# Automatic Writer Identification Using Fragmented Connected-Component Contours

Lambert Schomaker
AI Institute, Groningen,
The Netherlands,
schomaker@ai.rug.nl

Marius Bulacu
AI Institute, Groningen,
The Netherlands,
bulacu@ai.rug.nl

Katrin Franke
Fraunhofer IPK, Berlin,
Germany
franke@ipk.fhg.de

## Abstract

*In this paper, a method for off-line writer identification is presented, using the contours of fragmented connected-components in mixed-style handwritten samples of limited size. The writer is considered to be characterized by a stochastic pattern generator, producing a family of character fragments (fraglets). Using a codebook of such fraglets from an independent training set, the probability distribution of fraglet contours was computed for an independent test set. Results revealed a high sensitivity of the fraglet histogram in identifying individual writers on the basis of a paragraph of text. Large-scale experiments on the optimal size of Kohonen maps of fraglet contours were performed, showing usable classification rates within a non-critical range of Kohonen map dimensions. Further validation experiments on variable-sized random subsets from an independent set of 215 writers gives additional support for the proposed method. The proposed automatic approach bridges the gap between image-statistics approaches and manual character-based methods.*

## 1. Introduction

Writer identification on the basis of optically scanned handwritten samples enjoys a renewed interest [14, 5, 10, 9]. The *target performance* for writer-identification systems is a near-100% recall of the correct writer in a hit list of one hundred writers, computed from a database in the order of $10^4$ samples, the size of search sets in current European forensic databases. Recently, we have proposed the use of connected-component contours ($CO^3s$) and their occurrence histogram, i.e., discrete PDF, as a writer identification feature [12] in upper-case Western handwriting. In this approach, a codebook of $CO^3s$ was constructed with a Kohonen self-organized map on the basis of a sufficiently large sample set of upper-case script. The writer is assumed to act as a stochastic generator of ink-blob shapes, such that the probability distribution of shape usage is characteristic of each writer. The performance of this approach is very promising, especially if it is used in conjunction with a complementary feature set which is based on edge-directional histograms which cover yet another aspect of writing style [3]. However, large collections of handwritten samples usually contain a mixture of upper case, isolated hand print, connected-cursive and mixed-style script. Therefore, it would be most convenient if the $CO^3$ codebook approach could be generalized to free-style handwriting.

Isolated connected components (ink blobs) in upper-case handwriting are large in number but limited in complexity when compared to connected components which are present in cursive and mixed-style scripts. For cursive-script images, the construction of a $CO^3$ codebook by a Kohonen self-organizing map would amount to the storage of complete word and syllable patterns. This is undesirable from the point of view of writer identification, since the text content is a confounding factor. It seems clear that a robust segmentation into small ink objects is needed, yielding a compound writing-style characterization similar to the successful case of the upper-case $CO^3$ PDF as a writer feature. Thus, the main goal of the current paper is to test whether heuristic fragmentation of connected components in cursive and mixed-style script will allow for the construction of a PDF of fragmented connected-component contours ($FCO^3$) such that reliable writer identification is still possible. Furthermore, we will explore the needed code-book size and the sensitivity of the approach to the number of known writers in the comparison set, given an sample of unknown writer identity.

It is useful to make a distinction between four factors which cause variability in handwriting [11, 12]: *affine transforms*; *neuro-biomechanical variability*; *sequencing variability* and *allographic variation*. The fourth factor, *allographic variation*, refers to the phenomenon of writer-specific character shapes, which produces most of the

Proceedings of the 9th Int'l Workshop on Frontiers in Handwriting Recognition (IWFHR-9 2004)
0-7695-2187-8/04 $20.00 © 2004 IEEE

