This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIFS.2018.2833042, IEEE Transactions on Information Forensics and Security

Benchmarking Touchscreen Biometrics for Mobile Authentication

Julian Fierrez, *Member, IEEE,* Ada Pozo, Marcos Martinez-Diaz, Javier Galbally, and Aythami Morales

***Abstract*—We study user interaction with touchscreens based on swipe gestures for personal authentication. This approach has been analyzed only recently in the last few years in a series of disconnected and limited works. We summarize those recent efforts, and then compare them to three new systems (based on SVM and GMM using selected features from the literature) exploiting independent processing of the swipes according to their orientation. For the analysis, four public databases consisting of touch data obtained from gestures sliding one finger on the screen are used. We first analyze the contents of the databases, observing various behavioral patterns, e.g., horizontal swipes are faster than vertical independently of the device orientation. We then explore both an intra-session scenario where users are enrolled and authenticated within the same day; and an inter-session one, where enrollment and test are performed on different days. The resulting benchmarks and processed data are made public, allowing the reproducibility of the key results obtained based on the provided score files and scripts. In addition to remarkable performance thanks to the proposed orientation-based conditional processing, the results show various new insights into the distinctiveness of swipe interaction, e.g.: some gestures hold more user-discriminant information, data from landscape orientation is more stable, and horizontal gestures are more discriminative in general than vertical ones.**

***Index Terms*—Active authentication, biometrics, smartphone, touchscreen, human computer interaction**

I. Introduction

Nowadays everyone carries sensitive information, such as bank account details, emails or passwords in their smarthphones and tablets, which can be easily lost or stolen, resulting in information leaks. In addition, in our society, there is an increasing requirement to reliably authenticate individuals in new applications as, for example, electronic transactions. Traditional single entry point authentication schemes in such mobile scenarios can be improved, as passwords and PIN codes tend to be short and easy to remember, therefore easy to break [1], [2]. Other authentication alternatives such as secret touch patterns are encouraging [3], but have their limitations, e.g., they are vulnerable to attacks, such as following the residues left on the device's screen after entering the same pattern frequently [4].

In addition to the drawbacks pointed out above, probably the main limitation offered by traditional security systems comes from the fact that users only authenticate once at the beginning of the session. This authentication is not performed again until the next time the device needs to be unlocked. Therefore, if the security is compromised, it is compromised for a long period of time in which the attacker can potentially access multiple personal information. This situation has led to a growing body of literature looking for authentication based on continuous biometrics, which periodically authenticate the user and thus ensure security in the device beyond the entry point [1], [5].

Biometrics using touchscreen signals is one of the most active fields of investigation in continuous authentication, where the user is passively being authenticated in the background while normally interacting with a mobile device [1], [6], [7]. This is done by comparing the patterns of use to those of the legitimate subject and blocking the device if there are not enough coincidences. This way the system regularly verifies that the same person who enrolled is the one still interacting with the device.

Touchscreen biometrics allow a passive authentication of the user that does not need any extra sensors in the device. Data is obtained from the user normal interaction with the touchscreen, without needing any specific task to be done. The basis for this form of authentication is that every person behaves differently when interacting with a touchscreen, which results in different patterns of use [7], [10]–[12], [17]. These patterns have been shown to be discriminative, presenting a high inter-class variance which allows users to be recognized with them [7]. However, they also present the problem of having high intra-class variability [11], hence changing within the same user with time [20] or depending on the emotional

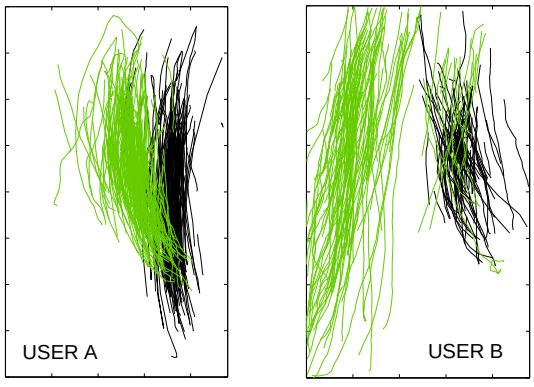


Fig. 1: Example of touch gestures for two different subjects, showing in green and black data captured in different days.

J. Fierrez, A. Pozo, M. Martinez-Diaz and A. Morales are with Universidad Autonoma de Madrid, C/Francisco Tomas y Valiente 11, EPS, Madrid 28049, Spain. J. Fierrez is the corresponding author for this paper.
E-mail: julian.fierrez@uam.es.

J. Galbally is with the European Commission, DG Joint Research Centre, Ispra, Italy.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIFS.2018.2833042, IEEE Transactions on Information Forensics and Security

TABLE I: Related works in touchscreen biometrics. Intra-session refers to experiments where the user is enrolled and authenticated within the same day, while inter-session relates to authentication in a different day. EER - Equal Error Rate, FAR - False Acceptance Rate, FRR - False Rejection Rate, Acc - Accuracy, FNMR - False Non Match Rate, FMR - False Match Rate. (No authentication experiments in Antal *et al.* [8], only age/gender estimation) (Only delays and probabilities of false detections in Perera and Patel [9])

Study	**# users**	**# strokes/user**	**# features**	**Classifiers**	**Performance (%)**	
					Intra-session	**Inter-session**
Frank *et al.* [7] (2013)	41	488	27	SVM, kNN	EER: 2.0-3.0	EER: 0.0-4.0
Serwadda *et al.* [10] (2013)	190	400	28	Ten different classifiers (best logistic regression, SVM and random forest)	-	EER: 13.8-36.0
Xu *et al.* [11] (2014)	32	200 (29 users), 1200 (3 users)	37	SVM	EER: 10.0	Acc: 70.0-100.0
Antal *et al.* [8] (2015)	71	3260	15	SVM, random forest, kNN	-	-
Zhang *et al.* [12] (2015)	50	-	27	SVM, sparsity-based classifiers	EER: 4.1-5.9 [7] EER: 0.4-0.9 -	EER: 4.9-14.4 - [10] EER: 19.2-23.8
Mondal *et al.* [13] (2015)			15	Artificial Neural Networks, Counter Propagation Artificial, Neural Networks	[8] FNMR: 0.0 FMR: 0.08	-
Murmuria *et al.* [14] (2015)	73	-	5	StrOUD	-	EER: 32.1-46.3
Kumar *et al.* [15] (2016)	28	175	5	kNN, random forest	Acc: 88.0-92.0	-
Mahbub *et al.* [16] (2016)	48	3482	24	kNN, SVM, GBM random forest	EER: 22.1-38.0	-
Shen *et al.* [17] (2016)	71	2002	22-27	SVM, random forest, kNN, neural networks	FAR: 1.9-7.4 FRR: 2.7-8.6	FAR: 4.7-10.9 FRR: 5.7-13.5
Sitová *et al.* [18] (2016)	90	-	22-27	SVM, Scaled Manhattan, Scaled Euclidean	-	EER: 15.0-16.0
Kumar *et al.* [19] (2018)				Elliptic envelope, SVM, Local Outliers Factor, isolation forest	Acc: 80.1-89.6	-
Perera *et al.* [9] (2018)			5	Bayesian and MiniMax QCD	-	-

state. An example of the touch gestures captured for two given subjects over two days can be found in Fig. 1. It can be observed that these gestures are very different between the users, having different length, inclination, etc., thus presenting high inter-user variability. However, the strokes captured also vary significantly within each subject over the two days, for example, changing the area used on the screen for the same task. Nevertheless, it is worth mentioning that the gestures tend to be stable within the same day.

The contributions of this paper are:

- Summary of recent efforts and resources for research in this topic (see Table I).
- Analysis of the contents of four touchscreen biometrics databasets, from which we observe various behavioral patterns (e.g., horizontal gestures are in general faster than vertical independently of the device orientation).
- Experiments with statistical and discriminative methods for swipe biometrics, including a novel architecture that independently processes swipes of different orientation (see Fig. 2).
- New knowledge about this biometric technology and in particular about the discriminative power of different types of swipe gestures in various realistic scenarios.

This paper is an extension of [21], where only a statistical system and one database were used, very limited experiments were reported, and the comprehensive summary of related works presented here was lacking.

More in detail, here we analyse three different systems: 1) a discriminative approach using Support Vector Machines (SVM) [7]; 2) a statistical approach, using Gaussian Mixture Models (GMM) adapted from Universal Background Models (UBM) [3], [22]; and 3) a third approach based on the fusion of the previous two [23], [24]. We explore different scenarios over four public benchmark datasets, where enrolment and authentication are performed on the same session and on different sessions, evaluating the performance and limitations in each of the systems. We also study whether combining both sessions' data reduces the variability within each subject and can improve the results. Additionally, we compare the performance of different types of touch operations (sliding upwards, downwards, leftwards and rightwards), considering which hold more user-relevant information and which are more distinctive for authentication. Resources for reproducing the benchmark results obtained are available online[1].

The rest of this paper is organized as follows. Section II

[1] https://atvs.ii.uam.es/atvs/databases.jsp

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIFS.2018.2833042, IEEE Transactions on Information Forensics and Security

summarizes related works in swipe biometrics. Section III includes the systems description, the feature vectors evaluated and the approaches followed for authentication. In Section IV we describe the databases. Experimental results are given in Section V, with additional analysis in Section VI. A short note on Active Authentication is then included in Section VII, before concluding in Section VIII.

II. Related works

A one-finger touch swipe gesture, or simply a *swipe*, is considered to be a touch gesture in which the user places one finger on a touchscreen and quickly moves it horizontally or vertically, typically for scrolling purposes. Despite the fact that swipes are not the only touch signals adequate for mobile biometrics (e.g., one may also use fling, press, or pinch signals), most, if not all works in touch biometrics so far, have demonstrated best results using one-finger swipe gestures. We will therefore concentrate in swipe signals, using the terms swipe biometrics and touch biometrics interchangeably in the rest of paper.

