person based on their style, foot pressure, angle of walking, angle of bending, acceleration of walk, and step-by-step distance. These parameters form a unique pattern to plot under unique identity for access authorization. This sanitized data of patterns is further passed to a residual deep convolution network that automatically extracts the hierarchical features of gait pattern signatures. The end layer comprises of a Softmax classifier to classify the final prediction of the subject identity. This state-of-the-art work creates a gait-based access authentication that can be used in highly secured premises. This work was specially designed for Defence Department premises authentication. The authors have achieved an accuracy of 90% ± 1.3% in real time. This paper mainly focuses on the assessment of the crucial features of gait patterns and analysis of gait patterns research.

1. Introduction

Fingerprints and faces are unique to everyone [1]; hence, they are well suited for current biometric systems. Similarly, gait refers to a person's walking pattern, which is a unique characteristic of each individual. Gait is a complex biological process and a distinctive walking style, which cannot be spoofed or imitated exactly. These unique patterns make it ideal for user identification and authentication [2]. Generally, a biometric system compares the information registered for proof of identity to a person's current features. This corresponds to the concept of one-to-one matching with a matching rate greater than 95%. Existing solutions achieve a decent accuracy; however, the concept of spoofing, which describes the mimicking or realistic imitation of the original pose of a person, constitutes a threat to access authentication. The latest facial recognition techniques for biometrics are threatened similarly in terms of facial spoof

attacks. Many datasets of possible multiple-scenario spoof attacks have been released by CASIA and the ROSE lab, but there remains an echo of the danger of spoofing. RFID or access cards and sensors [3] can easily break the red line in any authentication system. In this paper, the overcoming of the pitfalls of natural biometrics using an advanced method called gait biometrics is discussed.

However, gait features are purely dependent on visual appearance, and this can cause a problem if there are slight variations in color, contrast, walking speed, and low-resolution imagery. These problems are sensitive to gait computation to pool in the results. The background color contributes to the disruption and slows down the computation with regard to classification. These are some of the major problems that contribute to a back-step in using gait biometrics. The authors address the issues faced during implementation and resolve unnecessary features, such as isolating targets from the background, re-resolution of low-imagery by turning to high frames per second (FPS), and blur-free motion in the case of speed walking. All these techniques contribute to the accuracy of pose estimation. The design of gait features should be invariant with regard to clothing (color difference), viewing angle, and so on. Therefore, this work aims to disentangle gait features by isolating the target (walking person) from a background in visual appearance and converting the frames to a high frame rate, that is, 240 FPS for a motion blur-free visual to avoid pixel corruption or distortion, which degrades the accuracy [4, 5].

To avoid the overlapping issue concerning the human gait, existing solutions use a 2-stage extractor; the first one is person detection extraction with a unique ID tagged on him/her [6, 7]. The second stage is human pose extractions inside the boundary box. Unique ID tagging would erase pose overlapping problems [7, 8]. It has also been demonstrated that gait recognition performance is significantly influenced by different intraclass variations related to the subject, such as shadows, walking surfaces, angle variations, environment variations, clothing, and segmentation errors [5, 9].

A very common factor in human vision is that people are often able to identify a familiar person from a certain distance simply based on their walking style. It is even common to be able to mimic a person's walking style [10, 11]. Although walking styles can be mimicked, microscopic parameters, such as leg pressure, angle of the walk, and distance of each step, cannot be accurately imitated. Thus, there has been a growing interest in natural biometrics and researchers have exploited it for identification purposes. Initially, the inability of humans to mimic microscopic gait features aroused interest in defense research for secured access authentication. Each step may appear to be of a similar style on the macrolevel of observation; however, on the microlevel of vision, parameters have fewer variations in the gait. These microlevel variations are unique to each frame and position. This work proposes an end-to-end deep learning technique to extract temporal information from the gait in each frame and position, which is frame-level feature extractions from each silhouette independently [12].

The main contributions of this work include the following:

(1) Identify a registered "subject" through his/her gait patterns.

(2) Perform basic and powerful statistical methods such as mean, median, and max in attention mechanism rather than other activation layers. This is to make the feature levels as simple as possible since gait patterns would be biased and similar to each other often.

(3) Train the network with attention model rather than conventional CNN or transfer learning for high-level spatiotemporal feature extraction.

The list of abbreviations used in this manuscript are shown in Table 1.

2. Related Work

In this section, the authors will describe the cycle of gait biometrics and deep learning-based models.

2.1. Human Gait Analysis

2.1.1. Gait Cycle. A gait cycle describes the repetitive patterns of the human walking posture. A cycle outlines the postures between successive time instances of foot-to-floor contact. These contact points in relation to certain gait parameters are essential for gait analysis. The gait cycle mainly consists of 2 phase cycles: (i) stance and (ii) swing. These two giant parameters contribute approximately ~80% to a complete gait motion analysis. The gait stance covers 60% of the cycle, especially for the walking motion. However, when a person runs, the major proportion is in the swing phase. Feature extraction in the time domain includes variations in intrinsic properties while walking, such as velocity, motion, body length, width, and bend angle. This provides the intrinsic patterns of the walking cycle of a person, which is of extreme importance in recognition. This is how pattern information is extracted from walking cycles.

2.1.2. Characteristics of Gait Cycle. It has been demonstrated and sufficiently proven that human gait is unique to each person. Pattern differentiation of each individual, especially features, such as the pelvis and thorax, is completely different for each individual. This temporal information can be used to adapt computer vision-based biometrics without any intrinsic hardware equipment. The gait cycle of a subject can be broadly divided into 2 categories: right leg strikes and left leg strikes. Temporal information can be extracted from these 2 phases, mainly depicting heel strikes. This information contains much more discriminative information for the differentiation between 2 subjects. A complete heel strike begins with a lifting of the heel in the forward direction and a swing in the backward direction and then, the cycle is repeated.

2.1.3. Uniqueness of Gait. Similar to the uniqueness of fingerprints, the gait posture of each individual is distinctive. Often, the concern regarding spoof attacks arises in natural biometrics, but an exact mimicking of a "target subject's" gait posture is impossible. A subject's walking

Table 1: The list of abbreviations used in this manuscript.

2D	2-dimensional
3D	3-dimensional
CPU	Central processing unit
CNN	Convolutional neural network
FN	False negative
FP	False positive
FPS	Frame per second
GAN	Generative adversarial network
GPU	Graphics processing unit
ID	Identification
RAM	Random-access memory
ReLu	Rectified linear unit
RFID	Radio-frequency identification
RGB	Red-green-blue
RGB-D	Red-green-blue depth
SDTN	Spatial detransformer network
SPPE	Single-person pose estimator
STN	Spatial transformer network
TP	True positive
VAE	Variational autoencoders

style can be mimicked easily, but only at the macrolevel; the microlevel characteristics, such as foot pressure, heel strike angles, and distance maintained between each step, are impossible to mimic accurately. This fact is sufficient evidence for the expected success for a full implementation of gait biometrics for access authentication systems. There may be minute variations in walking posture, but this would not be a natural change; rather, this would be a forced change similar to that when the movie stars completely transform their posture [13].

2.2. Gait Representation in Biometrics. Unlike other vision-based models, gait does not entirely rely on pixel values. Metrics can be extracted from RGB, RGB-D, or binary frames. Metrics are purely based on the structure of the pixels but not on the intensity. Hence, for better results, the authors converted the frames into binary and computed the subject based on the posture. These frames are generally known as masks and are obtained by gait energy/entropy images, defined by GEnI [14]. These images extract the silhouette masks of a target. The next extracted metric is the kinematic 2D body joint points. Conventional drawbacks, such as clothing and walking speed, can be resolved by this approach given high-resolution frames for extracting rich information. This method of extraction is proven to be robust to covariates, such as clothing, walking speed, and view angle, because high-resolution frames are used for computation.

The authors of this work are investigating the training of patterns rather than completely training with data-based on pixel-based features. This is contrary to the CNN model, which extracts millions of parameters, rather than dealing with high-computation parameters. This novel approach can be utilized to train a model for different scenarios, such as pattern recognition. This work involves patterns, and the backbone is patterns extracted from the data of each subject.

