Automatic Handwriting Identification on Medieval Documents

Marius Bulacu Lambert Schomaker
Artificial Intelligence Institute, University of Groningen, The Netherlands
(bulacu, schomaker)@ai.rug.nl

Abstract

In this paper, we evaluate the performance of text-independent writer identification methods on a handwriting dataset containing medieval English documents. Applicable identification rates are achieved by combining textural features (joint directional probability distributions) with allographic features (grapheme-emission distributions). The aim is to develop an automatic handwriting identification tool that can assist the paleographer in the task of determining the authorship of historical manuscripts.

1. Introduction

The automatic identification of a person on the basis of scanned images of handwriting has received significant research interest in recent years (after 9/11 and the anthrax letters) primarily due to its forensic applicability [9, 2, 7, 4]. This paper explores an alternative use of automatic writer identification for establishing the authorship of old manuscripts for historical studies of paleography and codicology. This complementary application area has not been extensively studied until the present and the main purpose of the current paper is to provide a concrete contribution to the development of this research direction.

At present, in historical studies requiring manuscript authentication and/or dating, handwriting identification is performed by skilled human experts, paleographers, who can identify the writing peculiarities of a particular scribe and can recognize the type of calligraphy used in a given historical period. But given the large amounts of documents deposited in historical archives, a computer system that performs automatic manuscript indexing and retrieval based on script style can become a useful tool for the historian.

A *writer identification* system performs a one-to-many search in a database with handwriting samples of known authorship and returns a likely list of candidates (see Fig. 1). This list is further scrutinized by a human expert (paleographer in the case of historical documents) who takes the final decision regarding the identity of the author of the questioned sample. *Writer verification* involves a one-to-one comparison with an automatic decision whether or not the two samples were written by the same person.

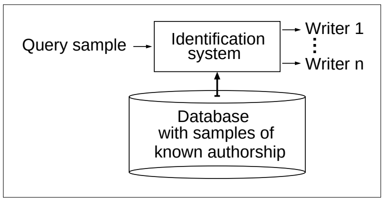

Figure 1. A *writer identification* system generates a hit list containing those handwritings from the database that are most similar to the query in terms of writing style.

Two important natural factors are in direct conflict in the attempt to identify a person based on samples of handwriting: *between-writer variation* as opposed to *within-writer variability*. In automatic writer identification, it is necessary to use computer representations (features) with the ability to maximize the separation between different writers, while remaining stable over samples produced by the same writer. In recent years, we proposed a number of new and very effective statistical features for automatic writer identification [4]. Our features are probability distribution functions (PDFs) extracted from handwritten text blocks and characterize writer individuality *independently of the textual content* of the written samples. In our methods, the computer is completely agnostic of the actual text written in the samples. In the current study, we will test the effectiveness of these features on a dataset of medieval English documents.

Until the present, studies concerning automatic writer identification applied on historical documents remain rather sparse. Run-length histograms were used in [5] for ancient Hebraic handwriting identification on a dataset containing 8 writers. We will also use run-lengths in the present work mainly for comparison purposes. In [2], an information retrieval approach is proposed using graphemes to encode handwriting individuality. The tests were performed on a dataset containing 39 writers from the correspondence of the 19th century French novelist Emile Zola. Our allograph-level technique is similar to the method described in this work, however we adopt a main-stream statistical pattern recognition approach. The Hermite transform is used in [3] for document denoising and handwriting identification on a patrimonial dataset containing 1400 documents from 189 authors, in different languages and alphabets. An engaging research project [1] aims to use automatic writer identifi-

[0]This paper was published as: Marius Bulacu, Lambert Schomaker, *Automatic handwriting identification on medieval documents*, Proc. of 14th Int. Conf. on Image Analysis and Processing (ICIAP 2007), IEEE Computer Society, 2007, pp. 279-284, 11 - 13 September, Modena, Italy

Table 1. The medieval English scribes contained in the experimental dataset and the number of documents produced by each scribe.

	Scribe	No. docs
1	Hengwrt-Ellesmere Scribe	7
2	Hammond Scribe	15
3	Slanting Hooked-G Scribe	7
4	Trinity Anthologies Scribe	7
5	Westminster Scribe	2
6	Beryn Scribe	6
7	Romances Scribe: or Baggehey	4
8	Edmund-Fremund Scribe	9
9	Selden Scribe	6
10	Thomas Hoccleve	7
Total: 10 scribes / 70 documents.		

cation methods for studying the production and dissemination of handwritten texts immediately preceding and after the invention of printing in the mid-15th century. From this project originates the dataset used in the present study. Our interest for writer identification on historical documents was announced in a previous publication [8].