Existing literature on swipe biometrics may be categorized in two main groups. The first one uses swipes made on the entry-point, i.e. a secret pattern, to authenticate the subject. Thus, the authentication is not being continuously performed. The second type of approaches explore active authentication methods, where swipes made on the screen during normal interaction with the device are continuously exploited for authentication. Since, as presented in the introduction, the main focus of the article is the use of swipe biometrics for active authentication, in this section we will review existing methods belonging to this type of technology. A summary of previous works in this field is presented in Table I. Other related works comparing image-based features and exploring factors such as experience, gender, and age, have been published, respectively, in [25] and [8].

Frank *et al.* [7] is one of the first and most comprehensive works using touch data for continuous authentication. They studied 41 subjects who provided data from single touch operations while comparing images and reading texts. Intra and inter-session authentication is studied, obtaining less than 4.0% Equal Error Rate (EER) using Support Vector Machines (SVM) with Radial-Basis Functions (RBF) and k-Nearest-Neighbors (kNN). In addition, it was observed that combining blocks of strokes for authentication results in better performance, conclusion also reached in other works, such as [10] and [17].

In Serwadda *et al.* [10], a benchmark of the best suited algorithms for active authentication using swipe biometrics was generated using a large dataset with touch data operations, acquired across two different sessions for 190 subjects. Extracting a 28-feature vector, they reported that the best performance, around 15% EER, was obtained using logistic regression, SVM and random forests.

Shen *et al.* [17] studied SVM, kNNs, neural networks and random forest classifiers for different applications (e.g. document reading or picture viewing), as well as with free tasks. They analysed four types of touch operations (up, down, left and right) and different feature sets were extracted in each of them. They concluded that: 1) swipes with a smaller active area (horizontal gestures) are more stable and discriminative; and 2) better results are obtained with specific tasks (1% EER) than free ones (5% EER) following the same methodology.

In Zhang *et al.* [12], SVM and dictionaries based on sparse representations were compared for three datasets, two of them public [7], [10], reporting that dictionaries perform slightly better than SVM, with EER ranging from 0.4% to 23.8%.

Mahbub *et al.* [16] studied touch data authentication over a dataset with a large number of samples per subject obtained with a more realistic application that allowed free interaction. kNN, SVM, random forest and Gradient Boosting Model (GBM) were exploited for authentication, resulting in EER ranging from 22% to 38%.

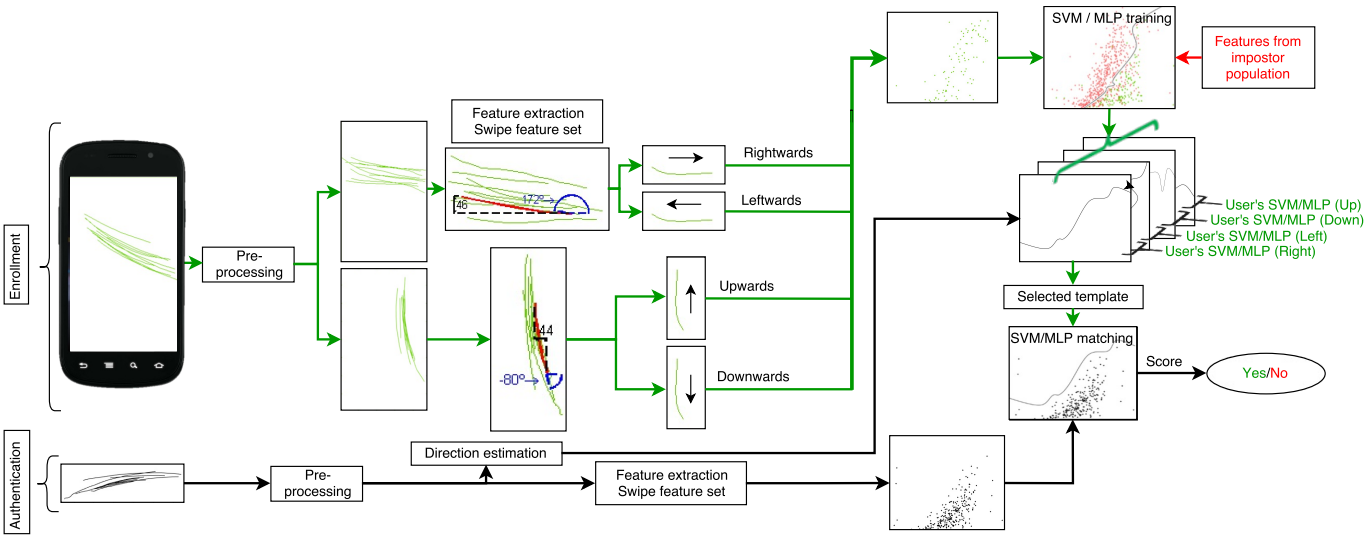
Additionally, fusion of single touch operations with other biometrics have been studied in Xu *et al.* [11] (keystroke, swipe, pinch), Kumar *et al.* [15] (keystroke, swipe, phone movement) and Sitová *et al.* [18] (hand movement, orientation, grasp, tap, keystroke features). The first one reported accuracies above 90% with SVM, whilst the second obtained 10% EER with kNN and random forest, using only swipe data for authentication. Finally, in the third one they compared the EER obtained while the user was walking and sitting, reporting a 7.16% and 10.05% EER, respectively, when they combined all the mentioned features.

Most of the literature summarized above assume the availability of both genuine and impostor samples for training. However, some applications impose restrictions that make difficult or even impractical the use of impostor samples for training. In order to deal with those cases, researchers have explored the use of one-class classifiers and anomaly detection techniques. Murmuria *et al.* [14] proposed Strangeness-based Outlier Detection (StrOUD) to monitor the user behavior based on power consumption, touch gestures, and physical movement. Those algorithms demonstrated competitive performance with EER under 7% when sufficient data is available to model each user (only genuine samples). Kumar *et al.* [19] analyzed three one-class classifiers for continuous authentication including one-class Support Vector Machines, Elliptic Envelop, and Local Outlier Factor algorithms. The results obtained suggested that it is possible to achieve comparable performance only with genuine data compared to using both genuine and impostor data for training. Anomaly detection methods have been also applied for continuous authentication based on touchscreen interaction. Perera and Patel [9] proposed different quick intrusion detection methods for mobile active user authentication. Their results show that it is possible to detect a high percentage of intrusions with a relatively small number of gestures. Similarly, in [13] Mondal and Bours proposed a trust model based on Counter Propagation Artificial Neural Networks.

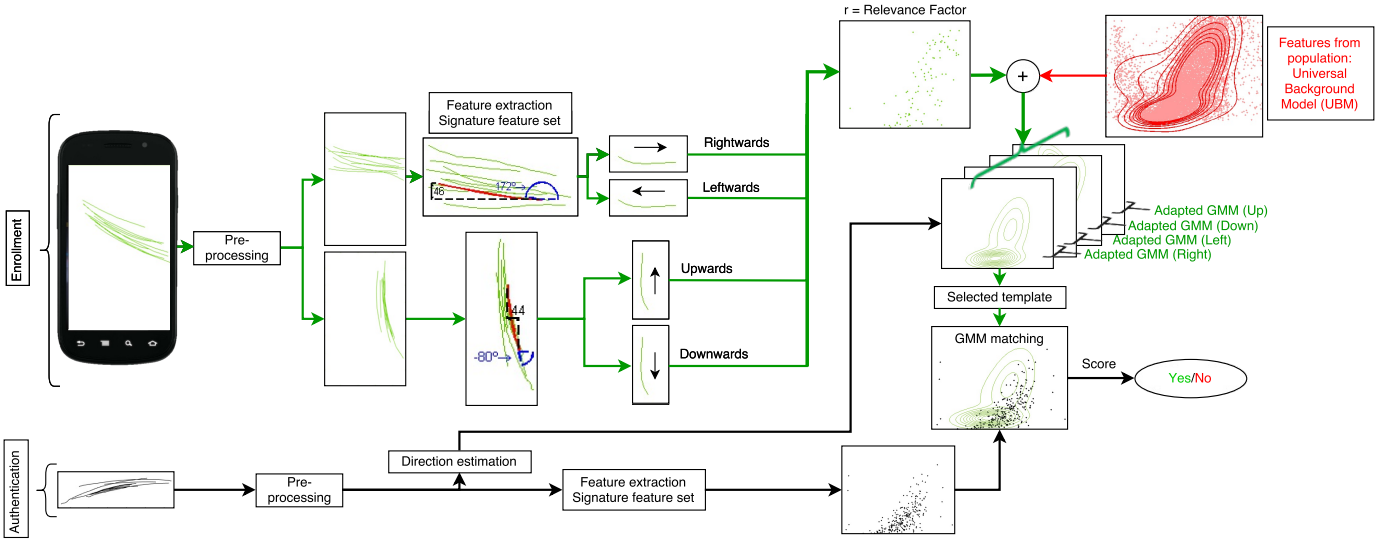
III. Systems description

As mentioned in the introduction, three different systems are studied in the present work: 1) Discriminative: based on a feature set previously used in swipe biometrics and an

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIFS.2018.2833042, IEEE Transactions on Information Forensics and Security



(a) Discriminative: Swipe feature set + SVMs (or MLP)



(b) Statistical: Signature feature set + GMMs

Fig. 2: Architecture of the touch biometrics authentication schemes studied.

