Pattern Recognition 143 (2023) 109798

Contents lists available at ScienceDirect

Pattern Recognition

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

Exploring transformers for behavioural biometrics: A case study in gait recognition

Paula Delgado-Santos[a,b,1,*], Ruben Tolosana[b,1], Richard Guest[a], Farzin Deravi[a], Ruben Vera-Rodriguez[b]

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

ARTICLE INFO

Article history:
Received 22 August 2022
Revised 26 June 2023
Accepted 2 July 2023
Available online 4 July 2023

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

ABSTRACT

Biometrics on mobile devices has attracted a lot of attention in recent years as it is considered a user-friendly authentication method. This interest has also been motivated by the success of Deep Learning (DL). Architectures based on Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) have established convenience for the task, improving the performance and robustness in comparison to traditional machine learning techniques. However, some aspects must still be revisited and improved. To the best of our knowledge, this is the first article that explores and proposes a novel gait biometric recognition systems based on Transformers, which currently obtain state-of-the-art performance in many applications. Several state-of-the-art architectures (Vanilla, Informer, Autoformer, Block-Recurrent Transformer, and THAT) are considered in the experimental framework. In addition, new Transformer configurations are proposed to further increase the performance. Experiments are carried out using the two popular public databases: whuGAIT and OU-ISIR. The results achieved prove the high ability of the proposed Transformer, outperforming state-of-the-art CNN and RNN architectures.

1. Introduction

Biometrics has become a relevant topic for security and authentication purposes [1]. Among the different biometric traits, gait behaviour biometrics has attracted considerable attention in recent years; for example, in surveillance scenarios where popular biometric traits such as face and fingerprint are hard or impossible to distinguish. Gait recognition uses the movement pattern of subjects by focusing on specific characteristics such as the arm swing amplitude, step frequency, and gait length [2]. Depending on the specific application scenario, gait pattern can be captured using visual sensors such as surveillance cameras [3] or inertial sensors such as the accelerometer and gyroscope included in wearable devices [4].

The popularity of gait recognition has also increased with the success of Deep Learning (DL) [5,6]. Architectures based on Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), such as Long Short-Term Memory (LSTM), have proven to be convenient for the task, improving performance and robustness compared to traditional machine learning techniques. However, these popular DL architectures still have several disadvantages that must be revisited and improved. The main drawbacks are [7,8]: *i)* Sequential computation, not allowing parallelisation within batches; *ii)* compression and condensation of the previous time samples, limiting the past information seen, and *iii)* vanishing gradients during back-propagation; the forget gate in a RNN removes a small portion of the previous state after each sample.

Transformers are more recently proposed DL architectures that have already garnered immense interest due to their effectiveness across a range of application domains such as language assessment, vision, and reinforcement learning [9]. Their main advantages compared with traditional CNN and RNN architectures are [7,8,10]: *i)* Transformers are feed-forward models that process all the sequences in parallel, therefore increasing efficiency; *ii)* They apply Self-Attention/Auto-Correlation mechanisms that allows them to operate in long sequences; *iii)* They can be trained efficiently in a single batch since all the sequence is included in every batch; and *iv)* They can attend to the whole sequence, instead of summarising all the previous temporal information. Recent studies

[*] Corresponding author at: School of Engineering, University of Kent, United Kingdom.
E-mail address: p.delgado-de-santos@kent.ac.uk (P. Delgado-Santos).
[1] Co-first author. Both authors contributed equally to this research.

https://doi.org/10.1016/j.patcog.2023.109798
0031-3203/

have successfully proved the advantages of Transformers for time-sequential data, outperforming traditional CNN and RNN architectures [11–13].

Several Transformer architectures have been recently proposed in the literature [9,14]. The original one, the Vanilla Transformer, was introduced in 2017 by Vaswani et al. [7]. It was based solely on Self-Attention mechanisms, dispensing with recurrence and convolutions layers entirely. Impressive results were achieved on the machine translation task, reducing also the training costs of the best models compared with the literature. Despite these improvements, the Vanilla Transformer has disadvantages for some applications based on time series: *i)* the computational complexity of the attention mechanism is quadratic $O(L^2)$ where L denotes the length of the input sequence; and *ii)* the total memory usage is $O(N \odot L^2)$ where N indicates the number of encoder/decoder layers, limiting the scalability of the model with long sequences. As a result, different Transformer architectures have recently emerged with the aim of addressing the shortcomings of the Vanilla Transformer, including: Informer [11], Autoformer [10], Block-Recurrent Transformer [8], and THAT [13], among others.

The present article intends to explore and propose novel behavioural biometric systems based on Transformers. The main contributions of the present study are as follows:

- An in-depth analysis of state-of-the-art deep learning approaches for gait recognition on mobile devices.
- An overview of the main concepts of Transformers, including the key differences between popular architectures proposed in the literature.
- To the best of our knowledge, this is the first study that explores the potential of Transformers for behavioural biometrics, in particular, gait biometric recognition on mobile devices. Several state-of-the-art Transformer architectures are considered in the evaluation framework (Vanilla, Informer, Autoformer, Block-Recurrent Transformer, and THAT), comparing them with traditional CNN and RNN architectures. In addition, new configurations of the Transformers are proposed to further improve the performance.
- An extensive experimental framework using popular public databases in gait biometric recognition. On the existing whuGAIT [15] and OU-ISIR [16,17] databases, the proposed Transformer outperforms traditional CNN and RNN architectures and achieves competitive results compared with the state of the art.
- We make our experimental framework available to the research community in order to advance mobile gait recognition research[2].

The exploration and analysis included in the present study can also be very useful for other research lines, for example: *i)* improving the authentication performance of other behavioural biometric traits such as handwritten signature and keystroke [18,19], among many others, *ii)* improving the prediction and monitoring of diseases [20], and *iii)* facilitating the training and synthesis of new data [21,22].

The remainder of the article is organised as follows. Section 2 summarises previous studies in the field of gait recognition on mobile devices. Section 3 explains the main concepts of Transformers and the key differences between the architectures considered in the study. Section 4 describes the databases and experimental protocol while Section 5 provides a description of the system details. Section 6 describes the results achieved and comparison with the state of the art. Finally, Section 7 draws the final conclusions and future research lines.

2. Related works

Gait biometric recognition enables subjects to be authenticated based on their walking patterns. Due to the exponential increase in the number of mobile devices and the high precision of their sensors, the interest in gait recognition based on mobile devices is on the increase [4]. One of the most popular approaches is based on the Inertial Measurement Units (IMU), e.g., accelerometer and gyroscope [28]. Table 1 provides a summary of the most relevant methodologies for gait biometric recognition on mobile devices based on DL methods. It is important to highlight that all approaches consider the same experimental protocol proposed in Zou et al. [15] for two popular public databases in the literature: *i)* whuGAIT [15], which comprises accelerometer and gyroscope data acquired from mobile devices, and *ii)* OU-ISIR [16,17], which includes accelerometer and gyroscope data obtained from IMU sensors.

In the past few years, the research community has focused on DL models to improve the robustness of gait recognition systems, extracting more discriminative features. As both the spatial and temporal information of the gait pattern is important for the task, DL architectures based on CNN and RNN have been utilised. One of the earliest systems based on DL models using CNNs was created by Gadaleta and Rossi [23]. The authors used CNNs for feature extraction and a Support Vector Machine (SVM) for the final classification with 0.15% misclassification rates. The score was obtained in less than five walking cycles with their own collected database. Their results proved how DL methods could extract more discriminative features compared with previous machine learning methods. The same model was evaluated in Zou et al. [15] following a predefined experimental protocol, obtaining an accuracy of 92.91% in the whuGAIT database [15], and 44.29% accuracy in the OU-ISIR database [16,17]. Another approach based on CNNs was presented by Delgado-Escaño et al. [24], dividing the data into two branches, according to each sensor (accelerometer and gyroscope). The output of both branches were concatenated to produce a joint feature vector. Cross-validation was used, achieving 95.20% accuracy with the OU-ISIR database using their own experimental protocol. Following the predefined experimental protocol presented in Zou et al. [15], results of 92.89% and 44.29% accuracy were achieved in the whuGAIT and OU-ISIR databases, respectively. However, by using only CNNs, the system focuses mainly on spatial characteristics, leaving out the temporal information.

To overcome this drawback, RNNs were proposed, extracting temporal features from the time sequences. Watanabe et al. created an end-to-end RNN with a softmax layer [25]. The model was tested with the experimental protocol presented in Zou et al. [15], achieving a 91.88% accuracy with whuGAIT database, and 66.36% accuracy with OU-ISIR database. Zou et al. evaluated RNNs in Zou et al. [15] over the OU-ISIR database achieving 78.92% accuracy. They also presented the whuGAIT database and proposed a predefined experimental protocol, achieving 93.14% accuracy.

