Available online at www.sciencedirect.com
ScienceDirect
Pattern Recognition Letters 28 (2007) 719–727
Pattern Recognition Letters
www.elsevier.com/locate/patrec

Using codebooks of fragmented connected-component contours in forensic and historic writer identification

Lambert Schomaker [a,*], Katrin Franke [b], Marius Bulacu [a]

[a] AI Institute, University of Groningen, Grote Kruisstr. 2/1, NL-9712 TS, Groningen, The Netherlands
[b] Fraunhofer IPK, Pascalstr. 8-9, D-10587, Berlin, Germany

Available online 11 September 2006

Abstract

Recent advances in 'off-line' writer identification allow for new applications in handwritten text retrieval from archives of scanned historical documents. This paper describes new algorithms for forensic or historical writer identification, using the contours of fragmented connected-components in free-style handwriting. 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. The proposed automatic approach bridges the gap between image-statistics approaches and purely knowledge-based manual character-based methods.

Keywords: Writer identification; Author identification; Cursive-script segmentation

1. Introduction

Writer identification on the basis of optically scanned handwritten samples enjoys a renewed interest (Srihari et al., 2002; Franke and Köppen, 2001; Said et al., 2000; Marti et al., 2001). The goal is to find in a large database a sample of a known writer (author) on the basis of an unknown or questioned handwritten document sample. The *target performance* for forensic writer-identification systems is a near-100% recall of the correct writer in a hit list of 100 writers, computed from a database in the order of 10^4 samples, the size of search sets in current European forensic databases. Another application which enjoys increased interest is writer verification. Here, the goal is to develop systems which are able to decide whether two handwritten samples are from the same writer or not. In the domain of the cultural heritage, writer identification and verification are becoming a realistic tool in information retrieval methods. Additionally, interesting new applications are emerging in this domain. Due to the fact that writing style of an individual author evolves over time, attempts are currently made at dating handwritten samples of a writer whose style evolution may be present in a large scanned archive of samples with a known date of writing (Bensefia et al., 2003). Examples are the scanned collections of manuscripts and letter correspondence by authors such as Zola and Flaubert (Bensefia et al., 2003). The manuscripts in such collections are often annotated in a handwritten script of which the author may not be the same person as the main, original author. Also here, automatic writer identification may act as a useful tool for humanities researchers. Fig. 1 shows a sample from an administrative Dutch collection, with handwriting of one particular scribe.

Clearly, these new applications necessitate the development of powerful shape descriptors of free-style handwriting which are designed to capture individual style

* Corresponding author. Tel.: +31 50 3636687.
E-mail address: schomaker@ai.rug.nl (L. Schomaker).

0167-8655/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.patrec.2006.08.005

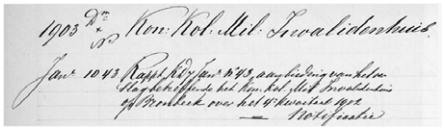

Fig. 1. An example of a paragraph from the Dutch National Archief (Kabinet der Koningin). Writer-identification tools will allow to search for particular scribes in a huge collection.

information. The problem is complex at a number of levels: (1) the degree of variability and variation of script; (2) the problem of foreground/background segmentation in highly textured and smudged documents; (3) the limited amount of text in unknown samples; (4) the differences in scanning technologies and image preprocessing. As a consequence, in forensic practice, a combination of statistical and knowledge-based techniques is used (Franke and Köppen, 2001). We have developed an ontology and XML format (WandaXML) for the systematic processing of forensic handwritten samples (Wanda, 2004). Elements of systematic style categorization can be entered in such a system to aid in boosting the performance of the pattern classification algorithms. It is to be expected that applications in historical writer identification and verification will similarly require a hybrid approach. In this paper, however, we will mainly focus on recent progress at the level of feature extraction in automatic, image-based (i.e., off-line) methods for writer identification.

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 (Schomaker and Bulacu, 2004) 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 (Bulacu et al., 2003). Fig. 2 shows a number of connected-component contours. Table 1 shows the raw identification rates in a set of 150

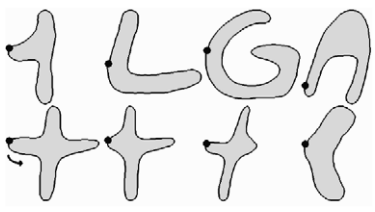

Fig. 2. A number of connected-component contours (CO^3s), with the body displayed in gray, and the starting point for the counter-clockwise contour coordinates (black border) depicted with black discs. Note that inner contours such as in the A-shape, upper right, are not incorporated in the CO^3 vector.

