IET Biometrics

Special Issue: Selected Papers from the International Workshop on Biometrics and Forensics (IWBF2016)

Signature authentication based on human intervention: performance and complementarity with automatic systems

ISSN 2047-4938
Received on 18th July 2016
Revised 26th December 2016
Accepted on 12th January 2017
doi: 10.1049/iet-bmt.2016.0115
www.ietdl.org

Aythami Morales[1] ✉, *Derlin Morocho*[2], *Julian Fierrez*[1], *Ruben Vera-Rodriguez*[1]
[1]*BiDA Lab – Biometrics and Data Pattern Analytics Laboratory, Universidad Autonoma de Madrid, Madrid, Spain*
[2]*Departamento de Electrica y Electronica, Universidad de las Fuerzas Armadas-ESPE, Sangolquí, Ecuador*
✉ *E-mail: aythami.morales@uam.es*

Abstract: This work explores human intervention to improve Automatic Signature Verification (ASV). Significant efforts have been made in order to improve the performance of ASV algorithms over the last decades. This work analyzes how human actions can be used to complement automatic systems. Which actions to take and to what extent those actions can help state-of-the-art ASV systems is the final aim of this research line. The analysis at classification level comprises experiments with responses from 500 people based on crowdsourcing signature authentication tasks. The results allow to establish a human baseline performance and comparison with automatic systems. Intervention at feature extraction level is evaluated using a self-developed tool for the manual annotation of signature attributes inspired in Forensic Document Experts analysis. We analyze the performance of attribute-based human signature authentication and its complementarity with automatic systems. The experiments are carried out over a public database including the two most popular signature authentication scenarios based on both online (dynamic time sequences including position and pressure) and offline (static images) information. The results demonstrate the potential of human interventions at feature extraction level (by manually annotating signature attributes) and encourage to further research in its capabilities to improve the performance of ASV.

1 Introduction

The signature is worldwide accepted as an identity authentication method and it has been used by several cultures over the past 2000 years. The signature is a behavioural biometric trait which comprises neuromotor characteristics of the signer (e.g. our brain and muscles among other factors define the way we sign) as well as socio-cultural influence (e.g. the western and asian styles). During centuries, the examination of signatures has been made by experts who determine the authenticity of the sample based on forensic analysis. Recently, automatic signature verification (ASV) systems have emerged as a feasible way to automate the traditional signature authentication method made by forensic document examiners (FDEs) [1–3]. The variety of AS authentication applications is large. The AS authentication literature is commonly divided into online and offline systems depending on the nature of the data and the applications [2]:

- *Offline or static signature authentication:* The signatures are performed using an ink-pen and the information is digitalised by optical scanners. The authentication is executed by analysing the visual characteristics of the signature including morphology, texture, and geometry. The potential applications are mostly related with document analysis.
- *Online or dynamic signature authentication:* The signatures are acquired with digital devices which capture the temporal sequences of the signing process. The authentication is executed based on global parameters (e.g. total time and number of pen-ups) or temporal functions derived from the acquired sequences (e.g. velocity and acceleration). The applications of this type include those related with automatic authentication systems (e.g. point of sales, delivery services and mobile authentication).

In most of these applications, humans usually supervise the signing process but their responsibilities are mostly limited to guaranteeing a valid acquisition without any contribution to authentication. These supervisors do not usually have the specific experience of FDEs and they will be referred to as laymen in the rest of this work. Without specific training and considering that signature authentication is not the principal job assignment of the above-mentioned laymen, their performance is an open question. Fig. 1 tries to illustrate the difficulties related with this task. The deployment of automated systems is eliminating human intervention in many authentication applications. However, the abilities of humans should not be undervalued and there is large room for improvement of automatic methods by incorporating human intervention in some scenarios. Some of these scenarios where a layman may help or contribute to AS authentication are banking, point of sales, notary public, or parcel delivery. We advocate for the consideration of human interaction in these scenarios due to the particularities of the signature as a behavioural biometric. As it has been demonstrated [3], the biometric information of the signature (used to recognise the authenticity) fluctuates severely for different users and acquisition conditions. Our aim in this research line of human interactions in automatic systems is to alleviate such fluctuations with simple actions a layman can take in many scenarios of practical importance. Which actions to take and to what extent those actions can help state-of-the-art ASV systems is the final aim of this research line.

Human-assisted schemes in biometrics take advantage of both human skills and automated system capabilities [4–8]. The human intervention on biometric systems can be done at different levels (see Fig. 2) according to the different tasks to be realised: at image level (e.g. quality assessment to discard samples with large distortions); at feature level (e.g. manual annotation of discriminative attributes); at matching or classification level (e.g. human ratings in the form of scalar values); and finally at decision level (e.g. binary decision, genuine, or fake). In this work, we will focus on two specific types of interventions: (a) human ratings that measure the perceived authenticity (intervention at classification level) and (b) manual annotation of attributes (intervention at feature level) used as input of an automatic classification system.

The study of human performance on biometric applications is not new and it helps to better understand the potential and capabilities of automatic systems [9, 10]. Human skills are commonly used as benchmark for the evaluation of automatic algorithms [11, 12]. Some biometric characteristics are more suitable than others for these evaluations. In general, biometric

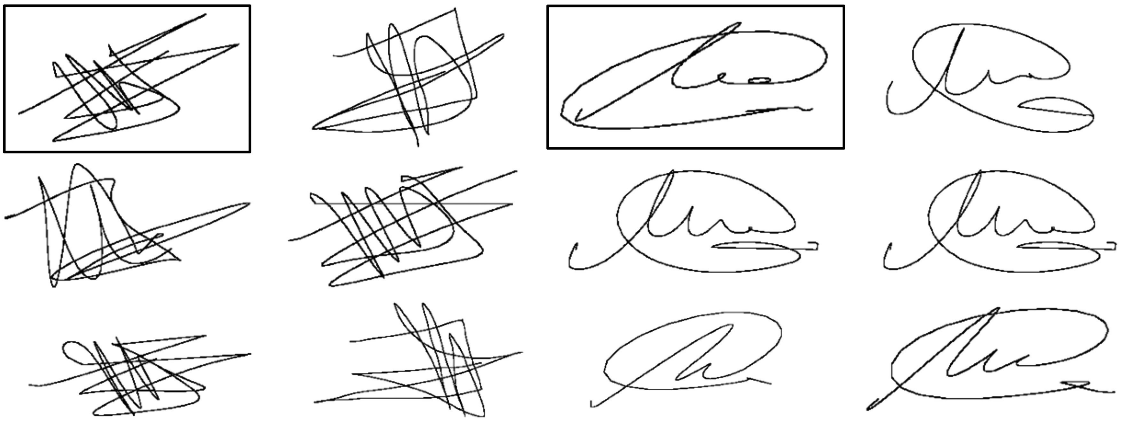

Fig. 1 *Mixed genuine signatures and forgeries (made by other people after practising for 2 min). Which signatures are genuine? The rectangles highlight two genuine signatures as gallery samples. See the rest of the labels at [Solution to Fig. 1: From left to right. Top: genuine, forgery, genuine, forgery; centre: forgery, genuine, forgery, forgery; and bottom: genuine, forgery, genuine, genuine.]*

