M. Santopietro, R. Vera-Rodriguez, R. Guest, A. Morales and A. Acien, "Assessing the Quality of Swipe Interactions for Mobile Biometric Systems," *2020 IEEE International Joint Conference on Biometrics (IJCB)*, (2020): 1-8

DOI: https://doi.org/10.1109/IJCB48548.2020.9304858

Assessing the Quality of Swipe Interactions for Mobile Biometric Systems

Marco Santopietro[1], Ruben Vera-Rodriguez[2], Richard Guest[1], Aythami Morales[2], and Alejandro Acien[2]

[1]University of Kent, Canterbury, CT2 7NZ, UK
[2]Biometrics and Data Pattern Analytics Lab, Universidad Autonoma de Madrid, 28049, Madrid, Spain
[*]e-mails: (ms2101, r.m.guest)@kent.ac.uk; (ruben.vera, aythami.morales, alejandro.acien)@uam.es

Abstract

Quality estimation is a key study in biometrics, allowing optimisation and improvement of existing authentication systems by giving a prediction on the model performance based on the goodness of the sample or the user. In this paper, we propose a quality metric for swipe gestures on mobile devices. We evaluate a quality score for subjects on enrollment and for swipe samples, we estimate three quality groups and explore the correlation between our quality score and a state-of-art biometric authentication classifier performance. A further analysis based on the combined effects of subject quality and the amount of enrollment samples is conducted, investigating if increasing or decreasing enrollment size affects the authentication performance for different quality groups. Results are shown for three different public datasets, highlighting how higher quality users score a lower equal error rate compared to medium and low quality users, while high quality samples get a higher similarity score from the classifier.

Keywords-Mobile biometrics, behavioural biometrics, swipe gestures, user quality, sample quality.

1 Introduction

Over recent years, aligned to the rise of mobile technologies, biometrics authentication became increasingly popular. Smartphone devices contain many sensors that can acquire different biometric signals. In particular, some datasets have been collected in the last years including behavioural biometrics (e.g. keystroke, swipe, signature) [1, 2, 3, 4]. In particular, swipe biometrics is focused on authenticating the user, continuously or on queries, based on behavioural informations gathered from fingers interaction with a device touch screen. Compared to signature recognition, it introduces more challenges due to a lower amount of time samples captured and the lack of visual information feedback. It is well known that there are several aspects that influence authentication performance [5], including quality of sample and ability of the subject. A poor quality user template might cause a lack of robustness against attacks (i.e. increasing the False Positives). On the other hand, poor quality samples might reduce the recognition rates of genuine users (i.e. increasing the False Negatives).

In this paper we propose a framework to evaluate quality of swipes at both single sample and user template levels, based on the spread of the population and the consistency of the user. We consider three quality groups and perform an analysis based on the effect of the quality and the amount of enrollment data in performance. The quality study we propose aims to improve the performance of an authentication system in a multitude of ways, such as selecting better quality samples for enrollment, requesting a longer or shorter enrollment depending on the user quality group or weighting the thresholds according to subject quality.

Results are obtained in terms of average similarity score and equal error rate (EER) per quality range from three different public swipe datasets. To perform the evaluation we used an existing fusion model for swipe authentication based on a combination of a discriminative non-linear classifier and a statistical mixture model [6].

The remainder of the paper is organized as follows. Section 2 highlights the recent studies in biometric quality. Section 3 describes the classifier configuration, the preprocessing of data and the methodologies used to estimate quality scores and to define quality ranges. In Section 4 we describe the public datasets used and the experimental protocols. Section 5 describes the experimental results achieved. Finally Section 6 draws the final conclusions and addresses future challenges.

2 Related Works

Over the past years, biometric quality has been a topic of interest for many research groups and has seen its definition changed multiple times. Most studies were focused on image quality for fingerprint, iris and face recognition; quality was assesed in terms of *extractability* of features or *suitability* of the sample or even as an estimation of degrading factors known to affect the classification.

In 2014 Bharadwaj *et al.* [7] reviewed the methodologies for quality assessment and explored factors that could affect quality for different modalities. Regarding fingerprint quality assessment, in 2005 NIST released the "Fingerprint Image Quality (NFIQ) Compliance Test" [8]. More recently, in 2016, Yao *et al.* published a review of quality assessments for fingerprints [9].