The future development of this area of research requires the cooperation with historians who can formulate pertinent research questions, the creation of sizeable test datasets and the development of functional systems that are able to deal with the complexity of the historical documents using a combination of automatic methods and human input.

2. Data

The dataset used in our experiments contains 10 late medieval English scribes (1375-1525) [1], with a variable number of documents per individual (see Table 1). There are a total of 70 documents and the authorship for each manuscript was ascertained by professor Linne Mooney, expert codicologist-paleographer at University of York, UK.

The gray-scale images (8-bit/pixel) have been collected from a wide variety of sources and have different resolution. The documents are complex, almost always containing graphical objects besides the handwritten text (see Fig. 2). The document layout varies significantly and, frequently, the background is not uniform across the whole manuscript due to aging, stains and noise. Overall, this is a rather difficult dataset that raises significant processing problems without immediate automatic solutions. Our approach was to manually select rectangular regions of homogeneous text that are large enough to obtain reliable statistical features, but at the same time avoid all graphics and have a background sufficiently uniform such that binarization would be possible using a global threshold. Two such regions (A and B), with no overlap, were selected on every manuscript. The two regions originating from the same document have been

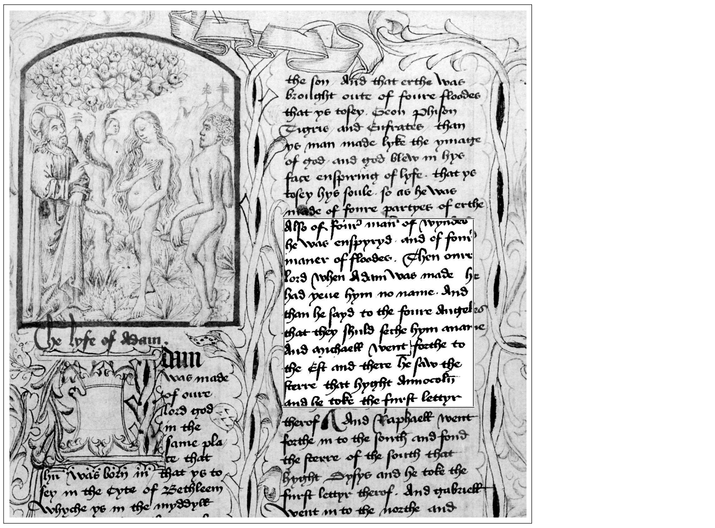

Figure 2. Example of a document from the dataset: the image shows part of a manuscript produced by the Trinity Anthologies Scribe, Cambridge - Trinity College R.3.21, folio 249 Middle English verse and prose. One of the two regions manually selected for writer identification experiments was binarized and appears highlighted.

put in different test suites (set A and set B). Results will be averaged over these two sets.

3. Text-independent statistical features

Two fundamental sources of information regarding the individuality of handwriting are exploited by our techniques functioning at two levels of analysis. First, handwriting slant, curvature and roundness, as determined by habitual pen grip, are captured by joint directional probability distributions operating at the *texture* level. Second, the personalized set of letter shapes, *allographs*, that a writer has learned to use under educational, cultural and memetic influences is captured by a grapheme-emission probability distribution operating at the character level. By combining texture-level and allograph-level features, we achieved very high writer identification and verification performance in extensive tests carried out using large datasets (containing up to 900 subjects) of contemporary Western handwriting [4].

The purpose of this work is to test the effectiveness of our features on historical manuscripts. An overview of all features used here is given in Table 2. We have designed features *f2*, *f3* and *f4*, while features *f1*, *f5* are classically known. The most discriminative features were selected here from a large number of features tested in previous studies.

The regions manually selected are extracted from the original documents and rescaled to obtain approximately the same height (50 pixels) for the handwritten lines. This is estimated using the spread of the peaks in the horizontal-projection profile. The rescaled images are binarized using

Table 2. Overview of features and their dimensionalities.

	Feature	Explanation	Dim
f1	$p(\phi)$	Contour-direction PDF	12
f2	$p(\phi_1, \phi_2)$	Contour-hinge PDF	300
		Direction co-occurrence PDFs	
f3h	$p(\phi_1, \phi_3)$ h	- horizontal run	144
f3v	$p(\phi_1, \phi_3)$ v	- vertical run	144
f4	$p(g)$	Grapheme emission PDF	400
		Run-length on white PDFs	
f5h	$p(rl)$ h	- horizontal run	50
f5v	$p(rl)$ v	- vertical run	50

Otsu's method [6] and then the binary images are further processed by extracting the connected components and their inner and outer contours (using Moore's algorithm).