2.3. Disentanglement Learning. Disentanglement learning is a new form of gait approach in feature extraction using intense computation resources. Existing models are mostly semantic latent vectors of data (features) from CNN architectures. Disentanglement learning is gaining popularity in the computer vision realm for its pure data-driven approach. One of the outstanding networks in disentanglement learning is DrNet, which uses pose vectors with a two-encoder architecture. The content information was removed by generative adversarial training. Another approach, which includes segmentation and conducts analysis, moderates foreground segment masks of body parts from 2D pose joints using a U-Net architecture. These body part segments are transformed into the desired motion with adversarial training. Esser et al. [15] utilized U-Net and variational autoencoders (VAE) to disentangle an image into appearance and shape. Tran et al. [16, 17] attained state-of-the-art performances on pose-invariant facial recognition by explicitly disentangling variations in a pose with a multitask GAN [18, 19]. Further, DR-GAN [17] implicates adversarial training with pose labels to disentangle pose features.

2.4. Single Image-Based Action Recognition. Bhandari et al. [20] implemented simple image-based action recognition based on an HRNet [21] human pose estimation network. HRNet [21] represents multitasking features with a set of feature maps extracted from an image in the decreasing order of resolution and increasing order of channels [20]. The model returns heat maps and human joints for action recognition [14].

2.5. Attention Image-Based Feature Stream. Bhandari et al. [20] and Fukui et al. [22] used attention mechanisms such as attention-based image feature extraction streams foreground analysis. This method uses an attention image-based feature stream [23], which is built on top of ResNet18 [24]. The shallow stream is aided by skip connections from the feature maps extracted by HRNet [21]. These feature maps are concatenated with each output of ResNet18 [24] layer using transition blocks.

2.6. Part-Image-Based Feature Stream. For accurate action recognition using each part, Bhandari et al. [20] proposed a part-image-based feature stream. This implies feature extraction from HRNet [21], which is used for classifying body parts through the "Conv for pose estimation" block. These individual parts are then fed to ResNet18 [8] for classification.

2.7. Gait Datasets. There are many conventional open-source datasets on gait. These are large in quantity and high in quality. A few examples are the SOTON large database, USF, CASIA-B, OU-ISIR, and TUM GAID [25]. The authors use CASIA-B [26], a custom-collected dataset for this work. Here, architectural performance is analyzed rather than compared with different state-of-the-art techniques in different databases. Hence, a largely custom-collected dataset is

used for testing and metrics extraction. CASIA-B is a multiview dataset with 3 variations of each subject in terms of view angle, clothing, and carrying. The dataset contains 11 different view angles of each subject in the walking posture. A sample of the CASIA-B dataset is shown in Figure 1.

2.8. Limitations. Most of the state-of-the-art methods, as mentioned above, have performed experiments based on subject-sequence frames image classification using conventional convolutional neural networks and transfer learning methods. Most of these methods achieved good results but not accurate results on all subjects. Few of them have carried out attention mechanism using pretrained transfer learning alone, which lacked in extracting required features but alongside extracted even unwanted features, which diluted the feature learning. The above limitations described are maximumly focussed on a single architecture image classification with a given dataset of different subjects. Hence, even if a slight variation is observed in the subject in terms of angle and distance, the predictions conclude as false negatives.

Keeping the above factors, their main drawbacks, the authors proposed a unique architecture that considers all necessary features and discards unnecessary features during the training phase, which is explained in detail in the coming sections.

3. Technical Approach

In this section, the authors define a technical approach for gait formulation for a model that learns discriminative information from gait silhouettes. The proposed technical framework is shown in Figure 2.

3.1. Problem Formulation. The proposed method mainly focuses on part-wise image feature extraction to understand the gait silhouette [27, 28] at a microlevel. The training dataset from CASIA-B, which contains information on "N" people with unique identities $y_i, i \in \{1, 2, 3, \ldots, N\}$, whereby the assumptions of the gait silhouettes are subjected to a probability distribution P_i that is proportional to its identity. All silhouettes of persons in one or more temporal sequences can be regarded as a set of n silhouettes $X_i = \{x_i^j | j = 1, 2, 3, \ldots, N\}$, where $x_i^j \sim P_i$. Hence, the gait recognition of a person can be started by mathematical modeling as follows:

$$f_i = H(G(F(X_i))), \tag{1}$$

where F is the set of convolutional layers aimed at extracting features at the frame level from each unique-identity gait silhouette. The function G represents a permutation of invariant function which maps frame-level features to a set-level feature extracted from each subject target. A set pooling operation is implemented for these layers. The function H represents the discriminative learning of the probability function P_i, which indicates set-level features. Input X_i is a tensor with four dimensions: set dimensions, image channel dimensions, image height dimensions, and image width dimensions.

3.2. Set Pooling. Set pooling [27] is specifically used to extract the gait features of all "N" identity sets. Mathematically, it can be formulated as $N = G(V)$, where N denotes set-level features and V denotes frame-level features, given by $V = \{v^j | j = 1, 2, 3, \ldots, n\}$. To formulate a much deeper function G, the permutation invariant function G is defined as

$$G\left(V = \left\{v^j \mid j = 1, 2, 3, \ldots, n\right\}\right) = G\left(V = \left\{v^{\pi(j)} \mid j = 1, 2, 3, \ldots, n\right\}\right), \tag{2}$$

where π is the permutation element [25]. Because this method is deployed in a real-time environment, function G takes each set of a person's gait silhouette with arbitrary cardinality. To work with invariant constraints present in function G in equation (2), statistical functions are applied to the set dimensions. Three powerful statistical functions are used for computation: max, mean, and median. The joint functions to combine these three functions are as follows:

$$G = \max + \text{mean} + \text{median}, \tag{3}$$

$$G = 1 \times 1\_\text{Conv}(\text{cat}(\max, \text{mean}, \text{median})), \tag{4}$$

where "cat" represents the concatenation of channel dimensions and 1×1_Conv represents a 1×1 convolutional layer. These three statistical functions, max, mean, and median, are applied to the set dimension. Equations (3) and (4) represent learning a proper weight to combine the information extracted by these three statistical functions.

3.3. Attention Mechanism. Visual attention networks are applied to extract spatiotemporal features at the frame level, which were proven to improve the set pooling performance in [4, 5, 7].

Local information often misses some crucial points in extracting temporal features from gait patterns. Hence, the authors present an element-wise attention map that extracts global information from the gait silhouettes. As shown in Figure 3, global information is first collected by statistical functions. These values are then fed to a 1×1 convolutional filter with a feature map to calculate the attention feature maps. However, the final set-level features [27] from each frame are extracted from the Max function and used to refine the frame-level features [29]. This residual structure can stabilize the convergence of the loss function of the network. In this work, the authors are considering three powerful statistical functions, that is, mean, median, and max, because they are easy to compute, and results driven through these parameters are easy to analyze the gait parameters. Since the parameters are in the form of a matrix from each parameter, these three statistical functions work well to compute and maintain a distance between each subject.

3.4. Pyramid Mapping. The extracted features are split into feature strips or feature vectors, which are then used for person reidentification, as proposed in [8, 10]. As in many networks, here too, the images are cropped and resized to a uniform size. The discriminative parts vary from one size to

Figure 1: Sample of the CASIA-B gait dataset used for training.