Regarding face, different studies assessed quality for face images considering different kinds of approaches and issues. Corsetti *et al.* [10] investigated how accessibility influences the quality of captured image and therefore the authentication process, revealing how users with accessibility issues struggle providing good samples compared to control population. Chen *et al.* [11] proposed a flexible ranking method to evaluate the quality of face images depending on the dataset and the authentication system in use, allowing to select the best performing images during the authentication process (when more than one is provided, for example in a video recording). Hernandez-Ortega *et al.* [12] developed a quality assessment approach for face recognition based on deep learning (FaceQnet). Very recently, NIST has released the "Face Image Quality Asssessment" on the Ongoing Face Recognition Vendor Test (FRVT) [13].

Regardless the common interest to develop a solid and consistent quality score for each modality, few studies have been conducted on behavioural biometrics, except for signature recognition. A big issue is the absence of ground truth when it comes to behavioural biometrics (such as swipe or keystroke dynamics). Manual labeling is also not possible in these cases due to lack of understanding, unlike in the case of pictures where visual samples are provided.

A number of studies have explored the impact of quality in signatures trying to find metrics or predictors to estimate sample quality and correlate to the classification performance. Müller *et al.* [14] described the *a priori* and *a posteriori* approaches to evaluate quality on handwritten signatures, identifying specific quality features descriptive of signature stability; Galbally *et al.* [15] applied the *Sigma-Longnormal* model as a quality estimator for handwritten signature, while Sae Bae *et al.* [16] proposed a quality metric for online signatures that measures the separation between intra-user and inter-users distributions.

In our study we have considered the approach of Sae Bae *et al.* [16] adapting it to the case of swipe biometrics. We consider both sample and user quality and extract behavioural features and evaluate consistency of samples.

3 Methods

There are a limited number of studies regarding quality of behavioural biometrics, mostly because of the lack of metrics and ground truth. Currently, there is no ISO standard that defines what makes a sample good or bad in terms of quality and the definition of biometric quality itself is not absolute. For our study we took as reference the quality definition from NIST [17]: *A sample should be of good quality if its suitable for automated matching. [] A quality measure could be tuned to predict the performance of one matcher or the more difficult case of one that generalizes to other matchers or classes of matchers.*

We applied this definition to elaborate a quality metric for both sample and user quality and for both cases we estimated three quality ranges (low, medium and high) and studied the correlation between the quality score and the performance of a classifier (in terms of similarity score and EER) on various datasets. The quality metric is based on the algorithm by Sae-Bae *et al.* [16], but we have modified it to take into account other factors like population, task and time lapse. The quality score we used is defined as:

$$Q = \frac{1}{N}\sum_{f=1}^{N}\frac{||\mu_{l_f} - \mu_{g_f}||}{\sqrt{\frac{\sigma_{l_f}^2+\sigma_{g_f}^2}{2}}} \tag{1}$$

with N = number of extracted features and $f = f^{th}$ feature from the sample; μ and σ are the mean and standard deviation operators; l and g subscripts stand for *local* and *global*. We refer with *local* to the single sample or subject, respectively for sample and user quality, and with *global* to the estimator of the population, which is a global value obtained by all samples in the considered dataset.

The classifier we used, based on [6], is a multimodal system that combines a Support Vector Machines (SVM) with non linear kernel and a Gaussian Mixture Model (GMM) at the score level. This system comprises the statistical modelling from the GMM and the good discriminate ability of SVM. From every swipe, two feature vectors are extracted and used as inputs for the two models and a single similarity score is evaluated by the system for each swipe. For authentication, ten consecutive swipes are used and their similarity scores are averaged.

3.1 User Template Quality

To estimate user quality we preprocessed the data extracting from each swipe two feature vectors of 28 and 5 dimensions each, according to [6]. The first 28 extracted features are

representative of the state-of-the-art swipe biometric recognition:

- *mean, standard deviation, first quartile, second quartile and third quartile* of velocity, acceleration, pressure and finger area.
- *x* and *y* coordinates of extreme points in the stroke.
- Distance between start and end of the stroke.
- Stroke duration.
- Distance traveled.

The last 5 features were proposed for on-line signature recognition. These features are selected from a larger set of 100 features as explained in [6]:

- θ (finger down to finger up)
- σ_{a_x} (std of the acceleration in x)
- $(x_{max} - x_{min})/x_{maxrange}$
- $(\overline{x} - x_{min})/\overline{x}$
- $(y_{max} - y_{min})/y_{maxrange}$

From the second set of features, the last four features are considered for vertical strokes, in case of horizontal ones *x* and *y* are swapped in the formulations.

A quality score is calculated from the extracted features using equation (1), but instead of considering a small subset of impostors, we calculated the global mean μ_g and standard deviation σ_g from all users and samples in the database with respect to each task and only on first session, if there was more than one. μ_l and σ_l are calculated for each feature over the enrollment samples of the *i-th* subject. The enrollment samples of each user are also included in the computation of the global mean and variance, to avoid them being automatically considered outliers from the population and, as a consequence, having a higher quality score than expected.

For each subject a single quality score is evaluated for each direction (meaning that the same user could provide better data for specific tasks) and the corresponding EER is obtained with the classifier for inter and intra sessions.

3.2 Sample Quality

Compared to subject quality, the estimation of a score for a single sample required further preprocessing. Also, new features had to be extracted with specific conditions to ensure that they were descriptive of a sample goodness and not of the unicity of the user. Those conditions are:

- Locally defined features.
- Screen independent.
- Device invariant.
- Position independent.

Considering these issues, we extracted six derivative features (*point to point velocity and acceleration on x-axis, y-axis and distance travelled*) and two physiological (*screen pressure and finger area*). From Equation 1, μ_l and σ_l are computed over the sample points for all the features. Global mean and standard deviations of the previously calculated features means are evaluated over all the samples in the database:

$$\mu_{g^f} = \frac{1}{S}\sum_{i=1}^{S}\mu_{l_f^i}$$
$$\sigma_{g^f} = \sqrt{\frac{\sum_{i=1}^{S}(\mu_{l_f^i} - \mu_{g^f})^2}{S-1}} \quad (2)$$

where *S* is the total number of samples in the database. A quality score is then assigned to each sample following Equation 1 taking into account these new μ_{g^f} and σ_{g^f}. Instead of the EER, for each sample used during the testing phase the similarity score is stored and only the genuine samples are considered. The reason is that the same genuine sample could be an impostor if compared to the profile of another subject during the evaluation, but it would still maintain the same quality score. Thus, in this case its better to just consider the correlation between quality and similarity score for genuine samples.

3.3 K-means Clustering and Quantile Normalization

Its important to notice that the quality score used in this study has no upper bound and as stated before there is no ground truth for quality. To define ranges and thresholds we propose two methodologies: K-means and quantile normalisation.

One approach is to use K-means algorithm [18] with K as the number of quality ranges set to 3 (*low, medium and high*) to cluster the data in an unsupervised mode. We ran the algorithm using just the quality score for both user and sample quality for each scenario/task. Once obtained the thresholds, the mean and the variance of the EER (or similarity scores for sample quality) are computed for every range and compared. This approach worked well when considering a large number of points, but with few sparse ones (like in datasets with few subjects in case of user quality or tasks with few samples in case of sample quality) the estimation of the cluster means was mostly random and not completely reliable.

The second proposed solution is a max score normalization with fixed thresholds at 0.33 and 0.66 and outliers removal. A standard scaler was not a good option (negative values, no constraints) and a normal min-max or max scaler would be biased by the presence of outliers. With our approach, we normalize the data based not on the maximum quality score in the dataset, but on the 95^{th} quantile (considering the 5% of the samples as outliers). This methodology works quite well with small datasets containing outliers,

the downside is that it also considers equal width for all the quality ranges (which might not necessarily be the real case). This method is still a valid option when K-means algorithm cannot be applied and highlights quite consistently the correlation between quality and classifier performance.

4 Datasets and Experiments

Before explaining the experimental protocol, we describe in this section the public swipe datasets used, to give an insight on the demographic and how data were collected.

4.1 Serwadda Dataset

This is a public database collected at Louisiana Tech University by Serwadda *et al.* [19] from 190 different subjects of different ages with one smartphone (*Google Nexus S.*). The data were collected through two applications, asking the subject multiple questions and allowing free interaction with the touch screen. Touch data were recorded considering only one finger touch and ignoring other interactions with multi-touch like zooms. Features collected were *x* and *y coordinates, timestamp, pressure, finger area*.

Data collection was split in two sessions at least one day apart, the overall number of strokes per user was around 80. Amongst the three datasets, the Serwadda database is the largest in terms of number of samples and subjects.

4.2 Frank Dataset

This database was collected by Frank *et al.* [20], and it is composed of swipe data collected in two sessions (1 week distance) from 41 subjects. Data have been acquired from several different android devices and two applications have been developed for the purpose. Users were free to interact with the screen. The applications captured *x and y coordinates, pressure, finger area, timestamp, device orientation and finger orientation*. It is also important to state that not all swipe directions count the same number of samples, with the *down* direction containing the most samples.

4.3 Antal Dataset

The last public database used in this study is composed of horizontal and vertical swipe data collected from 71 users on eight different mobile devices [21]. An application has been developed for the purpose and the strokes were task related (vertical to read text, horizontal to choose pictures). The data was collected in one single session with each user interacting with multiple devices. The same features of the previous database were collected. It is important to note that the majority of swipes in this datasets are horizontal, while the least amount is found in the *Up* direction.

