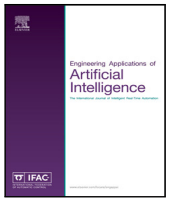
Engineering Applications of Artificial Intelligence 125 (2023) 106682

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

journal homepage: www.elsevier.com/locate/engappai



M-GaitFormer: Mobile biometric gait verification using Transformers

Paula Delgado-Santos [a,b,*], Ruben Tolosana [b], Richard Guest [a], Ruben Vera-Rodriguez [b], Julian Fierrez [b]

[a] School of Engineering, University of Kent, United Kingdom
[b] Biometrics and Data Pattern Analytics Lab, Universidad Autonoma de Madrid, Spain

ARTICLE INFO

Keywords:
Biometrics
Behavioural biometrics
Gait verification
Mobile devices
Deep learning
Transformers

ABSTRACT

Mobile devices such as smartphones and smartwatches are part of our everyday life, acquiring large amount of personal information that needs to be properly secured. Among the different authentication techniques, behavioural biometrics has become a very popular method as it allows authentication in a non-intrusive and continuous way. This study proposes M-GaitFormer, a novel mobile biometric gait verification system based on Transformer architectures. This biometric system only considers the accelerometer and gyroscope data acquired by the mobile device. A complete analysis of the proposed M-GaitFormer is carried out using the popular available databases whuGAIT and OU-ISIR. M-GaitFormer achieves Equal Error Rate (EER) values of 3.42% and 2.90% on whuGAIT and OU-ISIR, respectively, outperforming other state-of-the-art approaches based on popular Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs).

1. Introduction

The deployment of mobile devices in our society has become ubiquitous, interacting with them at anytime, anywhere. As a result, our mobile devices have become data hubs, storing all our personal information such as diary and financial information (Delgado-Santos et al., 2022a; Niknejad et al., 2020). As a consequence, it is crucial to protect the access to them using robust and user-friendly techniques (Melzi et al., 2022). One of the most popular authentication techniques is based on behavioural biometrics. Behavioural traits such as gait (Hadjkacem et al., 2020), keystroke dynamics (Stragapede et al., 2022), touch gestures (Acien et al., 2021), and handwritten signature (Tolosana et al., 2021) have recently shown impressive results in different security scenarios.

Gait biometrics is based on several patterns of the subject such as the arm swing amplitude, step frequency, and gait length (Wang et al., 2003). This trait can be considered in different authentication scenarios such as surveillance cameras capturing the data with visual sensors (Singh et al., 2018), or mobile devices where data are acquired using inertial sensors (i.e., accelerometer and gyroscope) (Marsico and Mecca, 2019). In this paper we focus on the latter, mobile gait biometrics through inertial sensors, considering the challenging verification scenario. This is an interesting scenario as: (i) subjects do not have to perform any specific task (i.e., interact with the mobile device), and (ii) subjects can be authenticated in a continuous non-intrusive way (Anon, 2018; Patel et al., 2016).

Recently, most biometric gait verification systems proposed in the literature have been based on popular Deep Learning (DL) architectures, such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs). However, some constraints remain part of the problem (Sepas-Moghaddam and Etemad, 2022; Filipi Gonçalves dos Santos et al., 2022). The main limitations are: (i) sequential computation; (ii) squeeze the previous data seen; and (iii) vanishing gradients during back propagation (Vaswani et al., 2017; Hutchins et al., 2022). In order to overcome these limitations, Transformer architectures have been proposed in recent years in several fields (e.g., machine translation, computer vision, time-series forecasting, etc.) (Tay et al., 2022). Their main advantages in comparison with traditional deep learning architectures are: (i) they are feed-forward models, processing all sequences in parallel; (ii) they apply self-attention mechanisms, operating over long sequences; (iii) they process all the sequences efficiently even in one batch; and (iv) they attend all the previous data simultaneously without the need to summarise them (Vaswani et al., 2017).

This study proposes M-GaitFormer, a novel mobile biometric gait verification system based on Transformers in order to overcome previous limitations and improve the state of the art. Fig. 1 provides a graphical representation of the proposed approach, including both learning and inference stages. First, we consider as input the accelerometer and gyroscope time sequences captured by the mobile device. After that, we can observe two main modules: (i) a feature extractor module based on an adaptation of our recently proposed Transformer architecture (Delgado-Santos et al., 2022b), which is trained on a learning stage

* Corresponding author at: School of Engineering, University of Kent, United Kingdom.
E-mail address: p.delgado-de-santos@kent.ac.uk (P. Delgado-Santos).

https://doi.org/10.1016/j.engappai.2023.106682
Received 7 November 2022; Received in revised form 24 April 2023; Accepted 18 June 2023
Available online xxxx

