SIGNATURE RECOGNITION:
ESTABLISHING HUMAN BASELINE PERFORMANCE VIA CROWDSOURCING

Derlin Morocho[1,2], Aythami Morales[2], Julian Fierrez[2], Ruben Tolosana[2]

[1]Departamento de Electrica y Electronica, Universidad de las Fuerzas Armadas-ESPE, Sangolquí, Ecuador
[2]ATVS- Biometric Recognition Group Universidad Autonoma de Madrid, Spain
dmorocho@espe.edu.ec, {aythami.morales,julian.fierrez,ruben.tolosana}@uam.es

ABSTRACT

This work explores crowdsourcing for the establishment of human baseline performance on signature recognition. We present five experiments according to three different scenarios in which laymen, people without Forensic Document Examiner experience, have to decide about the authenticity of a given signature. The scenarios include single comparisons between one genuine sample and one unlabeled sample based on image, video or time sequences and comparisons with multiple training and test sets. The human performance obtained varies from 7% to 80% depending of the scenario and the results suggest the large potential of these collaborative platforms and encourage to further research on this area.

Index Terms— ***Biometrics, Mechanical Turk, crowdsourcing, signature recognition, worker***

1. INTRODUCTION

The handwritten signature is one of the most accepted personal authentication methods and it has been used over the past 2000 years. As a behavioral biometric trait, one of the key characteristics of the signature is its high intra-person variability. The high variability between samples of the signature of the same person together with the ability of people to perform skilled forgeries make signature recognition a great challenge. Historically, signature recognition is made by Forensic Document Examiners (FDE) who have developed well-established protocols and methods to analyze the authenticity of a query signature. This is a time consuming and manual task which depends on the FDE training and experience. Therefore, the applications are limited to authentications without requirements of real time response (forensics and offline scenarios). Automatic signature verification systems (ASV) emerged as a feasible way to automate the traditional signature verification method made by FDEs [1][2] and extend the potential applications.

The variety of applications based on automatic signature recognition systems is large (e.g. banking, point-of-sales, parcel delivery, notary public). In most of these applications, humans supervise the signing process but his responsibilities are mostly limited to guarantee the correct record of the data (without any impact in the analysis of the authenticity). These supervisors do not have the specific FDE experience and they will be referred to as layman in the rest of this work. The deployment of automated systems is reducing the trustworthiness on human abilities. However, perception and analytic capability of humans should not be undervalued and there is large room for improvements exploiting both the efficiency of computers and the human abilities. Attribute annotation made by humans has emerged as a way to improve automatic recognition systems in face [3][4][5], gait [6][7] or security assessment [8]. One question raises: *How good is a layman in recognizing the authenticity of a query signature?* The use of our signature in our day-to-day life makes us good forgery detectors of imitation (made by others) of our own signature. We are capable of differentiating our intra-person variability from the variability of a forger (our brain models are trained with hundreds of samples made during years of practice). This ability can be extended to signatures from other people but it is expected a drop of performance caused by the lack of information about the variability of the owner. In addition, we have to consider motivation as an important factor to be considered. Without specific training and considering that signature recognition is not the principal job assignment of the laymen, their performance is an open question. Fig. 1 tries to illustrate the difficulties related with this task.

Crowdsourcing appears as a tool for experimentation on large population groups. This work explores crowdsourcing for the establishment of human baseline performance on signature recognition. We present different experiments involving 150 people. The contributions of this work are twofold: i) new insights on the use of crowdsourcing in signature recognition biometric studies and ii) the establishment of human baseline performance in signature recognition tasks according to different scenarios.