4.4 Experiments

4.4.1 Sample Quality Protocol

For each database and each swipe we extracted the three feature vectors, two for the classifier and the last one for the sample quality estimation. We calculated the global μ and σ for the quality features over all the samples in the first session for each stroke direction. Then we assigned to all samples in the test set (either the second session for *inter session* scenario or same for *intra session* scenario) the quality score using equation 1 and a similarity score evaluated with the classifier.

The K-means algorithm is then used to find quality thresholds on genuine samples considering the corresponding similarity score. Mean and variance of the similarity scores for the samples in the three quality ranges are calculated, expecting a correlation between quality and classifier performance (similarity score should be higher for high quality samples).

4.4.2 User Quality Protocol

After extracting the two quality vectors used for the classifier, we computed the quality score for each user during the training phase of the classifier, using 10 enrollment samples from every different subject. Then we evaluated subjects' EER for inter and intra session and for each dataset and stroke direction. We divided again users in three quality groups, using K-means for the Serwadda database and quantile normalization for the Antal and Frank datasets, due to the low amount of subjects compared to the other database. After estimating the quality groups, we repeated the classifier evaluation three more times, considering different numbers of enrollment samples for each user (5 , 15 and 20 samples) and comparing the mean EERs for the different quality groups for each enrollment size.

5 Results

For *Sample Quality*, Table 1 shows the *mean* and *standard deviation* (in brackets) of the similarity score for testing genuine samples, considering quality ranges, stroke directions, sessions and datasets. Higher scores represent a better classifier performance. An example is shown in Fig. 1.

Intra-session classification performs better in every circumstance, due to the increased consistency of the subjects. Overall, we can see an increase of the similarity score for higher quality samples, with some exceptions (especially in the Frank database) caused probably by a smaller number of samples or increased inconsistency towards certain stroke directions.

For *User Quality*, Tables 2, 3 and 4 show EER values (*mean* and *standard deviation*) for the three datasets, con-

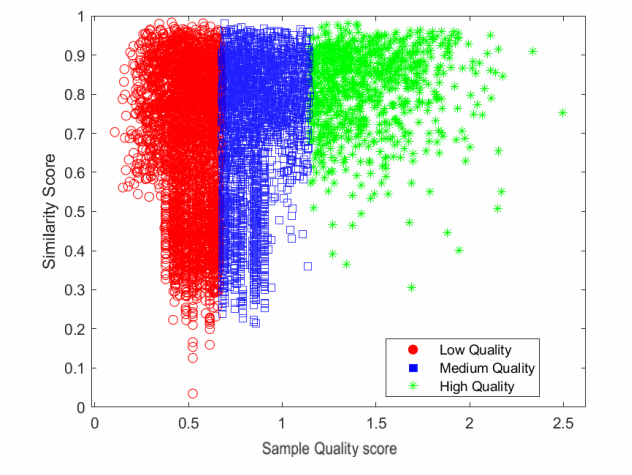

Figure 1: Sample quality score vs Similarity score, Serwadda database, intra session, right swipe. K-means algorithm has been used to cluster the three ranges.

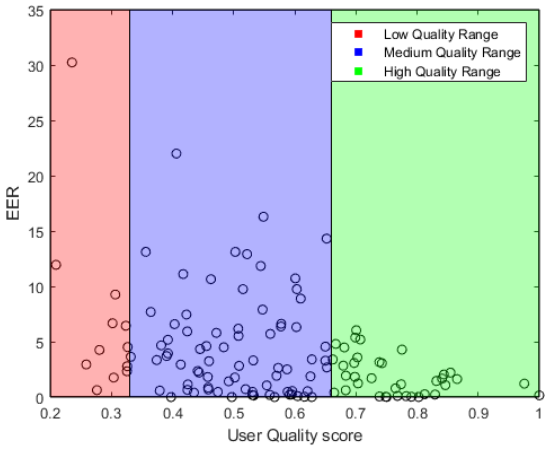

Figure 2: User quality vs EER in Serwadda database (Direction: down, Intra session). Here we used quantile normalisation to separate quality groups.

sidering directions, sessions and number of enrollment samples. Examples of quality clustering using quantile normalisation and corresponding distribution of EER per range are showed in Figure 2 and Figure 3. The Tables highlight not only the decreasing EER over the quality ranges, but also how the different number of training samples affect the performance for different quality subjects. In general, high quality users are quite consistent with their own samples and increasing or decreasing the number of enrollment samples does not impact on the classifier performance. On the other hand, for low quality users, increasing the number of training samples helps providing correct classification during testing, as showed in Figure 4.