Hybrid approaches have also been proposed in the literature, achieves a more complex structure, where the CNN extracts spatial features while the RNN obtains temporal features. Ordoñez and Roggen presented in Ordóñez and Roggen [27] DeepConvLSTM, which comprises convolutional layers, followed by recurrent and softmax layers. The model obtained 95.8% F1-score for the activity recognition task with the Opportunity database [29]. The system was also evaluated for gait recognition in Zou et al. [15], achieving 92.25% and 37.33% accuracy for the whuGAIT and OU-ISIR databases, respectively. Also, Zou et al. presented in [15] an hybrid approach with two-parallel branches, one CNN and one RNN. The extracted features were independent in each branch, obtaining a view of the raw data with both convolutional and recurrent layers. After each branch, the features were concatenated and

[2] https://www.github.com/BiDAlab/ExploringTransformers.

Table 1
Summary of the most relevant methodologies for gait biometric recognition based on DL methods.

Category	Year	Ref.	Description	Performance	Database
CNNs	2016	[23]	CNN Feature Extractor + SVM Classifier	92.91%	whuGAIT
				44.29%	OU-ISIR
	2019	[24]	Fusion CNN + Euclidean Distance	92.89%	whuGAIT
				40.60%	OU-ISIR
RNNs	2020	[25]	End-to-End RNN	91.88%	whuGAIT
				66.36%	OU-ISIR
	2020	[15]	End-to-End RNN	91.88%	whuGAIT
				66.36%	OU-ISIR
	2021	[26]	End-to-End Multi-RNN	93.14%	whuGAIT
				78.92%	OU-ISIR
CNNs + RNNs	2016	[27]	Cascaded CNN + RNN	92.25%	whuGAIT
				37.33%	OU-ISIR
	2020	[15]	2-Parallel Branches: CNN + RNN	93.52%	whuGAIT
	2021	[26]	2-Parallel Branches: CNN + Multi-RNN	94.15%	whuGAIT
				89.79%	OU-ISIR
Proposed Transformer	**2022**	**Present Work**	**2-Parallel Branches: Temporal and Channel Modules** **Temporal: Auto-Correlation + GBR CNN Layers and Recurrent Layer**	**94.25%**	**whuGAIT**
			Channel: Auto-Correlation + GBR CNN Layers **Gaussian Range Encoding in both Temporal and Channel Modules**	**93.26%**	**OU-ISIR**

fed into a fully connected layer. The authors achieved 93.52% accuracy on the presented whuGAIT database.

Previous approaches are based on prior gait cycle detection. The input of the DL models is an interval time between two consecutive occurrences of the gait pattern, i.e., putting the same foot on the ground [4]. Gait cycle detection is usually a tedious task that can induce to errors due to the sensor restrictions (e.g., noise-sensitive, sensor specification, body placement, etc.). To solve this problem, Tran et al. proposed in [26] a new approach using window-based data segment. The authors used a Multi-RNN model considering fixed-length segments as input, without the need to extract gait cycles. The authors achieved an accuracy of 93.14% for the whuGAIT database, and 78.92% for the OU-ISIR database. In addition, the same authors introduced an hybrid approach, achieving 94.15% and 89.79% accuracy for the whuGAIT and OU-ISIR databases, respectively.

Despite the success of CNN and RNN architectures, some of their limitations could still be revisited and improved, such as limited window sizes for RNNs. By summarising all previously observed information into one vector, these approaches miss temporal information that is relevant to gait biometric recognition. Due to the limitations highlighted, this article explores the potential of recently developed Transformer architectures for gait biometric recognition and proposes new configurations to further improve the results. Table 1 also includes the results achieved using our proposed Transformer.

3. Methods

This section provides an overview of the main concepts of Transformers, including the key differences between recent architectures proposed in the literature. To facilitate the understanding of this section, we include in Fig. 1 a graphical representation of the different Transformer architectures. As the present article is related to behavioural recognition, we focus only on the encoder part of the Transformer.

3.1. *Vanilla transformer*

The original Vanilla Transformer was presented in Vaswani et al. [7] for the task of machine translation. It was defined as a multi-layer encoder-decoder architecture with no recurrence and convolution layers. Fig. 1 *A.* provides a graphical representation of the encoder, which is composed of a stack of N identical layers. Each layer is mainly formed by two different sub-layers: *i)* a multi-head Self-Attention mechanism (Full-Attention), and *ii)* a point-wise feed-forward network. Subsequent of each sub-layer, a residual connection and a layer normalisation are considered (*Add & Norm* in Fig. 1). The input sequence is a matrix $X \epsilon \mathbb{R}^{c\times L}$ where c is the number of channels and L the length of the sequence.

The encoder maps each sample l of the input sequence $X = (x_0, x_1, \dots, x_l, \dots, x_L)$ into hidden states $Z = (z_0, z_1, \dots, z_l, \dots, z_L)$. The output of each sub-layer is $LayerNorm(X + sublayer(X))$, where $sublayer(X)$ is the function implemented by the multi-head Self-Attention mechanism (Full-Attention) or the point-wise feed-forward network. Both the input X and output Z have the same dimension L to facilitate the work of the residual connections. As no recurrence and convolutional layers are considered in the Vanilla Transformer, a previous encoding of the model is needed to keep certain information about the sample l of the sample in the input sequence. This is achieved using a positional encoding placed at the input of the model.

We describe next the key aspects of the positional encoding, multi-head Self-Attention mechanism (Full-Attention), and the point-wise feed-forward network for a better understanding of the Vanilla Transformer, and the later Transformer implementations.

3.1.1. *Positional encoding*

This stage encodes the relative and/or absolute position *pos* of the sample l of the input sequence. In the original work, Vaswani et al. [7] preserved the relative context using a fixed point encod-

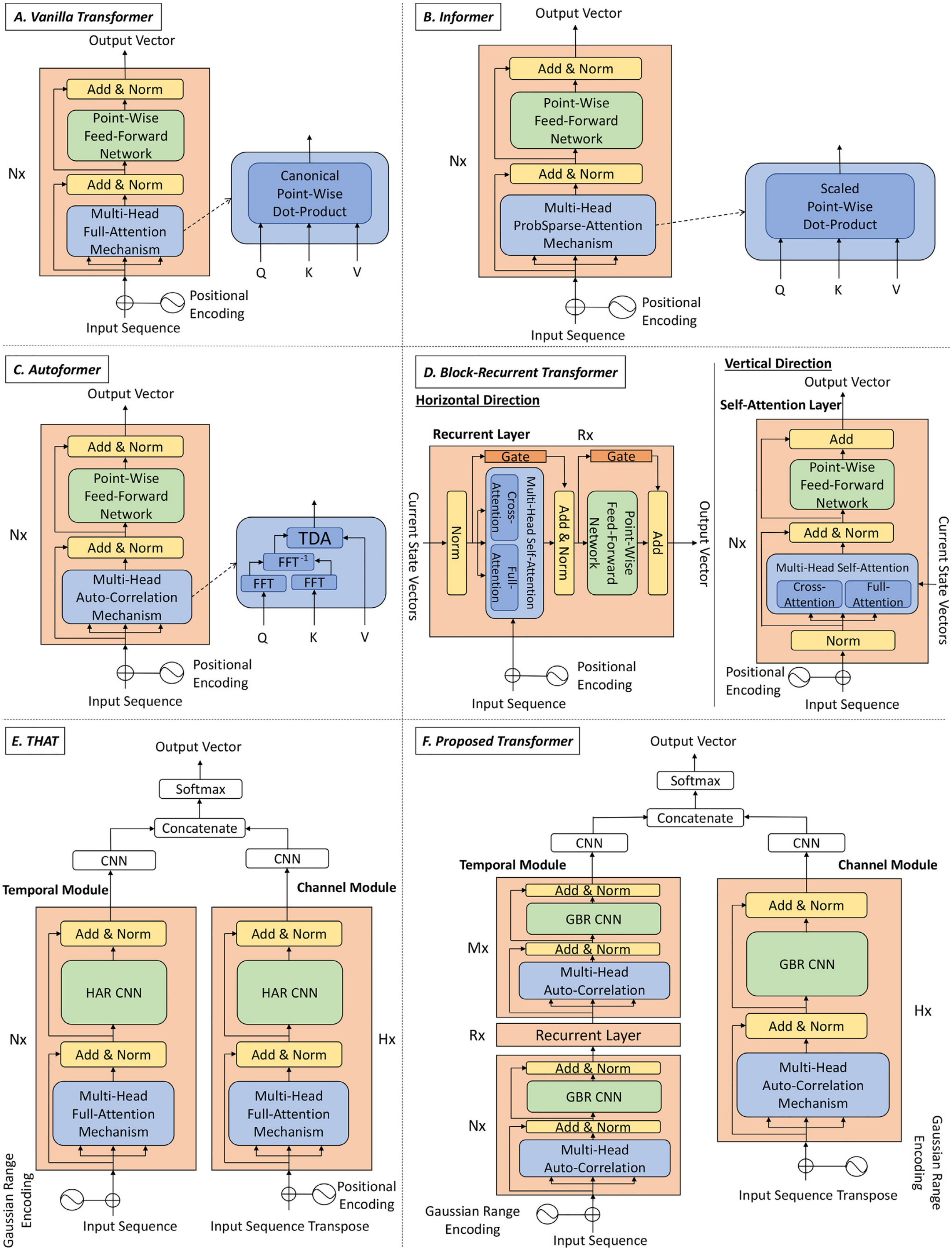


Fig. 1. Graphical representation of the Transformer architectures used in this study (Vanilla Transformer [7], Informer [11], Autoformer [10], Block-Recurrent [8], THAT [13], and our proposed Transformer). Q: Queries; K: Keys; V: Values; Nx,Hx,Rx: they refer to the number of layers of each type; FFT: Fast Fourier Transform; TDA: Time Delay Aggregation; HAR CNN: Human Activity Recognition CNN; GBR CNN: Gait Biometric Recognition CNN.