problems in automatic script recognition but at the same time provides useful information for automatic writer identification. In this paper, we will show how writer-specific allographic shape variation present in handwritten Western script allows for effective writer identification. A more thorough description of the rationale behind the approach is given in [12]. It is assumed that each writer produces a recognizable set of allographs, due to schooling and personal preferences. This implies that a histogram of used allographs would characterize each writer, and given a sufficient number of allographs in a text, such a histogram of allographic usage could function as a feature vector in writer identification. However, there exists no exhaustive and world-wide accepted list of allographs in Western handwriting. The problem then, is to generate automatically a codebook, which sufficiently captures allographic information in samples of handwriting, given a histogram of the usage of its elements. Since automatic segmentation into characters is an unsolved problem, we would need, additionally, a reliable method to segment handwritten samples to yield components for such a codebook. It was demonstrated that the use of the shape of connected components of upper-case Western handwriting (i.e., not using allographs but the contours of their constituting connected components) as the basis for codebook construction can yield high writer-identification performance. On the basis of these results in writer identification on upper-case handwriting, the natural step is to explore the possibilities of the approach in free, connected-cursive styles. Here, the connected components may encompass several characters or syllables. Therefore, a fragmentation of the ink trace would be necessary, yielding broken connected components (fraglets), the ensemble of which still captures the shape details of the allographs emitted by the writer. Fortunately there are several heuristics which might deliver the proper fragmentation of connected components. An example of a possible method (”SegM”, segment on Y-minima) is based on segmentation at each vertical lower-contour minimum which is one ink-trace width away from a corresponding vertical minimum in the upper contour of the connected component under scrutiny. A similar method of segmentation is known to be useful in the text recognition of connected-cursive script [1, 4]. In our case, for each vertical minimum in the lower contour, the nearest minimum in the upper contour is searched. If the path between these minima has a length in the order of the ink-trace width and covers a minimum amount of black (ink) pixels, a cut is generated in the trace such that the connected component may be fragmented (Fig. 1). The resulting fraglets will usually be of character size or smaller. Sometimes a fraglet will contain more than one letter. Another possible method is based on a segmentation at points of strong directional

**Figure 1. Fragmentation on the basis of proximal minima in the vertical contour (method ”SegM”). The Euclidean distance between the upper and lower minima in the XY-plane must be in the order of the ink-trace width.**

change, i.e., mostly at points of high curvature ”SegS” [6]. Both fragmentation heuristics, SegM and SegS can be supported by a rationale from the domain of handwriting movement control. Here, we will report the results for the SegM method. The basic question is whether sub-allographic fraglets might be usable for writer identification on the basis of free-style handwriting. This needs to be demonstrated empirically, as has been done in the case of original unbroken connected-component segmentation for upper-case script [12].

In the processing pipeline of the proposed method, the use of domain-specific heuristics is kept to a minimum. There are no rule-based image enhancements. The amount of image and contour normalizations will be kept to a minimum, as well. Simple distance computation will be used, avoiding expensive usage of weights (as in multi-layer perceptron or support-vector machine based trained similarity functions). The proposed approach is size invariant. However, in the case of forged handwriting, the forger tries to change the handwriting style, usually by changing the slant and/or the chosen allographs. In the current approach, we assume that all samples are produced with a comparable *natural* writing attitude.

### 1.1. Research questions

A number of questions will be addressed in the experimental evaluation hereafter. The main question is whether the concept of a PDF of basic shape occurrence can be used in cursive and mixed writing styles as has been shown in the case of upper-case script. We will explore a fragmentation heuristic for connected components in free styles, i.e., a segmentation based on vertical-minima in the contour which coincide with a corresponding vertical minimum in the upper contour (SegM). An important parameter to be explored concerns the dimensions of the Kohonen map for optimal writer identification, i.e., the required code-book size. Finally, the performance of the

writer identification as a function of the number of writers in a reference base will be studied.

## 2. Methods

### 2.1. Data

The Firemaker[1] set is a database of handwritten pages of 250 writers, four pages per writer: Page1 contains a *Copied* text in natural writing style; Page2 contains copied *Upper-case* text; Page3 contains copied *Forged* text (”please write as if to impersonate another person”) while Page4 contains a self-generated description of a cartoon image in *Free* writing style. The text content and amount of written ink varies considerably per writer in this latter page. All pages were scanned at 300 dpi gray-scale, on lineated paper with a vanishing line color. The text to be copied has been designed in forensic praxis to cover a sufficient amount of different letters from the alphabet while remaining conveniently writable for the majority of suspects. Of 100 writers which were set apart for system training purposes, the pages 1,3 and 4, i.e., the pages with mixed-style content, were used for determining a codebook (Kohonen self-organized map) of fragmented connected-component contours ($FCO^3$s). Page 2, copied upper case, was not used in the training. Data from the remaining set of 150 other writers were used for testing writer identification. Apart from the Firemaker data, a separate image set which was derived from the Unipen [7] collection was used, containing two paragraphs of text for each of 215 writers. This latter set is used to determine the effects of writer-set size on a multinational collection which is remote in content and (technical) origin from the Firemaker reference set. The experimental procedure is as follows:

$for$ a range of Kohonen network sizes NxN, where $N \in [2, 50]$ {

- Compute a single codebook of Fragmented Connected-component Contours ($FCO^3$s) for 100 writers, 3 pages each) by means of the Kohonen self-organized map
- Compute writer-specific feature vectors $P(FCO^3)$ using this codebook
- Evaluate writer-identification performance (150 other writers, split-page tests)

}

1 This data set was collected thanks to a grant of the Netherlands Forensic Institute for the NICI Institute, Nijmegen, Schomaker & Vuurpijl, 2000

### 2.2. Stage one: Computing a codebook of fragmented connected-component contours

The images of 100x3 pages were processed in order to extract the fragmented connected components representing the handwritten ink. The gray-scale image was blurred using a 3x3 flat smoothing window and subsequently binarized using the mid-point gray value. For each connected component, its contour was computed using Moore’s algorithm, starting at the left-most pixel in a counter-clockwise fashion. The resulting contour-coordinate sequence was resampled to contain 100 (X, Y) coordinate pairs. Subsequently, a fragmentation method (here, SegM) is applied to the connected components, using heuristics as described above. After applying the fragmentation, the original connected components are broken into several fraglets. For each fraglet, the Moore contour was computed, once again. The resulting fixed-dimensional (N=200) vector will be dubbed Fragmented COnnected-COmponent COntour ($FCO^3$). Figure 2, panel 2x2, shows the general shape of such patterns.

The 300 pages in the training set yielded 152k $FCO^3s$ using the SegM heuristic. The fragmented connected-component contour training set was presented to a Kohonen [8] self-organizing feature map (SOFM) as described elsewhere [12], using 500 epochs and a fast cooling schedule for learning rate and network bubble radius. Network size was varied from 2x2 to 50x50. Training was performed on a Beowolf high-performance Linux cluster with 128 nodes. Computing time varied from 7 hours (2x2 SOFM) to 122 hours (50x50 SOFM). Results are based on a total of 3000 cpu hours on 1.7GHz/0.5GB machines. The computational complexity is $O[N_{epochs} * N_{samples} * N_{cells} * N_{(X,Y)}]$.

At the end of training the resulting SOFM contained the patterns as shown in Figure 2. Each network is considered to constitute the codebook $C$ necessary for computing the writer-specific $FCO^3$ emission probabilities used for writer identification, as described earlier. Writer-identification performance levels will become interesting at codebooks of 15x15 and larger (cf. Fig. 3).

### 2.3. Stage two: Computing writer-specific feature vectors

The writer is considered as a signal-source generator of a finite number of basic patterns. In the current study, such a basic pattern consists of a $FCO^3$. An individual writer is assumed to be characterized by the discrete probability-density function for the emission of the basic patterns. Consequently, from a database of 150 writers, for each of the writers, a histogram was computed of the occurrence of the nodes in the Kohonen SOFM of $FCO^3$s in his/her

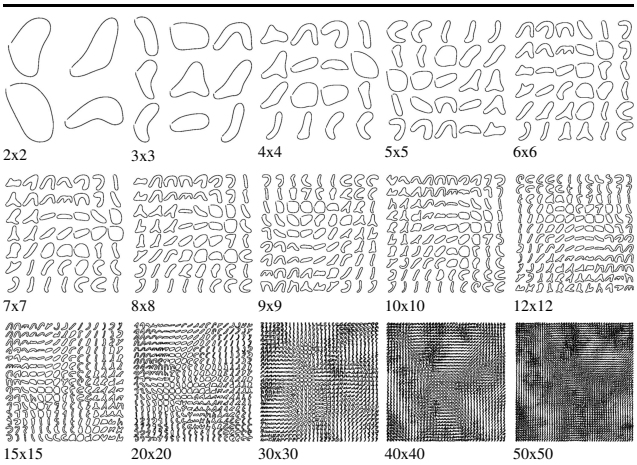


**Figure 2. A selection of Kohonen Self-organized maps of Fragmented Connected-Component Contours from the SegM(inima) fragmentation heuristic. Networks vary in size from 2x2 cells (upper left) to 50x50 feature vector cells (bottom right). Training data consisted of 300 pages by 100 different writers (152k sample vectors). Each $FCO^3$ is normalized in size to fit its cell.**

handwriting, as determined by Euclidean nearest-neighbor search of a handwritten $FCO^3$ to the patterns which are present in the SOFM. The pseudo-code for the algorithm is as follows:

$$
\begin{aligned}
&\vec{\xi} \leftarrow 0 \\
&\text{forall} \quad i \in \mathcal{K} \\
&\{ \\
&\qquad \vec{x}_i \leftarrow (\vec{x}_i - \mu_x)/\sigma_r \\
&\qquad \vec{y}_i \leftarrow (\vec{y}_i - \mu_y)/\sigma_r \\
&\qquad \vec{f_i} \leftarrow (X_{i1}, Y_{i1}, X_{i2}, Y_{i2} ..., X_{i100}, Y_{i100}) \\
&\qquad k \leftarrow argmin_l \, ||\vec{f_i} - \vec{\lambda}_l|| \\
&\qquad \Xi_k \leftarrow \Xi_k + 1/N \\
&\}
\end{aligned}
$$

Notation: $\vec{\xi}$ is the PDF of $FCO^3$s, $\mathcal{K}$ is the set of fragmented connected components in the sample. Scalar vector elements are shown as indexed upper-case capitals. Steps: First, the PDF is initialized to zero. Then each fragmented connected-component contour $(\vec{x}_i, \vec{y}_i)$ is normalized to an origin of $0, 0$ and a standard deviation of radius $\sigma_r = 1$, as reported elsewhere [13]. The $FCO^3$ vector $\vec{f_i}$ consists of the X and Y values of the normalized contour resampled to 100 points. In the table of pre-normalized Kohonen SOFM vectors $\lambda$, the index $k$ of the Euclidean nearest neighbor of $\vec{f_i}$ is sought and the corresponding value in the PDF $\Xi_k$ is updated ($N = |\mathcal{K}|$) to obtain, finally, $\vec{\xi}$, i.e., $p(FCO^3)$. This PDF is assumed to be a writer descriptor containing the connected-component shape-emission probability for characters by a given writer.

### 2.4. Stage three: Writer identification

Each of the 150 paragraphs of the 150 writers is divided into a top half (set $A$) and a bottom half (set $B$). Writer descriptors $p(FCO^3)$ are computed for set $A$ and $B$. For each writer sample $u$, its Hamming distance to all samples $v \neq u$ was computed where $v, u \in A \cup B$ (leave one out). A sorted hit list of samples $v_i$ with increasing distance to the query $u$ was constructed.

## 3. Results

As regards nearest-neighbor search, we will report the results on the Hamming distance only: Use of the Chi-square distance function [12] produced similar results, while Euclidean, Bhattacharya and Minkowski$_3$ distances performed much worse.

Figure 3 shows the Top-1 writer-identification performance as a function of Kohonen self-organized map dimensions. A point represents from 7hrs (2x2) to 122 hrs (50x50 network) training time time. However, training is an infrequent processing step. Performances are stable for Kohonen maps of dimension 15x15 units or larger. The highest performance is reached for the "Copied" text category: Using the 33x33 codebook as the measuring stick (cf. [12]), a Top-1 performance of 97% is reached. The performance of the "Upper case" category shows the generalization (70%) of a codebook trained on mixed lower-case styles to queries which are fully written in upper-case letters. The "Free" text

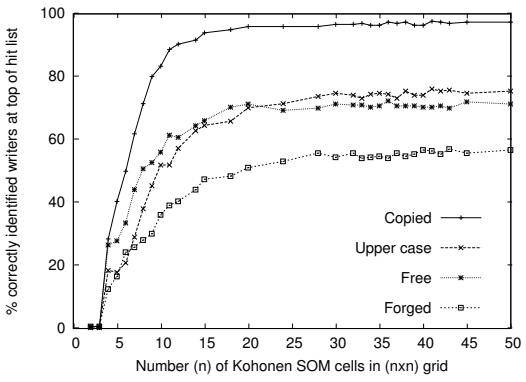


**Figure 3. Top-1 writer-identification performance as a function of Kohonen map dimensions (performance is % of correct writer identification at the first position of the hit list).**

