Multi-modality sensor fusion for gait classification using deep learning

Document Version
Accepted author manuscript

Link to publication record in Manchester Research Explorer

Citation for published version (APA):

Citing this paper
Please note that where the full-text provided on Manchester Research Explorer is the Author Accepted Manuscript or Proof version this may differ from the final Published version. If citing, it is advised that you check and use the publisher's definitive version.

General rights
Copyright and moral rights for the publications made accessible in the Research Explorer are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

Takedown policy
If you believe that this document breaches copyright please refer to the University of Manchester’s Takedown Procedures [http://man.ac.uk/04Y6Bo] or contact uml.scholarlycommunications@manchester.ac.uk providing relevant details, so we can investigate your claim.
Abstract— Human gait has been acquired and studied through modalities such as video cameras, inertial sensors and floor sensors etc. Due to many environmental constraints such as illumination, noise, drifts over extended periods or restricted environment, the classification f-score of gait classifications is highly dependent on the usage scenario. This is addressed in this work by proposing sensor fusion of data obtained from 1) ambulatory inertial sensors (AIS) and 2) plastic optical fiber-based floor sensors (FS). Four gait activities are executed by 11 subjects on FS whilst wearing AIS. The proposed sensor fusion method achieves classification f-scores of 88% using artificial neural network (ANN) and 91% using convolutional neural network (CNN) by learning the best data representations from both modalities.

Keywords— ambulatory inertial sensors (AIS), floor sensors (FS), deep learning (DL), multi-modality sensor fusion, artificial neural network (ANN), convolutional neural network (CNN).

I. INTRODUCTION AND BACKGROUND

Gait defines a unique pattern of human locomotion. Since the latter depends strongly on each person’s own biological characteristics such as age, weight, height and gender, people can be differentiated on the basis of their walking patterns. Gait has been extensively used in studies of walking behaviours among healthy people [1] and people suffering from neurological diseases such as Huntington’s disease (HD), Amyotrophic Lateral Sclerosis (ALS) and Parkinson’s disease (PD) [2]. Moreover, outside the area of healthcare, gait has been exploited by other fields such as biometric security systems [3].

Gait information can be collected non-invasively unlike any other biometrics i.e., iris, face, fingerprints and other biometric identifiers. Gait patterns are difficult to duplicate due to their individual nature. However, gait patterns get affected by many factors such as illness, fatigue, emotions, cognitive and motor tasks [4]. In addition, gait is also prone to influence from external factors such as clothing, shoes or carrying load [5].

Among the many ways to acquire gait data, vision based systems are dominant due to their non-intrusive nature [6]. As vision based systems make use of several cameras, they require demanding resolution and illumination. Floor sensors (FS) are unobtrusive and mainly based on resistive plates, capacitive plates, piezoelectric sensors or fiber optic cables. These systems are typically installed indoors, in controlled environments such as offices and buildings. Most FS have been employed to record physiologically defined features, such as centre of pressure, step length and cadence, rather than for collecting raw data over longer periods of time [7],[8]. Recently, inertial sensors have been actively and widely used to acquire gait information due to their small size, weight and cost [9]. An important factor to consider is that although ambulatory inertial sensors (AIS) are non-invasive, they require the individual to cooperate in wearing them on different body parts such as head, waist, chest, thigh, shank and foot to record gait signals [10].

In the past decade, an exponential rise in the efficiency and capabilities of sensor systems has enabled the extraction of more valuable information from sensing modalities. Advances in gait sensing instruments have resulted in the evaluation of many human locomotion characteristics obtained from high quality data. Furthermore, this corroborated that the complex nature of gait brings the need of multisource and multi-modality sensor fusion of data with diverse characteristics [11]. In this context, multi-modality sensor fusion uses information from multiple sources to produce new data representations which are unique to the collection of individual sources. DL has attained a lot of attention due to its improved accuracy when dealing with larger datasets and quality of data. It maximizes the use of data variance and removes the dependencies on handcrafted features. A survey of results for gait classification using multi-modality sensor fusion is presented in table 1.