SVM classifier (also tested with worse results with an MLP classifier, see Section V-E); 2) Statistical: based on a feature set previously used in signature recognition and a GMM classifier; 3) Multimodal: based on the score level fusion of the previous two. These three systems can all work under enrolment or authentication mode:

Enrolment. (See Fig. 2) First, in the pre-processing step short strokes of less than five data points are discarded, because they are likely to come from taps on the screen. Data from landscape and portrait orientations is processed separately. Afterwards, strokes are classified as vertical or horizontal. A stroke is considered vertical if there is a bigger deviation in the y axis than in the x axis. Likewise, if the biggest deviation is in the x axis, the stroke is considered horizontal. Then, a feature vector is extracted for each stroke. Gestures are separated depending on their direction as upwards, downwards, leftwards or rightwards. With this division, it is easier to exploit the potential individualities of each gesture, as each operation is performed differently and has its own characteristic features depending on its direction [17]. A template is obtained for each of the four operations.

Authentication. (See Fig. 2) Swipes on the screen are captured while the user interacts normally with the device. If the stroke captured has five or more data points, its direction (up, down, left, right) is estimated. Based on this direction, the corresponding user template is selected. Afterwards, the features of each stroke are extracted and a similarity score is computed comparing it to the selected template. Similarly to previous works [7], [10], [17], we combine 10 consecutive input strokes for authentication by averaging their individual scores. The resulting score is used to decide whether the person interacting with the device is the legitimate user.

We also conducted various tests exchanging feature sets and classifiers, i.e., swipe_features+GMMs and signature_features+SVMs with worse results, therefore we only report the mentioned couplings swipe_features+SVMs and

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIFS.2018.2833042, IEEE Transactions on Information Forensics and Security

TABLE II: Optimal 5-dimensional feature set selected by the SFFS algorithm. Notation similar to [26].

Stroke type	Best performing features
Vertical	θ(finger-down to finger-up), std of a_x, $(x_{max} - x_{min})/x_{max\_range}$, $(\bar{x} - x_{min})/\bar{x}$, $(y_{max} - y_{min})/y_{max\_range}$
Horizontal	θ(finger-down to finger-up), std of a_y, $(y_{max} - y_{min})/y_{max\_range}$, $(\bar{y} - y_{min})/\bar{y}$, $(x_{max} - x_{min})/x_{max\_range}$

signature_features+GMMs, in architectures similar to the most successful related works.

A. Feature extraction

Two different feature vectors are exploited in this work: one previously used for authentication with touch interaction in [10], and another one adapted from the feature vector presented in [26] for online signature verification. For both feature vectors normalization into (0-1) range is performed using tanh-estimators [27].

1) Swipe feature set: For each of the strokes in the dataset a 28-dimensional feature vector, previously employed in [10], is computed. A velocity vector and an acceleration vector are computed for every pair of adjacent points in a stroke. For these two vectors, as well as for the pressure and area measurements, the mean, the standard deviation, the first quartile, second-quartile and third quartile are calculated.

The eight remaining features are:

- The x and y coordinates of the most extreme points in a stroke, that is, the most leftward and rightward in horizontal strokes and the uppermost and bottommost in vertical ones. Each point contributes as two features.
- The distance between start and end-points.
- The angle of the straight line that joins the start and end-point.
- The total duration of the stroke.
- The summation of the distance between every pair of adjacent points.

Similar to [10] we evaluate this feature set with SVMs.

2) Signature feature set: The motivation for the use of this feature set is its distinctiveness both in signature biometrics [26], and in graphical touch passwords [3]. Nevertheless, swipes made on a screen are much simpler than handwriting or graphical passwords and hence, this feature set needs to be adapted, removing several features from the initial 100 [26] that do not apply to this problem, like pen-ups or number of direction changes. Of the remaining 61 features, we select the best subset using the Sequential Forward Floating Search (SFFS) algorithm [28]. The best features chosen by this algorithm (5 in total) are shown in Table II.

Similar to [26] we evaluate this feature set with GMMs.

B. Classifiers

Two different classifiers are used to compute the similarity scores: discriminative using SVM with RBF kernel [7], [10], [11], [17], and statistical using GMM with Universal Background Model (UBM) adaptation [3], [22], [23]. The fusion of both approaches is also evaluated.

1) Discriminative system: In this case, the objective is to find a boundary in the feature space that separates the legitimate user and the impostors using SVM with RBF-kernel. The regularization parameter C and the Kernel's variance σ^2 are chosen heuristically (as later explained in Section V). Similar to previous related works [8], this classifier is coupled with the 28-dimensional feature vector described in Section III-A1. Fig. 2a shows this system's architecture. For each of the touch operations, a SVM is trained using T randomly chosen training samples from the legitimate user and T samples from $T/10$ randomly chosen subjects from the impostor population, each contributing with 10 samples. LS-SVM lab Toolbox (Version 1.8) is used for coding.

2) Statistical system: This approach tries to model how the user's data is distributed. For this purpose, the subject model is derived from a UBM, an "average" user model describing the behaviour across a population. The system architecture can be found in Fig. 2b. The UBM is computed once for all subjects using full covariance matrices and all data from the training set. The legitimate user's training samples are used for the adaptation, deriving the user's model. The relevance parameter r is a trade-off that controls how much information from the current subject is used to adapt the UBM. If it is high, the UBM information will weight more, while if it is low it will be less affected by it. The steps followed to adapt the UBM can be found in [22]. This classifier is coupled with the 5-dimensional feature vector described in Section III-A2, similarly as related works [3], which exploit similar features with GMMs. Matlab's Statistics Toolbox (Version 8.3) was employed for the GMM-UBM implementation.

Fig. 3 depicts as an example the adaptation from the general UBM model using three GMM components and a relevance factor $r = 5.5$. In Fig. 3a the three Gaussian components of the UBM are plotted with their adaptation to the user's data. Each of the Gaussian components has moved towards the subject data to better represent it. Fig. 3b and 3c show respectively the UBM and its adaptation, obtained by calculating a weighted sum of each of the Gaussian components. The results show how the UBM covers most users' data, while the specific GMM model for the current user is located over his data at the same time it still covers areas that mostly contain data from other users. Even though there were no data points in the legitimate subject's training set in these areas, they are more likely to fall there, given the accumulation from other users. Nevertheless, the Gaussian component covering those areas has a smaller weight than in the UBM.

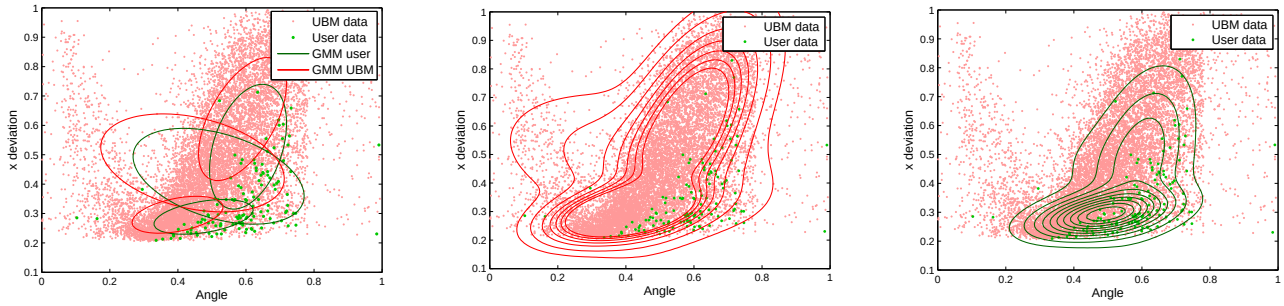
3) Fusion system: The two previous systems are combined, merging their information using score level fusion [29]. A similarity score is obtained in each system, s_{SVM} and s_{GMM}, and normalized to (0-1) range using tanh estimators, $\bar{s}_{SVM}$ and $\bar{s}_{GMM}$ [27]. The final score s_{fusion} is obtained as the average of the previous two: $s_{fusion} = (\bar{s}_{SVM} + \bar{s}_{GMM})/2$.

IV. Evaluation Datasets

A. Serwadda database

The public database from [10] is composed of contributions from 190 subjects, students, faculty or staff at Louisiana

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIFS.2018.2833042, IEEE Transactions on Information Forensics and Security



(a) UBM and GMM components (b) UBM model (c) User's model adapted from the UBM

Fig. 3: Example of the UBM model and its adaptation to obtain the user's model. Impostor data is plotted in red, while the legitimate user's is in green, overlapped with the UBM and the adapted user's model.

TABLE III: Analysis of Serwadda [10], Frank [7], Antal [8], and UMDAA-02 [16] datasets. Mean and standard deviation (in brackets) of the number of strokes per user, points per stroke in each session, and sessions per user (when available). UMDAA-02 is the only with free use of the mobile, the other three datasets are task-specific.

Database		Portrait				Landscape			
		Up	**Down**	**Left**	**Right**	**Up**	**Down**	**Left**	**Right**
Serwadda 190 users	**#users**	124	132	104	118	54	54	17	27
	#strokes/user	85.6 (26.5)	116.8 (32.5)	70.2 (31.2)	86.5 (34.0)	80.9 (28.5)	122.4 (38.4)	70.6 (32.4)	75.1 (34.1)
	#points/stroke	21.1 (7.1)	24.8 (8.6)	9.7 (3.7)	8.9 (4.5)	18.4 (5.8)	21.8 (7.5)	7.9 (3.4)	7.9 (4.5)
Frank 41 users	**#users**	40	41	41	41	7	9	2	1
	#strokes/user	33.0 (31.68)	234.4 (136.0)	112.0 (57.5)	108.1 (42.8)	19.6 (15.4)	107.4 (168.8)	2.5 (0.7)	4.0 (0.0)
Antal 71 users	**#users**	58	58	64	65	2	2	8	7
	#strokes/user	15.4 (11.7)	29.1 (22.4)	63.3 (38.8)	109.0 (60.8)	6.0 (0.0)	10.0 (8.5)	20.6 (23.4)	22.7 (25.2)
UMDAA-02 48 users	**#users**	32	33	33	31	29	28	26	29
	#strokes/user	5.0 (10.6)	15.9 (54.7)	6.0 (13.4)	9.5 (22.1)	1.7 (1.9)	1.9 (1.4)	2.0 (1.7)	1.7 (2.2)
	#points/stroke	21.4 (13.0)	21.6 (7.6)	18.4 (12.9)	13.8 (5.3)	15.0 (10.6)	17.7 (13.0)	14.1 (7.2)	11.9 (6.5)
	#sessions/user	34.0 (38.7)	43.7 (50.5)	28.1 (33.3)	34.2 (39.1)	8.5 (8.3)	8.4 (8.3)	5.2 (4.5)	10.3 (10.7)

Tech University. Two applications were developed for data collection, running on Android 4.0 and using one device model (*Google Nexus S*). In these applications, multiple choice questions were asked based on the images/texts one had to browse/read. Free interaction with the device was allowed, permitting both landscape and portrait orientation.