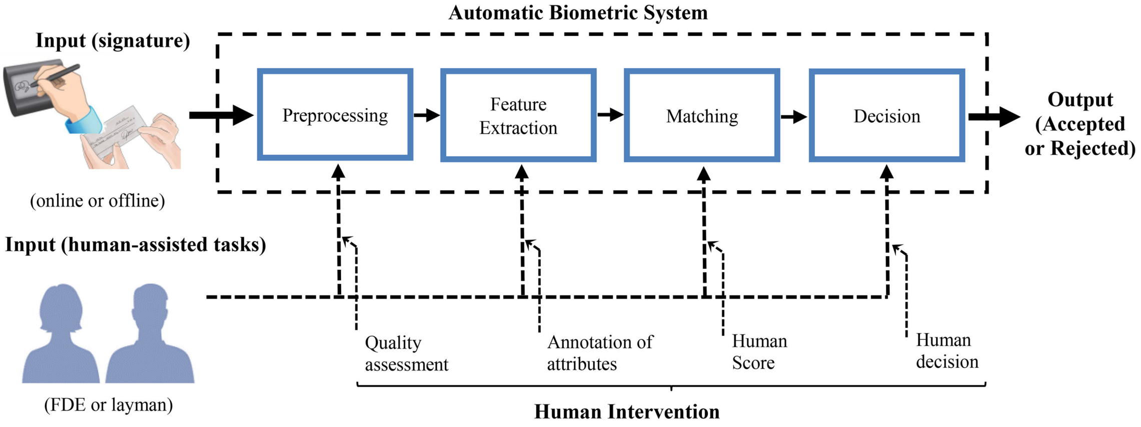

Fig. 2 *Example of human-assisted signature authentication scheme*

traits such as face, signature, or voice will be better recognised by humans than other characteristics such as fingerprint, iris, and palmprint. However, some studies reveal that humans can be highly inaccurate at recognising biometric characteristics such as faces from unfamiliar people [9, 12]. This means that we can easily recognise the face or voice of friends and our own signature but we fail more often at recognising the face, voice, or signatures from unfamiliar people.

The research community has investigated different ways to exploit the human skills in biometric applications. The use of human annotations in automatic biometric authentication systems has provided encouraging results in the literature [8]. Visual attribute annotation made by humans has emerged as a way to improve automatic authentication systems in face [4, 6–10], gait [5], and signature authentication [13]. The attributes can be defined as an extensive vocabulary of visual attributes (low-level image features) that can be used to label a large dataset of images. A set of attributes can be used to train models or classifiers and recognise classes (based on the presence or absence of these attributes). Some of these attributes used for biometric applications are known as soft biometrics (e.g. gender, ethnicity, age, hair colour, complexion, or height among many others). An attribute, as well as soft biometrics, reveals information about the individual but this is not able to authenticate him/her because of its lack of uniqueness and permanence [14]. However, the combination of multiple attributes can be used to improve the overall uniqueness and generate models capable to recognise users [15].

Human intervention in signature authentication has been historically related to forensic sciences. On the basis of their training and experience, FDEs analyse the authenticity of a given signature according to a set of evidences. The attribute annotation of signatures is a common task in FDEs analysis and it consists of either discrete labels (e.g. the signature has proper punctuation) or scalar measures of specific characteristics (e.g. a stroke length of 6 mm). Oliveira *et al.* [16] analysed the performance of graphology features in AS authentication with promising results over a dataset with 5600 signatures from 60 writers (including genuine samples, simple forgeries, and simulated forgeries). Malik *et al.* compared the performance of FDEs and AS authentication systems for disguised signatures when the owner of the signature introduces changes in his/her signature in order to mask his/her identity [4]. The results obtained in their study suggest that FDEs can achieve similar performance to automatic systems with accuracies over 90%. FDEs are well trained to analyse the authenticity of signatures and their performance is usually high classifying genuine and forged samples [11]. Although the experience of the expert is also exploited, the work of FDEs is mostly based on well-defined protocols and methodologies. The set of features proposed by FDEs is large [16, 17] and their evaluations are based on evidences that support their final opinion. The results of FDE evaluations are therefore a mix of experience, training, and personal subjectivity. It is reasonable to assume that the analysis performed by a non-FDE human (excluding the experience and training) is mostly based on the personal subjectivity of each subject. While the baseline performance of FDEs has been analysed in the literature [18, 19], to the best of our knowledge, the literature lacks of studies analysing the baseline performance of laymen [13, 20–22]. Crowdsourcing was employed in [21] in order to establish a human baseline performance on signature authentication. The experiments reported over responses from 150 laymen shown performances ranging from 7% (false acceptance rate) to 80% (false rejection rate) depending on the scenario and the information provided to the users. Signature authentication based on human annotations made by laymen (intervention at feature level) was proposed in [13]. The experiments included annotations of 13 attributes (inspired by FDE analysis) made by one layman on samples from 30 different signers. The results reported suggest the potential of human annotations to improve

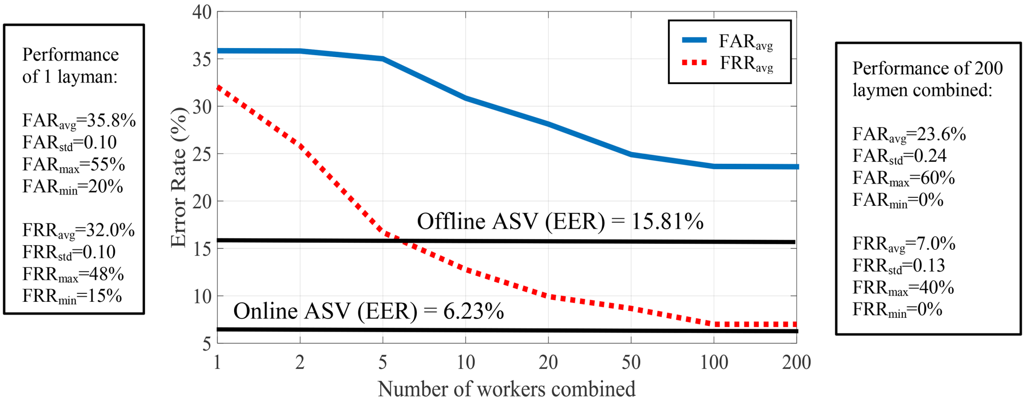

Fig. 3 *Human baseline performance (the curves present the averaged FAR and FRR) according to the number of workers combined and performance (in terms of EER) of ASV baseline systems*

ASV with improvements between 25 and 90%. Even though these works represent valuable contributions, the literature suffers from three important shortcomings: (i) lack of studies about the performance of laymen (non-expert humans) for large populations; (ii) comprehensive framework including both online and offline signatures; and (iii) proposals to improve the performance of ASV systems through the human intervention.

The present work is a step forward in the analysis of the potential of human intervention to improve AS authentication. The contributions of this work are three-fold: (i) we extended the experiments presented in [21] with responses from 500 laymen, providing new insights in the performance of humans in signature authentication tasks based on crowdsourcing experiments; (ii) we present a deep analysis of the attribute-based signature authentication system proposed in [13] via human interventions made by 11 different annotators and the complete BiosecurID-UAM corpus (3696 signatures from 132 different signers); and (iii) we analyse the performance of combined schemes incorporating the proposed attribute-based manual approach to online and offline state-of-the-art ASV systems.

The rest of this paper is organised as follows: Section 2 presents our work to establish a human baseline performance based on human interventions at classification level. Section 3 describes the proposed manual attribute-based signature authentication developed to analyse human intervention at feature extraction level. Section 4 reports the experiments and results. Finally, Section 5 summarises the conclusions and future works.