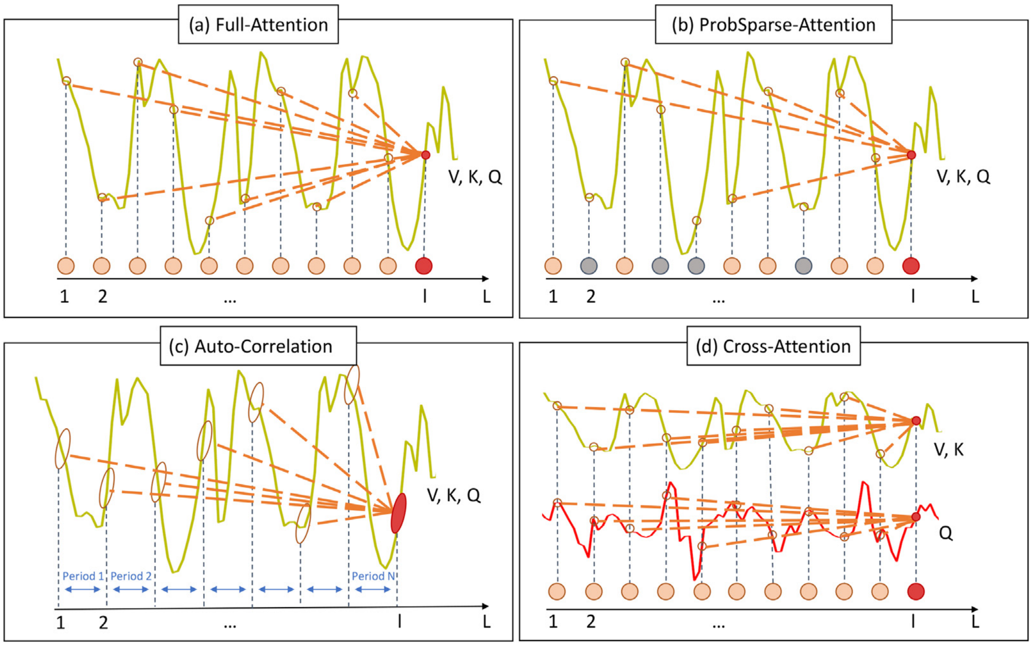


Fig. 2. Graphical representation of Attention and Auto-Correlation mechanisms. (a) Full-Attention (Vanilla Transformer [7]); (b) ProbSparse-Attention (Informer [11]); (c) Auto-Correlation (Autoformer [10]); and (d) Cross-Attention (Block-Recurrent Transformer [8]). The solid line represents the input sequence and the red one (second line) the current states in Cross-Attention. The red points/series are the sample l of the sequence of length L with V values, K keys, and Q queries. The orange points represent the mapped points/series along the entire sequence, while the grey ones are points not mapped. Figure adapted from [10].

ing with the sine and cosine functions:

$$PE_{(pos,2l)} = \sin(pos/10000^{2l/L}) \tag{1}$$

$$PE_{(pos,2l+1)} = \cos(pos/10000^{2l/L})$$

where L is the total length of the input sequence. The positional encoding has the same length L as the embeddings, so that the two can be summed. The output of the positional encoding is:

$$\hat{x_l} = x_l + PE_{(l)} \tag{2}$$

3.1.2. Multi-head self-attention mechanism

This mechanism is responsible for mapping scattered points along the entire sequence, studying the long-range dependencies. This mechanism avoids the limited time window problem of previous architectures (e.g., RNNs). The information aggregation is accomplished with a Full-Attention mechanism where the outputs are the weighted sum of the values V according to the canonical point-wise dot-product of the queries Q with the corresponding keys K. Fig. 2(a) provides a graphical representation of the Full-Attention mechanism. The solid line represents the input sequence with its values V, keys K, and queries Q. The red point represents the sample l in the sequence with length L. The orange points are the scattered points mapped in the Full-Attention mechanism for the red point at sample l. The Full-Attention mechanism can be defined as:

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \tag{3}$$

where d_k is the dimension of the queries Q and keys K, and $\sqrt{d_k}$ is a scaling factor that enables flatter gradients. $Q = XW_Q$, $K = XW_K$, $V = XW_V$ are the linear projections of X in the corresponding projection parameters d_k, d_k, and d_v respectively where $W_Q \in \mathbb{R}^{L\times d_k}$, $W_K \in \mathbb{R}^{L\times d_k}$, and $W_V \in \mathbb{R}^{L\times d_v}$. The computational cost is quadratic $O(L^2)$ where L denotes the length of the input sequence.

Alternatively to apply one single projection of the queries, keys, and values, better results can be achieved with h independent projections to d_k, d_k, and d_v respectively. The multi-head Self-Attention is based on a concatenation and final projection of the h independent heads:

$$MultiHead(Q, K, V) = [head_1, \dots, head_h]W^O \tag{4}$$

where $head_i = Attention(Q_i, K_i, V_i)$ and $W^O \in \mathbb{R}^{hd_v\times L}$ is the final attention matrix. To achieve the same length L of the input sequence, $d_v = L/h$. Therefore the attention matrix of Full-Attention is $L \times L$.

3.1.3. Point-wise feed-forward network

In addition to the multi-head Self-Attention sub-layer, the Vanilla Transformer has a point-wise feed-forward network. This consists of two linear transformations with a ReLU activation in between, operating in each position independently. The input and output dimensions are the same, L.

To summarise, the Vanilla Transformer has shown great advances in Natural Language Processing and Computer Vision applications but still needs to be adapted for time sequences. Aspects such as the periodicity or seasonality, and long- and short-range dependencies still need to be revisited [14]. To alleviate these drawbacks, different Transformers have been proposed in the research community, modifying aspects such as the multi-head Self-Attention sub-layer and the positional encoding.

3.2. Informer

Zhou et al. presented in [11] a new Transformer architecture named Informer. Informer is an adaptation of the Vanilla Transformer for Long Sequence Time-series Forecasting (LSTF). Some limitations of the Vanilla Transformer are the quadratic time complexity $O(L^2)$ and the high memory usage $O(L^2)$ for each encoder layer; and the inherent limitation of the encoder-decoder architecture. To overcome these drawbacks, the authors proposed several improvements. The multi-head Self-Attention mechanism based on Full-Attention was changed by ProbSparse-Attention to scattered points, as provides Fig. 1 *B*. The Full-Attention to the input sequence is reduced to half, more favourable handling long-range sequences. The canonical dot-product was replaced by a scaled dot-product. Informer reduces the time complexity to $O(L \log L)$ and the memory usage to $O(L \log L)$ for each layer. In addition, previous studies have shown a potential sparsity in Full-Attention. As a result, the authors decided to use a selective strategy on all probabilities, i.e., Sparse-Attention [30] (sparsity coming from separate spatial correlations) and LogSparse-Attention [31] (selecting points through exponentially increasing intervals). Fig. 2(b) provides a graphical representation of the ProbSparse-Attention mechanism. The solid line denotes the input sequence with the extracted values V, keys K, and queries Q. The red point represents the sample l in the input sequence. The ProbSparse-Attention mechanism, unlike the Full-Attention mechanism that looks at all previous points, chooses selected dominant points (orange) in the input sequence, while the grey ones are not used.

3.3. Autoformer

Autoformer was presented by Wu et al. [10] for the task of long-term forecasting. In this Transformer architecture, the original multi-head Self-Attention mechanism based on Full-Attention was changed by Auto-Correlation. Contrary to previous Transformers, where the proposed dot-product only establishes point connections, the Auto-Correlation mechanism not only utilises long-range dependencies but also periodicity-based dependencies. Using series-wise instead of point-wise connections, Autoformer achieves $O(L \log L)$ time complexity and $O(L \log L)$ memory usage for each layer, and breaks the information utilisation bottleneck. Fig. 2(c) shows a graphical representation of Auto-Correlation. It takes into consideration series of points in the same position during previous periods of the input sequence instead of scattered points.

Fig. 1 *C*. provides a graphical representation of Autoformer. The multi-head Auto-Correlation sub-layer comprises two main sub-blocks: *i)* an aggregated top-k similar sub-series, calculated by Fast Fourier Transform (FFT) and based on periodicity (instead of scattered points like the Self-Attention family), and *ii)* Time Delay Aggregation (TDA) among periods (instead of point-wise dot-product like in the Self-Attention family), used for the information aggregation.

