Towards Explainable Writer Verification and Identification Using Vantage Writers*

Axel Brink Lambert Schomaker Marius Bulacu
Dept. of Artificial Intelligence, University of Groningen
{a.a.brink, schomaker, bulacu}@ai.rug.nl

Abstract

In this paper, a new method for off-line writer verification and identification is proposed which encodes writer features as a mix of typical handwriting styles, written by so-called vantage writers. Since their handwriting can be shown to the user, the method provides a degree of transparency that is usually not present in automatic verification and identification systems. It acts as a dimensionality reduction of a precomputed basic feature vector. In this experiment, the hinge feature was used. The method was tested with unconstrained connected cursive text in two datasets: a known dataset of 252 writers and 1074 writers from a new, forensic dataset.

1. Introduction

Forensic document examination has always received scepticism. The main reason is that judgments of experts are considered to be subjective [10]: their performance varies [3] and they have no means of quantifying their findings [2]. In the courtroom, this is considered as a deficiency. Although it has been shown that forensic experts are skilled [3, 4], we think that forensic document examination can be strengthened with a more solid scientific basis.

Since the nineteen eighties, the advent of computers has inspired researchers to make computer programs to support (or even replace) the human expert's judgment. Such programs exist in two variants: programs for writer *verification* and writer *identification*. In verification, the computer program determines whether two documents have been written by the same person. This can be useful when a suspected writer of a questioned document has already been found, and he or she has written a sample text for comparison. On the other hand, writer identification is the process of finding possible suspects in a database. This can be applied when no suspected writer of a questioned document is available, but handwritten texts of known authorship already have been collected. Over time, several writer verification and identification systems have been implemented and impressive performances have been reported [7], but still no system exists that convinces forensic experts. One of the reasons is that the output of such systems is often hard to interpret. Experts consider them as black boxes of which the inner workings are unclear. These systems usually yield numbers without an intuitive explanation. That issue is addressed in this paper.

A method for automatic off-line writer verification and identification is proposed, which allows to generate reports that are quite comprehensible. The idea is that a person's handwriting can be seen as a mixture of handwritings of typical writers, the *vantage writers*. The degrees of dissimilarity between each of the vantage writer's handwritings and a fresh document form a small list of numbers, called a *vantage profile*. This profile is used to discriminate writers. The vantage writers are each represented by a document selected from the input data. This makes it possible to *show* their handwriting, which makes the new feature vectors more comprehensible than plain feature vectors.

To test how this approach performs, we have created setups that perform both writer verification and identification based on vantage profiles. Two datasets have been used separately; one of them is a newly introduced collection of confidential forensic data.

1.1. Related work

A vantage profile based on vantage writers is an implementation of a *dissimilarity representation*, as introduced in [6]. In this publication, classes are discriminated by a Bayes classifier that assumes normal distributions of the dissimilarity values. We used a different approach for the classification, because there are very few samples per writer.

A similar approach is described in [11], where hieroglyphic images are retrieved using *vantage objects*. On-line characters are hierarchically clustered in [12]; each cluster is represented by a prototypical input sample, and each

*Published in ICDAR 2007, pages 824-828.

writer is assigned a style vector that indicates the usage of each of the prototypical characters. In [13], characters from several copybooks were clustered in order to determine the nationality of writers based on their character usage. In [5], writing styles are expressed as usage of inferred copybooks.

In the next section, the method and experiments are described in detail. In section 3, the verification and identification performance of the vantage writer method are presented. Finally, in section 4, the results are discussed and future work is suggested.

2. Method

For each image dataset, basic features of the images were extracted first, resulting in two feature datasets. These were each split into a *train set* and a *test set*. The train set was used to select the vantage writers. It was also used to determine a classification threshold for verification. The test set was used to assess the performance independently. A more detailed description follows in the next subsections.

2.1. Dataset preparation

Two datasets were used separately to provide the input patterns for the experiments: Firemaker and NFI. The *Firemaker* [9] dataset consists of 1008 scanned pages of handwritten texts, written by 252 students, 4 pages each. The pages were lined with a color that vanished during scanning. A glance at the dataset gives the general impression of neatly written text. Only two pages per writer were used for the current experiment, page 1 and page 4, since these contain unconstrained connected cursive handwriting.

