2010 12th International Conference on Frontiers in Handwriting Recognition

Forensic Signature Verification Competition 4NSigComp2010 - Detection of Simulated and Disguised Signatures

Marcus Liwicki*, C. Elisa van den Heuvel†, Bryan Found‡, and Muhammad Imran Malik*

*German Research Center for AI (DFKI), Germany
Email: Firstname.Lastname@dfki.de
†Netherlands Forensic Institute, The Hague, The Netherlands
Email: E.van.den.Heuvel@nfi.minjus.nl
‡Victoria Police, La Trobe University, Melbourne, Australia

***Abstract*—This competition scenario aims at a performance comparison of several automated systems for the task of signature verification. The systems have to rate the probability of authorship and non-authorship of signatures. In particular they have to determine whether questioned signatures are simulated disguised or the normal signature of the reference writer. Furthermore, the results will be compared to forensic handwriting examiners (FHEs) opinions on the same tasks. As such, to the best of the authors' knowledge, this scenario will be the first attempt in literature to relate system performances to the performance of FHEs who gave their opinion on exactly the the same signatures.**

I. Introduction

The topic of writer identification and verification has been addressed in the literature for several decades [1], [2]. Usually the task is to identify the writer of a handwritten text or signature or to verify his or her identity. Work in writer verification can be differentiated according to the available data. If only a scanned image of the handwriting is available then writer classification is performed with *offline* data. Otherwise, if temporal and spatial information about the writing is available, writer classification is performed with *online* data. Usually, the former task is considered to be less difficult than offline classification [2].

Surveys covering work in automatic writer identification and signature verification until 1993 are given in [2]. Subsequent works up to 2000 are summarized in [3]. Most approaches are tested on specially collected data sets which were acquired in controlled environments. In the past, several competitions were organized to measure the detection rate of several classifiers:

- First international Signature Verification Competition (SVC 2004), online data, 5 reference signatures
- BioSecure Signature Evaluation Campaign 2009, online data, 5 reference signatures
- SigComp 2009 [4], online and offline data, 1 reference signature

Unfortunately, current research in the field of signature verification does not take the real needs of Forensic Handwriting Experts (FHEs) into account. In their real casework they often work with offline signatures produced in different environments. The most crucial fact is that they also have to deal with disguised signatures, where the author tries to disguise his or her handwriting in order to make it seem to be a simulated signature. To the best of the authors' knowledge there has been no reported signature verification competition where disguised signatures were also present in the testing data.

The task considered in this paper aims at a comparison between FHEs opinions on authorship of signatures and the systems performances to determine whether questioned signatures are simulated disguised or the normal signature of the reference writer.

II. Background

Forensic signature verification is done by visual comparison by trained FHEs. The authenticity of the questioned signature is estimated by weighing the particular similarities/differences observed between the features of the questioned signature and the features of several known signatures of a reference writer. Automated signature verification tools can help FHEs in evaluating the probability of the evidence in light of the two research hypotheses under investigation:

H1: The questioned signature is an authentic signature normally used by the reference writer;
H2: The questioned signature is not authentic but rather
  a: it is simulated by another writer than the reference writer;
  b: it is disguised by the reference writer;

The FHE weighs the observations in light of two hypotheses H1 vs. H2. The interpretation of the observed similarities/differences in signature analysis is not as straightforward as in other forensic disciplines such as DNA or fingerprint evidence, because signatures are a product of a behavioral

978-0-7695-4221-8/10 $26.00 © 2010 IEEE
DOI 10.1109/ICFHR.2010.116

process that can be manipulated by the reference writer himself, or by a person other than the reference writer. In signature verification research, a 100% spatial match does not necessarily support Hypothesis 1, because a perfect match can occur if a signature is traced. Also, differences between signatures do not necessarily support Hypothesis 2a, because slight changes can be put into a signature image by the reference writer when disguising his signature for the purpose of denial, or can occur due to a within-writer variation.