2 Human intervention at classification level: human baseline performance

It is relatively easy to recognise our signature from signatures made by others but several studies have probed how difficult it is to recognise the authenticity of signatures different to ours [11, 20–22]. In [11, 20], the ability of 22 individuals to recognise signatures was evaluated using 765 signatures from 51 writers (432 genuine and 333 forgeries). The results obtained suggest that people perform worst than state-of-the-art offline signature authentication automatic systems (Hidden Markov Models (HMM)-based equal error rate of ∼12%). The performance of human (FDE experts) and offline ASV systems was evaluated in [18]. The study found that even experts have difficulties to recognise some cases (especially disguised signatures). Using a common framework, the performance of human experts and automatic systems show similar accuracies. In [21], Amazon Mechanical Turk was employed to establish a human baseline performance in signature authentication. The results obtained from 150 subjects suggest that laymen performance is poor with False Acceptance Rate (FAR) and False Rejection Rate (FRR) up to 30%. However, the results improved when the responses of different laymen were combined, which suggests that errors (especially false rejection) are not equally distributed between people. In [22], the researchers analyse the combination of offline signature verification systems and human decisions. The authors propose a combined scheme based on a pool of optimal human–machine actions that minimises the error curves (in the form of Receiver Operating Curve (ROC)). About 23 amateur humans evaluated the authenticity (intervention at decision level) of 765 test signatures (same database as [11, 20]). The results presented suggest that combination of human responses and AS verifiers can be used to improve the unaided schemes (only humans or only machine).

In the present work, we develop an extension of the experiment proposed in [21]. The new experiment comprises 160 (80 genuine and 80 forgeries) signatures made by 20 different signers from BiosecurID database [23] and 500 workers who provide a confidence value between 1 and 10 related with the perceived authenticity of a query signature (1 = I am sure this is a forgery; 10 = I am sure this is a genuine signature). Four genuine samples (known authenticity) are shown in addition to the query sample (unknown authenticity). The decision threshold is set to 5. Fig. 3 shows the average performance of individual workers and the performance obtained when different responses from different workers are combined (using the mean rule). Comparing human performance by aggregate human ratings is a standard protocol [12]. The responses of workers can be fused to determine the complementarity potential of the human abilities. In addition to the human performance, the performance of two baseline ASV systems [24, 25] are evaluated (in terms of Equal Error Rate (EER)) on the same dataset. Note that only static images of the signatures are shown to the laymen. Human ranks and ASV scores are not combined in this experiment. A brief description of both systems is given below.

Online system [24, 26]: A function-based dynamic time warping algorithm (ranked among top three algorithms in international technology evaluations [27, 28]). The algorithm is based on Dynamic Time Warping (DTW) [26] applied to functions of time sequences extracted from each signature. A set of seven time functions are derived from $[\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{p}]$ sequences. The sequences were selected after feature selection (based on the performance of the feature set) from a larger set of sequences defined in [24]. The DTW algorithm matches two different sets of sequences based on the Euclidean distance between the time functions. The classification score is obtained as the average distance between one test signature and the enrolled set.

Offline system [25, 29]: Local binary patterns (LBPs) and local directional patterns (LDPs) are used to characterise the signature regions (12 overlapping blocks for each signature). Discrete cosine transform is applied to reduce the dimensionality of the feature vectors and two different least squares support vector machine (LSSVM) classifiers are trained using each of the feature sets (LBP and LDP features). The final score is computed as the sum of the two LSSVM scores coming from each of the classifiers. The offline system is applied on the offline versions of the signature samples (the BiosecurID database [23] comprises both the online and offline versions of the same signature samples).

The results show FAR and FRR up to 30% when the responses of individual workers are evaluated. However, the combination of

Table 1 Taxonomy of some of the most popular features used in FDE analysis: (a) morphological; (b) dynamics; (c) writer ability; and (d) writing style

Morphological	Dynamics	Writer ability	Writing style
proportionality	speed	hesitation	shape
slant or slope	overall pressure	enlargements	formatting
alignment to the baseline	local pressure	skill	method of production
text loops	slowness	tremor	embellishments
flourish characteristics	stops	arrangement	handedness
size	sudden endings	retouching	cross strokes and dots
character spacing	—	legibility	entry/exit strokes
stroke lengths	—	freedom of execution	punctuation
direction of strokes	—	simplification	connections
order	—	range of variation	emphasis
—	—	pen hold	capitalisation
—	—		handedness

different workers clearly outperforms the individual results with absolute improvements of 25 and 12% for FRR and FAR, respectively. These performance metrics are calculated averaging the performance obtained by different sets of laymen. Fig. 3 shows the maximum and minimum performances when the responses of 1 and 200 laymen are analysed. The results show how even combining the responses of laymen, the differences between the best and the worst set of users are clear (see minimum and maximum values). The large improvement obtained for the FRR suggests that errors produced by the authentication of genuine signatures are not equally distributed among different workers. Regarding the comparison with ASV, the individual responses of laymen are far in terms of performance (higher error rates). However, the performance obtained by aggregate opinions suggests the potential of human interventions when a large number of responses are combined (regardless of the practical limitations of these combinations in real applications). There are two important conclusions: (i) laymen are highly inaccurate on signature authentication when they do not have specific training or tools designed to improve their performance and (ii) the combination of responses from different laymen can be used to improve the performance, especially false rejection.

3 Human intervention at feature level: attribute-based signature authentication

The human baseline performance obtained suggests the necessity of specific training or guidance to improve the performance of laymen. The annotation of attributes inspired by FDE analysis is a way to provide information capable to authenticate signatures in a semi-automatic scheme. In this work, we employ the tool for the manual annotation of signature attributes presented in [13, 21]. The application is a MATLAB Graphic User Interface (GUI) self-developed for Windows computers (Intel Core-i3, 8 GB random access memory). The application is designed to be used by a human without the previous experience on FDE analysis or signature authentication tasks. About 13 attributes are annotated from a unique static binary image of the signature (each signature is annotated separately). While dynamic data is highly discriminative for AS authentication, our previous studies suggest that layman performs better on static images rather than dynamic videos or sequences [21]. The features annotated using the application are described below.

3.1 Signature attributes

The list of features of a signature used in signature authentication either on forensic or automatic scenarios is large [16, 17, 24]. These features can be classified according to different criteria as the school (e.g. mimic or symbolic), the nature of the information (e.g. graphology or graphometry), or the different signature parts (e.g. flourish, text), among others. Table 1 presents a taxonomy of the most popular features analysed in the FDE literature.

Owing to the large number and variety of existing features, we have selected a set of 13 attributes (inspired in the FDE analysis) on the basis of two characteristics: (i) efficiency: the annotation of the attributes must be fast for a layman without any FDE experience and (ii) performance: the attributes must be discriminative and useful for signature authentication. While efficiency is easy to estimate, the performance is certainly unknown. The final selection was based on a preliminary experiment performed with 1 annotator and 840 signatures (30 signers × 28 samples). Some of the attributes discarded because of its low performance in accordance with the time necessary for the labelling were connections, capitalisation or cross strokes. We divided the final set of features into two groups:

- *Categorical attributes (A1–A9):* denoted by discrete labels (e.g. spaced/concentrated signature).
- *Scalar measures (A10–A13):* which are calculated according to representative keypoints manually located (e.g. distance between characters or strokes). The keypoint selection reflects the human ability to highlight the most representative signature regions.