3.1. Texture-level features

In these features, handwriting is merely seen as an image texture described by probability distributions that capture the distinctive visual appearance of the written samples.

The most prominent visual attribute of handwriting that reveals individual writing style is slant. In fact, the whole distribution of directions in the script provides useful information for writer identification [4]. The directional PDF can be computed very fast using the contours by considering the orientation of local contour fragments determined by two contour pixels taken a certain distance apart (see Fig. 3). As the algorithm runs over the contours, the angle that the analyzing fragment makes with the horizontal is computed using eq. 1 and an angle histogram is built thereby. This histogram is then normalized to a probability distribution $p(\phi)$ that constitutes the feature used in writer identification.

$$\phi = arctan(\frac{y_{k+\epsilon} - y_k}{x_{k+\epsilon} - x_k}) \qquad (1)$$

In our implementation $\epsilon = 5$, this value was selected such that the length of the contour fragment is comparable to the thickness of the ink trace. The number of histogram bins spanning the interval 0° - 180° was set to $n = 12$. These settings will be used for all the directional features.

The directional PDF $p(\phi)$ was our starting point in designing more complex features that give a more intimate description of handwriting individuality and ultimately yield significant improvements in writer identification and verification performance. In order to capture, besides orientation, also the curvature of the ink trace, which is very discriminatory between different writers, we designed the "hinge" feature. The central idea is to consider, not one, but two contour fragments attached at a common end pixel and, subsequently, compute the joint PDF of the orientations of the two legs of the "contour-hinge" (see Fig. 3). The feature $p(\phi_1, \phi_2)$ is therefore a bivariate PDF capturing both

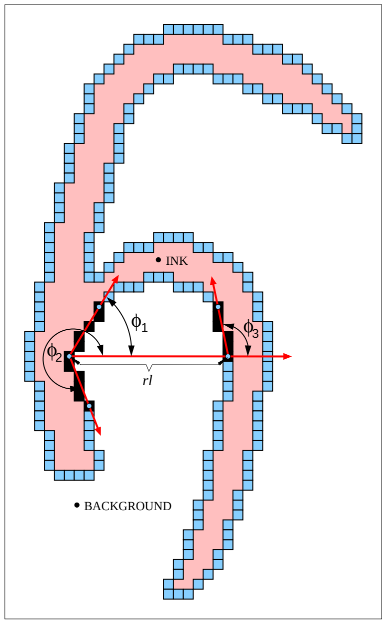

Figure 3. Schematic description for the feature extraction methods of directional and run-length PDFs.

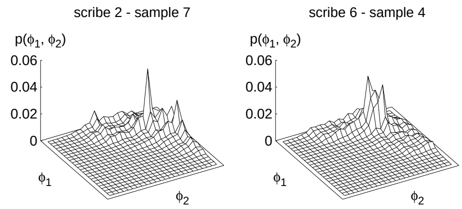

Figure 4. Surface plots of the contour-hinge PDF $p(\phi_1, \phi_2)$ for two writers. Every writer has a different "probability landscape". One half of the 3D plot (on one side of the main diagonal) is flat because we only consider angle combinations with $\phi_2 \geq \phi_1$.

the orientation and the curvature of contours. Examples of $p(\phi_1, \phi_2)$ for two writers are given in Fig. 4.

Building upon the same idea of combining oriented contour fragments, we designed another feature: the directional co-occurrence PDF. For this feature, we consider the combination of contour-angles occurring at the ends of run-lengths on the background (see Fig. 3). The joint PDF $p(\phi_1, \phi_3)$ of the two contour-angles occurring at the ends of a run-length on white captures longer range correlations between contour directions and gives a measure of the roundness of the written characters. Horizontal runs along the rows of the image generate *f3h* and vertical runs along the columns of the image generate *f3v*.

Run lengths are classically known features for writer identification [5]. They are determined on the binary image taking into consideration either the black pixels (the ink trace) or the white pixels (the background). We consider the white runs that capture the regions enclosed inside the letters and also the empty spaces between letters and words. There are two basic scanning methods: horizontal along the rows of the image (*f5h*) and vertical along the columns of the image (*f5v*). Similarly to the contour-based directional features presented above, the histogram of run lengths is normalized and interpreted as a PDF.

