Multimodal Gait Recognition With Inertial Sensor Data and Video Using Evolutionary Algorithm

Kumar, Pradeep; Mukherjee, Subham; Saini, Rajkumar; Kaushik, Pallavi; Roy, Partha Pratim; Dogra, Debi Prosad

Published in:
IEEE Transactions on Fuzzy Systems

DOI:
10.1109/TFUZZ.2018.2870590

Publication date:
2019

Multimodal Gait Recognition With Inertial Sensor Data and Video Using Evolutionary Algorithm

Pradeep Kumar, Subham Mukherjee, Rajkumar Saini, Pallavi Kaushik, Partha Pratim Roy, and Debi Prosad Dogra

***Abstract*—Evolutionary decision fusion has applications in biometric authentication and verification. Gray wolf optimizer (GWO) is one such evolutionary decision fusion approach that can be used to tune the fusion parameters in a multimodal data acquisition system. Human gait is a proven biometric trait with applications in security and authentication. However, acquiring human-gait data can be erroneous due to various factors and multimodal fusion of such erroneous gait data can be challenging. In this paper, we propose a new decision fusion-based approach to solve the above problem. Gait data is recorded simultaneously using motion sensors and visible-light camera. The signals of the motion sensors are modeled using a long short-term memory neural network and corresponding video recordings are processed using a three-dimensional convolutional neural network. GWO has been used to optimize the parameters during fusion. It has been chosen based on the underlying hunting strategy that leads to better approximation of the solution. Interestingly, in our case it converges quicker than other optimization techniques such as genetic algorithm or particle swarm optimization. To test the model, a dataset involving 23 males and females has been recorded while they perform four different types of walks, including, *normal walk, fast walk, walking while listening to music*, and *walking while watching multimedia content on a mobile*. An overall accuracy of 91.3% has been recorded across all test scenarios. Results reveal that the proposed study can further be explored to design robust gait biometric systems.**

***Index Terms*—Biometric, deep learning, gait analysis, gray wolf optimizer (GWO), Shadow Motion.**

I. Introduction

In RECENT times, gait is being considered as a useful biometric trait because of its unique advantages over other biometric modalities such as noncontact, hard to fake, and obtainable, from a distance [1]. In literature, researchers have shown that different features can be extracted by analyzing the walking patterns of individuals to prove their individuality [2]. Gait-recognition systems are equipped with multiple characteristics as follows.

1) Gait can be recognized with low-resolution images and at a distance from the camera, whereas the other biometric traits, such as face, iris, or fingerprint, require relatively higher image resolution to be applicable to authentication applications.

2) Gait can be recognized without subject cooperation as it can be recorded without a user's consciousness.

(3) Gait is an unconscious behavior, therefore, it is difficult to be spoofed.

There are a number of pioneering works in gait recognition with the help of various acquisition systems such as RGB cameras, floor sensors, wearable and depth sensors. Simple camera based systems involve usage of analog or digital cameras with suitable optics for acquiring the gait data. Such systems have several advantages over other biometric systems because almost all gait features such as stride, cadence, static parameters (e.g., distance between head and pelvis, pelvis, and feet) and the gait cycle can be directly extracted from the video [3]. However, these techniques suffer from imperfect segmentation due to clothing, view variations in cameras with respect to the walking, changes in gait because of carrying objects, mood or change in speed [4]. Therefore, various approaches have been proposed over time to tackle such problems by analyzing human silhouettes by dividing them into segments and assigning weights, background subtraction, and building three-dimensional (3-D) models from shadow [5]. Choudhury *et al.* [6] has proposed a gait-recognition method by analyzing human silhouettes using procrustes shape analysis (PSA) and elliptic Fourier descriptors. The authors have combined spatio–temporal features with statistical and physical parameters with silhouette contours. It has also been analyzed that existing gait-recognition systems assume that cameras are placed at locations such that the complete body shape of the person can be observed. However, the classification rate decreases if cameras are installed at a height from the ground plane, such as on the rooftops of tall buildings or on unmanned aerial vehicle (UAVs) for wide area security operations. This is because the human body area, which is used to extract gait features, may not be fully captured from such locations or angles. It becomes more difficult to acquire gait features when the subject engaged in activities like listening to music, watching videos in mobile phones, and walking fast in front of the acquisition device.

Alternatively, wearable sensors based techniques can be used to acquire data. However, such techniques require appropriate

Manuscript received December 31, 2017; revised May 29, 2018; accepted September 6, 2018. Date of publication September 17, 2018; date of current version May 1, 2019. *(Corresponding author: Pradeep Kumar.)*

P. Kumar, R. Saini, P. Kaushik, and P. P. Roy are with the Department of Computer Science & Engineering, IIT Roorkee, Roorkee 247667, India (e-mail: pradeep.iitr7@gmail.com; rajkumarsaini.rs@gmail.com; pallavikaushik27@gmail.com; proy.fcs@iitr.ac.in).

S. Mukherjee is with the Department of Electronics & Communication Engineering, Institute of Engineering and Management, Kolkata 700 091, India (e-mail: subhammukherjee61196@gmail.com).

D. P. Dogra is with the School of Electrical Sciences, IIT Bhubaneswar, Bhubaneswar 751013, India (e-mail: dpdogra@iitbbs.ac.in).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TFUZZ.2018.2870590

placement of the sensors over the body to measure different walking styles [7]–[9]. Several body locations have been proposed by researchers where sensors can be placed, e.g., hip, legs, arms, or other parts of the body. The wearable sensors can be of different types as per requirements such as accelerometers (measure acceleration), gyro sensors (measure rotation), pressure sensors (measure force when walking), etc. The main advantage of developing gait-recognition system using wearable sensor techniques is its provision of unobtrusive authentication for mobile devices (like mobile phones, personal digital assistant (PDAs), etc.) containing accelerometers or other sensors. Therefore, it can be utilized for continuous verification of user's identity without his/her intervention. Additionally, sensors readings are more accurate and does not need any preprocessing in comparison to the vision-based approaches. Therefore, in this paper, we have utilized a wireless body sensor suit popularly known as "Shadow Motion" along with the conventional video system. The suit consists of 17 inertial sensor nodes, where each node is equipped with accelerometer, gyroscope, and magnetometer.

Deep learning has been extensively used in computer vision for image/video classification, gesture recognition, face recognition, and gait recognition. In these approaches, especially convolutional neural network (CNN) has been widely used to extract features from images or videos that are proven to be better than the hand-crafted features [10]. In addition, long short-term memory (LSTM) neural networks have been used by researchers to model temporal sequences in speech recognition, handwriting recognition, language modeling, etc. Basically, LSTMs are a specific type of recurrent neural network (RNN) architecture that are designed to model long-range dependencies in temporal sequences more accurately [11]. Since we are using more than one modality to acquire gait data, it is necessary to adopt a decision fusion approach that can effectively tune the fusion parameters for each modality [12], [13]. Therefore, we have adopted an evolutionary algorithm to select the fusion parameters [14]. Basically, evolutionary algorithms are highly flexible, easy to follow, and are robust in responding to any change. Moreover, these algorithms generate global solution and can be applied to real-world problems where optimization techniques lead to unsatisfactory results by generating local optima. Therefore, these algorithms are gaining attention in the research community, particularly in real-life applications. Recently, Mirjalili *et al.* [15] has proposed gray wolf optimizer (GWO) algorithm that has been inspired from gray wolves (Canis lupus). The authors have shown the robustness of the algorithm on 29 test functions where the results outperformed with existing metaheuristics. Likewise, Emary *et al.* [16] has used GWO algorithm to propose a fitness function for finding the optimal subset of features for accurate classification. The feature subset with the least number of features and the highest classification accuracy is termed to be the most optimal. It has also been reported by the authors that GWO has rendered better classification accuracy and more reduced feature size than particle swarm optimization (PSO) and genetic algorithm (GA) optimizers by experimenting on different datasets and using three different initialization methods.