The features are strongly related to the signature content (text and flourish) and include both global information related with the whole aspect and local information related with specific strokes. The set of features allow to explore how the human perception can help to improve automatic systems. Guidelines (in the form of a few examples) are shown to the annotator to obtain more consistent features. However, the annotation of the attributes is a subjective task and some values can vary between annotators (see the stability analysis performed at Section 3.3). Listed below are the features chosen and a brief description based on the principles given above (see [16, 11, 17, 18] for details):

A1. Shape (*rounded strokes, vertical strokes, horizontal strokes, calligraphic model, vertical, and horizontal strokes,* or *unknown*): This attribute defines the most predominant orientation of the strokes of the signature including both text and flourish. This attribute is strongly related to the writing style of the signer.

A2. Proportionality (*proportional, unproportioned, mixed,* or *unknown*): The proportion defines the symmetry and size of the writing characters (typically the given name and the family name of the signer). It is strongly biased by the nature of the signature (e.g. text-based, text + flourish, and flourish-based).

A3. Text loops (*round, sharp,* or *unknown*): **P**redominant style of the loops (typical in letters such as ‘l, g, p, f, j, y’, and others) and directional changes (typical in uppercase letters such as ‘A, M, N’ and others).

A4. Order (clear order, confusing, concentrated, or *spaced)*: This attribute refers to the graphic distribution of the parts that form the signature. Some authors refer to this attribute as complexity and it is related with the number of trajectory intersections and density distribution of the information.

A5. Punctuation (the signature has proper punctuation, the signature has punctuation but in the wrong place, or *there is no punctuation)*: This attribute analyses any punctuation mark or distinctive stroke that can characterise the signature (e.g. ‘i’ or ‘j’ punctuation).

A6. Flourish symmetry (symmetric, asymmetric, or *unknown):* This attribute refers to the flourish strokes and their symmetry.

A7. Flourish weight (thin, wide, or *unknown)*: This attribute is related to the whole shape of the flourish which is commonly characterised by thin or wide strokes made by fast and very person-dependent movements.

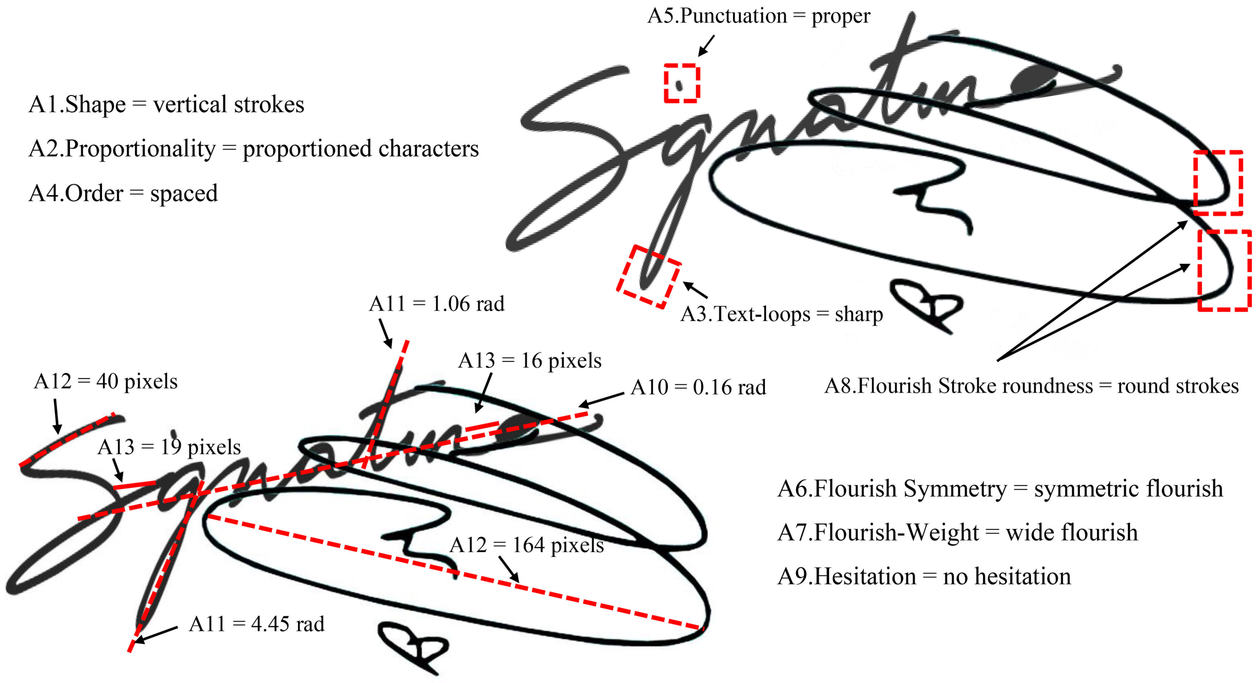
A5.Punctuation = proper
A1.Shape = vertical strokes
A2.Proportionality = proportioned characters
A4.Order = spaced
A3.Text-loops = sharp
A8.Flourish Stroke roundness = round strokes
A11 = 1.06 rad
A13 = 16 pixels
A10 = 0.16 rad
A12 = 40 pixels
A13 = 19 pixels
A12 = 164 pixels
A11 = 4.45 rad
A6.Flourish Symmetry = symmetric flourish
A7.Flourish-Weight = wide flourish
A9.Hesitation = no hesitation

Fig. 4 *Example of categorical attributes (left) and scalar measures (right) for a given signature*

A8. Flourish stroke roundness (round, sharp, or *unknown)*: This attribute is related to the style of the strokes of the flourish, which typically include changes of directions that can be classified into sharp (highly abrupt) or round (soft change of direction).

A9. Hesitation (the user did not hesitate while making the signature, the user did hesitate while making the signature, or *unknown)*: This attribute reveals the level of perceived hesitation in the signature. Hesitation produces enlargement of characters, tendency of curves to become angles, patching, and retouching, tremors, among others.

A10. Alignment to the baseline: It is defined as the angle (radians) between the main dominant axis of the signature and the horizontal baseline.

A11. Slant of the strokes: This attribute measures the slope (angle in radians with respect to the horizontal baseline) of up to three different characters or stroke segments. The annotator has to choose which are the most relevant strokes (if they exist; otherwise, the attribute is set to zero).

A12. Strokes-length: As in the slant measures, the annotator has to select up to three representative strokes (initial and ending points) to automatically calculate their lengths (in pixels).

A13. Character spacing: This attribute measures the separation (in pixels) between up to four most relevant characters in the signature (typically part of the given name and family name).

Note that A12 and A13 are measured in pixels. To improve the interoperability between different scanners, the values can be converted to the International System of Units by using the resolution parameter (e.g. 600 dpi of BiosecurID database). We recommend Fig. 4 shows an example of the 13 attributes associated to a given signature.

3.2 Attribute database and protocol

The database used in our experiments is the signature data in the BiosecurID multimodal database [23]. The database was acquired in five different universities using five different acquisition devices. To avoid any bias in the results (e.g. sensor interoperability), only the UAM subcorpus, which is the largest subcorpus within the database, will be considered in this work. The subcorpus employed comprises 132 signers of the UAM corpus acquired in four different sessions, with 16 genuine signatures (four per session) and 12 simulated forgeries (three per session) for every subject ($132 \times 28 = 3696$ signatures). Simulated signatures are made by writers different to the owner trying to imitate the natural style of a genuine signature. Signatures were performed on a marked area over paper templates (25 mm × 120 mm) with an inking pen which also captured the $\boldsymbol{x}$ and $\boldsymbol{y}$ trajectories and the pen pressure $\boldsymbol{p}$ during the signing process, with a sampling frequency of 100 Hz. The database includes both the dynamic sequence $[\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{p}]$ and the static image scanned (600 dpi) from the sheets.

