Man vs. Machine: A Comparative Analysis for Signature Verification

Muhammad Imran Malik[1], Marcus Liwicki[1,2], Andreas Dengel[1], Bryan Found[3]

Abstract. *Signatures have been used as a means to authenticate documents for centuries. From the outset, the focus of forensic examinations was to both objectively and subjectively establish whether they were genuine (written by the specimen author) or simulated (written by an imposter/forger). With the emergence of new computing technologies, additional objective examination techniques designed to determine the authenticity of questioned signatures became available. Although the opinions of Forensic Handwriting Examiners (FHEs) remain the most popular method of signature authenticity determinations, computer based techniques are attracting increasing interest within the forensic community. The question here is; which is better: man or machine? To address this question we focus on empirically comparing the performance of the two, on the same or similar material. The novelty of this work is that we have applied various state-of-the-art signature verification systems to questioned signature problems which had already been worked by FHEs and then performed a comparative analysis of the two.*

Reference: Muhammad Imran Malik, Marcus Liwicki, Andreas Dengel, Bryan Found (2014). Man vs. Machine: A Comparative Analysis for Signature Verification. J. Forensic Document Examination, Vol. 24, pp. 21-35.

Keywords: Signature verification, Signature Comparison, Forensic handwriting analysis, Performance comparison,Evaluation, 4NSigComp2010, 4NSigComp2012.

1. Introduction

Signatures are used as a seal of authenticity in our everyday life. There has always been a demand to authenticate this seal, particularly in cases where a signature became disputed. A large body of literature has developed which describes the theories, methods and techniques used to examine and evaluate the authenticity of questioned signatures (e.g., Osborn, 1929; Harrison, 1958; Conway, 1959; Hilton, 1982; Ellen, 1989; Huber & Headrick, 1999). In the contemporary forensic environment, signature authentication/verification is still universally carried out by humans. But outside of forensic science, computer based signature verification techniques continue to develop in response to commercial needs around quickly recording and authenticating an individual's mark (Impedovo & Pirlo, 2008). The aim of this paper is to make some inroads with respect to comparing the performance of the traditional human approach to that used by contemporary objective techniques.

Note that we do not envisage objective automated systems as replacements for FHEs; we believe that there is great potential for these systems to assist human experts in signature analysis and interpretation in the future. We consider that there are significant limitations when using machines, since machines are generally trained on the case specific training data (containing the specimen signature samples alone from a said case). This training enables them learn

1. German Research Center for Artificial Intelligence (DFKI)
Trippstadter Str. 122
67663 Kaiserslautern, Germany
2. University of Fribourg, Switzerland
marcus.liwicki@unifr.ch
3. Victoria Police Forensic Services Department,
31 Forensic Drive, Victoria, Australia
bryan.found@police.vic.gov.au

the genuine signing behavior of the specimen author and by virtue of this training, they develop statistical and/or structural models for providing judgments about the questioned signatures. These models are highly influenced by the training data provided to them and therefore the representativeness of the training material is critical to ensure they model the relevant elements of the task that they are to carry out. Further, generally machines do not consider the lessons they learned while analyzing other cases (although there exist various techniques which can be applied to enable machines to utilize the writing behaviors they learned from various cases other than the specific case at hand (e.g., methods by Weber et al., 2009)). Human experts, also rely on the case specific data, but also heavily rely on previous knowledge of predictive features associated with genuine and forged behaviors and routinely apply this knowledge to the specific case at hand. The difference in philosophy between human and machine based examination strategies limits the machines to acting as a potentially good assistant, rather than a complete replacement of the FHE. Having said this, we also recognize that commercially, outside forensic casework, machines are used in preference to humans (e.g., the banking industry) primarily due to issues associated with the volume of authentications that are required to be carried out, and the timeliness associated with task.

It is interesting that with respect to signature authentication, FHEs have traditionally made very limited use of automated tools e.g., CEDAR-FOX (Srihari et al., 2003), FISH (Philipp, 1996), WANDA project framework (Franke et al., 2004). This may be because many of these tools have been designed to perform comparison tasks and present results in a form that FHEs are not comfortable using. It is also the case that sometimes automated tools (e.g., FISH) are not available to FHEs outside of the agencies or specific organizations. With this in mind, this study is designed to directly compare state-of-the-art Pattern Recognition (PR) automatic methods to FHEs' performance such that the potential for the incorporation of more objective comparative techniques into the routine tasks of FHEs might be entertained in the future.

Today the PR community considers automatic signature verification to be a two-class pattern classification problem (Impedovo & Pirlo, 2008). In earlier PR studies it was defined differently where PR researchers also considered other genres of signatures such as disguised signatures (Plamondon & Lorette, 1989). As a two class classifier, an automated system has to decide whether a given signature belongs to a referenced authentic author or not. If a system finds *enough* evidence of genuine authorship from the questioned signature, it considers the signature as

Year	Reference Signatures	Disguised Signatures	Forged Signatures	Genuine Signatures	Total Signatures
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

Table1. Year-wise data breakup

genuine; otherwise it declares the signature forged/ simulated. Clearly, this is not the case when FHEs approach signature comparison tasks.