978-1-4673-9448-2/16/$31.00 ©2016 IEEE

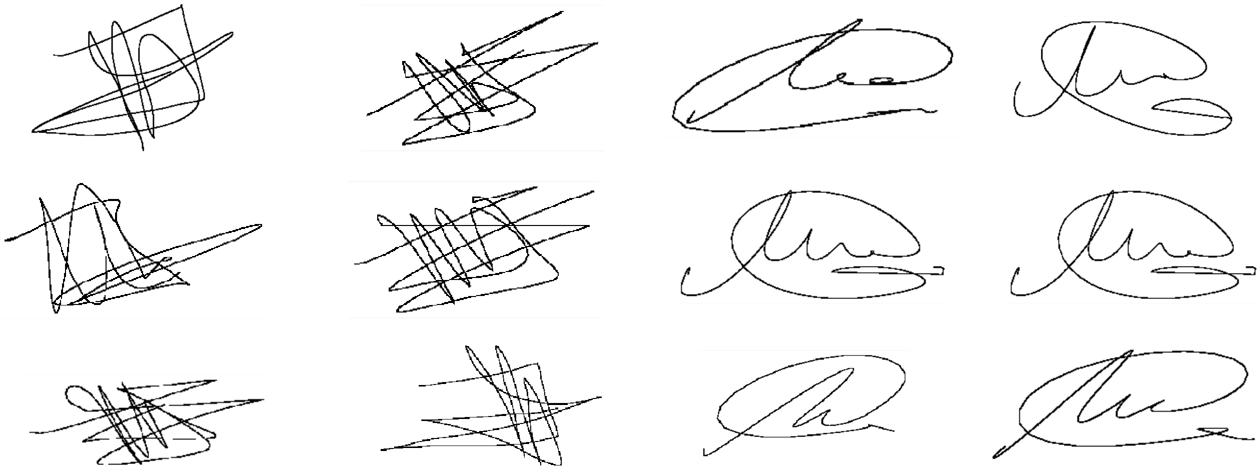


Figure 1. Genuine (from two different signers) and forgeries (made by other people after practicing for 2 minutes). Which signatures are genuine? See solution at[1]

The rest of the work is organized as follows: Section 1.1 analyze previous studies while Section 2 introduces the crowdsourcing methodology designed for this work. Section 3 presents the results and finally, Section 4 summarizes the conclusions and future works.

To the best of our knowledge, there are few previous works analyzing the signature recognition performance of laymen. In [9][10] the ability of 22 individuals (people staff and students from the university) was evaluated using 51 signatures (15 samples per signature mixing genuine and forged samples). There is not much information about the human performance but according to their results, it is strongly user dependent and varies from $1\% < False\ Rejection\ Rate > 25\%$ and $5\% < False\ Acceptance\ Rate > 65\%$. These results suggest the difficulties of individuals to recognize skilled forgeries. The human performance obtained was used as a baseline that was compared with automatic signature verification systems. Malik et al. [11] compared the performance of FDE and automatic signature recognition systems for disguised signatures. The results obtained in their study suggest that FDEs can achieve similar performance to automatic systems with average accuracies around 71% and the best accuracy of 91% (the best FDE).

2. ESTABLISHING HUMAN BASELINE PERFORMANCE VIA CROWDSOURCING

Crowdsourcing has become popular in science in the last decade. Massive human–assisted tasks take advantage of the human abilities and the benefits of a worldwide data sampling using internet [12][13][14]. The use of large-scale human annotations in automatic biometric recognition systems has been analyzed and encouraging results have been obtained in modalities like face recognition [4].

How good are humans (non FDE) analyzing the authenticity of a query signature? The main idea of applying crowdsourcing in the present work is the recruitment of amateur volunteers (people without FDE previous experience) to help in the establishment of a layman performance baseline for signature authentication. Amazon Mechanical Turk (MTurk) is a popular web-platform for the acquisition of data from large scale Human Intelligence Tasks (HITs) [13]. The participants (workers) work effectively together through pre-defined tasks [7][8][9].

We have designed five different tasks to be accomplished by different or the same set of workers. The proposed tasks are focused on analyzing the baseline performance of humans to recognize the authenticity of signatures according to different scenarios. The system is divided into Front-end which presents the tasks to the workers and captures their responses and Back-end which comprises the processes and algorithms running in the MTurk servers (see Fig. 2). It is important to note that the population can vary from one experiment to any other. It is recommendable to design simple tasks in order to be done in a short time. The motivation of the workers have an important impact on the results and complex or lengthy tasks directly influence this motivation [13][14].

Before the release of the tasks on the platform, there are some design parameters to be defined:

- **Qualification:** represents worker's skill, ability or reputation from previous HIT. In our experiments we have requested for regular and master workers. Masters are group of workers who have demonstrated superior performance in completing HITs across the Mechanical Turk marketplace.
- **Geographic location:** we did not request for specific geographic locations for our HITs. The workers

[1] Solution to Fig. 1: From left to right. Top: forgery, genuine, genuine, forgery; Center: forgery, genuine, forgery, forgery; Down: genuine, forgery, genuine, genuine.