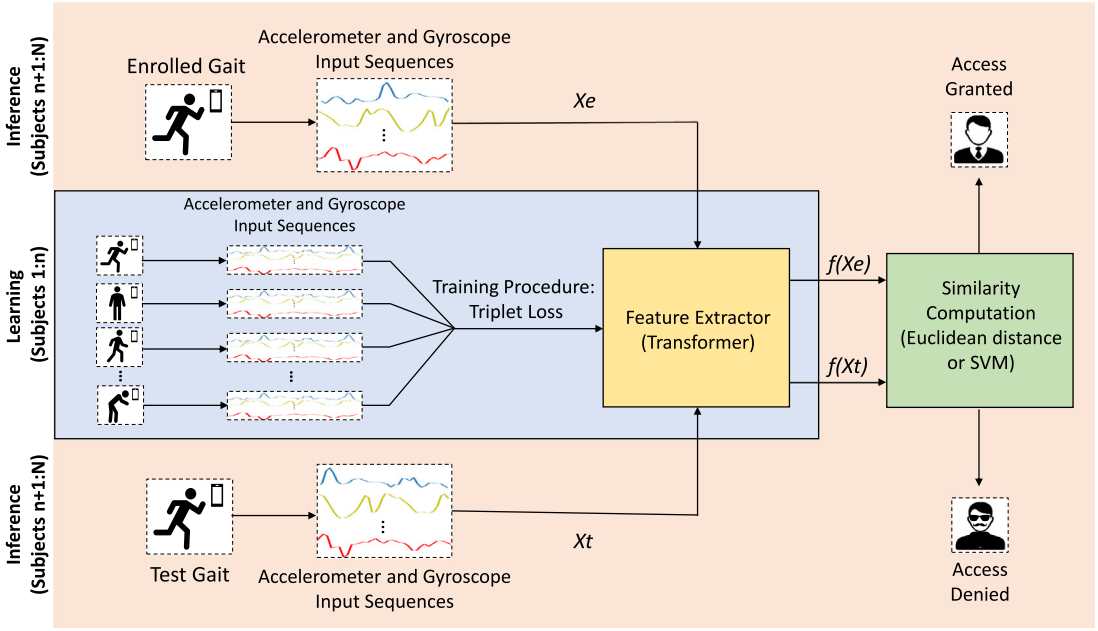


Fig. 1. Graphical representation of M-GaitFormer, the proposed mobile biometric gait verification system based on Transformers. N: total number of subjects; X_e: Enrolment input sequences; X_t: Test input sequences; $f(X_e)$: Enrolment feature vector; $f(X_t)$: Test feature vector.

using a development dataset, and (ii) a similarity computation module based on Euclidean distance or Support Vector Machine (SVM), which provides the final verification score of the comparison (inference stage).

The main contributions of this article are:

- An in-depth analysis of state-of-the-art deep learning approaches for mobile biometric gait verification, detailing key public databases and results.
- Proposal of M-GaitFormer, a novel mobile biometric gait verification system based on Transformers. In particular, we consider an adaptation of our recent Transformer architecture for feature extraction (Delgado-Santos et al., 2022b). Fig. 2 provides a graphical representation of the proposed Transformer architecture. To the best of our knowledge, this is the first study that explores the potential of Transformers for the task of mobile biometric gait verification. Finally, for the biometric similarity computation module, we explore different configurations such as the popular Euclidean distance or different SVM configurations, i.e., One-Class SVM (OC-SVM), and Binary SVM (B-SVM).
- A complete analysis of the proposed M-GaitFormer is carried out using the popular available databases whuGAIT (Zou et al., 2020) and OU-ISIR (Iwama et al., 2012; Ngo et al., 2014). M-GaitFormer achieves Equal Error Rate (EER) values of 3.42% and 2.90% on whuGAIT and OU-ISIR, respectively, outperforming other state-of-the-art approaches based on traditional CNNs and RNNs.

The remainder of the paper is organised as follows: Section 2 summarises the related work on gait verification on mobile devices. Section 3 describes the proposed M-GaitFormer system with the Transformer-based feature extractor and the different similarity computation configurations (i.e., Euclidean distance, OC-SVM, and B-SVM). Section 4 presents a detailed description of the popular whuGAIT (Zou et al., 2020) and OU-ISIR (Iwama et al., 2012; Ngo et al., 2014) databases whereas Section 5 describes the experimental setup. Section 6 contains the experimental results of the proposed M-GaitFormer, and the comparison with the state of the art. Finally, Section 7 draws conclusions and future research lines.

2. Related work

Biometric gait verification allows a corroboration as to whether a subject is who he/she claims to be based on the gait patterns. Due to the high deployment of mobile devices and their accurate sensors, mobile gait biometrics is becoming more and more popular nowadays (Marsico and Mecca, 2019). This is usually based on the background sensors data captured by mobile devices, in particular the accelerometer and gyroscope (Acien et al., 2020).

One of the most popular public databases is OU-ISIR, presented by (Ngo et al., 2014). This database comprises 744 subjects and considers the gyroscope and accelerometer data of the mobile device. In addition to the database, the authors implemented a Dynamic Time Warping (DTW) scheme with verification purposes achieving an EER of 13.5%. However, the experimental protocol considered in that paper might not be very realistic for operational conditions as the same subjects were considered for both training and testing the system. Despite this, other studies in the literature have followed similar experimental protocols, achieving EER values between 5% and 10% through handcrafted machine learning techniques, for example: Hoang et al. presented an approach based on Gait Dynamic Images (GDIs) and *i*-vector (Zhong and Deng, 2014), Sprager and Juric proposed a feature extractor based on Higher-Order Statistics (HOS) (Sprager and Juric, 2015), and Subramanian and Sarkar introduced an approach based on the Kabsch alignment (Subramanian and Sarkar, 2019).

In recent years, researchers have turned to DL techniques to extract more discriminative features. Delgado-Escaño et al. presented an approach based on CNNs (Delgado-Escaño et al., 2018). Data were divided into two branches, one for each sensor (accelerometer and gyroscope). CNN features extracted from each branch were concatenated into a common feature vector, and the Euclidean distance was finally computed in order to obtain the similarity score between enrolment and test samples. An EER of 1.1% was obtained over the OU-ISIR database, considering the same subjects for training and testing the gait verification systems. A similar approach was also presented by Tran and Choi in (Tran and Choi, 2020), considering CNNs as the feature extractor and OC-SVM for the similarity computation, achieving 4.49% EER in similar experimental protocol conditions.

As we have described, previous approaches in the literature tend to use the same subjects to train and test their gait verification systems.