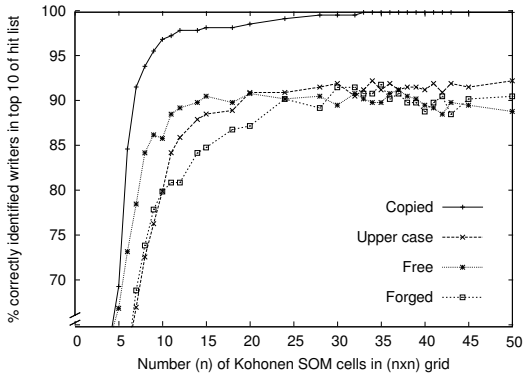


**Figure 4. Top-10 hit list performance (please note: the vertical axis is broken) as a function of Kohonen self-organized map dimensions (performance is % of correct writer identification in the Top-10 of the classifier hit list.**

category displays a similar performance (70%) which might be attributed to both the smaller number of characters and its variable text content. As was to be expected, the variability in the "Forged" category is highest, which can be inferred from a lower identification performance (50%). The number of writers in the reference set is 150, the number of distractor samples to a single query is 300-1=299 paragraphs of text.

Figure 4 displays the Top-10 writer-identification performance as a function of Kohonen self-organized map dimensions. As can be seen, the likelihood of finding the correct writer in a hit list of 10 best matching samples approach 100%, for Kohonen self-organized maps of 30x30 or larger, for the "Copied" set. The asymptote for the other categories, "Upper case", "Forged" and "Free" is about 90%. The number of writers in the reference set is 150, the number of distractor samples to a single query is 300-1=299 paragraphs of text.

In order to estimate the influence of the number of writers, a test was performed on a set of 210 writers. Images where derived from the Unipen database. The on-line $x_k, y_k$ coordinates where transformed to a simulated 300-dpi image using a Bresenham line generator and an appropriate brushing function. For each size of the writer set, ten tests on random selections of writers were performed up to 210 writers. The total set contains 215 writers, such that the randomness of sampling is reduced for larger set sizes. The results show a consistent but not dramatic decrease in performance on this data, starting at an average of about 95% on 10 writers and decreasing to 83% Top-1 performance on 210 writers (420-1=419 paragraphs of text) (Figure 5).

As an additional experiment, we adjoined the present feature vector with an edge-directional feature ("hinge") as reported elsewhere [12, 3]. By using a normalization of each PDF feature dimension and using Hamming distance, a Top-1 performance of 97% (Top-5: 99%; Top-10: 99.7%) could be reached on the *Copied* data set, as a "best result ever" exercise on the 150-writer (300 paragraph) set. Table 1 shows the results for features reported elsewhere on the same dataset (the size of the writer set varies among those experiments). Only the method "split hinge", i.e. computing edge-curvature histograms for the upper and lower parts of lines, separately, displays a performance which is in the same ballpark as the method proposed here.

| *Method/Feature* | *Nwriters* | *Top-1* (%) | *Top-10* (%) | Ref. |
|---|---|---|---|---|
| SysA | 100 | 34 | 90 | [12] |
| SysB | 100 | 65 | 90 | [12] |
| splitEdge | 250 | 29 | 69 | [2] |
| splitAla | 250 | 64 | 86 | [2] |
| splitHinge | 250 | 79 | 96 | [2] |

**Table 1. Performances of other features on data set "Copied".**

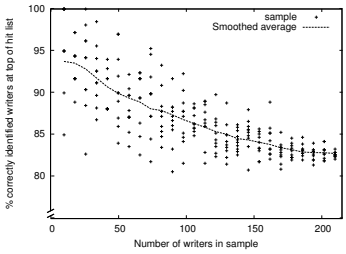


**Figure 5. Top-1 writer-identification performance as a function number of writers. Random writer subsets up to N=210 writers were generated, using ten tries per set size.**

## 4. Discussion

Results indicate that the use of fragmented connected-component contour shapes in writer identification on the basis mixed-style script yields valuable results. We think that the reason for this resides in the fact that writing style is largely determined by allographic shape variations. Small style elements which are present within a character are the result of the writer's physiological make up as well as education and personal preference. Experiences on style variation in on-line handwriting recognition show evidence that the amount of shape information at the

level of the characters is increasing as a function of the number of writers [16]. It should be noted that the essence of our method does not seem to be located in an exhaustive enumeration of all possible connected-component allographic part shapes. Rather, the $FCO^3$ codebook spans up a shape space by providing a finite set of nearest-neighbor attractors for the set of connected-component contours within a given handwritten sample. In literature, similar approaches are currently being reported. For example, in [1], normalized bitmap fragments are used, in conjunction with a clustering method for determining a base set of shapes, in an Information Retrieval framework. Future work will need to be directed at evaluating the differences between this image-based and our contour-based approach, including the use of other distance measures. As we have shown here and previously [12], the combination of character-shape elements and image properties such as the edge-hinge angular probability distribution function will yield enhanced classification rates.