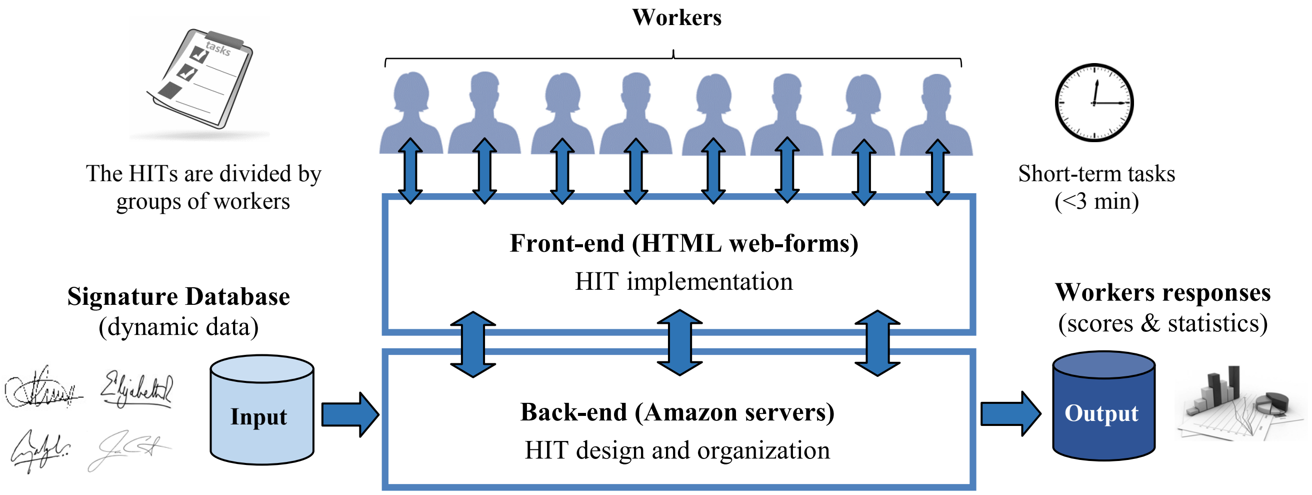

Workers
The HITs are divided by groups of workers
Short-term tasks (<3 min)
Front-end (HTML web-forms)
HIT implementation
Signature Database
(dynamic data)
Input
Back-end (Amazon servers)
HIT design and organization
Workers responses
(scores & statistics)
Output

Figure 2. Scheme of crowdsourcing studies applied to the establishment of human signature recognition performance baseline.

involved in our tasks are 47% from the United States of America, 34% from India and 19% from others. The ethnicity defines some characteristics of the signature but experiments included in this work do not consider the ethnic of the layman.

- **HIT runtime per worker:** represents the available time to perform the HIT. Parameters such as the complexity and the amount of data have direct impact on the time needed by the workers. Several experiments were performed (with a small set of workers) to evaluate the time needed to recognize the authenticity of a signature. Although some signatures require more time than others, the average time per worker to analyze one signature was around 10 seconds. Therefore, the HIT runtime of our experiments results on approximately 15 seconds per signature.

The different graphic interfaces implemented for each task has been programed in HTML language. These interfaces feature the following parts (see Fig. 3):