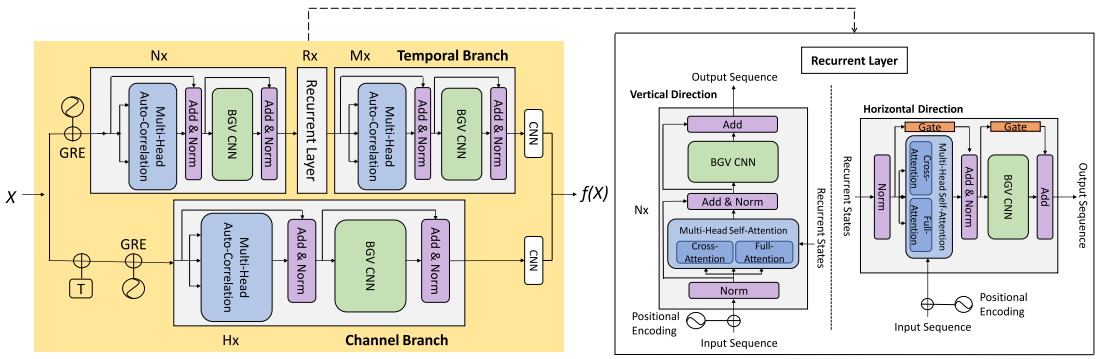


Fig. 2. Graphical representation of the Transformer-based Feature Extractor. X: Input sequences; $f(X)$: Feature vector; GRE: Gaussian Range Encoding; T: Transposition; Nx, Rx, Mx, Hx: Number of layers of each type; BGV CNN: Biometric Gait Verification CNN.

However, this scenario may not be realistic representing operational conditions, as the CNN feature extractor needs to be trained every time new subjects are available. Following this observation, Nguyen et al. presented an approach based on CNNs and SVM (Nguyen et al., 2017), considering different sets of subjects for training and testing. They evaluated their proposed approach with the OU-ISIR database achieving an EER of 10.43%, much higher compared with the case of using the same subjects for training and testing. Fernandez et al. incorporated in (Fernandez-Lopez et al., 2019) RNN-LSTMs to extract features and then compared them with Euclidean distance. A final EER value of 7.55% was obtained.

Apart from the popular OU-ISIR database, Zou et al. presented in (Zou et al., 2020) the whuGAIT database based on mobile data from gyroscope and accelerometer sensors. They also presented a standard experimental protocol considering different subjects for training and testing the systems. In addition, they proposed an approach based on two CNN-branches (one for each sensor) that are concatenated and introduced into a single RNN-branch. A final 6.5% EER was obtained using an OC-SVM classifier for the final similarity computation. Recently, Tran et al. presented in (Tran et al., 2021) a new approach based on a multi-CNN and multi-RNN system. In both databases, different subjects were used for training and testing the systems, achieving EER values of 4.52% and 3.36% for the whuGAIT and OU-ISIR databases, respectively.

To summarise this section, we can see that most approaches in the literature are based on popular CNN and RNN architectures, despite the limitations described in Section 1. In addition, it is important to remark the high variability of the performance results depending on the specific experimental protocol, i.e., training and testing the system with the same subjects or not. These aspects motivate the proposal of novel biometric gait verification systems based on Transformers and the evaluation of them under realistic experimental conditions.

3. Proposed system: M-GaitFormer

Fig. 1 shows the general diagram of the proposed gait verification system for mobile scenarios. First, the three-axis accelerometer and gyroscope time sequences, X, captured by the mobile device are considered as input. Each time sequence contains L samples. Then, the proposed system consists of two modules: (i) a feature extractor module based on a Transformer architecture, which is trained on a learning stage using a development dataset of subjects; and (ii) a similarity computation module, which provides the final verification score of the comparison (inference stage). We describe next each module in detail.

3.1. Feature extractor

Fig. 2 provides a graphical representation of the feature extractor based on a Transformer architecture. The Vanilla Transformer (Vaswani et al., 2017), which was introduced for machine translation showing impressive results. This architecture needs some adaptations to be used in different domains, such as time sequences. Several researchers have presented diverse modifications in order to improve this original Transformer architecture, reducing complexity, including periodicity-based dependencies or time-depending encoding (Tay et al., 2022). We provide next a description of our proposed Transformer architecture.

Following the idea presented in (Li et al., 2021), our proposed Transformer contains two branches: (i) Temporal Branch, which extracts information related to time from the input sequences (temporal-over-channel features); and (ii) Channel Branch, which extracts information related to space from the input sequences (channel-over-temporal features).

Analysing first the Temporal Branch, the input time sequences X are modelled using a Gaussian Range Encoding (GRE) to preserve temporal information. The Probability Density Functions (PDFs) of the G Gaussian distributions are L1-normalised and combined into a vector. In addition, the encoding can be applied over different ranges at each single point, acquiring a more complete context position. Consequently, the output of the GRE, X', is the weighted multiplication of the normalised PDFs, β, over the different ranges of the input time sequences X:

$$X' = \beta + X \tag{1}$$

After the GRE, the Temporal Branch contains a sequential stack of three types of layers: N, R, and M, being $f_T(X)$ the output feature vector. In particular, N and M are identical layers, comprising the following two sequential sub-layers: (i) a multi-head Auto-Correlation mechanism, proposed in (Wu et al., 2021), and (ii) a multi-scale Biometric Gait Verification (BGV) CNN specifically designed for the task. The multi-head Auto-Correlation sub-layer connects different sequences among estimated periods. A Time Delay Aggregation (TDA) block is introduced to align similar sub-sequences under different time delays, $\tau_1, \dots, \tau_k$. Specifically, the Auto-Correlation mechanism can be defined as:

$$\tau_1, \dots, \tau_k = \underset{\tau \in (1,\dots,L)}{\text{argTopK}}(R_{Q,K}(\tau))$$

$$\hat{R}_{Q,K}(\tau_1), \dots, \hat{R}_{Q,K}(\tau_k) = \text{SoftMax}(R_{Q,K}(\tau_1), \dots, R_{Q,K}(\tau_k))$$

$$\text{Auto-Correlation}(Q, K, V) = \sum_{i=1}^{k} \text{Roll}(V, \tau_i)\hat{R}_{Q,K}(\tau_i) \tag{2}$$

where argTopK is the output of TopK correlations of the Fast Fourier Transforms (FFTs) along L; $R_{Q,K}$ is the Auto-Correlation between the Q queries and the K keys; and $\text{Roll}(V, \tau_i)$ scroll X with a τ time delay over the V values, re-introducing the elements moved beyond the first position to the last one.

