2009 10th International Conference on Document Analysis and Recognition

The ICDAR 2009 Signature Verification Competition

V.L. Blankers [1,2], C.E. van den Heuvel [1],
K.Y. Franke [3] and L.G. Vuurpijl [4]

[1] Netherlands Forensic Institute, Den Haag, The Netherlands
[2] University of Amsterdam, The Netherlands
[3] Norwegian Information Security Laboratory, Gjøvik University College, Norway
[4] Donders Institute for Brain, Cognition and Behavior, Radboud University Nijmegen, The Netherlands

Abstract

Recent results of forgery detection by implementing biometric signature verification methods are promising. At present, forensic signature verification in daily casework is performed through visual examination by trained forensic handwriting experts, without reliance on computer-assisted methods. With this competition on on- and off-line skilled forgery detection, our objective is to make a first step towards bridging the gap between automated biometric performances and expert-based visual comparisons. We intent to combine realistic forensic casework with automated methods by testing systems on a forensic-like new dataset. The results achieved by the participating systems are promising: 2.85% Equal Error Rate (EER) on the online data and 9.15% on the offline data. From these results we indicate that automated methods might be able to support forensic handwriting experts (FHEs) to formulate the strength of evidence that needs to be reported in court in the future.

1. Introduction

In the last few years, many computational methods have been implemented to build applications that develop new procedures for criminal law and justice [10]. The applications often use a similarity score between a stored model and a presented biometric and use a corresponding threshold to decide whether authentication will be provided to a person. For example, a biometric signature verification system could be used for fraud detection. In forensic handwriting examination, experience-based conclusions are drawn by a forensic handwriting expert (FHE) based on the comparison between a questioned signature and several reference signatures. The comparison is done visually, without the aid of scores computed by an automated system.

In forensic casework, the use of a precise threshold is not desirable as evidence often cannot be presented as a binary truth value (i.e. true or false), which is why conclusions are presented in a probabilistic way [7]. Existing automated methods can be useful for FHEs to objectively report the strength of evidence. Similarity scores could be used to compute the probability that the specific similarities/differences will occur if the prosecution hypothesis is true (the suspect wrote the signature) or if the defence hypothesis is true (another person than the suspect wrote the signature). Thus, results of objective feature selection methods could be used to support the FHEs conclusions and express the strength of evidence numerically instead of verbally. In this competition, we would like to make a first step in bridging the gap between objective biometric methods and forensic expert-based opinions.

We provided the participants with a new signature dataset, containing both genuine and forged signatures. The objective of the systems will be to distinguish genuine from forged signatures. Different types of forgeries exist [8, 12]. The best type of forgeries are over-the-shoulder forgeries, which means that the forger has been able to view the signing process of the genuine writer. Another type of forgeries are random forgeries which are made when only the name of the authentic writer is provided to the forger, without any example of the original signature [5]. The forgeries in this dataset are home improved (skilled) forgeries, meaning that the forger received a paper copy of the genuine signatures and has had the opportunity to improve the forgery by practicing. This type of forgeries are usually quite challenging to detect.

FHEs normally need at least three reference signatures of a certain quality to verify a questioned signature. These references are used to determine the intervariability of the original writer. However, sometimes a case could occur where only one reference signature is available. In this competi-

978-0-7695-3725-2/09 $25.00 © 2009 IEEE
DOI 10.1109/ICDAR.2009.216

tion, the choice was made to present the systems with this worst case scenario. The systems are expected to calculate a similarity score between one questioned and one reference signature. Our objective is to explore to what extent systems provided with only one reference signature yield promising results.

Signatures were collected both in the online and offline domain (i.e. dynamic and static information of the signing behavior was collected). This gave the participants the opportunity to compete in either the on- or offline format, or in a combination of the two. As FHEs consider movement order the most important feature, a system that combines features from the offline image with semi-online data (for example if the FHE could trace the signature and this semi-online information is used for verification) would be a major contribution.

To our knowledge, this is the first competition on skilled forgery detection in offline signature verification and the first competition where combined formats are provided. Next to that, in our dataset FHEs have judged the authenticity and complexity of the signatures [4]. This makes our competition unique because it allows us to combine expert judgements with the performance of automated methods.