The *NFI* dataset is a new and very heterogeneous set of handwritten pages that have been collected by the NFI, the Dutch National Forensic Institute. The collection consists of 3501 scanned forms with handwritten material, on demand written by 1311 suspects in criminal cases. Most of the suspects wrote two pages, but there are many exceptions. Most of them wrote a dictated standard text; the first part in connected cursive, the second part in capitals. In many cases the transition from cursive to capitals occurs within a page. The handwriting in this dataset gives an impression of great sloppiness. This is due to the facts that the forms were not lineated, the subjects have a lower level of education and some of the subjects may have not been cooperative. Also, many pages contain erasures or visible perforator holes. Altogether, one could conclude that this is a "dirty" dataset. For examples, see Figure 1.

For performance evaluation, the Firemaker dataset could be used as-is, but the NFI data required some preprocessing. First, border effects like visible pieces of form fields and the paper edge were removed by cutting along straight horizontal and vertical lines. The positions of the cutting

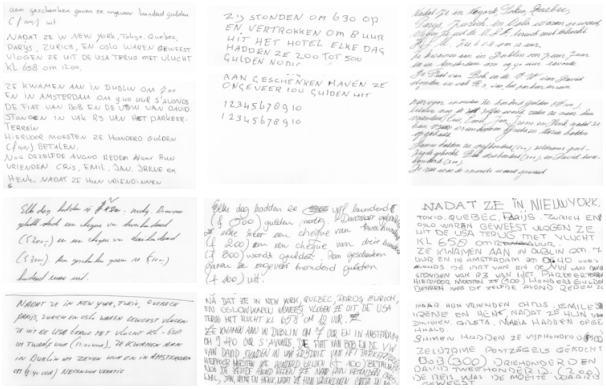

Figure 1. Example documents in the NFI dataset, each one cut in two parts.

lines were determined heuristically using a smoothed projection histogram. The same kind of histogram was used to split the page into an upper and a lower part, to increase the probability that every writer has at least two pieces of similar text in the database; the parts were treated as separate documents. This method is relatively simple, and visual inspection proved that it works reasonably well. However, as much as 1127 pages had to be rejected because the text baseline was too curved or sloped, making a horizontal cut between two text lines in the middle of the text impossible. Still, after splitting, 4748 page parts remained, written by 1074 persons. The fact that the pages were cut in two parts that are treated as separate documents is a weakness; it biases the view on the verification and identification results positively. On the other hand, the NFI data has not been collected with automatic processing in mind, and is contaminated in many ways.

2.2. Basic feature extraction

Creating a vantage profile for an image requires that a basic feature vector has already been computed. This can be done using virtually any feature extraction method. In our experiment, we computed feature vectors for all input documents using a method that has proved to be very effective: the *hinge* feature [1]. This technique captures the orientation and curvature of the ink trace as follows: On every point on the ink contour, two small connected edge fragments are are locally aligned with the contour and form a shape like a hinge. The angles of both legs of the hinge with respect to the horizontal are recorded into a 2D histogram. By normalization this histogram is converted into a 2D probability distribution, which is called the "hinge" feature.

2.3. Selecting vantage writers

Vantage profiles represent writer information as a mixture of the handwriting style of typical writers, the vantage writers. To create these profiles, the set of vantage writers must be defined first. This can be done in various ways. For example, they could be manually selected to represent styles from different countries, sexes or ages. In our implementation, the vantage writers were represented by a sample of input documents in the training part of the dataset, the *vantage-writer sample*. This was done by random sampling a fixed number of times, and picking the best choice afterwards. The number of random samples was set to 50 for the Firemaker dataset, and 25 for the NFI dataset, for reasons of computing time. For every sample, the performance was computed by creating vantage profiles and performing writer verification or identification on the train set as described in the next subsections. The sample yielding the highest performance on the train set was assigned to be the final set of vantage writers. Vantage writers were determined for verification and identification separately.