The output of the sub-layer is the concatenation of applying Auto-Correlation to F independent heads. The multi-scale BGV CNN sub-layer comprises a CNN with ReLU activations and unique kernels

for the different scales. Following each sub-layer, a residual connection and a layer normalisation are included (*Add & Norm* in Fig. 2).

Finally, R layers consist of the Block-Recurrent Transformer architecture presented in (Hutchins et al., 2022). Fig. 2 (right) provides a graphical representation of the recurrent layers in detail. First, the time sequences are modelled using a positional encoding. After that, a recurrent form of attention over time sequences is introduced, which comprises two directions: vertical and horizontal. The vertical direction consists of two sub-layers: (i) a multi-head Self-Attention mechanism and (ii) a multi-scale BGV CNN network. The multi-head Self-Attention mechanism contains two different attentions: (i) Full-Attention to the time sequences to extract the values V and the matching keys K (Vaswani et al., 2017); and (ii) Cross-Attention to the recurrent states (initialised to 0) to extract the Q queries. Consequently, a BGV CNN network with ReLU activations and unique kernels for the different scales. Previous to each sub-layer, a layer normalisation is included, while after each sub-layer a residual connection is added (*Add & Norm* in Fig. 2). Regarding the horizontal direction, it also contains the same two sub-layers. However, in contrast to the vertical direction, the multi-head Self-Attention mechanism applies: (i) Cross-Attention to the time sequences to extract the queries Q; and (ii) Full-Attention to the recurrent states to obtain the keys K and values V. In this direction, residual connections are replaced by forget gates which modify the recurrent states. In both directions the attention mechanism is replicated in F independent heads.

Analysing the Channel Branch, first the input sequences X are transposed and also modelled by a GRE, similar to the temporal branch. After that, the Channel Branch contains a stack of H identical layers. Each layer comprises two sub-layers: (i) a multi-head Auto-Correlation mechanism applied to F independent heads, and (ii) multi-scale BGV CNN. Following each sub-layer, a residual connection and a layer normalisation are included (*Add & Norm* in Fig. 2). At last, the features are merged into an output feature vector $f_C(X)$.

Finally, after the Temporal and Channel Branches we include a CNN, similar to (Li et al., 2021). Subsequently, all extracted features are concatenated, $f(X)$, and introduced into the similarity computation module:

$$f(X) = [\text{CNN}(f_T(X)); \text{CNN}(f_C(X))] \tag{3}$$

where $f_T(X)$ and $f_C(X)$ are the extracted features from the Temporal and Channel Branches, respectively.

3.2. Similarity computation

As described in Fig. 1, the similarity computation module receives as an input the features of the enrolled and test gait, $f(X_e)$ and $f(X_t)$, to obtain the final similarity score. Three different configurations are studied: (i) Euclidean distance, (ii) OC-SVM, and (iii) B-SVM.

3.2.1. Euclidean distance

This is a simple but very popular approach in biometrics based on the distance between the feature vectors $f(X_e)$ and $f(X_t)$:

$$d(X_e, X_t) = \left\| f(X_e) - f(X_t) \right\| \tag{4}$$

3.2.2. One-class support vector machine (OC-SVM)

This comprises of training a single specific SVM classifier per subject. In this particular configuration (one-class), only the enrolment samples of the subject are considered to train the SVM.

3.2.3. Binary support vector machine (B-SVM)

Similar to the OC-SVM, the B-SVM comprises of training one specific SVM classifier per subject. The main difference is that for each subject, one classifier is trained using both enrolment samples of the subject and also gait samples of other subjects (from a development dataset), acting as impostors.

4. Databases

Two of the most popular public databases in mobile gait verification are considered in the experimental framework of this study: (i) whuGAIT (Zou et al., 2020), and (ii) OU-ISIR (Ngo et al., 2014). These databases are used as they provide realistic and standard experimental protocols, considering specific development and evaluation datasets with different subjects, making possible a fair comparison with the state of the art (Fernandez-Lopez et al., 2019; Tran et al., 2021).

4.1. WhuGAIT database

The whuGAIT database was presented by Zou et al. in 2020 (Zou et al., 2020). This database includes accelerometer and gyroscope data collected with Samsung, Xiaomi, and Huawei smartphones in an unconstrained scenario. There is no record of when, how, and where smartphones have been used. In total, data were acquired from 118 subjects in both walking and non-walking scenarios with a sampling frequency of 50 Hz.

4.2. OU-ISIR database

The OU-ISIR database was introduced by Ngo et al. in 2014 (Ngo et al., 2014), being the largest public biometric gait database to date. This database comprises 744 total subjects, including accelerometer and gyroscope data captured by three Inertial Measurement Units (IMUs) and a smartphone Motorola ME860 around the waist of the subject. A single session was recorded per subject where four activities were performed (two flat walking, slope-up walking, and slope-down walking) with a sampling frequency of 100 Hz. The database was divided into 2 sub-sets: (i) data from 744 subjects obtained using the IMU sensor located in the middle of the subject's back waist (two flat walking); and (ii) data from 408 subjects recorded by the three IMUs and the smartphone (two flat walking, slope-up walking, and slope-down walking).

5. Experimental setup

This section provides the details of the experimental framework of the study. First, we describe in Section 5.1 the system configuration of the proposed M-GaitFormer. Then, Section 5.2 presents the standard experimental protocol considered for whuGAIT and OU-ISIR databases as a verification task.