Table 1
Nearest-neighbor writer-identification performance in % of correct writers, as a function of hit-list size (χ^2 distance), for basic feature f0 (edge orientation histogram) and the histogram (f1) of connected-component contour patterns in upper-case script

Feature hit-list size:	1	2	3	4	5	6	7	8	9	10
f0: $p(\phi)$	34	45	54	60	66	71	73	75	78	79
f1: $p(CO^3)$	72	78	83	85	88	89	91	91	92	93

The 95% confidence limits are ±3.5% for $N = 150$ at a performance of 95%.

writers, on the basis of a paragraph, comparing a basic edge-directional histogram feature (f0) and the proposed contour-based method (f1). Fig. 3 shows an example of an application of the method to upper-case script samples. Comparisons with other methods have been reported (Schomaker and Bulacu, 2004) and the proposed method appears to perform very well.

In spite of these promising results, a problem remains. 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 from upper-case style to free-style handwriting. However, isolated connected components (ink blobs) in upper-case handwriting are large in number but limited in complexity when compared to connected components

Fig. 3. An example of a successful hit list of a writer-identification method based on the histogram of connected-component contour shapes $P(CO^3)$ in upper-case Western handwriting (Schomaker and Bulacu, 2004). The query sample is at the top. The nearest neighbor is the sample directly below it, which is correctly from the same writer. The distance value increases with left-to-right reading order down the hit list.

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 a 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 in free-style script, a reliable writer identification is possible with similar performances as has been measured in the case of upper-case script samples. Furthermore, we will explore the code-book size parameter, the sensitivity of the approach to the number of reference writers in the comparison set, given an sample of unknown writer identity. Finally, we will also address the issue of small script samples and propose a method to improve writer-identification reliability.

1.1. Allographic style characterization which avoids letter segmentation

It is useful to make a distinction between four factors which cause variability in handwriting (Schomaker, 1998, 2004): *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 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 (Schomaker and Bulacu, 2004) (see Fig. 4).

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

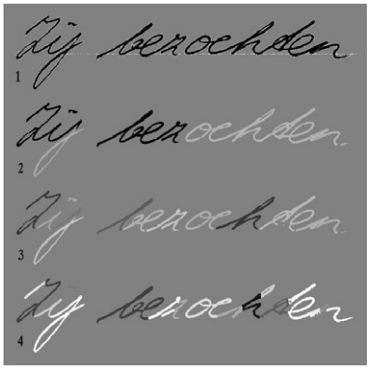

Fig. 4. Fragmentation methods: (1) raw input; (2) connected-components of cursive word parts; (3) fraglets based on related vertical minima in the lower and upper contours (method "SegM"); (4) fraglets based on "shadow" fragmentation (method SegS) (Franke et al., 2002). Method SegM will be evaluated in the current paper.

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 (Bensefia et al., 2003; El-Yacoubi et al., 1999). 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. 5). The resulting fraglets will usually be of character size or smaller. Sometimes a fraglet will contain more than one letter. Other methods are possible, such as fragmentation at points of strong directional change (Franke et al., 2002). However, in this study we will focus on a fragmentation based on spatial minima to find out whether the resulting sub-allographic fraglets might be as usable for writer identification on the basis of free-style handwriting as the unbroken connected-components are in the case of upper-case script (Schomaker and Bulacu, 2004).

Fig. 5. 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. The characters represent the first four letters of the Dutch word *veilingen* ("auctions"). The method is similar to segmentation approaches in (Bensefia et al., 2003; El-Yacoubi et al., 1999).

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 ("lease 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 writers. 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^3s). Page2, 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 (Guyon et al., 1994) 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 $N \times N$, where $N \in [2, 50]$ {

- compute a single codebook of fragmented connected-component contours (FCO^3s) for 100 writers, three pages each) by means of the Kohonen self-organized map;
- compute writer-specific feature vectors $P(FCO^3)$ using this $N \times N$ codebook;
- evaluate writer-identification performance (150 other writers, split-page tests).

}

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

The images of 100×3 pages were processed in order to extract the fragmented connected components representing the handwritten ink. The gray-scale image was blurred using a 3×3 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, the fragmentation method is applied to the connected components, using a heuristic 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).