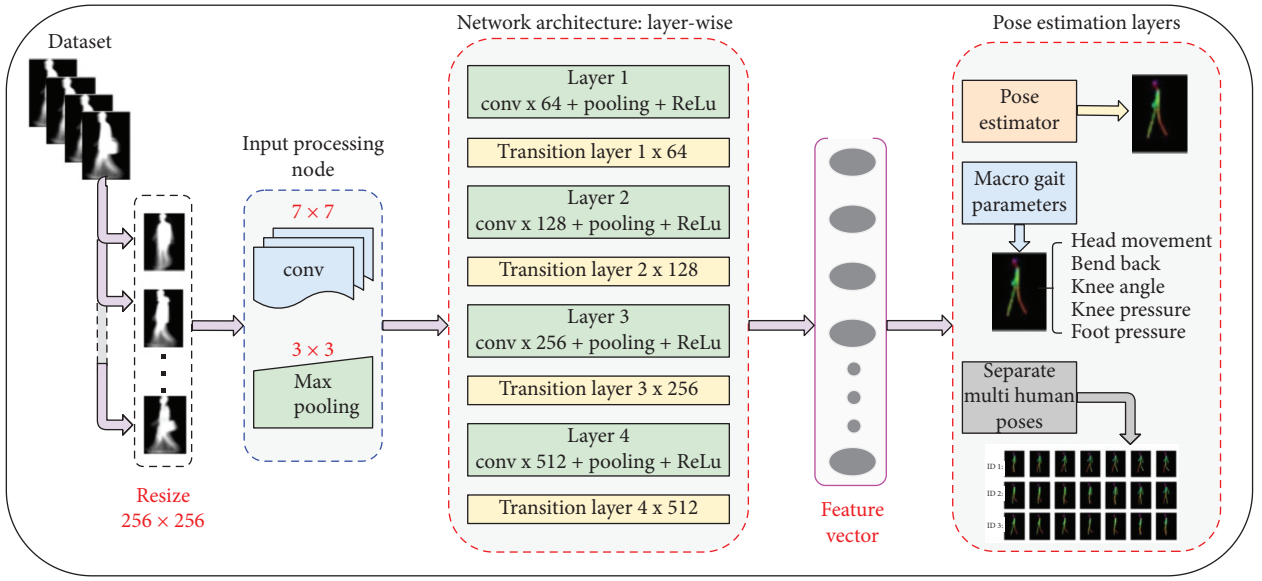


Figure 2: The framework of the proposed network using the CASIA-B dataset as training images and a few custom-collected datasets for training.

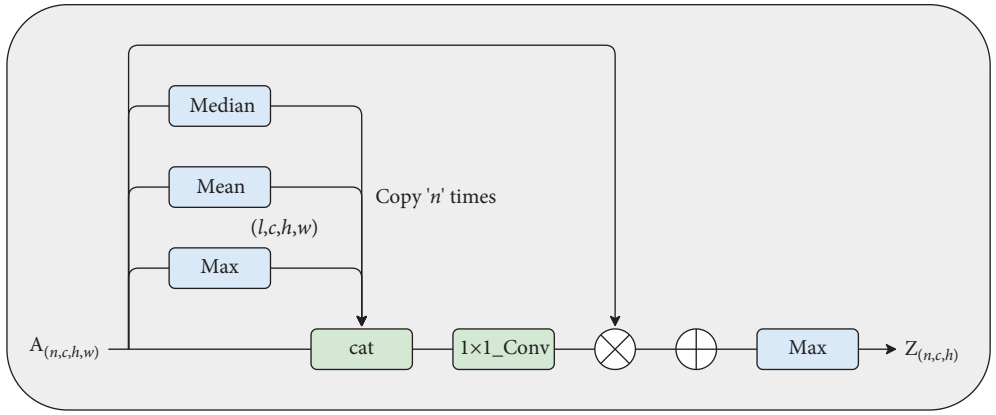


Figure 3: Attention mechanism in the set pooling pipeline. Matrix multiplication and addition are point-wise.

another based on the camera angle fixed at the top. Fu et al. [10] proposed a horizontal pyramid mapping network for local and global feature extractions. The pyramid network consists of 4 scales that help the network focus on gathering local and global features. Extending this technique in this work, a fully connected layer for each pooled feature is used to map discriminative information. Suppose the pyramid network has S scales. The scales can be defined as $s \in 1, 2, 3, \ldots, S$, where the feature maps are extracted by the set pooling function, which is then further split into 2^{s-1} strips based on image height dimensions. Thus, the total strips can be calculated as $\sum_{s=1}^{S} 2^{s-1}$. Global pooling is applied to the 3D strips to extract crucial information formulated in the 1D feature vector. For an individual strip $z_{s,t}$, where $t \in 1, 2, 3, \ldots, 2^{s-1}$. Here, t stands for the index of each strip. The final formulated global pooling is given by $f'_{s,t} = \text{maxpool}(z_{s,t}) + \text{avgpool}(z_{s,t})$, which represents global max pooling and global average pooling, respectively. The

final step includes the fully connected layers for mapping features f' into a feature vector in the discriminative space. The strip vector represents differently depicted features from different receptive fields in different spatial positions. These feature vectors are further given to the fully connected layers. The convolutional layers are deprecated with different receptive fields [30]. The deeper the extraction is, the larger the receptive field will be, and as the identity has to be identified with a deeper parameter, the deeper the layers are.

The pixels representing the features (feature vectors) [31] in the shallow layers in the convolutional network focus mainly on local fine-grained information rather than global feature information to avoid inappropriate features being used for training. However, global features are also important for isolating a person from the frame; hence, the deeper layers focus on global and coarse-grained information. The authors implemented a similar approach to that in [10, 27], a multilayer global pipeline to collect set-level information from the convolutional layers. These set-level features extracted from different layers are sent to the global pipeline. The mathematical modeling of the final feature map from the global pipeline is defined as $\sum_{s=1}^{S} 2^{s-1}$. This is mapped by the global pipeline onto pyramid mapping for a final feature set (vector) [10].

3.5. Pose Plotting. The feature vectors are used further to plot the pose plots and grab macro-gait features [32–34], such as pressure, knee angle, knee pressure, step size, and step angle. This information is fed to the network comprising of pose plotting single and multiple persons from a reference taken from the authors' previous work in [35, 36]. The pipeline for the pose plotting flows as follows.

3.5.1. Single-Pose Plotting. As [35] proved that a pure single-person pose estimator (SPPE) is unreliable due to localization errors, a hybrid network symmetric spatial transformer network (STN) and a single-person pose estimator (SPPE) are introduced for perfect human pose estimation. Furthermore, focusing on local features from the feature vectors given by the shallow layers, the spatial symmetry network (STN) and spatial detransformer network (SDTN) are then pinned to remap the original image to generate grids based on Gamma. The spatial affine transformation used for this pose prediction from the feature vector is given by

$$\begin{bmatrix} x_i^s \\ \overline{y_i^s} \end{bmatrix} = [\theta_1 \theta_2 \theta_3] \begin{bmatrix} x_i^t \\ \overline{y_i^t} \\ 1 \end{bmatrix}, \tag{5}$$

$$\begin{bmatrix} x_i^t \\ \overline{y_i^t} \end{bmatrix} = [\gamma_1 \gamma_2 \gamma_3] \begin{bmatrix} x_i^s \\ \overline{y_i^s} \\ 1 \end{bmatrix}. \tag{6}$$

The symmetric spatial transformer network receives the feature vector from the global pipeline, and the spatial detransformer network generates the pose proposals. The affine transformation used in the spatial symmetric network, as mentioned in equations (5) and (6), where θ_1, θ_2, and θ_3 are vectors and $\{x_i^t, y_i^t\}$ and $\{x_i^s, y_i^s\}$ are the coordinates of the transformation. As mentioned in [15], the spatial detransformer network is the inverse operation of the spatial transformer network. The operations γ are given by

$$\begin{aligned} [\gamma_1 \gamma_2] &= [\theta_1 \theta_2]^{-1}, \\ \gamma_3 &= -1 \times [\gamma_1 \gamma_2] \theta_3. \end{aligned} \tag{7}$$

3.5.2. Fine-Tuning Pose Plotting. For a better extraction of human-dominant regions after the final feature vector recommendation, a parallel single-person-pose estimator (SPPE) is added to the pose network of training [15], which shares its branch with the spatial transformer network. All layers of the SPPE are frozen during the training phase and the weights of this branch are fixed, which is further used as a backpropagation mechanism to modulate center-located pose errors to the spatial transformer network. If the extracted pose of the spatial transformer network is not center-located, pose plotting errors would occur in large quantities, resulting in the backpropagation of large errors in the parallel branch. Hence, the spatial transformer network focuses on the area that is of high quality and dominant, as estimated from the global pipeline. To maintain a higher effectiveness, a parallel pose estimator is turned off in the testing phase to avoid multiple overlaps.