The 3696 signatures were manually annotated according to Section 3.1. The annotation was made by 11 M.Sc. students (from Universidad de las Fuerzas Armadas – ESPE, Ecuador) without any previous experience on FDE analysis. No information about the authenticity (genuine or imitation) of the samples was provided to the annotator and all signatures were analysed separately. Therefore, the attribute database comprises more than 800,000 attributes [The full attribute database will be available here: http://www.atvs.ii.uam.es/databases.jsp.] (11 annotators × 132 signers × 28 samples × 20 annotations).

3.3 Discriminability and stability of the human attribute annotation

The first experiment is carried out to evaluate the discriminative power of manually annotated attributes. First of all, the categorical labels are transformed into numerical values from 1 to the number of possible labels for each attribute (e.g. six for A1 or four for A2). Let $\boldsymbol{A}_i^j$ be a matrix with the values of the attribute $i \in \{1, 2, \ldots, 20\}$ annotated by the annotator $j \in \{1, 2, \ldots, 11\}$ for the whole database (note that A11, A12, and A13 have more than one annotation). All the values $A_i^j(n, p)$ are first normalised as

$$\hat{A}_i^j(n, p) = \frac{1}{2}\left(\tanh\left(0.01\left(\frac{A_i^j(n, p) - \mu_i}{\sigma_i}\right)\right) + 1\right) \qquad (1)$$

where μ_i and σ_i are, respectively, mean and standard deviation of the attribute i from all the genuine signatures across all annotators j. We have used the tanh normalisation function in order to reduce the impact of outliers on the models generated from the labelled features [30]. Index $n \in \{1, \ldots, N = 132\}$ is the signer and $p \in \{1, \ldots, P = 16\}$ is the sample number. We define two discriminability indices D_{R} and D_{F} for random and forgery comparisons, respectively. In D_{R}, the model of signature n is evaluated against signatures samples of different signer. In D_{F}, the

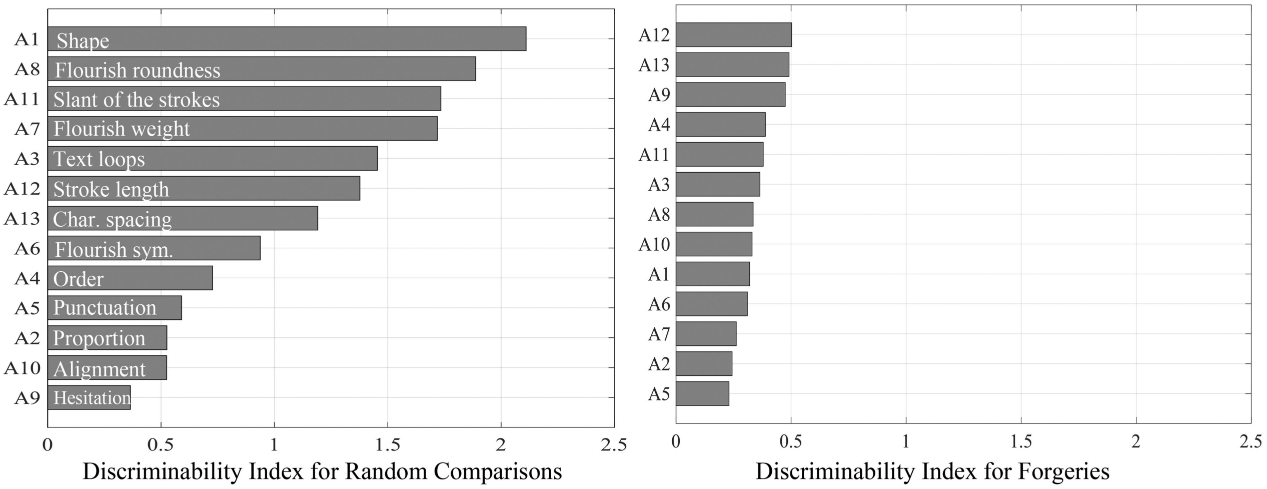

Fig. 5 *Discriminability index of the different attributes for random (left) and simulated forgery comparisons (right)*

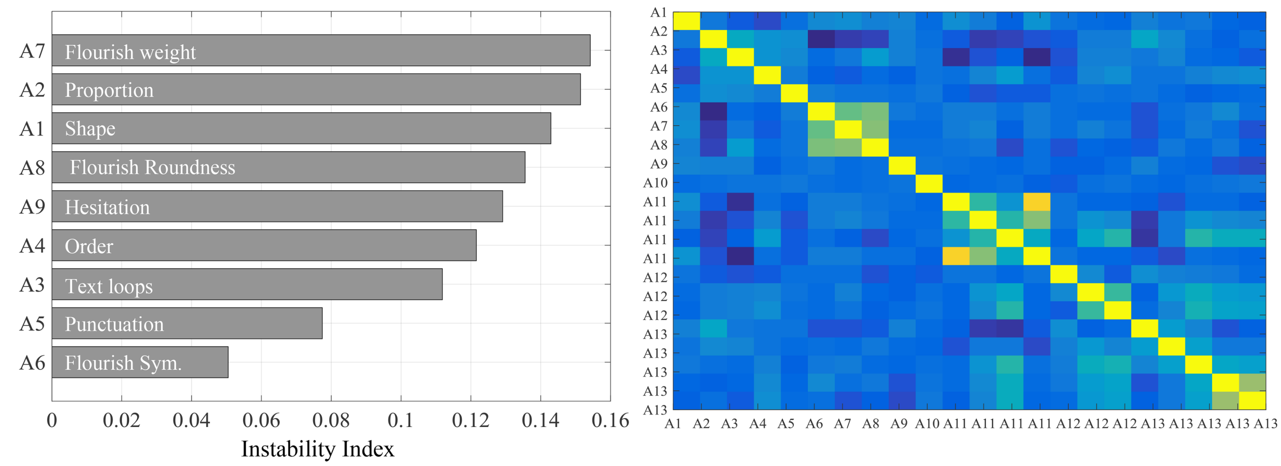

Fig. 6 *Instability index of the categorical attributes for genuine signatures (left) and correlation matrix of the attributes (right)*

model of signature n is evaluated with imitations made by other signers. D_R is computed for a specific attribute i as

$$D_R(i) = \frac{1}{(N-1)N}\sum_{n=1,\, n\neq m}^{N}\sum_{m=1}^{N}\frac{|\mu_i(n)-\mu_i(m)|}{\sigma_i(n)+\sigma_i(m)} \tag{2}$$

where $\mu_i(n)$ is the attribute i mean for signer n computed across the 16 available genuine signatures for that signer (and all annotators). Similarly $\sigma_i(n)$ is also the attribute i standard deviation for signer n. The discriminability index of simulated forgeries D_F for attribute i is computed as

$$D_F(i) = \frac{1}{N}\sum_{n=1}^{N}\frac{|\mu_i(n)-\tilde{\mu}_i(n)|}{\sigma_i(n)+\tilde{\sigma}_i(n)} \tag{3}$$

where $\tilde{\mu}_i(n)$ and $\tilde{\sigma}_i(n)$ are the mean and standard deviation of the simulated forgeries of the signer n computed across the 12 available forgeries for that signer (and all annotators). In the case of attributes with more than one annotation (A11, A12, and A13), the annotations are processed separately and then combined into one value by averaging. As it is expected, the discriminability of attributes is higher in random forgeries, see Fig. 5. However, the results suggest that depending on the scenario (random or simulated forgeries), some attributes can be more discriminant than others. As an example, the Hesitation (A9) is more discriminant for simulated forgeries than for random. This is because of the vacillations of the forger which are not present in genuine signatures (used for the random comparisons). On the other hand, the Shape (A1) is highly discriminative for random comparisons but not for forgeries.

The annotation of attributes depends of the perception of the annotator and it can vary between annotators. It is expected that some attributes will be more stable (similar among different annotators) than others. We calculated the index $\bar{S}(i)$ to measure the instability of an attribute i as

$$\bar{S}(i) = \frac{1}{NPT}\sum_{n=1}^{N}\sum_{p=1}^{P}\frac{1}{11}\sum_{j=1}^{11}\left|A_i^j(n,p) - \mathrm{mode}(\boldsymbol{A}_i^{:}(n,:))\right| \tag{4}$$

where N, P, T are the number of signers, samples, and number of labels of each attribute, respectively ($N = 132$, $P = 16$; T varies for each attribute, see Section 3.1). $A_i^j(n,p)$ is the value of the attribute i by the annotator j for the sample p of the signer n. $\boldsymbol{A}_i^{:}(n,:)$ is a matrix (dimension 176×1) with all the values of the attribute i from all the samples of the signer n and all the annotators ($176 = 11$ annotators $\times$ 16 genuine samples per signer).

The instability index for all the nine categorical attributes can be seen in Fig. 6 (left-hand side). We did not include the measured attributes in the analysis because the measures are strongly dependent on the keypoints selected by the annotators. Therefore, the instability indexes of measured attributes show values much greater than categorical attributes. The results show how some attributes such as Flourish weight (A7), Proportion (A2), Shape (A1), and Flourish roundness (A8) are less stable than others such as Flourish symmetry (A6), Punctuation (A5), or Text loops (A3). More instructive guidelines or training can be used to improve the stability of the attribute annotations. Fig. 6 (right-hand side) shows the correlation matrix of all attributes. In general, there is a low correlation between features, except for the three attributes related to the flourish characteristics (A6, A7, and A8) and the four measures of the slant of the strokes (A11).

Table 2 Performance for the different systems on the BiosecurID database (improvement with respect to the baseline systems added as subscript)

System	EER, %		FRR (FAR = 10%)	
	Random	Forgeries	Random	Forgeries
offline system (baseline)	4.72	20.27	2.13	34.31
online system (baseline)	1.85	6.85	1.21	6.12
attribute-based (average)	6.89	24.22	4.64	54.23
attribute-based (best annotator)	4.25	22.31	2.04	46.32
offline + attributes (average)	$2.63_{\downarrow 44\%}$	$16.80_{\downarrow 17\%}$	$0.84_{\downarrow 60\%}$	$30.21_{\downarrow 12\%}$
offline + attributes (best annotator)	$1.66_{\downarrow 65\%}$	$15.55_{\downarrow 23\%}$	$0.46_{\downarrow 78\%}$	$29.80_{\downarrow 13\%}$
online + attributes (average)	$0.72_{\downarrow 61\%}$	$5.98_{\downarrow 13\%}$	$0.10_{\downarrow 92\%}$	$4.65_{\downarrow 24\%}$
online + attributes (best annotator)	$0.20_{\downarrow 89\%}$	$5.55_{\downarrow 19\%}$	$0.01_{\downarrow 99\%}$	$4.24_{\downarrow 31\%}$

4 Experiments

The experiments are designed to answer the following questions: What is the performance of manual annotated signature attributes? What is the complementarity (in terms of performance) between human attribute-based authentication and traditional automatic online signature authentication? The experiments are divided into two categories:

Scenario 1 —– random comparisons: The model of the user is evaluated using genuine samples from other users (different to the owner) as impostor attacks (simulation of users who try to spoof the identity of the user with their own signature).

Scenario 2 – forgery comparisons: Also known as skilled forgeries, the model of the user is evaluated using imitations made by other users (with different level of skill, see the database description for details [23]).

The training set is composed of the four genuine signatures from the first session of each user. Genuine scores are obtained comparing the training model to the remaining 12 genuine samples of each user (sessions 2–4) for a total number of genuine scores equal to 1584 (132 × 12). Impostor scores for the random scenario are obtained comparing the training model to the first genuine samples from all users (different to the owner) for a total number of random impostor scores equal to 17,292 (132 × 129 × 1). The 1584 impostor scores for the simulated forgery scenario are obtained comparing the training samples to the 12 simulated forgeries available for each user (132 × 12).

The attribute-based matching proposed in this work can be used in both online and offline signature authentication applications. To compare the performance of attribute-based signature authentication and ASV systems, we have used two state-of-the-art systems (described in Section 2) based on online features (dynamic sequences derived from the signing process) and offline features (obtained from the static image of the signature).

The distances between categorical features and scalar measures are obtained separately. The distance between two scalar attribute vectors (attributes A10–A13) is calculated using the Manhattan distance normalised by the average absolute deviation of each attribute. Assume $\boldsymbol{f} = [f_1, f_2, \ldots, f_I]$ as the feature vector (with I features) of a given test sample and $\boldsymbol{g}^p = [g_1^p, g_2^p, \ldots, g_I^p]$ $p \in \{1, \ldots, P\}$ as an enrollment set with P samples. The distance between the feature vector $\boldsymbol{f}$ of the test sample and the enrollment set $\{\boldsymbol{g}^p\}_{p=1}^{P}$ of a given signer is calculated as

$$d = \sum_{i=1}^{I} \frac{|f_i - \bar{g}_i|}{\sigma_i} \qquad (5)$$

where $\bar{\boldsymbol{g}}$ is the mean of the enrollment set and $\boldsymbol{\sigma} = [\sigma_1, \sigma_2, \ldots, \sigma_I]$ is the standard deviation of the enrollment features. In our experiments P is equal to 4 and $I = 11$ (note that attributes A11–13 comprise ten measures). In the case of categorical attributes (attributes A1–A9), we consider a fixed distance equal to 1 when the label of the feature vector and the mode of the gallery vectors (most frequent value of the attribute for this signer) are not equal. Therefore, the distance between categorical attributes ranges from 1 to 9 (number of attributes between query sample and gallery set with different labels). Both distances (categorical and scalar) are normalised similar to (1). The final score is obtained as the sum of both distances.

4.1 Attribute-based matcher performance

The rest of the experiments try to evaluate the performance of the manually annotated signature attributes (detailed in Section 3) and the improvement obtained when they are combined with the two signature authentication systems (detailed in Section 2). Note that BiosecurID database includes both online information (captured using a digital tablet) and the static image of the signatures (scanned at 150 dpi). The static images are used as input of the offline system, while the online sequences are used as input of the online system and the tool for the attribute annotation (the image shown to the annotator is a synthetic version derived from the $[\boldsymbol{x}, \ \boldsymbol{y}, \ \boldsymbol{p}]$ sequences). Both static real signature and synthesised version can be used to annotate the attributes proposed in this work. We have chosen the synthetic version (generated from dynamic sequences) because digital devices are common in real applications (e.g. points of sales, banks, and postal).

The results are reported in Table 2 and Fig. 7 in terms of average performance (EER and FRR when FAR is equal to 10% across all annotators) and best performance (EER and FRR when FAR is equal to 10% for the best annotator).

The performance obtained by the proposed attribute-based matcher is similar to the performance obtained by the offline ASV baseline system. The better performance of the online matcher is caused by the more discriminant information available in the dynamic sequences in comparison with the features obtained from single static images (both the attribute-based and offline systems are based on static information).