The 300 pages in the training set yielded 152 k FCO^3s using the SegM heuristic. The fragmented connected-component contour training set was presented to a Kohonen (Kohonen, 1988) self-organizing feature map (SOM) as described elsewhere (Schomaker and Bulacu, 2004), using 500 epochs and a fast cooling schedule for learning rate and network bubble radius. Network size was varied from 2×2 to 50×50. Training was performed on a Beowolf high-performance Linux cluster with 128 nodes. Computing time varied from 7 h (2×2 SOM) to 122 h (50×50 SOM). Results are based on a total of 3000 cpu hours on 1.7 GHz/0.5 GB machines. The computational complexity is $O[N_{\text{epochs}} * N_{\text{samples}} * N_{\text{cells}} * N_{(X,Y)}]$.

At the end of training the resulting SOM contained the patterns as shown in Fig. 6. Each network is considered to constitute the codebook 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 15×15 and larger (cf. Fig. 7).

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 SOM of FCO^3s in his/her handwriting, as determined by Euclidean nearest-neighbor search

[1] This data set was collected thanks to a grant of the Netherlands Forensic Institute for the NICI Institute, Nijmegen.

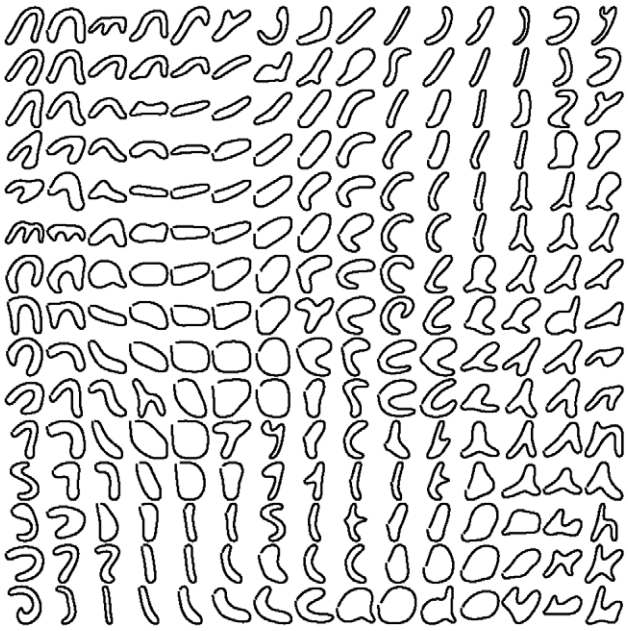

Fig. 6. A Kohonen self-organized map (SOM) of fragmented connected-component contours from the SegM(inima) fragmentation heuristic. The network size of 15×15 was selected for display because writer-identification performances start to be useful at this dimension and contour details of all cells can still be discerned. In the evaluation, network size varied from 2×2 to 50×50 feature-vector cells. Training data consisted of 300 pages by 100 different writers (152 k sample vectors). Each contour is normalized in size to fit its cell.

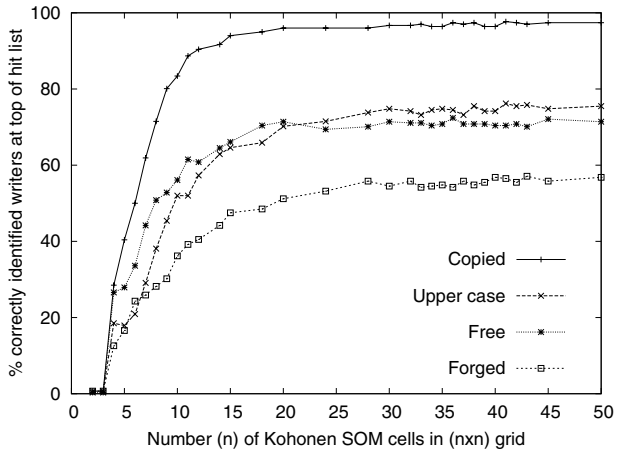

Fig. 7. Top-1 writer-identification performance as a function of Kohonen map dimensions (performance is % of correct writer identification at the first position of the hit list).

of a handwritten FCO^3 to the patterns which are present in the SOM. The pseudo-code for the algorithm is as follows:

$\vec{\xi} \leftarrow 0$
forall $i \in K$
{
$\vec{x}_i \leftarrow (\vec{x}_i - \mu_x)/\sigma_r$
$\vec{y}_i \leftarrow (\vec{y}_i - \mu_y)/\sigma_r$
$\vec{f}_i \leftarrow (X_{i1}, Y_{i1}, X_{i2}, Y_{i2} \ldots, X_{i100}, Y_{i100})$
$k \leftarrow \text{argmin}_l, \|\vec{f}_i - \vec{\lambda}_l\|$
$\Xi_k \leftarrow \Xi_k + 1/N$
}

Notation: $\vec{\xi}$ is the PDF of FCO^{3}s, 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 (Schomaker, 1993). 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 SOM vectors λ, the index k of the Euclidean nearest neighbor of $\vec{f}_i$ is sought and the corresponding value in the PDF Ξ_k is updated ($N = |K|$) to obtain, finally, $\vec{\xi}$, i.e., $p(\text{FCO}^3)$. 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 $p(\text{FCO}^3)$ 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 (Schomaker and Bulacu, 2004) produced similar results, while Euclidean, Bhattacharya and Minkowski$_3$ distances performed much worse. Fig. 7 shows the Top-1 writer-identification performance as a function of Kohonen self-organized map dimensions. A point represents from 7 h (2×2) to 122 h (50×50 network) training time. However, training is an infrequent processing step. Performances are stable for Kohonen maps of dimension 15×15 units or larger. The highest performance is reached for the "Copied" text category: Using the 33×33 codebook as the measuring stick (cf. Schomaker and Bulacu, 2004), 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 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 $300 - 2 = 298$ paragraphs of text.

Fig. 8 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 30×30 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, 10 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) (Fig. 9).

As an additional experiment, we adjoined the present feature vector with an edge-directional feature ("hinge") as reported elsewhere (Schomaker and Bulacu, 2004; Bulacu et al., 2003). 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 2 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.

Table 3 shows performances of a number of features on the upper-case data set, in leave-one-out mode. Feature *e* represents a one-dimensional feature, i.e., the number of

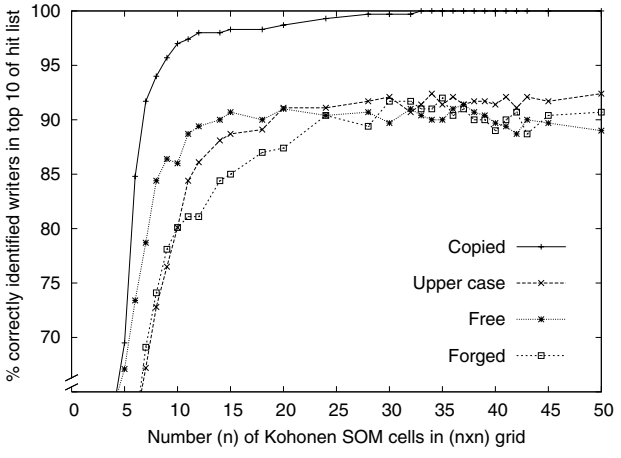

Fig. 8. Top-10 hit list performance (please note: the vertical axis is broken) as a function of Kohonen self-organized map dimensions (performance is % of correct writer identification in the Top-10 of the classifier hit list).

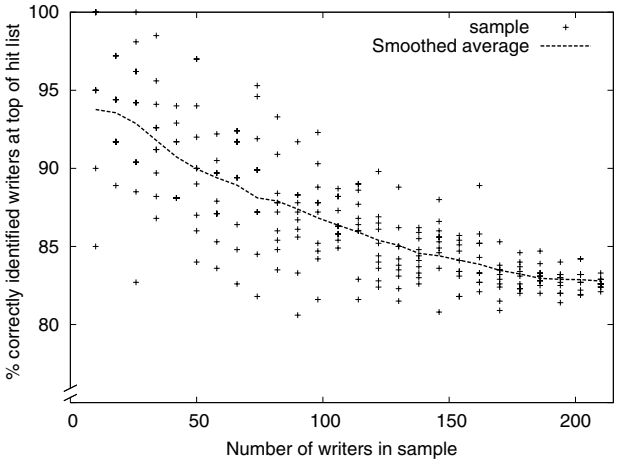

Fig. 9. Top-1 writer-identification performance as a function number of writers. Random writer subsets up to N = 210 writers were generated, using ten tries per set size.