Empty values occur when there are no samples in the related quality ranges; this happens only for the Frank and Antal datasets, having less subjects and majority of strokes in specific directions.

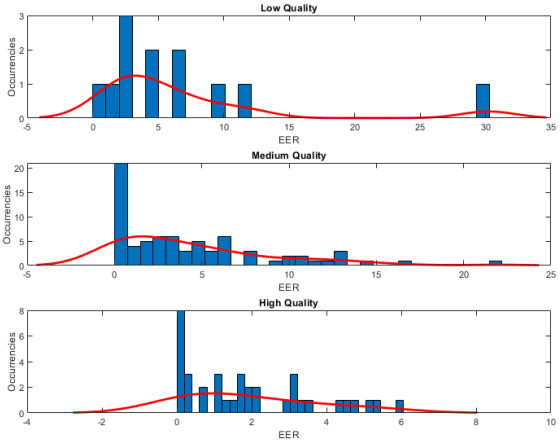

Figure 3: Histograms of the subjects' EERs distribution per quality group in Serwadda database (Direction: down, Intra session).

6 Conclusions

We have conducted a study to analyse authentication performance on swipe based on user and sample quality on our proposed metric, in order to reduce EER and False Negative Rate (FNR) once the quality is assessed. In particular, we defined three ranges (low, medium and high quality) for both sample and subject, depending on which a larger number of training data could be asked (in case of low quality user) or another stroke attempt could be acquired (in case of low quality sample) to improve the performance of the classifier.

In most of the cases, the results proved our assumptions: higher quality user on average score a lower EER and are less affected by varying the number of enrollment samples, while increasing it for low quality users leads to a considerable improvement in classification.

In case of the sample quality, we show how widely spread are the similarity scores for low quality samples compared to high quality, leading to misclassification and false rejections. Applying a quality threshold at the acquisition could result in better performance for many systems.

In addition to these protective measurements, other uses of this study involve:

- Monitoring low quality users to identify and prevent attacks.
- In case of devices with multiple authentication systems, select the fittest one for the given subjects de-

Serwadda Database		Quality Range		
		Low	Medium	High
INTRA SESSION	Down	0.73 (0.12)	0.74 (0.13)	**0.78 (0.14)**
	Up	0.71(0.14)	0.75 (0.13)	**0.79 (0.13)**
	Left	0.63 (0.23)	0.78 (0.10)	**0.78 (0.15)**
	Right	0.67 (0.19)	0.70 (0.18)	**0.82 (0.1)**
INTER SESSIONS	Down	0.63 (0.14)	0.65 (0.14)	**0.67 (0.15)**
	Up	0.61(0.14)	0.63 (0.14)	**0.66 (0.15)**
	Left	0.64 (0.15)	0.67 (0.13)	**0.69 (0.15)**
	Right	0.60 (0.14)	0.63 (0.15)	**0.68 (0.15)**

Frank Database		Quality Range		
		Low	Medium	High
INTRA SESSION	Down	0.69 (0.12)	0.60 (0.19)	**0.68 (0.2)**
	Up	-	-	-
	Left	0.73 (0.14)	**0.77 (0.12)**	0.74 (0.15)
	Right	0.71 (0.12)	0.76 (0.12)	**0.77 (0.15)**
INTER SESSIONS	Down	**0.62 (0.15)**	0.49 (0.16)	0.52 (0.18)
	Up	-	-	-
	Left	0.63 (0.15)	**0.69 (0.14)**	0.65 (0.14)
	Right	0.66 (0.14)	**0.70 (0.14)**	0.65 (0.17)

Antal Database		Quality Range		
		Low	Medium	High
INTRA SESSION	Down	0.62 (0.12)	0.74 (0.10)	**0.83 (0.01)**
	Up	0.70 (0.15)	0.68 (0.13)	**0.74 (0.14)**
	Left	0.65 (0.18)	0.71 (0.17)	**0.74 (0.17)**
	Right	**0.74 (0.12)**	0.71 (0.16)	0.52 (0.19)
INTER SESSIONS	Down	-	-	-
	Up	-	-	-
	Left	-	-	-
	Right	-	-	-

Table 1: Results for Sample Quality Analysis. Mean Similarity score (standard deviation in brackets) for genuine samples in different quality ranges, evaluated for each direction and each dataset on both intra and inter sessions. values are left blank when missing.