3.5.3. Pose Distance Plots. The distance function for the pose plots is modulated by $d\text{pose}(P_i, P_j)$. The box for pose P_i is B_i. The soft function for the pose plots is then defined as

$$K_{\text{sim}}(P_i, P_j \mid \sigma_1) = \sum_n \tanh \frac{c_i^n}{\sigma_1} \cdot \tanh \frac{c_j^n}{\sigma_1} \tag{8}$$

If k_j^n is in the range of $B(k_i^n)$; otherwise, it is 0. $B(k_i^n)$ is the center of the box around the pose plot at k_i^n, within each dimension of $B(k_i^n)$, which is 1/10th of the original box B_i. However, there may be some instances with poses having low confidence scores, which are probably not the correct poses, but the inaccurate plots of mismatches, whereby the tanh operation eliminates poses with low confidence scores. The joint confidence score would be close to 1. The distance plots indicate the number of joints in a complete pose. The spatial distance between each part is plotted as

$$H_{\text{sim}}(P_i, P_j \mid \sigma_2) = \sum_n \exp\left[-\frac{\left(k_i^n - k_j^n\right)^2}{\sigma_2}\right]. \tag{9}$$

Combining all equations mentioned above for a complete final distance plotting between human poses, they are articulated as

$$d(P_i, P_j \mid \Lambda) = K_{\text{sim}}(P_i, P_j \mid \sigma_1) + \lambda H_{\text{sim}}(P_i, P_j \mid \sigma_2), \tag{10}$$

where λ is the weight value between two distances and $\Lambda = \{\sigma_1, \sigma_2, \lambda\}$.

In the case of a multihuman pose, the spatial detransformer network is remapped with the estimated human pose back to its original image coordinate system. To avoid redundant pose deduction, a nonmaximum suppression network is used.

3.6. Training Details. The CASIA-B dataset was used for training the proposed model. The dataset comprises 124 subjects with 3 different walking conditions and 11 angle views varying in the range of $0-180^0$. Each subject contained 6 sequences for the normal condition, 2 sequences for the walking condition, and 2 sequences for wearing a coat/jacket. Summing up all conditions, there were 110 sequences for each subject. As there is no official split into training and test datasets, the authors split the datasets in an 80–20 ratio.

4. Experiments and Results

4.1. Network Construction. The proposed algorithm is built with two main blocks. The first block with deep convolutional layers [37] was used to train the CASIA-B dataset to extract the feature vectors. It is built on top of the ResNet18 module. This acts as a human detector for the frames. The images are resized to 256×256 in the first block for a part-image-based stream. The resized images are then fed to a 7×7 Conv matrix with 64 filters with a stride size of 2. Then, max pooling with a matrix size of 3×3 and a stride of 2 is computed. This is entered in the first hidden layer, with two layers of 3×3 conv size and 64 filters that recur twice and are then passed on to the transition layer of 3×3 size with 64 filters. Downsampling is followed in the first layer and ReLu nonlinearity. The second layer is the same as the first layer, but the filter size is 128, followed by the ReLu activation function and then, following the transition layer with a 3×3 filter size with 128 filters. The third layer is the same as described above, but with a filter size of 256 with the ReLu function, followed by a transition layer of 3×3 size with 256 filters, the ReLu function, and the final layer with 512 filters followed by the transition layer with 3×3 size and 512 filters. This attention network is provided to the statistical layers, called set pooling layers, rather than a fully connected layer with several classes, and finally a softmax activation function.

In the attention image-based feature stream, the network configuration, as shown in Table 2, there are additional transition layers unlike those present in an ResNet18 module. These transition layers are the feature maps $\{F_1, F_2, F_3, F_4\}$, which consist of the extracted features, that is, the feature vector, which is further fed into the pose estimator block for extracting statistical pose features. F_4 is crucial in action-based feature extraction. As shown in Table 2, the network contains 4 layers, each layer consists of two basic conv blocks. Pictorial representation of the technical architecture of classification model is shown in Figure 4. The further block that is the pose estimator, which is taken from as a base reference

Table 2: Network configuration of first block.

Layer name	Attention network (ResNet18 module)
Conv	$[7 \times 7,\ 64]$ stride = 2
Max pooling	$[3 \times 3]$ stride = 2
Layer 1	$\begin{bmatrix} 3 \times 3 & 64 \\ 3 \times 3 & 64 \end{bmatrix} \times 2$
Transition	$[3 \times 3,\ 64]$
Layer 2	$\begin{bmatrix} 3 \times 3 & 128 \\ 3 \times 3 & 128 \end{bmatrix} \times 2$
Transition	$[3 \times 3,\ 128]$
Layer 3	$\begin{bmatrix} 3 \times 3 & 256 \\ 3 \times 3 & 256 \end{bmatrix} \times 2$
Transition	$[3 \times 3,\ 256]$
Layer 4	$\begin{bmatrix} 3 \times 3 & 512 \\ 3 \times 3 & 512 \end{bmatrix} \times 2$
Transition	$[3 \times 3,\ 512]$
Set pooling	Mean + median + max pooling
Fully connected	#Classes-n Softmax

form the previous work in [35], has 6 convolutional layers with a pretrained pose model from VGG [38]. The pose blocks have each body part to identify and plot pose distance between each of them with respect to the confidence score. The pose block plots all the joints, a 2D coordinate system in each frame and plot all pose points and joining the lines between the joints [39]. The pose coordinates will now give the space to extract macro crucial features for maintaining a unique patterns values for each identity.

4.2. Performance Metrics. The proposed network was trained with a batch size of 16, and training was performed with the Adam optimizer [40] with an initial learning rate of 0.0001. The decaying rate factor was maintained at 0.1 for the first 200 epochs. The training was active until 36 h and 1000 epochs on a Nvidia 1080 TI GPU.

The model evaluation metrics, as shown in Figure 5, are used to validate that the performances are of average precision and F1 scores. The pose estimator compares the predicted joint coordinates and pose distances inside it (human region) according to the intersection over union (IOU), where its parameters are updated at each epoch. The F1 score is used to calculate the success rate of precision and recall. Precision is the ratio of actual matches, and recall is defined as the ratio of correct predictions to that of total ground truths. However, neither is sufficient to measure the performance of the network [9, 41]. The F1 score is calculated with the precision and recall as dependent parameters to compute the evaluation of the network on data. F1 score is termed as true positive (TP) for correct predictions, false negative (FN) for false nondetections and false positive (FP) for false correct predictions [9, 41]. The mathematical computations for the abovementioned parameters are as follows:

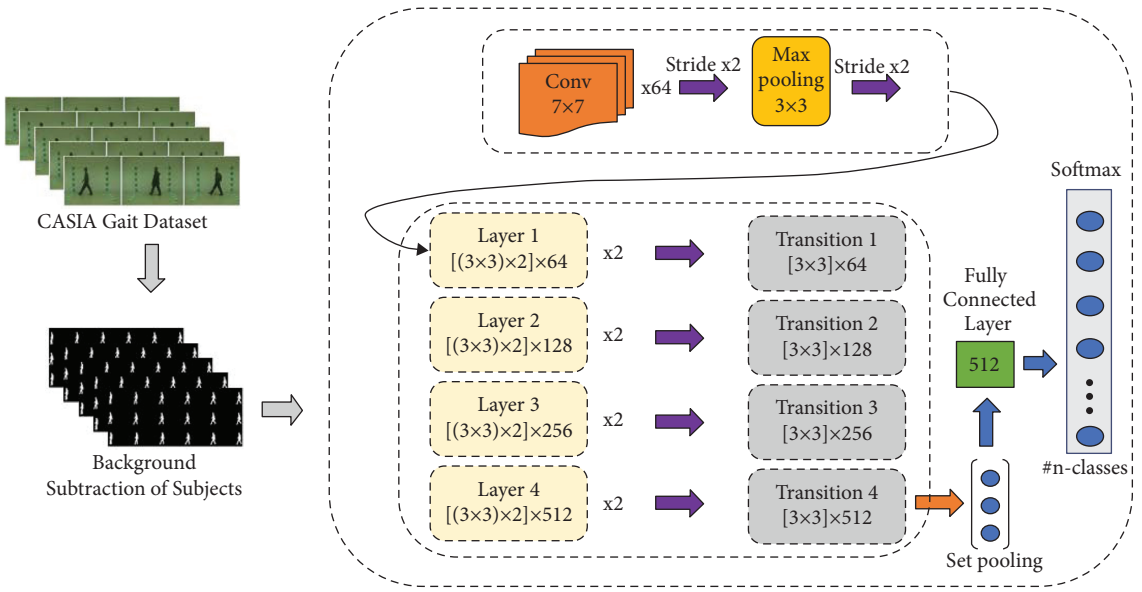


Figure 4: Pictorial representation of the proposed network from architecture Table 1 for image classification.