2. Participants

After the competition announcement on the 20th of January 2009, 24 teams (21 from academia and three from industry) showed their interest in participating in our competition. Of those 24 teams that were registered and received the trainingset, 7 teams submitted their programs for the offline signature verification competition, and 12 teams submitted systems for the online competition. Some teams participated in both tasks and some teams submitted multiple systems for a single scenario. Finally, 15 online and 8 offline systems were evaluated, together with one system that combined both on- and offline formats. The teams are from 8 different countries (France, Germany, Japan, The Netherlands, Spain, Switserland, Tunesia and the United States of America). Table 1 shows the participating teams. After the results were announced, a few teams decided to remain anonymous.

3. Signature databases

For this competition, two different datasets are used. These datasets both contain corresponding online and offline signatures. The offline datasets contain only static information while the online datasets also contain dynamic information, which refers to the recorded temporal movement of the handwriting process. Detailed information about these datasets can be found in the README files on the SigComp09 website [1].

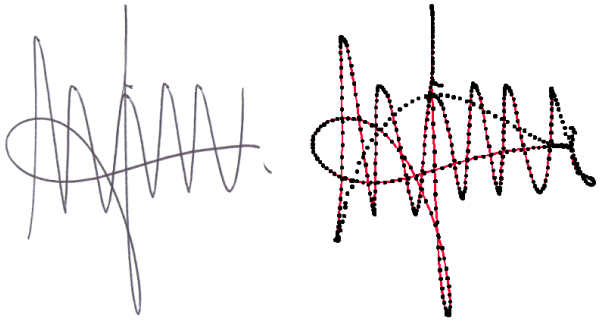

Figure 1. An off- and online signature sample

For training, the NISDCC signature collection was made available to all participants on the competition website [1]. This dataset was acquired in the framework of the WANDA project [9], a cooperative effort by Franke, Schomaker and Vuurpijl. The offline training NISDCC dataset is composed of 1920 images from 12 authentic writers (5 authentic signatures per writer) and 31 forging writers (5 forgeries per authentic signature). Due to an error during recording, the online training dataset consists of 1905 files.

The dataset collected at the Netherlands Forensic Institute (NFI) was not provided to the participants before the evaluation of the systems and consists of authentic signatures from 100 newly introduced writers (each writer wrote his signature 12 times) and forged signatures from 33 writers (6 forgeries per signature). Each authentic signature was forged by 4 writers. It has 1953 signatures for both the online and the offline dataset.

Each signature is stored in a separate file. For the online signatures, these are text files. The offline signatures are saved as PNG images. The naming convention of the files is SA_PA_xx where SA is the simulating author, PA the primary author and xx the signature ID. Genuine signatures have the same value for SA and PA. For the evaluation process, the files were renamed to avoid the class information from being revealed by the file names.

In each online signature file, the first line of the file contains a header and an integer describing the total number of coordinates in the signature. Each point contains five recorded pen-tip coordinates (x-coordinate, y-coordinate, pen pressure, azimuth angle, and elevation angle).

The offline datasets are segmented, visually inspected and then preprocessed to provide color, greyscale and binary formatted images, in both 300 and 600 dpi. An example of a colored offline signature and a plotted matching online signature can be found in Figure 1.

The online information includes pen pressure values, which measures the amount of pressure executed during writing the signature. In the NISDCC set, the pen pressure was recorded in raw values (pressure levels of the calibrated

Table 1. SigComp09 participating teams

ID	Institution	Country	Member(s)	Online	Offline	Combined
1	Biometric Recognition Group - ATVS, Univ. Autonoma de Madrid	Spain	F. Alonso-Fernandez, M. Martinez-Diaz, J. Fierrez and J. Ortega-Garcia	x	x	x
2	Parascript, LLC	USA	A. Filatov, I. Kil, T. Strunkov	x		
3	Dept. Electrical and Mechanical Engineering Seikei University	Japan	D. Muramatsu	x		
4	anonymous			x		
5	Laboratoire d'Analyse des Systmes du Littoral (LASL) Université du Littoral Côte d'Opale, Calais	France	E. Caillault	x		
6	Technical University of Munich	Germany	A. Graves	x		
7	Artificial Intelligence Department, University of Groningen	Netherlands	M. Bulacu	x		
8	University of Sfax, Faculty of Sciences of Sfax	Tunesia	M. Kherallah, L. Haddad, A.M. Alimi	x		
9	anonymous			x		
10	anonymous			x		
11	PatternLab, Lausanne	Switserland	J. Richiardi	x		
12	anonymous			x		
13	Centre de Morphologie Mathématique	France	A. Hassaïne, E. Decencière		x	
14	anonymous				x	
15	anonymous				x	
16	Center of Excellence for Document Analysis and Recognition (CEDAR), University at Buffalo, New York	USA	G. Ball, D. Pu		x	
17	Computer Vision Center, Autonomous University of Barcelona	Spain	E. Sala, E. Valveny		x	
18	German Research Center for Artificial Intelligence	Germany	M. Liwicki		x	

