



Repositorio Institucional de la Universidad Autónoma de Madrid

<https://repositorio.uam.es>

Esta es la **versión de autor** del artículo publicado en:
This is an **author produced version** of a paper published in:

2023 IEEE 17th International Conference on Automatic Face and
Gesture Recognition (FG). IEEE, 2023. 1-6

DOI: <https://doi.org/10.1109/FG57933.2023.10042710>

Copyright: © 2023 IEEE

El acceso a la versión del editor puede requerir la suscripción del recurso

Access to the published version may require subscription

“Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.”

Mobile Keystroke Biometrics Using Transformers

Giuseppe Stragapede¹, Paula Delgado-Santos^{2,1}, Ruben Tolosana¹, Ruben Vera-Rodriguez¹,
Richard Guest², Aythami Morales¹

¹ Biometrics and Data Pattern Analytics Lab, Universidad Autonoma de Madrid, Spain

² School of Engineering, University of Kent, UK

Abstract—Among user authentication methods, behavioural biometrics has proven to be effective against identity theft as well as user-friendly and unobtrusive. One of the most popular traits in the literature is keystroke dynamics due to the large deployment of computers and mobile devices in our society. This paper focuses on improving keystroke biometric systems on the free-text scenario. This scenario is characterised as very challenging due to the uncontrolled text conditions, the influence of the user’s emotional and physical state, and the in-use application. To overcome these drawbacks, methods based on deep learning such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) have been proposed in the literature, outperforming traditional machine learning methods. However, these architectures still have aspects that need to be reviewed and improved. To the best of our knowledge, this is the first study that proposes keystroke biometric systems based on Transformers. The proposed Transformer architecture has achieved Equal Error Rate (EER) values of 3.84% in the popular Aalto mobile keystroke database using only 5 enrolment sessions, outperforming by a large margin other state-of-the-art approaches in the literature.

I. INTRODUCTION

Due to the increasing number of online transactions and fraudulent activities in sectors such as Banking, Financial Services and Insurance (BFSI), healthcare, e-commerce, and government, among many others, the demand and the investments for more secure and reliable digital authentication methods are rising [1]. Such trend is particularly relevant with regard to mobile devices, given their popularity.

Recent authentication methods propose to increase the security through an additional transparent layer based on the user’s behavioural biometric information¹, overcoming potential identity theft in a user-friendly and continuous way [2], [14]. Among the different behavioural biometric traits, keystroke dynamics is one of the most popular authentication methods in the literature [4], [10]. The information considered is the timestamp of the actions of pressing and releasing a key, together with the information of the key typed.

Keystroke biometric systems are typically divided into two groups [12]: *fixed-text*, where the keystroke sequence typed by the user is prefixed, such as a username or password, and *free-text*, where the keystroke sequence is arbitrary, such as writing an email. In the latter, typing errors are common, and the keystroke sequences between the enrolment

and test samples are different, contrary to fixed-test scenarios. Consequently, the performance achieved with free-text keystroke systems are traditionally lower than their fixed-text counterparts due to the higher intra-subject variability and complexity of the task.

In addition, focusing on keystroke biometrics on mobile scenarios, many challenges must be considered to develop robust authentication systems. In this particular scenario, keystroke is typically acquired under uncontrolled circumstances, which can be affected by the user’s activity, body position, emotional state, and the acquisition device [10], [17]. The performance might also be affected if the same subject is able to speak different languages [3].

In the present work, we explore and propose a Transformer architecture to overcome the challenges commented before and improve the authentication performance of free-text keystroke biometric systems on mobile scenarios. Originally proposed in [20], Transformers are defined by an encoder-decoder architecture. They have quickly gained the attention of the scientific community given their ability to model a number of different processes, in fields such as computer vision, machine translation, reinforcement learning, time-series analysis for classification and prediction, etc. [16]. These new architectures have shown several advantages compared with Convolutional Neural Network (CNNs) and Recurrent Neural Networks (RNNs): (i) they process all sequences in parallel being feed-forward models, (ii) they operate over long-distance sequences applying self-attention mechanism, (iii) they undergo a more efficient training allowing the process of all samples in one batch, and (iv) they attend to all of the previous sequence at the same time without the need to summarise the seen information [20].

