Signature Verification Competition for Online and Offline Skilled Forgeries (SigComp2011)

Marcus Liwicki*, Muhammad Imran Malik*, C. Elisa van den Heuvel[†], Xiaohong Chen[‡],
Charles Berger[†], Reinoud Stoel[†], Michael Blumenstein[§], and Bryan Found[¶]
*DFKI, Germany, Email: firstname.lastname@dfki.de
[†]Netherlands Forensic Institute, The Hague, The Netherlands
Email: E.van.den.Heuvel@nfi.minjus.nl, c.berger@nfi.minjus.nl, reinoud@holmes.nl
[‡]Forensic Science Institute, Ministry of Justice, Shanghai, China
Email: ccpccxh@hotmail.com
[§]Griffith University, Australia, Email: m.blumenstein@griffith.edu.au
[¶]La Trobe University, Melbourne, Australia, Email: Email b.found@latrobe.edu.au

***Abstract*—The Netherlands Forensic Institute and the Institute for Forensic Science in Shanghai are in search of a signature verification system that can be implemented in forensic casework and research to objectify results. We want to bridge the gap between recent technological developments and forensic casework. In collaboration with the German Research Center for Artificial Intelligence we have organized a signature verification competition on datasets with two scripts (Dutch and Chinese) in which we asked to compare questioned signatures against a set of reference signatures. We have received 12 systems from 5 institutes and performed experiments on online and offline Dutch and Chinese signatures. For evaluation, we applied methods used by Forensic Handwriting Examiners (FHEs) to assess the value of the evidence, i.e., we took the likelihood ratios more into account than in previous competitions. The data set was quite challenging and the results are very interesting.**

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 or a camera captured 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 the 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

Most of the current research in the field of signature verification does not take the real needs of Forensic Handwriting Experts (FHEs) into account. In their casework they often work with signatures produced in various real world environments. These signatures are more difficult to analyze compared to the signatures produced in controlled environments. FHEs also have to deal with possibly disguised signatures, where the author tries to disguise his or her handwriting in order to make it seem to be a forgery. The 4NSigComp2010 [5] was the first signature verification competition focusing explicitly the classification of disguised, simulated and genuine signatures.

We have now organized the Signature Verification Competition for Online and Offline Skilled Forgeries (SigComp2011). The major emphasis of this competition is not the possibility of disguised signatures but to motivate the signature verification community to enable their systems to compute the likelihood ratios instead of just computing the evidence (for more details see [6]). This is very important as it allows one to combine the FHE's evidence (from the results of an automated system) with other evidence presented in a court of law. In this competition we ask to produce a comparison score (e.g. a degree of similarity or difference), and the evidential value of that score, expressed as the ratio of the probabilities of finding that score when the questioned signature is a genuine signature and when it is a forgery (i.e. the likelihood ratio). Note that this competition has introduced a paradigm shift from the "decision paradigm" to an evidential value that impacts the task in the competition.

1520-5363/11 $26.00 © 2011 IEEE
DOI 10.1109/ICDAR.2011.294

The issue is not the pure classification, since

- The FHE cannot and was never asked to decide on authorship,
- the FHE cannot know the probability of authorship based on handwriting comparison alone, and
- classification brings with it the probability of an error of which the cost is undefined.

The true issue is to find the likelihood ratio (LR) for a comparison: the probability of finding a particular score given that Hypothesis $H1$ is true, divided by the probability of finding the score when the alternative Hypothesis $H2$ is true. $H1$ corresponds to intra-source scores (same author) and $H2$ to inter-source scores (different authors).

The relevant graphs therefore show histograms of some measure of similarity (or difference; or any continuous measure that used to be compared to some threshold in a classification task) for intra-source and inter-source comparisons. Such graphs make it possible to assess the value of the evidence given both hypotheses, which is of major importance to forensic experts and the courts. Therefore, in this competition we have had a closer look at the likelihood ratios.

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.

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 process that can be manipulated by the writer. In this competition only such cases of $H2$ exist, where the forger is not the reference writer. In signature verification research, a 100% perfect match does not necessarily support $H1$, because a perfect match can occur if a signature is traced. Also, differences between signatures do not necessarily support $H2$, because slight changes can occur due to a within-writer variation.