5.1. M-GaitFormer: System details

Regarding the Transformer-based feature extractor, the GRE contains $G = 20$ Gaussian distributions. The Temporal Branch contains $N = 9$, $R = 1$, and $M = 2$ layers, and $F = 8$ independent heads for each layer whereas the Channel Branch comprises $H = 1$ layer and $F = 6$ independent heads for each layer. In both branches, the BGV CNN contains 3 convolutional layers with L units each, ReLU activation functions, and kernel sizes 1, 3, and 5, respectively, followed by dropout layers with a rate of 0.1. Finally, after the Temporal and Channel Branches we consider 2 convolutional layers with L units each, ReLU activation functions, and kernel sizes of 512 and 256, respectively.

The Transformer-based feature extractor is trained with a triplet loss function, using Euclidean distance with a margin $\alpha = 1.0$. Adam optimiser is considered with a learning rate of 0.001. It is trained using a stop condition: if the feature extractor does not achieve better results in the validation dataset during 15 epochs, the training stops. Regarding the similarity computation module, the OC-SVM has an *RBF* kernel and $\gamma = 1.0$ while B-SVM has an *RBF* kernel and $\gamma = 0.5$. Experiments are implemented in `PyTorch`.

Table 1
Results of our proposed M-GaitFormer in terms of EER (%) for the whuGait and OU-ISIR evaluation datasets and for the different similarity computation configurations considered: Euclidean distance, One-Class SVM (OC-SVM), and Binary SVM (B-SVM). In addition, for completeness, we include: (i) the results achieved by the Vanilla Transformer (Vaswani et al., 2017), and (ii) the contributions in the performance of each of the branches considered in M-GaitFormer.

Method		Databases	
Feature extractor	Similarity computation	whuGAIT	OU-ISIR
Vanilla Transformer (Vaswani et al., 2017)	Euclidean distance	14.67	10.17
	OC-SVM	10.15	6.55
	B-SVM	5.82	5.30
M-GaitFormer (Temporal Branch w/o Recurrent Layer)	Euclidean distance	13.03	9.15
	OC-SVM	8.02	6.89
	B-SVM	4.02	5.13
M-GaitFormer (Temporal Branch w/ Recurrent Layer)	Euclidean distance	11.97	8.59
	OC-SVM	7.70	6.78
	B-SVM	4.11	4.79
M-GaitFormer (Channel Branch)	Euclidean distance	13.97	9.80
	OC-SVM	7.38	7.05
	B-SVM	4.25	4.15
M-GaitFormer (Temporal + Channel Branches)	Euclidean distance	9.62	5.69
	OC-SVM	4.54	3.73
	B-SVM	**3.42**	**2.90**

5.2. Experimental protocol

We describe next the details of the experimental protocol considered for each database and stage (learning and inference):

- WhuGAIT: We follow the experimental protocol proposed by Tran et al. in (Tran and Choi, 2020). From the 118 total subjects, 98 are used in the learning stage for training the feature extractor while the remaining 20 unseen subjects are only considered for the final evaluation (inference stage). Regarding the learning stage, we build triplets using the 98 subjects of the development dataset. Each triplet comprises two genuine samples of the same subject (enrolment and genuine test), and a third one from a different subject (impostor test). The genuine and impostor test samples included in the triplets are selected randomly with a uniform distribution. Considering all possible triplets, a total of 284,030 triplets are included for the feature extractor training.
- OU-ISIR: We follow the same experimental protocol presented in (Fernandez-Lopez et al., 2019) and (Tran et al., 2021). From the 744 total subjects, 520 are used in the learning stage for training the feature extractor while the remaining 224 unseen subjects are part of the final evaluation (inference stage). Concerning the learning stage, we build triplets using the 520 subjects of the development dataset, following the same approach described for the WhuGAIT database. A total of 229,543 triplets are considered for training the feature extractor.

Regarding the inference stage, we consider in both WhuGAIT and OU-ISIR databases the same experimental protocol. For each unseen subject of the evaluation dataset, we follow the same experimental protocol presented in (Tran and Choi, 2020; Fernandez-Lopez et al., 2019) and (Tran et al., 2021) 50% of the samples of the subject selected randomly are used as enrolment, while the remaining 50% are considered for testing in order to obtain the genuine scores. Impostor scores are obtained comparing the enrolment samples of the subject with samples of the remaining subjects of the evaluation dataset (same number of genuine and impostor comparisons). Depending on the similarity computation approach considered (i.e., Euclidean distance, OC-SVM, B-SVM), the final score is calculated differently. For the Euclidean distance, the final score is the average of the scores obtained when comparing one test sample (genuine/impostor) with each enrolment sample. For the SVM approaches, the final score is obtained when comparing one test sample (genuine/impostor) with the specific SVM model created with all enrolment samples of that subject.

6. Experimental results

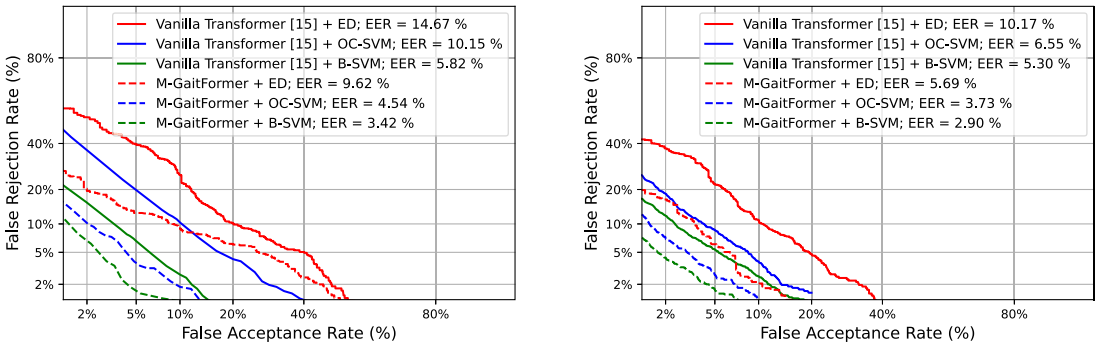
Section 6.1 provides an analysis of the proposed M-GaitFormer and each of its modules, for both whuGAIT and OU-ISIR databases, and also for the different similarity computation configurations (i.e., Euclidean distance, OC-SVM, and B-SVM). Finally, we compare in Section 6.2 our proposed M-GaitFormer with the state of the art using the same experimental protocol.