Data was captured over two sessions, at least one day apart, recording the x and y coordinates, the timestamp, the area covered by the finger, the pressure on the screen and the device orientation. Only gestures obtained by swiping one finger on the screen were recorded. Multi-touch gestures, e.g. zooms, were ignored.

For each of the different gestures included in the database, the first row of Table III summarizes: the number of users with that type of data, the mean number of strokes per user and the mean number of points per stroke in each session. It can be observed that most subjects have data for portrait orientation, while only 54 different users out of the 190 subjects have strokes from landscape orientation. The most frequent gestures are those made downwards, with around 120 strokes per user, while all the rest have a similar frequency of around 80 strokes per user. It is worth noting that, in both orientations, the number of points per stroke is only of about 8 points for horizontal strokes and 20 for vertical ones. This means that the active area available does not affect the number of data points. Therefore, given the same sample frequency, horizontal strokes must be performed, in general, faster.

B. Frank database

This database is composed of swiping data generated by 41 users over two sessions, one week apart [7]. Two Android applications were deployed for data acquisition, one for comparing images and another for reading texts, allowing the subject to move and interact freely with the screen. Both phone orientations were allowed. Multiple devices (operating on Android 2.3.x) with different sampling frequencies were employed, recording for each data point the x and y coordinates, the timestamp, the area covered by the finger, the pressure, the device orientation and the finger orientation.

The second row of Table III presents a summary of this database. Due to the multiple devices used to capture the database, each with a different sample frequency, the number of points per stroke are not shown in this case. All subjects have data in portrait orientation, with downwards strokes being the most frequently performed and upwards the least. On the other hand, as also happened in the previous database, only a smaller proportion of users have data in landscape orientation, where downwards strokes are the only ones with a significant amount of samples.

A summary of biometric authentication results on this database from [7] appears in Table I. That work implements a direction- and orientation-independent approach, so comparison with our methods developed here is not straightforward.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIFS.2018.2833042, IEEE Transactions on Information Forensics and Security

C. Antal database

In this case, eight different devices were used, including tablets, with varying screen sizes, for 71 users [8]. An application was developed for the acquisition, where subjects had to read texts, which required vertical strokes, and choose their favourite picture, which required horizontal strokes. The data was obtained during 4 weeks (not separated in sessions in the database), where each subject interacted with multiple devices, recording for each data point the same information as in the previous databases and allowing both phone orientations.

This database summary can be found in the third row of Table III. As happened in the Frank database, the analysis of the number of points per stroke is not conducted because each of the devices employed in the acquisition has a different sampling frequency. In this database, the most frequently performed gestures are horizontal, specially rightwards. It should be mentioned that, the same as in Frank database, upwards strokes are the least frequent. A very small number of users swiped the screen in landscape orientation.

As can be seen in Table I, the authors in [8] describe this dataset and report results on age and gender estimation, but not on biometric authentication.

D. UMDAA-02 database

According to [16] this dataset contains samples from 48 volunteers captured using Nexus 5 phones over two months. On the contrary to the other databases, free use of the devices was allowed during these two months, without requiring any concrete task to be performed. Thus, more data from each user is present, divided in sessions from the unlocking of the device until it's locked again. Each session may include data from several days until the device is locked naturally by the user.

In the last row of Table III a summary of the data present can be found. This summary shows that the most used gestures during the normal use of the phone are downwards. The number of sessions per type of gesture also indicates that the majority of times, portrait orientation is preferred, although most of the users used both orientations.

V. Experiments

A. Statistical system configuration

The best performing parameters for the statistical system are tuned following the procedure explained in [21], considering the inter-session scenario (see Section V-D). Serwadda *et al.* database [10] is employed for this purpose. The optimum parameters change depending on the type of gesture, but in general the best results are obtained using a bigger number of components and samples, and an intermediate value for the relevance factor r, balancing out the weight given to the UBM and user models. The best parameters obtained are $N = 4$ GMM components, $r = 30$, 40 training samples. Fig. 4 shows an example of the tuning of r with vertical strokes in portrait orientation. More examples and details about how these parameters were selected can be found in [21].

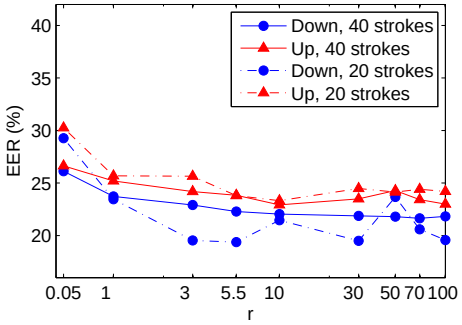
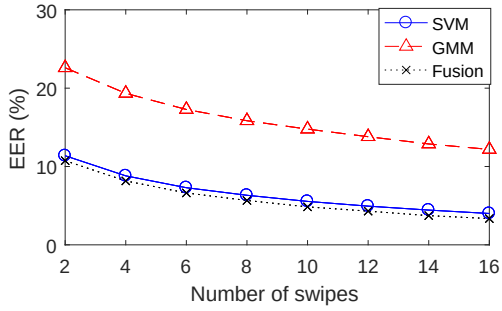
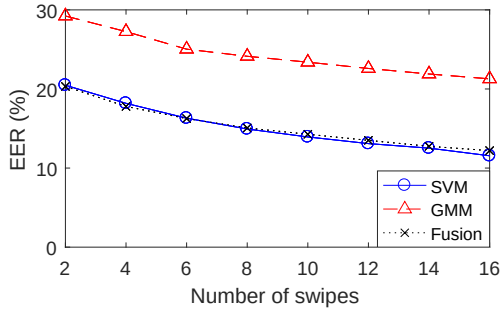


Fig. 4: Trade-off between user-specific information and UBM (increasing r) for statistical user modeling.



(a) Serwadda database



(b) UMDAA-02 database

Fig. 5: Accuracy of the three systems depending on the number of test swipes averaged to obtain one final score.

B. Influence of the number of test swipes

Following the same protocol proposed in [16], we compare the effect of averaging a different number of swipes to obtain the final score. For this experiment, 70% of the data is used for training while 30% is used for testing using downwards swipes from portrait orientation. We carry out this experiment with Serwadda and UMDAA-02 databases.

Fig. 5a and 5b show the results using Serwadda and UMDAA-02 databases, respectively. It can be observed that in both cases the EER improves when increasing the number of averaged swipes, with a small slope. This slope is similar for all methods, with GMM performing the worst. It should also be noted that despite having more training data and less impostors, the performance is much worse on UMDAA-02. For all other experiments the number of swipes averaged is set to 10.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIFS.2018.2833042, IEEE Transactions on Information Forensics and Security

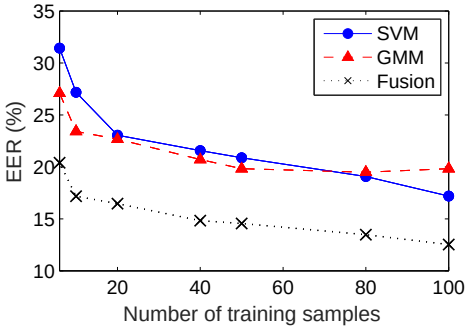


Fig. 6: Influence of the number of training swipes on the accuracy of the three considered systems.

C. Influence of the training swipes

Employing the database from Serwadda *et al.* [10], the influence of the number of training swipes is evaluated for each of the systems. Fig. 6 shows the results for the inter-session scenario with downwards strokes from portrait orientation, the most commonly performed. It can be observed that performance improves with the number of training samples, with an initial steep decrease of the error. In the GMM system, it stabilizes around 40-50 training samples, whereas in the SVM and fusion based system it does not stabilize. Therefore, the statistical system needs a smaller number of training samples to reach its optimal performance and has less computational cost than the other two. It can be emphasized that the fusion based system outperforms the statistical and discriminative ones in all cases. The discriminative system obtains a better EER than the statistical with a large number of training samples, 80 or more, but does not when the number of samples is lower.

We also replicated most of the results that will be presented in Section V-D, comparing three different training sets: consecutive samples from the session start, from the session end, and chosen randomly across the session. Results in all cases are very similar: always below 5% relative performance difference among the three training strategies, being always a bit better the random selection.

D. Experiments across sessions

Following the results presented in V.B and V.C, 40 swipes are used to the train the users' models and 10 swipes are averaged to obtain one single authentication score.

In the following set of experiments we analyse the performance of all the systems presented in Section III-B for each of the four datasets described in Section IV, considering three different scenarios depending on the origin of the train and test data:

Intra-session scenario. A user is enrolled and authenticated within the same day. The evaluation is performed employing 40 genuine training samples to train from the legitimate user and the rest to test the system.

Inter-session scenario. The user's model is obtained with training data from the first session, while it is evaluated using the second session's data.

Combined sessions. The data from both sessions is combined. Following the same procedure as with the intra-session scenario, the train set comprises 40 training swipes from the legitimate user, whereas the remaining samples are used for testing.