Because forensic signature verification is performed in a highly subjective manner, the discipline is in need for a scientific, objective base. The use of automatic signature verification tools can objectify the FHEs opinion about the authenticity of a questioned signature. However, to our knowledge, signature verification algorithms are not widely accepted by the FHEs. The objective of this competition is to compare automatic signature verification performances on new unpublished forensic-like datasets to bridge the gap between recent technology developments and the daily casework of the forensic examiner. We consider the opportunity to conduct a performance evaluation of algorithms a basic contribution in establishing the scientific basis for the discipline of forensic signature comparison.

III. Data

The collection contains offline signature samples. The signatures were collected under supervision of Bryan Found and Doug Rogers in the years 2002 and 2006, respectively. The images were scanned at 600dpi resolution and cropped at the Netherlands Forensic Institute for the purpose of this competition.

A. Data Description

The La Trobe signature collection for training contains 209 images. The signatures comprise 9 reference signatures by the same writer A and 200 questioned signatures. The 200 questioned signatures comprise 76 genuine signatures written by the reference writer in his/her normal signature style; 104 simulated signatures (written by 27 forgers freehand copying the signature characteristics of the reference writer); 20 disguised signatures written by the reference writer.

The La Trobe signature collection for testing contains 125 signatures. The signatures comprise 25 reference signatures by the same writer B and 100 questioned signatures. The 100 questioned signatures comprise 3 genuine signatures written by the reference writer in his/her normal signature style; 90 simulated signatures (written by 34 forgers freehand copying the signature characteristics of the reference writer); 7 disguised signatures written by the reference writer.

All writings were made using the same make of ball-point pen and using the same make of paper. The disguise process comprises an attempt by the reference writer to purposefully alter his/her signature in order to avoid being identified or for him/her to deny writing the signature. The simulation process comprises an attempt by a writer to imitate the reference signature characteristics of a visual or mental model.

B. Training Set

The participants were provided with the following training set.

Collection of genuine signatures of the reference writer A: The following signatures were supplied by the reference writer:

- 15 normal signatures per day over a seven day period; 9 signatures were chosen from this subset as reference set to which the questioned signatures are to be compared.
- 6 disguised signatures per day over a seven day period. In addition to these signatures, the reference writer provided an additional 81 genuine signature samples (27 pages containing three signatures per page). Signatures from this supplementary pool were provided to the forgers as examples of the signature they were required to forge.

Generation of simulated signatures: The 27 'forgers' were volunteers drawn from groups such as secondary school teachers and professional organizations. Each of the forgers was provided with 3 normal samples of the signature written by the reference writer. Forgers were instructed that they could use any or all of the supplied reference signatures as models for their simulations. Forgers were also instructed that their simulations must be unassisted (not tracings). Each forger was asked to complete the following task:

- Inspect the genuine signature and, without practice, immediately attempt to forge it 3 times.
- Practice simulating the genuine signature 15 times then simulate the signature an additional 3 times.

C. Test Set

Collection of genuine signatures of the reference writer B: Similar to the training set data collection, the reference writer provided a set of signatures over a five day period; 25 signatures were chosen from this subset as reference set to which the questioned signatures are to be compared. The test data contains 3 genuine signatures and 7 disguised signatures.

Generation of simulated signatures: For the generation of simulated signatures a 34 adult 'forgers' were used. These individuals were volunteers. The forgers were either 'lay' persons or calligraphers. The test data contains 90 simulations. Note the huge difference between authentic data (3 genuine +7 disguised signatures) vs. simulations (90 signatures). This is not a problem for the evaluation of the system performances because we evaluate the equal error rate in Section V.

IV. Submitted Systems

In total, we received six systems for the competition. In the following we will list the participants and a small description of their systems if we were provided with a description.