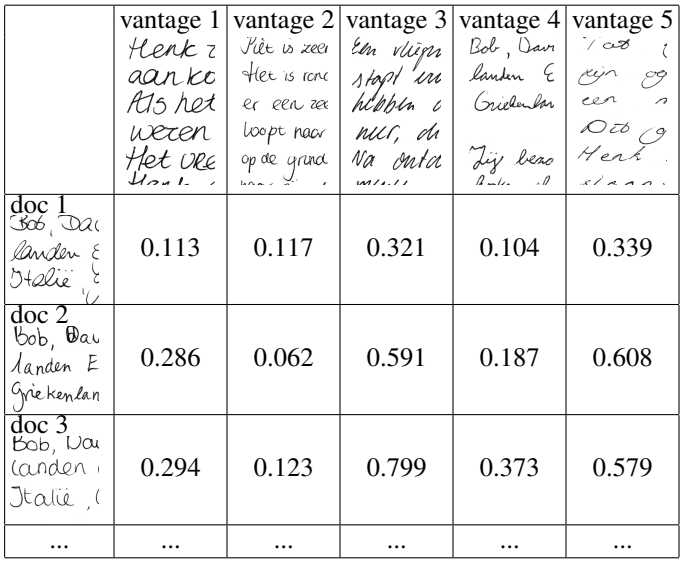

Figure 2. Example vantage profiles based on five vantage writers in the Firemaker dataset.

2.4. Creating vantage profiles

The vantage profile $\mathbf{q}'$ of a document is acquired by computing the *distance* between its basic feature vector $\mathbf{q}$ and the basic feature vectors $\mathbf{v}_j$ of the sample documents of each of the vantage writers j. The distance can be computed using various measures, such as Euclidean, Hamming, χ^2, Bhattacharyya, etc. Since a hinge feature vector is essentially a probability distribution and the χ^2 measure has proved to be effective for this kind of feature vectors [8], this measure was used. It is defined (after renaming) as:

$$d_{\chi^2}(\mathbf{q}, \mathbf{v}_j) = \sum_{i=1}^{|\mathbf{q}|} \frac{(\mathbf{q}_i - \mathbf{v}_{ji})^2}{\mathbf{q}_i + \mathbf{v}_{ji}}$$

where i is an index to the elements of $\mathbf{q}$ and $\mathbf{v}_j$.

The distances together form a vector of distances $\mathbf{q}'$, which we call a *vantage profile*. This can be written as:

$$\mathbf{q}' = (d(\mathbf{q}, \mathbf{v}_1), ..., d(\mathbf{q}, \mathbf{v}_n))$$

where d is a dissimilarity measure, $\mathbf{v}_j$ are the feature vectors of the vantage writers and n is the number of vantage writers. It is seen as a new, indirect, feature vector that is used to discriminate writers. As such, the computation of vantage profiles can be seen as a form of dimensionality reduction. See Figure 2 for an example.

2.5. Writer verification

Writer verification means that a decision must be made whether two documents have been written by the same person. In this experiment, such a decision was made by testing whether the dissimilarity between two vantage profiles was below a certain threshold θ. This can be symbolically expressed as:

$$h(\mathbf{q}', \mathbf{r}') = \begin{cases} true & \text{if } d(\mathbf{q}', \mathbf{r}') < \theta \\ false & \text{otherwise} \end{cases}$$

where h is the classifier function. If the dissimilarity was below the threshold, the verdict of the verification was *true*, or *same writer*, otherwise *different writer*. The dissimilarity can be computed by virtually any distance measure; we used Euclidean distance.

The threshold was learned from the documents in the train set that were not used in a vantage sample. This was done by modeling the distances between vantage profiles within the "same-writer" and "different-writer" classes. These distances were found by comparing every document with every other and computing the Euclidean distance between them. Smoothed-probability distributions of the distances in both classes were created using Parzen windowing with a Gaussian kernel. For an example, see Figure 3. Based on these probability distributions, the threshold was positioned for the highest expected performance in terms of the *true positive* (TP) and *true negative* (TN) percentages (in other words, the lowest *false negative* and *false positive* percentages). The expected TP and TN could be balanced according to the desire of the end user, but as there was no such information available yet, the threshold was selected such that TP = TN: the *equal-error rate* (EER). The

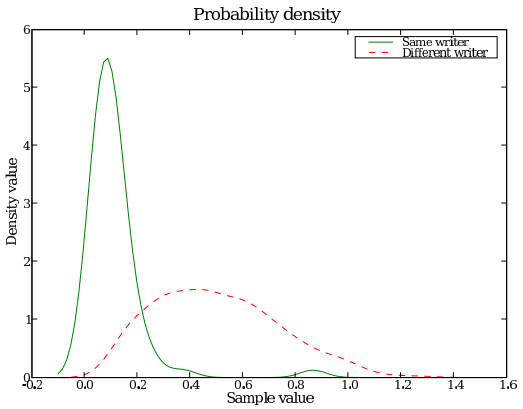

Figure 3. Model of hinge-5-vantage profile distances in Firemaker dataset, smoothed using Parzen windowing with a Gaussian kernel.