In the three scenarios presented above, to train the discriminative system, 4 users contributing each with 10 training samples are chosen randomly to compose the impostor population.

1) Serwadda database: The performance obtained with this database is shown in Table IV.

Intra-session scenario. For both portrait and landscape orientations the best performance, ranging from 3% to 6% EER, is obtained with the fusion based system, which merges the different information of the strokes exploited by the discriminative and statistical systems. The statistical system performs slightly worse than the discriminative. The results are better for portrait orientation in comparison to landscape. In addition, horizontal strokes outperform vertical ones.

Inter-session scenario. The performance deteriorates in comparison to the intra-session scenario, with the EER raising to around 15%. This is caused by the lack of stability of users behaviour across sessions. However, the fusion based system is better than the other two again, and thus less affected by this variability. As happened before, horizontal gestures outperform vertical ones and, in this case, landscape orientation obtains a better result. It is worth mentioning that upward strokes in both device orientations are the ones whose results degrade more in this scenario. Thus, it can be hypothesized that these gestures are especially unstable.

Combined sessions. In this scenario the data coming from both sessions is combined to form the train and test sets. Thus, the intra-user variability is better represented here in comparison to the inter-session experiment. Yet, the variability is bigger than in the intra-session scenario. As a result, the EER found improves the inter-session scenario, but not the intra-session, with values between the ones found in each of them. Landscape and portrait orientation perform similarly and horizontal strokes obtain the best result. It can be emphasized that upward strokes, which were the worst in the inter-session scenario, present the largest qualitative improvement in comparison.

2) Frank database: Even though both phone orientations were allowed, very few users swiped the screen in landscape orientation, so these data (i.e., landscape) are not employed in the experiments. Table V summarizes the results obtained.

Intra-session scenario. As happened before, the best performing system is the fusion based, with EER around 3%. Vertical strokes perform worse than horizontal ones and the best performing gestures are rightwards.

Inter-session scenario. This scenario results are worse than the ones found in the intra-session. It should be noted that, in the statistical system, rightwards strokes, that performed the best in the intra-session scenario, obtain a much worse 27.8% EER. The most likely cause is that, although they are discriminative and performed differently in each user, they do not remain stable, and thus the model found in the first session is not representative. Additionally, in this scenario, downward

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIFS.2018.2833042, IEEE Transactions on Information Forensics and Security

TABLE IV: Performance in terms of EER (%) with Serwadda dataset [10]. 40 swipes for training, average of 10 swipes for authentication in all experiments. P = Portrait; L = Landscape. Mean EER across users. (Standard deviation in brackets.)

		Present Work						**Serwadda *et al.* [10]**
Scenario		**Portrait**			**Landscape**			
		SVM	**GMM**	**Fusion**	**SVM**	**GMM**	**Fusion**	**SVM**
Intra-session	**Up**	5.7 (6.2)	10.7 (10.0)	**3.5 (5.3)**	7.1 (6.6)	7.7 (6.2)	**3.3 (3.6)**	-
	Down	7.2 (6.9)	11.4 (8.3)	**4.2 (4.6)**	10.8 (7.8)	12.0 (7.0)	**6.4 (5.8)**	
	Left	4.4 (5.0)	8.6 (7.4)	**2.9 (3.8)**	5.5 (5.6)	7.9 (7.2)	**3.3 (3.4)**	
	Right	5.8 (5.0)	8.0 (6.4)	**2.9 (3.4)**	4.8 (6.6)	6.9 (6.7)	**4.8 (6.6)**	
Inter-session	**Up**	23.8 (17.3)	22.8 (13.4)	**17.4 (14.5)**	14.7 (12.4)	17.4 (10.2)	**10.7 (9.1)**	P: 18.0; L: 14.7
	Down	22.9 (15.8)	19.2 (13.4)	**15.9 (12.0)**	16.1 (10.7)	22.3 (11.9)	**12.2 (10.0)**	
	Left	21.6 (17.7)	20.6 (18.0)	**13.9 (14.1)**	13.2 (11.3)	20.7 (17.0)	**12.1 (11.1)**	P: 15.7; L: 13.1
	Right	22.9 (18.3)	18.1 (17.1)	**16.1 (13.4)**	**11.8 (10.0)**	19.2 (17.1)	12.0 (10.5)	
Combined sessions	**Up**	10.1 (7.0)	16.0 (9.5)	**7.5 (8.6)**	10.8 (7.1)	13.1 (7.2)	**7.1 (4.9)**	-
	Down	11.4 (6.6)	15.6 (8.4)	**7.7 (5.3)**	14.4 (8.5)	14.7 (7.9)	**9.3 (5.7)**	
	Left	7.6 (3.5)	12.0 (7.2)	**5.9 (5.0)**	9.8 (6.3)	12.8 (7.0)	**6.9 (5.1)**	
	Right	8.9 (6.8)	13.4 (7.4)	**6.8 (4.5)**	8.6 (6.7)	11.8 (7.4)	**7.1 (5.7)**	

TABLE V: Performance in terms of EER (%) for portrait strokes on the Frank dataset [7]. Mean EER across users (standard deviation in brackets).

Scenario		**SVM**	**GMM**	**Fusion**
Intra-session	**Up**	6.9 (6.5)	11.0 (10.3)	**3.5 (4.9)**
	Down	6.9 (4.8)	10.7 (7.3)	**4.3 (3.5)**
	Left	5.6 (4.4)	8.4 (5.7)	**3.3 (3.2)**
	Right	5.3 (5.6)	7.5 (5.1)	**3.1 (3.0)**
Inter-session	**Up**	11.6 (12.5)	16.0 (15.1)	**8.1 (7.2)**
	Down	11.8 (12.8)	15.5 (15.3)	**8.4 (7.6)**
	Left	12.3 (7.8)	17.9 (10.2)	**10.2 (7.9)**
	Right	**9.9 (8.2)**	27.8 (27.0)	12.8 (15.6)
Combined sessions	**Up**	10.6 (11.3)	12.6 (9.2)	**5.9 (5.9)**
	Down	8.2 (6.6)	12.0 (7.6)	**5.2 (4.3)**
	Left	5.8 (4.7)	8.1 (5.7)	**3.1 (2.9)**
	Right	6.6 (5.2)	8.9 (7.0)	**3.8 (3.4)**

strokes perform better than horizontal and are the ones which deteriorate less.

Combined sessions. The best performing system is again the fusion based. The EER improves in comparison to the inter-session scenario and is in a close range to the intra-session, even obtaining a better EER for leftwards strokes. Rightwards strokes, which performed badly in the inter-session scenario using the statistical system, improve significantly, dropping the EER to 8.9%.

The results are better than the ones found with Serwadda *et al.* database [10], shown in Table IV. In the intra-session scenario, the results are in a similar range, but in the inter-session one they are improved, even though the parameters were fixed for the other database. Combining both sessions, the results improve more and are closer to the intra-session scenario, because this way the real variability is better represented.

3) Antal database: As this dataset is not divided in days, the inter-session scenario cannot be evaluated. In the same line as with the Frank database, due to the small number of users, landscape orientation is not evaluated. Likewise, upward strokes in portrait orientation cannot be evaluated, as only four subjects have at least 40 training samples. Table VI depicts the mean EER obtained across all users.

Intra-session scenario. The fusion system outperforms the other approaches in this situation. It is worth emphasizing that, despite still performing worse, the statistical system results are closer to the discriminative system's. This is probably caused by subjects having data more scattered, due to the use of multiple devices, instead of just one, which makes more difficult to separate them from impostors. Contrary to what happened in the previous databases in this scenario, downwards strokes obtain the best result.

TABLE VI: Performance in terms of EER (%) for portrait strokes on the Antal dataset [8]. Mean EER across users (standard deviation in brackets).

Scenario		**SVM**	**GMM**	**Fusion**
Intra-session	**Up**	-	-	-
	Down	4.4 (4.2)	8.5 (5.0)	**2.6 (2.5)**
	Left	9.7 (7.5)	10.3 (6.6)	**6.0 (5.3)**
	Right	10.9 (8.4)	12.8 (7.8)	**8.2 (6.7)**

TABLE VII: Performance (EER in %) for portrait strokes on the UMDAA-02 dataset [16]. Mean EER across users (standard deviation in brackets). For reference, in similar conditions compared to Combined Sessions, Mahbub et al. [16] obtain ca. 30% EER with another SVM-based approach.

Scenario		**SVM**	**GMM**	**Fusion**
Intra-session	**Up**	16.6 (10.8)	**10.2 (9.9)**	11.1 (10.2)
	Down	13.0 (9.6)	9.8 (7.1)	**9.5 (7.14)**
	Left	-	-	-
	Right	10.9 (9.2)	**3.6 (4.0)**	4.1 (4.5)
Combined sessions	**Up**	19.9 (8.0)	23.7 (8.4)	**16.3 (7.5)**
	Down	21.8 (6.6)	23.4 (6.3)	**17.8 (6.0)**
	Left	18.3 (10.4)	25.7 (9.3)	**18.2 (10.4)**
	Right	21.8 (11.4)	26.1 (10.4)	**18.5 (10.2)**

4) UMDAA-02 database: In this case the sessions are not grouped by days, so it is not possible to study the inter-session scenario (i.e., we cannot be sure that training-testing data come from different days). Similarly to the previous databases, very limited data is available from each user in landscape orientation, reason why it is not evaluated. For the same reason, upward swipes are not used in the intra-session scenario. In Table VII we show the performance obtained with this database.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIFS.2018.2833042, IEEE Transactions on Information Forensics and Security

Intra-session scenario. On the contrary to the other databases, the best-performing system is the statistical one. This is probably an indication, due to the longer use of the devices and getting used to them, of a small intra-user variability. Overall, the performance is in similar ranges as in other databases.