A. Biometric Recognition Group - ATVS EPS - Univ. Autonoma de Madrid

The off-line system submitted is based on the fusion of two machine experts, one based on **local** analysis of the image [5] and a second approach based on **allographic** analysis [6]. The **local** matcher uses contour level features [5]. It is based on features proposed for writer identification and verification using images of handwriting documents [6]. It computes the orientation of local contour fragments, as well as its curvature. The contour-direction distribution f_1 is extracted by considering the orientation ϕ_1 of local contour fragments and computing its probability distribution. Curvature of the signature contour f_2 is computed by considering two contour fragments attached at a common end pixel and computing the joint probability distribution of the directions ϕ_1 and ϕ_2 between that pixel and both fragments. As the algorithm runs over the contour, the two histograms of f_1 and f_2 are built, which are then normalized to a probability distribution. To compute the similarity between two signature images, the χ^2 distance is used. This matcher outputs two distances, one for f_1 and another one for f_2. The matcher based on **allographic** analysis considers a signature as an stochastic pattern of handwritten shapes [6]. The probability distribution function (PDF) of these shapes in a given signature image is used to characterize the identity of the writer, which is computed using a common codebook of shapes obtained by means of clustering techniques. The codebook is generated using an external database of handwritten signatures [7]. This way, the codebook provides a common shape space and the PDF captures the individual shape usage preference of the signer. To compute the similarity between two signature images, the χ^2 distance is used. Finally, **fusion** of the two machine experts is performed via linear combination of the individual scores [8]. Linear regression is used to compute the optimal fusion weights.

B. Université du Littoral Cote d'opale LISIC

To compare two signatures, we compute a DTW similarity on their projections obtained by Mojette transform. Each image is first pretreated like that:

- Compute the (gray level) luminance matrix from the RGB image
- Reduce the matrix in a bounding box
- Compute the inverse video matrix

Then, we compute all projections of the image according to a 2-rank Farey series extended to [0,Pi]. These projections are computed according to Mojette transform algorithm. This encoding is derived of the Radon transform. Each projection can be interpreted as a spatial histogram of the luminance in a fixed axis.

To answer the nature of the signature (naturally written or not), we compute the similarity matrix on the given reference signatures. ¿From this matrix, we obtain a luminance threshold vector based on 1-NN algorithm. If there is a sufficient number of similarities between the questioned signature and the reference signature higher than this threshold, we consider the questioned signature according to a naturally handwriting process.

C. NifiSoft, Saint-Etienne, France

The proposed method computes several features based on the number of connected components, number of holes, moments, projections, distributions, position of barycenter, number of branches in the skeleton, Fourier descriptors, tortuosities, directions, curvatures and chain codes. Each feature F_i is computed for the questioned signature $F_i(q)$ and the N reference signatures $F_i(r), (r = 1, \ldots, N)$. The average absolute difference between the value of the feature F_i in the questioned signature and its values in the reference signatures is then computed. The obtained differences are combined via a logistic regression classifier trained either on the 4NSigComp2010 database (partial training method) or on both 4NSigComp2010 and SigComp09 databases (full training method).

D. Parascript LTD, USA

This software is described at `http://www.parascript.com/`

E. Sabanci University, Turkey

After preprocessing and size normalization steps, we tesselate the image into a fixed number of zones using polar coordinate representation and extract gradient information in each zone. The extracted features of the query signature classified using a user-dependent SVM that is trained with the reference signatures of the user and negative examples.

We also present a combination classifier, which does score level combination of the user-dependent SVM classifier described above, with one based on normalized correlation and another similar to the first one, but using a user-independent SVM classifier.

F. Anonymous

This submission did not include a detailed description.

V. Comparison Experiments

Basically the underlying aim here is to compare the performance of automated systems against the judgements given by professional Forensic Handwriting Experts (FHEs). The systems presented their opinion by means of the following two output values for each of the questioned signatures.

1: A Probability Value P between 0 and 1.

Table I
INTERPRETATION OF THE OUTPUT

Decision	Probability		
Value	$P > t$	$P < t$	$P = t$
1	authentic	misleading	inconcl.
2	disguise	simulation	inconcl.
3	inconcl.	inconcl.	inconcl.