FHEs consider signature verification as a multiclass (at least three class) classification problem (Malik & Liwicki, 2012). Along with genuine and forged signatures, they also consider the possibility that the observed combination of any similar and dissimilar features might result from disguise behavior, or might result from a myriad of other factors that could impact on the writing act and which may not be captured in the population of specimen material used in the comparison (for example illness, drug effects, writing surface effects, writer position etc.). For the purpose of this comparative study we only considered automated systems that could look into the possibility of questioned signatures being genuine, forged, or disguised. The PR systems, as well as human experts, were required to classify the given signatures in one of the three following classes, or to the class inconclusive (when they were unable to conclude anything about a signature's authenticity);

Genuine signatures: normal signatures written by the specimen writer.

Forged signatures: written by some person other than the specimen writer where that person has tried to imitate the genuine signature of the specimen writer. Note that FHEs prefer using the term "simulated" rather than "forged" as the later term implies intent. However, in this paper, we use the terms 'forged' and 'simulated' interchangeably since we know which of the signatures were intentionally simulated, and since in the PR community the term "forgery" is already in widespread usage.

Disguised signatures: written by the specimen writer where there has been a deliberate attempt to change the features of the signature for the purpose of later denial. Typically the strategy associated with this behavior is either to introduce gross changes to the form that can easily be referred to, or to make the signature appear to be a forgery by executing the signature in a way that introduces feature changes that the genuine writer believes would be present if the signature was forged by another person (also referred to as auto-simulation behavior).

2. Data and PR Systems

For the purpose of this study we used blind test data collected by a La Trobe program run over the years 2001, 2002, 2004, 2005, and 2006, respectively. Although the year by year data has not been published, the approach used and summary statistics have been presented (Found & Rogers, 2003; Found &Rogers, 2008). All the signatures were in the form of static images. The original signatures were scanned at 600dpi resolution and cropped at the Netherlands Forensic Institute for the purpose of this study. More information can be found in C. Bird et al., 2007 and C. Bird et al., 2009.

For the year 2001, one specimen writer wrote three normal signatures per day (written with a ball point pen) over a fifteen day period, six disguised signatures per day (written with a ball point pen) over a fifteen day period, and six normal signatures per day (written with a pencil) over a three day period. From the normal signatures pool ,the genuine questioned signatures and the reference signatures (the set formed to which the questioned signatures would be compared) were constructed. Two 'forgers' were selected from the academic staff at La Trobe University to forge the specimen writers' signatures. Each of the forgers was provided with six normal samples of the specimen writer's signature. The forgers were instructed that they could use any or all of the supplied specimen signatures as models for their forgeries. The forgers were also instructed that their forgeries must be unassisted (not tracings). Each forger was asked to complete the following task each day over a 10 day period.

25 practice signatures (ball point pen)
5 forgeries (ball point pen)
5 forgeries (pencil)

The forgeries, other than the practice attempts, were used as a pool from which the questioned forged signatures were selected. All the questioned samples were numbered randomly, scanned and inkjet or laser printed into a booklet. For the year 2001, the total selected corpus contained 270 signatures belonging to different signature categories as given in Table 1.

For the year 2002, one specimen writer wrote fifteen normal and six disguised signatures per day

over a seven day period. In addition to these signatures, the specimen writer provided an 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. For forging the signatures of the specimen writer, 27 'forgers' were selected from 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 specimen writer. Forgers were instructed that they could use any or the entire supplied reference signatures as models for their forgeries. Forgers were also instructed that their forgeries must be unassisted (not tracings). Each forger was asked to complete the following tasks.

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

The total selected corpus, for the year 2002, contained 209 signatures belonging to the different signature categories as given in Table 1.

For the year 2004, one specimen writer wrote the normal and disguised signatures over a ten days period. From the normal signature pool the genuine and reference signatures were drawn for the specimen writer. Thirty one adult 'forgers' were used to generate the forgery pool. These individuals were volunteers drawn from a single private company. Each of the forgers was provided with three genuine samples of the signatures written by the specimen writer. These forgers were instructed similarly to the forgers from years 2001 and 2002. The total selected corpus, for the year 2004, contained 116 signatures belonging to the different signature categories as given in Table 1.

For the year 2005, one specimen writer wrote the normal and disguised signatures over a ten days period. Six adult volunteer 'forgers' were used to generate the forgery pool. Each of the forgers was provided with 3 original normal samples of the genuine signature written by the specimen writer. These forgers were instructed similarly to the forgers from previous years. The total selected corpus, for the year 2005, contained 115 signatures belonging to the different signature categories as given in Table 1.

For the year 2006, one specimen writer wrote the normal genuine and disguised signatures over a five day period. Seven disguised signatures and 25 normal genuine signatures were chosen from this subset. Thirty-four adult volunteer 'forgers' contributed to the forgery set. The forgers were either 'lay' persons or calligraphers. Similar instructions were given to the forgers as given in the previous years. The total selected corpus, for the year 2006, contained 125 signatures belonging to the different signature categories as given in Table 1.