Motivated by the recent developments in evolutionary algorithms and sophisticated deep learning architecture, and with the emergence of full-body sensors, we have proposed a new biometric-authentication approach by analyzing human gait motion captured using shadow device and video camera. Most existing systems perform in single modality by extracting complex handcrafted features from the raw inputs. We have used 3-D CNNs, which are able to extract features from spatial as well as temporal dimensions by performing 3-D convolutions; hence, they capture the motion information from multiple adjacent frames. This multichannel information is then combined into a final feature representation, which is fed to the LSTM layers for modeling sequential data. Similarly, data from inertial sensors are processed using a four-stacked LSTM architecture. The classification has been performed using Softmax function and the scores are then fused using GWO-guided evolutionary algorithm.

The main contributions of the paper are as follows.

1) First, we present a multimodal biometric gait-recognition method for different walking conditions by fusing video and 3-D sensor data.
2) Second, deep learning frameworks such as 3-D CNNs and LSTMs have been utilized to train the models with one type of walking patterns and tested on three different gait recordings.
3) Third, an evolutionary algorithm scheme (GWO) has been implemented to tune the fusion parameters of each modality to boost the recognition performance of the system. Finally, a comparison with other optimization schemes has been presented.

The rest of this paper is organized as follows. We present a review on the recent development on gait recognition in Section II. The proposed multimodal architecture is presented in Section III. Results are discussed in Section IV. Finally, we conclude in Section V by highlighting some of the future extensions of the presented work.

II. Related Work

In this section, we discuss different gait-recognition systems that have been proposed so far using vision-based or sensor-based techniques.

A. Vision-Guided Gait-Recognition Approaches

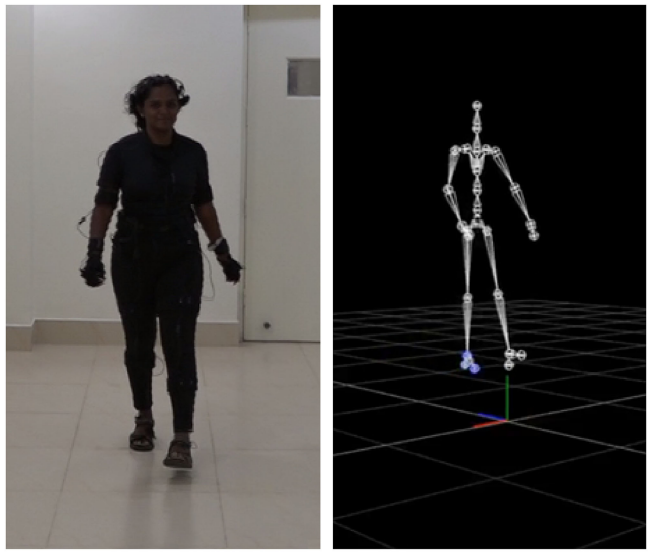
Wang *et al.* [17] has proposed an approach of personal identification from videos based on the gait information. As both appearance and walking play a crucial role in the recognition of an individual through gait, they have combined the static features like body weight and height, and combined them with the dynamic features like trajectory and joints angles of the limbs while moving for person identification. Silhouettes of a person have been obtained using a simple method of background subtraction and then the silhouettes are analyzed by PSA method to obtain the static features. J. Man and B. Bhanu [18] has proposed gait-analysis methodology that preserves the temporal information as a single image, thus reducing the requirement storage space and minimizing the susceptibility to silhouette noise. Finally, recognition has been performed by merging the features from original and synthesized data. A model-based gait-recognition system has been proposed in [19] with the help of leg and arm movements. The authors have shown that the features obtained

from the movement of various body parts can be efficient and discriminate during recognition. Majority of the research work involving gait pertain to its usage as a biometric feature for recognizing a person from a distance as typically done in visual surveillance. For example, a viewing angle variation-based gait recognition has been proposed by Kusakunniran *et al.* [20]. They have proposed a regression-based view transformation method to address varying view angles. It has been observed that correlation exists between gait features across the views captured from various angles, and a regression has been employed to represent this correlation.

Recently, deep learning approaches have been successfully used by researchers to extract gait characteristics from videos. Wu *et al.* [21] has proposed a deep CNN model that is able to recognize changes in gait patterns that help in validating the change in individuals. The model has been trained using a small dataset, and it has been evaluated on cross-view and cross-walk scenarios as well. In [22], a CNN model with seven convolutional and pool layers followed by a fully connected layer with 4096 units and a softmax classifier has been proposed for gait recognition. Their model is robust to viewing angle's change. With the growing popularity of CNNs and deep learning, these techniques are being experimented and often applied these days. Hammerla *et al.* [23] has explored CNN- and LSTM-based approaches for developing a human-activity recognition system with gait as a biometric feature. They have experimented on three datasets with data captured during the subjects wearing different inertial sensors. A simulation-based methodology to generate synthetic video frames for data augmentation of gait sequences has been proposed in [24]. It has been observed that the synthetically generated data retains the identification traits of the subjects. Aforementioned existing works are either dependent on handcrafted features or consider data from single trial to build a gait-recognition system. In this paper, visual features are extracted with the help of 3-D CNNs by analyzing the spatial and temporal dimensions of the gait data. Our methodology has been tested on four different walk sequences that are commonly performed by humans.

B. Sensor-Guided Gait-Recognition Systems

Preliminary work in sensor-based gait recognition can be found in [25] and [26] where the authors have utilized accelerometer sensor to identify the individuals. Mntyjrvi *et al.* [25] has tied the accelerometer device on the subject's back to record the gait data and then processed the data using correlation, histogram, and frequency-based techniques. The authors have reported signal correlation method as the best way for gait recognition. L. Rong *et al.* [26] has recognized the gait patterns with the help of dynamic time warping matching algorithm. The authors have divided the acquired signals into gait cycles and extracted the relevant features where an equal error rate of 6.7% has been recorded while the subjects walk at normal speed. The use of inertial sensors can also be found in the medical domain. For example, researchers have developed solutions for various rehabilitation and diagnosis systems using gait analysis [27], [28]. Yang *et al.* [29] has developed a poststroke analysis system using two inertial sensors attached at the midpoint of each shank.



(a) (b)

Fig. 1. Pictorial representation of the Shadow Motion body sensor. (a) Sensor suit wore by a user. (b) 3-D skeleton view of the user.