In this work, we propose a reliable and efficient approach to fuse spatio-temporal information obtained from two different gait modalities, i.e. AIS and FS, using DL. This direction is also

<table>
<thead>
<tr>
<th>References</th>
<th>Features Classification</th>
<th>Models</th>
<th>Modalities</th>
<th>Test Measures</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vera-Rodriguez et al., 2013 [21]</td>
<td>Foot step and gait</td>
<td>SVM</td>
<td>Video &amp; floor sensors</td>
<td>EER = 4.8</td>
</tr>
<tr>
<td>Mazumder et al., 2017 [22]</td>
<td>Stride time and gait phase</td>
<td>RB-FNN</td>
<td>Inertial, EMG &amp; pressure sensors</td>
<td>Accuracy = 0</td>
</tr>
<tr>
<td>Ding et al., 2018 [23]</td>
<td>Gait phase</td>
<td>MLP</td>
<td>Inertial &amp; foot switch sensors</td>
<td>Accuracy = 91.8%</td>
</tr>
<tr>
<td>Mun et al. 2018 [24]</td>
<td>Gait phase</td>
<td>ANN</td>
<td>Inertial &amp; pressure sensors</td>
<td>Accuracy = 95%</td>
</tr>
<tr>
<td>Vu et al., 2018 [25]</td>
<td>Gait phase</td>
<td>ED-FNN</td>
<td>Inertial &amp; force resistors sensors</td>
<td>MSE = 0.006</td>
</tr>
<tr>
<td>Kamaret al., 2019 [26]</td>
<td>Gait</td>
<td>3d CNN &amp; LSTM</td>
<td>Video, inertial &amp; pressure sensors</td>
<td>Accuracy = 91.3%</td>
</tr>
</tbody>
</table>
motivated by understanding the characteristics of sensor fusion using DL algorithms. The state of the art is clearly described in Table 1. Experimental results show that our proposed solutions can produce better results in terms of f-scores and training time. The effectiveness of this approach has been tested and verified using DL based methods such as ANN [12]-[13] and CNN [14]. This paper is organized as follows: Section II describes the acquisition and pre-processing of gait information obtained from AIS and FS. Section III explains the deep learning based fusion of information using multi-modalities. Results have been further discussed in section IV, followed by conclusions in section V.

II. METHODOLOGY

Two systems i.e., AIS and FS were used in this research to acquire data for the purpose of subsequent analysis. Experiments conducted used both modalities simultaneously, requiring the subject to attach AIS on their body whilst performing different types of gait on FS. Signals were collected as shown in figure 1.

The intermediate processors for both modalities are Raspberry Pi-s (RPi), single-board computers used to calibrate, acquire and save the spatio-temporal information obtained from each modality. Further, this information is transferred through Wi-Fi from both RPi’s to a workstation for real-time/offline processing and analysis. Our proposed multi-modal follows as:

A. AIS modality

A portable AIS system, comprising a RPi with a Sense-HAT board [15] and two IMU sensor boards, has been developed and deployed for the purpose of this research. RPi III (Model b+) is a small credit card sized computer with a quad core 1.4GHz processor, 1GB RAM, Bluetooth and Wi-Fi. Sense-HAT is an additional board compatible with RPi series featuring a built-in +/−16g 3D accelerometer and a +/−2000dps gyroscope. The same gyroscope and accelerometer specifications are also deployed on each of the other two IMU boards (9DoF Razor from SparkFun [16]) each with an on-board 32 bit Atmel SAMD21 cortex-M0+ microprocessor. The 9DoF IMUs, each of dimensions 15.24x10.16x2.54 cm and weight of 9.07 g, are attached through USB cables to the RPi.

AIS is built with the aim to study the gait parameters when attached to the lower half of the body. As shown in figure 2, the RPi and Sense-HAT, together with a battery bank, are worn on the pelvis area of the subject’s body; the two additional IMU sensors are attached to the ankle of each leg to capture maximum gait dynamics. AIS is capable of saving spatio-temporal gait information from all three sensors on the RPi’s memory and transfer that information further to a workstation over Wi-Fi.

From AIS, raw data for acceleration and angular velocity values is obtained from sensors 1-3. The default sampling frequency of sensor 1 is around 30Hz while sensor 2 and sensor 3 are sampled at 100Hz. After filtering and re-sampling, the spatio-temporal information from all three sensors of AIS is synchronized at 20Hz. Raw acceleration values are in two’s compliment format, therefore they are converted into values between +16g and -16g. To calculate the angle ($\theta$) from raw angular velocity ($\omega$) values, the following formula is used:

$$\theta = \omega \cdot \Delta t + \theta$$

where $\Delta t$ is the time step. The nature of experiments requires subjects to start walking from one end of the FS to the other in forward direction whilst wearing the AIS. Therefore, all IMUs are aligned so that the highest acceleration (in forward direction) is represented by X-axis; weaker acceleration (vertical, in up/down direction between heel strike and toe off events of each foot) is represented by Z-axis; weakest acceleration (lateral, in left/right direction) is represented by Y axis, as shown in figure 2.