The results of the FHEs analysis of the data summarized in Table 1 were known. To determine how well PR approaches performed compared to FHEs, two offline (static data only) signature verification competitions were organized. These were titled the '4NSigComp2010' and the '4NSigComp2012'. These two competitions were managed though the 12th and 13th International Conferences on Frontiers in Handwriting Recognition (ICFHR). For the purpose of brevity, we provide here only a framework of the techniques that were developed and used by the participants. Further details about the methods used here can be sourced from the references provided.

In the 4NSigComp2010 competition, seven different systems were submitted and applied to the La Trobe signature test data. The first system used a fusion system of signature local analysis (Gilperez et al., 2008) and allograph feature analysis (Bulacu & Schomaker, 2007); the second system computed a dynamic time warping similarity measure on signature projections obtained by Mojette transform (Guedon & Nicolas, 1997); the third system used logistic regression based classification applied on signatures' geometric features(Hassaine et al., 2011); the fourth system was a commercial classifier and the participant chose not to provide any details about the applied approach; the fifth system applied support vector machines on zone-features(Yilmaz et al., 2011); the participant submitting the sixth system chose not to provide information as to the approach used, and the seventh system used logistic regression in conjunction with signature's connected components, moments, number of branches in the signature skeleton, directions and

curvatures, etc. (Hassaine et al., 2011). In addition to these seven systems, we added two further systems to our experiments specifically for this paper (they were not part of the 4NSigComp2010 competition). Our system 8 used various mixtures of global and grid based features including grid cell size, center of gravity, respective angles on different axes, etc.(Malik et al., 2011) and our system 9 used Gaussian mixture models for classification while utilizing various local features extracted through a sliding window approach (Liwicki & Malik,2011).

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 for classification (Nguyen & Blumenstein, 2011); the second system combined, through logistic regression, a large number of geometrical features (number of holes, signature projections, distributions, position of barycenter, number of branches in the skeleton, Fourier descriptors, tortuosities, directions, curvatures and chain codes etc.) (Hassaine et al., 2011); the third system used histograms of oriented gradient and local binary patterns (Yilmaz et al., 2011); the fourth system applied Gaussian mixture models and the fifth system used various global features like cell size, centroid, angle of inclination, gradients etc. More details about these systems can be found in (M. Liwicki et al., 2012).

3. Results of PR Systems

The La Trobe signature images from the year 2002 were provided to the participants of the 4NSigComp2010 competition for use as training data. Signature images from the year 2006 La Trobe trial were used for evaluation. In the 4NSigComp2012competition, the training and evaluation set(i.e., complete data from the 4NSigComp2010, i.e., the years 2002 and 2006 data) were provided as the training set to the participants, and the La Trobe data from the years, 2001, 2004, and 2005 were used for evaluation. We report here on the performance of the participating automated systems in these competitions on a year-wise basis. We report the evaluation results on the data from years 2001, 2004, 2005, and 2006 as the data from year 2002 were only used for training and were not included in any evaluation set.

All the participating systems had to classify signatures as genuine, forged, disguised, or whether they were unable to classify (this option was given to the systems as it is always present within the FHE paradigm). It is interesting to note, however, that the automated systems, in spite of being provided with an inconclusive option, generally provided a classification. Although the participants never explicitly stated the reasons for this, we consider that a majority of automatic systems were initially developed for applications other than forensic. In such applications, e.g., in banking, automatic systems usually only accept or reject a signature. If rejected, authentication may further be carried out by bank staff where they can consider other proofs as well, e.g., passwords/ids, to allow the signer complete the transaction--in fact without the actual successful verification of signatures. Having said that, we envisage that in the future, as more automatic systems are developed for forensic applications, the option to report inconclusive opinions might also be considered.

Table 2 shows the results when we evaluated the automated systems for the years 2001, 2004, and 2005 data (in the 4NSigComp2012 competition) and Table 3 shows the results when we evaluated the automated systems for the year 2006 data (in the 4NSigComp2010 competition). In both the tables, 2 and 3, we also report the results when we removed disguised signatures from the evaluation set and repeated the experiments. This was done to analyze the effect of presence of disguised signatures on the performance of automatic systems. Given in the tables, 2 and 3, accuracy represents the percentage of correctly classified signatures with respect to all the questioned signatures. The False Rejection Rate (FRR), also named Type I error or miss probability, occurs when a genuine signature is rejected by the system as being forged. The False Acceptance Rate (FAR), also known as Type II error or false alarm probability, occurs when a simulated signature is accepted as being genuine. The Equal error Rate (EER) is the point/value at which FRR equals FAR. Note that, along with accuracy, we also report the FRR and FAR in order to uncover the actual performance of the systems. In fact, accuracy is insufficient in representing the actual performance of a system (and to that effect also humans) when there are unequal number of different types of signatures, i.e.,

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

Table 2: Results of automated systems on the years 2001, 2004, and 2005 data.
FAR: False Acceptance Rate, FRR: False Rejection Rate, EER: Equal Error Rate.
*: When disguised signatures were not included in the test set.

