Man vs. Machine: A Comparative Analysis for Forensic Signature Veri cation

Conference Paper · January 2013

CITATIONS
4

READS
266

4 authors:

Muhammad Imran Malik
Deutsches Forschungszentrum für Künstlic…
29 PUBLICATIONS **126** CITATIONS
SEE PROFILE

Marcus Liwicki
Deutsches Forschungszentrum für Künstlic…
189 PUBLICATIONS **1,322** CITATIONS
SEE PROFILE

Andreas Dengel
Deutsches Forschungszentrum für Künstlic…
467 PUBLICATIONS **2,420** CITATIONS
SEE PROFILE

Bryan Found
Victoria Police Forensic Services Department
33 PUBLICATIONS **243** CITATIONS
SEE PROFILE

Man vs. Machine: A Comparative Analysis for Forensic Signature Verification

Muhammad Imran MALIK [a] , Marcus LIWICKI [a] , Andreas DENGEL [a] , and Bryan FOUND [b]

[a] *German Research Center for Artificial Intelligence (DFKI)*
Trippstadter Str. 122, 67663 Kaiserslautern, Germany
firstname.lastname@dfki.de

[b] *Victoria Police Forensic Services Department*
31 Forensic Drive, Macleod, Victoria Australia
bryan.found@police.vic.gov.au

Abstract. Traditionally, human experts authenticated/verified the authorship of signatures. With the emergence of modern computing technologies, there is a push to enable machines/computers deciding on this problem. Today we have both machines and specifically trained human experts to check the authenticity of signatures. The question here is; which is better: man or machine? The answer can be quite subjective; we however, in this paper focus on empirically comparing the performance of the two on the same/similar basis. The novelty of this work is that we applied various state-of-the-art signature verification systems; gathered results and provided the same data to Forensic Handwriting Examiners (FHEs) and finally performed a comparative analysis of the two on the basis of accuracy and error rate. This is an ongoing research and the current paper reports on various interesting results we obtained during our experiments.

1. Introduction

Signatures are considered as a seal of authenticity in our everyday life. There has always been a demand to authenticate this seal of authenticity. Today this authentication/verification is done by humans as well as machines. The aim of this paper is to compare the performance of the two with respect to each other. Note that we do not see machines/automated systems as a replacement of humans; rather we perform this comparison to highlight the potential of machines to assist human experts in verifying signatures. We also note that there are certain limitations with machines, e.g., available training data, contrary to human experts who carry their previous experience of judging signatures from case to case along with case specific data. This renders machines as a good assistant rather than a replacement of human experts. Having said that, we also state that there are some areas, e.g., banking, where sometimes machines are seen as a replacement of human signature verifiers but we exclude those areas from our study and confine it to forensic handwriting analysis. This is required since today Forensic Handwriting Examiners (FHEs) make a very limited use of automated tools e.g., CEDAR-FOX (Srihari & al., 2003), FISH (M. Philipp, 2003), WANDA project framework (K. Franke, 2004) due to various limitations and it is required to present some study where we show FHEs the potential of the state-of-the-art Pattern Recognition (PR) methods so that they can include them in their routine casework.

The rest of the paper is organized as follows. In section 2 we will define signature verification from the PR and FHEs perspectives. Section 3 describes the data used for this study. Section 4 provides a brief description about the state-of-the-art automated systems we used in this study. Section 5 details our experiments where we provided the automated systems with data for verification. Section 6 describes how we provided the similar data to FHEs and conducted the so-called proficiency tests with human experts and finally performed the man vs. machine comparison. Section 7 concludes this paper and provides some insights for future work.

2. Signature Verification

Today the PR community moves by defining automatic signature verification as a two-class pattern classification problem (D. Impedovo, 2008). Note that in earlier PR studies it was defined differently where PR researchers also considered other genres of signatures such as, disguised signatures (R. Plamondon, 1989). As a two class classifier, an automated system has to decide whether or not a given signature belongs to a referenced authentic author. If a system could find enough evidence of genuine authorship from the questioned signature's feature vector, it considers the signature as genuine; otherwise it declares the signature as forged. Contrary to this, FHEs take signature verification as a multi-class (at least three classes) classification problem (M. I. Malik, 2012). Along with genuine and forged signatures, they also look into the possibility of disguised signatures. We, therefore, for this study also made automated systems to detect disguised signatures and both the systems as well as human experts had to classify the given signatures in one of the following classes or to the class inconclusive: (when they were unable to say anything about potential authorship due to lack of evidence/information).