In summary, the main contributions of this work can be listed as follows:

- Novel keystroke verification system for the challenging free-text mobile scenario based on Transformers. Fig. 1 provides a graphical representation of the proposed Transformer architecture. To the best of our knowledge, this is the first study that explores Transformers for keystroke biometrics.
- Comparison of the proposed Transformer with previous approaches in the literature using the popular and public Aalto mobile keystroke database [13]. The proposed approach achieves Equal Error Rate (EER) values of 3.84% using only 5 enrolment sessions, outperforming by a large margin previous state-of-the-art approaches.

¹In contrast to *physiological* biometrics, which pertains to the biological characteristics of an individual, such as face or fingerprint, all means that enable or contribute to differentiating between individuals throughout the way they perform activities are labelled as *behavioural*, i.e., gait, keystroke dynamics, handwritten signature, etc.

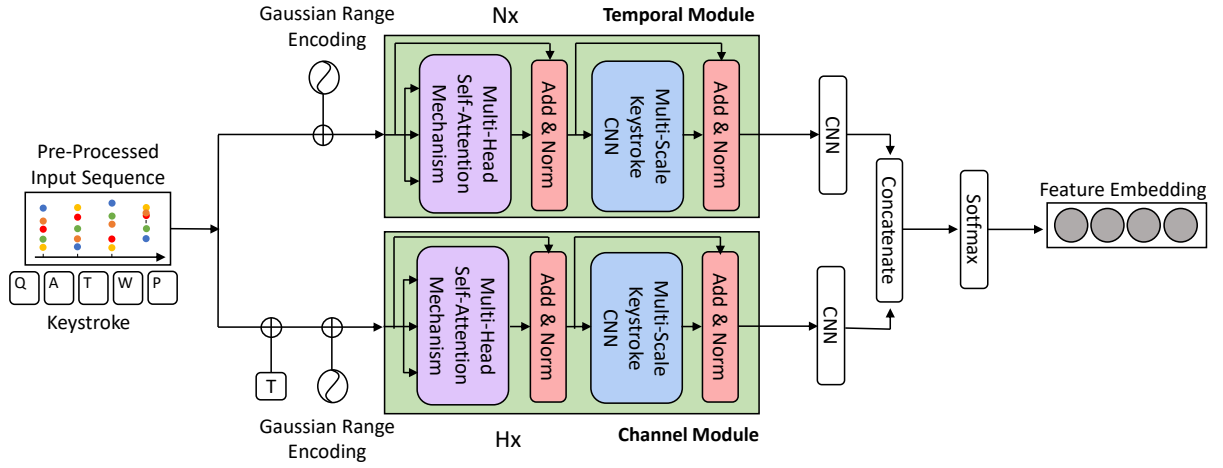


Fig. 1. Graphical representation of the proposed Transformer architecture. T: Transposition; N, H: Number of layers of each of the modules.

- We make our proposed approach and experimental framework available to the research community in order to advance the state of the art of keystroke biometrics in free-text mobile scenarios².

The remainder of the paper is organised as follows: Sec. II summarises the Aalto mobile keystroke database. Sec. III describes the feature extraction process and the proposed Transformer architecture. Then, in Sec. IV, we present a detailed description of the experimental setup. Sec. V contains the experimental results of the proposed approach, and the comparison with the state of the art. Finally, Sec. VI draws some conclusions and future research lines.

II. THE AALTO MOBILE KEYSTROKE DATABASE

The Aalto mobile keystroke database comprises free-text keystroke dynamics data from around 260,000 subjects [13]. A mobile web application was implemented for the data acquisition, in a totally unsupervised way. The subjects were asked to read English sentences, and to type them as rapidly and accurately as possible in their own smartphones. The provided sentences were randomly withdrawn from a set of 1,525 sentences obtained from the Enron mobile mail [21] and the Gigaword Newswire corpora [7]. The requisite for each of the sentences was containing at least 3 words and at most 70 characters. Around 68% of the participating subjects were English native speakers. The raw data recorded consists in the acquisition of key press and key release events from the browser, with the resolution of 1ms. In the present work, we select all subjects (62,454) that completed at least 15 acquisition sessions.

III. PROPOSED SYSTEM

This section provides the details of the proposed keystroke verification system for free-text mobile scenarios.

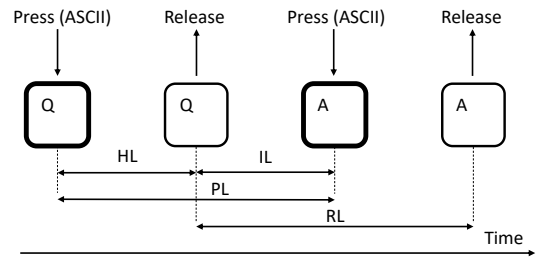


Fig. 2. Example of the keystroke features extracted from the Aalto mobile keystroke database [13]. HL: Hold Latency; IL: Inter-key Latency; PL: Press Latency; RL: Release Latency; ASCII: Key Pressed.

A. Feature Extraction

The raw data are pre-processed following the approach described in [4]. Data consist in the timestamp of the action of pressing and releasing a key, together with the ASCII code typed. The set of 5 features reported below are extracted per each press-release action:

[hold latency, inter-key latency, press latency, release latency, key pressed]

The five considered features are illustrated in Fig. 2. Since the length of the text considered in each of the acquisition sessions is not constant (free-text scenario), they are sliced or zero-padded to obtain a sequence of $L = 50$ samples, aiming to minimise the system input duration. The ASCII code (key pressed) is normalised in the $[0, 1]$ interval.

B. Transformer Architecture

Fig. 1 provides a graphical representation of the proposed Transformer, based on an adaptation of the encoder part of the Vanilla Transformer [20]. The Vanilla Transformer was tested in several fields showing impressive results but needs some adaptations in order to be used in time sequences. Several researchers introduced new aspects such as reduced complexity, periodicity-based dependencies, or time-depending encoding [16]. We describe next the key aspects of our proposed Transformer.

²<https://github.com/BiDALab/TypeFormer>

The pre-processed input sequence $\mathbf{X} = (\mathbf{x}_0, \mathbf{x}_1, \dots, \mathbf{x}_L, \dots, \mathbf{x}_L)$ is introduced into the Transformer model. Adopting the idea presented in [8], we have first changed the original positional encoding by a *Gaussian range encoding*. Fig. 3 provides a graphical representation of the Gaussian range encoding. The pre-processed input sequence is modelled with G Gaussian distributions where the Probability Density Function (PDF) vector is a L1-normalized vector of the Gaussian PDFs. Furthermore, more than one range can be used at the same time, obtaining a more complex context position of each sample compared with the original positional encoding. The global Gaussian range encoding is the pondered multiplication of the PDF vector in the different ranges. It is important to highlight that, contrary to [8], we have considered the Gaussian range encoding in both branches of the Transformer (Temporal and Channel Modules).

After the Gaussian range encoding, the proposed Transformer changes the original layer considered in the Vanilla Transformer by the two different modules considered in [8]: (i) Temporal Module, and (ii) Channel Module. The Temporal Module extracts information from the original input sequence (temporal-over-channel features), while the Channel Module transposes the input sequence to extract channel-over-temporal features. The Temporal Module contains a stack of N identical layers while the Channel Module contains a stack of H identical layers. Each module comprises two sub-layers: (i) a multi-head self-attention mechanism, and (ii) a multi-scale keystroke CNN. Then, each sub-layer is followed by a residual connection and a layer normalisation (Add & Norm in Fig. 1). We provide next the essential details of the *multi-head self-attention mechanism* and the *multi-scale keystroke CNN* sub-layers.

The *multi-head self-attention mechanism* is responsible for linking each of the samples along the entire input sequence. The procedure extracts long-range dependencies without limiting the time window size. The output of the sub-layer is the weighted summation of the values V in accordance with the dot-product of the queries Q and the matching keys K [20]. The output of the sub-layer is the concatenation of applying the attention mechanism to a F independent heads. The *multi-scale keystroke CNN* comprises convolutional layers with ReLU activation and different kernel sizes. A batch normalisation and a dropout layer are introduced in between.

A convolutional block is placed after each module. The CNN features are then concatenated and introduced to a softmax layer. The feature embedding obtained is $\mathbf{P} = (\mathbf{p}_0, \mathbf{p}_1, \dots, \mathbf{p}_s, \dots, \mathbf{p}_S)$, where S is the number of output features. Finally, for the verification task considered in the present study, the feature embeddings of the enrolment and test samples are compared using the Euclidean distance. It is important to highlight that the architecture configuration and model hyperparameters of the proposed Transformer have been adapted to free-text keystroke verification systems on mobile devices. The specific details of the proposed Transformer are described in Sec. IV-A.

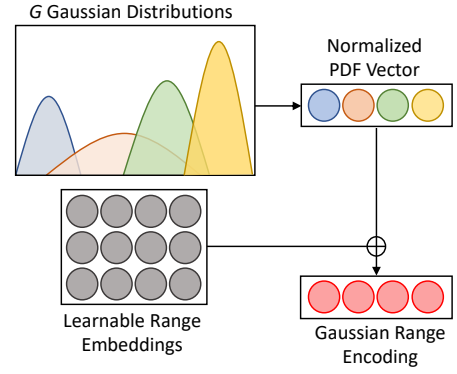


Fig. 3. Graphical representation of the Gaussian range encoding. PDF: Probability Density Function.

IV. EXPERIMENTAL SETUP

A. Transformer Hyperparameters

This section describes the optimal hyperparameters of the proposed Transformer. These hyperparameters have been selected using the development experimental protocol described in the following section, Sec. IV-B.

The Gaussian range encoding relies on $G = 20$ Gaussian distributions. The Temporal Module comprises $N = 10$ identical layers whereas the Channel Module comprises $H = 1$ layer. Regarding the multi-head self-attention sub-layer, $F = 10, 5$ heads are considered respectively for the Temporal and Channel Modules. In both modules, the multi-scale keystroke CNN comprises 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. Then, after the Temporal and Channel Modules, we consider 2 convolutional layers (L units each, ReLU activation functions, and kernel sizes 128 and 32 respectively, followed by dropout layers with a rate of 0.5) with max-pooling, and a linear layer with softmax activation function. The size of the final feature embedding is $S = 64$.

B. Model Development

We follow the public experimental protocol presented by Acien *et al.* in [4], considering the same 30,000 subjects for training the models and 400 for validation. In total, 15 sessions per subject are considered. The proposed Transformer is implemented in PyTorch. Its training relies on a triplet loss function based on Euclidean distance with a margin $\alpha = 1.0$. Adam optimiser with default parameters and learning rate value of 0.001 is considered. We train the Transformer for 1,000 epochs in total, considering 29 batches sized 1024 per epoch, i.e., 29,696 triplets. The selection of triplets takes place randomly with a uniform distribution. At the end of each epoch, the model is evaluated on the entire validation set, and when achieving a lower EER value, the corresponding model is saved. Fig. 4 provides a graphical representation of the training/validation results of the proposed Transformer along the number of epochs. In general, we can observe a smooth training curve in time.

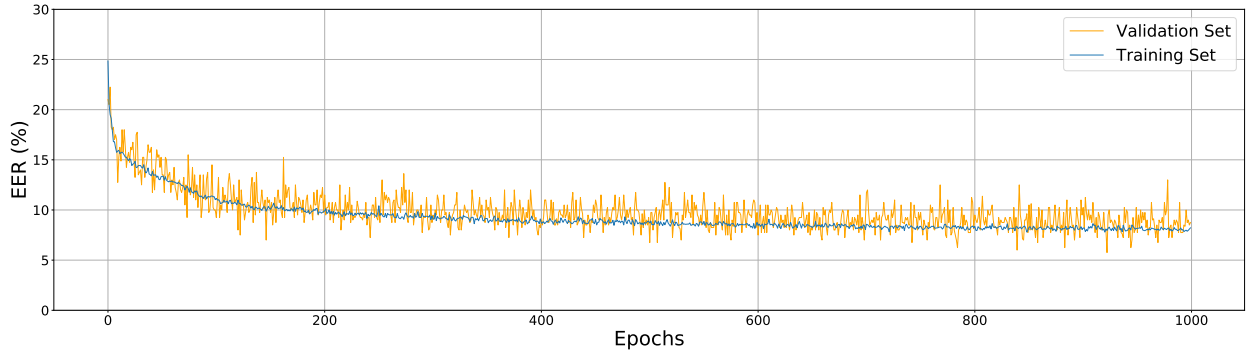


Fig. 4. The EERs [%] achieved on the training and validation sets at the end of each epoch of the training process are displayed above.

C. Model Evaluation

The best model selected in the development stage is finally evaluated using the public experimental protocol considered in [4]. This evaluation consists of 1,000 unseen subjects, not considered in the development stage. In addition, from the total 15 sessions available per subject, we perform experiments using different configurations of enrolment sessions ($E = 1, 5, 10$), similar to [4], in order to assess the system adaptation to reduced availability of enrolment data. To obtain the genuine score distribution, for each user, we always consider the last 5 sessions for testing. Each of them is compared with the E enrolment sessions obtaining $5 \times E$ scores, which are averaged over the E enrolment sessions, leading to 5 final genuine scores per user. Regarding the impostor score distribution, we follow the same approach, but this time each user is compared with one test session of the remaining users, leading to 999 impostor scores per user. Furthermore, it is important to highlight that from here two different approaches can be adopted to compute the EER: (i) selecting an individual threshold for each of the enrolled subjects obtaining an EER per subject and then obtaining their average value as the final EER value; (ii) selecting a unique threshold for all subjects. Their use depends on the particular application and scenario. The advantage of (i) consists in a better performance in terms of EER, as a specific threshold of the system is adapted to each subject. This is the approach adopted in the comparison work [4]. However, the drawback is related to the fact that a large impostor embedding distribution has to be compared with each of the individual subjects' feature embedding to tune each specific threshold. Such solution might be in conflict with the mobile environment resource constraints in terms of memory, or with privacy and security concerns. In fact, it has been shown that a significant amount of information can be obtained from data of mobile devices such as keystroke [6]. On the other hand, if the system is trained *offline* on a large database, deploying the system with a fixed pre-determined unique threshold (ii) is undoubtedly more convenient due to the described aspects. In the present work, both scenarios are considered in the experimental framework, naming them respectively "Average" EER for case (i), and "Global" EER for case (ii).

TABLE I
SYSTEM PERFORMANCE COMPARISON BETWEEN THE
STATE-OF-THE-ART TYPENET [4] AND THE PROPOSED TRANSFORMER.

Number of Enrolment Sessions	Average EER (%)		Global EER (%)	
	TypeNet [4]	Proposed Transformer	TypeNet [4]	Proposed Transformer
1	12.60	6.99	18.19	10.68
5	9.20	3.84	14.39	7.23
10	8.00	3.15	13.16	6.26

V. EXPERIMENTAL RESULTS

A. Comparison with the state of the art

Table I provides the system performance results in terms of EER (%) of the proposed Transformer for the different number of enrolment sessions ($E = 1, 5, 10$) and the two different threshold configurations, i.e., computing an individual threshold per subject and obtaining the mean EER over all subjects (Average EER), and a unique threshold for all subjects (Global EER). In addition, to provide a better comparison of the proposed Transformer with recent state-of-the-art keystroke biometric systems, we include the results achieved by TypeNet [4]. TypeNet is based on a Long Short-Term Memory (LSTM) RNN architecture, achieving state-of-the-art performance results in both physical and touchscreen keyboards. Several learning approaches were also studied using different loss functions (softmax, contrastive, and triplet loss). It is important to highlight that we opt for the comparison with TypeNet as (i) they considered one of the largest mobile free-text keystroke databases available up to date, the Aalto mobile keystroke database [13], (ii) their experimental protocol is publicly available in GitHub³, so we can rigorously follow it, considering the same sets of subjects and metrics, for development and evaluation, and (iii) TypeNet has outperformed previous approaches in keystroke biometrics.

As can be seen in Table I, the proposed Transformer significantly outperforms TypeNet in all cases on the same evaluation set of 1,000 subjects from the Aalto mobile keystroke database [13]. Analysing the number of enrolment sessions, the proposed Transformer achieves a 6.99%

³<https://github.com/BiDALab/TypeNet>

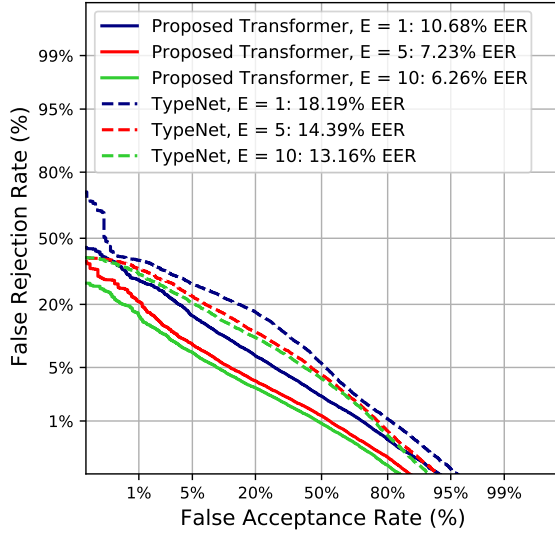


Fig. 5. DET curves comparing the performance of the proposed Transformer with TypeNet [4]. E corresponds to the number of enrolment sessions considered. The reported EERs (%) are for the global threshold.

EER when considering $E = 1$ single enrolment session. This is an absolute improvement of 5.61% EER compared with TypeNet (12.60% EER), proving the high potential of the proposed Transformer compared with traditional deep learning architectures such as RNNs. Increasing the number of enrolment sessions per subject, we can see a general improvement of the proposed Transformer, with values of 3.84% and 3.15% EERs for $E = 5, 10$ enrolment sessions, respectively. This trend also shows the large improvement of the proposed Transformer with the number of enrolment sessions, outperforming TypeNet in both scenarios (absolute improvement of around 5% EER).

Analysing the two different threshold scenarios (Average and Global), we can observe better results for the Average case regardless of the number of enrolment sessions, e.g., for $E = 1$ single enrolment session, values of 6.99% and 10.68% EERs are achieved by the Proposed Transformer for the Average and Global cases, respectively. A similar trend is observed for TypeNet. These results make sense as one specific threshold is adapted to each subject in the Average case, at the expense of more computational efforts.

For completeness, we include in Fig. 5 the Detection Error Trade-off (DET) curves computed for the different number of enrolment sessions available for the Global EER threshold case. As can be seen, the proposed Transformer outperforms TypeNet even in the case of having fewer enrolment sessions available, i.e., $E = 1$ enrolment session (proposed Transformer achieves 10.68% EER) vs. $E = 10$ (TypeNet achieves 13.16% EER).

Finally, to provide a better comparison of the proposed Transformer with the literature, we include in Table II the EER results obtained by other state-of-the-art systems in keystroke biometrics: digraphs and SVM [5], Partially Observable Hidden Markov Model (POHMM) [11], and a combination of RNNs and CNNs [9]. All of them are

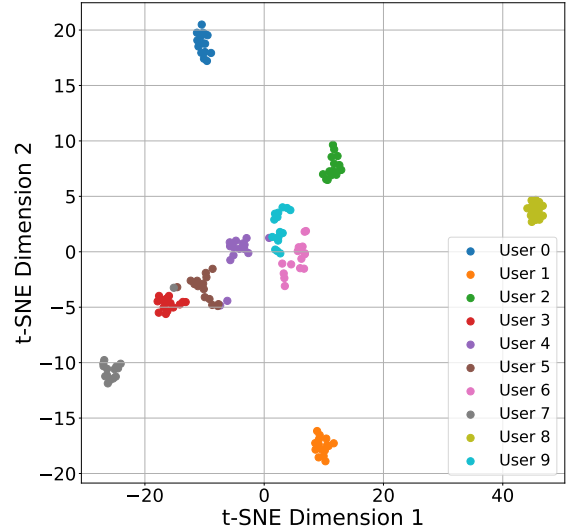


Fig. 6. 2D graphical visualisation of the latent space through t-SNE considering 15 sessions of 10 subjects [19]. Selected parameters⁵: perplexity = 14, init = 'pca', n_iter = 1000.

TABLE II
COMPARISON OF THE PERFORMANCE ACHIEVED BY THE PROPOSED TRANSFORMER WITH RELATED SYSTEMS ($E = 5$).

System	Average EER (%)
POHMM [11]	40.40
Digraphs [5]	29.20
CNN+RNN [9]	12.20
TypeNet [4]	9.20
Proposed Transformer	3.84

trained under the same experimental protocol and evaluated on the same set of 1,000 subjects in terms of Average EER considering $E = 5$ enrolment sessions. Our proposed Transformer outperforms all previous approaches with EER absolute improvements of 36.56% (POHMM [11]), 25.36% (Digraphs [5]), 8.36% (CNN + RNN [9]), and 5.36% (TypeNet [4]). These results evidence the success and potential of the proposed Transformer for the challenging free-text mobile scenario considered in the study.

B. Analysis of the feature embeddings

Fig. 6 provides a graphical representation of the feature embedding space achieved with the proposed Transformer for 10 different subjects of the Aalto mobile keystroke database (15 acquisition sessions per subject). We consider the popular mathematical method t-SNE [19] to visualise data points in high dimensional spaces. Apart from few outliers (one sample of users 4 and 7), most of the embeddings of each of the subjects are clearly separated. Fig. 6 demonstrates how the proposed Transformer is able to group clearly the feature embeddings belonging to the same subject, achieving small intra-class variability, and to distance as much as possible between the feature embeddings of the different subjects, increasing the inter-class variability.

⁵`sklearn.manifold.TSNE -- scikit-learn 1.1.1 documentation`. Accessed: 2022-07-13.

C. Discussion

The system performance improvement achieved with our proposed Transformer in relation with previous approaches is due, in our opinion, to the following reasons: (i) our model applies the self-attention mechanism, being able to operate over long-distances in the input sequence; (ii) our model attends to all the prior samples of the time sequence at the same time, without summarising the previous seen information; (iii) the features are extracted from two different perspectives, from the time and the channel modules, providing more complex information; and (iv) the Gaussian range encoding together with the multi-scale keystroke CNN allow to obtain a perspective of each sample in different environments, as different ranges are treated at the same time.

VI. CONCLUSIONS AND FUTURE WORK

The present study has explored and proposed novel keystroke verification systems based on Transformers. To the best of our knowledge, this is the first attempt to apply Transformers to keystroke biometrics. We have focused on the task of free-text mobile keystroke authentication, traditionally far more challenging than its fixed-text counterpart. Our proposed Transformer has greatly reduced the performance gap existing between the two scenarios, reaching numbers as low as 3.15% EER with 10 short enrolment sessions of 50 samples each, and 3.93% EER with only 1 enrolment session. Furthermore, for the popular and public Aalto mobile keystroke database considered in the study, the proposed Transformer has achieved remarkable improvements with the same experimental protocol considered in the recently state-of-the-art TypeNet [4] (3.15% EER vs. 8.00% EER). Finally, it is important to remark that we will make our proposed approach and experimental framework available to the research community in order to advance the state of the art of keystroke biometrics in free-text mobile scenarios⁶.

Future work will be oriented in several directions: (i) improvement of the Transformer architecture; (ii) an optimised training approach considering hard triplet mining. Forcing the model to learn from harder comparisons has in fact proved to be an effective strategy in many applications [15]; (iii) a more sophisticated mechanism than the traditional Euclidean distance for the comparison of feature embeddings in the latent space, such as Support Vector Machines (SVM); (iv) investigating the subject information contained in the feature embeddings, i.e., gender, age, etc., to determine if keystroke data should be treated as privacy-sensitive biometric data. The metadata available in the Aalto mobile keystroke database can be used to shed some light; and (v) applying Transformers to other biometric modalities [18].

VII. 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. R. Tolosana and R. Vera-Rodriguez are also supported by INTER-ACTION (PID2021-126521OB-I00 MICINN/FEDER).