System	Accuracy (%)	FAR (%)	FRR (%)	EER (%)	EER*(%)
1	90	1.1	90	80	34
2	54	41.1	90	58	41
3	75	20	70	70	8
4	92	0	80	70	0
5	80	13.3	80	55	28
6	20	87	10	60	21
7	91	1.1	80	70	8
8	80	20	78	56	33
9	80	20	20	20	33

Table 3: Results of automated systems on the year 2006 data.
FAR: False Acceptance Rate, FRR: False Rejection Rate, EER: Equal Error Rate.
*: When disguised signatures were not included in the test set.

genuine, forged, and disguised, among the questioned data. For example, we had 90 forged, 3 genuine, and 7 disguised signatures among the questioned signatures in year 2006 data (Table 2). Now if an automatic system or even a human expert blindly and falsely declared all questioned signatures as forged, it would seem 90% accurate which is actually not correct. Therefore, we must also consider the FRR and FAR to further characterize an automated system's performance. For the same reason we also plot the human performance in FRR/FAR space (see Figures 4 and 5). Furthermore, the EER (the point where FRR equals FAR) is also important as a single objective measure to rank any systems' performances when tested on the same data. This EER is not directly correlated with the accuracy and systems with varying accuracies can have the same EER, as shown in Table 3. We can also measure the system performance by putting different weight penalties when a system makes errors in identifying different types of signatures, i.e., genuine, forged,

Type	Genuine	Disguised	Forged	Total
Correctly Classified	1628	571	2840	5039
Misleading (Errors)	30	461	265	756
Reported Inconclusive	105	895	3455	4455
Total	1763	1927	6560	10250

Table 4: Results of the blind trial conducted with FHEs on the data from year 2001.

or disguised. This allows us to mold the FRR/FAR metric with respect to our preferences as whether we consider a misclassified forgery a greater error or a misclassified genuine signature a greater error and vice versa. We however in our experiments, treated all the misclassifications/errors equally by giving them a penalty weight of 1. For further details and background issues about these metrics, please refer to Fawcett, 2006.

As given in Tables 2 and 3, the accuracy and error rates varied among systems. In general, the systems faced difficulties in classifying disguised signatures (considered most of the disguised signatures as forgeries) and nearly in all the cases (except for system 9 of in Table 2, reasons for this in Liwicki & Malik, 2011) the systems performance increased when we removed the disguised signature samples from the questioned signatures. In Table 3, system 4 reached an error rate of 0 % when disguised signatures were not considered in evaluation. In fact, this system was 100% correct in classifying forgeries and genuine signatures, but misclassified all the disguised signatures as forged.

Furthermore, we applied various evaluation metrics, such as likelihood ratios, and cost of log likelihood ratios etc., on the automated systems. Here we only report the results in terms of error rates and accuracy. This is done to later compare the performance of the automated systems with that of FHEs. Note that we can evaluate the results of an automatic system by varying the numerical thresholds according to which an automatic system objectively classifies a signature as genuine, forged, or disguised. In contrast, there are no numerical thresholds for humans with respect to their opinions regarding the category of a signature, i.e., genuine, forged, or disguised. For example, a human expert cannot have a numerical objective threshold below which s(he) can consider a signature as forged and above which she can consider the same signature as genuine (and vice versa) and keep varying that objective threshold to give opinions about the signatures making certain signatures fall into one category on one threshold and into the other category at another threshold.

4. Comparison of PR Systems with FHEs

As previously stated the evaluation of FHEs opinion data on the La Trobe trials had been carried out as part of the program offered over the trial years. For each of the yearly La Trobe trials, FHEs were provided with a hardcopy image 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. For one of the trials they were also informed that a calligrapher group was used in the production of some of the simulations (Dewhurst et al., 2008). FHEs were asked to express their opinion as to authenticity of each of the questioned signatures on a five-point scale. For simplicity, we did not consider the levels of opinions in this study.

For the year 2001 La Trobe signature data, 51 answer booklets were submitted, comprising 10 peer reviewed responses (cross-checked by a second FHE), 31 individual responses (not peer-reviewed), and 10 experimental responses (from individuals and trainees). A total of 10250 authorship opinions were expressed by the group. Of these opinions 49.2%were correct, 7.4% were misleading and 43.5% were inconclusive.

Type	Genuine	Disguised	Forged	Total
Correctly Classified	990	69	343	1402
Misleading (Errors)	1	13	9	23
Reported Inconclusive	9	78	488	575
Total	1000	160	840	2000

Table 5: Results of the blind trial conducted with FHEs on the data from year 2004.

Type	Genuine	Disguised	Forged	Total
Correctly Classified	587	73	1263	1923
Misleading (Errors)	1	52	174	227
Reported Inconclusive	32	154	764	950
Total	620	279	2201	3100

Table 6: Results of the blind trial conducted with FHEs on the data from year 2005.

This translates into an error rate of 13.0% on the decisions (accuracy of 87.0%) when those opinions that were inconclusive were disregarded. The opinion data associated with these results is given in Table 4.

For the year 2004 La Trobe signature data, 21 answer booklets were submitted, comprising 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%) when those opinions that were inconclusive were disregarded. A detailed breakdown of these results is given in Table 5.