		SERWADDA DATABASE							
		INTRA SESSION				INTER SESSIONS			
		Enrollment samples				Enrollment samples			
	Quality Ranges	*5 samples*	*10 samples*	*15 samples*	*20 samples*	*5 samples*	*10 samples*	*15 samples*	*20 samples*
Down	*Low*	14.61 (9.6)	12.97 (9.5)	8.02 (5.4)	**7.75 (6.0)**	25.31 (14.1)	23.56 (14.7)	22.22 (13.7)	**19.79 (13.9)**
	Medium	10.48 (7.9)	7.45 (6.5)	**5.25 (5.3)**	5.28 (5.0)	22.41 (16.6)	17.88 (14.0)	20.10 (14.4)	**16.89 (13.7)**
	High	3.55 (4.4)	2.11 (2.5)	1.74 (2.2)	**1.50 (1.6)**	12.57 (12.5)	11.27 (10.4)	**10.16 (12.5)**	15.76 (22.6)
Up	*Low*	13.86 (8.8)	10.45 (7.9)	7.77 (5.8)	**7.19 (6.4)**	30.12 (14.9)	25.97 (15.1)	24.71 (18.2)	**22.86 (14.6)**
	Medium	7.93 (6.3)	4.38 (4.3)	3.95 (4.1)	**3.18 (3.5)**	21.85 (13.8)	21.85 (15.9)	**17.56 (16.1)**	18.51 (14.7)
	High	1.83 (2.1)	0.51 (0.7)	0.18 (0.3)	**0.11 (0.2)**	15.30 (11.6)	12.14 (12.8)	**10.24 (9.1)**	11.97 (10.4)
Right	*Low*	11.26 (7.2)	9.25 (6.9)	6.09 (5.2)	**5.14 (3.8)**	25.88 (16.4)	24.70 (15.7)	21.59 (13.4)	**20.43 (12.8)**
	Medium	5.63 (5.7)	3.85 (3.8)	2.90 (3.3)	**2.74 (2.6)**	16.42 (14.4)	16.24 (13.8)	15.47 (14.4)	**13.78 (12.4)**
	High	2.19 (2.1)	1.35 (1.4)	**0.85 (1.2)**	1.08 (1.2)	21.27 (21.4)	**18.45 (17.3)**	19.35 (17.4)	18.49 (23.2)
Left	*Low*	10.36 (9.4)	8.25 (7.0)	**5.84 (6.2)**	6.21 (5.5)	19.29 (14.5)	19.67 (13.6)	17.86 (15.6)	**17.49 (12.3)**
	Medium	6.51 (6.9)	5.20 (5.2)	2.78 (2.8)	**2.12 (2.5)**	18.34 (17.3)	17.00 (16.9)	17.41 (18.4)	**13.87 (15.1)**
	High	3.24 (4.6)	2.47 (3.5)	2.16 (2.8)	**1.43(2.8)**	17.82 (16.8)	18.30 (19.5)	**17.35 (20.7)**	18.49 (23.2)

Table 2: Results for User Quality Analysis for Serwadda database. Mean EER in % (standard deviation in brackets) for subjects in quality ranges. In addition to directions and intra/inter sessions, different number of training samples are considered in the evaluation.

		FRANK DATABASE							
		INTRA SESSION				INTER SESSIONS			
		enrollment samples				enrollment samples			
	Quality Ranges	*5 samples*	*10 samples*	*15 samples*	*20 samples*	*5 samples*	*10 samples*	*15 samples*	*20 samples*
Down	*Low*	40.06 (0.)	15.13 (0.)	**13.96 (0.)**	16.90 (0.)	-	-	-	-
	Medium	17.69 (11.3)	9.19 (2.4)	**4.67 (4.5)**	6.6 (4.9)	31.61 (27.9)	**19.24 (5.4)**	27.22 (26.2)	23.08 (25.2)
	High	13.3 (8.6)	8.71 (6.9)	8.83 (6.4)	**7.95 (7.3)**	5.66 (5.1)	2.95 (4.8)	4.82 (8.4)	**2.83 (3.9)**
Right	*Low*	17.46 (0.)	12.42 (0.)	19.58 (0.)	**8.28 (0.)**	-	-	-	-
	Medium	15.27 (17.7)	10.06 (7.8)	**6.46 (3.9)**	7.61 (4.6)	11.31 (9.1)	9.25 (10.9)	9.48 (11.1)	**4.08 (4.1)**
	High	2.09 (1.7)	**1.00 (1.1)**	1.94 (2.4)	1.14 (2.5)	11.24 (12.1)	12.00 (12.1)	**11.29 (11.8)**	12.43 (12.5)
Left	*Low*	-	-	-	-	-	-	-	-
	Medium	9.31 (10.1)	5.55 (4.9)	9.46 (10.5)	**5.16 (7.1)**	19.91 (10.1)	18.88 (10.3)	13.33 (12.2)	**11.28 (5.1)**
	High	7.36 (6.1)	4.29 (3.4)	4.71 (4.3)	**3.29 (3.1)**	17.33 (20.5)	11.80 (9.3)	**8.58 (11.1)**	9.46 (9.5)