The next step is to explore the complementarity between baseline systems and the proposed attribute-based matcher. Once again, the scores are normalised similar to (1) and combined using a weighted sum. The weights are heuristically selected based on the performance achieved in the previous experiment: 0.8 × online score + 0.2 × attribute-based score, and 0.8 × offline score + 0.2 × attribute-based score. The results (see Table 2) suggest that the proposed attribute-based matcher can be used to significantly improve the performance of baseline systems either in random and simulated forgery scenarios. In the random comparison scenario, it is possible to observe improvements from 44% (average annotators, offline + attribute-based matcher) to 90% (best annotator, online + attribute-based matcher). In the case of simulated forgeries, the improvements range from 16% (average annotators, online + attribute-based) to 23% (best annotator, offline + attribute-based).

4.2 Ranked performance

As it was mentioned, comparing human performance by aggregate human ratings is a standard protocol for the evaluation of human-assisted schemes [4, 12]. Similar to the experiment included in Section 2, we propose a combination of laymen responses based on sum rule at score level (scores obtained by different laymen are combined). We analyse the performance for the combination of an increasing number of annotators: 2, 5, and 10. The experiments with 2 and 5 annotators are repeated 50 times (with random selection of annotators) and the experiment with 10 annotators is repeated 11 times using the 11 different possible combinations. Table 3 shows the averaged results and the improvement obtained by the combination of laymen.

The results suggest the complementarity of annotations made by different laymen with improvements of the EER ranging from 27 to 75% for random scenarios and 2 to 42% for forgeries

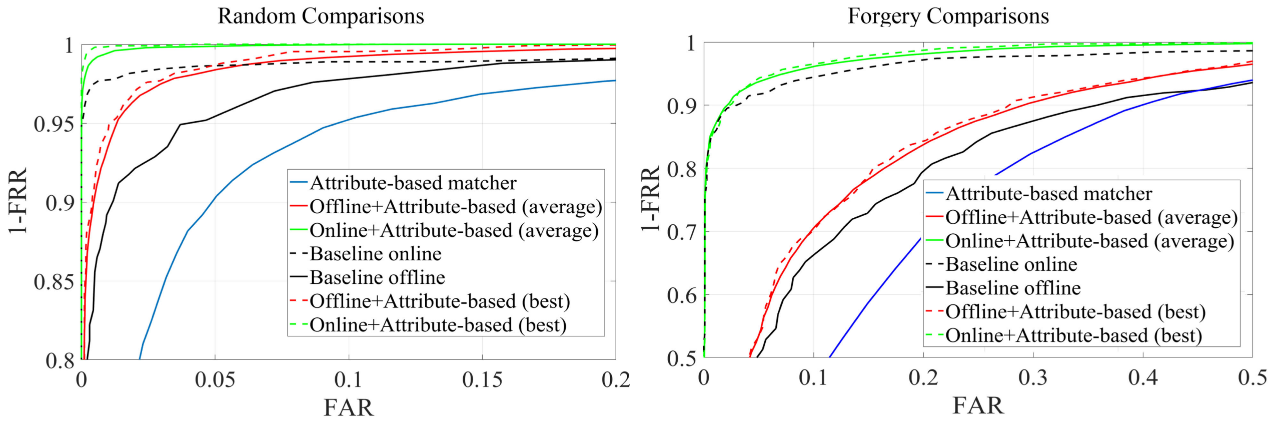

Fig. 7 *ROC curves for the different forgery scenarios and systems*

Table 3 Performance combining scores from different number of annotators (improvement with respect to the baseline systems added as subscript)

System	# Annotators	EER, %		FRR (FAR = 10%)	
		Random	Simulated	Random	Simulated
attribute-based	1	6.89	24.22	4.64	54.23
attribute-based	2	$3.96_{\downarrow 42\%}$	$21.14_{\downarrow 13\%}$	$2.11_{\downarrow 54\%}$	$42.96_{\downarrow 21\%}$
attribute-based	5	$2.38_{\downarrow 65\%}$	$18.66_{\downarrow 23\%}$	$1.53_{\downarrow 67\%}$	$33.43_{\downarrow 38\%}$
attribute-based	10	$1.68_{\downarrow 75\%}$	$18.21_{\downarrow 24\%}$	$1.02_{\downarrow 78\%}$	$31.78_{\downarrow 41\%}$
offline + attributes	1	2.63	16.80	0.84	30.21
offline + attributes	2	$1.92_{\downarrow 27\%}$	$15.75_{\downarrow 6\%}$	$0.34_{\downarrow 59\%}$	$22.48_{\downarrow 25\%}$
offline + attributes	5	$1.65_{\downarrow 37\%}$	$14.79_{\downarrow 12\%}$	$0.27_{\downarrow 68\%}$	$21.31_{\downarrow 29\%}$
offline + attributes	10	$1.33_{\downarrow 49\%}$	$13.98_{\downarrow 17\%}$	$0.18_{\downarrow 74\%}$	$20.09_{\downarrow 33\%}$
nline + attributes	1	0.72	5.98	0.10	4.65
online + attributes	2	$0.47_{\downarrow 34\%}$	$5.88_{\downarrow 2\%}$	$0.01_{\downarrow 99\%}$	$3.81_{\downarrow 18\%}$
online + attributes	5	$0.33_{\downarrow 54\%}$	$5.61_{\downarrow 6\%}$	$0.01_{\downarrow 99\%}$	$3.23_{\downarrow 30\%}$
online + attributes	10	$0.28_{\downarrow 61\%}$	$5.57_{\downarrow 6\%}$	$0.01_{\downarrow 99\%}$	$2.88_{\downarrow 38\%}$

scenarios. These improvements are even higher for the FRR with values ranging from 54 to 99% for random scenarios and 18 to 41% for forgeries scenarios. As in previous experiments, the improvement is larger in offline applications than in online applications. The higher error rates obtained in offline systems offer a larger margin for performance improvement.

5 Conclusions

This work explores human intervention on signature authentication systems at two different levels. The first scheme considers intervention at classification level, with an analysis of how humans perform at signature authentication tasks. The experiments based on the analysis of the response of 500 people help to establish a human baseline performance. The results suggest that laymen perform worst than ASV systems and highlight the difficulties associated to this task. The average error rate of laymen is around 30% but aggregated opinions show the potential of human capabilities when responses from different people are combined.

The second scheme evaluates the human intervention at feature level based on attributes inspired in the work of FDEs. The experiments include 11 different annotators, 3696 signatures, and more than 800,000 labelled attributes. The results suggest the potential of human capabilities to improve automatic authentication systems in both offline and online applications. The combination of attribute-based intervention and ASV systems at score level shows improvements ranging from 16 to 90% depending on the scenario.

The results reported in this work reveal new insights on how humans perform on signature authentication, and some ways in which ASV systems can be improved with human intervention. Our methods and experimental framework were developed for that purpose and not for direct practical application. For practical applications, we would recommend to obtain a reduced set of the most discriminative features either automatically or manually labelled in a short amount of time (e.g. <10 s for a point of sales or <1 min for an important banking operation). In addition, the human intervention evaluated in this work is focused on static information of the signature and future work should also explore how dynamic information could be integrated into these human annotations. The discriminative power of the dynamic information of the signature could be used to increase the differences between genuine and forged samples. Previous works suggest that more information showed to the laymen does not necessarily imply better performance [21]. How to integrate the dynamic information into the human evaluations is not trivial and further research is needed.