Table 2
Performances of other features on data set "Copied"

Method/ Feature	Nwriters	Top-1 (%)	Top-10 (%)	Ref.
SysA	100	34	90	Schomaker and Bulacu (2004)
SysB	100	65	90	Schomaker and Bulacu (2004)
splitEdge	250	29	69	Bulacu and Schomaker (2003)
splitAla	250	64	86	Bulacu and Schomaker (2003)
splitHinge	250	79	96	Bulacu and Schomaker (2003)

Table 3
Nearest-neighbor performance of other features on upper-case script: leave-one out (1 vs. 299 samples), $N = 150$ writers, as before

Feature	Description	Ndim	Top-1 (%)	Top-10 (%)
e	Normalized entropy	1	2	19
*w*1	Wavelets, Haar	99	5	14
*w*2	Wavelets, Odegard	99	14	28
*w*3	Wavelets, Daubechies 14	99	15	29
*w*4	Wavelets, Villasenor 2	99	15	30
v	Vertical run-length PDF	100	21	61
r	Horizontal autocorrelation	100	25	61
h	Horizontal run-length PDF	100	26	66
f0	Edge-angular PDF	16	34	79
b	Brush feature, 15×15	225	69	93
f1	CO^3 PDF	1089	72	93
f2	Hinge-angular PDF	464	80	97

Given are the dimensionality Ndim of the feature vectors and the Top-1 and Top-10 percentages of the correct writer found in a sorted hit list of size 1 and 10, respectively.

bytes in the Lempel–Ziv compressed 1-byte gray-scale image of a paragraph sample, divided by the number of black (ink) pixels after contrast normalization. This simple feature with a value range of 2–15 bits/inkpixel provides a baseline performance well above chance level (Top-10: 19%). The wavelet-based features ($w1 - w4$) are computed on the basis of Davis' wavelet package (Davis and Nosratinia, 1998), using coefficients $HL_1, HH_1, LH_1, \ldots, HL_{11}$, HH_{11}, LH_{11}, yielding 33 rectangles with coefficients per paragraph of written text. For each coefficient rectangle, the relative energy, skew and kurtosis were computed, yielding a 99-dimensional feature vector. Only best results per feature group are shown, such as Daubechies 14 (Table 3, *g*). The performance of the wavelet (energy and distribution) features is low. It may be predicted that compute-intensive Gabor wavelets (not tested) may perform better than the 'technical' wavelets used here, as Gabor wavelets are more similar to our edge-angular features. However, it is as yet unclear whether the periodicity in the Gabor wavelet would provide an additional source of information in writer identification. Features *v, r, h, b* are described elsewhere (Bulacu et al., 2003; Schomaker et al., 2003). The "brush" feature (Schomaker et al., 2003) shows an interesting performance (Top-1: 69%). However, unlike the features proposed in the current paper, the brush feature requires that the same type of pen is used for writing the known and unknown sample, due to its focus on the ink deposition pattern at stroke endings. Clearly, such a feature will not be applicable in historical collections where a single writer uses different types of writing implements.

3.1. Smearing of fraglet occurrences over the Kohonen codebook

A conspicuous characteristic of the performance on the "Copied lower case" condition in comparison to the other script types is its high performance (Figs. 7 and 8). Although this level of performance may be due to (1) the more regular handwriting style during copying text as well as (2) a better fit of the codebook content with this type of data, there is (3) a third important factor which plays a role in explaining this difference. The "Copied" lower-case text contains 126 words, whereas upper-case, forged and self-generated free-style samples contained 65, 45 and 59 ± 16 words respectively. Note that a sample contains about half of these amounts of words. In comparison to the dimensionality of the codebook, the amount of data is limited and a smoothing method on the histograms seems appropriate. In order to increase the reliability of the histogram (PDF), the occurrence of a nearest-neighbour FCO^3 was smeared out over N_{smear} Kohonen cells in the codebook, where N_{smear} is a control parameter. Neighbourhood is defined in shape space, not in the network topology. In collecting the counts for the codebook histogram, not only the best candidate but the set of N_{smear} neighbours receives a tally in this procedure. An unseen test set of 215 writers, two paragraphs/writer, was used on a codebook which