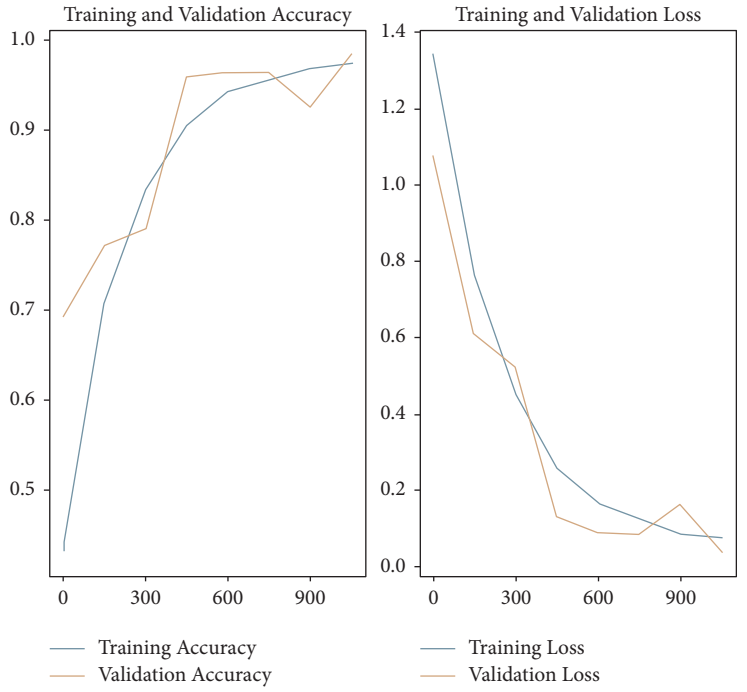


Figure 5: Performance graph of training on CASIA-B dataset.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}},$$
$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \tag{11}$$
$$F1\ \text{Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}.$$

As a heavyweight model with a combination of 2 pretrained models, VGG and ResNet18, the inference on GPU was ~ 1.5 ms, and on CPU, it was ~ 5 s in real time. Table 3 describes the results of some crucial gait patterns contributing to gait identity. The qualitative and quantitative results are plotted in the real-time video feed captured as a testing phase on the trained model of the proposed algorithm. Figure 6 visualizes the gait pattern motions graphically, which are unique to each identity subject and focus mostly on knee elements [32, 42], such as angle, size, and pressure. During the training, the model extracts the motion recordings from the sequences of each subject under their respective subject ID. These are fed as values to the feature vector. Figure 7 comprises the results of a test subject,

Table 3: F1 score results on crucial gait parameters at the macrolevel validated in real time.

Crucial gait parameters	Precision	Recall	*F*1 score
Foot pressure	0.99	0.98	0.98
Knee angle	0.96	0.93	0.94
Step size	0.94	0.89	0.91
Knee pressure	0.97	0.92	0.95
Foot motion	0.81	0.75	0.77
Joints motion	0.80	0.65	0.71

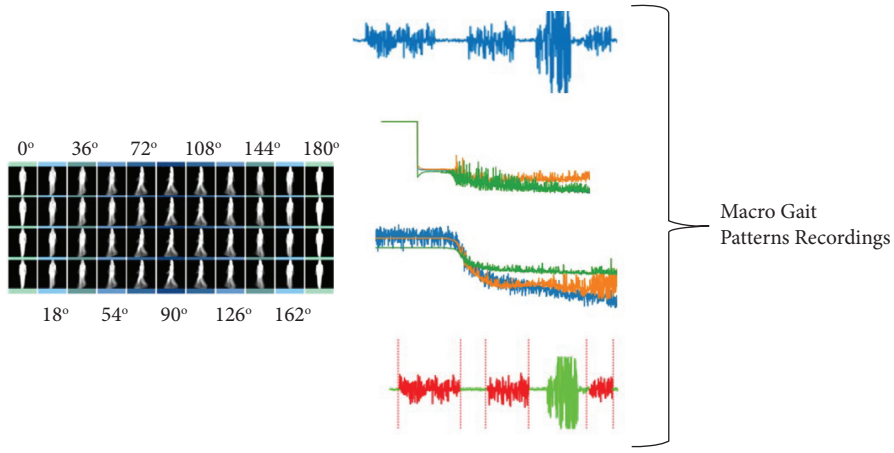


Figure 6: Extraction of macro-gait patterns, such as foot pressure, knee motion, knee angle, and step size, from CASIA-B dataset.

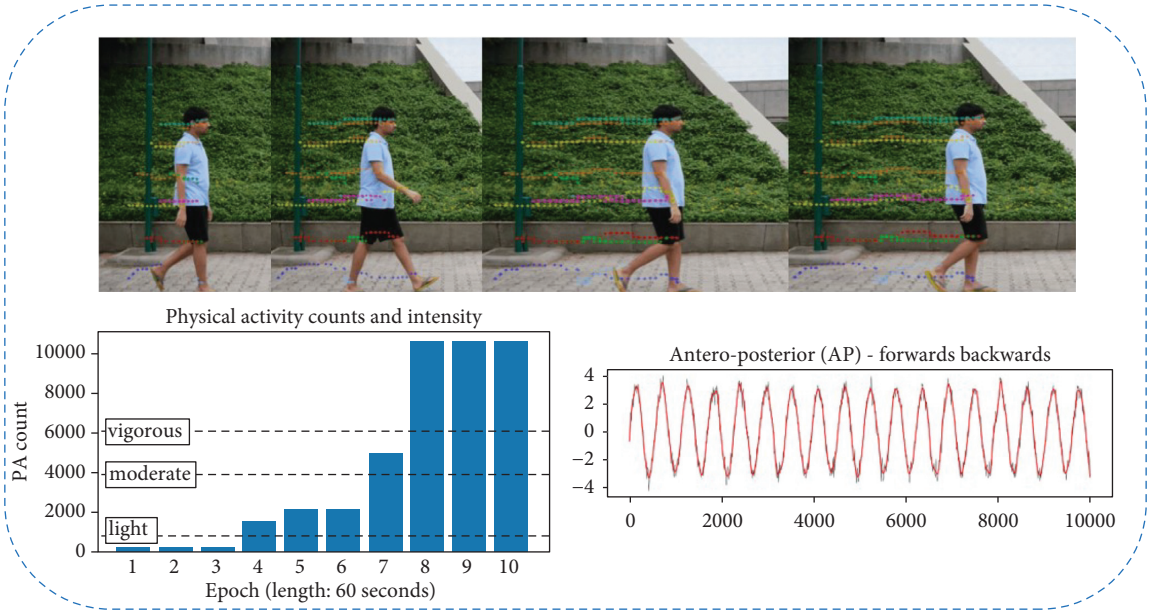
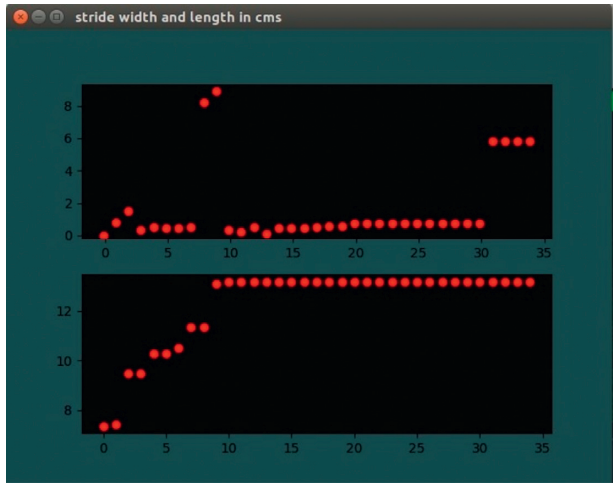


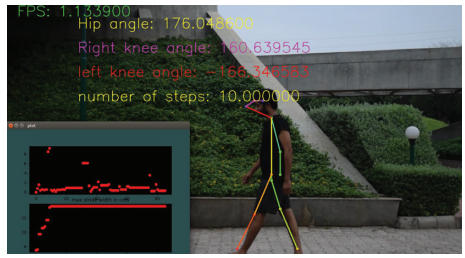
Figure 7: Testing on real-time feed with a plotting motion graph on gait patterns of knee motion.

representing the motion graphically throughout the feed. The variation pattern, which is a microlevel wave, is unique for each frame for all subjects. Figure 8, in contrast, is another test subject for which the camera is closer to the subject, with the pose plotted and the results of the gait motion displayed. Figure 8(b) specifically blacks out the frame and visualizes the graphical motion pattern. These gait patterns are used to save the weights for each unique subject. These weights are used to classify each subject according to their identity. In this way, the proposed algorithm pins the gait biometrics based on gait motion patterns.