The *aggregated top-k similar sub-series* presents series-wise connections based on period-based dependencies. The sub-series are correlated between them at the same position in previous periods, which are congenitally sparse. For an input sequence $X = (x_0, x_1, \dots, x_l, \dots, x_L)$, $X \in \mathbb{R}^{c \times L}$ where c is the number of channels and L the length of the input sequence, the Auto-Correlation $R_{XX}(\tau)$ can be obtained by FFT based on Wiener–Khinchin theorem as:

$$\begin{aligned} S_{XX}(f) &= FFT(X)FFT^*(X) \\ R_{XX}(\tau) &= FFT^{-}1(S_{XX}(f)) \end{aligned} \tag{5}$$

where FFT^* is the conjugate operation, FFT^{-1} its inverse, and $S_{XX}(f)$ is the Auto-Correlation obtained in the frequency domain.

The *Time Delay Aggregation (TDA)* sub-block links the sub-series over the selected time delays $\tau_1, \dots, \tau_k$. This operation aligns sub-series in the same phase of the predicted periods, contrary to point-wise dot-product in the Self-Attention family. Finally, the sub-series are aggregated by softmax normalised function. The Auto-Correlation mechanism can be defined as:

$$\begin{aligned} \tau_1, \dots, \tau_k &= \underset{\tau \ \in \ (1,\dots,L)}{argTopK}(R_{Q,K}(\tau)) \\ \hat{R}_{Q,K}(\tau_1), \dots, \hat{R}_{Q,K}(\tau_k) &= SoftMax(R_{Q,K}(\tau_1), \dots, R_{Q,K}(\tau_k)) \\ Auto-Correlation(Q, K, V) &= \sum_{i=1}^{k} Roll(V, \tau_i)\hat{R}_{Q,K}(\tau_i) \end{aligned} \tag{6}$$

where *argTopK* takes the output of *topK* Auto-Correlations along l, $R_{Q,K}$ is the Auto-Correlation between Q and K series, and $Roll(V, \tau_i)$ scroll X with a τ time delay, re-introducing the elements moved beyond the first position to the last one.

3.4. Block-recurrent transformer

Hutchins et al. introduced the Block-Recurrent Transformer in Hutchins et al. [8] for the task of auto-regressive language modelling. This Transformer introduces a recurrent form of attention. It is presented as an alternative to using the dot-product or periodicity-based series mechanism, which fix an attention window size. The Block-Recurrent Transformer summarises the sequence that the model has previously seen. The time complexity is linear $O(L)$ for each layer. The recurrent layers operate on series-wise connections as in the Autoformer, achieving linear memory consumption $O(L)$ in each layer. The Block-Recurrent Transformer is based on a sliding-window attention mechanism [32]. Given an input X with length L, a causal mask is applied by a sliding window with size W where every sample can attend only to the previous W samples. Being the attention matrix of Full-Attention $L \times L$, the Block-Recurrent Attention matrix is $W \times W$, where $W << L$. The sliding-window attention processes multiple blocks of size W at the same time.

Fig. 1 *D*. provides a graphical representation of the Block-Recurrent Transformer architecture, which comprises two main directions: *i)* vertical direction (Self-Attention Layer in Fig. 1 *D*.), where layers are placed in the usual way; and *ii)* horizontal direction (Recurrent Layer in Fig. 1 *D*.), where layers contain recurrence. Both directions attend to the input sequence X and to the current states S.

The *vertical direction* presents a multi-head Self-Attention sub-layer with two attentions: *i)* Full-Attention to the input sequence X as shown in Fig. 2(a); and *ii)* Cross-Attention applied in a similar way to the original Vanilla Transformer [7], with the main difference being that the queries Q come from the current states S, which are initialised to 0, whereas the keys K and values V are extracted from the input sequence X, Fig. 2(d).

The *horizontal direction* also presents a multi-head Self-Attention sub-layer with two attentions: *i)* Cross-Attention to the input sequence X to extract the queries Q while the keys K and values V are extracted from the current states S, Fig. 2(d), and *ii)* Full-Attention to the current states S, Fig. 2(a). The horizontal direction applies recurrence where the residual connections are replaced by gates, allowing the model to forget. Also, the gates help the model to apply Full-Attention and Cross-Attention in parallel. For the recurrence, the current states S are modified by residual connection gates. The input of the state at the next window (s_{w+1}) depends on the output of the state at the actual window (s_w):

$$\begin{aligned} s_{w+1} &= s_w \odot g + z_w \odot (1 - g) \\ g &= \sigma(b^{(g)}) \\ z_w &= W^{(z)}h_w + b^{(z)} \end{aligned} \tag{7}$$

where $\odot$ is the point-wise multiplication, g the gate, z_w the learned convex combination, $b^{(g)}$ and $b^{(z)}$ are trainable bias vectors (learned functions between the distance of the query Q and key K), W the weight matrix, h_w the output of the corresponding sub-layer (i.e., multi-head Self-Attention mechanism or point-wise feed-forward network), and σ the sigmoid function.

The Block-Recurrent Transformer applies layer normalisation before the multi-head Self-Attention sub-layer, and before the point-wise feed-forward network. Dropout is also introduced before the multi-head Self-Attention sub-layer and after the point-wise feed-forward network.

3.5. THAT

Contrary to images, which have spatial information in two dimensions (2D), temporal sequences might consider spatial information in one dimension (1D) in each time position. Furthermore, they can extract temporal information for each time position in a second dimension. The spatial information is available in the same way, between the different channels of each time sample, which can be called as channel-over-time features. On the other hand, being a temporal sequence, there are time-over-channel features, which need to be treated as a temporal sequence.

Based on this idea, the *Two-stream Convolution Augmented Human Activity Transformer* (THAT) model was proposed by Li et al. [13]. The authors proposed a new Transformer architecture for Human Activity Recognition (HAR). Fig. 1 *E.* provides a graphical representation of the THAT Transformer. The model contains two parallel modules for the feature extraction: *i)* Temporal Module (in charge of time-over-channel features), and *ii)* Channel Module (in charge of channel-over-time features). Subsequently, all extracted features are concatenated for the prediction task.

The authors claimed that the original positional encoding considered in the Vanilla Transformer [7] might not be sufficient to capture all the temporal information along the sample as it is defined on a single point. As a result, the authors proposed a Gaussian range encoding, suggesting the use of a range of points rather than just one. Furthermore, several ranges g can be used at the same time, allowing to have different contexts of the sample x_l.

Assuming $g \in \mathbb{R}^G$ different ranges, $\mathcal{N}(\mu^g, \sigma^g) \in \mathbb{R}^{L\times G}$ is a Gaussian distribution with the probability $p^g(l)$. Being $p_l = (\frac{p^1(l)}{\zeta}, \ldots, \frac{p^G(l)}{\zeta})$ the distribution over the G ranges with a normalisation factor ζ, $V = (v_1, \ldots, v_G)$ is the values vector over the ranges. All μ, σ, and V variables are initialised randomly and readjusted with the training of the whole model. To summarise, the output of the Gaussian range encoding at the position of sample l is:

$$\hat{x}_l = x_l + V^T p_l \tag{8}$$

In addition, as the point-wise feed-forward layer proposed in the Vanilla Transformer [7] focuses attention on a single point in time, the authors implemented a multi-scale CNN with adaptive Scale-Attention in both Temporal and Channel Modules. They replaced the linear transformations of the original feed-forward layer with a HAR CNN. Also, by introducing Scale-Attention Adaptive, the training can be adjusted to the different ranges introduced by the Gaussian range encoding.

Finally, THAT has quadratic time complexity $O(L^2)$ and the high memory usage $O(L^2)$ for each encoder layer, since the model uses Self-Attention (i.e., Full-Attention similar to the Vanilla Transformer).

3.6. Proposed transformer

Finally, Fig. 1 *F.* presents the new proposed Transformer based on a selection of the best components presented in previous Transformer architectures. First, we consider a parallel two-stream architecture with Temporal and Channel Modules, similar to the THAT approach presented in Li et al. [13]. Unlike the THAT model, we consider a Gaussian range encoding as input of both Temporal and Channel Modules. In addition, for the Temporal Module (left branch), we consider a combination of multi-head Auto-Correlation layers, proposed in Autoformer [10], and a recurrent layer in between, proposed in Block-Recurrent Transformer [8]. For the multi-head Auto-Correlation layer, we design a specific multi-scale Gait Biometric Recognition (GBR) CNN sub-layer. Regarding the Channel Module (right branch), we consider a multi-head Auto-Correlation sub-layer together with a multi-scale GBR CNN sub-layer. After each sub-layer, a residual connection is applied followed by a normalisation of the layer, similar to the Vanilla Transformer [7]. The time complexity and memory usage of each layer with Auto-Correlation is $O(L\ log\ L)$, whereas for the recurrent layer this is $O(L)$.

4. Experimental protocol

Two popular public databases used for research in gait recognition on mobile devices are considered in the evaluation framework of the present study: *i)* whuGAIT [15], and *ii)* OU-ISIR [17]. These databases have been selected as they also contain predefined experimental protocols for the identification task (i.e., development and evaluation datasets), allowing for a fair comparison between existing state-of-the-art approaches.