Table 1. Year-wise data breakup

Year	Reference	Disguised	Forged	Genuine	Total
2001	20	47	160	43	270
2002	9	20	104	76	209
2004	16	8	42	50	116
2005	15	9	71	20	115
2006	25	7	90	3	125
Overall	85	91	467	192	835

Table 2. Summary results of automated systems: (a) when applied on La Trobe 2001, 2004, and 2005 data collectively. (b): when applied on La Trobe 2006 data. *: Results without disguised signatures in the dataset.

(a)

System	Accuracy	FAR	FRR	EER	EER*
1	85.11	14.29	15.82	15.82	14.16
2	77.88	21.61	23.16	23.16	16.81
3	78.89	20.88	21.47	21.47	13.19
4	30.67	73.63	62.71	70.24	68.14
5	71.11	28.94	28.81	28.81	20.51

(b)

System	Accuracy	FAR	FRR	EER	EER*
1	90.0	1.1	90	80	34
2	54.0	41.1	90	58	41
3	75.0	20.0	70	70	8
4	92.0	0.0	80	70	0
5	80.0	13.3	80	55	28
6	20.0	87.0	10	60	21
7	91.0	1.1	80	70	8

- Genuine signatures: written by an authentic reference author.
- Forged signatures: written by some other person than the authentic reference author where that person has tried to imitate the genuine signatures of the authentic reference author.
- Disguised signatures: written by the authentic reference author where (s) he has deliberately tried to make the signatures look like a forgery, i.e., imitated a forgery. This is usually done by authentic authors with the purpose of denying their signatures at a later date, e.g., on bank checks, false wills etc.

3. Data

For the purpose of this study we used the La Trobe signature data collected under the supervision of Bryan Found and Doug Rogers in the years 2001, 2002, 2004, 2005, and 2006, respectively (More information can be found in (C. Bird, 2007)). The images were scanned at 600 dpi resolution and cropped at the Netherlands Forensic Institute for the purpose of this study. A detailed breakdown of the data used in our study from the year wise La Trobe data collection is given in Table 1.

4. State-of-the-Art Automated Signature Verification Systems Used

For analyzing the results of various state-of-the-art signature verification systems, we organized two signature verification competitions. These are the 4NSigComp2010 and 4NSigComp2012 organized with the 12th and 13th International Conferences on Frontiers in Handwriting Recognition (ICFHR).

In the 4NSigComp2010 competition, seven systems were submitted. The first system used a fusion system of local analysis (A. Gilperez & al., 2008) and allographic analysis (M. Bulacu & al., 2007); the second system computed a DTW similarity measure; the third and the seventh systems were both from the same participant where they applied logistic regression on a selected set of global features with full and partial training, respectively; the fourth system was a commercial classifier; the fifth system applied SVMs on zone-features; the sixth system decided to stay anonymous. More details about these systems are provided in (M. Liwicki & al., 2010).

In the 4NSigComp2012 competition, five systems were submitted. The first system employed the Gaussian grid feature extraction technique by taking signature contours as input and used Support Vector Machines (SVM) for classification (V. Nguyen & al., 2011); the second system combined through logistic regression a large number of geometrical features like number of holes, moments, projections, distributions, position of barycenter, number of branches in the skeleton, Fourier descriptors, tortuosities, directions, curvatures and chain codes etc. (A. Hassaine & al., 2011); the third system used Histogram of Oriented Gradient (HOG) and Local Binary Patterns (LBP) (M. Yilmaz & al., 2011); the fourth system implied Gaussian Mixture Models (GMM) and the fifth system used various global features like cell size, centroid, angle of inclination, number of holes etc. More details about these systems are provided in (M. Liwicki & al., 2012).

5. Experiments with Automated Systems

For the participants of 4NSigComp2010, we provided the signature data from year 2002 as training data and used the signature data from year 2006 for evaluation. For the 4NSigComp2012 participants, we provided the training and evaluation set (i.e., complete data from the 4NSigComp2010) as training set

Table 3. Results of FHEs' proficiency tests from different years. G: Genuine, D: Disguised, and F: Forged signatures.

(a) Year 2001 data.

Result	G	D	F	Total
Correct	1628	571	2840	5039
Errors	30	461	265	756
Inconclusive	105	895	3455	4455
Net Opinions	1763	1927	6560	**10250**

(b) Year 2004 data.