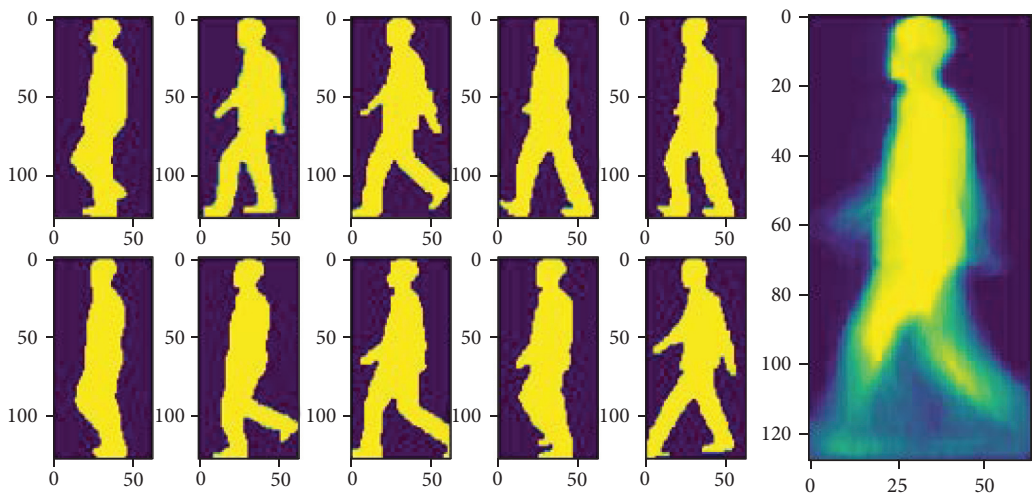
The environment set up for the real-time testing is taken with a normal Logitech camera with 1080p resolution; subjects in Figures 7 and 8 are taken from Logitech 1080p camera. The camera is placed 10m away from the subject, in normal natural sunlight. The view angle taken for real-time testing is side angles in order to capture leg movements throughout the motion. The frontal view would sometimes



(a)



(b)



(c)

Figure 8: Continued.

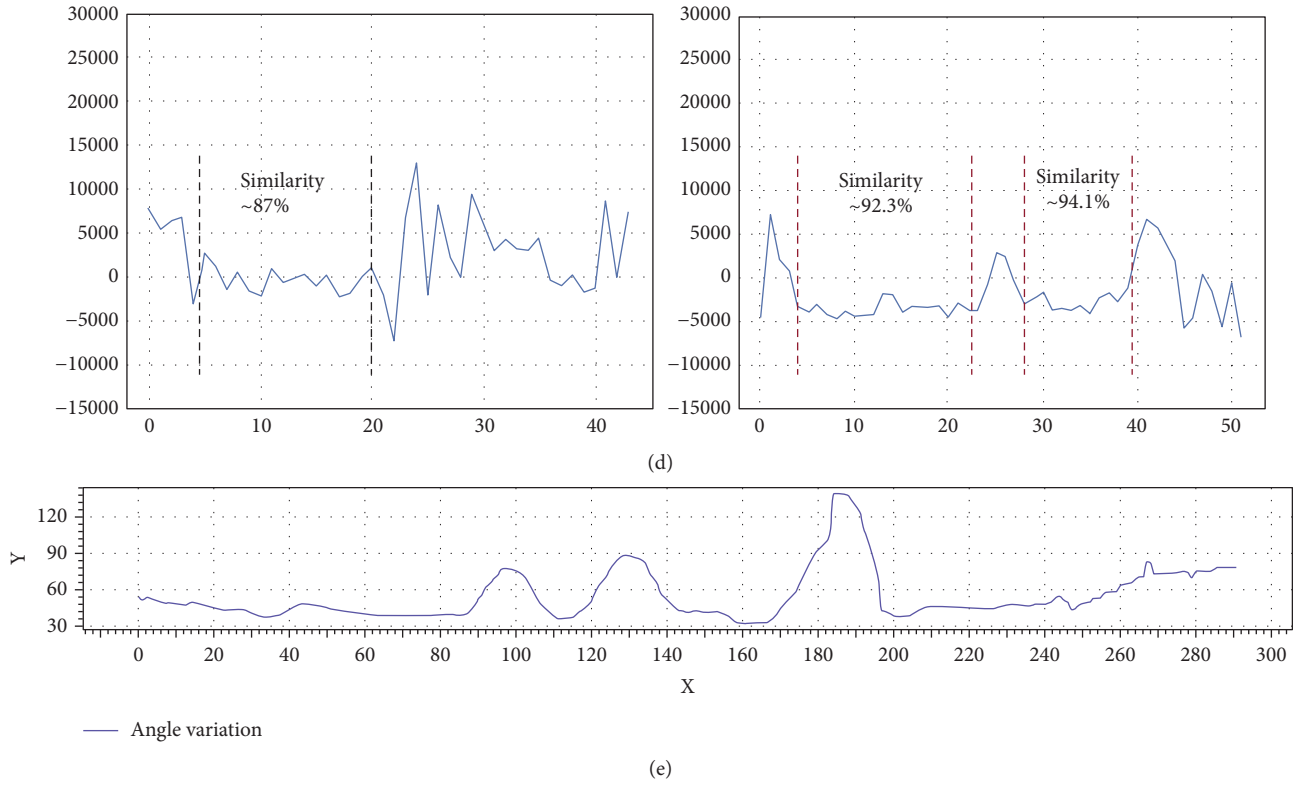
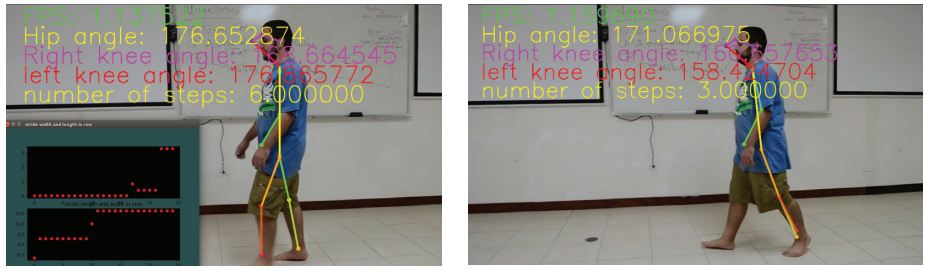


Figure 8: Testing on real-time dataset. (a) Testing on real-time dataset to visualize crucial gait patterns of angles. (b) Graphical plot of (a) of knee angles and their motions. (c) Gait silhouettes of (a), isolating the background to extract gait parameters. (d) Trained parameters (A) and testing parameters (B). Gait parameters tested were body posture (backbone motion). (e) Knee angle variation of subject presented in (a).

dilute the concentration of the leg movements with limited numbers since the leg backward movements would be missing in the frontal view. Hence, side angle views are considered for real-time testing to capture all crucial points from the legs. Another real-time testing with camera placed nearer to the subject and different lighting conditions is shown in Figure 9. Figure 10 represents a comparison between 2 test subjects. Figure 10(a) represents the same subjects with different scenarios, and the subjects in Figure 10(b) are different test subjects.

4.3. Comparison with the State-of-the-Art Methods. A comparison table has been drawn as shown in Table 4 against the proposed method GaitVision. The comparison is made with respect to the real-time evaluation. Appearance based method includes method done in GEI-SVR [44], where gait entropy image (GEI), extracted silhouettes, and energy image are defined by their silhouette masks. This method draws a major drawback from sizeable intrasubject appearance changes due to covariates such as clothing, carrying, view angles, and walking speed. The methods described in [43, 47] extract gait features from RGB images via conditional random field. Zhang et al. [47] is a CNN-based approach with discriminating representation from data with multicovariates. The main drawback of these two methods is their low performance in real time. Wu et al. [46] is a low computational cost method that can handle low-resolution images. However, it is sensitive to clothes change, view angles, and walking speed, which makes it inappropriate to real-time deployment. Hu et al. [11] proposed view invariant human gait identification, but in terms of view angles, GaitNet [47] surpasses [11]. Kusakunniran's [45] method describes spatiotemporal information extraction of features. This method extracts the crucial spatiotemporal information while subject is in motion, but many false and unnecessary information can be captured from the data. Kusakunniran et al. [48] proposed gait subjects recognition through various angles through correlation motion. In real time, however, view angles and motion speed make this method unreliable.

