Automatic Writer Identification Using Connected-Component Contours and Edge-Based Features of Uppercase Western Script

Lambert Schomaker, *Member*, *IEEE*, and Marius Bulacu, *Student Member*, *IEEE*

Abstract—In this paper, a new technique for offline writer identification is presented, using connected-component contours (COCOCOs or CO^3s) in uppercase handwritten samples. In our model, the writer is considered to be characterized by a stochastic pattern generator, producing a family of connected components for the uppercase character set. Using a codebook of CO^3s from an independent training set of 100 writers, the probability-density function (PDF) of CO^3s was computed for an independent test set containing 150 unseen writers. Results revealed a high-sensitivity of the CO^3 PDF for identifying individual writers on the basis of a single sentence of uppercase characters. The proposed automatic approach bridges the gap between image-statistics approaches on one end and manually measured allograph features of individual characters on the other end. Combining the CO^3 PDF with an independent edge-based orientation and curvature PDF yielded very high correct identification rates.

Index Terms—Writer identification, connected-component contours, edge-orientation features, stochastic allograph emission model.

1 Introduction

Automatic, offline writer identification enjoys a renewed interest [1], [2], [3], [4], [5]. Leading a worrisome life among the "harder" forms of biometric person identification such as DNA typing [6], [7], fingerprint classification [8], [9], and iris identification [10], it appears that the identification of a person on the basis of a handwritten sample still remains a useful application. Contrary to other forms of biometric person identification used in forensic labs, automatic writer identification often allows for determining identity in conjunction with the intentional aspects of a crime, such as in the case of threat letters. This is a fundamental difference from other biometric methods, where the relation between the evidence material and the details of an offense can be quite remote. The **target performance** for writer-identification systems is less impressive than is the case in DNA or iris-based person identification. In forensic writer identification, as a rule of thumb, one strives for a near-100 percent 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. A hit-list size of one hundred suspects is based on the pragmatic consideration that such a number of cases is just about manageable in the criminal-investigation process.

Recent advances in image processing, pattern classification, and computer technology at large allow for a substantial improvement of current procedures in forensic practice. There exist three groups of script-shape features which are derived from scanned handwritten samples in forensic procedures:

1. Fully automatic features computed from a region of interest (ROI) in the image.
2. Interactively measured features by human experts using a dedicated graphical user-interface tool.
3. Character-based features which are related to the allograph subset which is being generated by each writer.

Of these features, the first group has been treated with some skepticism by practitioners within the application domain, given the complexity of real-life scanned samples of handwriting which are collected in practice. Indeed, automatic foreground/background separation will often fail on the smudged and texture-rich fragments, where the ink trace is often hard to identify. However, there are recent advances in image processing using "soft computing" methods, i.e., combining tools from fuzzy logic and genetic algorithms, which allow for advanced semi-interactive solutions to the foreground/background separation process [2]. Under these conditions, and assuming the presence of sufficient computing power, the use of automatically computed image features (group 1, above), is becoming feasible. Before dealing with the methods and results in detail, we will introduce the rationale and the general model of the proposed approach.

It is generally assumed that uppercase characters contain less writer-specific information than does, e.g., connected-cursive handwritten script. This assumption is corroborated by the observation that the automatic classification of uppercase isolated characters is easier than the recognition of connected cursive script. However, much of the difference in recognition performance between uppercase

• *The authors are with the AI Institute, Groningen University, Grote Kuisstraat 2/1, 9712 TS, Groningen, The Netherlands.*
E-mail: {schomaker, bulacu}@ai.rug.nl.

Manuscript received 23 May 2003; revised 7 Jan. 2004; accepted 10 Jan. 2004. Recommended for acceptance by V. Govindaraju.
For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference IEEECS Log Number TPAMI-0101-0503.

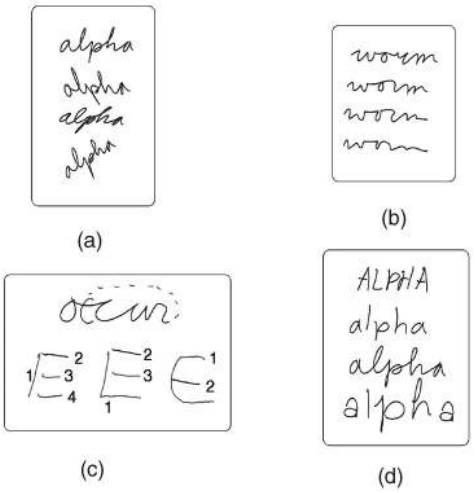

Fig. 1. Factors causing handwriting variability: (a) Affine transforms are under voluntary control. However, writing **slant** constitutes a habitual parameter which may be exploited in writer identification. (b) Neurobiomechanical variability refers to the amount of effort which is spent on overcoming the low-pass characteristics of the biomechanical limb by conscious cognitive motor control. (c) Sequencing variability becomes evident from stochastic variations in the production of the strokes in a capital *E* or of strokes in Chinese characters, as well as stroke variations due to slips of the pen. (d) Allographic variation refers to individual use of character shapes. Factors (b) and (c) represent system state more than system identity. In particular, **allographic variation** (d), is a most useful source of information in forensic writer identification.

characters versus free-style words can be attributed to the character segmentation problem, proper. Fig. 1 shows four factors causing variability in handwriting [11].

The first factor concerns the **affine transforms** (Fig. 1a), which are under voluntary control by the writer. Transformations of size, translation, rotation, and shear are a nuisance, but not a fundamental stumbling block in handwriting recognition or writer identification. In particular, *slant* (shear) constitutes a habitual parameter determined by pen grip and orientation of the wrist subsystem versus the fingers [12].

The second factor concerns the **neurobiomechanical variability** (Fig. 1b) which is sometimes referred to as "sloppiness space:" the local context and physiological state determines the amount of effort which is spent on character-shape formation and determines the legibility of the written sample. In realizing the intended shape, a writer must send motor-control patterns which compensate for the low-pass filtering effects of the biomechanical end effector. This category of variability sources also contains tremors and effects of psychotropic substances on motor-control processes in writing. As such, this factor is more related to system state than system identity.

The third factor is also highly dependent on the instantaneous system state during the handwriting process and is represented by **sequencing variability** (Fig. 1c): the stroke order may vary stochastically, as in the production of a capital *E*. A four-stroked *E* can be produced in $4! * 2^4 = 384$ permutations. In the production of some Asian scripts, such as Hanzi, stochastic stroke-order permutation are a well-known problem in handwriting recognition (even though the training of stroke order at schools is rather strict). Finally, spelling errors may occur and lead to posthoc editing strokes in the writing sequence. Although sequencing variability is generally assumed to pose a problem only for handwriting recognition based on temporal (online) signals, the example of posthoc editing (Fig. 1c) shows that static, optical effects are also a possible consequence of this form of variation.

The fourth factor, **allographic variation** (Fig. 1d), 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 the information for automatic writer identification. In this paper, we will show how writer-specific allographic shape variation present in handwritten uppercase characters allows for effective writer identification.

1.1 Theory

There exist two fundamental factors contributing to the individuality of script, i.e., allographic variation: *genetic* (biological) and *memetic* (cultural) factors.

The first fundamental factor consists of the *genetic* make up of the writer. Genetic factors are known or may be hypothesized to contribute to handwriting style individuality:

- The biomechanical structure of the hand, i.e., the relative sizes of the carpal bones of wrist and fingers and their influence on pen grip.
- The left or right handedness [13].
- Muscular strength, fatiguability, peripheral motor disorders [14].
- Central nervous system (CNS) properties, i.e., aptitude for fine motor control and the CNS stability in motor-task execution [15].

The second factor consists of *memetic* or culturally transferred influences [16] on pen-grip style and the character shapes (allographs) which are trained during education or are learned from observation of the writings of other persons. Although the term *memetic* is often used to describe the evolution of ideas and knowledge, there does not seem to be a fundamental objection to view the evolution and spreading of character shapes as a memetic process: the fitness function of a character shape depends on the conflicting influences of 1) legibility and 2) ease of production with the writing tools [17], which are available within a culture and society. The distribution of allographs over a writer population is heavily influenced by writing methods taught at school, which in turn depend on factors such as geographic distribution, religion, and school types. For example, in The Netherlands, allographic differences may be expected between Protestant and Catholic writers, writers of different generations, and immigrant writers.

Together, the genetic and memetic factors determine a habitual writing process, with recognizable shape elements at the local level in the writing trace, at the level of the character shape as a whole, and at the level of character placement and page layout. In this paper, we will focus on the local level in the handwritten trace and on the character level.

The writer produces a pen-tip trajectory on the writing surface in two dimensions (x, y), modulating the height of the pen tip above the surface by vertical movement (z). Displacement control is replaced by force control (F) at the moment of landing. The pen-tip trajectory in the air between two pen-down components contains valuable writer-specific information, but its shape is not known in the case of offline scanned handwritten samples. Similarly, pen-force information is highly informative of a writer's identity, but is not directly known from offline scans [18]. Finally, an important theoretical basis for the usage of

handwritten shapes for writer identification is the fact that handwriting is not a feed-back process which is largely governed by peripheral factors in the environment. Due to neural and neuromechanical propagation delays, a handwriting process based upon a continuous feed-back mechanism alone would evolve too slowly [19]. Hence, the brain is continuously planning series of ballistic movements ahead in time, i.e., in a feed-forward manner. A character is assumed to be produced by a "motor program" [20], i.e., a configurable movement-pattern generator which requires a number of parameter values to be specified before being triggered to produce a pen-tip movement yielding the character shape [21], [22], [23] by means of the ink deposits [24], [25]. Although the process described thus far is concerned with continuous variables such as displacement, velocity, and force control, the linguistic basis of handwriting allows for postulating a discrete symbol from an alphabet to which a given character shape refers.

1.2 A Model

Assume there exists a finite list S of allographs for a given alphabet L. Each allograph s_{li} is considered to be the ith allowable shape (style) variation of a letter $l \in L$ which should, in principle, be legible at the receiving end of the writer-reader communication line [26]. The source of allographic variation may be located in teaching methods and individual preferences. The human writer is thus considered to be a pattern generator, stochastically selecting each allograph shape s_{li} when a letter l is about to be written. It is assumed that the probability density function $p_w(S)$, i.e., the probability of allographs being emitted by writer w, will be informative in the identification of writer w if it holds that

$$w \neq v \Rightarrow p_w(S) \neq p_v(S), \tag{1}$$

where w and v denote writers, S is a common allograph codebook, and $p(.)$ represents the discrete PDF for allograph emission. This (1) will be realizable if, for handwritten samples u emitted by w and characterized by

$$\vec{x}_{wu} = p_{wu}(S), \tag{2}$$

and assuming that the sample u is representative

$$\vec{x}_{wu} \approx p_w(S), \tag{3}$$

it holds that

$$\forall a, b, c, w, v \neq w : \Delta(\vec{x}_{wa}, \vec{x}_{wb}) < \Delta(\vec{x}_{wa}, \vec{x}_{vc}), \tag{4}$$

where Δ is an appropriate distance function on PDFs $\vec{x}$, v, and w denote writers, as before, and a, b, c are handwriting-sample identifiers. Equation (4) states that, in feature space, the distance between any two samples of the same writer is smaller than the distance between any two samples by different writers. In ideal circumstances, this relation would always hold, leading to perfect writer identification. Note that, in this model (1), the implication is unidirectional: in case of forged handwriting, $p_w(S)$ does not equal $p_v(S)$, but writer w imposes as v ($w = v$).

A problem at this point is that an exhaustive list S of allographs for a particular script and alphabet is difficult to obtain in order to implement this stochastic allograph-emission model. Clustering of character shapes with a known letter label is possible and has been realized [27]. However, the amount of handwritten image data for which no character ground truth exists vastly exceeds the size of commercial and academic training sets which are labeled at the level of individual characters. At this point in time, a commonly accepted list of handwritten allographs (and their commonly accepted names, e.g., in Latin, such as in the classification of species in the field of biology) does not exist, as yet. In this respect, it is noteworthy that for machine-print fonts, with their minute shape differences in comparison to handwriting variation, named font categories exist (e.g., Times-Roman, Helvetica, etc.), whereas we do not use generally agreed names for handwritten character families.

Therefore, it would be conducive to use an approach which avoids expensive character labeling at both training and operational stages. Contrary to character segmentation in handwriting, connected components can be detected reliably and in a nonparametric manner. The question then, is whether such suballographic text fragments might be usable for writer identification.

If each allograph s_{li} is composed of a nonempty set of connected components c_j, i.e., $s_{li} = \{c_1, c_2, \ldots, c_m\}$, then let us assume that a finite set or codebook C of connected components for all possible allographs can be estimated. If we assume, additionally, that the shape of a connected component is informative of the allographic character variant of which it is an element, then, for the probability function

$$\vec{\xi}_{wu} = p_{wu}(C) \tag{5}$$

of connected components derived from handwritten samples u by writer w it holds, analogously to (4), that

$$\forall a, b, c, w, v \neq w : \Delta(\vec{\xi}_{wa}, \vec{\xi}_{wb}) < \Delta(\vec{\xi}_{wa}, \vec{\xi}_{vc}) \tag{6}$$

again, under the assumption that samples u will be representative:

$$\vec{\xi}_{wu} \approx p_w(C), \tag{7}$$

which needs to be demonstrated empirically. A potential problem concerns the phenomenon of touching characters. For the approach proposed in this paper, this would not constitute a real problem if the tendency to produce connecting or overlapping letter combinations is typical for a writer. An exploration of the available data is needed in any case. In the next section, we will describe the construction of a connected-component codebook C, the computation of an estimate of the writer-specific pattern-emission PDF $p_w(C)$, and an appropriate distance function Δ for PDFs.

1.3 Design Considerations

In the application domain, a sparse-parametric approach has several advantages [28] because new data can easily be incorporated without retraining. In the current study, this goal is not met due to the use of a codebook which will be based on a self-organized map containing a considerable number of parameters. However, in the processing pipeline, 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 multilayer perceptron or support-vector machine based trained similarity functions). As regards the target application, it should be noted that the proposed approach is size invariant. However, in the case of forged handwriting, the forger tries to change the handwriting style, usually by

TABLE 1
Uppercase Dutch Text Containing All Letters of the Alphabet and All Digits

NADAT ZE IN NEW YORK, TOKYO, QUÉBEC, PARIJS, ZÜRICH EN OSLO WAREN GEWEEST, VLOGEN ZE UIT DE USA TERUG MET VLUCHT KL 658 OM 12 UUR.

ZE KWAMEN AAN IN DUBLIN OM 7 UUR EN IN AMSTERDAM OM 9.40 UUR 'S AVONDS. DE FIAT VAN BOB EN DE VW VAN DAVID STONDEN IN R3 VAN HET PARKEERTERREIN. HIERVOOR MOESTEN ZE HONDERD GULDEN (F 100,-) BETALEN.

changing the slant and/or the chosen allographs. Using detailed and manual analysis, forensic experts are sometimes able to correctly identify a forged handwritten sample. However, the proposed algorithm aims at recovering the correct known sample from a database for a query sample of which the writer is unknown, under the assumption that both were produced with a comparable and natural writing attitude.

2 Methods

2.1 Data

From the Firemaker[1] database of handwritten pages of 250 writers, "*Page 2*" was being used, i.e., the set which consisted of a copied text in uppercase handwriting. This text consists of two sentences, with a total of 65 text chunks, i.e., words and money amounts (Table 1), scanned at 300 dpi gray scale, on lineated paper with a vanishing line color (yellow). The number of words amounts to a paragraph of text. The text has been designed in forensic praxis to cover a sufficient amount of different letters from the alphabet while remaining writable for the majority of suspects. Fig. 2 shows a fragment of such a paragraph by a single writer.

A set of 100 paragraphs by as much writers was used for training purposes. The remaining set of 150 paragraphs by as much but different writers was used for testing writer identification. Processing entails three steps:

- Stage 1. Computing a codebook of Connected-component Contours in uppercase handwriting.
- Stage 2. Computing writer-specific feature vectors.
- Stage 3. Writer identification.

Whenever the word "feature" is used in the sequel, it should be interpreted as meaning "writer-feature vector."

2.2 Stage 1: Computing a Codebook of Connected-Component Contours in Uppercase Handwriting

The images of 100 paragraphs were processed in order to extract the connected components representing the handwritten ink. The gray-scale image was blurred using a 3×3 flat smoothing window and, subsequently, binarized using the midpoint 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. The resulting fixed-dimensional (N = 200) vector

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

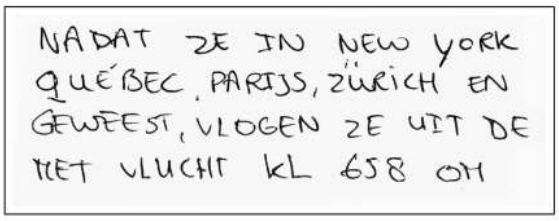

Fig. 2. An example fragment of a paragraph by one writer (female, age 22, right handed, Dutch, black ball-point pen).

will be dubbed COnnected-COmponent COntour (COCO-CO or CO^3). Fig. 3 shows a number of such patterns.

The 100 paragraphs yielded 26,896 CO^3s. These were presented to a Kohonen [29] self-organizing feature map (SOFM) of 33×33 (1,089) nodes, thus yielding an a priori uniform coverage of about 25 samples per Kohonen node. The goal of this procedure is to yield an accurate table of CO^3 shapes, rather than aiming at topology preservation. Hence, an ample number of 500 epochs was used to train the network. The network bubble size varied from a radius of 33 (full network) at the beginning of training, to 0 (one node) at the end of training. The learning rate was 0.9 at the beginning of training, ending at 0.015 at the end of training. Usually, in training Kohonen self-organizing maps, linear cooling schedules are used. However, if the goal is to obtain a veridical, least rms-error between the ensemble of possible patterns and the finite set of Kohonen cells, it has proved to be beneficial to use a steeply decaying temperature [30]. A Kohonen relaxation process can be roughly divided into three stages: 1) chaotic oscillation, 2) structural consolidation, and 3) fine tuning (Fig. 4). The use of a linear temperature cooling schedule is useful for obtaining maps with topology-preserving characteristics on a limited number of epochs. However, using a nonlinearly and steeply decaying function of bubble radius and learning rate results in a prolonged fine-tuning stage, yielding a reliable codebook after the presentation of a sufficiently large number of training epochs.

It should be noted that overfitting is not an issue here: In Kohonen self-organized maps, the degree of overfitting is mainly determined by the number of cells. Taking these considerations into account, a fast cooling schedule was used, on the basis of the following power function (8):

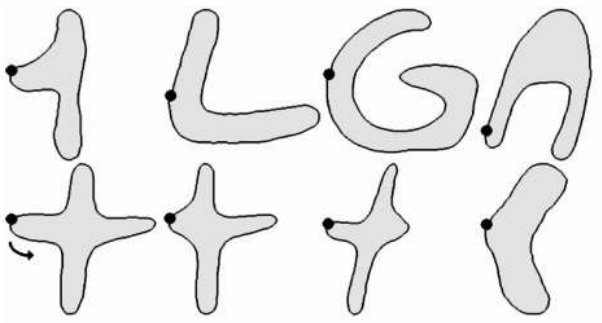

Fig. 3. A number of Connected-Component Contours (COCOCOs), with the body displayed in gray, and the starting point for the counter-clockwise contour coordinates (black border) depicted with black discs. Note that inner contours such as in the A-shape, upper right, are not incorporated in the CO^3 vector.

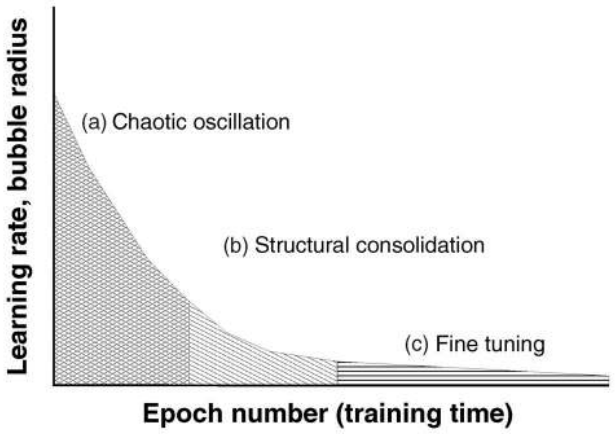

Fig. 4. Three conceptual "stages" during the training of a Kohonen self-organized map: (a) Chaotic oscillation, (b) structural consolidation, and (c) fine tuning. If the goal is to obtain a codebook vector set, the fine-tuning stage can be prolonged relative to the number of epochs by using a power function for the learning rate decay (8). This will lead to lower final rms error values than is the case in using a linear decay, provided an ample number of epochs is used.

$$r_k = \left((r_m^{1/s} - r_0^{1/s}) \frac{k}{m} + r_0^{1/s} \right)^s, \tag{8}$$

where $s(>0)$ is the steepness factor, r is a decreasing training parameter (here, learning rate or Kohonen bubble radius), $k = [0, m]$ is the epoch counter, and m is the last training epoch. If $s = 1$, r_k is a linear function. A steepness factor of $s = 5$ was used. This relatively high steepness speeds up the self-organizing process by reducing the duration of the initially irregular state space evolution.

At the end of training, the resulting SOFM contained the patterns as shown in Fig. 5. This table is considered to constitute the codebook C necessary for computing the writer-specific CO^3 emission probabilities used for writer identification, as described in Section 1. The training procedure lasted 28 hours and 19 minutes on a personal computer with a 600 MHz CPU. The computational complexity is $O[N_{epochs} * N_{samples} * N_{cells} * N_{(X,Y)}]$. The Kohonen training reduced the initial rms error of 0.036 per coordinate x or y of the contour to an rms error of 0.010 at 500 epochs. When using the resulting codebook for a nearest-neighbor search of connected-components contours of all writers, a PDF can be computed for this Kohonen network as a communication channel with $33 * 33 = 1,089$ discrete symbols, yielding an overall entropy of $\sum_{i=1}^{1,089} -\xi_i \, log(\xi_i) = 9.8$ bits.

2.3 Stage 2: Computing Writer-Specific Feature Vectors

Similar to an approach reported elsewhere [31], 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 CO^3. An individual writer is assumed to be characterized by the discrete probability-density function for the emission of the basic stroke 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 CO^3s in his/her handwriting, as determined by Euclidean nearest-neighbor search of a handwritten CO^3 to the patterns which are present in the SOFM. The pseudocode for the algorithm is as follows:

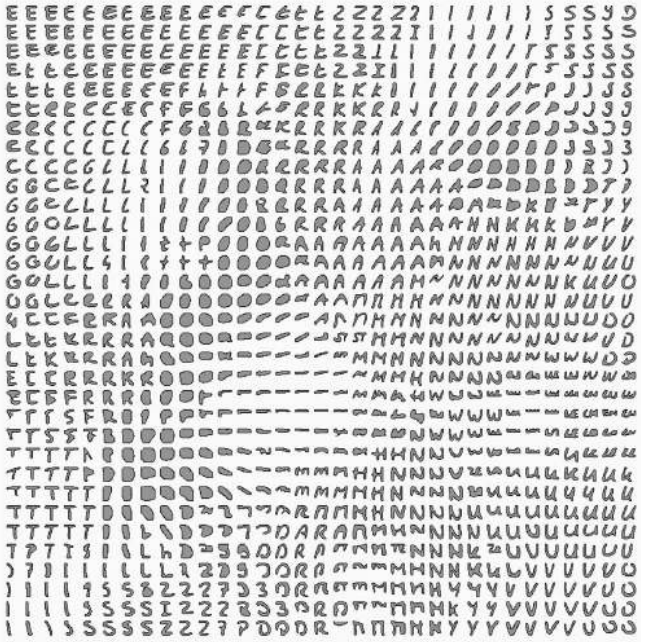

Fig. 5. A Kohonen self-organized map of 33×33 Connected-Component Contours (COCOCOs) from 26k samples, derived from the text written in Table 1 by 100 different writers. Some CO^3 represent whole uppercase characters whereas others represent character fragments. