Self-Supervised-based Multimodal Fusion for Active Biometric Verification on Mobile Devices

Youcef Ouadjer [1], Chiara Galdi[2], Sid-Ahmed Berrani[1,3], Mourad Adnane[1] and Jean-Luc Dugelay[2]
[1]*École Nationale Polytechnique, 10 rue des Frères Oudek, 16200 El Harrach, Algiers, Algeria*
[2]*Department of Digital Security, EURECOM, 450 Route des Chappes, 06410 Biot, France*
[3] *National School of Artificial Intelligence, Route de Mahelma, 16201 Sidi Abdellah, Algiers, Algeria*

Keywords: Active biometric verification, Multimodal fusion, Self-supervised learning.

Abstract: This paper focuses on the fusion of multimodal data for an effective active biometric verification on mobile devices. Our proposed Multimodal Fusion (MMFusion) framework combines hand movement data and touch screen interactions. Unlike conventional approaches that rely on annotated unimodal data for deep neural network training, our method makes use of contrastive self-supervised learning in order to extract powerful feature representations and to deal with the lack of labeled training data. The fusion is performed at the feature level, by combining information from hand movement data (collected using background sensors like accelerometer, gyroscope and magnetometer) and touch screen logs. Following the self- supervised learning protocol, MMFusion is pre-trained to capture similarities between hand movement sensor data and touch screen logs, effectively attracting similar pairs and repelling dissimilar ones. Extensive evaluations demonstrate its high performance on user verification across diverse tasks compared to unimodal alternatives trained using the SimCLR framework. Moreover, experiments in semi-supervised scenarios reveal the superiority of MMFusion with the best trade-off between sensitivity and specificity.

1 INTRODUCTION

Active biometric verification on mobile devices consists in verifying the identity of a user frequently (Stylios et al., 2021). In active biometric systems, samples are collected continuously, as an example: Face, motion gestures, walking, typing, and scrolling can be used for active biometric verification on mobile devices (Stragapede et al., 2023) (Fathy et al., 2015) (De Marsico et al., 2014). In a realistic scenario, active biometric systems can combine multiple modalities with fusion at different levels (Stragapede et al., 2022). Specifically, multimodal fusion can be achieved with touch screen interactions and hand movement. Mobile users produce touch screen interactions while moving the device with their hands. This typical way of mobile usage provides unique patterns from touch screen and motion sensors for active biometric verification (Sitová et al., 2016).

In the dynamic field of biometric research, deep neural networks have attracted considerable attention. Previous studies explored state-of-the-art architectures, including Convolutional Neural Networks (CNN) (Tolosana and Vera-Rodriguez, 2022) and Long Short-Term Memory Networks (LSTM) (Tolosana et al., 2018) (Delgado-Santos et al., 2022), focusing on behavioral biometric attributes. While these architectures have delivered valuable insights, there is a notable gap in the literature regarding the integration of multimodal data for active biometric verification. In (Zou et al., 2020), the authors demonstrated that Convolutional Recurrent Neural Netwrok (CNN-RNN) can extract robust feature representations using inertial sensors for active mobile verifcation. This work reported impressive performance, achieving accuracy rates of 93.5% for person identification and 93.7% for verification. Similary, Giorgi et al. (Giorgi et al., 2021) introduced a verification framewrok for mobile devices that leveraged gait patterns, combining inertial sensors with a recurrent neural network. Their approach demonstrated remarkable effectiveness in terms of biometric verification and real-time efficiency, substantiated through a series of practical experiments. Recently, Stragapede et al. (Stragapede et al., 2023) proposed a verification system that relies on an LSTM model and, importantly, incorporates modality fusion at the score level. They reported consistent results, with an Area Under the Curve (AUC) score ranging from 80% to 87% for random impostor verification and an AUC score ranging from 62% to 69% for skilled impostor verification. In a similar work (Stragapede

et al., 2022), the authors explored the fusion of touch screen and motion sensors for active biometric verification while users engaged in various tasks such as typing, scrolling, and tapping. Notably, they adopted a weighted fusion strategy at the score level using different sensors and demonstrated that the combination of keystroke touch data and magnetometer sensor data consistently yielded the most favorable results in terms of verification scores. Nevertheless, even though these studies have made significant progress in advancing active biometric verification, there is a lack of research examining the fusion of multimodal data. The unexplored potential of integrating multiple sensory inputs in enhancing the accuracy and reliability of biometric verification represents an important path for future research.

In mobile biometrics, data annotation challenges persist, primarily due to the sensitivity of biometric data. Self-supervised learning, such as contrastive pre-training, offers a solution by utilizing unlabeled data for pre-training and fine-tuning on labeled data. We propose a self-supervised multimodal fusion framework inspired from (Chen et al., 2020), uniquely capitalizing on the rich sources of touch screen data and sensor data. Leveraging unlabeled data from these modalities, we construct feature representations based on positive and negative pairs, allowing for the effective transfer of these representations in the context of user verification.

The proposed multimodal fusion (MMFusion) framework is trained on the Hand Movement Orientation and Grasp (HMOG) database (Sitová et al., 2016), and Human Interaction (HuMIdb) database (Acien et al., 2021). The evaluation is performed on different touch screen tasks such as: Scrolling, tapping and swipping. The performance of MMFusion is evaluated against the original contrastive learning (SimCLR) (Chen et al., 2020) using the hand movement modality and touch modality separately. Interestingly, the verification performance of our MMFusion approach outperforms the SimCLR approach with a significant margin. Moreover, the verification performance is assessed when the MMFusion is finetuned on different fraction of labels, demonstrating high performance using a small set of labeled examples. We summarize our contributions as follows:

- We propose a self-supervised multimodal fusion framework for active biometric verification by combining touch screen and hand movement data.
- Performance of MMfusion framework is evaluated on different touch screen tasks and compared to the SimCLR approach with touch screen and sensor modalities.
- Extensive experiements conducted on two benchmark databases, report high verification performance when both touch and hand movement modalities are combined, in addition to the scenario where the fusion model is fine-tuned on limited labeled data.

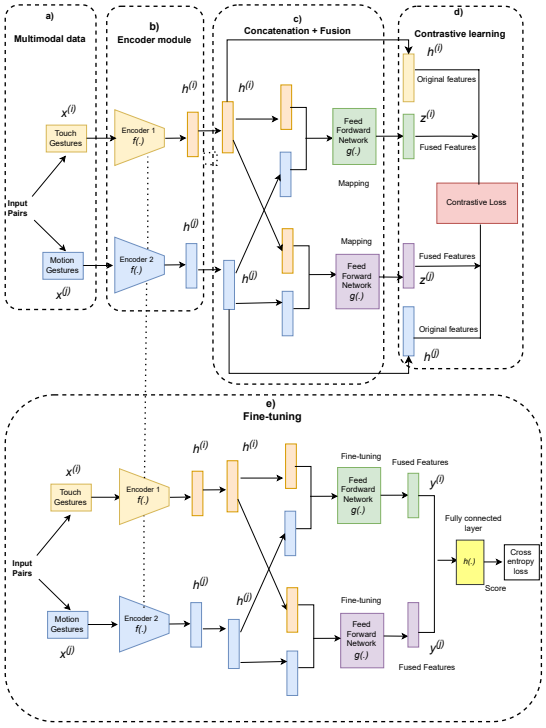


Figure 1: The proposed multimodal fusion framework. **a) Multimodal data**: Data is sampled from touch and motion gesture modalities. **b) Encoders**: the encoders produce features from motion and touch gesture inputs. **c) Fusion**: features produced by the encoders are passed to the feed-forward networks for fusion. **d) Contrastive learning**: both original and fused features are used for computing similarities and loss. **e) Fine-tuning**: the fully connected layer *h(.)* is trained with labeled data.

2 PROPOSED METHOD

In this section the proposed MMFusion approach based on contrastive learning is described, following different stages, from contrastive pre-training with multimodal data to fine-tuning on limited labeled training data.

2.1 MMFusion Framework

The proposed MMFusion is inspired from the work of SimCLR (Chen et al., 2020), initially introduced

for representation learning in images. But the research community followed-up with multiple variants of SimCLR, where different regularizations schemes, and input modalities were used to learn similarities from pairs of data (Jing and Tian, 2021). Figure 1 illustrates our solution, it is composed of five steps: multimodal data, DNN encoder, fusion module, contrastive loss and finally fine-tuning. Further details will be given about each component in the following subsections.

2.1.1 Multimodal Data

The data module represents the first part of multimodal contrastive learning, in the original work of SimCLR. (Chen et al., 2020) the authors proposed data augmentation to generate similar input pairs from a single modality, however in our case we simply extract pairs of input from different modalities i.e. touch screen and hand movement modality. Following Figure 1.(a), each input pair is composed of two elements (X^1, X^2) denoting the sensor and touch gesture vectors of a given mobile user. The input feature vector of hand movement modality S_q is a sequence composed of 3 inertial sensors, each one describing the acceleration (a_x, a_y, a_z), angular velocity (g_x, g_y, g_z) and ambient magnetic field (m_x, m_y, m_z) using 3 physical axes. It results in an input vector of 9 elements: S_q = $\{a_x, a_y, a_z, g_x, g_y, g_z, m_x, m_y, m_z\}$. Similarly the input vector for the touch modality is a sequence Tq = $\{x, y, p, s\}$ where x and y are the coordinates of the finger on the screen, while p and s denote the pressure and surface of the finger on the mobile screen. The trainig batch of positive pairs used during contrastive learning is then expressed as follows:

$$D_P = \{(X_1^{(1)}, X_1^{(2)}), (X_2^{(1)}, X_2^{(2)}), ..., (X_N^{(1)}, X_N^{(2)})\} \quad (1)$$

The negative pairs of contrastive learning are sampled following the equation2:

$$D_N = \{(X_i^{(1)}, X_k^{(2)})\}_{i \neq k}, \quad (2)$$

X_i and X_k are input feature vectors from two different mobile users.

2.1.2 Encoder Module

The encoder module represented in Figure 1.(b), takes the input feature vector of the sensor and touch screen data, performs a series of linear and non-linear transformations to map each feature vector (X^1, X^2) into a representation h^1 and h^2. The encoder module is based on convolutional recurrent neural network (Conv-RNN), proposed in (Tolosana and Vera-Rodriguez, 2022). The architecture of the encoder is based on two CNN layers with maxpooling in between, and two recurrent layers. Dropout is used after the second convolutional layer, and the first recurrent layer respectively. In the end, a linear layer is used to produce representations which are then passed to the fusion module.

$$\begin{aligned} h^{(1)} &= f(X^{(1)}), \\ h^{(2)} &= f(X^{(2)}) \end{aligned} \quad (3)$$

2.1.3 Multimodal Fusion Module

The multimodal fusion module is a feed-forward neural newtork with two hidden layers, described by the function *g(.)* in Figure 1.(c). This module takes as input concatenated features (h^1, h^2) from touch screen and sensor data, performs joint feature fusion with linear and non-linear operations to transform the features into representations z^1 and z^2.

$$\begin{aligned} z^{(1)} &= g(h^{(1)}, h^{(2)}), \\ z^{(2)} &= g(h^{(2)}, h^{(1)}) \end{aligned} \quad (4)$$

2.1.4 Contrastive Task

Contrastive learning represents the final part of the MMFusion module (Figure 1.(d)), the representations obtained after the fusion are used for contrastive learning, following equation 5:

$$l_{i,j} = -\log \frac{\exp(sim(z^{(i)}, z^{(j)})/\tau)}{\sum_{k=1}^{2N} \mathbb{1}_{[k \neq i]} \exp(sim(z^{(i)}, z^{(k)})/\tau)} \quad (5)$$

The loss $l_{i,j}$ is the cross-entropy loss computed for similarities, and $sim(z^{(i)}, z^{(j)})$ is the cosine similarity function between a pair of feature vectors:

$$sim(z^{(i)}, z^{(j)}) = \frac{z^{(i)\top} z^{(j)}}{\tau \|z^{(i)}\| . \|z^{(j)}\|} \quad (6)$$

- N: Number of samples in the training batch.
- $\|z^{(i)}\|$ and $\|z^{(j)}\|$: The ℓ_2 norms of the two feature vectors $z^{(i)}$ and $z^{(j)}$ respectively.
- τ: Hyperparameter set to 0.5 in our case; to scale the cosine similarity to a range of [-1, +1].
- $\mathbb{1}_{[k \neq i]} \in \{0, 1\}$: Binary indicator which is equal to : 1 if $k \neq i$, and 0 if not.

The loss L is evaluated for (i, j) and (j, i) pairs in the positive training batch D_P following Equation 7:

$$L = \frac{1}{2N} \sum_{k=1}^{N} [l(2k-1, 2k) + l(2k, 2k-1)] \quad (7)$$

The final loss L_{total} is computed for both original and fused features as expressed in Equation 8. By

combining the original and fused features, the MMFusion framework builds powerful representations to discriminate between similar and dissimilar pairs of multimodal data.

$$L_{total} = L(z_i, z_j) + L(y_i, y_j) \quad (8)$$

2.2 Fine-tuning

In the fine-tuning step, MMFusion is trained on labeled training data following a user verification scenario. Input pairs from the same user are considered as genuine samples, while pairs from different users are set to be impostor samples. Fine-tuning process is illustrated in Figure 1.(e), the two encoders are used as feature extractors and the prediction layer *h(.)* is trained to output a probability score for genuine or impostor class. The prediction network is composed of a fully-connected layer with a sigmoid as an activation function.

3 EXPERIMENTAL EVALUATION

3.1 Databases

Two databases are used in this study, namely: the Hand Movement Orientation and Grasp (HMOG) database (Sitová et al., 2016), and the Human Machine Interaction database (HuMIdb) (Acien et al., 2021). The data pipeline starts by segmenting raw hand movement (accelerometer, gyroscope, magnetometer) data and touch data into 50% overlapping time-windows of approximately 1 second length, for the two databases, the training and validation sets are generated by subjects.

HMOG database is a freely available database proposed in the context of multimodal active user verification for mobile devices. It comprises hand movement data (from background sensors: accelerometer, gyroscope, magnetometer) and different touch gesture interactions. A total of 100 users participated in the data collection process, which makes it one of the largest benchmark databases used in the litterature. Data is splitted by selecting 80% of the subjects for training and 20% of the remaining users for testing.

HuMIdb is a publiculy available database collected with 14 sensors while users interacted naturally with their mobile phones (Acien et al., 2021). Compared to other databases, HuMIdb includes 600 users, it is the largest database up-to-date. In our study, we have selected 100 users from the 600 corpus, and we have used hand movement data (given by three background sensors: accelerometer, gyroscope, magnetometer) combined with touch gesture interactions. Here also, data is splitted similarly to HMOG database.

3.2 Implementation Details

Multimodal pre-training is conducted following the framework of the Figure 1, separately, for HMOG and HuMIdb database. The encoder *f(.)* and the feed forward network *g(.)* are trained for 1000 epochs without labeled training data using stochastic gradient descent (SGD) and momentum of 0.9. Batch size is set to 128 and learning rate to 10^{-3}. In order to compare the performance of the MMFusion framewrok, we replicated the original (unimodal) SimCLR (Chen et al., 2020) framework by performing contrastive learning with the hand movement modality and touch gesture modality separately. The pre-training routine and hyperparameters used for SimCLR are the same as for MMFusion. We used the Pytorch library and a GPU NVIDIA GTX 1060 for pre-training routine.

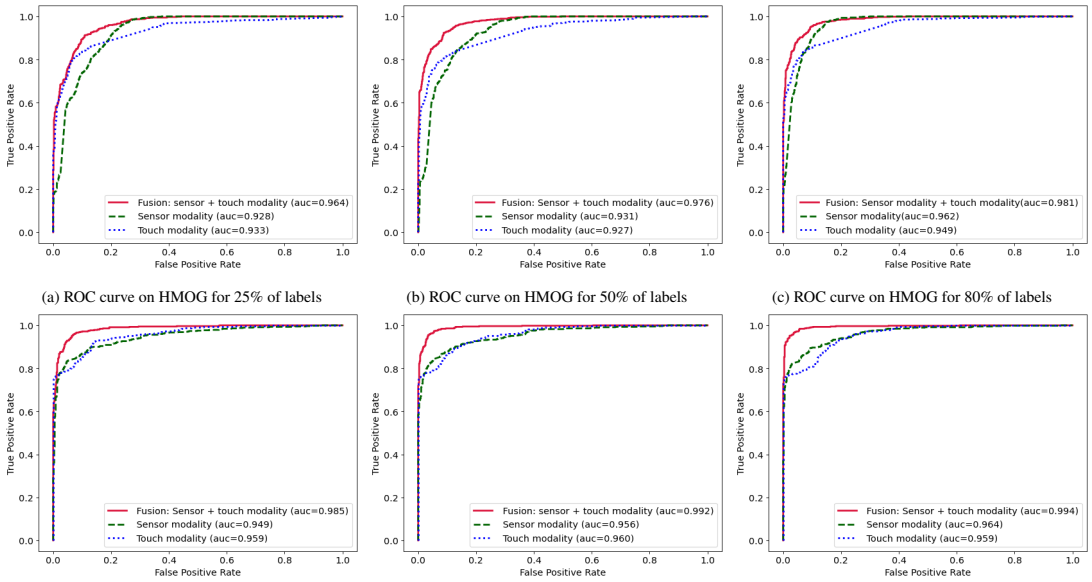
In the fine-tuning phase, the encoder *f(.)* and the fully connected layer *h(.)* are fine-tuned with labeled training data for 200 epochs using SGD with a learning rate of 10^{-2}. Similarly to MMFusion, we conducted the same fine-tunning routine for the SimCLR framework using hand movement modality and touch modality.

3.3 User Verification

As explained in the section 2.2, MMFusion is evaluated on user verification, results are reported in Table 1 for HMOG database, in terms of EER and AUC. In the case of MMFusion and SimCLR applied to the touch modality, AUC and EER scores are computed for 3 different touch gesture tasks: Tapping, scrolling and swipping. However, for the hand movement modality, only the tapping task is used for computing AUC and EER. This is due to the data collection process of HMOG, where sensor data is collected once for all the touch gesture tasks.

Table 1: Results of user verification on HMOG database.

Modality	**Task**	**Performance**	
		AUC (%)	**EER (%)**
Fusion Ours	Tap	97.48	10.71
	Scrollling	83.62	28.95
	Swipping	75.08	48.80
Sensor SimCLR	Tap	93.25	13.2
Touch SimCLR	Tap	93.56	12.11
	Scrolling	82.21	26.91
	Swipping	37.46	50.34



(a) ROC curve on HMOG for 25% of labels (b) ROC curve on HMOG for 50% of labels (c) ROC curve on HMOG for 80% of labels

(d) ROC curve on HumIdb for 25% of labels (e) ROC curve on HumIdb for 50% of labels (f) ROC curve on HumIdb for 80% of labels

Figure 2: Receiver operating chracteristic curve of the three evaluated models: MMFusion, sensor model, and touch model. All the models are fine-tuned on test set for three different fractions of labels: 25%, 50% and 80%.

As it can be seen in Table 1, the proposed model outperforms the hand movement modality and touch modality on the different tasks. Specifically, using the tapping task, MMFusion outperforms the hand movement modality by 4.23% and 2.49% margin on AUC and EER respectively. MMFusion also outperforms the touch modality by 3.92% margin on AUC and 1.4% on EER. Similarly for scrolling and swipping tasks, the MMFusion framework is the best performing method. It is clear that the combination of touch and sensor data brings significant improvement for user verification compared to a single modality.

Table 2: Results of user verification on HuMIdb database.

Modality	**Task**	**Performance**	
		AUC(%)	**EER (%)**
Fusion Ours	Scroll up	99.82	1.70
	Scroll down	99.15	3.57
	Tap	87.04	19.55
	Swipe	64.92	38.60
Sensor SimCLR	Scroll down	96.74	10.20
	Scroll up	91.8	15.98
	Swipe	81.38	29.58
	Tap	81.17	29.76
Touch SimCLR	Tap	95.90	10.26
	Swipe	48.35	50.16

Results of user verification on HuMIdb are reported in Table 2, the proposed multimodal fusion outperforms the hand movement modality on: scroll up, scroll down, and tapping tasks. The best verification performance is obtained with scroll up task, where MMFusion outperforms hand movement modality by 8.02% margin on AUC and 14.28% on EER. However, hand movement modality outperforms the proposed model on the swipe task, and touch modality outperforms MMFusion on tap task. This means that some touch gesture tasks add more complexity to the fusion system, and would preferably be used in a single modality for user verification.

3.4 Semi-supervised Evaluation

As it was stated before, collecting and labeling data for biometric applications raises privacy concerns because annotations include sensitive information. Therefore to reduce the annotations, we evaluate the performance of the proposed MMFusion on user verification when different fractions of annotated data are available. We conduct experiments on MMFusion by fine-tuning the model on three different fractions of labeled data per class: 25%, 50%, and 80% of labels. The same experiments are conducted using SimcLR framework for sensor and touch gesture modality. For each of the three models (MMFusion, sensor and touch) evaluated on user verification, we selected the best performing one in each task to perform semi-supervised evaluation. Therefore, on HMOG data the three models are selected according to the tapping task. While for HuMIdb, MMFusion is selected according to scrolling up task, the sensor model is selected according to scroll down task and finally the

touch model is selected according to tapping task.

In the semi-supervised scenario, the encoder *f(.)* and linear classifier *h(.)* are fine-tuned according to the number of labels available for each configuration (25%, 50%, 80%).

Resutls of the semi-supervised experiementation are reported in Figure 2, interestingly, MMFusion outperforms sensor and touch modality even on a very limited portion of annotated train set of 25% (Figure 2.(a)) by showing the best trade-off between sensitivity and specificity. Moreover, we observe consistent improvement with more annotated data available, in the case of 50% and 80% of train labels (Figure 2.(b) and Figure 2.(c)). Regarding HumIdb database, MMFusion gives the best trade-off between sensitivity and specificity when it is fine-tuned on 25% labels of train data (Figure 2.(d)), and improvement is consistent with more available labels (Figure2.(e) and Figure2.(f)). Unlike sensor and touch modality, MMFusion shows the best trade-off in semi-supervised evaluation, both on HMOG and HumIdb databses, which makes it a feasible solution when limited annotated data is used for fine-tuning a self-supervised model.

4 CONCLUSIONS

In this paper a powerful multimodal fusion is proposed within the context of active biometric verification on mobile devices. It relies on self-supervised learning, and combines touch screen and hand movement data collected from mobile users while performing natural interactions. MMFusion builds strong feature representations at the contrastive learning level by leveraging the complementary information of sensor and touch data.

Extensive experiments on two benchmark databases show that the proposed model outperforms contrastive learning with SimCLR when applied on hand movement and touch modality separately. In addition, during semi-supervised evaluation where labeled data is very limited, MMFusion gives the best trade-off compared to hand movement and touch models. In a future work, we aim to evaluate the vulnerability of self-supervised models on active biometric verification and suggest a method to defend against adversarial attacks.

REFERENCES

Acien, A., Morales, A., Fierrez, J., Vera-Rodriguez, R., and Delgado-Mohatar, O. (2021). Becaptcha: Behavioral bot detection using touchscreen and mobile sensors benchmarked on humidb. *Engineering Applications of Artificial Intelligence*, 98:104058.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. (2020). A simple framework for contrastive learning of visual representations. Vienna, AUSTRIA.

De Marsico, M., Galdi, C., Nappi, M., and Riccio, D. (2014). Firme: Face and iris recognition for mobile engagement. *Image and Vision Computing*, 32(12):1161–1172.

Delgado-Santos, P., Tolosana, R., Guest, R., Vera-Rodriguez, R., Deravi, F., and Morales, A. (2022). Gaitprivacyon: Privacy-preserving mobile gait biometrics using unsupervised learning. *Pattern Recognition Letters*, 161:30–37.

Fathy, M. E., Patel, V. M., and Chellappa, R. (2015). Face-based active authentication on mobile devices. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1687–1691.

Giorgi, G., Saracino, A., and Martinelli, F. (2021). Using recurrent neural networks for continuous authentication through gait analysis. *Pattern Recognition Letters*, 147:157–163.

Jing, L. and Tian, Y. (2021). Self-supervised visual feature learning with deep neural networks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(11):4037–4058.

Sitová, Z., Šeděnka, J., Yang, Q., Peng, G., Zhou, G., Gasti, P., and Balagani, K. S. (2016). Hmog: New behavioral biometric features for continuous authentication of smartphone users. *IEEE Transactions on Information Forensics and Security*, 11(5):877–892.

Stragapede, G., Vera-Rodriguez, R., Tolosana, R., and Morales, A. (2023). Behavepassdb: Public database for mobile behavioral biometrics and benchmark evaluation. *Pattern Recognition*, 134:109089.

Stragapede, G., Vera-Rodriguez, R., Tolosana, R., Morales, A., Acien, A., and Le Lan, G. (2022). Mobile behavioral biometrics for passive authentication. *Pattern Recognition Letters*, 157:35–41.

Stylios, I., Kokolakis, S., Thanou, O., and Chatzis, S. (2021). Behavioral biometrics & continuous user authentication on mobile devices: A survey. *Information Fusion*, 66:76–99.

Tolosana, R. and Vera-Rodriguez, R. (2022). Svc-ongoing: Signature verification competition. *Pattern Recognition*, 127:108609.

Tolosana, R., Vera-Rodriguez, R., Fierrez, J., and Ortega-Garcia, J. (2018). Exploring recurrent neural networks for on-line handwritten signature biometrics. *IEEE Access*, 6:5128–5138.

Zou, Q., Wang, Y., Wang, Q., Zhao, Y., and Li, Q. (2020). Deep learning-based gait recognition using smartphones in the wild. *IEEE Transactions on Information Forensics and Security*, 15:3197–3212.