writing tablet), while for the NFI set, it was calculated in grams. Therefore, equations 1 (for values $0 < x \leq 25$) and 2 (for values $25 < x < 625$), where y represents the raw pen pressure and x the pen pressure in grams, were used to convert the pressure in grams of the NFI set to the raw pressure values as used in the NISDCC set. A value of $x = 0$ relates to $y = 64$, whereas $x \geq 625$ matches $y = 1024$. The correlation between y and the raw pen pressure data is 0.957 for the first equation and 0.995 for the second.

$$y = 181.550 * Ln(x) - 217.134 \quad (1)$$

$$y = 229.916 * Ln(x) - 479.731 \quad (2)$$

4. Performance evaluation

Each team was required to submit a software tool that is able to compare one questioned signature against one reference signature that outputs a similarity score (a high score indicates a high similarity between the two signatures, while a low score indicates a low similarity) and a binary decision score (0 for non match and 1 for match). The tool had to be made available as Linux or Windows-win32 command line application or a standalone java J2SE application. Some participants delivered a tool based on Matlab code.

The testing is done as follows. Each tool was evaluated by tests with the NFI signature database. The evaluation phase consisted of 6374 online matching and 9378 online non matching comparisons. The offline systems were supposed to do 15887 comparisons (6527 matches and 9360 non-matches). However, from the 8 offline systems, 4 did not reach the total number of offline comparisons. These 4 systems seemed to skip certain signature files, for different reasons (e.g. issues on the image sizes). However, it was beyond the competition to further investigate these reasons. It was decided to still evaluate these systems for the competition as not too much information was lost (the missing number of comparisons was at most 2.35%).

The number of comparisons is based on our initial assumption that the similarity score of signature A and B is the same as the similarity score for signatures B and A. The matching was therefore only done one way (A with B, A with C, B with C). However, it must be noted that not all systems have this symmetry.

The evaluation for the competition was only based on the Equal Error Rate (EER) computed on the NFI dataset. On request, EERs of the NISDCC dataset were provided to the participants to cross compare their results. Based on the similarity scores on the NFI set, we computed false rejection rates (FRR) and false acceptance rates (FAR) for different threshold values. Detection Error Trade-off (DET) curves [11] were then obtained using DETcurve plotting Matlab software described by NIST [2]. Corresponding EERs were then calculated by measuring where the line $x = y$ crosses the DET curve. We must stress that is not straightforward whether an EER of for example 3% is significantly better than an EER of 5%. Therefore, non-parametric bootstrapping [6] was used to obtain 95% confidence intervals for the EERs, as earlier applied by [3].

Table 2. Evaluation of the systems

ID	EER online NFI set (%)	ID	EER offline NFI set (%)
2	2.85	13	9.15
1a	8.33	14	15.50
7	8.41	18b	15.78
3	8.65	18a	16.10
1b	9.15	1	18.27
11a	10.18	16	23.00
1c	11.29	17	41.12
4	11.29	15	43.02
11b	12.50		
5	13.70		
10	14.39		
9	15.61		
6	19.65		
12	20.68		
8	24.23		

5. Results

Figures 2 and 3 show the corresponding DET curves for the offline and online systems respectively, as evaluated on skilled forgeries of the NFI dataset. Table 2 shows the EER results corresponding to the curves. The ranking of the systems cannot be done on these EERs alone, as some systems have almost similar EERs and may not be significantly different on a 95% confidence interval, which was tested by non-parametric bootstrapping [6]. The two winning systems are significantly better than all other systems that participated for that task. For the online task, the system submitted by Parascript, LLC (system number 2) gives the lowest EER value (2.85%) when tested with the NFI skilled forgeries, while for the offline task, the system from Centre de Morphologie Mathématique (system number 13) gives the lowest EER value (9.15%) on the forgery data from the evaluation set. Only the Biometric Recognition Group (Univ. Autonoma de Madrid) participated in the combined version of the competition. They submitted a system that loaded both an on- and offline reference signature, and compared that with both an on- and offline questioned signature. Their EER is 8.17%.