T. T. Ngo *et al.* [30] has proposed a methodology to overcome the sensor orientation inconsistency and to segment action signal. Support vector machine (SVM) classifier has been used to classify five gait cycles with an accuracy of 93.36%. Integration of inertial and depth sensors (RGB-D) sensors for human gait identification has been proposed in [31]. The authors have extracted gait features from accelerometer readings in the eigenspace and by analyzing 3-D dense trajectories for RGB-D sensor. The recognition has been performed with the help of SVM classifier. Deep learning approaches have also been used in [27] and [32]. A CNN-based methodology has been proposed for gait assessment in multiple sclerosis using inertial (gyroscope, accelerometer) sensors in [27]. The training data consists of eight healthy participants who have performed 6-min walk in five different conditions. Cohen-D (Effect size) and *t*-test (p value) techniques have been used to compare the performance of different features in separability of the three groups in the testing dataset. Likewise, an integration of CNNs and RNNs has been proposed in [32] to automatically extract features from different motion sensors for various classification tasks including car tracking, activity recognition, and gait recognition. Majority of the existing methods use limited number of inertial sensors, which are positioned at specified body locations, hence, do not cover the complete gait aspect. Therefore, we have utilized the full body sensor setup that can capture large variations in human gait. Also, sensor fusion with the help of optimization has not been tried earlier. This makes our proposed method distinct and accurate.

III. Proposed System

To the best of our knowledge, no such work exists that fuses sensor as well as video data with the help of an evolutionary framework to recognize human gaits. Therefore, we have proposed a methodology that uses multiple criteria obtained using individual learning framework and fuse them to achieve the

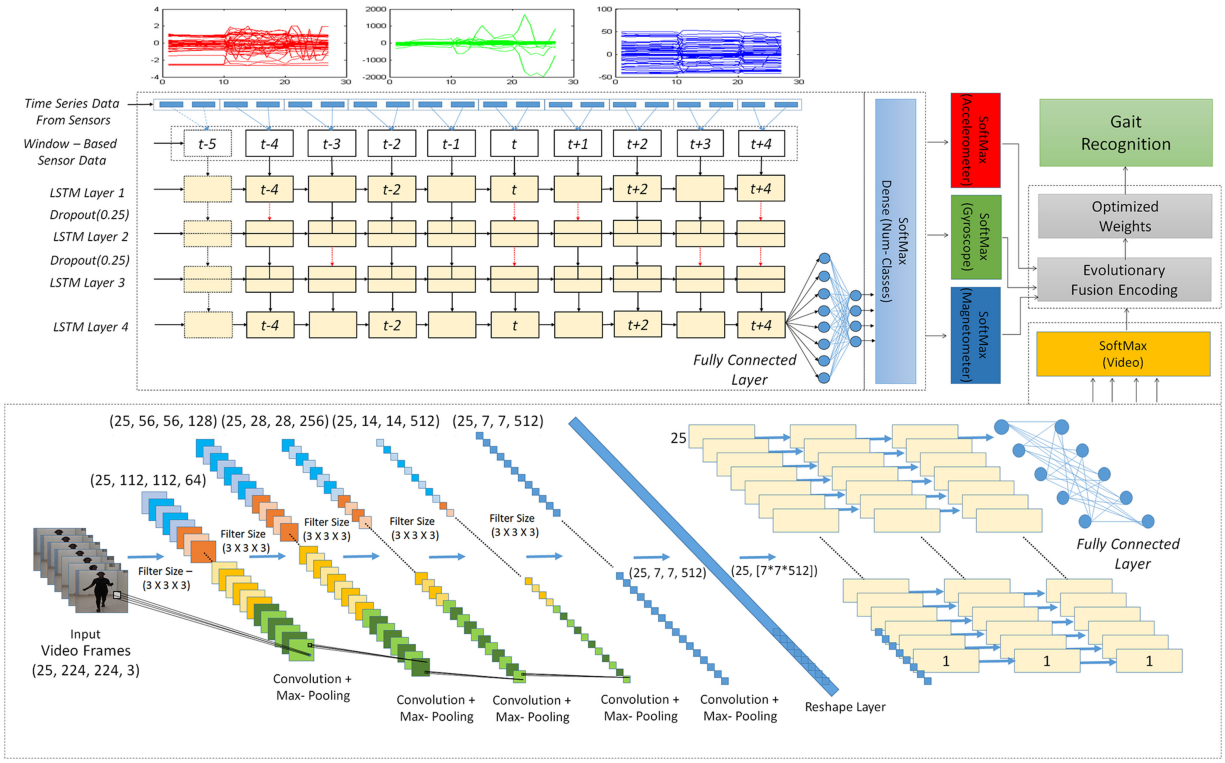


Fig. 2. Architectural detail of the proposed multimodal human-gait recognition using video and sensor fusion.

objective. We present the details of the proposed gait-recognition system that has been built using inertial sensors and video data. For inertial sensor data, we have used Shadow Motion wireless body suit. The suit is a complete solution for many applications such as animation, sports, biomechanics research, gait analysis, and posture analysis. The body-sensor suit is portable and wireless, and it does not require cameras or a permanent studio, thus making it easy to use.

A pictorial representation of the complete setup is shown in Fig. 1(a) where a user is walking by wearing the full body suit, and the corresponding 3-D representation of the human skeleton is depicted in Fig. 1(b).

The suit consists of 17 precision inertial sensors nodes and 2 pressure insoles that provide full body tracking.[1] Each node is equipped with three different sensors, namely accelerometer, gyroscope, and magnetometer that provide access to the 3-D raw data at a sampling rate of 100 Hz with the help of the software development kit. We have used this setup to acquire 3-D skeleton information from the sensors, and simultaneously a video camera is used to acquire the gait sequences of every user. Thus, the system has a total of four modalities, i.e., three inertial sensors and a video recorder, which are used to develop a gait biometric system. The flow diagram of the proposed framework is shown in Fig. 2, where we have combined multiple modalities to improve gait analysis. A 3-D CNN with stacked convolutional layers and a stack of LSTM layers is used for the analysis of optical flow fields in successive frames in video data sequences. In this architecture, we have considered 25 frames of size 224 $\times$ 224 as inputs to the 3-D CNN model. Next, we have applied 3-D convolutions with a 3 $\times$ 3 $\times$ 3 kernel. To increase the number of feature maps, multiple convolution operations are applied at each location starting from 64 with a multiplication of 2 at each layer. Similarly, we have applied Max pooling of size 2 $\times$ 2 with a stride of 2 in each layer. With max pooling, the size of the resultant image is reduced while retaining the valuable information. It also reduces the number of parameters within the model. The operation is used to make feature detection invariant to scale and orientation changes [33]. The flow fields extracted from successive frames justify the use of LSTMs as they are well known for their capability to model sequential data. Sensor data for each corresponding video are collected using accelerometers, gyroscope, and magnetometers, each of which has 3-D information and processed with the help of four stacked LSTM layers. The classification has been performed using Softmax function that returns the probability score of each class in the range [0,1] that add up to 1. Next, we have used these scores and implemented GWO algorithm to improve the gait-recognition performance by combining all modalities of the architecture. In the subsequent sections, we present each component separately.