3.2. Allograph-level features

Our allograph-level method assumes that every writer is a stochastic generator of ink-blob shapes, or graphemes [8]. The PDF of grapheme usage in a given sample is characteristic of each writer and is computed using a common shape codebook obtained by clustering. To make this approach applicable to free-style handwriting (cursive and isolated), a segmentation method is used yielding graphemes (sub- or supra-allographic fragments) that often will not overlap a complete character. This method is similar to the approach described in [2]. Three processing stages are involved:

1) Handwriting segmentation: the ink is cut at the minima in the upper contour for which the distance to the lower contour is comparable to the ink-trace width (see Fig. 5). Graphemes are then extracted as connected components, followed by size normalization to 30x30 pixel bitmaps.

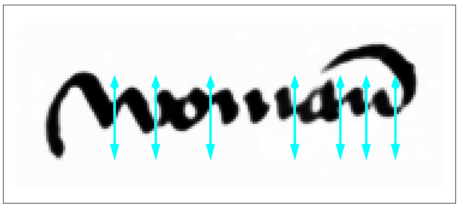

Figure 5. Handwriting segmentation at the minima in the upper contour that are proximal to the lower contour.

2) Shape codebook generation: a 20x20 Kohonen self-organizing map was used to cluster a set of 9k graphemes extracted from 30 samples. These samples used for training the codebook do not overlap, as image regions, with those used in writer identification tests. However, the important separation at the level of writers was not possible due to the reduced size of our experimental dataset.

The codebook graphemes (see Fig. 6) act as prototype shapes representative for the types of shapes to be expected as a result of handwriting segmentation. In [4], we show that the writer identification technique described here is robust to design choices regarding the size of the codebook and the clustering algorithm used to generate it.

3) Grapheme-usage PDF computation: one bin is allocated to every grapheme in the codebook and a shape occurrence histogram is computed for every handwritten sample. For every ink fraglet extracted from a sample after segmentation, the nearest codebook grapheme g is found using Euclidean distance and this occurrence is counted into the

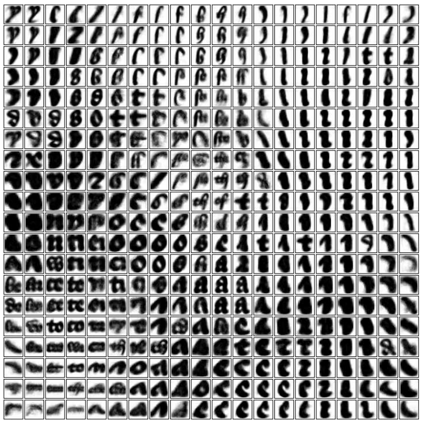

Figure 6. Grapheme codebook generated using a 20x20 Kohonen self-organizing map.

corresponding histogram bin. In the end, the histogram is normalized to a PDF $p(g)$ that acts as the writer descriptor used for identification.

The perfect segmentation of individual characters in free-style script is still unachievable and this represents a fundamental problem for handwriting recognition. Nevertheless, the ink fraglets generated by our imperfect segmentation procedure can still be effectively used for writer identification. The essential idea is that the ensemble of these simple graphemes still manages to capture the shape details of the allographs emitted by the writer.

4. Feature matching and feature fusion

Writer identification is performed using nearest-neighbor classification in a "leave-one-out" strategy: one sample is chosen as the query and all the other samples from the test set (70 - 1 = 69) are ordered with increasing distance from the query, using a selected feature. Ideally the first ranked (Top-1) sample should be one of the other samples produced by the writer of the query. If a longer hit list is considered (Top-10) the chance of finding the correct writer increases. The χ^2 distance is used for matching a query sample q and any other sample i from the test set:

$$\chi^2_{qi} = \sum_{n=1}^{Ndims} \frac{(p_{qn} - p_{in})^2}{p_{qn} + p_{in}} \qquad (2)$$

where p are entries in the PDF, n is the bin index and $Ndims$ is the number of bins in the PDF. The χ^2 distance represents a natural choice for our PDF features.