For the year 2005 La Trobe signature data, in total, 31 answer booklets were submitted, comprising 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%) when those opinions that were inconclusive were disregarded. A detailed breakdown of these results is given in Table 6.

For the La Trobe data collection of year 2006, in total, 33 answer booklets were submitted, comprising 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 40.5% were correct,7.2% were misleading and 52.3%were inconclusive. This translates into an error rate of15.2% on the decisions (accuracy of 84.8%) when those opinions that were inconclusive were disregarded. The opinion data associated with these results is given in Table 7.

In addition to the collective results, various tests were performed to analyze the errors made by individual examiners. Figure 1 shows the examiner scores (inconclusive and misleading/incorrect opinions) by questioned signature category for the 2006 trial. The percentage inconclusive opinions are colored yellow and the percentage incorrect opinions are colored red. The x-axis depicts the examiners' anonymous identification

Type	Genuine	Disguised	Forged	Total
Correctly Classified	93	10	1151	1254
Misleading (Errors)	0	111	113	224
Reported Inconclusive	0	96	1526	1622
Total	93	217	2790	3100

Table 7: Results of the proficiency tests conducted with FHEs on the data from year 2006.

code. It can be seen that a large number of FHEs were either inconclusive or they misclassified the disguised signatures and forgeries. They, on the other hand, were quite good at identifying the genuine signatures individually. The results from the other years show a similar trend, however for brevity, they are not included here.

Since we might predict that FHEs will exhibit a much wider range of performance success as compared to automatic systems, several other tests were performed to characterize the FHE data. The relationship between examiners experience and the total number of opinion errors (see Figure 2), and the relationship between the time examiners took to complete the trials and the total number of opinion errors (see Figure 3) is presented here for interest. For brevity, we present these results for the 2001 data only. Both for Figure 2 and Figure 3, no simple correlation was found to exist between the two variables (at x and y axis). The experiments show that there is no support for the notion that the validity of a trained examiner's 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 examiner's

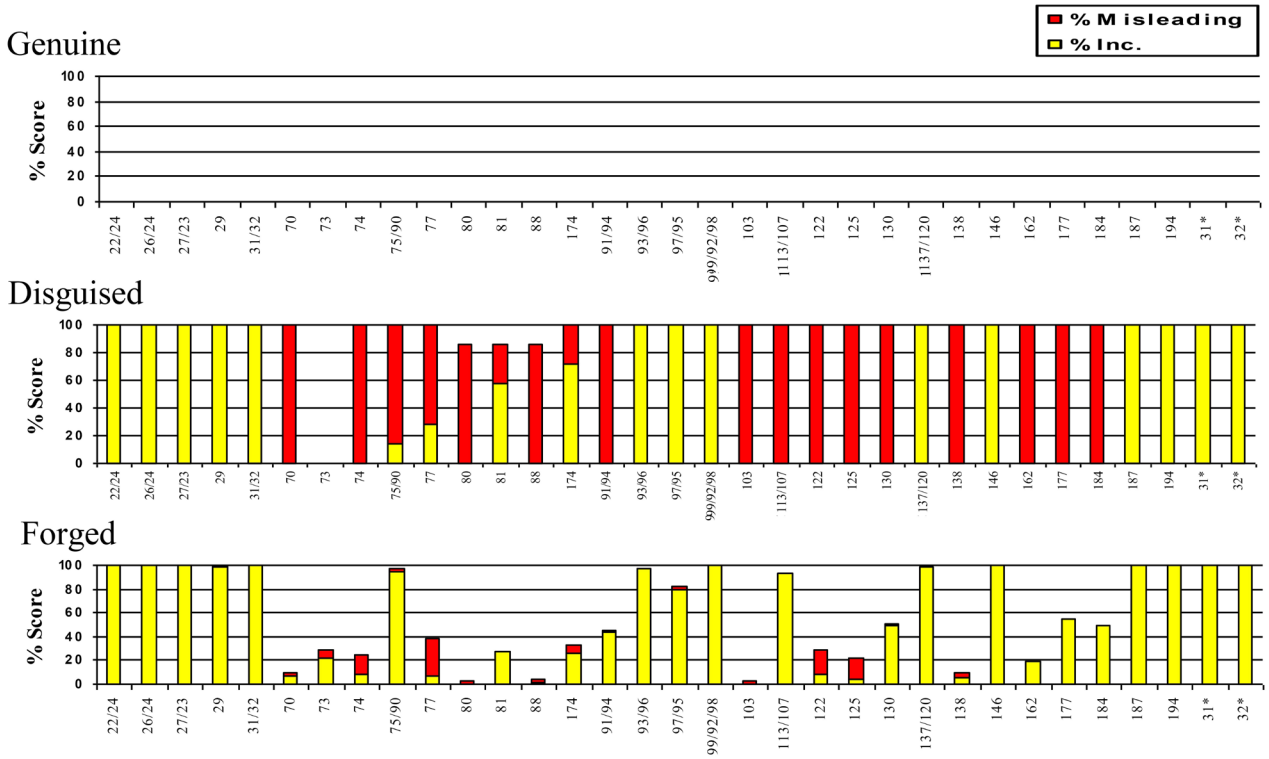

Figure 1: Results of individual FHEs on the La Trobe 2006 data. Total examiners: 33.
X-axis: IDs of examiners, Y-axis: percentage score of each examiner.

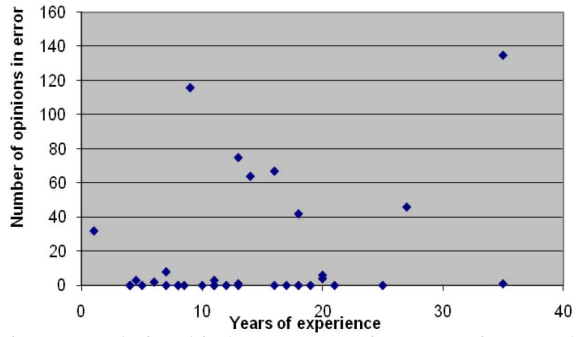

Figure 2: Relationship between examiners experience and the total number of opinion in errors.
Points indicate the years of experience of FHEs and the corresponding number of opinions expressed in error.

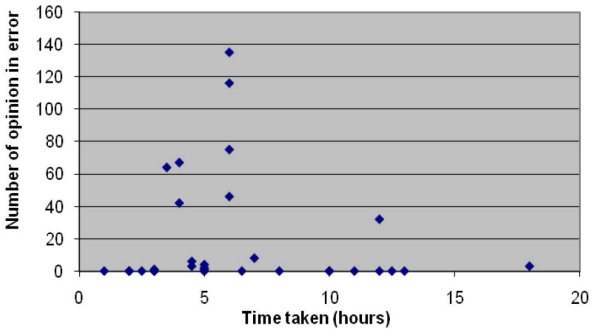

Figure 3: Relationship between the time taken by FHEs and the total number of opinions in error.
Points indicate the time taken by FHEs to complete the trial and the corresponding number of opinions in error.

opinion can be referenced by the amount of time the examiner spent performing the task.

We also measured the time taken by automatic systems to complete the verification task for the data sets from different years. Most of the automatic systems were able to complete the task very efficiently, e.g., the most efficient automatic system was able to output results for the whole data of year 2006 in less than 100 seconds. The important point here is that the state-of-the-art automatic systems usually look only at specific information/evidence present in the signatures (and in fact machines are quite efficient in processing specific information in this way). The humans in our study, on the other hand, use complex perceptual and cognitive processes to assess all of the features of the questioned signature trace and not surprisingly take vastly longer to perform the task. Taking this into consideration, we are not reporting a direct time comparison between man and machine in this paper.

The overall man vs. machine comparison was initially performed on the basis of collective accuracies of the two, man and machine. Table 8 provides the overall results of this performance comparison taking into account the complete signatures; i.e., along with genuine and forged signatures, the disguised signatures were also considered while computing these results. We report the average as well as the best performances, of man and machines, to provide a clear comparison between the two. As shown in Table 8, there is much variation in human performance from trial to trial when compared to that of machines. For example in the year 2001, similar trends can be seen for the results from other years, the average human performance is at 44.8% although the best performance by any human expert is at 100%. For humans in general we can assume large variance in performance whereas for machines the average performance is at 70.8% and the best at 93.6% showing comparatively less variance. Here we may infer that most of the state-of-art automatic methods (applying different classification approaches) perform close to each other, and humans carry great

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
2006	38.8	71.7	91	92

Table 8: Man vs. machine results for the data collections from various years

performance diversity, both in terms of accuracy and speed. It is clear that automatic systems could provide a good supplementary objective tool for FHEs as they provide quite consistent results. Further, these systems can also be used to cut down a large population into a smaller population (due to their speed) when examining real world signature cases.

In order to measure the total performance capabilities of automatic systems, we generated Receiver Operating Characteristic (ROC) curves (Fawcett, 2006). These curves are given in Figure 4(combined data from 2001, 2004, and 2005) and Figure 5 (data from 2006). Note that an automatic system has many possible points/thresholds on which it can operate to reach an opinion on classifying signatures as genuine, forged, or disguised. Therefore, it is preferred to represent the complete performance behavior in the form of so-called ROC curves. Generally, these curves are developed by considering the FRR on one axis and FAR on the other axis while varying the thresholds on the basis of which a system gives an opinion about signature type (genuine, forged, or disguised). This generates the complete behavior of a system on the given data in the form of a curve in the FRR/FAR space. In Figures 4 and 5, FHEs' performance, unlike automated systems, is represented by single points. These points are calculated by looking into the overall experts' performance. The false acceptance is calculated by taking the ratio of the forged questioned signatures which were misclassified as genuine by the examiners and the total forged questioned signatures. The inconclusive opinions were not considered. The false rejection was computed by taking the ratio of the sum of disguised and genuine questioned signatures which were misclassified as forged, and the total disguised and genuine questioned signatures. The inconclusive opinions were again neglected. These are plotted as single points in the FRR/FAR space (the same containing the ROC-curves for the automatic systems). Note that, unlike machines, a complete ROC curve of human performance is impossible since there are no objective numerical thresholds for humans who they can vary with respect to their opinions regarding signature classification (Malik et al., 2013).

As can be observed from Figures 4 and 5, humans outperformed nearly all the machines. An important reason for these results is that humans used a possibility to note their opinion as inconclusive when they were unable to find enough evidence of genuine or forged authorship as per their analysis. The machines were also given this possibility but none of the machines used this, and nearly in all the cases came up with an opinion. None the less, the humans also carried a great deal of previous knowledge in terms of their experiences in solving forensic cases, but machines relied on the case specific data alone. This might have also affected the performance of machines.

Note that currently there are some limitations associated with automatic systems which are required to be overcome in order for these systems to be applicable in real world forensic cases. The main challenges are that these systems need to train on more forensic data captured from real casework to improve system learning (Liwicki et al., 2012), that some forensic environments require report outputs in the form of likelihood ratios (according to the Bayesian inference) to be acceptable as a laboratory output (Gonzalez et al., 2005; Malik & Liwicki, 2012), and that automatic systems should provide explanations of their outputs such that FHEs can weigh the probative value of their outputs accordingly. It is the case that no state-of-the-art automatic system is currently capable of completely fulfilling these desired requirements (Malik & Liwicki, 2012). Having said this, since automatic systems are extremely efficient when compared to humans, they have the potential to serve as assistants for human experts where they may potentially be guided by fast objective data. Many methods can be devised to enable machines to automatically incorporate knowledge from previous cases, e.g., using case based reasoning (Weber et al., 2009). The challenge here is investigate whether it is really helpful, required, or even recommended to consider the incorporation of previous knowledge when automatically classifying signatures by machines.

5. Conclusions and Future Work

In this paper we have provided the results of a detailed study performed in order to compare signature verification performance of FHEs against automated systems. As the technology around automated systems develops, the potential applications for these objective systems in forensic casework grow. This paper shows that the performance of automatic systems, although

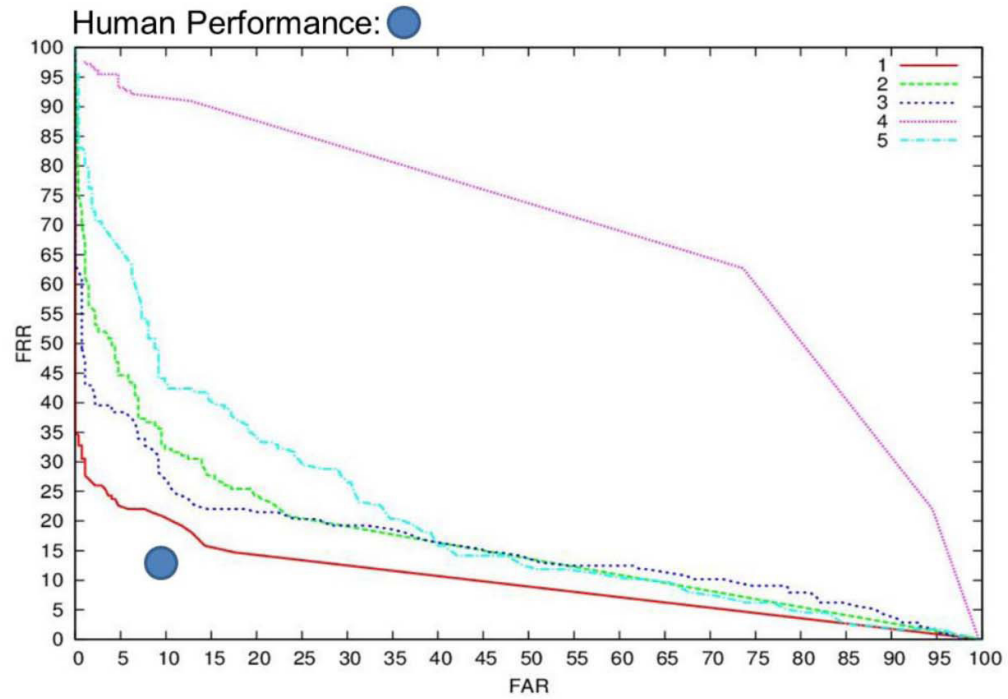

Figure 4: Man vs. machine comparison in the FAR/FRR space (on combined 2001, 2004, and 2005 data).

Systems 1-5: participants of the 4NSigComp2012 competition.