6.1. M-GaitFormer results

Table 1 shows the results of our proposed M-GaitFormer in terms of EER (%) for the whuGAIT and OU-ISIR evaluation datasets and for the different similarity computation configurations considered: Euclidean distance, One-Class SVM (OC-SVM), and Binary SVM (B-SVM). In addition, for completeness, we include: (i) the results achieved by the Vanilla Transformer (Vaswani et al., 2017), and (ii) the contributions in the performance of each of the branches considered in M-GaitFormer.

First, we analyse the impact in the system performance of each of the branches considered in the proposed M-GaitFormer. To provide a better understanding of the results, we focus now on the Euclidean distance configuration, as the proposed Transformer is used as feature extractor. The Temporal Branch (without the recurrent layer) achieves values of 13.03% and 9.15% EER for the whuGAIT and OU-ISIR databases, respectively. These results are further improved if we include the recurrent layer in the Temporal Branch, i.e., 11.97% and 8.59% EER for the whuGAIT and OU-ISIR databases, respectively. On the other hand, we can see that the Channel Branch is also able to extract discriminative features for the task, achieving EER values of 13.97% and 9.80% for the whuGAIT and OU-ISIR databases. Finally, we can see in Table 1 how the combination of both Temporal and Channel Branches achieves the best results in both whuGAIT (9.62% EER) and OU-ISIR databases (5.69% EER). These results prove the potential of our proposed Temporal and Channel Branches for the feature extraction.

Analysing the impact of the similarity computation configuration, we can see that in general, the Euclidean distance provides worse results compared to the case of training classifiers such as SVM. For example, focusing on the M-GaitFormer (Temporal + Channel Branches) in Table 1, the Euclidean distance achieves values of 9.62% and 5.69% EER for the whuGAIT and OU-ISIR databases, respectively. These results are further improved when considering the B-SVM (3.42% and 2.90% EER, respectively), with relative improvements of 64.45% and 49.03% EER. These results evidence the importance of using classifiers



(a) whuGAIT **(b)** OU-ISIR

Fig. 3. DET curves and EER (%) results on the (a) whuGAIT and (b) OU-ISIR evaluation datasets for the Vanilla Transformer (Vaswani et al., 2017) and the proposed M-GaitFormer in the three similarity computation configurations considered: (i) Euclidean distance (ED), (ii) One-Class SVM (OC-SVM), and (iii) Binary SVM (B-SVM).

such as SVM to better adapt the features extracted by the Transformer to each specific subject. This is in accordance with related works that have shown subject-adaptation (Fierrez et al., 2018a) to be very useful in behavioural biometrics (Fierrez-Aguilar et al., 2005).

Finally, we compare in Table 1 the results achieved by our proposed M-GaitFormer (Temporal + Channel Branches) with the original Vanilla Transformer (Vaswani et al., 2017). In addition, for completeness, we include in Fig. 3 the Detection Error Trade-Off (DET) curves of the proposed M-GaitFormer and the Vanilla Transformer (Vaswani et al., 2017) for both whuGAIT and OU-ISIR databases. In general, we can observe that M-GaitFormer outperforms the Vanilla Transformer in all configurations (Euclidean distance, OC-SVM, and B-SVM), proving the potential of the proposed method. For the Euclidean distance configuration, M-GaitFormer achieves relative improvements of 34.42% and 44.05% EER for the whuGAIT and OU-ISIR databases whereas for the B-SVM configuration the relative improvements are 41.24% and 45.28% EER, respectively. Finally, the best results achieved by the proposed M-GaitFormer are 3.42% and 2.90% EER for the whuGAIT and OU-ISIR databases.

6.2. Comparison with the state of the art

Table 2 provides a comparison in terms of EER (%) of the proposed M-GaitFormer with others state-of-the-art approaches in the literature for both whuGAIT and OU-ISIR evaluation datasets. Note that in some cases, indicated in Table 2 with the symbol *, the studies do not use the standard experimental setup considered in the literature, therefore the results must be interpreted carefully. Despite of that, it is patent that the proposed M-GaitFormer achieves state-of-the-art results in both whuGAIT and OU-ISIR databases.

Analysing the results on the whuGAIT database, our proposed M-GaitFormer achieves an EER of 3.42%, a relative EER improvement of 54.40% and 41.24% compared with systems based on LSTM architectures (Zou et al., 2020; Tran et al., 2021). Furthermore, M-GaitFormer outperforms previous approaches in the literature based on CNN & LSTM (Zou et al., 2020; Tran et al., 2021) with relative improvements of 47.39% and 24.34% EER.

A similar trend can be observed for the OU-ISIR database. Our proposed M-GaitFormer achieves an EER of 2.90%. This is a relative EER improvement of 72.20% and 35.41% compared with traditional CNN architectures (Nguyen et al., 2017; Tran and Choi, 2020). In addition, M-GaitFormer achieves a relative EER improvement of 61.59% and 56.26% in comparison with LSTM architectures (Fernandez-Lopez et al., 2019; Tran et al., 2021). Finally, M-GaitFormer reaches a relative improvement of 13.69% EER in comparison with the CNN & LSTM architecture presented in (Tran et al., 2021).