EER was also used as the performance measure for vantage writer selection.

2.6. Writer identification

Writer identification is the process where a list of documents is yielded with handwriting similar to the handwriting in a questioned document. In this experiment, each document in the train set was once treated as a questioned document, and its *hit list* was constructed. This is a sorted list containing the s nearest neighbors of the questioned document based upon the vantage profiles, with $s = 1, 10, 100$. During the selection of vantage writers, the writer identification performance was simply assessed as the top-1 performance ($s = 1$).

3. Results

To test the performance of the vantage writer method, several runs of K-fold cross validation were carried out, varying the number of selected vantage writers: $n = 2, 4, 5, 50$. In each run, the cross validation iterated four times ($k = 4$), where in every iteration 25% of the data was available for training and 75% for testing. The data was split such that all pages of each writer were always in the same part. Testing was done for verification and identification separately.

For verification, the train part of the dataset was used to select vantage writers, compute vantage profiles and determine a verification threshold as described in 2.3, 2.4 and 2.5. Then, the found vantage writers and threshold were used to perform writer verification in the same way, but on the *test* set. The TP and TN were recorded. The results for the Firemaker and NFI dataset are shown in Table 1. For comparison, the table also includes a line with the performance of the bare hinge feature vector, without the indirection induced by vantage writers. The tables show that the performance using vantage profiles is similar to the performance using the hinge feature directly, and that the number of vantage writers does not have much influence. The ratio between TP and TN shifts a bit as the number of vantage writers increases; this is probably an effect of the Parzen smoothing. The effect of smoothing can still be optimized by changing its parameters.

	Firemaker		NFI	
	TP	TN	TP	TN
hinge	96.1%	83.5%	79.0%	73.9%
hinge-2-vantage profile	91.4%	85.9%	74.5%	75.8%
hinge-4-vantage profile	92.0%	89.7%	75.1%	77.8%
hinge-5-vantage profile	90.8%	90.5%	75.1%	77.8%
hinge-50-vantage profile	88.7%	91.3%	75.0%	77.6%

Table 1. Average verification results for the Firemaker dataset (378 pages, 189 writers) and NFI dataset (3561 pages, 805 writers).

	Top-1	Top-10	Top-100
hinge	67.3%	89.0%	98.3%
hinge-2-vantage profile	12.4%	48.9%	94.8%
hinge-4-vantage profile	30.6%	72.3%	96.2%
hinge-5-vantage profile	36.5%	75.3%	97.3%
hinge-50-vantage profile	41.0%	76.8%	96.5%

Table 2. Average identification results for the Firemaker dataset (378 pages, 189 writers).

Similarly, for identification, the train part of the dataset was used to select vantage writers and compute vantage profiles as described in 2.3, 2.4 and 2.6. Then, the found vantage writers were used to perform writer identification in the same way on the test set. Top-1, top-10 and top-100 performance were recorded. The results for the Firemaker and NFI dataset are shown in Table 2 and 3 respectively. For comparison, the table also includes a line with the performance of the bare hinge feature vector, without the indirection induced by vantage writers. It is clear that the vantage writer method does not work as well for identification as it does for verification.

	Top-1	Top-10	Top-100
hinge	53.5%	77.1%	92.2%
hinge-2-vantage profile	2.8%	17.4%	61.2%
hinge-4-vantage profile	14.1%	42.8%	80.6%
hinge-5-vantage profile	18.2%	48.2%	82.6%
hinge-50-vantage profile	28.3%	58.5%	85.6%

Table 3. Average identification results for the NFI dataset (3561 pages, 805 writers).