The proposed method GaitVision clearly surpasses all the state-of-the-art methods in real time working with different view angles, difference backgrounds, and in decent motion. Most of the current methods lack their presence in real time since they are constrained to specific environments. Even though methods have trained their model with a huge number of classes, they have imitations to fail to run in real-time. The authors here proposed a unique way to collect the subject motions for at least 5 minutes, train them, and deploy in real time. The subjects collected in real time are trained and tested with various view angles and clothes with a decent motion with camera 10 meters away from the subject. Figure 8(c) represents the isolation of background to focus on the subject for a clean gait patterns extraction. Figure 8(d) represents the similarity of body posture of the test case of a subject.



(a) (b)

Figure 9: Real-time results with different environment setups (a, b), lighting conditions, angles (side), camera distance: ~2.5 m from the subject.

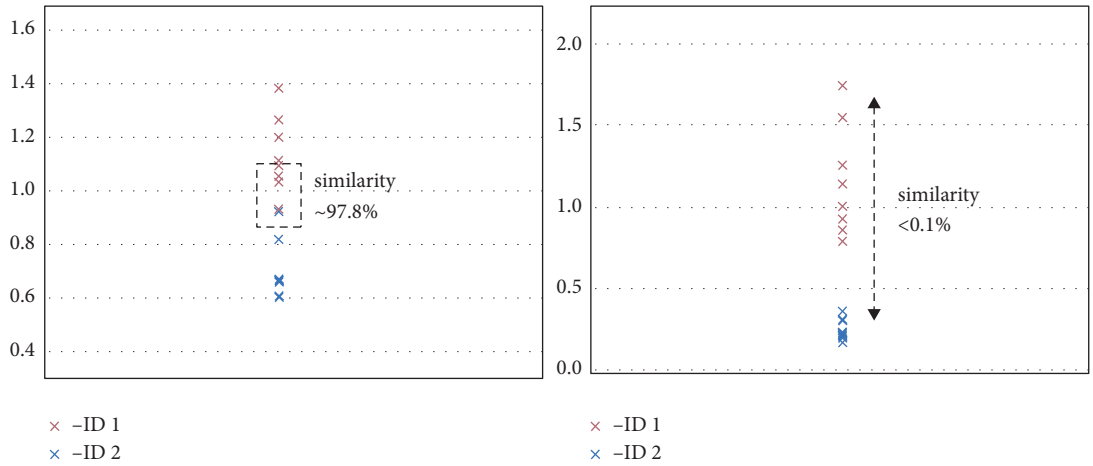


Figure 10: Red *X* represents trained subject; blue *X* represents test subject. (a) Image represents two similar subjects and (b) two different subjects. Gait parameter here was footstep speed.

Table 4: Comparison of GaitVision (ours) with the state-of-the-art method.

Method	Average accuracy
CPM [43]	24.1
GEI-SVR [44]	42.0
CMCC [45]	43.9
ViDP [11]	45.4
LB [46]	56.9
L-CRF [43]	67.8
GaitNet [47]	81.8
STIP + NN [48]	84.0
GaitVision (proposed)	90.3

Figure 8(a) depicts the knee motion graph (similarity with respect to training set of the subject) of the same test case (as shown in Figure 8). The essential advantage of the proposed GaitVision algorithm is its method of training the subjects with different backgrounds and angles. Unlike the method proposed in [46], GaitVision algorithm can detect the trained subject with different clothes and backgrounds. This is possible because of training of crucial features and parameters extracted from the subject's motion. Another essential component in the proposed algorithm is the real-time deployment, unlike most of the gait methods work on prerecorded videos. Hence, using these steps in extracting crucial gait features, GaitVision surpasses the rest of the state-of-the-art methods in real-time deployment.

5. Conclusion

The proposed algorithm is the backbone for classifying each subject based on its unique natural patterns. Gait motion is a unique natural pattern that can be used to train for each distinctive subject in any authentication system. With training using deep convolutional layers and the results then being fed to feature maps to extract individual gait motion patterns, the accuracy of detection of any trained subject with at least 120 s of their motion provided at diverse angles is high. The environment is not hugely influential, as the subjects are initially converted to grayscale and feature maps are extracted from the deep layers; subsequently, this vector suggests the local regions of human presence to the pose estimator, which then plots the pose and extracts the gait motion patterns. Mainly, the core novelty of this work lies in the pose estimation followed by the training of the pose patterns in frame-wise for each subject. This thereby avoids the extraction of unwanted

features from the pose. The second major part is the extraction of gait parameters from the pose patterns. These parameters are trained frame-wise for each subject. Combining all gives a better result than compared to the state-of-the-art methods.

6. Limitations and Future Scope

The core aim of this work is to conduct an extensive research on contactless gait biometrics using a simple camera. Hence, edge devices would be the optimal case to deploy it in the production. However, this work uses VGG pretrained network and ResNet152 module, which are heavy weight and some of the edge devices [33] like Raspberry Pi is not suitable for the production deployment. The edge device should be computationally capable of having more than 4 GB of RAM. Another drawback is the constrained environments. In camera angle-wise, top-view angle is a major drawback and the top-view angles are not suitable because the head and body pose would dominate the bottom pose gait parameters. Since bottom pose gait parameters play a crucial role in this work, top-view angles fail to capture the correct gait parameters. Hence, feature extraction would be difficult for top-view angles and leads to false positives. Another problem in feature extractions is the camera distance from the subject. As the camera focus is kept beyond 10 meters from the subject, the motion parameters get merged and lead to false parameter training and result in major inaccuracies.

Algorithm-wise, the major drawback is the FPS. As shown in the Results section (Figures 8 and 9), the maximum FPS is ~2 FPS on CPU, and on GPU, the maximum FPS is 20 FPS, provided with at least 4 GB RAM, and with 12 GB RAM, the maximum FPS is 55 FPS. The results provided in this work have taken CPU to show the complexity of the algorithm.

The core vision would be to deploy the algorithm in an edge device and deploy it as a standalone; hence, this research can be used in further enhancements of cutting off the computational loads and preparing a light-weight model for an easy edge deployment in the production level. The model can be further processed by collecting larger-scale datasets on larger subjects with at least 5 minutes of their motion in all possible angles and movements and train the model for a large-scale deployment.

Data Availability

The data were collected from cite investigation and can be provided upon request from the corresponding author.

Ethical Approval

The manuscript is conducted within the ethical manner advised by the Complexity journal.

Conflicts of Interest

The authors declare no conflicts of interest to any party.

Acknowledgments

The authors acknowledge the support by the National Institute of Technology, Warangal, India.

References

[1] M. F. Hashmi, B. K. K. Ashish, V. Sharma et al., "Larnet: real-time detection of facial micro expression using lossless attention residual network," *Sensors*, vol. 21, no. 4, p. 1098, 2021.

[2] L. Sudha and D. R. Bhavani, "Biometric authorization system using gait biometry," 2011, http://arxiv.org/abs/1108.6294.

[3] S. Muzaffar and I. M. Elfadel, "Self-synchronized, continuous body weight monitoring using flexible force sensors and ground reaction force signal processing," *IEEE Sensors Journal*, vol. 20, 2020.

[4] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7794–7803, Salt Lake City, UT, USA, June 2018.

[5] K. Xu, J. Ba, R. Kiros et al., "Show, attend and tell: neural image caption generation with visual attention," in *Proceedings of the 32nd International Conference on Machine Learning*, pp. 2048–2057, Lille, France, July 2015.

[6] A. Hermans, L. Beyer, and B. Leibe, "In defense of the triplet loss for person re-identification," 2017, http://arxiv.org/abs/1703.07737.

[7] W. Li, X. Zhu, and S. Gong, "Harmonious attention network for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2285–2294, Salt Lake City, UT, USA, June 2018.