The acceleration and angular velocity values obtained by AIS for different gait patterns are pre-processed and organised into spatio-temporal frames for further analysis. Routines to observe and verify the linear acceleration and the angular velocities values obtained from three sensors during individual experiments. For real time analysis, coding in Python has been implemented as shown in figure 3. Data obtained from AIS is analysed for fault finding and sanity checks before being handed over for multi-modality sensor fusion.
B. FS

In this work, FS (size: 2 m x 1 m approx.) is used to record gait parameters comprising 116 plastic optical fiber (POF) cables (each terminated with an LED as a light source and a photodiode as detector) [17]. The three-ply arrangement of POF cables on FS is shown in figure 4. These POF cables are further connected through an umbilical cord to a RPi. FS is selected to operate at 20 Hz which is also the synchronized frequency of AIS. FS is capable to save spatio-temporal information on RPi and also send/receive this information to a workstation over Wi-Fi as shown in figure 1.

The average of the spatial information obtained from FS is calculated using the formula:

$$SA[t] = \frac{1}{n} \sum_{k=1}^{n} F_k[t]$$

(2)

where n is the total number of sensors and $F_k$ is the amplitude of the individual sensor response. The average signal amplitude of 116 sensors over time for different gait patterns is shown in figure 5. This spatial information for all the time patterns of each experiment is converted into spatio-temporal frames to be analysed by the DL algorithm.

![Fig. 4. 3-Ply arrangement of 116 POF cables on FS; Ply-1 with 22 sensors, Ply-2 with 47 sensors and Ply-3 with 47 sensors.](image)

Some of the gait events are possible to identify by visual inspection of data obtained separately from both modalities. For illustration purposes, in figure 6 only HS is indicated on time patterns from the FS signal (as the mean of all sensor values) and the AIS signal (as the root sum of squared maximum accelerations in all 3 directions, given by (3)):

$$A_{max} = \sqrt{A_x^2 + A_y^2 + A_z^2}$$

(3)

![Fig. 6. (a) Mean of sensor values vs time frames obtained from FS; (b),(c),(d) Root sum of maximum accelerations vs time frames from AIS sensor-1,sensor-2,sensor-3 for the normal walk gait pattern.](image)

D. Experiments and Data Collection

A total of 11 healthy young male and female volunteers in the age bracket 20-35 years took part in data collection. Gait activities comprised: normal walk, fast walk, subtracting number 7 from a given random number whilst walking and texting on a personal mobile. Each activity was repeated 10 times. The dataset comprises 52,800 samples with 116 input patterns from FS and another 52,800 samples with 18 input values from AIS [18 input values = 3 sensors x (3 linear acceleration values + 3 angular velocity values)].

III. DEEP LEARNING BASED SENSOR FUSION

There are several fields, such as healthcare, sport, biometric security etc., where deep learning (DL) is a candidate to fuse spatio-temporal gait data to obtain new representations using multi-modalities [11]. In this work we focus on fusion of data, potentially containing gait anomalies, from the two modalities using DL. The DL approaches used for this purpose are briefly outlined below:

C. Visual Gait Analysis

The full gait cycle can be represented (see figure 1 in [11]) as 5 events in the stance phase, starting with a heel strike (HS) and finishing with the HS of the opposite foot.

![Fig. 5. Average of the FS signals vs time frames during gait patterns (a) normal walk (b) fast walk (c) walk while subtracting 7 (d) walk while texting on a mobile](image)
1) Artificial Neural Networks (ANN)

ANNs are employed to learn complex patterns [18]. The basic architecture of an ANN model consists of an input layer, few hidden layers and an output layer of neurons. In a fully connected layer, the neurons in one layer are fully connected to the previous layer through synapses or assigned weights. The output of a \( k_{th} \) neuron is given by

\[ L_k = \sum_{i=0}^{n} w_{ik} x_i + b_k \]  

where \( i \) enumerates all neurons in the layer, \( n \) is the total number of neurons in the layer, \( x \) is the input, \( w \) is the weight and \( b \) is the bias of the \( k_{th} \) neuron. The effect of every weight is limited by the activation function. To address the non-linearity of the spatio-temporal gait patterns in our dataset a Rectified Linear Unit (ReLU) activation function

\[ Z_k = \max(L_k, 0) \]  

with input \( L \), was used in all the hidden layers. In the output layer, \( Z \) is the input vector for a linear classifier, namely a Softmax function given by

\[ f(z) = \frac{e^{z_i}}{\sum_{k} e^{z_k}} \]  