Combined sessions scenario. This scenario shows a much worse performance than the previous one and than other databases. A likely cause may be that there is bigger variation across sessions, as they span through two months. This is a motivation to not use all previous sessions from a user to train a model, but instead adapt the model only with recent data [20].

E. Experiments with Neural Networks

For comparative purposes we also studied the performance using a basic MultiLayer Perceptron Neural Network (MLP) on the Serwadda database in the inter-session scenario, following the same protocol to choose the impostor and genuine samples as with the discriminative system. We first tried as input the same 28 features used in the discriminative approach (Section III-A1) and observed the performance with different number of layers (1 or 2) and hidden units (5 to 30 in steps of 5). We obtained performances around 40% EER, with the best one being two hidden layers with 25 hidden units each (36% EER). This performance is much worse compared to the other systems, which in the worst case were around 20% EER and in the best were close to 11%.

We also used directly as input the x and y position coordinates of the touch trajectories and the velocity between each adjacent time sample, after size normalization (linear interpolation) in order to have fixed-length descriptors. The resulting performance is ca. 45% EER with similar MLP architectures as explored before. These results show that more complex architectures are needed for exploiting neural networks in this problem [30], [31].

VI. Analysis of the results

A. Error analysis across users

The statistical system classifies users modelling their behaviour and how their data is distributed, while the discriminative system only tries to separate the data without taking into account how it is distributed. Despite the fact that the discriminative system presents a slightly better performance than the statistical, one of the most important limitations it presents is not being able to authenticate well users who, although they are stable, have their data very dispersed. Fig. 7 shows the typical behaviour of one of these users, who obtains an EER around 50% in the SVM system, over two sessions. It can be observed that the train and test data are distributed similarly. Therefore, it's stable, but it has great variability and is not so condensed as other users' data.

Other example users are shown in Fig. 8. Top of Fig. 8 shows two users with good performance in the GMM system, in scatter plots of the two most discriminant features from Table 5. It can be observed that, for both users, their train and test data are distributed similarly. Nevertheless, one of them has data more dispersed than the other one, which leads us to believe that scattered data do not affect much this system's performance. On the other hand, in Fig. 8c and 8d we show the GMM model and data for two users who obtain a bad performance. In this case, even though the user model represents correctly the train data, the test data is distributed over other areas. This lack of stability makes the user model for the considered features not representative of the second session, and these subjects result in a bad EER in both SVM and GMM systems.

TABLE VIII: Comparison of the mean EER found in the inter-session scenario for the 10% worst performing users with SVM in the statistical and discriminative systems.

Scenario		SVM	GMM adaptation
Portrait	Up	57.76	32.56
	Down	56.14	25.93
	Left	41.75	20.07
	Right	53.91	30.15

With the Serwadda *et al.* database [10], employing the same 40 training samples for both systems, the performance in the inter-session scenario across the 10% worst users in the SVM system is compared with the performance in the system based on GMM adaptation. Only data from portrait orientation is used in this comparison. Table VIII depicts the results obtained. It can be observed that, while worse than the average for the statistical system, the performance is much better than in the discriminative system. Therefore, the statistical system is capable of better authenticating subjects that cannot be authenticated well with the discriminate system.

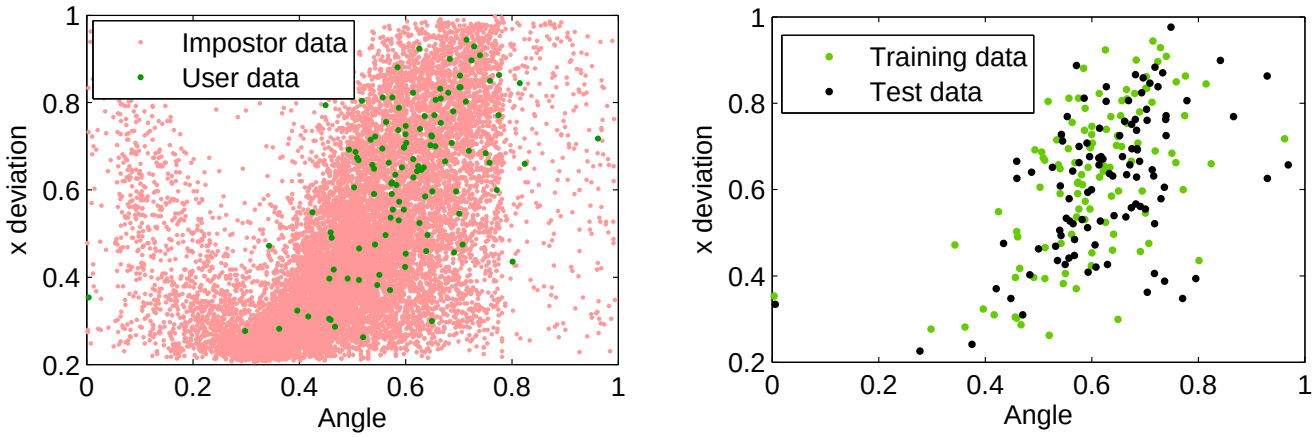
Hence, two types of users can be distinguished: those who are modelled better with SVM (well-behaved stable users) and those who are modelled better with GMM (ill-behaved unstable users). As the fusion system uses the information present in both of them, it is capable of representing these two types of users despite their different behaviours, obtaining an overall better performance.

These large differences between users [32] are in line with other behavioural biometrics like handwriting [33], and may be exploited with user-dependent processing techniques [24].

B. Session scenario analysis

The best performance is obtained in the intra-session scenario, where subjects are enrolled and authenticated within the same day. This is likely caused by users behaviour remaining stable in the same day and as a result, the variance is smaller than the one found across different days. In this situation, the fusion system always obtained the best results. Therefore, as explained in Section VI-A, the discriminative and statistical systems complement each other well. This was also observed in the inter-session scenario, where the individual systems obtained significantly worse results. The evaluation of the performance when both sessions data is combined and both form the training and test sets, shows that the EER improves, as the variability is smaller than across both sessions separated.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIFS.2018.2833042, IEEE Transactions on Information Forensics and Security



(a) User's data compared to other users.

(b) User's data in the training and test set.

Fig. 7: Typical behavior in users with an EER around 50% in the SVM system.

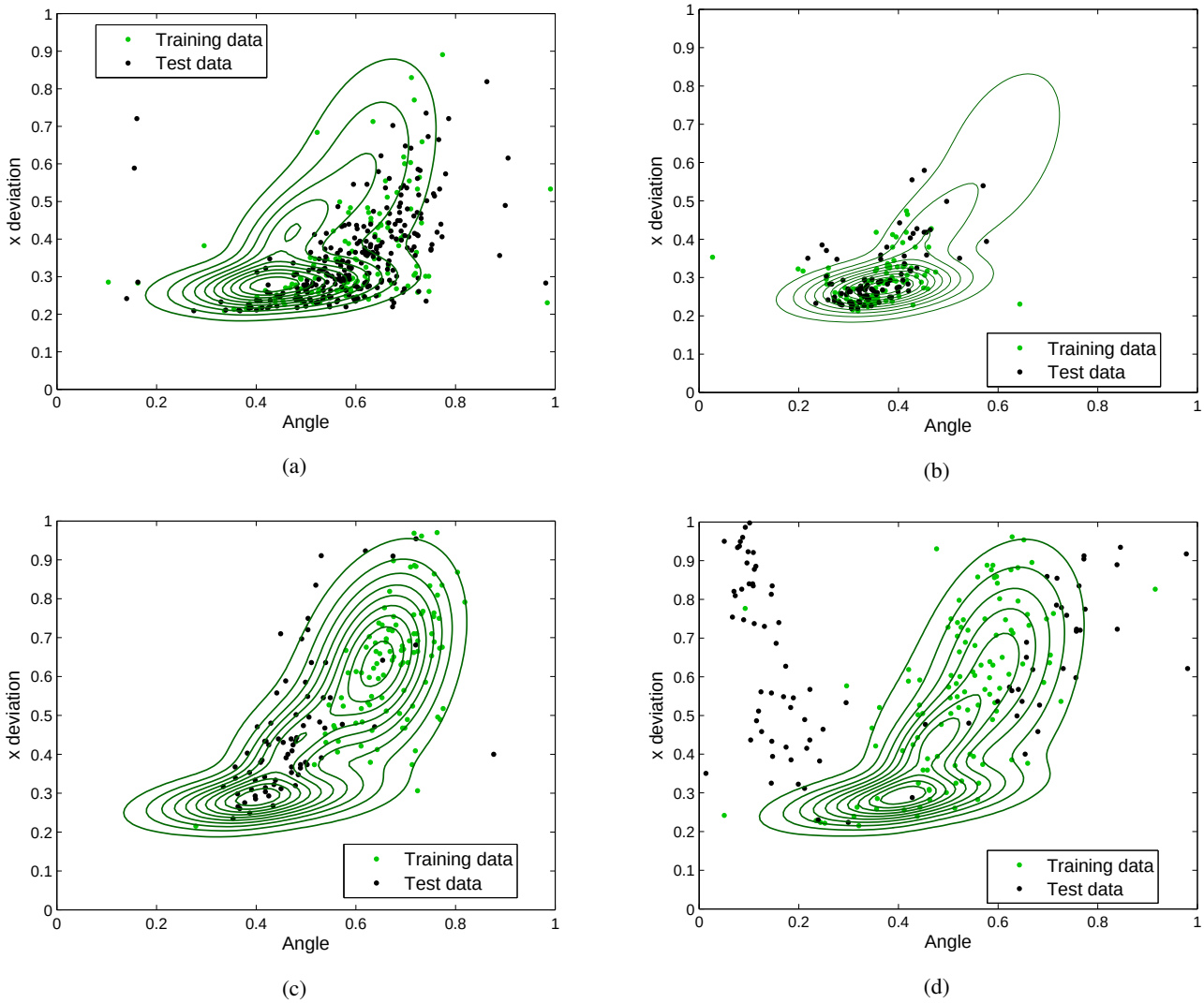


Fig. 8: Example of the GMM user model estimated for four different users overlapped with the data used for train (green) and test (black) in the inter-session scenario (i.e., train and test data were acquired in different sessions). (a) and (b) depict two users with good performance, while (c) and (d) depict two users who perform badly.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIFS.2018.2833042, IEEE Transactions on Information Forensics and Security

C. Computational complexity analysis

Despite its worse performance, the statistical system had the least computational cost, needing a smaller number of training swipes. In a real case scenario, where the user's model has to be obtained as fast as possible to protect the device, the time required to obtain the training swipes for each kind of operation should be, thus, as low as possible. Therefore, reducing the number of swipes to train the model gains importance.