Result	G	D	F	Total
Correct	990	69	343	1402
Errors	1	13	9	23
Inconclusive	9	78	488	575
Net Opinions	1000	160	840	**2000**

(c) Year 2005 data.

Result	G	D	F	Total
Correct	587	73	1263	1923
Errors	1	52	174	227
Inconclusive	32	154	764	950
Net Opinions	620	279	2201	**3100**

(d) Year 2006 data.

Result	G	D	F	Total
Correct	93	10	1151	1254
Errors	0	111	113	224
Inconclusive	0	96	1526	1622
Net Opinions	93	217	2790	**3100**

and used the data from years, 2001, 2004, and 2005 for evaluation. All the participating systems had to classify signatures as genuine, forged or disguised or whether they were unable to classify. We note that the automated systems we used nearly in all the cases came out with a conclusion).

Table 2(a) shows the results when we applied the automated systems (from the 4NSigComp2012 competition) for the years 2001, 2004, and 2005 data, collectively. Table 2(b) shows the results when we applied the automated systems (from the 4NSigComp2010 competition) for the year 2006 data. Note that, in both the cases, we also report the results when we removed disguised signatures from the evaluation set and repeated the experiments. It was done since sometimes FHEs also face problems in disguise detection in their real casework (J. sita, 2002).

Furthermore, we applied various evaluation metrics, such as likelihood ratios, cost of log likelihood ratios etc., on the automated systems. We, however here, report the results in terms of False Acceptance Rate(FAR), False Rejection Rate (FRR), Equal Error Rate (EER) and accuracy. It is done to later compare the performance of the automated systems with that of human experts as we can only use accuracy and/or error rate for man vs. machine comparison since in case of humans there is no threshold that can be balanced (M. Liwicki & al., 2011).

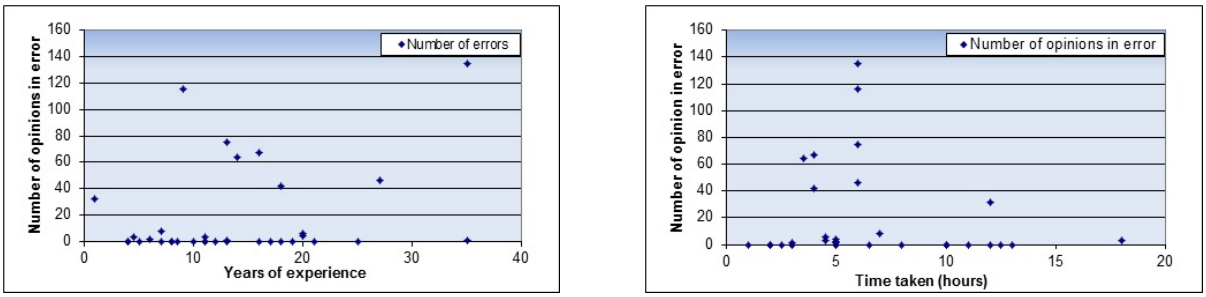
6. Comparison and Results

The evaluation of FHEs opinions has been carried out 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 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/forgeries. FHEs are asked to express their opinion on authenticity on a five-point scale (C. Bird, 2007). 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 (scanned signature images to machines and signature photocopies to human experts).

For evaluating human experts/FHEs on the La Trobe data collection of year 2001, in total, 51 answer booklets were submitted, thereof 10 peer reviewed responses (cross-checked by a second FHE), 31 individual responses (not peer-reviewed), and 10 experimental responses (from trainees). A total of 10250 authorship opinions were expressed by the group. Of these opinions 5039 (49.2%) were correct, 756 (7.4%) were misleading and 4455 (43.5%) were inconclusive. This translates into an error rate of 13.0% on the decisions (Accuracy of 87.0%). %) by disregarding the cases which were inconclusive. Detailed breakdown of these results is given in Table 3(a).