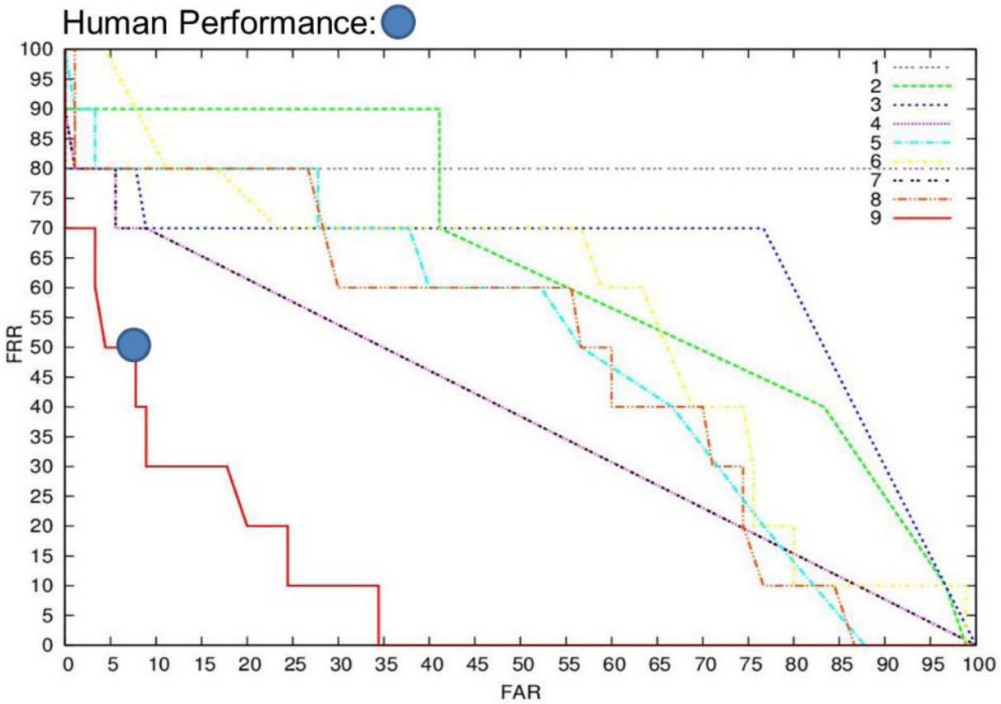

Figure 5: Man vs. machine comparison in the FAR/FRR space (on 2006 data). Systems 1-7 are participants of the 4NSigComp2010 competition while system 8 and 9 are added later (details in Malik et al., 2011)

in many respects unlike that of FHEs, can result in characteristics approaching the average for FHEs participating in the La Trobe trials. Our study suggests that different automatic systems, just like humans, were better on different data. However, there was not much variance in the performances of automatic systems which is unlike that of humans, as humans' performance showed a great degree of variations from average to the best case performance.

The automated systems/machines encountered difficulties in correctly classifying disguised signatures. When disguised signatures were removed from the test data, some automatic systems could reach an EER of nearly 0% in one of the datasets. Similar to machines, FHEs also faced difficulties when they attempt to classify questioned signatures that are a product of disguise behavior. This is likely to result from the mixed signal that disguised signatures provide to both FHEs and the automated machines. Similarities may exist with the genuine signature, since it was written by the writer of the authentic signatures, and dissimilarities may exist due to the conscious changes to the signature made by the genuine writer in order to introduce features where denial can be claimed. The human experts faced problems in correctly classifying disguised signatures; however, they had used a possibility to declare their findings inconclusive on the basis of not being able to find *enough* evidence of genuine, forged, or disguised authorship from the signatures (the automatic systems were also provided this possibility but no participating system used this). In fact, a large number of human trials reported disguised and forged signatures as inconclusive. Furthermore, both humans and machines were in many cases accurate in identifying genuine signatures.

Performance comparisons of the type described here offer promise regarding the future of objective techniques in forensic casework. We must be careful however not to overestimate the potential of automated techniques since this study is based on data derived not from casework, but carefully constructed blind trials. The signature sets, both specimen and questioned, are therefore very 'clean' with respect to controlling variables which are not normally able to be controlled in casework (e.g., controlling the representativeness of the population of specimen signatures, the type of writing instrument, the writing medium, the writing conditions etc.). In many cases casework samples suffer from a lack of specimen signatures, or specimen signatures that may not be representative of the writer's normal behavior. It is not known to what extent these sorts of limitations impact on the potential of automated systems to produce accurate or useful results.

In the future we plan to perform analyses on data with more reference writers and forgers. In particular, we plan to gather larger datasets containing disguised signatures and making them publicly available so that the performance of automatic systems can be further improved for this important category of signatures. It is planned to organize various competitions and workshops on the topic of automated forensic signature analysis.

Acknowledgements

The authors would like to thank Elisa van den Heuvel and Linda Alewijnse from the Netherlands Forensic Institute for providing us with the data and their support during the course of this study.

References

Bird, C., Found, B., &Rogers, D. (2007). Forensic document examiners' opinions on process of production of disguised and simulated signatures. The 13th Int. Graphonomics Society Conf., 71-174.

Bird, C., Found, B., &Rogers, D. (2009). Forensic handwriting examiners' skill in detecting disguise behavior from handwritten text samples. The Conf. of the European Network of Forensic Handwriting Experts.

Bulacu, M., & Schomaker, L. (2007). Text-independent writer identification and verification using textural and allographic features. IEEE Trans. Pattern Analysis and Machine Intelligence, 29 (4), 701-717.

Conway, J. V. P. (1959). Evidential documents. Charles C Thomas, Illinois.

Dewhurst T., Found, B., & Rogers, D. (2008). Are expert penmen better than lay people at producing simulations of a model signature? Forensic Science International, 180(1), 50-53.