Table 3: Results for User Quality Analysis for Frank database. Mean EER in % (standard deviation in brackets) for subjects in quality ranges. In addition to directions (excluding Up direction) and intra/inter sessions, different number of training samples are considered in the evaluation.

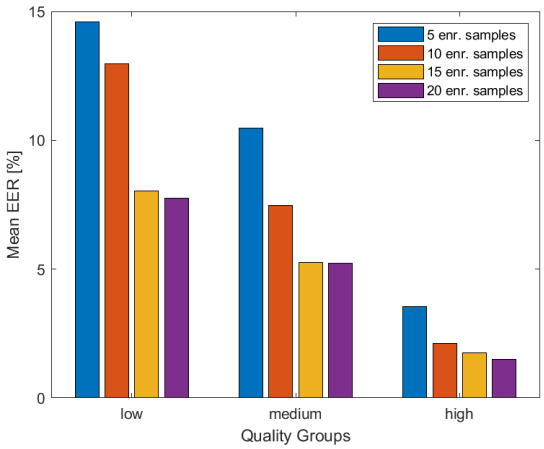

Figure 4: Mean subjects' EERs with varying enrollment sizes for each quality group. The quality score has been evaluated considering 10 enrollment samples.

		ANTAL DATABASE			
		INTRA SESSION			
		enrollment samples			
	Quality Ranges	*5 samples*	*10 samples*	*15 samples*	*20 samples*
Down	*Low*	-	-	-	-
	Medium	13.77 (7.2)	12.75 (8.8)	**5.61 (4.9)**	7.01 (3.7)
	High	5.66 (5.0)	7.83 (6.1)	3.97 (4.7)	**0.26 (0.4)**
Up	*Low*	-	-	-	-
	Medium	-	-	-	-
	High	9.68 (15.8)	5.03 (6.2)	5.51 (6.2)	**0.85 (0.5)**
Right	*Low*	26.17 (5.0)	23.31 (11.3)	17.83 (5.4)	**17.75 (8.3)**
	Medium	18.15 (10.3)	14.59 (8.9)	9.95 (6.6)	**9.49 (5.8)**
	High	11.82 (8.3)	10.81 (8.5)	6.84 (5.8)	**6.25 (5.0)**
Left	*Low*	16.91 (4.6)	29.70 (6.7)	15.88 (6.5)	**8.64 (8.5)**
	Medium	18.90 (10.9)	14.59 (8.3)	11.70 (8.6)	**10.05 (7.6)**
	High	11.00 (6.1)	8.82 (6.5)	6.48 (5.3)	**4.72 (4.6)**

Table 4: Results for User Quality Analysis for Antal database. Mean EER in % (standard deviation in brackets) for subjects in quality ranges. In addition to directions, different number of training samples are considered in the evaluation.

pending on their quality scores.

- Explore the variation in quality and performance over long periods of time and estimate time windows for new enrollments.

In addition to these points, a follow up research could analyse different thresholds or a combination of the two scores in order to select the best samples for enrollment. Also, the recent HuMiDB database [22] that comprises user mobile interaction data from 600 users will be taken into account in follow up works.

7 Acknowledgements

Thanks to support by projects: PRIMA (H2020-MSCA-ITN-2019-860315), TRESPASS-ETN (H2020-MSCA-ITN-2019-860813), BIBECA (RTI2018-101248-B-I00 MINECO/FEDER), and BioGuard (Ayudas Fundacion BBVA a Equipos de Investigacion Cientifica 2017).

References

[1] A. Acien, J. V. Monaco, A. Morales, R. Vera-Rodriguez, and J. Fierrez, "Typenet: Scaling up keystroke biometrics," in *Proc. IEEE/IAPR Intl. Joint Conf. on Biometrics (IJCB)*, September 2020.

[2] A. Acien, A. Morales, J. Fierrez, R. Vera-Rodriguez, and I. Bartolome, "BeCAPTCHA: Detecting Human Behavior in Smartphone Interaction using Multiple Inbuilt Sensors." AAAI Workshop on Artificial for Cyber Security, 2020.

[3] R. Tolosana, R. Vera-Rodriguez, J. Fierrez, and J. Ortega-Garcia, "BioTouchPass2: Touchscreen Password Biometrics Using Time-Aligned Recurrent Neural Networks," *IEEE Trans. on Information Forensics and Security*, 2020.

[4] R. Tolosana, R. Vera-Rodriguez, J. Fierrez, and J. Ortega-Garcia, "Deepsign: Deep on-line signature verification," *arXiv preprint arXiv:2002.10119*, 2020.

[5] M. Boakes, R. Guest, F. Deravi, and B. Corsetti, "Exploring Mobile Biometric Performance Through Identification of Core Factors and Relationships," *IEEE Transactions on Biometrics, Behavior, and Identity Science*, vol. 1, no. 4, p. 278291, 2019.

[6] J. Fierrez, A. Pozo, M. Martinez-Diaz, J. Galbally, and A. Morales, "Benchmarking Touchscreen Biometrics for Mobile Authentication," *IEEE Transactions on Information Forensics and Security*, vol. 13, no. 11, p. 27202733, 2018.

[7] S. Bharadwaj, M. Vatsa, and R. Singh, "Biometric Quality: A Review of Fingerprint, Iris, and Face," *EURASIP Journal on Image and Video Processing*, vol. 2014, Feb 2014.

[8] E. Tabassi, "NIST Fingerprint Image Quality (NFIQ) Compliance Test," 2005.

[9] Z. Yao, J.-M. L. Bars, C. Charrier, and C. Rosenberger, "Literature Review of Fingerprint Quality Assessment and its Evaluation," *IET Biometrics*, vol. 5, p. 243251, Jan 2016.

[10] B. Corsetti, R. Sanchez-Reillo, R. M. Guest, and M. Santopietro, "Face Image Analysis in Mobile Biometric Accessibility Evaluations," in *Proc. 2019 International Carnahan Conference on Security Technology (ICCST)*, pp. 1–5, 2019.

[11] J. Chen, Y. Deng, G. Bai, and G. Su, "Face Image Quality Assessment Based on Learning to Rank," *IEEE Signal Processing Letters*, vol. 22, no. 1, p. 9094, 2015.

[12] J. Hernandez-Ortega, J. Galbally, J. Fierrez, R. Haraksim, and L. Beslay, "FaceQnet: Quality Assessment for Face Recognition based on Deep Learning," in *Proc. 2019 International Conference on Biometrics (ICB)*, pp. 1–9, 2019.

[13] P. Grother, A. Hom, M. Ngan, and K. Hanaoka, "Ongoing Face RecognitionVendor Test (FRVT). Part 5: Face Image Quality Asssessment," tech. rep., NIST, 2020.

[14] S. Mller and O. Henniger, "Evaluating the Biometric Sample Quality of Handwritten Signatures," *Advances in Biometrics, Lecture Notes in Computer Science*, p. 407414, 2007.

[15] J. Galbally, J. Fierrez, M. Martinez-Diaz, and R. Plamondon, "Quality Analysis of Dynamic Signature based on the Sigma-Lognormal Model," in *Proc. ICDAR*, pp. 633–637, 2011.

[16] N. Sae-Bae and N. Memon, "Quality of Online Signature Templates," in *Proc. IEEE Intl. Conf. on Identity, Security and Behavior Analysis (ISBA 2015)*, pp. 1–8, 2015.

[17] E. Tabassi and P. Grother, "Biometric Quality. The Last 1%-Biometric Quality Assessment for Error Suppression," tech. rep., 2009.

[18] D. Pollard, "Strong Consistency of K-Means Clustering," *The Annals of Statistics*, vol. 9, no. 1, p. 135140, 1981.

[19] A. Serwadda, V. V. Phoha, and Z. Wang, "Which Verifiers Work?: A Benchmark Evaluation of Touch-based Authentication Algorithms," in *Proc. IEEE Intl. Conf. on Biometrics: Theory, Applications and Systems (BTAS)*, pp. 1–8, 2013.

[20] 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, p. 136148, 2013.

[21] M. Antal, Z. Bokor, and L. Z. Szab, "Information Revealed from Scrolling Interactions on Mobile Devices," *Pattern Recognition Letters*, vol. 56, p. 713, 2015.

[22] A. Acien, A. Morales, J. Fierrez, R. Vera-Rodriguez, and O. Delgado-Mohatar, "Becaptcha: Bot detection in smartphone interaction using touchscreen biometrics and mobile sensors," *arXiv preprint arXiv:2005.13655*, 2020.