The considered features capture different aspects of handwriting individuality and operate at different scales. Combining features yields improved performance. In our feature combination scheme, the final unique distance between any two handwritten samples is computed as the average (simple or weighted) of the distances due to the indi-

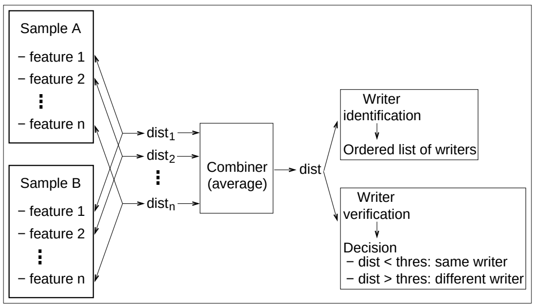

Figure 7. Feature fusion method: the distances due to the individual features are averaged (simple or weighted average) and the result is used in identification and verification.

vidual features participating in the combination (see Fig. 7). In feature combinations, the Hamming distance was used:

$$H_{qi} = \sum_{n=1}^{Ndims} |p_{qn} - p_{in}| \tag{3}$$

5. Experimental results

Table 3 gives the writer identification performance of the individual features considered in this paper. The two sets (A and B) of manually selected regions have been kept separate such that, in the tests, there is always only one sample for every document contained in the original dataset. The results were averaged in the end over the two sets. The best performing features are the contour-hinge PDF (feature *f2*: Top-1 81%, Top-10 96%) and the grapheme PDF (feature *f4*: Top-1 73%, Top-10 97%).

The angle-combination features *f2*, *f3h* and *f3v* perform better than the basic directional PDF *f1*, confirming that joint PDFs capture more individuality information from the handwriting. Despite their higher dimensionality, reliable probability estimates can be obtained from the selected regions containing a few handwritten text lines. The run length PDFs, while having the worst performance among the considered features, provide additional information that will be used in feature combinations.

The identification rates obtained by combining features are given in Table 4. Features *f3* and *f5* (first two rows of the table) merge the two orthogonal directions of scanning the input image and perform better than their single horizontal or vertical counterparts.

The features considered here can be grouped into 3 broad categories (see Table 2): contour-based directional PDFs (*f1*, *f2*, *f3*), grapheme-emission PDF (*f4*) and run-length PDFs (*f5*). The results show that improvements are obtained by combining features from different groups. As stated earlier, feature fusion is performed by distance averaging. Assigning distinct weights for each feature yielded significant performance improvements only for the combinations involving *f2*, which is the strongest individual feature and requires more weight. For the other mixtures, we preferred simplicity and used plain distance averaging. The best performing feature combination fuses directional, grapheme and run-length information yielding identification rates of Top-1 89% and Top-10 97%. Fig. 8 shows a successful hit list generated by our system named GRAWIS (Groningen Automatic Writer Identification System).

It is important to observe that, if we put together the test sets A and B and then run our leave-one-out writer identification search, we obtain a near-100% Top-1 identification rate for the best-performing features and combinations. Almost always sample A will find, as its nearest-neighbor, sample B extracted from the same original document. This demonstrates the discriminatory power of our features, because samples A and B are not overlapping and have different textual content. Nevertheless, we chose to report the more conservative results in tables 3 and 4, because keeping the A and B samples separate is a more realistic situation.

The difference in performance can be attributed to a number of reasons. The historical documents can span a large time period within the life of a scribe and gradual modifications in handwriting style and changes in writing instrument lead to lower overall identification rates. This opens, in principle, the possibility of manuscript dating if labeled samples are available across the time interval of interest. The large within-writer variability of our dataset is

Table 3. Writer identification performance of individual features on the dataset of handwritten medieval documents (10 writers with a variable number of sample per writer - see Table 1). The features are explained in Table 2.

	Feature	Identification rate (%)	
		Top 1	Top 10
f1	$p(\phi)$	47	94
f2	$p(\phi_1, \phi_2)$	81	96
f3h	$p(\phi_1, \phi_3)$ h.	71	93
f3v	$p(\phi_1, \phi_3)$ v.	56	94
f4	$p(g)$	73	97
f5h	$p(rl)$ h.	33	86
f5v	$p(rl)$ v.	44	92

Table 4. Writer identification performance of feature combinations on the dataset of historical documents.