6. Discussion

As expected, the Equal Error Rate (EER) results of the online systems (yielding an EER of 2.85% for the best system) are much better than those of the offline systems (reaching an EER of 9.15% for the best system). Nevertheless, the offline systems also reach very promising results, especially when taking into account the fact that only one reference sample was given. With only one reference sample, the systems are not able to compute the within variability of a writer. Differences between two signatures of one writer occur because of the neural biomechanical vari-

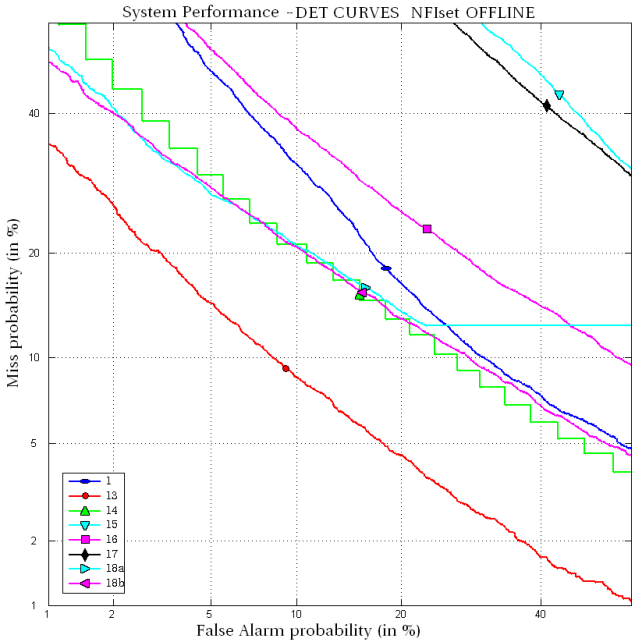

Figure 2. DET curves of the offline systems

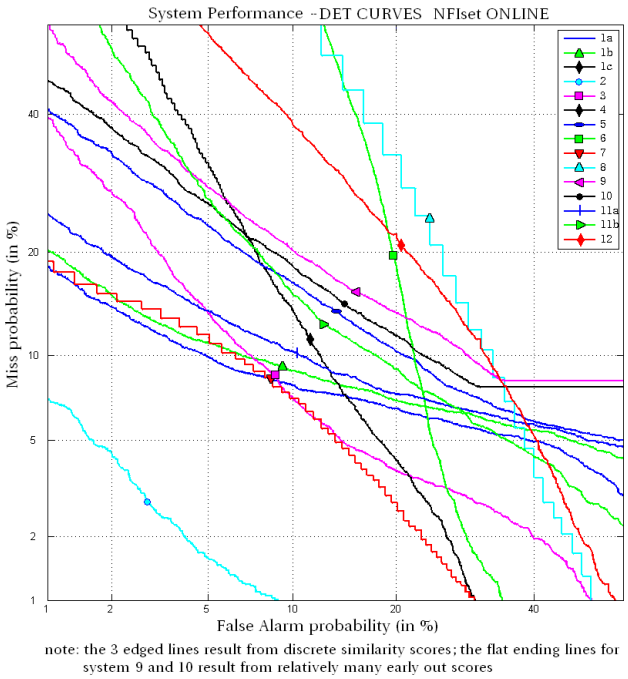

Figure 3. DET curves of the online systems

ations that underlie the (normal) signing process. The systems are therefore generally expected to yield better results when presented with three reference signatures or more.