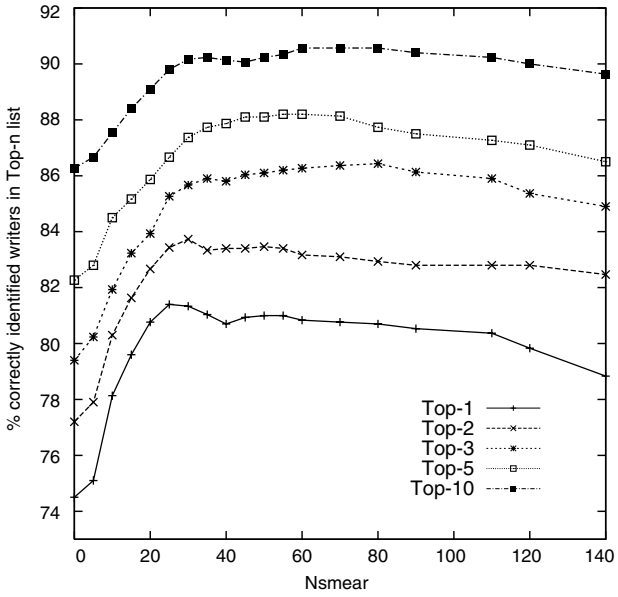

Fig. 10. Top-*n* writer-identification performance as a function number of the Nsmear parameter, which allows for smearing fraglet occurrence over its neighbourhood in shape space. The use of such smearing may increase the robustness of writer identification up to neighbourhood sizes of up to 80 Kohonen cells (7% of 33×33) as can be seen in this case of unseen, new data ($N_{writers} = 215$) of variable-content samples.

was trained on samples from 100 writers, four pages each, to provide a free-style codebook. The N_{smear} parameter was varied and Top-*n* identification performances were measured. The case $N_{smear} = 0$ corresponds to "single nearest-neighbour only", as in (Schomaker and Bulacu, 2004).

Results indicate (Fig. 10) that "smearing" of shape occurrence over the codebook, which increases the probability of overlap between similar histograms may improve the raw results ($N_{smear} = 0$) considerably. As an example, $N_{smear} = 30$ raises the Top-1 performance from 71% to 82%, while the Top-10 performance increases from 86% to 91%. Such increments are statistically significant ($N = 429$, $\alpha = 0.05$).

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 (Vuurpijl et al., 2003). 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. This interpretation is supported by the observation that a smearing of fragment occurrences over their neighbourhood in the codebook may actually improve rather than deteriorate identification performance, as one might expect with such a smooth operator. In literature, similar code-book approaches are currently being reported. For example, in (Bensefia et al., 2003), normalized bitmap fragments are used, in conjunction with a clustering method for determining a base set of shapes, in an information retrieval framework. More work is needed to evaluate the differences between this image-based and our contour-based approach. As we have shown here and previously (Schomaker and Bulacu, 2004), the combination of character-shape elements and image properties such as the edge-hinge angular probability distribution function will yield further enhanced classification rates. It is important to note also the recent advances in writer identification (Srihari et al., 2002; van Erp et al., 2003) that have been made at the detailed allographic level. Such methods, however, require some form of detailed and elaborate user interaction, contrary to the method proposed here.

5. Conclusion

We have presented an overview of recently developed methods which use 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 histogram of usage has a number of advantages. No detailed manual measuring on text details is necessary, representing an advantage over interactive methods in forensic feature determination. This convenience can be exploited in the case of writer retrieval from historical collections, as well. The contour-based 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. The FCO^3 approach itself is, in principle, generic and could easily be applied to other, non-Western scripts. Automatic approaches in this application domain will allow for convenient search in large sample databases. By reducing the size of a target set of writers to a manageable dimension, a detailed analysis becomes feasible. Although the approach described in this paper is of a statistical nature, its relation with knowledge-based approaches is twofold. In the first place, the design of the algorithm is inspired by a long tradition of handwriting recognition research, using explicit knowledge-based allographic modeling (Schomaker, 1993; Vuurpijl and Schomaker, 1997). In the second place, the approach easily allows for a partitioned training of codebooks for particular styles, particular historical periods, or particular life stages of an individual author. Future work will be directed at two areas. In the historical archive applications, writer verification may be even more important than identification. In order to achieve this, a detailed analysis on the distributions of distances within and between classes needs to be undertaken. It is not guaranteed a priori that a good feature for identification purposes will produce similar results in verification. However, the proposed feature appears to extract useful style-specific information. A second area of research will be directed at an analysis of the sensitivity of this methods with respect to the amount of text. In the current paper, we propose a smearing method for the codebook usage probability distribution. We will compare this approach to using the Kullback–Leibler distance measure which may compensate for an unbalance in the reliability of the probability estimates (Bensefia et al., 2003). Current research concerns large sets of writers ($N > 900$). Fresh data collection processes with forensic and cultural-heritage institutions are in progress.