A. 3D CNNs

CNNs are driving advances not only in fields like whole image classification but also in localization tasks, bounding box recognition, and pixel–pixel image segmentation, etc. CNNs usually consist of two parts: 1) stacked convolutional layers, which act as feature extractors and 2) fully connected networks, which classify based on the feature maps extracted by the convolutional layers. In 2-D CNNs, 2-D convolution is performed in the convolutional layers to extract features maps from a local

[1]https://www.motionshadow.com//

neighborhood from the feature maps consisting of the previous layer. Each feature map is then passed through a nonlinear activation and then passed on to the next layer. It is to be noted that 2-D CNNs consider all inputs as images, and they extract spatial features maps based on the objects inside the image. However, when applying to video analysis, it is desirable to capture the motion information encoded in successive frames. Ji *et al.* [34] has achieved excellent classification results for human-action recognition in videos using 3-D CNNs. 3-D convolutions compute features from both spatial and temporal dimensions. The 3-D convolution is achieved by convolving a 3-D kernel to the cube formed by stacking multiple successive frames together. By this construction, the feature maps in the convolution layer are connected to the successive frames in the previous layers, and hence, capture motion information. The value at position (x, y, z) on the jth feature map in the ith layer is given in (1), where f is the activation function, C_i is the size of the 3-D kernel along the temporal dimension, and w_{ijm}^{abc} is the (a, b, c)th value of the kernel connected to the mth feature map in the previous layer

$$V_{ij}^{xyz} = f\left(b_{ij} + \sum_{m}\sum_{a=0}^{Ai-1}\sum_{b=0}^{Bi-1}\sum_{c=0}^{Ci-1} w_{ijm}^{abc} v_{(i-1)m}^{(x+a)(y+b)(z+c)}\right). \tag{1}$$

It is to be noted that one 3-D kernel can extract only a single spatio–temporal feature since the kernel weights are replicated across the cube; hence, multiple filters are stacked together for extraction of multiple feature maps at each layer. The number of filters is increased with successive layers for extraction of multiple types of features from a set of low-level features extracted at the initial layers.

B. Long Short-Term Memory

LSTMs are a special form of recurrent neural network (RNN). RNNs can use their neural feedback connections to store representations of their recent inputs in the form of activations. However, the main disadvantage of an RNN lies in its inability to model long term dependencies as compared to LSTMs. LSTMs are powerful sequence classifiers and are often used to model complex sequential data. Like RNNs, an LSTM can also be unrolled in time where each time-step represents a separate input state. It is to be noted that each input state is related to its previous and future states and is not mutually exclusive, i.e., sequential. An LSTM node has four distinct blocks as shown in Fig. 3, which provide the ability to learn long-term and short-term dependencies.

1) *Cell state:* The cell state *c(t)* in Fig. 3 is identical to a conveyor belt that runs through every time-step with only few linear interactions. The cell state carries information from the previous time-steps in the form of activations. LSTMs have the ability to add or remove necessary information its cell state.
   These operations are done via gates, namely input and forget gates. They are composed of a sigmoid activation layer followed by pointwise operation.
2) *Forget Gate:* The forget gate consists of a dense network followed by a sigmoid activation. The dense layer is trained to output a weighted representation based on the information available, which is then converted into a binary representation by the sigmoid activation. This is followed by a pointwise multiplication, which masks unnecessary data on the cell state but allows the necessary values.
3) *Input Gate:* The input gate decides what data are to be added to the cell state. A tanh activation function is applied on the input data. The gate like the forget gate consists of a dense network followed by sigmoid activation. Based on the input data, the dense network outputs a weighted representation adaptively, which is then converted to a binary format by the sigmoid layer. These weights are then multiplied with the input data and added to the cell state.
4) *Output Gate:* This gate provides an output based on the cell state data and the input into the corresponding time-step. The dense layer followed by the sigmoid activation is used to perform a weighted representation of the input data, which is then multiplied with the cell states data. The output of the pointwise multiplication operation is the output of the corresponding cell state.

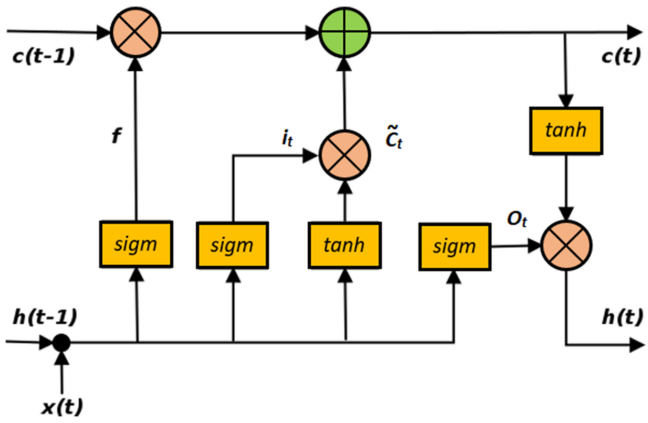


Fig. 3. Representation of a LSTM node with three gates and cell state.

The activation of different gates is calculated using (2)–(7), where h_{t-1} is the previous output and x_t denotes the current input to the LSTM node. The terms W_i, W_f, W_C, and W_o denote the weight matrices corresponding to input gate i, forget gate f, cell C, and output gate o activation vectors, respectively. Similarly, b_i, b_f, b_C, and b_o denote biases corresponding to i, f, C, and o gate activation vectors, respectively. σ is the sigmoid function, $\tan$ h is the hyperbolic tangent activation function, and h_t is the current output. This way, LSTMs can adapt themselves to any kind of sequence and learn from them

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \tag{2}$$

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \tag{3}$$

$$\widetilde{C}_t = \tan \mathrm{h}(W_C \cdot [h_{t-1}, x_t] + b_C) \tag{4}$$

$$C_t = f_t * C_{t-1} + i_t * \widetilde{C}_t \tag{5}$$

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \tag{6}$$

$$h_t = o_t * \tan \mathrm{h}(C_t). \tag{7}$$

In our implementation, we have flattened all the feature maps $F \in \mathbb{R}^{7\times7\times512}$ for each corresponding frame in the 3-D convolutional layers mentioned in Section III-A to form a vector $D \in \mathbb{R}^{1\times(7*7*512)}$. Each vector D is then fed into the corresponding time-steps of an LSTM layer. Since the number of frames in each video sample is 25, hence, there are 25 corresponding time-steps in the LSTM layer. It has been experimentally noted that LSTMs tend to overfit to the training data without proper regularization and parameter tuning due to the large number of parameters involved in the learning procedure. Srivastava *et al.* [35] has introduced an efficient method to prevent these networks from overfitting known as dropout regularization. In a dropout layer, some time-step outputs inside the network layer are randomly dropped along with their connections and no longer considered to be a part of training process. This method is used only during the training process. It forces the classifier to learn more robust features such that it is able to classify with almost no drop in performance even when dropout is applied on it.

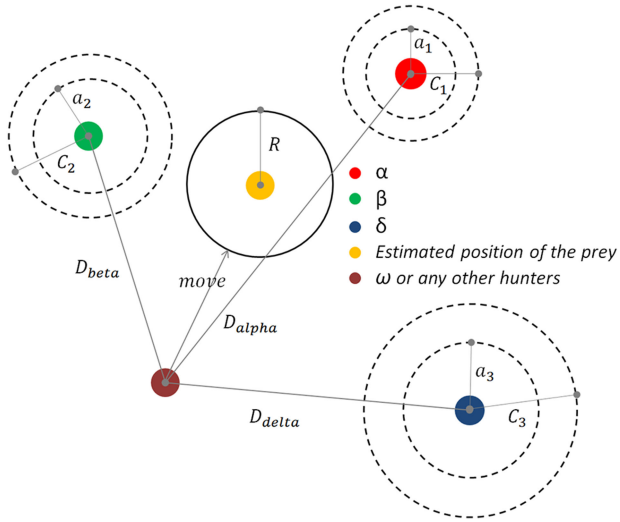


Fig. 4. Position update of wolves toward the prey with respect to the values of alpha, beta, delta, and omega.