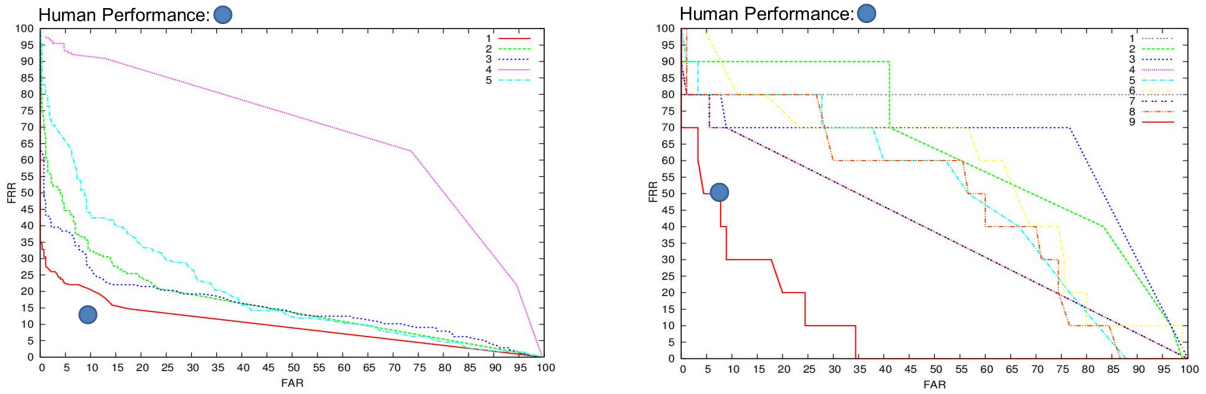
For the year 2004 La Trobe signature data, in total, 21 answer booklets were submitted, thereof 7 peer reviewed responses (cross-checked by a second FHE), and 14 individual responses (not peer-reviewed). A total of 2000 authorship opinions were expressed by the group. Of these opinions 1402 (70.1%) were correct, 23 (1.2%) were misleading and 575 (28.8%) were inconclusive. This translates into an error rate of 1.6% on the decisions (Accuracy of 98.4%). %) by disregarding the cases which were inconclusive. Detailed breakdown of these results is given in Table 3(b).

For the year 2005 La Trobe signature data, in total, 31 answer booklets were submitted, thereof 5 peer reviewed responses (cross-checked by a second FHE), and 26 individual responses. A total of 3100 authorship opinions were expressed by the group. Of these opinions 1923 (62.0%) were correct, 227 (7.3%) were misleading and 950 (30.6%) were inconclusive. This translates into an error rate of 10.6% on the decisions (Accuracy of 89.4%). %) by disregarding the cases which were inconclusive. Detailed breakdown of these results is given in Table 3(c).



(a) (b)

Figure 1. Analysis of FHEs performance on La Trobe 2001 data. (a): Relationship between examiner experience and total number of opinion errors. (b): Relationship between the time examiners took to complete the trial and total number of opinion errors.



(a) (b)

Figure 2. Man vs. machine comparison in ROC space. (a): On combined 2001, 2004, and 2005 data. Systems 1-5: participants of the 4NSigComp2012 competition. (b): On 2006 data. Systems 1-7 are participants of the 4NSigComp2010 competition while system 8 and 9 are added later on. Details about these systems are provided in (M. I. Malik & al., 2011).

For the year 2006 La Trobe signature data, 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%) by disregarding the cases which were inconclusive. Detailed breakdown of these results is given in Table 3(d).

Note that several other tests were performed to analyze the performance of FHEs, since FHEs usually exhibit a much wider range of performance with respect to automatic systems. Due to space limitations, we are only presenting the two cases of FHEs performance analysis where we examined the relationship between examiners experience and the total number of opinion errors (see Figure 1a), and the relationship between the time examiners took to complete the trials and the total number of opinion errors (see Figure 1b). We present these results for the 2001 data. Detailed results will be provided in an extended journal article.

Both for Figure 1a and Figure 1b, no simple correlation was found to exist between the two variables (at x and y axis). There is, therefore on these data, no support for the notion that the validity of a trained examiners opinion can be referenced by the number of years the examiner has been practicing and also no support for the notion that the validity of a trained examiners opinion can be referenced by the amount of time the examiner spent performing the task.

Finally, we used accuracy and error rate for comparing the performances of humans/FHEs and machines/systems. Table 4 provides these results on the basis of accuracy. We report both the average as well as best man and machine accuracies. Later, we used the error rates produced by humans/FHEs to place their performance mark on the FRR/FAR (ROC) space so that to have a graphical representation of human performance against machines. This is given in Figure 2a (combined data from 2001, 2004, and 2005) and Figure 2b (data from 2006).

Table 4. Detailed man vs. machine results for the data collections from various years.