The Netherlands Forensic Institute (NFI) dataset used for this competition has been partly evaluated by Forensic Handwriting Experts (FHEs) of the NFI. With six reference signatures, the FHEs gave an incorrect conclusion in 3,13% of the verifications [4]. However, we must stress that this percentage is not an EER. It must be noted that in this experiment, the FHEs were not given the possibility to give an inconclusive answer. In real forensic casework, FHEs need to decide whether differences between two signatures are the result of a disguise (the authentic writer changing his own signature with intent of denial) or a forgery. Therefore, FHEs are very careful when judging differences between signatures. As a wrong conclusion can have a large impact for either party, FHEs will not draw conclusions when they are not certain about the source of the questioned signature.

For the near future, we are especially interested in the performance of the offline systems, as only offline information is available to the FHEs. We want to extend our investigations by testing how the systems perform on non-matching genuine signatures. A non-matching genuine is a comparison between signatures from two different authentic writers. The expectation is that non-matching genuines are easier to distinguish than skilled forgeries, as the probability that the two signatures differ on dynamic as well as visual features is higher. It is interesting to test the systems on this scenario because this will give insight in signature characteristics that occur in a large population.

Further, we plan to combine the expert judgements on authenticity and complexity of the signatures with performance of the systems. FHEs from the NFI gave a complexity judgement for all 100 types of different genuine signatures in the evaluation dataset [4]. FHEs generally believe that complex signatures are more difficult to forge and are more unique. Chances for finding a random match with another signature decrease with complexity. Therefore, we are particularly interested to investigate whether the systems are able to detect skilled forgeries of less complex signatures.

To conclude, the results of this competition gave us more insight in the performances of the participating systems and has already served as a step towards collaboration with developers in the field. We hope this collaboration will lead to more objective results in the future, that can be used to support the experts judgements on authenticity of questioned signatures. To do this, systems should also produce explainable results so that experts can explain to judges in court how they have reached their conclusions.

Acknowledgments

The authors would like to thank Niels van Eijck, Reinoud Stoel (Netherlands Forensic Institute) and Cor Veenman (University of Amsterdam) for their contributions to the competition. We would also like to thank Linda Alewijnse (NFI) for collecting the evaluation data.

References

[1] http://sigcomp09.arsforensica.org, April 2009.

[2] http://www.itl.nist.gov/iad/mig//tools/DETware_v2.1.targz.htm, April 2009.

[3] I. Alberink and A. Ruifrok. Repeatability and reproducability of earprint acquisition*. *Journal of Forensic Science*, 53(2):325–330, 2008.

[4] L. C. Alewijnse, C. E. van den Heuvel, R. D. Stoel, and K. Franke. Analysis of signature complexity. In *Proc. of the 14th Conference of the Int. Graphonomics Society, Dijon*, 2009, Accepted.

[5] J. G. A. Dolfing, E. H. L. Aarts, and J. J. G. M. V. Oosterhout. On-line signature verification with hidden markov models. *Proc. of the 14th International Conference on Pattern Recognition*, 2:1309–1312, 1998.

[6] B. Efron and R. J. Tibshirani. *An Introduction to the Bootstrap*. Chapman & Hall, New York, 1993.

[7] I. Evett. Bayesian inference and forensic science: Problems and perspectives. *The Statistican*, 36:99–105, 1987.

[8] J. Fierrez-Aguilar, J. Ortega-Garcia, D. Torre-Toledano, and J. Gonzalez-Rodriguez. Biosec baseline corpus: A multimodal biometric database. *Pattern Recognition*, 40, 2007.

[9] K. Franke, L. Schomaker, C. Veenhuis, C. Taubenheim, I. Guyon, L. Vuurpijl, M. van Erp, and G. Zwarts. A generic framework applied in forensic handwriting analysis and writer identification. In *Proc. 3rd International Conference on Hybrid Intelligent Systems*, pages 927–938. Press, 2003.

[10] K. Franke and S. N. Srihari. Computational forensics: Towards hybrid-intelligent crime investigation. *Information Assurance and Security, International Symposium on*, 0:383–386, 2007.

[11] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki. The DET curve in assessment of detection task performance. In *Proc. Eurospeech '97*, pages 1895–1898, 1997.

[12] D.-Y. Yeung, H. Chang, Y. Xiong, S. George, R. Kashi, T. Matsumoto, and G. Rigoll. SVC2004: First International Signature Verification Competition. In *Proceedings of the International Conference on Biometric Authentication (ICBA), Hong Kong*, pages 16–22. Springer, 2004.