C. GWO Scheme

GWO has gained a lot of popularity in recent years and is being used in a varied number of domains. The optimizer has been proposed by Mirjalili *et al.* [15] to give competitive or better performance in terms of exploration and exploitation of the search space, avoidance of the local minima, and faster convergence in comparison with existing metaheuristic algorithms like PSO, differential evolution, gravitational search algorithm, etc. Basically, GWO is a metaheuristic optimization technique that derives its motivation from the 4-level leadership hierarchy, namely, alpha, beta, delta, and omega, and the social behavior followed by the gray wolves while catching prey. The alpha wolves are the highest in the leadership hierarchy and their decisions are followed by the entire pack of wolves. The next in hierarchy are the beta wolves, which help the alphas in decision making or other pack-related activities. They obey the alphas and make sure the rest of the pack also does the same. The last in hierarchy are the omega wolves that form the lowest level of wolves. They have to obey all their superiors and are the last ones to eat in the pack. The wolves that do not fall under the category of either alpha, beta, or omega are termed as the delta wolves. These obey the alpha and beta wolves but command the omega wolves.

Mathematical model for social hierarchy: The range of possible population is generated in the feasible domain. The solution in the population with the best fitness is termed as alpha, the second best as beta, the third best as delta, and the rest are considered to be as omega. Another important behavior mimicked by GWO algorithm is their hunting behavior. In order to hunt the gray wolves first encircle the "prey," which is defined as the optimum point. Then hunt the prey, which is usually guided by the alpha wolf. However, the beta and delta wolves might also take part in hunting occasionally. It is supposed that the alpha, beta, and delta wolves have better knowledge of the prey and, hence, their positions is used by the other wolves to update their positions as well, which is nearer to the prey. An overview of the position update of wolves toward the prey with respect to the values of alpha, beta, delta, and omega is depicted in Fig. 4. The encircling procedure of prey by gray wolves during the hunt is defined in (8)–(11), where t represents the current iteration; $X(t)$ and $X_p(t)$ are the position vectors of the wolf and the prey, respectively. The terms A, C, and D are the coefficient and difference vectors, and a is the vector that linearly decreased from 2 to 0. r_1 and r_2 are the random vectors, which allow wolves to reach any position

$$\overrightarrow{D} = |\overrightarrow{C} \cdot \overrightarrow{X_p}(t) - \overrightarrow{X}(t)| \tag{8}$$

$$\overrightarrow{X}(t+1) = \overrightarrow{X_p}(t) - \overrightarrow{A} \cdot \overrightarrow{D} \tag{9}$$

$$\overrightarrow{A} = 2\overrightarrow{a} \cdot \overrightarrow{r_1} - \overrightarrow{a} \tag{10}$$

$$\overrightarrow{C} = 2 \cdot \overrightarrow{r_2} \tag{11}$$

The positions of the prey are estimated from the positions of the alpha, beta, and delta wolves' positions and the rest of the wolves update their positions around the prey. In this paper, we have implemented GWO to tune the fusion parameters of the proposed multimodal gait recognition. Since four different modalities have been fused together, therefore, four weights are optimized using the algorithm. Thus, the search space for optimization becomes 4-D in nature. The optimization function O_f is defined in (12), where W_1, W_2, W_3, and W_4 are the weight parameters corresponding to $V_{\text{acc}}, V_{\text{mag}}, V_{\text{gyro}}$, and V_{vid}, which defines the validation accuracies of accelerometer, magnetometer, gyroscope, and video data, respectively

$$O_f = W_1 * V_{\text{acc}} + W_2 * V_{\text{mag}} + W_3 * V_{\text{gyro}} + W_4 * V_{\text{vid}} \tag{12}$$

Similarly, the fitness function Fit is designed to minimize the classification error and is defined in (13). The vector a is formulated as per the number of iterations defined in (14). The complete procedure of GWO-based weight optimization is

Algorithm 1 GWO-Based Weight Optimization.

Require: Size of the population of wolves (Popsize), Number of iterations (Numitr), Number of variables in the function to be optimized (Novar) (in this study Novar $= 4$, i.e., $W_1, W_2, W_3,$ and W_4)
Ensure: Find global optima for the function being optimized

```
for i = 1: Popsize
  for j = 1: Novar
    Generate a random number in the feasible space
  end
  Add the generated wolf to Population
end
for q = 1: Numitr
  Evaluate fitness Fit_k for each wolf k using (13)
  Find the alpha (a), beta (b) and delta (d) wolves
  for i = 1: Popsize
    Evaluate the D_a, D_b, D_d for the ith wolf using (8) and (11) //D_a, D_b, D_d are the difference vectors from the ith wolf to a, b, and d wolves, respectively.
    Evaluate Upd_a, Upd_b, and Upd_d using (9), (10), and (14) //Upd is update for corresponding wolf.
    Evaluate NewPos = ((Upd_a + Upd_b + Upd_d)/3) // NewPos is the new position
    Evaluate NewFit (New fitness) of the ith wolf using (13)
      if Fit_i < NewFit
        Update the ith wolf with NewPos
  end
end
return alpha //global optima for the function
```

defined in Algorithm 1

$$Fit = \frac{\text{Total Number of Samples}}{\text{Correct Prediction}} \tag{13}$$

$$a = 2 - 2 * \frac{\text{Iterations}}{\text{max Iterations}} \tag{14}$$

IV. Results

We first discuss the dataset that has been recorded using Shadow Motion suit and video camera. Next, we present the results obtained using the multimodal architecture. Finally, evolutionary-algorithm-based gait-recognition results are presented.

A. Dataset Description

A total of 23 participants including 19 males and 4 females have been enrolled for dataset collection. Four different walking styles that are usually performed in daily routine have been included in the study. None of the participants has been trained prior to the recording. Therefore, the actions performed by the volunteers are unsupervised, natural, and without any constraints. The length of the walk has been set to 6 m, and each walk has been performed by every participant after wearing the Shadow Motion suit. Corresponding video has been recorded using a front-facing RGB camera. Calibrated data from all sensor nodes have been recorded through accelerometer, gyroscope, and magnetometer. Sample frames while walking are presented in Fig. 5. Corresponding readings from each sensor and the video for each type of walk are plotted in Fig. 6 when performed by a user.

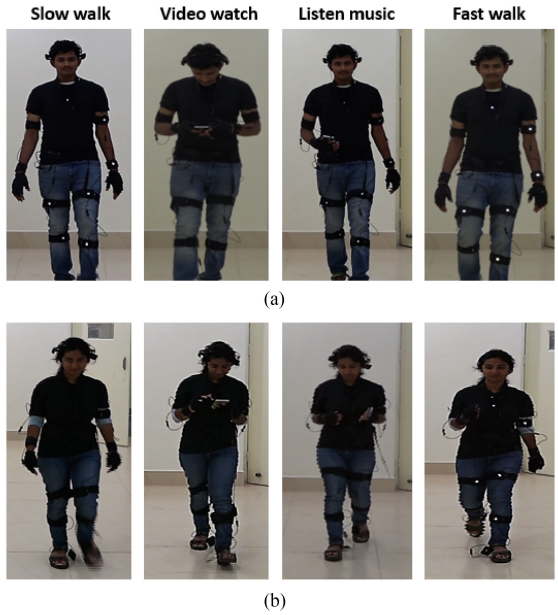


Fig. 5. Examples of the four walks (columnwise) considered in the dataset when performed by a: (a) male participant and (b) female participant.