We have presented a new approach which uses a connected-component contour codebook for the characterization of a writer of mixed-style Western letters. The use of the fragmented connected-component contour ($FCO^3$) codebook and its probability-density function of shape usage has a number of advantages. No manual measuring on text details is necessary, representing an advantage over interactive forensic feature determination. The feature is largely size invariant. A codebook has to be computed over a large set of samples from a wide range of writers, but this is an infrequent processing stage. Writer-identification performance on this new feature is promising, and could be improved using better distance measures. The $FCO^3$ approach itself is in principle generic and could easily be applied to other, non-Western scripts. We think that automatic approaches in this application domain will allow for convenient search in large sample databases, with less human intervention than is current practice. By reducing the size of a target set of writers, a detailed manual and microscopic forensic analysis becomes feasible. It is important to note also the recent advances [14, 15] that have been made at the detailed allographic level, requiring, however some form of detailed user interaction. In any case, in the foreseeable future, the tool box of the forensic expert will have been thoroughly modernized and extended.

## References


[1] A. Bensefia, T. Paquet, and L. Heutte. Information retrieval-based writer identification. In *7th International Conference on Document Analysis and Recognition (ICDAR 2003), 3-6 August 2003, Edinburgh, Scotland, UK*, pages 946–950. IEEE Computer Society, 2003.

[2] M. Bulacu and L. Schomaker. Writer style from oriented edge fragments. In *Proc. of the 10th Int. Conference on Computer Analysis of Images and Patterns*, pages 460–469, 2003.

[3] M. Bulacu, L. Schomaker, and L. Vuurpijl. Writer identification using edge-based directional features. In *Proc. of ICDAR'2003: International Conference on Document Analysis and Recognition*, pages 937–941. IEEE Computer Society, 2003.

[4] A. El-Yacoubi, R. Sabourin, C. Y. Suen, and M. Gilloux. An hmm-based approach for off-line unconstrained handwritten word modeling and recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 21(8):752–760, 1999.

[5] K. Franke and M. Köppen. A computer-based system to support forensic studies on handwritten documents. *International Journal on Document Analysis and Recognition*, 3(4):218–231, 2001.

[6] K. Franke, Y.-N. Zhang, and M. Köppen. Static signature verification employing a Kosko-Neuro-Fuzzy approach. In N. Pal and M. Sugeno, editors, *Advances in Soft Computing - AFSS 2002, LNAI 2275*, pages 185–190. Springer Verlag, 2002.

[7] I. Guyon, L. Schomaker, R. Plamondon, R. Liberman, and S. Janet. Unipen project of on-line data exchange and recognizer benchmarks. In *Proceedings of the 12th International Conference on Pattern Recognition, ICPR'94*, pages 29–33, Jerusalem, Israel, October 1994. IAPR-IEEE.

[8] T. Kohonen. *Self-Organization and Associative Memory*. Springer Verlag, Berlin, second edition, 1988.

[9] U.-V. Marti, R. Messerli, and H. Bunke. Writer identification using text line based features. In *Proc. of the Sixth International Conference on Document Analysis and Recognition (ICDAR '01)*, pages 101–105. IEEE Computer Society, 2001.

[10] H. Said, T. Tan, and K. Baker. Writer identification based on handwriting. *Pattern Recognition*, 33(1):133–148, 2000.

[11] L. Schomaker. From handwriting analysis to pen-computer applications. *IEE Electronics Communication Engineering Journal*, 10(3):93–102, 1998.

[12] L. Schomaker and M. Bulacu. Automatic writer identification using connected-component contours and edge-based features of upper-case western script. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(6):787–798, 2004.

[13] L. R. B. Schomaker. Using Stroke- or Character-based Self-organizing Maps in the Recognition of On-line, Connected Cursive Script. *Pattern Recognition*, 26(3):443–450, 1993.

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

[15] M. van Erp, L. Vuurpijl, K. Franke, and L. Schomaker. The WANDA measurement tool for forensic document examination. In *Proc. of the IGS'2003, Scottsdale, Arizona*, pages 282–285, 2003.

[16] L. Vuurpijl, L. Schomaker, and V. Erp. Architecture for detecting and solving conflicts: two-stage classification and support vector classifiers. *International Journal of Document Analysis and Recognition*, 5(4):213 – 223, 2003.