Data from the year	Accuracy (%)			
	Avg. human	Avg. machine	Best human	Best machine
2001	44.8	70.8	100	93.6
2004	66.2	70.4	97	87
2005	62	59.8	100	68
2005	38.8	71.7	91	92

7. Conclusions and Future Work

In this paper we have provided the results of a detailed analysis we performed in order to compare the performance of human experts/FHEs against automated systems with respect to signature verification. It is required since even today there are areas, e.g., forensic signature analysis, where automated signature authentication/verification systems find very limited applicability. This paper shows the power of automated systems to assist human experts in assessing the authorship of signatures. It is shown that the performance of automated systems remained on par with that of human experts in many cases. Noteworthy is the fact that different automated systems, just like humans, were better on different data. Both the human experts as well as automated systems/machines encountered difficulties in correctly classifying disguised signatures. This is probably because of limited disguised training data availability. However, the results provided are encouraging and we hope that automated systems will become better with time and with access to more forensically relevant data especially involving disguised behaviors.

In future we plan to perform analyses on data with much more reference writers and skilled forgers. It is planned to organize more competitions, as well as workshops on the particular topic of automated forensic handwriting analysis.

Acknowledgment

We would like to thank Elisa van den Heuvel and Linda Alewijnse for providing us the data for this study.

References

S. N. Srihari, B. Zhang, C. Tomai, S. Lee, Z. Shi, Y. C. Shin, A system for handwriting matching and recognition, in: Symp. on Document Image Understanding Technology, 2003, pp. 67–75.

M. Philipp, Fakten zu FISH, das forensische informations-system handschriften des bundeskriminalamtes — eine analyse nach ber 5 jahren wirkbetrieb, Tech. rep., Bundeskriminalamt, Germany, in German (1996).

K. Franke, L. Schomaker, C. Veenhuis, L. Vuurpijl, I. Erp, M. van Guyon, WANDA: A common ground for forensic handwriting examination and writer identification, ENFHEX News (2004) 23–47.

D. Impedovo, G. Pirlo, Automatic signature verification: The state of the art, IEEE Transactions on Systems, Man, and Cybernetics 38 (5) (2008) 609–635.

R. Plamondon, G. Lorette, Automatic signature verification and writer identification – the state of the art, in: Pattern Recognition, Vol. 22, 1989, pp. 107–131.

M. I. Malik, M. Liwicki, From Terminology to Evaluation: Performance Assessment of Automatic Signature Verification Systems, in: ICFHR, 2012, pp. 609–614.

J. Sita, B. Found, D. Rogers, Forensic handwriting examiners' expertise for signature comparison, Journal of Forensic Sciences 47 (2002) 1117–1124.

C. Bird, B. Found, and D. K. Rogers, Forensic Handwriting Examiners' Skill in Detecting Disguise Behavior from Handwritten Text Samples , Vol. 22, 2012.

A. Gilperez, F. Alonso-Fernandez, S. Pecharroman, J. Fierrez, and J. Ortega-Garcia, Off-line Signature Verification using Contour Features, in: ICFHR, 2008.

M. Bulacu, L. Schomaker, Text-independent writer identification and verification using textural and allographic features, IEEE TPAMI 29 (4) (2007) 701–717.

M. Liwicki, C. E. van den Heuvel, B. Found, M. I. Malik, Forensic signature verification competition 4nsigcomp2010 - detection of simulated and disguised signatures, in: ICFHR, 2010, pp. 715–720.

V. Nguyen, M. Blumenstein, , An application of the 2d Gaussian filter for enhancing feature extraction in off-line signature verification, in: ICDAR, 2011, pp. 339–343.

V. Nguyen, M. Blumenstein, , The icdar2011 Arabic Writer Identification Contest, in: ICDAR, 2011, pp. 1470–1474.

M. B. Yilmaz, B. Yanikoglu, C. Tirkaz, and A. Kholmatov, Offline signature verification using classifier combination of HOG and LBP features, in: IJCB, 2011, pp. 1–7.

M. Liwicki, M. I. Malik, L. Alewijnse, C. E. van den Heuvel, B. Found, ICFHR2012 Competition on Automatic Forensic Signature Verification (4NsigComp 2012), in: ICFHR, 2012, pp. 819–824.

M. Liwicki, M. I. Malik, C. E. van den Heuvel, X. Chen, C. Berger, R. Stoel, M. Blumenstein, B. Found, Signature verification competition for online and offline skilled forgeries (SigComp2011), in: ICDAR, 2011, pp. 1480–1484.

M. I. Malik, M. Liwicki, A. Dengel, Evaluation of local and global features for offline signature verification, in: AFHA, Beijing, China, 2011, pp. 26–30.