Since forensic signature verification is performed in a highly subjective manner, the discipline is in need for scientific, objective methods. The use of automatic signature verification tools can objectify the FHE's opinion about the authenticity of a questioned signature. However, to our knowledge, signature verification algorithms are not widely used by the FHEs. The objective of this competition is to compare automatic signature verification performances on new, unpublished, forensically relevant datasets to bridge the gap between recent technological developments and the daily casework of FHEs.

Table I
Number of Authors (A) and number of Genuine (G) (reference (GR) and Questioned (GQ)) and Forged (F) Signatures in the Chinese data set

Training Set	Training			Test			
	A	G	F	A	GR	GQ	F
Offline	10	235	340	10	116	120	367
Online	10	230	430	10	120	125	461

Table II
Number of Authors (A) and number of Genuine (G) (reference (GR) and Questioned (GQ)) and Forged (F) Signatures in the Dutch data set

Training Set	Training			Test			
	A	G	F	A	GR	GQ	F
Offline	10	240	123	54	648	648	638
Online	10	330	119	54	648	648	611

III. Data

Data collected from realistic, forensically relevant situations were used in this competition. Signature samples were collected while writing on a paper attached to a digitizing tablet. The collected signature data were made available in an online and offline format. Participants could choose to compete on the online data or offline data, or on both data formats.

The collection contains offline and online signature samples. Signatures were either genuine: written by the reference writer, or a simulation: simulated by another writer than the reference writer. The offline data sets consisted of PNG images, scanned at 400 dpi, RGB color. The online datasets consisted of ascii files with the format: X, Y, Z (per line) (sampling rate: 200 Hz, resolution: 2000 lines/cm, precision: 0.25 mm). For collection of these samples we used a WACOM Intuos3 A3 Wide USB Pen Tablet and collection software: MovAlyzer. A preprinted paper was used with 12 numbered boxes (width: 59 mm, height: 23 mm). The preprinted paper was placed underneath the blank writing paper. Four extra blank pages were added underneath the first two pages to obtain a soft writing surface.

Besides the detection of skilled forgeries of Western signatures, this competition also introduced a novel set of Chinese signatures. The purpose of using these two data sets was to evaluate the validity of the participating systems on both Western and Chinese signatures.

A. Data Sets

For both the online and offline cases, the data was divided in training and test sets having different naming conventions. Further details about the number of contributing authentic authors, forgers, number of authentic reference signatures and forgeries for both the training and test sets of Chinese and Dutch are provided in Tables I and II respectively. Note that due to minor problems during the acquisition the numbers of signatures in the online data sets differ from

Table III
OVERVIEW OF THE SUBMITTED SYSTEMS

ID	Institution	Mode
1	Sabanci	both
2	Anonymous-1	offline
3	HDU	offline
4	xyzmo(1)	online
5	xyzmo(2)	online
6	Qatar (Chinese optimization)	both
7	Qatar (Dutch optimization)	both
8	DFKI	offline
9	Anonymous-2	both

those in the offline data sets. However, this issue has no impact on the systems' performance, since 12 reference signatures could always be used (see below). Furthermore, while the training signatures were provided without restrictions on which signatures were used as reference signatures, the testing signatures have been divided by us into reference and questioned signatures.

IV. SUBMITTED SYSTEMS

In total, we received thirteen systems from six institutions for this competition. In the following we will list the participants and their brief descriptions. Participants were allowed to be anonymous upon request.

A. Sabanci University

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 are classified using a user-dependent support vector machine (SVM) that is trained with the reference signatures of the user and negative examples.

B. Anonymous-1

The method utilizes a modified direction feature and microstructure feature, both of which are based on the signature's boundary. The modified direction feature not only extracts direction information but also detects transitions between background and foreground pixels. For each transition, the location of the transition and the direction values are stored. The grid microstructure feature records the positions of some special contour pixel pairs in every local grid, which are used to calculate the appearance probability of different position pairs and express the writing style by the probability density distribution. Then the system adopts an SVM as classifier. In the training stage, the positive samples are authentic signatures from the reference writer; the negative samples are all the offline forgery signatures. In the verification stage, using the "-b" parameter of libsvm, it will get the similarity score P1 for genuine signatures and score P2 for forgeries. Then it uses log(P2)-log(P1) as log-likelihood-ratio.