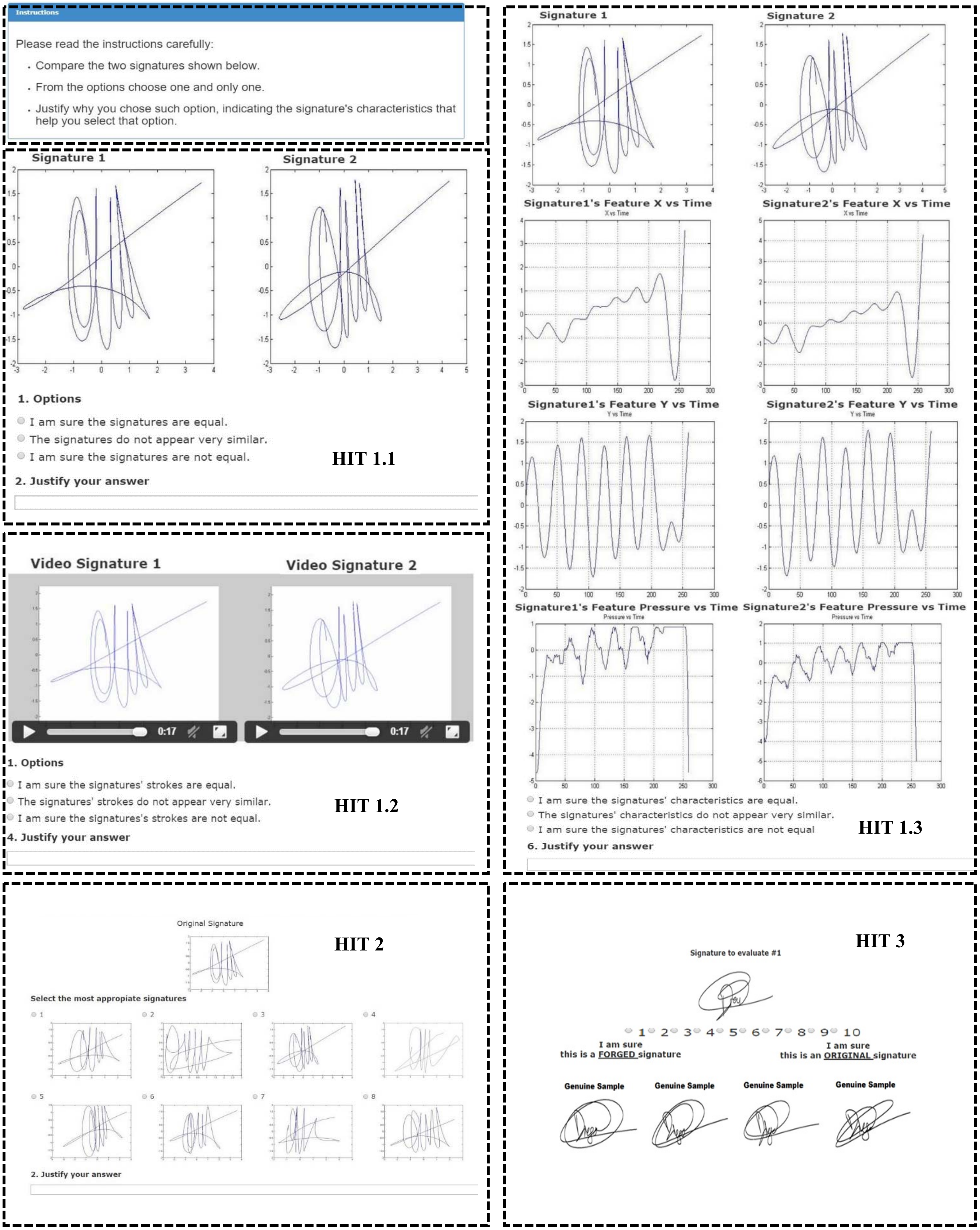
- **Instructions:** in this area the process and guidelines to be followed by each worker are detailed.
- **Signatures:** in this area a different set of labeled (genuine signatures) and unlabeled (genuine and forgery samples that workers have to label) signatures are showed. The information depicted varies according to the task: images (static versions of the signature), videos (showing the way the signature is done) and time sequences related with three dynamic signature features (x-axis, y-axis and pressure sequences). The signatures shown to the workers are synthesized on the screen according the dynamic information available on the databases employed (BIOSECURE-DS2 and BiosecurID).
- **Options:** this area includes the response of the workers to the predefined HIT. In our HITs, these responses are related with the authenticity analysis (genuine or forgery) made by the worker.
- **Justification:** in this area the worker must indicate the reasons why he/she chose the different options. This information helps to better understand the obtained results.

2.1. Human Intelligence Task design

In order to analyze the signature recognition human performance, we designed five different HITs divided into three experiments:

Experiment 1. One Training-One Test (HIT1.1-HIT1.3): the workers have to define the authenticity of a given sample using as reference only one genuine signature. This experiment is divided into three different tasks depending on the information showed: only the image of the signature (T1.1), image and video of the dynamic signing process (T1.2) and finally the image and time sequences of the signature that is the typical case in online dynamic recognition systems (T1.3). The aim of this experiment is to analyze the performance of the workers according to the different information available. These tasks include 12 different signers (with 2 genuine and 1 forgery samples per signer) from Biosecure-DS2 database [16]. The possible responses are indicated in Fig. 3. The execution time for this task is 2 minutes.