This comparison with the state of the art proves the potential of our proposed M-GaitFormer architecture for the task of mobile biometric gait verification, outperforming previous approaches based on CNN, LSTM, or a combination. Some of the advances that have been achieved by applying Transformers are: (i) application of Auto-Correlation and attention mechanisms, which allow operating over long sequences; (ii) operation on all previous samples of the time sequence at the same time, without the need to summarise; (iii) extraction of features from two different perspectives (Channel and Temporal Branches), obtaining more discriminative information; and (iv) inclusion of a GRE together with the Auto-Correlation and the Block-Recurrent attention to extract features over the entire time sequence instead of single points, considering important aspects in time sequences such as the samples distribution over the time.

Table 2
Comparison of the proposed M-GaitFormer system with state-of-the-art approaches in mobile biometric gait verification in terms of EER (%) for the whuGait and OU-ISIR evaluation datasets (Tran and Choi, 2020; Ngo et al., 2014). Note that the symbol * indicates those studies that do not use the standard experimental setup considered in the literature.

Study	Method	Database	
		whuGAIT	OU-ISIR
(Nguyen et al., 2017)	CNN	–	10.43
(Fernandez-Lopez et al., 2019)	LSTM	–	7.55
(Subramanian and Sarkar, 2019)	Kabsch alignment	–	> 6.00
(Tran and Choi, 2020)	CNN	–	4.49
(Zou et al., 2020)	LSTM CNN & LSTM	7.50 6.50	– –
(Tran et al., 2021)	LSTM CNN & LSTM	5.82 4.52	6.63 3.36
M-GaitFormer	**Transformer**	**3.42**	**2.90**

7. Conclusions

This article has proposed M-GaitFormer, a novel mobile gait verification system based on Transformers. To the best of our knowledge, this is the first time that Transformers have been applied to the mobile biometric gait verification task.

M-GaitFormer consists of two modules (i) a Transformer-based feature extractor trained on a learning stage using a development dataset; and (ii) a similarity computation module (Euclidean distance or SVM) which gives the final verification score of the comparison (inference stage). Our experiments are conducted with two popular public databases in mobile gait verification (whuGAIT and OU-ISIR), considering the same experimental protocol proposed in the state of the art. Due to the progress shown by the Transformers with respect to

CNNs and RNNs, our proposed M-GaitFormer system outperforms the state of the art achieving results of 3.42% and 2.92% EER on whuGAIT and OU-ISIR databases, respectively.

Future work will explore and analyse Transformers in other behavioural biometric modalities such as swipe (Fierrez et al., 2018b) and handwritten signature (Tolosana et al., 2022), and also implement privacy-preserving techniques for sensitive data in biometric scenarios (Melzi et al., 2022; Delgado-Santos et al., 2022c).

CRediT authorship contribution statement

Paula Delgado-Santos: Conceptualization, Investigation, Writing – original draft, Writing – review & editing, Visualization, Funding acquisition. **Ruben Tolosana:** Conceptualization, Investigation, Writing – original draft, Writing – review & editing, Visualization, Funding acquisition. **Richard Guest:** Conceptualization, Writing – review & editing, Visualization, Funding acquisition. **Ruben Vera-Rodriguez:** Conceptualization, Writing – review & editing, Visualization, Funding acquisition. **Julian Fierrez:** Conceptualization, Writing – review & editing, Visualization, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No data was used for the research described in the article.

Acknowledgements

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 860315. With support also from projects INTER-ACTION (PID2021-126521OB-I00 MICINN/FEDER), HumanCAIC (TED2021-131787B-I00 MICINN), and Comunidad de Madrid (ELLIS Unit Madrid).

References

Acien, A., Morales, A., Fierrez, J., Vera-Rodriguez, R., Delgado-Mohatar, O., 2021. BeCAPTCHA: Behavioral bot detection using touchscreen and mobile sensors benchmarked on HuMIdb. Eng. Appl. Artif. Intell. 98, 104058.

Acien, A., Morales, A., Vera-Rodriguez, R., Fierrez, J., 2020. Smartphone sensors for modeling human-computer interaction: General outlook and research datasets for user authentication. In: Proc. IEEE Annual Computers, Software, and Applications Conference.

Anon, 2018. ISO 9241-11:2018(en): Ergonomics of human-system interaction. Part 11: Usability: Definitions and Concepts.

Delgado-Escaño, R., Castro, F.M., Cózar, J.R., Marín-Jiménez, M.J., Guil, N., 2018. An end-to-end multi-task and fusion CNN for inertial-based gait recognition. IEEE Access 7, 1897–1908.

Delgado-Santos, P., Stragapede, G., Tolosana, R., Guest, R., Deravi, F., Vera-Rodriguez, R., 2022a. A survey of privacy vulnerabilities of mobile device sensors. ACM Comput. Surv. 54 (11).

Delgado-Santos, P., Tolosana, R., Guest, R., Deravi, F., Vera-Rodriguez, R., 2022b. Exploring transformers for behavioural biometrics: A case study in gait recognition. arXiv Preprint arXiv:2206.01441 (2022).

Delgado-Santos, P., Tolosana, R., Guest, R., Vera, R., Deravi, F., Morales, A., 2022c. GaitPrivacyON: Privacy-preserving mobile gait biometrics using unsupervised learning. Pattern Recognit. Lett. 161, 30–37.

Fernandez-Lopez, P., Liu-Jimenez, J., Kiyokawa, K., Wu, Y., Sanchez-Reillo, R., 2019. Recurrent neural network for inertial gait user recognition in smartphones. Sensors 19 (18), 1–16.

Fierrez, J., Morales, A., Vera-Rodriguez, R., Camacho, D., 2018a. Multiple classifiers in biometrics. Part 2: Trends and challenges. Inf. Fusion 44, 103–112.

Fierrez, J., Pozo, A., Martinez-Diaz, M., Galbally, J., Morales, A., 2018b. Benchmarking touchscreen biometrics for mobile authentication. IEEE Trans. Inf. Forensics Secur. 13 (11), 2720–2733.