C. Hong Duc University (HDU)

The system HDUSigVerify includes two main phases: the evidence estimation phase and the calibration phase. For every two signatures, we compute two types of descriptors (a local one and a global one) in order to gain robustness as well as precision. The local descriptors are locally computed at every sampled point based on the gradient in the gray scale image. The global descriptors are computed in a skeleton image by using ShapeContext [7]. The matching step is carried out by using the technique from the Linear Assignment Problem (LAP) [8]. Particularly, we carry out the following stages in the first phase: Pre-processing: to remove noises, small blobs and the rectangle surrounding the signature (if any). The Hough transforms are employed to remove the rectangles in signature images. Binarization and thinning: we employed Otsu's technique to do binarization and then the thinning step is carried out to obtain a skeleton image of signature. The skeleton is then smoothed to remove "unwanted" segments (e.g. very short branches connecting to main lines). Sampling: there are typically about 2000-2500 pixels for each skeleton image and in order to employ the ShapeContext descriptor to find candidate matches between two signatures, we sample the skeleton image to obtain about 300-500 pixels (i.e. sampled pixels). ShapeContext descriptors and matching: The 1D matching technique (DWT) is often used in literature for signature matching. One advantage of this technique is that it is able to find optimal matches by dynamically wrapping the signals over time. However, in order to use DWT we need to transform the signature image from 2D space into 1D space. For offline signatures, this step is not reliable and causes information loss. In order to take advantage of the DWT for 2D matching, we adapt the ShapeContext descriptors and propose a postprocess to refine matches as follows. We sparsely sample for one signature and densely sample for the other one. (1) Compute ShapeContext descriptors for every sampled pixel. (2) Compute a cost-matching matrix based on ShapeContext descriptors and then apply LAP to find candidate matches. (3) Apply RANSAC [9] to remove geometry-inconsistent matches (4) Compute a cost-matching matrix based on the RANSAC model and then apply LAP again to find optimal matches. Subsequently, we compute local descriptors: A circular window is placed at every sampled pixel in the gray scale image to build up a histogram of orientation and magnitude gradients. The radius of the window is the thickness of signature at every sampled pixel (this makes the descriptors scale invariant). The histograms are then normalized to unit length in order to obtain illumination changes. For rotation invariance, the orientation of every pixel within the window is computed relative to the orientation of the sampled pixel. Combination: Compute the evidence score for optimal matches by combining three scores: the matching score based on local descriptors, the matching score based on ShapeContext descriptors, and the matching score based on the RANSAC model. In addition, to deal with the intra-variation of each writer, these scores are normalized by using Z-Score computed from the genuine set of each writer. In the second phase, we employ the framework FoCal [10] to calibrate the evidence score.

D. xyzmo

The tool is based on a signature verification algorithm using statistical respectively empirical models and calculates the evidence score by comparing reference signatures and the questioned signature taking into account only features of the actual signatures without prior knowledge and does not require any training steps in advance as it is the case in other approaches. Mainly a biometric comparer is spanning a mathematical multi dimensional room (the tolerance border) built from extracted dynamic features of the reference signatures and evaluates the distance of the questioned signature to this room by correlation methods which is than expressed and formulated into a score with a range from 0 to 1 expressing the similarity, e.g., 1 means highest similarity possible. Input parameters into the algorithm are the native signature data because extraction and comparison steps will be done internally in the comparison component when signatures are loaded for being compared. Usually the underlying algorithm supports an extra enrollment step respectively checks which cannot be applied in the given test scenario. All signatures used as reference signatures are in the evaluated systems to fulfill defined minimum consistency criteria and a signature will be refused to go into a profile (the set of reference signatures) in case it fails to do so. In the test scenario the reference signatures will be enforced from outside and preselecting the reference set may not be allowed.