Experiment 2. One Training-Multiple Test (HIT2): in this case eight unlabeled images (five genuine and three forgeries) and only one genuine signature are showed. The aim of this experiment is to analyze the performance when the workers have contextual information (they can compare the variability of the different samples). Information about the number of genuine and forgery images is not shared with the workers. This task includes 6 different signers (with 5 genuine and 3 forgeries per signer) from Biosecure-DS2



Instructions

Please read the instructions carefully:

- Compare the two signatures shown below.
- From the options choose one and only one.
- Justify why you chose such option, indicating the signature's characteristics that help you select that option.

1. Options

○ I am sure the signatures are equal.
○ The signatures do not appear very similar.
○ I am sure the signatures are not equal.

2. Justify your answer

1. Options

○ I am sure the signatures' strokes are equal.
○ The signatures' strokes do not appear very similar.
○ I am sure the signatures's strokes are not equal.

4. Justify your answer

HIT 1.3
Signature 1
Signature 2
Signature1's Feature X vs Time
X vs Time
Signature2's Feature X vs Time
X vs Time
Signature1's Feature Y vs Time
Y vs Time
Signature2's Feature Y vs Time
Y vs Time
Signature1's Feature Pressure vs Time
Pressure vs Time
Signature2's Feature Pressure vs Time
Pressure vs Time

○ I am sure the signatures' characteristics are equal.
○ The signatures' characteristics do not appear very similar.
○ I am sure the signatures' characteristics are not equal

6. Justify your answer

Original Signature

Select the most appropiate signatures

○ 1 ○ 2 ○ 3 ○ 4
○ 5 ○ 6 ○ 7 ○ 8

2. Justify your answer

Signature to evaluate #1

○ 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ 6 ○ 7 ○ 8 ○ 9 ○ 10

I am sure this is a FORGED signature

I am sure this is an ORIGINAL signature

Genuine Sample Genuine Sample Genuine Sample Genuine Sample

Figure 3. Graphic interfaces of the different Mechanical Turk HITs programed in HTML.

database [15]. The execution time for this task is 1 minute. The execution time was reduced because of the statistics obtained from experiment 1, in which most of the workers invested less than 1 minute in the evaluation. Therefore, in this experiment we were interested to evaluate the performance of quick responses inspired in real operational scenarios in which the layman must provide a response in a short-time.

Experiment 3. Multiple Training-One Test (HIT3): this experiment imitates the traditional evaluation protocol of automatic signature recognition systems in which one query sample is compared with its corresponding training set. In this case, four labeled genuine signatures and one unlabeled signature (genuine or forgery) are showed to the worker. In order to improve the information available on previous experiments, the response of the worker is provided as a confidence value between 0 and 10 in which 0 means "I'm sure this is a forgery signature" and 10 means "I'm sure this is a genuine signature". This task includes 20 different signers (with 5 and 3 genuine and forgeries signatures respectively per signer) from BiosecurID database [16]. The 20 signers are divided into 5 sets of 4 signers each one ($4 \times 8 = 32$ signatures) which are equally distributed among workers (each worker process 32 signatures). The maximum execution time for this task is 4 minutes.

The images, videos and time sequences used in all interfaces have been created using MatLab and two public databases: BIOSECURE-DS2 (experiment 1 and 2) and BiosecurID (experiment 3). The signers were selected among those signers with lower performance (in terms of EER) according a state-of-the-art online signature verification system based on DTW algorithm and seven time functions derived from $[\mathbf{x}, \mathbf{y}, \mathbf{p}]$ sequences [17]. The number of workers varies for the different HITs: 60 workers (HIT1.1-HIT1.3), 30 workers (HIT2) and 60 workers (HIT3). The total number of signatures used in the experiments is $12 + 6 \times 8 + 12 \times 20 = 300$.

3. RESULTS

Different experiments imply different workers. As a human behavioral analysis, the response is strongly user dependent and therefore, performance obtained in different experiments must be compared with care. Table 1 shows the average results (among different workers) obtained for HIT1.1-1.3. The results are provided in terms of False Rejection Rate (FRR), False Acceptance Rate (FAR) and percentage of samples with No Defined response (ND).