Fierrez-Aguilar, J., Garcia-Romero, D., Ortega-Garcia, J., Gonzalez-Rodriguez, J., 2005. Bayesian adaptation for user-dependent multimodal biometric authentication. Pattern Recognit. 38 (8), 1317–1319.

Filipi Gonçalves dos Santos, C., Oliveira, D.d.S., A. Passos, L., Gonçalves Pires, R., Felipe Silva Santos, D., Pascotti Valem, L., P. Moreira, T., Cleison S. Santana, M., Roder, M., Paulo Papa, J., et al., 2022. Gait recognition based on deep learning: A survey. ACM Comput. Surv. 55 (2), 1–34.

Hadjkacem, B., Ayedi, W., Ayed, M.B., Alshaya, S.A., Abid, M., 2020. A novel gait-appearance-based multi-scale video covariance approach for pedestrian (Re)-identification. Eng. Appl. Artif. Intell. 91, 103566.

Hutchins, D., Schlag, I., Wu, Y., Dyer, E., Neyshabur, B., 2022. Block-recurrent transformers. arXiv Preprint arXiv:2203.07852 (2022).

Iwama, H., Okumura, M., Makihara, Y., Yagi, Y., 2012. The OU-ISIR gait database comprising the large population dataset and performance evaluation of gait recognition. IEEE Trans. Inf. Forensics Secur. 7 (5), 1511–1521.

Li, B., Cui, W., Wang, W., Zhang, L., Chen, Z., Wu, M., 2021. Two-stream convolution augmented transformer for human activity recognition. In: Proc. AAAI Conference on Artificial Intelligence.

Marsico, M.D., Mecca, A., 2019. A survey on gait recognition via wearable sensors. ACM Comput. Surv. 52 (4), 1–39.

Melzi, P., Rathgeb, C., Tolosana, R., Vera-Rodriguez, R., Busch, C., 2022. An overview of privacy-enhancing technologies in biometric recognition. arXiv preprint arXiv:2206.10465.

Ngo, T.T., Makihara, Y., Nagahara, H., Mukaigawa, Y., Yagi, Y., 2014. The largest inertial sensor-based gait database and performance evaluation of gait-based personal authentication. Pattern Recognit. 47 (1), 228–237.

Nguyen, K.-T., Vo-Tran, T.-L., Dinh, D.-T., Tran, M.-T., 2017. Gait recognition with multi-region size convolutional neural network for authentication with wearable sensors. In: Proc. International Conference on Future Data and Security Engineering.

Niknejad, N., Ismail, W.B., Mardani, A., Liao, H., Ghani, I., 2020. A comprehensive overview of smart wearables: The state of the art literature, recent advances, and future challenges. Eng. Appl. Artif. Intell. 90, 103529.

Patel, V.M., Chellappa, R., Chandra, D., Barbello, B., 2016. Continuous user authentication on mobile devices: Recent progress and remaining challenges. IEEE Signal Process. Mag. 33 (4), 49–61.

Sepas-Moghaddam, A., Etemad, A., 2022. Deep gait recognition: A survey. IEEE Trans. Pattern Anal. Mach. Intell. (2022).

Singh, J.P., Jain, S., Arora, S., Singh, U.P., 2018. Vision-based gait recognition: A survey. IEEE Access 6, 70497–70527.

Sprager, S., Juric, M.B., 2015. Inertial sensor-based gait recognition: A review. Sensors 15 (9), 1–39.

Stragapede, G., Vera-Rodriguez, R., Tolosana, R., Morales, A., 2022. BehavePassDB: Benchmarking mobile behavioral biometrics. Pattern Recognit..

Subramanian, R., Sarkar, S., 2019. Evaluation of algorithms for orientation invariant inertial gait matching. IEEE Trans. Inf. Forensics Secur. 14 (2), 304–318.

Tay, Y., Dehghani, M., Bahri, D., Metzler, D., 2022. Efficient transformers: A survey. ACM Comput. Surv. 55 (6), 1–28.

Tolosana, R., Delgado-Santos, P., Perez-Uribe, A., Vera-Rodriguez, R., Fierrez, J., Morales, A., 2021. DeepWriteSYN: On-line handwriting synthesis via deep short-term representations. In: Proc. AAAI Conference on Artificial Intelligence.

Tolosana, R., Vera-Rodriguez, R., Gonzalez-Garcia, C., Fierrez, J., Morales, A., Ortega-Garcia, J., Ruiz-Garcia, J.C., Romero-Tapiador, S., Rengifo, S., Caruana, M., et al., 2022. SVC-ongoing: Signature verification competition. Pattern Recognit. 127, 1–14.

Tran, L., Choi, D., 2020. Data augmentation for inertial sensor-based gait deep neural network. IEEE Access 8, 12364–12378.

Tran, L., Hoang, T., Nguyen, T., Kim, H., Choi, D., 2021. Multi-model long short-term memory network for gait recognition using window-based data segment. IEEE Access 9, 23826–23839.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I., 2017. Attention is all you need. In: Proc. Advances in Neural Information Processing Systems.

Wang, L., Tan, T., Ning, H., Hu, W., 2003. Silhouette analysis-based gait recognition for human identification. IEEE Trans. Pattern Anal. Mach. Intell. 25 (12), 1505–1518.

Wu, H., Xu, J., Wang, J., Long, M., 2021. Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting. In: Proc. Advances in Neural Information Processing Systems.

Zhong, Y., Deng, Y., 2014. Sensor orientation invariant mobile gait biometrics. In: Proc. IEEE International Joint Conference on Biometrics.

Zou, Q., Wang, Y., Wang, Q., Zhao, Y., Li, Q., 2020. Deep learning-based gait recognition using smartphones in the wild. IEEE Trans. Inf. Forensics Secur. 15, 3197–3212.