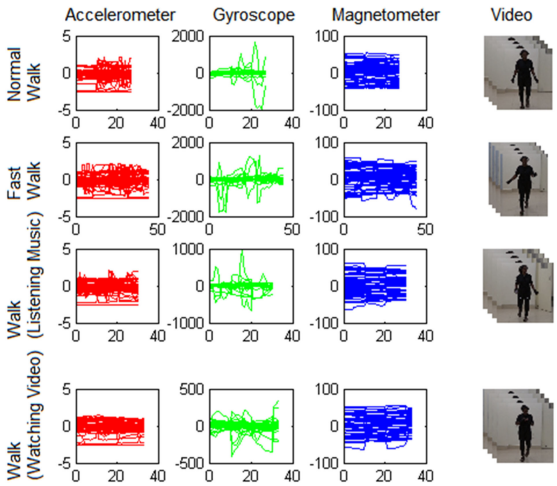


Fig. 6. Plots of different walks when performed by a user. The plots are corresponding to the readings of the different sensors and the video.

A total of 92 (i.e., 23 $\times$ 4) walking sequences have been recorded for carrying out the analysis. Next, a windowing technique has been used to process the data for each modality to represent the temporal sequence. The sensor data have been divided into multiple files with the help of a window consisting of 26 frames. Similarly, video data have been processed with a window consisting of 25 frames. The window size has been determined experimentally. A pictorial representation of the data segmentation is shown in Fig. 7.

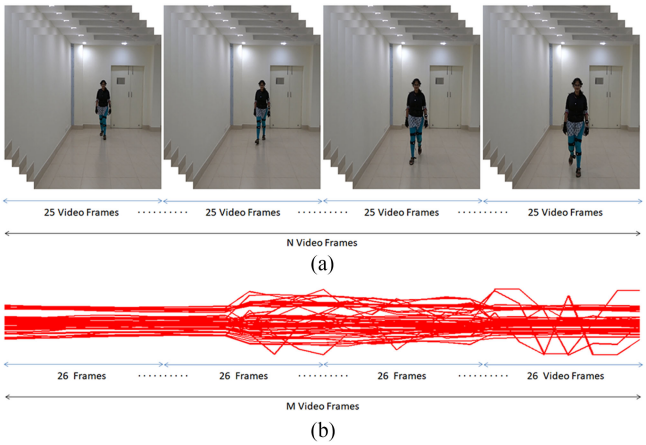


Fig. 7. Segmentation of each type of walk for processing. (a) Video data have been processed with a window of size 25 frames. (b) Sensor data have been partitioned with 26 frames.

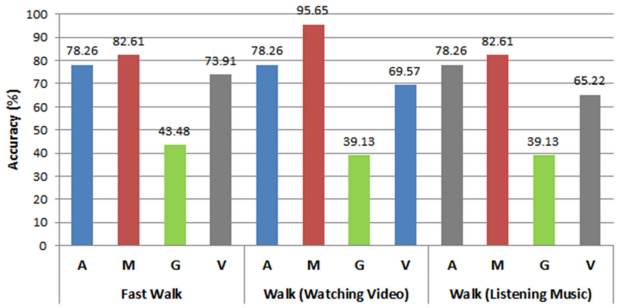


Fig. 8. Gait-recognition performance for each type of walk with unimodalities. Note: A-Accelerometer, M-Magnetometer, G-Gyroscope, and V-Video.

B. Gait Recognition Using Single Modality

The training of both networks has been performed using the data acquired by "normal walk" only, whereas the robustness of the architecture has been tested on different walk sequences. For video, 3-D CNN has been utilized to extract gait features, which are then fed into the LSTM classifier for gait recognition. A learning rate of $1e-3$ has been used with RMSprop optimizer and a decay of $1e-6$ too. The weights of the network have been initialized using the Xavier initialization that initializes the weights in such a manner that the variance remains the same. This helps in keeping the signal from exploding to a high value or vanishing to zero. Similarly, a 4-layer LSTM stack has been utilized to process the gait readings acquired from different sensors of Shadow Motion. The results of each modality as per the different walk scenarios are depicted in Fig. 8, where maximum accuracies of 82.61%, 95.65%, and 82.61% have been recorded for fast walk, video-watching walk, and music-listening walks using magnetometer sensor, respectively. The minimum accuracy has been recorded with gyroscope sensor readings across all modalities for all types of walk.

Since we have used windowing technique to process the video and sensors data, experiments have also been carried out by varying window size to decide an optimum window size for obtaining discriminative gait features. Likewise, the impact of the network depth has also been investigated to analyze gait-recognition performance. The analysis has been shown in Fig. 9, where it can be seen from the Fig. 9(a) that a window size of 26 frames for sensor data and 25 frames for videos results in higher performance. Depth of the network has been varied by stacking different LSTMs. It can be seen from Fig. 9(b) that, a 3-layer stack in video sequences and 4-layer stack in sensors results in higher recognition. The reason of low accuracies in small windows is due to the lack of temporal information to learn an effective gait model. However, while increasing the depth of the network, the accuracy remains almost the same or varies insignificantly because the network can learn more robust features with larger depths.

C. GWO-Based Multimodal Gait Recognition

Here, we present the gait-recognition rates by performing a weighted fusion of all the four modalities including sensors and the video using GWO optimizer. The optimization function and the fitness criteria are evaluated as per the formulas given in (12) and (13), respectively. The initial population has been set between 2 to 70 and the number of iterations varied from 2 to 15. Fast convergence has been noticed in the GWO algorithm where it has taken only 7 iterations to reach optimum solution with population as 9. The values of the four weights have been found as 0.8163, 0.239, 0.8541, and 0.45 for accelerometer, gyroscope, magnetometer, and video data, respectively. The results of the optimization are shown in Fig. 10, where the GWO optimizer performs better with a margin of 8.69% and 4.35% in fast walk and music-listening walk, whereas in the video-watching walk no improvement has been recorded when compared with the magnetometer-based gait recognition.

D. Comparative Study

To show the effectiveness of GWO algorithm, a comparison with other evolutionary algorithms has been performed. For this, we have tuned the fusion parameters with PSO [36] and GA [37]. GA is a metaheuristic algorithm that derives its inspiration from the Darwin's theory of evolution. It makes use of the criteria of inheritance and the survival of the fittest. It works on a population of chromosomes, i.e., solutions in the feasible search domain and updates the population via crossover and mutation operators. Crossover helps in creating children chromosomes by taking the aspects from their parents. Mutation helps in inducing a change in the chromosomes in order to increase the possibility of avoidance of local minima or maxima. Similarly, PSO is an optimization algorithm that helps to mimic the behavior of the swarm of birds or school of fish. These creatures wander in group while maintaining a safe distance from one another and also have the tendency to reach a fixed destination. In the swarm, each bird or fish helps its neighbors to update their positions with respect to them, thus, ensuring that the whole group travels together and in the right direction. This algorithm also works on a population of solutions in the feasible domain and the value of each particle is updated so that an optimal point is reached. The optimization and fitness functions are same as those for GWO optimizer. We have recorded similar performance of all the three algorithms. However, the convergence rate varies for all the algorithms. Therefore, we have shown a comparison among the convergence