References

Bensefia, A., Paquet, T., Heutte, L., 2003. Information retrieval-based writer identification. In: 7th Internat. Conf. on Document Analysis and Recognition (ICDAR 2003), 3–6 August 2003, Edinburgh, Scotland, UK. IEEE Computer Society, pp. 946–950.

Bulacu, M., Schomaker, L., Vuurpijl, L., 2003. Writer identification using edge-based directional features. In: Proc. ICDAR'2003: Internat. Conf. on Document Analysis and Recognition. IEEE Computer Society, pp. 937–941.

Bulacu, M., Schomaker, L., 2003. Writer style from oriented edge fragments. In: Proc. 10th Internat. Conf. on Computer Analysis of Images and Patterns, pp. 460–469.

Davis, G., Nosratinia, A., 1998. Wavelet-based image coding: an overview. Applied and Computational Control, Signals, and Circuits 1 (1), 205–269.

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

Franke, K., Köppen, M., 2001. A computer-based system to support forensic studies on handwritten documents. Internat. J. Document Anal. Recognition 3 (4), 218–231.

Franke, K., Zhang, Y.-N., Köppen, M., 2002. Static signature verification employing a Kosko-Neuro-Fuzzy approach. In: Pal, N., Sugeno, M. (Eds.), Advances in Soft Computing—AFSS 2002, LNAI, 2275. Springer-Verlag, pp. 185–190.

Guyon, I., Schomaker, L., Plamondon, R., Liberman, R., Janet, S., 1994. Unipen project of on-line data exchange and recognizer benchmarks, In: Proc. 12th Internat. Conf. on Pattern Recognition, ICPR'94, IAPR-IEEE, Jerusalem, Israel, pp. 29–33.

Kohonen, T., 1988. Self-Organization and Associative Memory, second ed. Springer Verlag, Berlin.

Marti, U.-V., Messerli, R., Bunke, H., 2001. Writer identification using text line based features. In: Proc. 6th Internat. Conf. on Document Analysis and Recognition (ICDAR '01). IEEE Computer Society, pp. 101–105.

Said, H., Tan, T., Baker, K., 2000. Writer identification based on handwriting. Pattern Recognition 33 (1), 133–148.

Schomaker, L.R.B., 1993. Using stroke- or character-based self-organizing maps in the recognition of on-line, connected-cursive script. Pattern Recognition 26 (3), 443–450.

Schomaker, L., 1998. From handwriting analysis to pen-computer applications. IEE Electron. Commun. Eng. J. 10 (3), 93–102.

Schomaker, L., Bulacu, M., 2004. Automatic writer identification using connected-component contours and edge-based features of upper-case

western script. IEEE Trans. Pattern Anal. Machine Intell. 26 (6), 787–798.

Schomaker, L., Bulacu, M., van Erp, M., 2003. Sparse-parametric writer identification using heterogeneous feature groups. In: Proc. IEEE Internat. Conf. on Image Processing(ICIP'03), vol. I. IEEE Computer Society, pp. 545–548 (I).

Srihari, S., Cha, S., Arora, H., Lee, S., 2002. Individuality of handwriting. J. Forensic Sci. 47 (4), 1–17.

van Erp, M., Vuurpijl, L., Franke, K., Schomaker, L., 2003. The WANDA measurement tool for forensic document examination. In: Proc. IGS'2003, Scottsdale, Arizona, pp. 282–285.

Vuurpijl, L., Schomaker, L., 1997. Finding Structure in Diversity: A Hierarchical Clustering Method for the Categorization of Allographs in Handwriting. In: ICDAR. IEEE Computer Society, pp. 387–393.

Vuurpijl, L., Schomaker, L., Erp, V., 2003. Architecture for detecting and solving conflicts: two-stage classification and support vector classifiers. Internat. J. Document Anal. Recognition 5 (4), 213–223.

Wanda: A generic framework applied in forensic handwriting analysis and writer identification. In: Proc. 9th IWFHR, Tokyo, Japan, IEEE Computer Society, 2004.