[8] G. Wang, Y. Yuan, X. Chen, J. Li, and X. Zhou, "Learning discriminative features with multiple granularities for person re-identification," in *Proceedings of the 26th ACM International Conference on Multimedia*, pp. 274–282, Seoul, Korea, October 2018.

[9] M. F. Hashmi, B. K. K. Ashish, A. G. Keskar, N. D. Bokde, and Z. W. Geem, "Fashionfit: analysis of mapping 3d pose and neural body fit for custom virtual try-on," *IEEE Access*, vol. 8, pp. 91603–91615, 2020.

[10] Y. Fu, Y. Wei, Y. Zhou et al., "Horizontal pyramid matching for person re-identification," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, pp. 8295–8302, 2019.

[11] M. Hu, Y. Wang, Z. Zhang, J. J. Little, and D. Huang, "View-invariant discriminative projection for multi-view gait-based human identification," *IEEE Transactions on Information Forensics and Security*, vol. 8, no. 12, pp. 2034–2045, 2013.

[12] Q. Chen, Y. Wang, Z. Liu, Q. Liu, and D. Huang, "Feature map pooling for cross-view gait recognition based on silhouette sequence images," in *Proceedings of the 2017 IEEE International Joint Conference on Biometrics (IJCB)*, pp. 54–61, IEEE, Denver, CO, USA, October 2017.

[13] E. R. Isaac, S. Elias, S. Rajagopalan, and K. Easwarakumar, "Trait of gait: a survey on gait biometrics," 2019, http://arxiv.org/abs/1903.10744.

[14] J. Han and B. Bhanu, "Individual recognition using gait energy image," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 2, pp. 316–322, 2005.

[15] P. Esser, E. Sutter, and B. Ommer, "A variational u-net for conditional appearance and shape generation," in *Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8857–8866, Salt Lake City, UT, USA, June 2018.

[16] L. Tran, X. Yin, and X. Liu, "Disentangled representation learning gan for pose-invariant face recognition," in *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1415–1424, Honolulu, HI, USA, July 2017.

[17] L. Tran, X. Yin, and X. Liu, "Representation learning by rotating your faces," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 12, pp. 3007–3021, 2018.

[18] I. Goodfellow, J. Pouget-Abadie, M. Mirza et al., "Generative adversarial nets," in *Proceedings of the 27th International Conference on Neural Information Processing Systems*, pp. 2672–2680, Montreal, Canada, December 2014.

[19] Y. He, J. Zhang, H. Shan, and L. Wang, "Multi-task gans for view-specific feature learning in gait recognition," *IEEE Transactions on Information Forensics and Security*, vol. 14, no. 1, pp. 102–113, 2018.

[20] B. Bhandari, G. Lee, and J. Cho, "Body-part-aware and multitask-aware single-image-based action recognition," *Applied Sciences*, vol. 10, no. 4, p. 1531, 2020.

[21] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *Proceedings of the 2019 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5693–5703, Long Beach, CA, USA, June 2019.

[22] H. Fukui, T. Hirakawa, T. Yamashita, and H. Fujiyoshi, "Attention branch network: learning of attention mechanism for visual explanation," in *Proceedings of the 2019 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 10705–10714, Long Beach, CA, USA, June 2019.

[23] J. Krause, J. Johnson, R. Krishna, and L. Fei-Fei, "A hierarchical approach for generating descriptive image paragraphs," in *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 317–325, Honolulu, HI, USA, July 2017.

[24] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, Las Vegas, NV, USA, June 2016.

[25] M. Zaheer, S. Kottur, S. Ravanbakhsh, B. Poczos, R. R. Salakhutdinov, and A. J. Smola, "Deep sets," in *Proceedings of the 31st Conference on Advances in Neural Information Processing Systems*, pp. 3391–3401, Long Beach, CA, USA, 2017.

[26] Casia Gait Database, http://www.cbsr.ia.ac.cn/english/Gait2020-03-05.

[27] H. Chao, Y. He, J. Zhang, and J. Feng, "GaitSet: regarding gait as a set for cross-view gait recognition," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, pp. 8126–8133, 2019.

[28] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, "Pointnet: deep learning on point sets for 3d classification and segmentation," in *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 652–660, Honolulu, HI, USA, July 2017.

[29] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, "Pointnet++: deep hierarchical feature learning on point sets in a metric space," 2017, http://arxiv.org/abs/1706.02413.

[30] K. Bashir, T. Xiang, and S. Gong, "Gait recognition without subject cooperation," *Pattern Recognition Letters*, vol. 31, no. 13, pp. 2052–2060, 2010.

[31] J. S. Matthis, S. L. Barton, and B. R. Fajen, "The critical phase for visual control of human walking over complex terrain," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 114, no. 32, pp. E6720–E6729, 2017.

[32] X. Wu, Y. Ma, X. Yong, C. Wang, Y. He, and N. Li, "Locomotion mode identification and gait phase estimation for exoskeletons during continuous multi-locomotion tasks," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 13, 2021.

[33] C. Wu, F. Zhang, Y. Hu, and K. R. Liu, "Gaitway: monitoring and recognizing gait speed through the walls," *IEEE Transactions on Mobile Computing*, vol. 20, 2021.

[34] A. Kumar, A. Godiyal, P. Joshi, and D. Joshi, "A new force myography-based approach for continuous estimation of knee joint angle in lower limb amputees and able-bodied subjects," *IEEE Journal of Biomedical and Health Informatics*, vol. 25, 2021.

[35] M. F. Hashmi, B. K. K. Ashish, and A. G. Keskar, "Gait analysis: 3d pose estimation and prediction in defence applications using pattern recognition," in *Proceedings of the Twelfth International Conference on Machine Vision*, Amsterdam, Netherlands, 2019.

[36] M. F. Hashmi, K. K. Ashish, S. Katiyar, and A. G. Keskar, "Accessnet: a three layered visual based access authentication system for restricted zones," in *Proceedings of the 2020 21st International Arab Conference on Information Technology (ACIT)*, pp. 1–7, IEEE, Giza, Egypt, November 2020.

[37] K. Shiraga, Y. Makihara, D. Muramatsu, T. Echigo, and Y. Yagi, "Geinet: view-invariant gait recognition using a convolutional neural network," in *Proceedings of the 2016 International Conference on Biometrics (ICB)*, pp. 1–8, IEEE, Halmstad, Sweden, June 2016.

[38] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, http://arxiv.org/abs/1409.1556.

[39] W. Hamilton, Z. Ying, and J. Leskovec, "Inductive representation learning on large graphs," pp. 1024–1034, 2017, http://arxiv.org/abs/1706.02216.

[40] D. P. Kingma and J. Ba, "Adam: a method for stochastic optimization," 2014, http://arxiv.org/abs/1412.6980.

[41] M. F. Hashmi, B. K. K. Ashish, A. G. Keskar, N. D. Bokde, J. H. Yoon, and Z. W. Geem, "An exploratory analysis on visual counterfeits using conv-lstm hybrid architecture," *IEEE Access*, vol. 8, pp. 101293–101308, 2020.

[42] M. Song and J. Kim, "An ambulatory gait monitoring system with activity classification and gait parameter calculation based on a single foot inertial sensor," *IEEE Transactions on Biomedical Engineering*, vol. 65, no. 4, pp. 885–893, 2017.

[43] X. Chen, J. Weng, W. Lu, and J. Xu, "Multi-gait recognition based on attribute discovery," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 7, pp. 1697–1710, 2017.

[44] W. Kusakunniran, Q. Wu, J. Zhang, and H. Li, "Support vector regression for multi-view gait recognition based on local motion feature selection," in *Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 974–981, IEEE, San Francisco, CA, USA, June 2010.

[45] W. Kusakunniran, "Recognizing gaits on spatio-temporal feature domain," *IEEE Transactions on Information Forensics and Security*, vol. 9, no. 9, pp. 1416–1423, 2014.

[46] 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 Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 2, pp. 209–226, 2016.

[47] Z. Zhang, L. Tran, X. Yin et al., "Gait recognition via disentangled representation learning," in *Proceedings of the 2019*

IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4710–4719, Long Beach, CA, USA, June 2019.

[48] W. Kusakunniran, Q. Wu, J. Zhang, H. Li, and L. Wang, "Recognizing gaits across views through correlated motion co-clustering," *IEEE Transactions on Image Processing*, vol. 23, no. 2, pp. 696–709, 2013.