Feature combination	Identification rate (%)	
	Top 1	Top 10
f3: *f3h* & *f3v*	75	94
f5: *f5h* & *f5v*	54	91
f1 & *f5*	69	93
f2 & *f4*	87	99
f3 & *f4*	83	97
f3 & *f5*	81	96
f2 & *f4* & *f5*	89	97
f3 & *f4* & *f5*	89	96

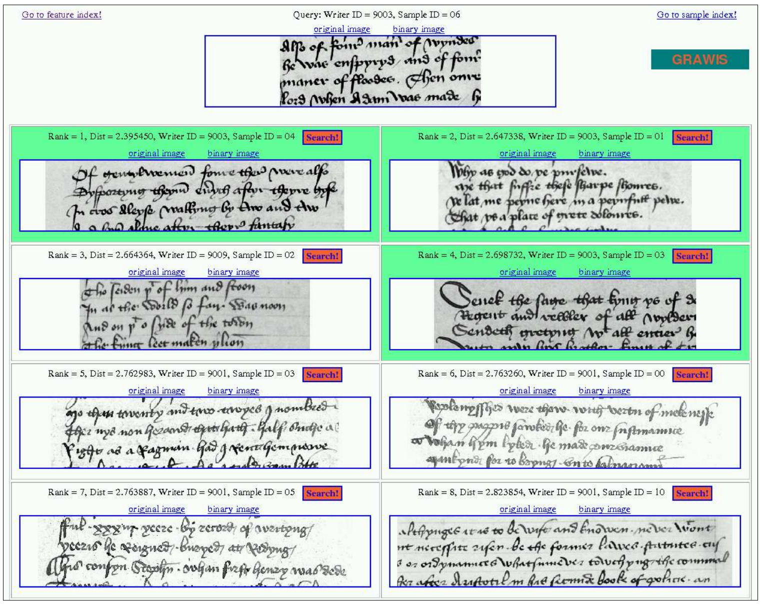

Figure 8. A successful writer identification hit list generated by GRAWIS using feature combination *f2* & *f4* & *f5*. The query is at top-center and other samples produced by the same writer are marked with a darker/green border (ranks 1, 2 and 4 in the hit list).

also testified by our writer verification results where we obtained an equal-error-rate (EER) around 25%. The between-writer variability for historical documents is limited by the fact that the scribes had to adhere to more strict conditions regarding text legibility. Primary image processing difficulties require more elaborate forms of human involvement to mitigate the complexity and noise of historical manuscripts.

We note that the described methods can be equally applied to machine print (font identification), for example in determining the printing house for historical books.

6. Conclusions

The goal of our research is to provide the paleographer with a effective tool that can assist in the process of establishing the authorship of a historical manuscript. After a region of interest from the document has been interactively selected, a hit list is automatically generated using the described text-independent methods. The hit list will contain samples of known authorship that pictorially look similar to the query document, thus providing clues to the human user and allowing him/her to focus on the likely candidates. Concrete efforts are under way to build an appropriate graphical user interface for our system and to extend our study, in collaboration with historians, to another, larger set of medieval documents concerning Dutch nobility titles.

The purpose of the current paper was to directly test, on historical documents, the functioning of the underlying pattern recognition engine. Our statistical methods generated robust and stable results: combining textural and allographic features yields usable writer identification rates.

7. Acknowledgement

The authors thank professor John Daugman (University of Cambridge, UK) and professor Linne Mooney (University of York, UK) who kindly provided the historical dataset on which the current study was performed.

References

[1] `http://www.cl.cam.ac.uk/~jgd1000/scribes.html` .

[2] A. Bensefia, T. Paquet, and L. Heutte. Information retrieval based writer identification. In *Proc. of 7th ICDAR*, pp 946–950, Edinburgh, Scotland, 3-6 August 2003.

[3] S. Bres, V. Eglin, and C. Volpilhac-Auger. Evaluation of handwriting similarities using Hermite transform. In *Proc. of 10th IWFHR*, pp 575–580, La Baule, France, 23-26 October 2006.

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

[5] I. Dinstein and Y. Shapira. Ancient hebraic handwriting identification with run-length histograms. *IEEE Trans. Syst., Man and Cybernetics*, SMC-12(3):405–409, 1982.

[6] N. Otsu. A threshold selection method from gray-level histogram. *IEEE Trans. Syst., Man and Cybern.*, 9:62–69, 1979.

[7] A. Schlapbach and H. Bunke. Using HMM-based recognizers for writer identification and verification. In *Proc. of 9th IWFHR*, pp 167–172, Tokyo, Japan, 26-29 October 2004.

[8] L. Schomaker, K. Franke, and M. Bulacu. Using codebooks of fragmented connected-component contours in forensic and historic writer identification. *Pattern Recognition Letters*, 28(6):719–727, 15 April 2007.

[9] S. Srihari, S. Cha, H. Arora, and S. Lee. Individuality of handwriting. *J. of Forensic Sciences*, 47(4):1–17, July 2002.