E. Qatar University and Northumbria University

The proposed method uses edge-based directional probability distribution features [11] and grapheme features [12]. These methods have previously been applied for Arabic writer identification and have shown interesting results. The classification step is performed using a logistic regression classifier trained separately on each dataset (Chinese and Dutch). The online tool combines the most discriminant features described in Nalwa's method [13] also trained separately on each dataset using a logistic regression classifier. All the tools use the proposed z-calibration method.

F. German Research Center for Artificial Intelligence

The system is based on the methods introduced in [14] However, we have modified/optimized it in order to fit in the scenarios presented in this signature verification competition. First, the signature image is spatially smoothed followed by a binarization. In the optimized version of this approach we used various combinations of local and global binarization methods and evaluated the results. After these preprocessing steps the operations of [14] have been performed.

Table IV
RESULTS FOR THE CHINESE OFFLINE COMPETITION

ID	Accuracy(%)	FRR	FAR	$\widehat{C}_{llr}$	$\widehat{C}_{llr}^{min}$
1	80.04	21.01	19.62	0.757712	0.693347
2	73.10	27.50	26.70	3.062735	0.765021
3	72.90	27.50	26.98	1.125224	0.789918
6	56.06	45.00	43.60	1.260461	0.890711
7	51.95	50.00	47.41	3.222468	0.951274
8	62.01	37.50	38.15	1.573580	0.926571
9	61.81	38.33	38.15	6.227011	0.918450

Table V
RESULTS FOR THE DUTCH OFFLINE COMPETITION

ID	Accuracy(%)	FRR	FAR	$\widehat{C}_{llr}$	$\widehat{C}_{llr}^{min}$
1	82.91	17.93	16.41	0.730387	0.573175
2	77.99	22.22	21.75	2.456203	0.674031
3	87.80	12.35	12.05	0.415796	0.386128
6	95.57	4.48	4.38	0.714976	0.133917
7	97.67	2.47	2.19	0.900352	0.075223
8	75.84	23.77	24.57	1.664745	0.722033
9	71.02	29.17	28.79	4.133458	0.794021

We used means and variances for thresholds' computations. Next, the nearest neighbor approach is applied to decide on the result of each feature vector and finally a voting based classification is made. In the optimized version different voting strategies were applied that improved the overall performance.

G. Anonymous-2

This institution did not provide us with any details.

V. EXPERIMENTS AND EVALUATION

The systems have been evaluated on the four testing sets described above, i.e., the offline and online Chinese and Dutch data set. The task was to determine if a given questioned signature has been written by the author of the n reference signatures or of it was forged by another writer. In all experiments the number of reference signatures was $n = 12$, i.e., twelve known reference signatures were presented to the systems.

We evaluated our systems according to several measurements. First, we generated ROC-curves to see at which point the equal error rate is reached, i.e., the point were the false acceptance rate (FAR) equals the false rejection rate (FRR). At this specific point we also measured the accuracy, i.e., the percentage of correct decisions with respect to all questioned signatures. Next, we measured the cost of the log-likelihood ratios $\widehat{C}_{llr}$ (see [10]) using the FoCal toolkit. Finally, the minimal possible value of $\widehat{C}_{llr}$, i.e., $\widehat{C}_{llr}^{min}$ as a final assessment value. Note that a smaller value of $\widehat{C}_{llr}^{min}$ denotes a better performance of the method.

The results of the offline competitions appear in Tables IV and V. Those of the online competitions appear in Tables VI

Table VI
RESULTS FOR THE CHINESE ONLINE COMPETITION

ID	Accuracy(%)	FRR	FAR	$\widehat{C}_{llr}$	$\widehat{C}_{llr}^{min}$
1	84.81	12.00	16.05	0.565146	0.351142
4	93.17	6.40	6.94	0.413413	0.217915
5	93.17	6.40	6.94	0.418631	0.217915
6	82.94	16.80	17.14	1.049099	0.503151
7	85.32	13.60	14.97	0.905516	0.461140
9	80.89	9.26	8.14	6.210251	0.733883