4.1. WhuGAIT database

The whuGAIT database was introduced in Zou et al. [15]. This database comprises accelerometer and gyroscope data acquired using Samsung, Xiaomi, and Huawei smartphones in unconstrained scenarios. The sampling frequency of the accelerometer and gyroscope sensors is 50 Hz. A total of 118 subjects participated in the acquisition, and both walking and non-walking sessions were considered.

Regarding the experimental protocol of the whuGAIT database, Zou et al. proposed in [15] a predefined division of the database into development and evaluation datasets in order to facilitate the comparison among approaches. For each subject, 90% of the samples are considered for development while the remaining 10% for the final evaluation. In total 33,104 samples are considered for the development dataset whereas the remaining 3,740 samples are used for the final evaluation.

4.2. OU-ISIR database

The OU-ISIR database was presented in Ngo et al. [17]. This database comprises 745 subjects; the largest public mobile device gait biometric database to date. Data from accelerometer and gyroscope sensors were collected using three IMUs and a Motorola ME860 smartphone around the waist of the subject. The sampling frequency of the sensors is 100 Hz. Subjects had to perform 4 different activities (two flat walking, slope-up walking, and slope-down walking). The database is divided into two different subsets. The first subset includes data from 744 users collected by one IMU located in the middle of the subject's back at waist-height. The second one contains data from 408 subjects collected by the three IMUs and the smartphone.

Regarding the experimental protocol of the OU-ISIR database, we consider the predefined division of the database into development and evaluation datasets proposed by Zou et al. [15]. For each subject, 87.5% of the samples are considered for development while the remaining 12.5% for the final evaluation. In total 13,212 samples are considered for the development dataset whereas the remaining 1,409 samples are used for the final evaluation.

5. Systems details

This section provides the system configuration details of the Transformers and traditional DL architectures (i.e., CNNs and RNNs) considered in the experimental framework of the study.

The same inputs to the models is used for all approaches. For the whuGAIT database, a total of 80 time signals (around 1.5 s each) are extracted from the 3-axis accelerometer and gyroscope sensors following the approach presented in Tran et al. [26]. Also, we consider an overlapping of 97% between samples in training.

For the OU-ISIR database, 128 time signals (around 1.5 s each) are extracted from the 3-axis accelerometer and gyroscope sensors following the approach presented in Zou et al. [15]. Also, we consider an overlapping of 61% between samples in training.

For a better comparison of Transformer architectures with popular DL architectures, we consider the following approaches: *i)* CNNs, *ii)* RNNs, and *iii)* a hybrid configuration based on the combination of CNNs and RNNs. These DL models are widely considered for gait biometric recognition, achieving state-of-the-art results as described in Section 2. CNNs have shown advantages in capturing spatial dependencies, while RNNs are better to capture the temporal dependencies.

We provide next a description of the networks parameters:

- *CNN*: we consider four 1D convolutional layers with 6 units each and kernel size 5, followed by one dense layer with $\frac{3}{2}L$ units (where L is the length of the time sequence), and one softmax layer. After every 2 convolutional layers, we use max-pooling and dropout with a 0.5 rate. ReLU activation functions are used in both convolutional and dense layers. The total number of model parameters is 57.3K and 35.4M for the whuGAIT and OU-ISIR databases, respectively.
- *RNN*: we consider three LSTM layers with 3 units each followed by one dense layer with $\frac{3}{2}L$ units, and one softmax layer. The total number of model parameters is 785.2 K and 3.6 M for the whuGAIT and OU-ISIR databases, respectively.
- *CNN-RNN*: it comprises two parallel modules, *i)* four convolutional layers with 6 units each and kernel size 5, and *ii)* three LSTM layers with 3 units each. After both modules, a feature concatenation is applied, followed by one dense layer with $\frac{3}{2}L$ units, and one softmax layer. We also consider dropout with 0.5 rate after each convolutional layer. The total number of model parameters is 870.5 K and 40.4 M for the whuGAIT and OU-ISIR databases, respectively.
- *Vanilla Transformer [7]*: we consider the positional encoding together with the encoder part of the Vanilla Transformer. The model consists of $N = 5$ layers. For the multi-head Self-Attention sub-layer, 8 heads are considered with Full-Attention whereas for the point-wise feed-forward network we consider two linear layers (layer 1 with L units and layer 2 with $L * 4$ units) with ReLU activation and dropout in between. The total number of model parameters is 705.7 K and 3.4 M for the whuGAIT and OU-ISIR databases, respectively.
- *Informer [11]*: we consider the same structure as the Vanilla Transformer but changing in the multi-head Self-Attention sub-layer the Full-Attention to ProbSparse-Attention. The model is composed of $N = 5$ layers. For the multi-head Self-Attention sub-layer, 8 heads are considered whereas for the point-wise feed-forward network we consider two linear layers (layer 1 with L units and layer 2 with $L * 4$ units) with ReLU activation and dropout in between. The total number of model parameters is 705.7 K and 3.4 M for the whuGAIT and OU-ISIR databases, respectively.
- *Autoformer [10]*: the same structure as the Vanilla Transformer is considered but changing the Self-Attention mechanism for the Auto-Correlation mechanism. The model comprises $N = 5$ layers with 8 heads in the multi-head Auto-Correlation sub-layer. For the point-wise feed-forward network we consider two linear layers (layer 1 with L units and layer 2 with $L * 4$ units) with ReLU activation and dropout in between. The total number of model parameters is 1.0M and 3.4M for the whuGAIT and OU-ISIR databases, respectively.
- *Block-Recurrent Transformer [8]*: it comprises 12 layers: $N = 9$ multi-head Self-Attention layers with Cross-Attention and Full-Attention (8 heads), followed by $R = 1$ recurrent layer, and $M = 2$ more multi-head Self-Attention layers with Cross-Attention and Full-Attention (8 heads). In each layer, the point-wise feed-forward network is composed of two linear layers (layer 1 with L units and layer 2 with $L * 4$ units) with ReLU activation and dropout in between. The total number of model parameters is 2.4 M and 3.3 M for the whuGAIT and OU-ISIR databases, respectively.
- *THAT [13]*: this is a two-stream convolution Transformer architecture. In the first stream (Temporal Module) the time-over-channel features are analysed. To this aim, Gaussian range encoding is used together with the original multi-head Self-Attention sub-layer (Full-Attention with 8 heads). The HAR CNN sub-layer is based on a multi-scale CNN (3 convolutional layers with L units each, ReLU activation functions, and kernel sizes 1, 3, and 5 respectively, followed by dropout layers). The Temporal Module contains $N = 9$ layers. For the second stream (Channel Module) the data is transposed to extract the channel-over-time features, adopting the original Vanilla Transformer structure with positional encoding. The multi-head Self-Attention sub-layer contains Full-Attention with 6 heads. The HAR CNN sub-layer is based on a multi-scale CNN (3 convolutional layers with L units each, ReLU activation functions, and kernel sizes 1, 3, and 5 respectively, followed by dropout layers). The Channel Module contains $H = 1$ layer. The total number of model parameters is 611.7 K and 4.7 M for the whuGAIT and OU-ISIR databases, respectively.
- *Proposed Transformer*: we consider a two-stream Transformer based on Temporal and Channel Modules. Both modules use Gaussian range encoding. The Temporal Module comprises 12 layers: $N = 9$ multi-head Auto-Correlation layers (8 heads), followed by $R = 1$ recurrent layer (8 heads), and $M = 2$ multi-head Auto-Correlation layers (8 heads). In each layer, the GBR CNN sub-layer is based on a multi-scale CNN (4 convolutional layers with L units each, ReLU activation functions, and kernel sizes 1, 3, 5, and 7 respectively, followed by dropout layers). The Channel Module comprises $H = 1$ layers. In all of them we consider multi-head Auto-Correlation mechanism with 6 heads. The GBR CNN sub-layer is based on a multi-scale CNN (4 convolutional layers with L units each, ReLU activation functions, and kernel sizes 1, 3, 5, and 7 respectively, followed by dropout layers). These parameters have been selected according to the performance achieved with the proposed Transformer. The total number of model parameters is 2.6 M and 6.7 M for the whuGAIT and OU-ISIR databases, respectively.

For the training of the models, we use cross-entropy and Adam optimiser with default parameters (learning rate of 0.001). All models are adapted to the gait biometric recognition task. To this aim, after the models we include 2 convolutional layers (L units each, ReLU activation functions, and kernel sizes 128, followed by dropout layers) with max-pooling and a linear layer with softmax activation function. For the THAT and proposed Transformer, we also consider feature concatenation of the Temporal and Channel Modules as described in Fig. 1 *E.* and *F.*

6. Experimental results

This section aims to analyse the performance of the different state-of-the-art Transformer architectures considered in this study (i.e., Vanilla, Informer, Autoformer, Block-Recurrent Transformer, THAT, and our proposed architecture) for the topic of gait biometric recognition on mobile devices. Section 6.1 provides a comparison of Transformer architectures with traditional DL architectures such as CNNs and RNNs. Finally, Section 6.2 provides a comparison of the proposed Transformer architectures with the state of the art.

Table 2
Comparison in terms of accuracy of traditional DL models (CNN, RNN) and recent Transformers for biometric gait recognition. GRE: Gaussian Range Encoding; N, M: Number of multi-head Auto-Correlation layers before and after the recurrent layer, respectively; R: Number of recurrent layers.

Model		**Database**	
		whuGAIT	**OU-ISIR**
CNN		75.31%	32.51%
RNN		82.42%	44.15%
CNN + RNN		84.54%	46.63%
Vanilla Transformer [7] (Positional Encoding + Full-Attention)		87.73%	54.51%
Informer [11] (Positional Encoding + ProbSparse-Attention)		89.26%	59.40%
Autoformer [10] (Positional Encoding + Auto-Correlation)		89.44%	63.10%
Block-Recurrent Transformer [8] (Positional Encoding + Full- and Cross-Attention)		91.78%	64.52%
THAT [13]: Temporal Module (GRE + Full-Attention + w/o Recurrent Layer), Channel Module (Positional Encoding + Full-Attention)		92.99%	85.74%
Proposed Transformer			
Temporal Module	GRE + Full-Attention + w/o Recurrent Layer	90.96%	57.06%
	GRE + ProbSparse-Attention + w/o Recurrent Layer	91.07%	59.48%
	GRE + Auto-Correlation + w/o Recurrent Layer	91.15%	60.61%
	GRE + Auto-Correlation + w/ Recurrent Layer ($N = 8$, $R = 1$, $M = 2$)	92.23%	59.20%
	GRE + Auto-Correlation + w/ Recurrent Layer ($N = 9$, $R = 1$, $M = 2$)	**92.45%**	**68.20%**
	GRE + Auto-Correlation + w/ Recurrent Layer ($N = 10$, $R = 1$, $M = 2$)	91.16%	53.73%
	GRE + Auto-Correlation + w/ Recurrent Layer ($N = 9$, $R = 1$, $M = 1$)	92.30%	56.50%
	GRE + Auto-Correlation + w/ Recurrent Layer ($N = 9$, $R = 1$, $M = 3$)	91.10%	57.06%
Channel Module	Positional Encoding + Full-Attention	91.68%	70.55%
	GRE + Full-Attention	92.28%	90.77%
	GRE + ProbSparse-Attention	93.26%	91.20%
	GRE + Auto-Correlation	**93.64%**	**92.19%**
Temporal + Channel Modules	**Temporal (GRE + Auto-Correlation + w/ Recurrent Layer) Channel (GRE + Auto-Correlation)**	**94.25%**	**93.33%**

6.1. Transformers vs. traditional DL architectures

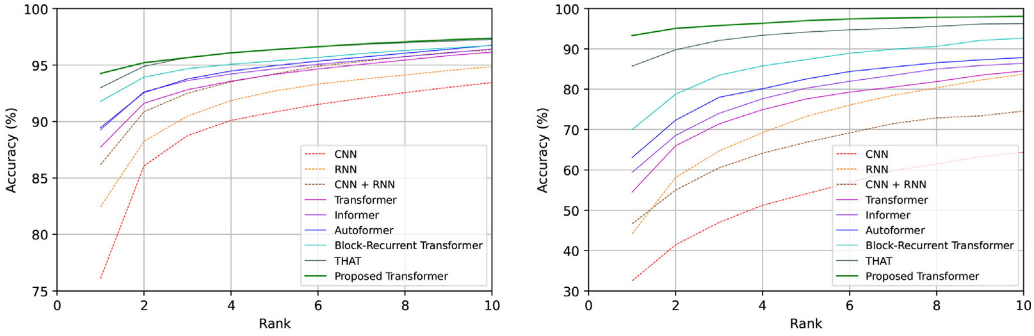
Table 2 provides a comparison of traditional DL models and recent Transformers for the whuGAIT and OU-ISIR databases. The best results achieved for each database and module configuration (Temporal and Channel) are remarked in bold. First, we can see that the Vanilla Transformer outperforms the traditional DL models (CNN, RNN, and CNN + RNN) in both databases. The Vanilla Transformer achieves an accuracy of 87.73% in the whuGAIT database (absolute improvement of 3.19% accuracy compared with the CNN + RNN approach), and 54.51% in the OU-ISIR database (absolute improvement of 7.88% accuracy compared with the CNN + RNN approach). These performance improvements demonstrate the advantages of Transformers compared with traditional CNN and RNN architectures, for example, the ability to train the model using large time sequences, attending to all the previous samples at the same time. In addition, we can also observe a considerable gap in the results between the whuGAIT and OU-ISIR databases. This is due to the OU-ISIR comprising more challenging database including many more subjects, sensors, and walking styles. This trend is also observed in the original article for traditional CNN and RNN architectures [17].

The Vanilla Transformer architecture [7] was improved using ProbSparse-Attention (Informer [11]) and Auto-Correlation (Autoformer [10]). Analysing the results included in Table 2, we can observe that both Informer and Autoformer outperform the Vanilla Transformer in both whuGAIT and OU-ISIR databases. In particular, for the whuGAIT database, the Informer and Autoformer achieve 89.26% and 89.44% accuracy, respectively, in comparison with the 87.73% accuracy achieved for the Vanilla Transformer (absolute improvement of around 1.6% accuracy). Regarding the OU-ISIR database, much better results are achieved by Informer and Autoformer compared with the Vanilla Transformer (59.40%, 63.10%, and 54.51% accuracy, respectively). Also, Autoformer outperforms Informer in both databases, proving the potential of the multi-head Auto-Correlation mechanism, replacing the point-wise connections for series-wise connections.

The Block-Recurrent Transformer [8] was presented as an alternative to use the dot-product or periodicity-based series mechanism, which fixes an attention window size. Analysing the results of Table 2, the Block-Recurrent Transformer outperforms previous Transformers for both whuGAIT (91.78% accuracy) and OU-ISIR (64.52% accuracy) databases. This is an absolute improvement of 2.34% and 1.42% accuracy compared with Autoformer for the whuGAIT and OU-ISIR databases, respectively.

The THAT Transformer [13] proposed a two-stream approach based on Temporal and Channel Modules. This Transformer architecture outperforms all previous Transformers, achieving accuracies of 92.99% and 85.74% for the whuGAIT and OU-ISIR databases, respectively. The improvement is much higher for the OU-ISIR database with an absolute improvement of 21.22% accuracy compared with the Block-Recurrent Transformer. The main reason for this improvement is the proposed Gaussian range encoding in the Temporal Module, better capturing the temporal information of the sample in comparison with the positional encoding considered in all previous Transformers. Moreover, by having multi-scale convolutions instead of feed-forward linear layers, more discriminative patterns of each subject are captured. THAT also demonstrates how, by obtaining features from two points of view (time-over-channel features and channel-over-time features), complementary information can be captured, achieving better performance.

In addition, we show in Table 2 the results achieved by our proposed Transformer under different configurations. First, we analyse the impact in the system performance of each of the modules individually. The Temporal Module with Self-Attention (Full- and ProbSparse-Attention) and without recurrent layer (“w/o recurrent layer” in Table 2) achieves values of 90.96% and 91.07% accuracy for the whuGAIT database and 57.06% and 59.48% accuracy for the OU-ISIR database, respectively. These results are further improved by replacing the Self-Attention mechanism with the Auto-Correlation mechanism (91.15% and 60.61% accuracy for the whuGAIT and OU-ISIR databases, respectively). In addition, when including the recurrent layer (“w/ recurrent layer” in Table 2), the Temporal Module achieves better results (92.45% and 68.20% ac-



(a) whuGAIT Database

(b) OU-ISIR Database

Fig. 3. Cumulative Match Characteristic (CMC) curves of the traditional DL models (CNN, RNN, CNN + RNN) and recent Transformers (Vanilla, Informer, Autoformer, Block-Recurrent, THAT, and the proposed Transformer) for both whuGAIT (top) and OU-ISIR (bottom) databases.

curacy for the whuGAIT and OU-ISIR databases, respectively), being the best configuration $N = 9$ Multi-head Auto-Correlation layers, $R = 1$ recurrent layer, and $M = 2$ Multi-head Auto-Correlation layers. On the other hand, we can see that the Channel Module with Full-Attention is also able to extract discriminative features for the task, achieving accuracy values of 91.68% and 70.55% for the whuGAIT and OU-ISIR databases, respectively. Moreover, including the Gaussian range encoding (instead of positional encoding), the Channel Module improves the results (92.28% and 90.77% for the whuGAIT and OU-ISIR databases), becoming even better when the Self-Attention mechanism with Full-Attention is replaced by ProbSparse-Attention or Auto-Correlation, (93.26% and 93.64% accuracy for the whuGAIT and 91.20% and 92.19% accuracy for the OU-ISIR databases, respectively).

Finally, we can see how the combination of both Temporal and Channel modules ("Temporal + Channel Modules" in Table 2) outperforms all previous Transformer architectures for both whuGAIT (94.25% accuracy) and OU-ISIR (93.33% accuracy) databases. In particular, the proposed Transformer achieves absolute improvements of 2.47% (Block-Recurrent Transformer), 4.81% (Autoformer), 4.99% (Informer), and 6.52% (Vanilla Transformer) accuracy for the whuGAIT database. This improvement is even higher for the OU-ISIR database with absolute improvements of 28.81% (Block-Recurrent Transformer), 30.23% (Autoformer), 33.93% (Informer), and 38.82% (Vanilla Transformer) accuracy. It is important to highlight that in the OU-ISIR database, which is far more challenging than whuGAIT in terms of number of subjects and walking activities, the proposed Transformer achieves considerable improvements in comparison with the THAT approach (93.33% vs. 85.74% accuracy), an absolute improvement of 7.59% accuracy. These results highlight the high potential of the proposed Transformer which are produced for several reasons. First, the Gaussian range encoding allows to introduce in each sample details about its relative position with respect to the contiguous samples (before the Temporal Module) and about the different channels (before the Channel Module), obtaining more complex information. Another advantage is the two-stream architecture, where each of the modules extracts different features (the Temporal Module extracts time features while the Channel Module extracts spatial features). By extracting features from two different perspectives, a more global view of each sample is obtained. In addition, the application of Auto-Correlation in the multi-head Self-Attention mechanism together with the Gaussian range encoding in both Temporal and Channel Modules allow the extraction of series-wise connections in each range of the encoding, analysing the different behaviour of each sample in different environments. Furthermore, including the recurrent layer proposed in the Block-Recurrent Transformer to the Temporal Module offers a comprehensive analysis. The module summarises all the information seen previously, giving a more global view of each sample with respect to the rest. In addition, by including a multi-scale CNN instead of the original feed-forward network, the whole model is series-wise: from the Gaussian range encoding that extracts the position of each sample based on a range of points, multi-head Auto-Correlation with Block-Recurrent Attention, which extracts information periodically based on series, and multi-scale CNN that applies convolutions with different kernels to test the behaviour of samples in different ranges. Finally, the proposed Transformer achieves an absolute improvement of 0.92% in the whuGAIT database (94.25% accuracy) compared with the OU-ISIR database (93.33% accuracy). Some of the differences between the databases that may produce this improvement are: *i)* number of subjects (118 for whuGAIT and 745 for OU-ISIR); *ii)* amount of data available per subject (33,104 training samples for whuGAIT and 13,212 for OU-ISIR); *iii)* different devices (Samsung, Xiaomi, and Huawei smartphones for whuGAIT and three IMUs and a Motorola smartphone for OU-ISIR); and *iv)* different types of walking (walking and non-walking for whuGAIT and walking, slope-up and -down for OU-ISIR).

Previous results correspond to the Rank-1 accuracy. Nevertheless, in some applications we might be interested in having a ranked list of possible subjects of interest (e.g., in forensic applications). Fig. 3 shows the Cumulative Match Characteristic (CMC) curve of the traditional DL models commonly used in biometric recognition (CNN, RNN, CNN + RNN) and recent Transformers (Vanilla, Informer, Autoformer, Block-Recurrent, THAT, and the proposed Transformer) for both whuGAIT and OU-ISIR databases. In general, we can see the same trend in both databases for all approaches, improving the accuracy results with the Rank values. For example, for the proposed Transformer, the accuracy increases from 94.25% (Rank-1) to 97.37% (Rank-10) for the whuGAIT database whereas for the OU-ISIR database this value increases from 93.33% (Rank-1) to 98.08% (Rank-10).

6.2. Comparison with the state of the art

Finally, we compare in Table 3 the Rank-1 accuracy results achieved by our proposed Transformer with other state-of-the-art approaches presented in the literature for gait biometric recognition: CNNs + SVM [23], RNNs [15,26], and CNNs + RNNs [15,26,27]. The best results achieved for each database are remarked in bold. It is important to highlight that all studies consider the same experimental protocol [15] for both whuGAIT and OU-ISIR databases, allowing a straightforward and fair comparison between approaches.

Table 3
Comparison of the proposed Transformer with state-of-the-art gait biometric recognition approaches in terms of accuracy.

Study	Method	Database	
		whuGAIT	OU-ISIR
Ordóñez et al. [27]	CNN + RNN	92.25%	37.33%
Gadaleta et al. (2018) [23]	CNN + SVM	92.91%	44.29%
Zou et al. [15]	RNN	91.88%	–
	CNN + RNN	93.52%	–
Tran et al. [26]	RNN	93.14%	78.92%
	CNN + RNN	94.15%	89.79%
Proposed transformer	**Transformer**	**94.25%**	**93.33%**

In general, our proposed Transformer has outperformed previous approaches in both databases. For the whuGAIT database, the proposed Transformer achieves 94.25% accuracy, showing better results compared with the CNNs + RNNs approach presented in [26]. Analysing the OU-ISIR database, the proposed Transformer further improves the results achieved by previous approaches with 93.33% accuracy. This is an absolute improvement of 3.54% accuracy compared with the best previous approach (CNNs + RNNs [26]). The authors improved the CNN + RNN architecture using an RNN to process each channel, combined in parallel with a CNN with two channels, one for each sensor. These results support the high potential of the proposed Transformer for gait biometric recognition. In addition, it is important to highlight the better time complexity and memory usage of the proposed Transformer compared with traditional DL models.

7. Conclusions

This article has explored and proposed novel behavioural biometric systems based on Transformers. To the best of our knowledge, this is the first study that presents a complete framework for the use of Transformers in gait biometrics. Several state-of-the-art Transformer architectures (Vanilla, Informer, Autoformer, Block-Recurrent Transformer, and THAT) are considered in the experimental framework, together with a new proposed configuration. Two popular public databases are considered in the analysis, whuGAIT and OU-ISIR.

The proposed Transformer has outperformed previous Transformer architectures and traditional DL architectures (i.e., CNNs, RNNs, and CNNs + RNNs) when evaluated using both databases. In particular, for the challenging OU-ISIR database, the proposed Transformer achieves 93.33% accuracy, resulting in accuracy absolute improvements compared with other techniques of 7.59% (THAT), 28.81% (Block-Recurrent Transformer), 30.23% (Autoformer), 33.93% (Informer), and 38.82% (Vanilla Transformer). The proposed Transformer has also been compared with state-of-the-art gait biometric recognition systems, outperforming the results presented in the literature. In addition, it is important to highlight the enhanced time complexity and memory usage of the proposed Transformer compared with traditional DL models.

However, our proposed system still has some aspects that can be addressed as future work. First, our proposed Transformer has been analysed in segment-based data. Therefore, in order to reproduce the system in continuous environments [33] it is necessary to adapt the system. Furthermore, our proposed Transformer has been analysed on gait biometrics. Therefore, future work will be oriented towards analysing the potential of the proposed Transformer architecture for other behavioural biometric modalities such as handwritten signature [34,35], electrocardiograms [36], and keystroke [19]. In addition, privacy aspects of mobile authentication have not been considered yet [37], being still a problem to address. Therefore, future work will be oriented to improve both authentication and privacy at the same time [38].

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

The data is already public

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) and HumanCAIC (TED2021-131787B-I00 MICINN), and from Comunidad de Madrid (ELLIS Unit Madrid).