Table II
ASSESSMENT OF THE OUTPUT

True	Probability		
Answer	$P > t$	$P < t$	$P = t$
authentic	correct	incorr.	incorr./ignored
disguise	correct	incorr.	incorr./ignored
simulation	incorr.	correct	incorr./ignored

2: A Decision Value D which could be either 1, 2 or 3.

The Probability Value P was compared to a predefined threshold t. A higher value ($P > t$) indicated that the questioned signature was most likely a genuine one. A lower value ($P < t$) indicated that the questioned signature was not genuine, meaning that it was not written by the reference author. A probability value of ($P = t$) was considered as inconclusive.

The Decision Value D represents the system's decision about the process by which the questioned signature was most likely generated. A Decision Value of 1 means that the underlying writing is natural: there was no or not enough evidence of any simulation or disguise attempt and the signature was written by the reference author. Decision Value 2 represented that the underlying writing process was unnatural: there was evidence of either a simulation or disguise attempt. Whereas a Decision Value 3 showed that the system was unable to decide if the underlying process was natural or unnatural: no decision could be made wether the signature was genuine, simulated or disguised.

The output reference table is provided in Table I. It presents the various output possibilities. In this table, a value of P greater than t with output 1 means correct genuine authorship, with output 2, on the other hand, means that the author has made an attempt to disguise her/his identity. If the Decision Value is 3 then with any value of probability it is simply inconclusive. Any value of P less than t with decision value 2 indicates that the questioned signature was a result of a simulation or disguise process. The final assessment of the output values is given in Table II. Note that we have performed two experiments, one where we ignored the inconclusive ratings and another where we counted them as errors. There was no significant difference of the results and especially no change in the ranking of the systems. Therefore we just report on the results obtained with counting them as errors, because of space limitations in this paper.

We performed all the tests at a machine with following specifications

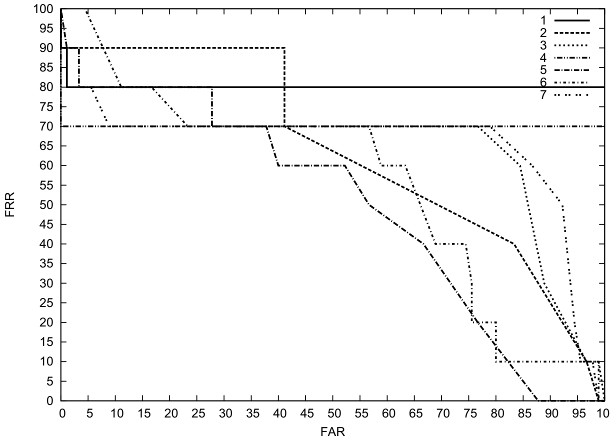

Figure 1. ROC-curve with respecting the disguised signatures

- Processor: Intel Dual Core 1.73 GHz
- Memory: 1GB
- OS: WinXP Professional

We took all the 25 reference signatures for performing all the tests.[1]

The results will be reported in receiver operating characteristic (ROC) curves, containing the error rates based on varying the threshold t (see Figs. 1 and 2). The x-Axis shows the false accept rate (FAR), i.e., the percentage of wrongly accepted simulations. The y-Axis shows the false rejection rate (FRR), i.e., the percentage of signatures by the reference writer which have been wrongly interpreted as simulations. Note that by drawing a line with slope 1, we can read the Equal Error Rate of the several systems.

The summary of the results is shown in Table III. The IDs presented in this table are also used in the Figures 1 and 2. Note that NifiSoft submitted two systems which are now denoted as 3 and 7.

A crucial observation is that most of the systems could not handle disguised signatures. There was only one system that performed well on the disguised signatures, but this system showed a large error rate in detecting simulated signatures. We made a second set of experiments where we excluded the disguised signatures and compiled the results. These results are quite encouraging (see Fig. 2 and Table III).