Table VII
RESULTS FOR THE DUTCH ONLINE COMPETITION

ID	Accuracy(%)	FRR	FAR	$\widehat{C}_{llr}$	$\widehat{C}_{llr}^{min}$
1	93.49	7.56	7.69	0.492844	0.237550
4	96.27	3.70	3.76	0.258932	0.122596
5	96.35	3.86	3.44	0.351189	0.122596
6	91.82	8.33	8.02	0.534542	0.290940
7	92.93	7.25	6.87	0.604641	0.241201
9	88.56	11.11	11.27	6.433622	0.408429

and VII. As can be seen, different systems performed best on different tasks. Interestingly, the system with the best FRR and FAR always turned out to have the best value of $\widehat{C}_{llr}^{min}$. The winners for the offline competitions are System 1 for Chinese data and System 7 for Dutch data. The winner for both online competitions is System 4.

Several interesting observations can be made when having a closer look at the tables. First, it is interesting, that a good EER does not always result in a good $\widehat{C}_{llr}^{min}$, e.g., System 9 performs quite well on online Chinese data when looking at the EER but has the worst $\widehat{C}_{llr}^{min}$. This might be explained by the fact that a few large errors might spoil the overall performance with $\widehat{C}_{llr}^{min}$. Second, surprisingly System 7 performs better on Chinese online data than System 6, even if System 6 has been optimized to Chinese data. Finally, the results on Chinese data are much worse than those on Dutch data. This indicates that a lot of research has to be performed on Chinese scripts and maybe that this data is more challenging.

ACKNOWLEDGEMENTS

We would like to thank Falco Schneider for performing parts of the evaluation.

REFERENCES

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

[2] F. Leclerc and R. Plamondon, "Automatic signature verification: the state of the art 1989–1993," in *Progress in Automatic Signature Verification*, R. Plamondon, Ed. World Scientific Publ. Co., 1994, pp. 13–19.

[3] R. Plamondon and S. N. Srihari, "On-line and off-line handwriting recognition: a comprehensive survey," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 22, no. 1, pp. 63–84, 2000.

[4] V. L. Blankers, C. E. van den Heuvel, K. Y. Franke, and L. G. Vuurpijl, "Icdar 2009 signature verification competition," 2009, pp. 1403–1407.

[5] M. Liwicki, C. E. van den Heuvel, B. Found, and M. I. Malik, "Forensic signature verification competition 4nsigcomp2010 - detection of simulated and disguised signatures," in *12th International Conference on Frontiers in Handwriting Recognition*, 2010, pp. 715–720.

[6] J. Gonzalez-Rodriguez, J. Fierrez-Aguilar, D. Ramos-Castro, and J. Ortega-Garcia, "Bayesian analysis of fingerprint, face and signature evidences with automatic biometric systems," *Forensic Science International*, vol. 155, no. 2-3, pp. 126–140, 2005.

[7] S. Belongie, J. Malik, and J. Puzicha, "Shape matching and object recognition using shape contexts," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, pp. 509–522, 2002.

[8] R. Jonker and A. Volgenant, "A shortest augmenting path algorithm for dense and sparse linear assignment problems," *Computing*, vol. 38, pp. 325–340, 1987.

[9] M. A. Fischler and R. C. Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," *Commun. ACM*, vol. 24, pp. 381–395, 1981.

[10] N. Brümmer and J. du Preez, "Application-independent evaluation of speaker detection," *Computer Speech & Language*, vol. 20, no. 2–3, pp. 230–275, 2006.

[11] S. Al-Ma'adeed, E. Mohammed, and D. Al Kassis, "Writer identification using edge-based directional probability distribution features for arabic words," in *IEEE/ACS International Conference on Computer Systems and Applications*, 2008, pp. 582–590.

[12] S. Al-Ma'adeed, A.-A. Al-Kurbi, A. Al-Muslih, R. Al-Qahtani, and H. Al Kubisi, "Writer identification of arabic handwriting documents using grapheme features," in *IEEE/ACS International Conference on Computer Systems and Applications*, 2008, pp. 923–924.

[13] V. Nalwa, "Automatic on-line signature verification," *kluwer International series in engineering and computer science*, pp. 143–164, 1999.

[14] P. I. S. D. D. Samuel, "Novel feature extraction technique for off-line signature verification system," *International Journal of Engineering Science and Technology*, vol. 2, pp. 3137–3143, 2010.