The complete ANN works as follows:

1. Weights are initialized with a value close to zero.
2. In the input layer, each input value from the dataset is represented by an input node.
3. Forward Propagation: Neurons get activated the outputs from each layer are passed in forward direction to the next layer. The weight of every neuron is multiplied by the input and passed through the activation function. Propagation continues until a prediction \( y \) is achieved.
4. The predicted results is compared with the actual truth and the error is quantified.
5. Backpropagation: The error is back propagated in the form of updated weights send to the neurons layer-wise in backward direction. Learning rate determines new weights.
6. Steps 1-5 are repeated and weights are updated after each observation (reinforcement learning) or a batch of observations (batch learning) from the training set.
7. One epoch is completed when one whole training set passes through the ANN. In this work, epochs were repeated for 100 times with batch size of 10 in our work.

The layer-wise implementation in this work of ANN for AIS, FS and multi-modal AIS+FS can be seen in figure 7. For AIS, the training set is In AIS, the training set is a 2D vector(42240x18). 18 Input values are passed to the fully connected layers of size 1x16, 1x10, 1x8 and an output layer of size 1x4. For FS, the training set is also a 2D vector(42240 x 116). 116 input values are passed to the fully connected layers of sizes 1x64, 1x16, 1x8 and an output layer of size 1x4. For the fused case, in order to create a comparison with individual modalities, the last fully connected layer of size 1x8 from each modality is merged into a single fully connected layer. Multiple cases for merging are discussed in section VI.

2) Convolutional Neural Networks (CNN)

CNN is a typical DL model which uses different levels of abstraction to learn the hierarchical representations of patterns existing in the dataset [19]. In this work, CNN consists of an input layer, convolution layers, down-sampling or pooling layers, flattening layers, fully connected layers and an output layer. The CNN functionality can be summarized as:

1. Convolution: A convolution operation is applied to automatically extract the unique variability features from the dataset. The output of a convolution layer is given by:

\[ c_i^k = \sigma(b_k + \sum_{i=0}^{n} w_{ik} x_{i+i-1}) \]  

where \( \sigma \) is the activation function, \( b_k \) is the bias, \( w_{ik} \) is weight of the \( k_{th} \) feature map and \( n \) is the size of the convolution kernel.

2. Pooling: Max pooling is used in this work to down-sample the large volume of data after convolution. Max-pooling outputs the maximum value from the nearby input values as given by:

\[ m_i^k = \max(m_{i+i-1}) \]  

where \( L \) is the stride length and \( o \) is the pooling size.

3. Flattening: Extracted features are the 2D feature maps and need to get aligned in the form of a 1D feature vector of inputs for fully connected layers:

\[ f(z) = [f_1, f_2, \ldots f_k] \]  

where \( k \) is the number of outputs from last pooling layer.

4. Full connection: A fully connected layer is a specific type of hidden layer which is fully connected to the input layer in a convolutional neural network.
ReLU (see V.1) is the activation function used to handle the non-linearity at the convolution layers and the fully connected layers. Softmax is used at the output layer as discussed earlier. The implementation of CNN using AIS, FS, and multi-modal AIS+FS can be seen in figure 8.

In FS, the training set is a 3D vector (352x120x18). 18 Input values are passed as frames (120x18) to the convolution layer1 (16 filters, kernel size 3), max-pooling layer 1 (kernel size 2), convolution layer 2 (10 filters, kernel size 3), max-pooling layer 2 (kernel size 2), convolution layer 3 (8 filters, kernel size 3), max-pooling layer 3 (kernel size 2), flattening layer (1x120), fully connected layer (1x8) and the output layer (1x4), respectively. In FS, the training set is also a 3D vector (352x120x116). 116 input values are passed as frames (120x116) to the convolution layer 1 (64 filters, kernel size 3), max-pooling layer 1 (kernel size 2), convolution layer 2 (16 filters, kernel size 3), max-pooling layer 2 (kernel size 2), convolution layer 3 (8 filters, kernel size 3), max-pooling layer 3 (kernel size 2), flattening layer (1x120), fully connected layer (1x120) and the output layer (1x4), respectively. For the fused case, in order to create a comparison with individual modalities the last flattening layer (1x120) from each modality is merged into a single fully connected layer (1x120). Multiple cases for merging are discussed in section VI.