References

[1] A.K. Jain, K. Nandakumar, A. Ross, 50 years of biometric research: accomplishments, challenges, and opportunities, Pattern Recognit. Lett. 79 (2016) 80–105.
[2] L. Wang, T. Tan, H. Ning, W. Hu, Silhouette analysis-based gait recognition for human identification, IEEE Trans. Pattern Anal. Mach. Intell. 25 (12) (2003) 1505–1518.
[3] J.P. Singh, S. Jain, S. Arora, U.P. Singh, Vision-based gait recognition: a survey, IEEE Access 6 (2018) 70497–70527.
[4] M.D. Marsico, A. Mecca, A survey on gait recognition via wearable sensors, ACM Comput. Surv. 52 (4) (2019) 1–39.
[5] A. Sepas-Moghaddam, A. Etemad, Deep gait recognition: a survey, IEEE Trans. Pattern Anal. Mach. Intell. 45 (1) (2022) 264–284.
[6] C. Filipi Gonçalves dos Santos, D.d.S. Oliveira, L.A. Passos, R. Gonçalves Pires, D. Felipe Silva Santos, L. Pascotti Valem, T.P. Moreira, M.C.S. Santana, M. Roder, J. Paulo Papa, et al., Gait recognition based on deep learning: a survey, ACM Comput. Surv. 55 (2) (2022) 1–34.
[7] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, in: Proc. Advances in Neural Information Processing Systems, vol. 30, 2017.
[8] D. Hutchins, I. Schlag, Y. Wu, E. Dyer, B. Neyshabur, Block-recurrent transformers, in: Proc. Advances in Neural Information Processing Systems, 2022.
[9] Y. Tay, M. Dehghani, D. Bahri, D. Metzler, Efficient transformers: a survey, ACM Comput. Surv. 55 (6) (2022).
[10] H. Wu, J. Xu, J. Wang, M. Long, Autoformer: decomposition transformers with auto-correlation for long-term series forecasting, in: Proc. Advances in Neural Information Processing Systems, 2021.
[11] H. Zhou, S. Zhang, J. Peng, S. Zhang, J. Li, H. Xiong, W. Zhang, Informer: beyond efficient transformer for long sequence time-series forecasting, in: Proc. AAAI Conference on Artificial Intelligence, 2021.
[12] N. Zhang, J. Wang, Z. Hong, C. Zhao, X. Qu, J. Xiao, DT-SV: a transformer-based time-domain approach for speaker verification, in: Proc. International Joint Conference on Neural Networks, IEEE, 2022, pp. 1–7.
[13] B. Li, W. Cui, W. Wang, L. Zhang, Z. Chen, M. Wu, Two-stream convolution augmented transformer for human activity recognition, in: Proc. AAAI Conference on Artificial Intelligence, 2021.
[14] Q. Wen, T. Zhou, C. Zhang, W. Chen, Z. Ma, J. Yan, L. Sun, Transformers in time series: a survey, arXiv preprint arXiv:2202.07125 (2022).
[15] Q. Zou, Y. Wang, Q. Wang, Y. Zhao, Q. Li, Deep learning-based gait recognition using smartphones in the wild, IEEE Trans. Inf. Forensics Secur. 15 (2020) 3197–3212.
[16] H. Iwama, M. Okumura, Y. Makihara, Y. Yagi, The OU-ISIR gait database comprising the large population dataset and performance evaluation of gait recognition, IEEE Trans. Inf. Forensics Secur. 7 (5) (2012) 1511–1521.
[17] T.T. Ngo, Y. Makihara, H. Nagahara, Y. Mukaigawa, Y. Yagi, The largest inertial sensor-based gait database and performance evaluation of gait-based personal authentication, Pattern Recognit. 47 (1) (2014) 228–237.
[18] R. Tolosana, R. Vera-Rodriguez, J. Fierrez, J. Ortega-Garcia, BioTouchPass2: touchscreen password biometrics using time-aligned recurrent neural networks, IEEE Trans. Inf. Forensics Secur. 5 (2020) 2616–2628.
[19] S. Mondal, P. Bours, Person identification by keystroke dynamics using pairwise user coupling, IEEE Trans. Inf. Forensics Secur. 12 (6) (2017) 1319–1329.
[20] P. Melzi, R. Tolosana, A. Cecconi, A. Sanz-Garcia, G. Ortega, L. Jimenez-Borreguero, R. Vera-Rodriguez, Analyzing artificial intelligence systems for the prediction of atrial fibrillation from sinus-rhythm ECGs including demographics and feature visualization, Sci. Rep. 11 (2021).
[21] Y. Sun, J. Tang, X. Shu, Z. Sun, M. Tistarelli, Facial age synthesis with label distribution-guided generative adversarial network, IEEE Trans. Inf. Forensics Secur. 15 (2020) 2679–2691.
[22] R. Tolosana, P. Delgado-Santos, A. Perez-Uribe, R. Vera-Rodriguez, J. Fierrez, A. Morales, DeepWriteSYN: on-line handwriting synthesis via deep short-term representations, in: Proc. AAAI Conference on Artificial Intelligence, 2021.

[23] M. Gadaleta, M. Rossi, IDNet: smartphone-based gait recognition with convolutional neural networks, Pattern Recognit. 74 (2018) 25–37.

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

[25] Y. Watanabe, M. Kimura, Gait identification and authentication using LSTM based on 3-axis accelerations of smartphone, Procedia Comput. Sci. 176 (2020) 3873–3880.

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

[27] F.J. Ordóñez, D. Roggen, Deep convolutional and LSTM recurrent neural networks for multimodal wearable activity recognition, Sensors 16 (1) (2016) 115.

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

[29] R. Chavarriaga, H. Sagha, A. Calatroni, S.T. Digumarti, G. Tröster, J.d.R. Millán, D. Roggen, The opportunity challenge: a benchmark database for on-body sensor-based activity recognition, Pattern Recognit. Lett. 34 (15) (2013) 2033–2042.

[30] R. Child, S. Gray, A. Radford, I. Sutskever, Generating Long Sequences with Sparse Transformers, URL https://www.openai.com/blog/sparse-transformers (2019).

[31] S. Li, X. Jin, Y. Xuan, X. Zhou, W. Chen, Y.-X. Wang, X. Yan, Enhancing the locality and breaking the memory bottleneck of transformer on time series forecasting, Adv. Neural Inf. Process. Syst. (2019) 5243–5253 471.

[32] I. Beltagy, M.E. Peters, A. Cohan, Longformer: The Long-Document Transformer, arXiv preprint arXiv:2004.05150 (2020).

[33] I. Papavasileiou, Z. Qiao, C. Zhang, W. Zhang, J. Bi, S. Han, GaitCode: gait-based continuous authentication using multimodal learning and wearable sensors, Smart Health 19 (2021) 100162.

[34] R. Tolosana, R. Vera-Rodriguez, et al., SVC-onGoing: signature verification competition, Pattern Recognit. 127 (2022) 1–14.

[35] R. Tolosana, R. Vera-Rodriguez, J. Fierrez, J. Ortega-Garcia, DeepSign: deep on-line signature verification, IEEE Trans. Biom., Behav., Identity Sci. 3 (2) (2021) 229–239.

[36] P. Melzi, R. Tolosana, R. Vera-Rodriguez, ECG biometric recognition: review, system proposal, and benchmark evaluation, IEEE Access 11 (2023) 15555–15566.

[37] P. Delgado-Santos, G. Stragapede, R. Tolosana, R. Guest, F. Deravi, R. Vera-Rodriguez, A survey of privacy vulnerabilities of mobile device sensors, ACM Comput. Surv. 54 (11) (2022) 1–30.

[38] P. Delgado-Santos, R. Tolosana, R. Guest, R. Vera, F. Deravi, A. Morales, GaitPrivacyON: privacy-preserving mobile gait biometrics using unsupervised learning, Pattern Recognit. Lett. 161 (2022) 30–37.

Paula Delgado-Santos received the M.Sc. degree in Telecommunications Engineering from Universidad Autonoma de Madrid, Spain, in 2020. At the same time, she was working in a scholarship of IBM. In 2019/2020 she was working at a Swiss University, HEIG-VD, as a Data Scientist. In 2020 she began her Ph.D. studies with a Marie Curie Fellowship within the PriMa (Privacy Matters) EU project, supervised by Doctor Ruben Tolosana (Universidad Autonoma de Madrid) and Professor Richard Guest (University of Kent). She will study the richness of background sensor data elements obtained from mobile devices in a continuous authentication scenario. Her research interests include signal and image processing, pattern recognition, machine learning, biometrics and data protection.

Ruben Tolosana received the M.Sc. degree in Telecommunication Engineering, and his Ph.D. degree in Computer and Telecommunication Engineering, from Universidad Autonoma de Madrid, in 2014 and 2019, respectively. In 2014, he joined the Biometrics and Data Pattern Analytics - BiDA Lab at the Universidad Autonoma de Madrid, where he is currently collaborating as an Assistant Professor. Since then, Ruben has been granted with several awards such as the FPU research fellowship from Spanish MECD (2015), and the European Biometrics Industry Award (2018). His research interests are mainly focused on signal and image processing, pattern recognition, and machine learning, particularly in the areas of DeepFakes, HCI, and Biometrics. He is author of several publications and also collaborates as a reviewer in high-impact conferences (WACV, ICPR, ICDAR, IJCB, etc.) and journals (IEEE TPAMI, TCYB, TIFS, TIP, ACM CSUR, etc.). Finally, he is also actively involved in several National and European projects.

Richard Guest obtained his Ph.D. in 2000. He is Professor of Biometric Systems Engineering and Head of the School of Engineering at the University of Kent. His research interests lie broadly within biometric and forensic systems, particularly in the areas of image and behavioural information analysis, standardisation and mobile systems.

Farzin Deravi received the B.A. degree in Engineering Science and Economics from the University of Oxford, U.K., in 1981, the M.Sc. degree in Communications Engineering from Imperial College, U.K., in 1982, and the Ph.D. degree in Electronic Engineering from the University of Wales, Swansea, U.K., in 1988. He is currently with the School of Engineering and Digital Arts, University of Kent, Canterbury, U.K., where he is the Emeritus Professor of Information Engineering. His current research interests include the fields of pattern recognition and signal processing and their application in security and healthcare.

Ruben Vera-Rodriguez received the M.Sc. degree in telecommunications engineering from Universidad de Sevilla, Spain, in 2006, and the Ph.D. degree in electrical and electronic engineering from Swansea University, U.K., in 2010. Since 2010, he has been affiliated with the Biometric Recognition Group, Universidad Autonoma de Madrid, Spain, where he is currently an Associate Professor since 2018. His research interests include signal and image processing, pattern recognition, machine learning, and biometrics. He is the author of more than 150 scientific articles published in international journals and conferences, and 3 patents. He is actively involved in several National and European projects focused on biometrics. He has served as Program Chair for some international conferences such as: IEEE ICCST 2017, CIARP 2018, ICBEA 2019 and AVSS 2022.