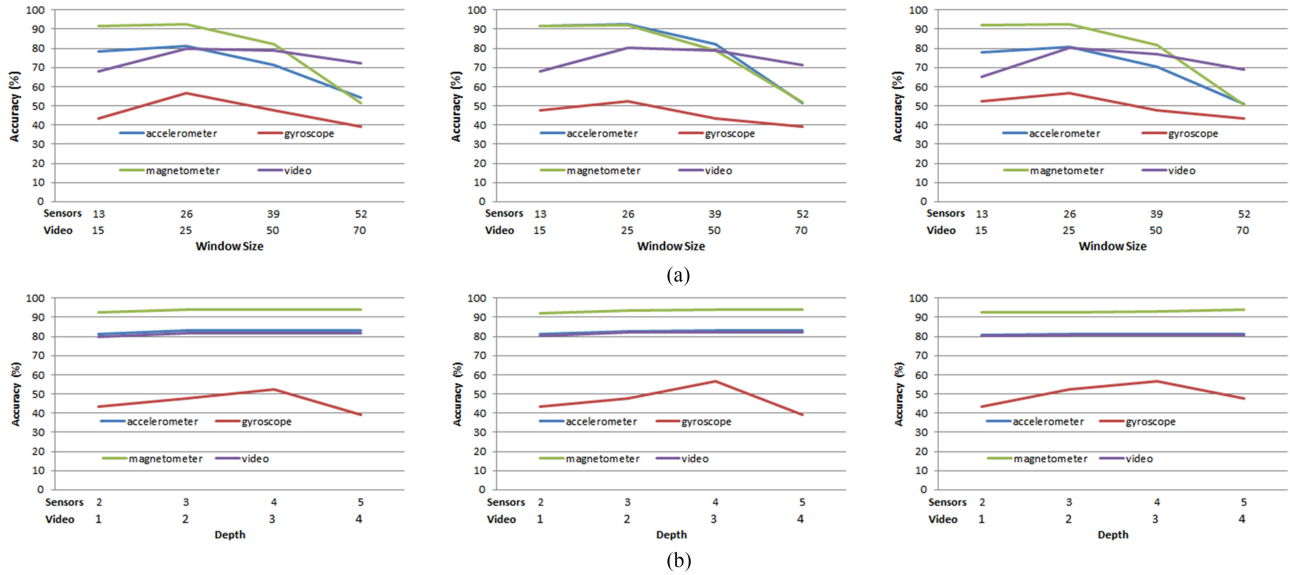


Fig. 9. Gait-recognition rates by varying: (a) window size to process the input frames and (b) depth of the network in terms of LSTM layers.

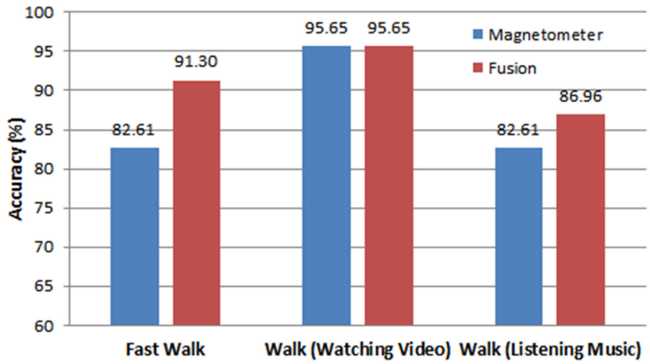


Fig. 10. Gait-recognition performance improvement using the proposed Gray wolf based algorithm when compared with magnetometer.

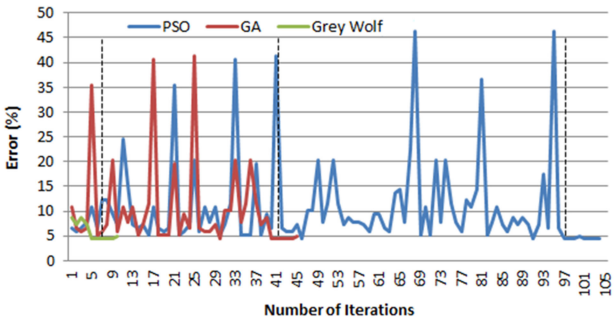


Fig. 11. Convergence comparison between different evolutionary algorithms.

to minimize classification error in different iterations. Fig. 11 depicts the variation in error as per the number of iterations for the 3 evolutionary algorithms, where GWO takes only 7 iterations to reach optimum solution in comparison to GA and PSO optimizers that have converged in 42 and 97 iterations, respectively.

TABLE I
COMPARATIVE ANALYSIS WITH EXISTING GAIT-RECOGNITION SYSTEMS

Author	CASIA B Gait Dataset	Accelerometer Gait Dataset[31]	Proposed Methodology
Liu et al. [38], 2016	83.87	-	**89.71**
Zou et al. [31], 2017	-	89.77	**92.47**

A comparison of the proposed system has been performed with other classification schemes such as SVM and random forest (RF). Since these classifiers are not sequential, we have extracted three statistical features, i.e., mean, standard deviation, and sum from the sensors data. The training has been performed using "normal-walk" data while other walks have been used for testing. Average accuracies of 83.57% and 71.34% have been recorded using SVM and RF classifiers with magnetometer readings, whereas the proposed GWO-based system performs with an average accuracy of 91.3%.

In addition, a comparison with state-of-the-art gait-recognition techniques has been performed. For this experiment, publicly available CASIA Gait Dataset B[2] has been used. The dataset contains videos of 124 subjects captured from different view angles. We have used 90° angle recordings for comparative analysis, and experiments have been performed by focusing on two normal and two carrying sequences [38]. Likewise, Zou *et al.* [31] has proposed a gait-recognition system using accelerometer and RGB-D sensors. The authors have used SVM classifier for gait identification. Their dataset contains gait sequences of normal and fast walks performed by 10 individuals. The comparison is shown in Table I. It may be observed that the proposed method outperforms both of the above-mentioned state-of-the-art techniques.

[2]http://www.cbsr.ia.ac.cn/english/Gait%20Databases.asp

V. Conclusion

In this paper, we have proposed a multimodal gait-recognition approach using Shadow Motion sensor and video sequences. We have found that the walking pattern of every individual person depends on the activity in which the user is engaged. Therefore, we have analyzed different walking patterns of individuals with four different walks, namely, *normal walk*, *fast walk*, *walking while listening to music,* and *walking while watching video on mobile*. Data from three different sensors, i.e., accelerometer, gyroscope, magnetometer, and video have been acquired. 3-D CNNs have been utilized to process the videos and to extract spatial and temporal features, which are then processed using LSTMs. Similarly, another LSTM architecture has been used to model the inertial sensors data for gait recognition. Finally, an evolutionary algorithm (GWO) has been implemented to fuse all modalities to enhance the gait-recognition performance. An average accuracy of 91.3% has been recorded using the GWO optimizer on all walk sequences when training is performed on *normal-walk* data. In future, the work can be extended by exploring other evolutionary algorithms that can tune the fusion parameters effectively.

References

[1] S. Yu, H. Chen, Q. Wang, L. Shen, and Y. Huang, "Invariant feature extraction for gait recognition using only one uniform model," *Neurocomputing*, vol. 239, pp. 81–93, 2017.

[2] H. Stolze, J. P. Kuhtz-Buschbeck, H. Drücke, K. Jöhnk, M. Illert, and G. Deuschl, "Comparative analysis of the gait disorder of normal pressure hydrocephalus and Parkinson's disease," *J. Neurol., Neurosurg. Psychiatry*, vol. 70, no. 3, pp. 289–297, 2001.

[3] J.-H. Yoo and M. Nixon, "Feature extraction and selection for recognizing humans by their gait," *Adv. Vis. Comput.*, 2006, pp. 156–165.

[4] M. A. Hossain, Y. Makihara, J. Wang, and Y. Yagi, "Clothing-invariant gait identification using part-based clothing categorization and adaptive weight control," *Pattern Recognit.*, vol. 43, no. 6, pp. 2281–2291, 2010.