**3) Sensor Fusion**

Features that contribute more towards gait classification are extracted and the rest are dropped by the DL algorithms. In our work, the extracted gait features from both modalities represent movements in lower parts of the body using AIS and the foot falls on the floor using FS respectively. Fusion of such a valuable information helps to capture most of the gait dynamics from individual modalities. Different techniques have been proposed and adapted about sensor fusion using DL [20]. Some of the strategies implemented in this work, are as follows:

1. **Add**: Adds two same sized inputs vectors (element wise) into a single vector of same size as inputs.
2. **Multiply**: Multiply two inputs vectors same as add.
3. **Average**: Computes the average of two input vectors.
4. **Maximum**: Computes the maximum of the two input vectors (element wise) into a single vector of same size.
5. **Minimum**: Computes the minimum of the two input vectors.
6. **Concatenate**: Combines two inputs vectors into a single long vector, so that that the second input comes after the first.

The approach in this work has been to use concatenation to fuse the unique gait features obtained from both modalities using DL. Fused data is further tested and verified across the validation and testing datasets. The concatenation operation was performed utilising the Keras library in Python.

**IV. RESULTS AND DISCUSSION**

We have presented an end to end approach of multi-modal sensor fusion for gait classification that takes values from AIS as inputs. The effectiveness of this sensor fusion has been tested and verified using DL based ANN and CNN. 10-fold cross validation has been employed to attain the best hyper-parameter values for all algorithms. Results obtained using the discussed fusion methods are shown in Table II.

**TABLE II. F-SCORES OF GAIT CLASSIFICATION USING DIFFERENT FUSION METHODS**

<table>
<thead>
<tr>
<th>Method</th>
<th>ANN</th>
<th>CNN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Add</td>
<td>88%</td>
<td>89%</td>
</tr>
<tr>
<td>Multiply</td>
<td>85%</td>
<td>87%</td>
</tr>
<tr>
<td>Average</td>
<td>85%</td>
<td>86%</td>
</tr>
<tr>
<td>Maximum</td>
<td>86%</td>
<td>90%</td>
</tr>
<tr>
<td>Minimum</td>
<td>85%</td>
<td>91%</td>
</tr>
<tr>
<td>Concatenate</td>
<td>88%</td>
<td>91%</td>
</tr>
</tbody>
</table>

In the case of ANN the maximum f-score achieved is 88% using ‘concatenation’ as well as ‘add’ methods. However, in the case of CNN the maximum f-score achieved is 91% through ‘concatenation’ as well as ‘minimum’ methods. Therefore the ‘concatenation’ method appears to be the best single choice for the utilised dataset. Table III shows the f-scores obtained from single modalities i.e. FS and AIS in comparison with the ‘concatenation’ based multi-modal system.
TABLE III. OVERALL F-SCORES OF GAIT CLASSIFICATION FOR SINGLE AND MULTI-MODALITY SYSTEM

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>FS</th>
<th>AIS</th>
<th>Multi-Modal</th>
<th>Execution Time (min:sec)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ANN</td>
<td>85%</td>
<td>74%</td>
<td>88%</td>
<td>00:31.10</td>
</tr>
<tr>
<td>CNN</td>
<td>76%</td>
<td>68%</td>
<td>91%</td>
<td>00:25.57</td>
</tr>
</tbody>
</table>

From table III, it can be seen that multi-modality fusion by concatenation is effective and gives higher f-scores as compared to individual f-scores obtained from single modalities i.e., FS and AIS. CNN has higher f-score of 91% and consumes much lesser time to train the model than ANN with an f-score of 88%. ANN and CNN based automatic feature extraction has been used to differentiate four gait activities such as normal walk, fast walk, walking while subtracting 7 and typing on the mobile phone. A classwise f-score comparison of four gait activities for single and fused multi-modality is shown in table IV. Using sensors fusion all four gait activities have higher f-scores than single modality scores using ANN or CNN.

TABLE IV. CLASSWISE F-SCORES OF GAIT CLASSIFICATION FOR SINGLE AND MULTI-MODAL SYSTEMS

V. CONCLUSION

It is evident that CNN outperforms ANN in terms of lesser classification time and higher accuracies. Feeding input data as 2-D arrays to CNN architecture and fully connected merging strategy is able to extract gait features in a robust manner. Therefore our multi-modality fusion yields the best results positioning itself as a candidate to set a new state-of-the-art in the case under study. This can be taken up further by expanding the dataset with a larger number of subjects, varying the number/position of the AIS sensors, as well as defining additional manners of walking.

REFERENCES