A. Comparison with Human Experts

The evaluation of FHEs opinions has been carried out in 2006 by Bryan Found and Doug Rogers. FHEs can validate their opinions by participating in the so-called proficiency tests. Often, this is the only way for FHEs to check their opinions with true scores. The experts were provided with

[1] System 6 only produced results when taking 9 reference signatures and it did not work with more than nine reference signatures. Therefore we have presented the first 9 reference signatures and include the results of System 6 for the purpose of completeness.

Table III
SUMMARY OF THE RESULTS

System	ID	Time (s)	#Correct	#Errors			Acc.	FAR	FRR	EER	EER w/o disguised
				D	F	G					
AVTS	1	312	90	7	1	2	90.0	1.1	90	80	34
LISIC	2	1,944	54	7	37	2	54.0	41.1	90	58	41
NifiSoft (full)	3	85	75	7	18	0	75.0	20.0	70	70	8
Parascript	4	19	92	7	0	1	92.0	0.0	80	70	0
Sabanci	5	45	80	7	12	1	80.0	13.3	80	55	28
Anonymous	6	730	20	1	79	0	20.0	87.0	10	60	21
NifiSoft (partial)	7	65	91	7	1	1	91.0	1.1	80	70	8

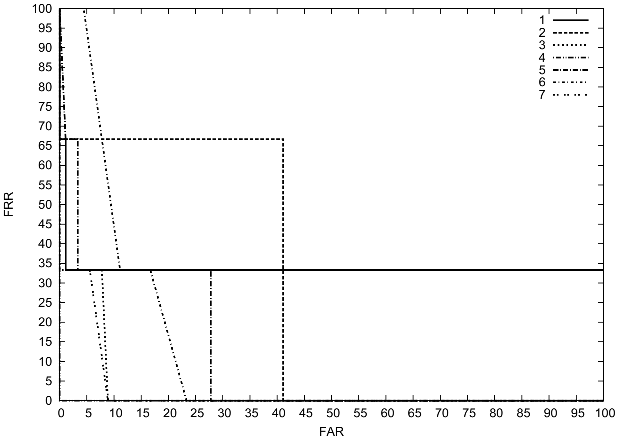

Figure 2. ROC-curve without respecting the disguised signatures

Table IV
RESULTS OF FHEs OPINIONS

	Genuine	Disguise	Simulation
correct	93	10	1151
misleading	2	111	113
inconclusive	0	96	1,526

a hardcopy photograph of each signature and an answer booklet. Examiners were informed that the date range over which the reference material was taken was around the time that the questioned samples were written. They were also informed that a calligrapher group was used for producing the simulations. FHEs are asked to express their opinion on authenticity on a five-point scale. A score of 1 means the opinion that the questioned signature was written by the reference writer. A score of 2 means that there are indications that there are indications that the signature was written by the reference writer. A score of 3 means inconclusive. A score of 4 means that there are indications the signature was written by another writer than the reference writer. A score of 5 means the opinion that the questioned signature was written by another writer. Next to that, they were asked to produce a decision score on the underlying writing process. We provided similar conditions to the automated systems as were given to the human experts.

In total, 33 answer booklets were submitted, thereof 11 peer reviewed responses (cross-checked by a second FHE) and 22 individual responses (not peer-reviewed). A total of 3100 authorship opinions were expressed by the group. Of these opinions 1254 (40.5 %) were correct, 224 (7.2 %) were misleading and 1622 (52.3 %) were inconclusive. This translates into an error rate of 15.2 % on the decisions (Accuracy of 84.8 %).