D. Type of swipes analysis

The best performing operations were, in general, the horizontal gestures. Hence, these gestures probably hold more discriminative information and patterns are more distinctive across users. A possible cause is that they are performed more frequently and are more stable as a result. It can be also emphasized that the best EER is found with leftwards strokes in general. In addition, upwards strokes presented a higher lack of stability than other gestures. Considering that in most databases there was a small number of samples from these gestures, the most probable cause is that they are not performed often.

E. Device orientation analysis

Landscape oriented strokes obtained slightly worse results compared to portrait ones in the intra-session scenario. Nevertheless, landscape performance was better in the inter-session situation. This entails that swipes made in landscape orientation tend to be more stable across sessions. One possible cause is that users who swipe on the device in this orientation have developed more stable and consistent habits due to probably swiping almost always employing landscape orientation.

VII. Active Authentication

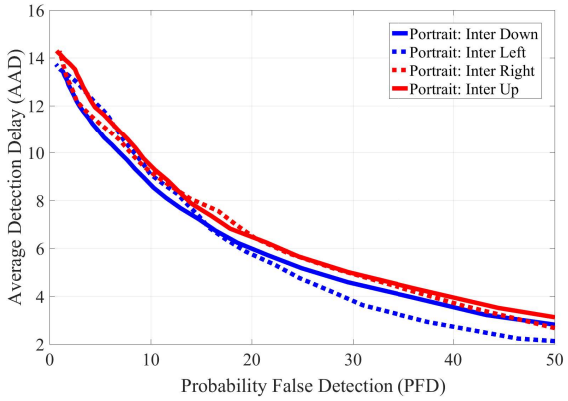
All previous results followed the same experimental structure: a user model was trained with various swipes, and then authentication was performed at a later stage by comparing the model to one or more swipes at the same time. Authentication error rates were then reported and analyzed using Equal Error Rates. That experimental framework, typically used in most related works (see Table I), represents a typical scenario in which the authentication is conducted at a given time instant. We can refer to these schemes as one-shot authentication.

On the other hand, active authentication schemes consider user authentication in a continuous manner [9]. Typical one-shot authentication (as studied in the present paper) can help in designing and evaluating active authentication schemes, but additional tools and performance measures are also needed to design and evaluate such continuous schemes.

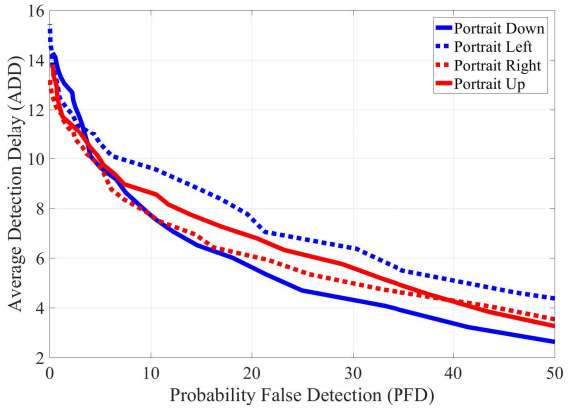
A comprehensive evaluation for active authentication of the datasets is out of the scope of the present paper. However, for completeness, we report a few selected results following the methodology for active authentication developed by Perera and Patel [9]. For evaluating active authentication schemes, they define the Average Detection Delay (AAD) as the number of samples necessary to detect an intruder, while the Probability of False Detection (PFD) represents the percentage of False Detections obtained for such AAD value. A curve PFD-AAD can be obtained by varying the detection threshold applied over the continuous detection score. The detection score is obtained by integrating the individual scores obtained for individual swipes. Several methods for such integration are explored by Perera and Patel [9], of which minimax Quickest Change Detection (QCD) performed best. Minimax QCD computes the detection score for M swipes as:

$$\max(\sum_{n=1}^{M} L_n, 0), \quad L_n = \frac{f_g(x_n)}{f_i(x_n)}, \tag{1}$$

where f_g and f_i are probability density functions of genuine and impostors scores calculated using the training set, and x_n is an individual score, computed as described in previous sections. Fig. 9 shows results of this approach on two of the databases studied.



(a) Serwadda dataset



(b) UMDAA-02 dataset

Fig. 9: Active authentication: Number of test swipes vs Probability of False Detection (in %)

Serwadda database was captured under more supervised and controlled conditions while UMDAA-02 is a more realistic dataset acquired in the wild. However, the results show similar performances for both databases. The larger number of samples available to model the users of UMDAA-02 helps to reduce the impact of the unsupervised scenario. The results

show that using 9 swipes, it is possible to detect intruders with a 10% False Detection Rate. These performances are encouraging and suggest the potential of swipe biometrics for active authentication of users.

VIII. Conclusions

This work has studied swipe biometrics employing usual interaction with a touchscreen. We began summarizing key recent related works from which we selected four benchmarks representative of practical applications. Three different systems have been explored, one discriminative using SVM, one statistical using GMM adapted from UBM and a third one based on their fusion. Their performance has been evaluated using the four public databases in three situations: intra-, inter-session and combining two sessions. It has been found that the fusion method is less sensible to intra-user variation and is capable of modelling all users, despite the differences in their behaviour. The experimental results show that an enrolment strategy that incorporates new session's data can mitigate this variance typical of behavioral biometrics and improve results.

The performance across the different touch operations (up, down, left and right) has been studied, reaching the conclusion that horizontal strokes tend to be more discriminative than vertical ones. In landscape orientation, the performance obtained for the inter-session scenario is better than in portrait orientation. Additionally, various other experiments have been conducted providing insights on the best recognition approaches and their practical results in the four selected realistic benchmarks. Resources for reproducing the benchmark results obtained are available online[2].

Future work includes studying the lack of stability of touch data [11] and template update techniques [20], better methods for integrating in time the information provided by individual strokes [1], [24], the use of gestures made with multiple fingers [34], and evaluating possible attacks to these systems [4], [35]. Additionally, touchpad input is also of interest [36], and new learning architectures capable of exploiting this kind of touch signals are encouraging [30], [31].

Acknowledgment

Funding from Cecabank, project CogniMetrics (TEC2015-70627-R) and contract IJCI-2015-24742 (MINECO/FEDER). Part of the work was conducted during a research stay of J.F. at Imperial College London (PRX16/00580).

References

[1] 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*, vol. 33, no. 4, pp. 49–61, July 2016.

[2] J. Galbally, I. Coisel, and I. Sanchez, "A new multimodal approach for password strength estimation. part i: Theory and algorithms," *IEEE Transactions on Information Forensics and Security*, vol. 12, no. 12, pp. 2829–2844, Dec 2017.

[3] M. Martinez-Diaz, J. Fierrez, and J. Galbally, "Graphical password-based user authentication with free-form doodles," *IEEE Trans. on Human-Machine Systems*, vol. 46, no. 4, pp. 607–614, August 2016.

[4] A. Hadid, N. Evans, S. Marcel, and J. Fierrez, "Biometrics systems under spoofing attack: an evaluation methodology and lessons learned," *IEEE Signal Processing Magazine*, vol. 32, no. 5, pp. 20–30, September 2015.

[5] T. Sim, S. Zhang, R. Janakiraman, and S. Kumar, "Continuous verification using multimodal biometrics." *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 4, pp. 687–700, 2007.

[6] Z. Akhtar, A. Hadid, M. Nixon, M. Tistarelli, J.-L. Dugelay, and S. Marcel, "Biometrics: in search of identity and security: (q & a)," *IEEE MultiMedia*, June 2017.

[7] M. Frank, R. Biedert, E. Ma, I. Martinovic, and D. Song, "Touchalytics: On the applicability of touchscreen input as a behavioral biometric for continuous authentication," *IEEE Transactions on Information Forensics and Security*, vol. 8, no. 1, pp. 136–148, 2013.

[8] M. Antal, Z. Bokor, and L. Z. Szabó, "Information revealed from scrolling interactions on mobile devices," *Pattern Recognition Letters*, vol. 56, pp. 7–13, April 2015.

[9] P. Perera and V. M. Patel, "Efficient and low latency detection of intruders in mobile active authentication," *IEEE Trans. on Information Forensics and Security*, vol. 13, no. 6, pp. 1392 – 1405, June 2018.

[10] A. Serwadda, V. V. Phoha, and Z. Wang, "Which verifiers work?: A benchmark evaluation of touch-based authentication algorithms," in *Proc. IEEE BTAS*, 2013, pp. 1–8.

[11] H. Xu, Y. Zhou, and M. R. Lyu, "Towards continuous and passive authentication via touch biometrics: An experimental study on smartphones," in *Proc. SOUPS*, 2014, pp. 187–198.

[12] H. Zhang, V. M. Patel, M. Fathy, and R. Chellappa, "Touch gesture-based active user authentication using dictionaries," in *Proc. of the IEEE Winter Conference on Applications of Computer Vision*, 2015, pp. 207–214.

[13] S. Mondal and P. Bours, "Swipe gesture based continuous authentication for mobile devices," in *Proc. IAPR Intl. Conf. on Biometrics*, 2015.