REFERENCES

- [1] Behavioral Biometrics Market Size & Share Report, 2020-2027. <https://www.grandviewresearch.com/industry-analysis/behavioral-biometric-market>. Accessed: 2022-07-11.
- [2] ISO 9241-11:2018(en): *Ergonomics of Human-System Interaction*, 2018. Part 11: Usability: Definitions and Concepts.
- [3] M. Abuhamad, A. Abusnaina, D. Nyang, and D. Mohaisen. Sensor-Based Continuous Authentication of Smartphones' Users Using Behavioral Biometrics: A Contemporary Survey. *IEEE Internet of Things Journal*, 2021.
- [4] A. Acien, A. Morales, J. V. Monaco, R. Vera-Rodriguez, and J. Fierrez. TypeNet: Deep Learning Keystroke Biometrics. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 2021.
- [5] H. Çeker and S. Upadhyaya. User Authentication with Keystroke Dynamics in Long-text Data. In *Proc. IEEE Int'l Conf. on Biometrics Theory, Applications and Systems*, 2016.
- [6] P. Delgado-Santos, G. Stragapede, R. Tolosana, R. Guest, F. Deravi, and R. Vera-Rodriguez. A Survey of Privacy Vulnerabilities of Mobile Device Sensors. *ACM Comput. Surv.*, 2022.
- [7] D. Graff and C. Cieri. English Gigaword LDC2003T05. *Philadelphia: Linguistic Data Consortium*, 2003.
- [8] B. Li, W. Cui, W. Wang, L. Zhang, Z. Chen, and M. Wu. Two-stream Convolution Augmented Transformer for Human Activity Recognition. In *Proc. AAAI Conf. on Artificial Intelligence*, 2021.
- [9] X. Lu, S. Zhang, P. Hui, and P. Lio. Continuous Authentication by Free-text Keystroke based on CNN and RNN. *Computers & Security*, 2020.
- [10] E. Maiorana, H. Kalita, and P. Campisi. Mobile Keystroke Dynamics for Biometric Recognition: An Overview. *IET Biometrics*, 2021.
- [11] J. V. Monaco and C. C. Tappert. The Partially Observable Hidden Markov Model and its Application to Keystroke Dynamics. *Pattern Recognition*, 2018.
- [12] S. Mondal and P. Bours. A Study on Continuous Authentication Using a Combination of Keystroke and Mouse Biometrics. *Neurocomputing*, 2017.
- [13] K. Palin, A. M. Feit, S. Kim, P. O. Kristensson, and A. Oulasvirta. How Do People Type on Mobile Devices? Observations from a Study with 37,000 Volunteers. In *Proc. Int'l Conf. on Human-Computer Interaction with Mobile*, 2019.
- [14] V. M. Patel, R. Chellappa, D. Chandra, and B. Barbello. Continuous User Authentication on Mobile Devices: Recent Progress and Remaining Challenges. *IEEE Signal Processing Magazine*, 2016.
- [15] F. Schroff, D. Kalenichenko, and J. Philbin. FaceNet: A Unified Embedding for Face Recognition and Clustering. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2015.
- [16] Y. Tay, M. Dehghani, D. Bahri, and D. Metzler. Efficient Transformers: A Survey. *ACM Comput. Surv.*, 2022.
- [17] P. S. Teh, N. Zhang, A. B. J. Teoh, and K. Chen. A Survey on Touch Dynamics Authentication in Mobile Devices. *Computers & Security*, 2016.
- [18] R. Tolosana, R. Vera-Rodriguez, C. Gonzalez-Garcia, J. Fierrez, A. Morales, J. Ortega-Garcia, J. C. Ruiz-Garcia, S. Romero-Tapiador, S. Rengifo, M. Caruana, et al. SVC-onGoing: Signature Verification Competition. *Pattern Recognition*, 127:108609, 2022.
- [19] L. Van der Maaten and G. Hinton. Visualizing Data using t-SNE. *Journal of Machine Learning Research*, 2008.
- [20] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is All you Need. In *Proc. Advances in Neural Information Processing Systems*, 2017.
- [21] K. Vertanen and P. O. Kristensson. A Versatile Dataset for Text Entry Evaluations Based on Genuine Mobile Emails. In *Proc. Int'l Conf. on Human Computer Interaction with Mobile Devices and Services*, 2011.

⁶<https://github.com/BiDALab/TypeFormer>