More details of the results appear in Table IV. As can be seen, in the test of 2006 FHEs had significant difficulties with the disguised signatures. Unfortunately, it is impossible to rate the EER of the human experts, since there is no threshold which could be balanced. In real casework, forensic scientist cannot report conclusions based on biometric identification techniques that make use of thresholds, because making use of thresholds allows the forensic scientist to take the actual decision that belongs to court. During the last decade a common framework for evidence evaluation and its interpretation in court has been discussed amongst forensic scientist. The Bayesian approach has been proposed as a theoretical framework, and Tippett plots are argued to be used to represent forensic system performances. Systems that use ROC curves to suite performance evaluation in detection tasks can be adapted into a forensic system according to the Bayesian approach [9].

B. Summary

Considering the results of the 4NSigComp2010 and the importance of forensic handwriting verification we can say that computer scientists should also focus on disguised signatures, since it is a crucial aspect in real FHEs' casework. For a next competition at ICFHR we plan to use a larger test set to investigate the diversity of the recognizers more thoroughly. Regarding simulations systems produced quite good results. Regarding genuine signatures, large and diverse test sets where signatures are produced by the different authors under various different psychological and physical conditions may also yield interesting results.

An interesting observation of this contest is that the performance of the automated systems is not so far away from human decisions. A more detailed analysis will be performed in future to directly assess the strengths and weaknesses of several classifiers. Also, in regard to the fact that FHEs cannot make use of thresholds but need to provide the court the likelihood of the two competing hypothesis, in a next competition we will reference existing system scores into a forensic system using within-source and between-source variabilities according to the Bayesian approach.

REFERENCES

[1] R. Plamondon and G. Lorette, "Automatic signature verification and writer identification – the state of the art," *Pattern Recognition*, vol. 22, pp. 107–131, 1989.

[2] F. Leclerc and R. Plamondon, "Automatic signature verification: The state of the art 1989–1993," *Int. Journal of Pattern Recognition and Artificial Intelligence*, vol. 8, no. 3, pp. 643–660, 1994.

[3] R. Plamondon and S. N. Srihari, "On-line and off-line handwriting recognition: A comprehensive survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, pp. 63–84, 2000.

[4] V. Blankers, C. van den Heuvel, K. Franke, and L. Vuurpijl, "The ICDAR 2009 signature verification competition," in *Proc. Int. Conf. Document Analysis and Recognition*, vol. III, 2009, pp. 1403–1407.

[5] A. Gilperez, F. Alonso-Fernandez, S. Pecharroman, J. Fierrez, and J. Ortega-Garcia, "Off-line signature verification using contour features," in *International Conference on Frontiers in Handwriting Recognition*, 2008.

[6] S. L. Bulacu, M., "Text-independent writer identification and verification using textural and allographic features," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 29, no. 4, pp. 701–717, 2007.

[7] J. Fierrez, J. Galbally, J. Ortega-Garcia, A. F. Freire, M., D. Ramos, D. Toledano, J. Gonzalez-Rodriguez, J. Siguenza, J. Garrido-Salas, E. Anguiano-Rey, G. Gonzalez-de-Rivera, R. Ribalda, M. Faundez-Zanuy, J. Ortega, V. Cardenoso-Payo, V. C. Viloria, A., I. J. Moro, Q., J. Sanchez, I.Hernaez, C. Orrite-Urunuela, F. Martinez-Contreras, and J. Gracia-Roche, "Biosecurid: A multimodal biometric database," *Pattern Analysis and Applications*, 2009.

[8] F. Alonso-Fernandez, J. Fierrez, D. Ramos, and J. Ortega-Garcia, "Dealing with sensor interoperability in multi-biometrics: The UPM experience at the Biosecure Multimodal Evaluation 2007," in *Proc. SPIE Defense and Security Symposium, Biometric Technologies for Human Identification*, vol. 6944, 2008, pp. 69 440J1–69 440J12.

[9] J. Gonzalez-Rodriguez, J. Fierrez-Aguilar, D. Ramos-Castro, and J. Ortega-Garcia, "Bayesian analysis of fingerprint, face and signature evidences with automated biometric systems," *Forensic Science International*, vol. 155, pp. 126–140, 2005.