[14] R. Murmuria, A. Stavrou, D. Barbará, and D. Fleck, "Continuous authentication on mobile devices using power consumption, touch gestures and physical movement of users," in *International Workshop on Recent Advances in Intrusion Detection*. Springer, 2015, pp. 405–424.

[15] R. Kumar, V. V. Phoha, and A. Serwadda, "Continuous authentication of smartphone users by fusing typing swiping and phone movement patterns," in *Proc. IEEE BTAS*, 2016, pp. 1–8.

[16] U. Mahbub, S. Sarkar, V. M. Patel, and R. Chellappa, "Active user authentication for smartphones: A challenge data set and benchmark results," in *Proc. IEEE BTAS*, 2016.

[17] C. Shen, Y. Zhang, X. Guan, and R. A. Maxion, "Performance analysis of touch-interaction behavior for active smartphone authentication," *IEEE Transactions on Information Forensics and Security*, vol. 11, no. 3, pp. 498 – 513, March 2016.

[18] Z. Sitov, J. ednka, Q. Yang, G. Peng, G. Zhou, P. Gasti, and K. S. Balagani, "Hmog: New behavioral biometric features for continuous authentication of smartphone users," *IEEE Transactions on Information Forensics and Security*, vol. 11, no. 5, pp. 877–892, May 2016.

[19] R. Kumar, P. P. Kundu, and V. V. Phoha, "Continuous authentication using one-class classifiers and their fusion," in *Proc. IEEE Intl. Conf. on Identity, Security, and Behavior Analysis*, 2018.

[20] J. Galbally, M. Martinez-Diaz, and J. Fierrez, "Aging in biometrics: An experimental analysis on on-line signature," *PLOS ONE*, vol. 8, no. 7, p. e69897, July 2013.

[21] A. Pozo, J. Fierrez, M. Martinez-Diaz, J. Galbally, and A. Morales, "Exploring a statistical method for touchscreen swipe biometrics," in *Proc. International Carnahan Conference on Security Technology, ICCST 2017*, October 2017, pp. 1–4.

[22] M. Martinez-Diaz, J. Fierrez, and J. Ortega-Garcia, "Universal background models for dynamic signature verification," in *Proc.IEEE BTAS*, September 2007, pp. 1–6.

[23] J. Fierrez-Aguilar, D. Garcia-Romero, J. Ortega-Garcia, and J. Gonzalez-Rodriguez, "Bayesian adaptation for user-dependent multimodal biometric authentication," *Pattern Recognition*, vol. 38, no. 8, pp. 1317–1319, August 2005.

[24] J. Fierrez, A. Morales, R. Vera-Rodriguez, and D. Camacho, "Multiple classifiers in biometrics. part 2: Trends and challenges," *Information Fusion*, vol. 44, pp. 103–112, November 2018.

[25] X. Zhao, T. Feng, W. Shi, and I. A. Kakadiaris, "Mobile user authentication using statistical touch dynamics images," *IEEE Trans. Information Forensics and Security*, vol. 9, no. 11, pp. 1780–1789, 2014.

[26] M. Martinez-Diaz, J. Fierrez, R. P. Krish, and J. Galbally, "Mobile signature verification: Feature robustness and performance comparison," *IET Biometrics*, vol. 3, no. 4, pp. 267–277, December 2014.

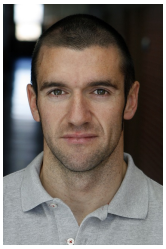
[2]https://atvs.ii.uam.es/atvs/databases.jsp

[27] A. Jain, K. Nandakumar, and A. Ross, "Score normalization in multimodal biometric systems," *Pattern Recognition*, vol. 38, no. 12, pp. 2270 – 2285, 2005.
[28] J. Galbally, J. Fierrez, and J. Ortega-Garcia, "Performance and robustness: a trade-off in dynamic signature verification," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing, ICASSP*, March-April 2008, pp. 1697–1700.
[29] J. Fierrez, A. Morales, R. Vera-Rodriguez, and D. Camacho, "Multiple classifiers in biometrics. part 1: Fundamentals and review," *Information Fusion*, vol. 44, pp. 57–64, November 2018.
[30] X. Y. Zhang, G. S. Xie, C. L. Liu, and Y. Bengio, "End-to-end online writer identification with recurrent neural network," *IEEE Transactions on Human-Machine Systems*, vol. 47, no. 2, pp. 285–292, April 2017.
[31] R. Tolosana, R. Vera-Rodriguez, J. Fierrez, and J. Ortega-Garcia, "Exploring recurrent neural networks for on-line handwritten signature biometrics," *IEEE Access*, vol. 6, pp. 5128–5138, February 2018.
[32] N. Yager and T. Dunstone, "The biometric menagerie," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, pp. 220–230, 2008.
[33] J. Fierrez-Aguilar, J. Ortega-Garcia, and J. Gonzalez-Rodriguez, "Target dependent score normalization techniques and their application to signature verification," *IEEE Trans. on Systems, Man and Cybernetics - Part C*, vol. 35, no. 3, pp. 418–425, August 2005.
[34] N. Sae-Bae, N. Memon, K. Isbister, and K. Ahmed, "Multitouch gesture-based authentication," *IEEE Trans. on Information Forensics and Security*, vol. 9, no. 4, pp. 568–582, April 2014.
[35] H. Khan, U. Hengartner, and D. Vogel, "Targeted mimicry attacks on touch input based implicit authentication schemes," in *Proc. of the 14th Annual International Conference on Mobile Systems, Applications, and Services*, 2016, pp. 387–398.
[36] A. Chan, T. Halevi, and N. Memon, "Touchpad input for continuous biometric authentication," in *Proc. of the Communications and Multimedia Security*, 2014, pp. 86–91.

Julian Fierrez received the MSc and the PhD degrees in telecommunications engineering from Universidad Politecnica de Madrid, Spain, in 2001 and 2006, respectively. Since 2002 he has been affiliated with the Biometric Recognition Group, first at Universidad Politecnica de Madrid, and since 2004 at Universidad Autonoma de Madrid, where he is currently an Associate Professor. From 2007 to 2009 he was a visiting researcher at Michigan State University in USA under a Marie Curie fellowship. His research interests include signal and image processing, pattern recognition, and biometrics, with emphasis on multi-biometrics, biometric evaluation, system security, forensics, and mobile applications of biometrics. Prof. Fierrez has been actively involved in multiple EU projects focused on biometrics (e.g. TABULA RASA and BEAT), has attracted notable impact for his research, and is the recipient of a number of distinctions, including: EAB European Biometric Industry Award 2006, EURASIP Best PhD Award 2012, Miguel Catalan Award to the Best Researcher under 40 in the Community of Madrid in the general area of Science and Technology, and the 2017 IAPR Young Biometrics Investigator Award. He is Associate Editor for IEEE Trans. on IFS and IEEE Trans. on Image Processing.

Ada Pozo received the BSc in Electrical Engineering from Universidad Autonoma de Madrid, Spain, in 2017. Since fall 2017 she will be studying the MSc in Computer Science at EPFL, Swizerland, thanks to the 2017 scholarship for abroad postgraduate studies from Mutua Madrileña Foundation. In 2016, she was awarded a grant as Research Assistant from MECD with the Biometric Recognition Group at Universidad Autonoma de Madrid. Her current research interests include pattern recognition and signal processing.

Marcos Martinez-Diaz received the MSc and PhD degrees in telecommunication engineering from the Universidad Autonoma de Madrid, Spain, in 2006 and 2015 respectively. From 2008 and 2016 he has worked in different IT management positions as a consultant and in a top-tier telecom company. He has now joined the Spanish Publich Administration as public servant focusing on IT and Telecom functions. Since 2005 he has collaborated with the Biometric Recognition Group at Universidad Autonoma de Madrid. His research interests include biometrics, pattern recognition, and signal processing primarily focused on signature verification and graphical passwords. He is the recipient of a number of distinctions such as the Honeywell Honorable Mention as Best Student at IEEE Intl. Conf. on BTAS 2007 and the EAB European Biometrics Industry Award 2014.



Javier Galbally received the MSc degree in Electrical Engineering from the Universidad de Cantabria, Spain, in 2005, and the PhD degree in Electrical Engineering from the Universidad Autonoma de Madrid, Spain, in 2009, where he was an Assistant Professor until 2012. In 2013, he joined the European Commission in DG Joint Research Centre, where he is currently a Scientific Project Officer. His research interests are mainly focused on the security evaluation of biometric systems, pattern and biometric recognition, synthetic generation of biometric traits, and inverse biometrics. He is the recipient of a number of distinctions, including the IBM Best Student Paper Award at Intl. Conf. on Pattern Recognition 2008, finalist of the EBF European Biometric Research Award 2009, Best PhD Thesis Award by the Universidad Autonoma de Madrid 2010, and Best Paper Award at IAPR/IEEE Intl. Conf. on Biometrics 2015.

Aythami Morales received the MSc degree in telecommunication engineering from the Universidad de Las Palmas de Gran Canaria in 2006 and the PhD degree from the La Universidad de Las Palmas de Gran Canaria in 2011. Since 2017 he is an interim Associate Professor at the Universidad Autonoma de Madrid. He has conducted research stays at the Biometric Research Laboratory - Michigan State University, the Biometric Research Center - Hong Kong Polytechnic University, the Biometric System Laboratory - University of Bologna, and the Schepens Eye Research Institute. His research interests are focused on pattern recognition, computer vision, machine learning, and biometrics signal processing. He is author of more than 70 scientific articles published in international journals and conferences. He has received awards from ULPGC, La Caja de Canarias, SPEGC, and COIT. He has participated in National and EU projects in collaboration with other universities and private entities such as UAM, UPM, EUPMt, Indra, Union Fenosa, Soluziona or Accenture.