[5] L. Wang, T. Tan, H. Ning, and W. Hu, "Silhouette analysis-based gait recognition for human identification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 12, pp. 1505–1518, Dec. 2003.

[6] S. D. Choudhury and T. Tjahjadi, "Silhouette-based gait recognition using procrustes shape analysis and elliptic Fourier descriptors," *Pattern Recognit.*, vol. 45, no. 9, pp. 3414–3426, 2012.

[7] M. Balazia and P. Sojka, "Learning robust features for gait recognition by maximum margin criterion," in *Proc. 23rd Int. Conf. Pattern Recognit.*, 2016, pp. 901–906.

[8] T. T. Ngo, Y. Makihara, H. Nagahara, Y. Mukaigawa, and Y. Yagi, "The largest inertial sensor-based gait database and performance evaluation of gait-based personal authentication," *Pattern Recognit.*, vol. 47, no. 1, pp. 228–237, 2014.

[9] S. Sprager and M. B. Juric, "Inertial sensor-based gait recognition: A review," *Sensors*, vol. 15, no. 9, pp. 22089–22127, 2015.

[10] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *Proc. Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 1725–1732.

[11] H. Sak, A. Senior, and F. Beaufays, "Long short-term memory recurrent neural network architectures for large scale acoustic modeling," in *Proc. 15th Annu. Conf. Int. Speech Commun. Assoc.*, 2014, pp. 338–342.

[12] Y. Deng, Z. Ren, Y. Kong, F. Bao, and Q. Dai, "A hierarchical fused fuzzy deep neural network for data classification," *IEEE Trans. Fuzzy Syst.*, vol. 25, no. 4, pp. 1006–1012, Aug. 2017.

[13] A. J. Pinar, J. Rice, L. Hu, D. T. Anderson, and T. C. Havens, "Efficient multiple kernel classification using feature and decision level fusion," *IEEE Trans. Fuzzy Syst.*, vol. 25, no. 6, pp. 1403–1416, Dec. 2017.

[14] A. Kumar and A. Kumar, "Adaptive management of multimodal biometrics fusion using ant colony optimization," *Inf. Fusion*, vol. 32, pp. 49–63, 2016.

[15] S. Mirjalili, S. M. Mirjalili, and A. Lewis, "Grey wolf optimizer," *Adv. Eng. Softw.*, vol. 69, pp. 46–61, 2014.

[16] E. Emary, H. M. Zawbaa, C. Grosan, and A. E. Hassenian, "Feature subset selection approach by gray-wolf optimization," in *Proc. Afro-Eur. Conf. Ind. Advancement*, 2015, pp. 1–13.

[17] L. Wang, H. Ning, T. Tan, and W. Hu, "Fusion of static and dynamic body biometrics for gait recognition," *IEEE Trans. Circuits Syst. for Video Technol.*, vol. 14, no. 2, pp. 149–158, Feb. 2004.

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

[19] F. Tafazzoli and R. Safabakhsh, "Model-based human gait recognition using leg and arm movements," *Eng. Appl. Artif. Intell.*, vol. 23, no. 8, pp. 1237–1246, 2010.

[20] W. Kusakunniran, Q. Wu, J. Zhang, and H. Li, "Gait recognition under various viewing angles based on correlated motion regression," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 22, no. 6, pp. 966–980, Jun. 2012.

[21] 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.

[22] T. Wolf, M. Babaee, and G. Rigoll, "Multi-view gait recognition using 3D convolutional neural networks," in *Proc. Int. Conf. Image Process.*, 2016, pp. 4165–4169.

[23] N. Y. Hammerla, S. Halloran, and T. Ploetz, "Deep, convolutional, and recurrent models for human activity recognition using wearables," in *Proc. Twenty-Fifth Int. Joint Conf. Artificial Intell.*, Jul. 2016, pp. 1533–1540.

[24] C. C. Charalambous and A. A. Bharath, "A data augmentation methodology for training machine/deep learning gait recognition algorithms," in *Proc. British Mach. Vis. Conf.*, R. C. Wilson, E. R. Hancock, and W. A. P. Smith, Eds. BMVA Press, Sep. 2016, pp. 110.1–110.12.

[25] J. Mantyjarvi, M. Lindholm, E. Vildjiounaite, S.-M. Makela, and H. A. Ailisto, "Identifying users of portable devices from gait pattern with accelerometers," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, vol. 2, 2005, pp. ii/973–ii/976.

[26] L. Rong, D. Zhiguo, Z. Jianzhong, and L. Ming, "Identification of individual walking patterns using gait acceleration," in *Proc. 1st Int. Conf. Bioinformat. Biomed. Eng.*, 2007, pp. 543–546.

[27] J. Gong, M. D. Goldman, and J. Lach, "Deepmotion: A deep convolutional neural network on inertial body sensors for gait assessment in multiple sclerosis," in *Proc. Wireless Health*, 2016, pp. 164–171.

[28] A. Rampp, J. Barth, S. Schülein, K.-G. Gaßmann, J. Klucken, and B. M. Eskofier, "Inertial sensor-based stride parameter calculation from gait sequences in geriatric patients," *IEEE Trans. Biomed. Eng.*, vol. 62, no. 4, pp. 1089–1097, Apr. 2015.

[29] S. Yang, J.-T. Zhang, A. C. Novak, B. Brouwer, and Q. Li, "Estimation of spatio-temporal parameters for post-stroke hemiparetic gait using inertial sensors," *Gait Posture*, vol. 37, no. 3, pp. 354–358, 2013.

[30] T. T. Ngo, Y. Makihara, H. Nagahara, Y. Mukaigawa, and Y. Yagi, "Similar gait action recognition using an inertial sensor," *Pattern Recognit.*, vol. 48, no. 4, pp. 1289–1301, 2015.

[31] Q. Zou, L. Ni, Q. Wang, Q. Li, and S. Wang, "Robust gait recognition by integrating inertial and RGBD sensors," *IEEE Trans. Cybern.*, vol. 48, no. 4, pp. 1136–1150, Apr. 2018.

[32] S. Yao, S. Hu, Y. Zhao, A. Zhang, and T. Abdelzaher, "Deepsense: A unified deep learning framework for time-series mobile sensing data processing," in *Proc. 26th Int. Conf. World Wide Web*, 2017, pp. 351–360.

[33] C.-Y. Lee, P. W. Gallagher, and Z. Tu, "Generalizing pooling functions in convolutional neural networks: Mixed, gated, and tree," in *Proc. Artif. Intell. Statist.*, 2016, pp. 464–472.

[34] S. Ji, W. Xu, M. Yang, and K. Yu, "3D convolutional neural networks for human action recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 221–231, Jan. 2013.

[35] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, 2014.

[36] J. Kennedy, "Particle swarm optimization," in *Encyclopedia of Machine Learning*. New York, NY, USA: Springer-Verlag, 2011, pp. 760–766.

[37] K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan, "A fast and elitist multiobjective genetic algorithm: NSGA-II," *IEEE Trans. Evol. Comput.*, vol. 6, no. 2, pp. 182–197, Apr. 2002.

[38] D. Liu, M. Ye, X. Li, F. Zhang, and L. Lin, "Memory-based gait recognition," in *Proc. Brit. Mach. Vis. Conf.*, 2016, pp. 82.1–82.12.

Authors' photographs and biographies not available at the time of publication.