6 Acknowledgments

A.M. is supported by a Juan de la Cierva (JdC) contract (JCI-2012-12357) from the Spanish MINECO. This work was partially funded by CECABANK and project CogniMetrics (TEC2015-70627-R) from Spanish spanish Government agencies (MINECO)/European Comission grant (FEDER).

7 References

[1] Plamondon, R., Srihari, S.N.: 'On-line and off-line handwriting recognition: a comprehensive survey', *IEEE Trans. Pattern Anal. Mach. Intell.*, 2000, **22**, pp. 63–84

[2] Impedovo, D., Pirlo, G.: 'Automatic signature verification: the state of the art', *IEEE Trans. Syst. Man Cybern. C*, 2008, **38**, (5), pp. 609–635

[3] Fierrez, J., Ortega-Garcia, J.: 'On-line signature verification', in Jain, A.K., Ross, A., Flynn, P. (EDs.): '*Handbook of biometrics*', (Springer, New York, NY 10013, USA, 2008), pp. 189–209

[4] Kumar, N., Berg, A.C., Belhumeur, P.N., *et al.*: 'Describable visual attributes for face verification and image search', *IEEE Trans. Pattern Anal. Mach. Intell.*, 2011, **33**, (10), pp. 1962–1977

[5] Reid, D., Nixon, M., Stevenage, S.V.: 'Soft biometrics; human identification using comparative descriptions', *IEEE Trans. Pattern Anal. Mach. Intell.*, 2014, **36**, (6), pp. 1216–1228

[6] Klare, B.F., Klum, S., Klontz, J., *et al.*: 'Suspect identification based on descriptive facial attributes'. Proc. of Int. Joint Conf. on Biometrics, Clearwater, FL, USA, 2014, pp. 1–8

[7] Samangouei, P., Patel, V.M., Chellappa, R.: 'Continuous user authentication on mobile devices based on facial attributes', *IEEE Signal Process. Mag.*, 2016, **33**, (4), pp. 49–61

[8] Tome, P., Fierrez, J., Vera-Rodriguez, R., *et al.*: 'Soft biometrics and their application in person recognition at a distance', *IEEE Trans. Inf. Forensics Sec.*, 2014, **9**, (3), pp. 464–475

[9] Best-Rowden, L., Bisht, S., Klontz, J.C., *et al.*: 'Unconstrained face recognition: establishing baseline human performance via crowdsourcing'. Proc. of the Int. Joint Conf. on Biometrics, Tampa, USA, 2014, pp. 1–6

[10] Han, H., Otto, C., Liu, X., *et al.*: 'Demographic estimation from face images: human vs. machine performance', *IEEE Trans. Pattern Anal. Mach. Intell.*, 2015, **37**, (6), pp. 1148–1161

[11] Coetzer, J., Herbst, B.M., Du Preez, J.A.: 'Off-line signature verification: a comparison between human and machine performance'. Proc. Tenth Int. Workshop on Frontiers in Handwriting Recognition, La Baule, France, 2006, pp. 481–485

[12] Phillips, P.J., Hill, M.Q., Swindle, J.A., *et al.*: 'Human and algorithm performance on the PaSC face recognition challenge'. Proc. Int. Conf. on Biometrics: Theory, Applications and Systems, Arlington, USA, 2015, pp. 1–8

[13] Morocho, D., Morales, A., Fierrez, J., *et al.*: 'Towards human-assisted signature recognition: improving biometric systems through attribute-based recognition'. Proc. IEEE Int. Conf. on Identity, Security and Behavior Analysis, Japan, 2016, pp. 1–6

[14] Jain, A.K., Dass, S.C., Nandakumar, K., *et al.*: 'Soft biometric traits for personal recognition systems'. Proc. Int. Conf. Biometric Authentication, Hong Kong, 2004, pp. 731–738

[15] Dantcheva, A., Velardo, C., D'angelo, A., *et al.*: 'Bag of soft biometrics for person identification: new trends and challenges', *Mutimedia Tools Appl.*, 2010, **10**, pp. 1–36

[16] Oliveira, L., Justino, E., Freitas, C., *et al.*: 'The graphology applied to signature verification'. Proc. 12th Conf. of the Int. Graphonomics Society, Salerno, Italy, 2005, pp. 286–290

[17] Burkes, T.M., Seiger, D.P., Harrison, D.: 'Handwriting examination: meeting the challenges of science and the law', *Forensic Sci. Commun.*, 2009, **11**, (4)

[18] Malik, M.I., Liwicki, M., Dengel, A., *et al.*: 'Man vs. machine: a comparative analysis for forensic signature verification'. Proc. of the 16th Int. Graphonomics Society Conf., 2013, pp. 9–13

[19] Malik, M.I., Liwicki, M., Dengel, A.: 'Part-based automatic system in comparison to human experts for forensic signature verification'. Proc. Int. Conf. on Document Analysis and Recognition, Washington, DC, USA, 2013, pp. 872–876

[20] Coetzer, H., Sabourin, R.: 'A human-centric off-line signature verification system'. Proc. Int. Conf. on Document Analysis and Recognition, Curitiba, Brazil, 2007, pp. 153–157

[21] Morocho, D., Morales, A., Fierrez, J., *et al.*: 'Signature recognition: establishing human performance via crowdsourcing'. Proc. Fourth Int. Workshop on Biometrics and Forensics, Limassol, Cyprus, 2016, pp. 1–6

[22] Coetzer, J., Swanepoel, J., Sabourin, R.: 'Efficient cost-sensitive human-machine collaboration for offline signature verification', *IS&T/SPIE Electron. Imaging*, 2012, **8297**, pp. 1–8

[23] Fierrez, J., Galbally, J., Ortega-Garcia, J., *et al.*: 'BiosecurID: a multimodal biometric database', *Pattern Anal. Appl.*, 2010, **13**, (2), pp. 235–246

[24] Martinez-Diaz, M., Fierrez, J., Krish, R.P., *et al.*: 'Mobile signature verification: feature robustness and performance comparison', *IET Biometrics*, 2014, **3**, pp. 267–277

[25] Galbally, J., Diaz-Cabrera, M., Ferrer, M.A., *et al.*: 'On-line signature recognition through the combination of real dynamic data and synthetically generated static data', *Pattern Recognit.*, 2015, **48**, pp. 2921–2934

[26] Martinez-Diaz, M., Fierrez, J.: 'Signature databases and evaluation', in Li, S.Z., Jain, A.K. (EDs.): '*Encyclopedia of biometrics*' (Springer, New York, NY 10013, USA, 2015)), pp. 1367–1375

[27] Malik, M.I., Liwicki, M., Alewijnse, L., *et al.*: 'ICDAR2013 competitions on signature verification and writer identification for on- and offline skilled forgeries (SigWiComp2013)'. Proc. of Int. Conf. on Document Analysis and Recognition, Tunisia, 2013, pp. 1108–1114

[28] Houmani, N., Mayoue, A., Garcia-Salicetti, S., *et al.*: 'Biosecure signature evaluation campaign (BSEC2009): evaluating online signature algorithms depending on the quality of signatures', *Pattern Recognit.*, 2012, **45**, pp. 993–1003

[29] Ferrer, M., Vargas, J., Morales, A., *et al.*: 'Robustness of offline signature verification based on gray level features', *IEEE Trans. Inf., Forensics Sec.*, 2012, **7**, (3), pp. 966–977

[30] Jain, A.K., Nandakumar, K., Ross, A.: 'Score normalization in multimodal biometric systems', *Pattern Recognit.*, 2005, **38**, (12), pp. 2270–2285