4. Discussion

In this paper a method was proposed that is a step towards explainable automatic writer verification and identification. This transforms the "black box" that automatic systems usually are into a more transparent one, since basic components can be shown to the user: a piece of text written by the vantage writers on which the output is based. The output is currently a *same writer* / *different writer* verdict (for verification) or a hit list (for identification), accompanied by vantage profiles. These show the degree of dissimilarity to the handwriting of each of the vantage writers. The system could be made even more transparent when the dissimilarities are replaced by probabilities. This could be done when the measurements are repeated in multiple pieces of text from the same writer. This is left as future work.

The results show that the method works very well for writer *verification*, even with as few as two vantage writers, yielding only a small performance drop relative to the original input hinge feature vector, while representing writers in only two dimensions. This is an enormous dimensionality reduction, which by itself can be a reason to use this method. The performance results for *identification* do not give rise to that much optimism, but our approach may still be useful because of its explainable basis. In any case, there is room for improvements. We plan to improve the preprocessing to be able to use more of the authentic forensic NFI samples. Furthermore, the vantage profile might be constructed based upon one or more other basic features; the vantage writers may be selected more thoroughly using more extensive stochastic sampling and other distance measures may be evaluated. The set of vantage writers could also be manually selected by experts, representing specific groups of writers distinguished by criteria like sex, handedness, age, nationality and script type. We have already tried unsupervised clustering using a Kohonen self-organizing map (SOM), but that did not result in better performance than using stochastically sampled optimal vantage writers. Additionally, the clustering-based vantage centroids did not contribute much to the explainability of results as was evident from discussions with a forensic expert. Transforming current writer vantage-distances to reliable probability estimates will hopefully further improve practical usability. Concluding, while there is still room for improvement, the proposed system could be the basis for a computer program that can be used in forensic practice.

References

[1] M. Bulacu and L. Schomaker. Text-independent writer identification and verification using textural and allographic features. *IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI)*, 29(4):701–717, 2007.

[2] B. Found and D. Rogers. A consideration of the theoretical basis of forensic handwriting examination: The application of 'complexity theory' to understanding the basis of handwriting identification. *International Journal of Forensic Document Examiners*, 4:109–118, 1998.

[3] B. Found and D. Rogers. Problem types of questioned handwritten text for forensic document examiners. In *Advances in Graphonomics: Proceedings of the 12th Biennial Conference of the International Graphonomics Society*, pages 8–12, 2005.

[4] M. Kam, G. Fielding, and R. Conn. Writer identification by professional document examiners. *Journal of Forensic Sciences*, 42:778–785, 1997.

[5] R. Niels and L. Vuurpijl. Generating copybooks from consistent handwriting styles. In *Proc. of the 9th International Conference on Document Analysis and Recognition (ICDAR)*, 2007. In press.

[6] E. Pekalska and R. Duin. Dissimilarity representations allow for building good classifiers. *Pattern Recognition Letters*, 23(8):943–956, 2002.

[7] R. Plamondon and G. Lorette. Automatic signature verification and writer identification - the state of the art. *Pattern Recog.*, 22(2):107–131, 1989.

[8] L. Schomaker and M. Bulacu. Automatic writer identification using connected-component contours and edge-based features of uppercase western script. *IEEE Trans. Pattern Anal. Mach. Intell.*, 26(6):787–798, 2004.

[9] L. Schomaker and L. Vuurpijl. Forensic writer identification: A benchmark data set and a comparison of two systems. Technical report, NICI, Nijmegen, 2000.

[10] S. N. Srihari and G. Leedham. A survey of computer methods in forensic document examination. In *Proc. Int. Graphonomics Soc. Conf.*, pages 2–5, 2003.

[11] J. Vleugels and R. C. Veltkamp. Efficient image retrieval through vantage objects. In *Proc. of the 3rd Int. Conf. on Visual Information and Information Systems VISUAL'99*, LNCS 1614, pages 575–584, 1999.

[12] V. Vuori. Clustering writing styles with a self-organizing map. In *Proc. of the 8th IWFHR*, pages 345–350, 2002.

[13] S. Yoon, S. Choi, S. Cha, and C. Tappert. Writer profiling using handwriting copybook styles. In *Proceedings of the 8th ICDAR*, pages 600–604. IEEE, 2005.