Results suggest that more information does not necessarily imply better performance for laymen. Analyzing the comments provided by the workers, the image and its characteristics (shape, size, and strokes) are the main focus of attention. Therefore, the time sequences displayed in HIT1.3 can have negative impact on the performance as the workers are not trained to use such information. This suggests the importance of guiding the workers in their analysis in order to exploit the discriminative characteristics of dynamic sequences. The workers focus their attention on the image characteristics because this is what they do in day-to-day life. However, the drop of the ND responses is important if we compare HIT1.1 and HIT1.3. The additional information of HIT1.3 increases the confidence of the worker in his response (although it is wrong in some cases).

Table 1. Experiment1: Human performance results obtained from MTurk responses considering 60 workers to HIT1.1-1.3

Task	FRR (%)	FAR (%)	ND (%)
HIT 1.1	26.7	30.0	33.3
HIT 1.2	40.0	30.0	25.0
HIT 1.3	43.3	33.3	21.7

Table 2. Experiments 2 and 3: Human performance results obtained from MTurk considering responses to HIT2 (30 workers) and HIT3 (60 workers)

Task	FRR (%)	FAR (%)
HIT 2	80.0	7.8
HIT 3	31.1	40.3

Table 2 shows the averaged results (among different workers) for the tasks HIT2 and HIT3. In these cases we eliminated the ND response in the questionnaires. In HIT2 the workers labeled the samples either as genuine or forgery. In HIT3 the workers estimate a score based on their confidence in the response (values greater than 5 will be labeled as genuine and lower or equal as forgeries). The results obtained in HIT2 suggest that the ability of workers to recognize forgeries improves when they have contextual information. However, this improvement in the FAR imply a worsening in the FRR results. The workers report 80% of FRR which means that most of genuine samples are considered as forgeries. The unbalanced performance between genuine recognition (FRR) and forgery recognition (FAR) suggests the importance of contextual information (information about the natural variability of the signer). The inclusion of forgery and genuine samples together increases the ability of the workers to detect the forgeries but not the genuines. The results obtained for HIT3 show the effects of including a training set (four signatures in our case) to guide the decision of the worker. The more information about intra-personal variability (variability between samples of the same signer) is available the lower false rejection is obtained (but also higher false acceptance). It means more signatures are labeled as genuine probably due to the differences between the samples of the training set. Depending on the application better FAR can be obtained (with longer training sets) perhaps at a cost of higher FRR.

Comparing human performance by aggregate human ratings is a standard protocol [3][18]. The responses of workers can be fused to determine the complementarity potential of the human abilities. Fig. 4 shows the performance obtained for HIT3 by averaging the responses of different number of workers. The results show how FRR can be drastically reduced (from 32% to 10%) when the

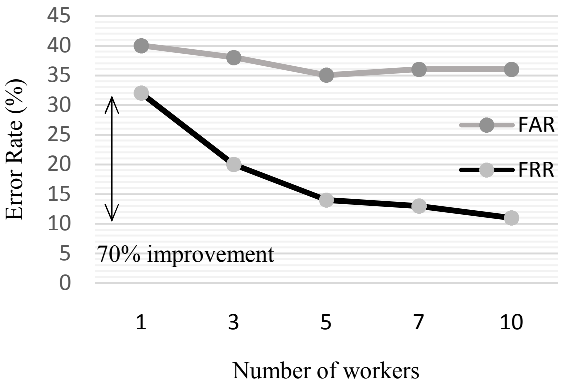


Figure 4. Evolution of FAR and FRR according to the number of workers fused in HIT3.

ranks of different workers are fused. However, this improvement cannot be observed for the FAR. Note that the signatures images are chosen among the most challenging samples in the BiosecurID database. The room for improvements in the way the information is fused is large and these results encourage to further research in this line.

4. CONCLUSIONS

This work explores crowdsourcing to establish a human performance baseline for signature recognition. We have presented a general framework and three different preliminary experiments. The potential of platforms such as Mturk in biometric recognition research is large and new insights can be developed on the basis of massive human collaborative tasks. The results obtained in this work suggest that more information (signature features) does not mean necessarily better performance (without any specific training). The results of HIT2 and HIT3 show the impact of the contextual information in the performance. The worker decision is closely related to the information provided and better FAR or FRR can be obtained depending on the samples showed. Finally, fusion of human ratings has shown a great potential in terms of FRR improvement (70% improvement when responses from 10 workers are fused) and further research is needed to export these results to the FAR.

5. ACKNOWLEDGMENT

A.M. is supported by a JdC contract by the Spanish MECD (JCI-2012-12357). This work has been partially supported by projects: Bio-Shield (TEC2012-34881) from Spanish MINECO, BEAT (FP7-SEC-284989) from EU and CECABANK e-BioFirma contract.

6. REFERENCES

[1] D. Impedovo and G. Pirlo. Automatic signature verification: The state of the art. IEEE Trans. on Systems, Man, and Cybernetics (Part C), 38(5):609-635, 2008.

[2] M. Martinez-Diaz, J. Fierrez and S. Hangai. Signature Matching", Stan Z. Li and Anil K. Jain (Eds.), Encyclopedia of Biometrics, Springer, pp. 1382-1387, 2015 (ISBN 978-1-4899-7487-7, re-edited from 2009).

[3] N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar. Describable visual attributes for face verification and image search. IEEE Trans. Pattern Analysis and Machine Intelligence, 33(10):1962–1977, 2011.

[4] L. Best-Rowden, et al.. Unconstrained Face Recognition: Identifying a Person of Interest from a Media Collection. IEEE Trans. on Information Forensics and Security, 9(12):2144-2157, 2014.

[5] L. Best-Rowden, S. Bisht, J. Klontz and A. K. Jain. Unconstrained Face Recognition: Establishing Baseline Human Performance via Crowdsourcing. Proc. Int. Conference on Biometrics, Clearwater, Florida, USA, Sept. 29-Oct. 2, 2014

[6] D. Reid, M. Nixon and S. V. Stevenage. Soft Biometrics; Human Identification using Comparative Descriptions. IEEE Trans. on Pattern Analysis and Machine Intelligence, 36(6): 1216-1228, 2014.

[7] D. Martinho-Corbishley, M. S. Nixon, Mark, J. N. Carter. Soft Biometric Recognition from Comparative Crowdsourced Annotations. Proc. Int. Conf. on Imaging for Crime Prevention and Detection, London, UK, pp. 1-6, 2015.

[8] S. Panjwani, A. Prakash. Crowdsourcing Attacks on Biometric Systems. Proc. Tenth Symposium On Usable Privacy and Security, Menlo Park, California, pp. 257-269, 2014.

[9] J. Coetzer, B.M. Herbst, J.A. Du Preez. Off-line signature verification: A comparison between human and machine performance. Proc. 10th Int. Workshop on Frontiers in Handwriting Recognition, La Baule, France, pp. 481-485, 2006.

[10] M. I. Malik, M. Liwicki, A. Dengel, and B. Found. Man vs. Machine: A Comparative Analysis for Forensic Signature Verification. Proc. of the 16th International Graphonomics Society Conference, pp. 9–13, 2013.

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

[12] J. Howe. The Rise of Crowdsourcing, Wired, 14(6), 2006.

[13] A. Kittur, E. H. Chi, B Suh. Crowdsourcing user studies with Mechanical Turk. Proc. of the SIGCHI conference on human factors in computing systems, pp. 453-456), 2008.

[14] M. Buhrmester, T. Kwang, S. D. Gosling. Amazon's Mechanical Turk a new source of inexpensive, yet high-quality, data? Perspectives on psychological science, 6(1), 3-5, 2011.

[15] J. Ortega-Garcia, et al. The Multi-Scenario Multi-Environment BioSecure Multimodal Database (BMDB). IEEE Trans. on Pattern Analysis and Machine Intelligence, 32(6):1097-1111, 2010.

[16] J. Fierrez, et al. BiosecurID: a multimodal biometric database. Pattern Analysis and Applications. 13(2):235-246, 2010.

[17] M. Martinez-Diaz, J. Fierrez, R.P. Krish and J. Galbally. Mobile signature verification: feature robustness and performance comparison. IET Biometrics, 3:267–277, 2014.

[18] P. J. Phillips, M. Q. Hill, J. A. Swindle, A. J. O'Toole. Human and Algorithm Performance on the PaSC Face Recognition Challenge. Proc. Int. Conference on Biometrics: Theory, Applications and Systems, Arlington, USA, pp. 1-8, 2015.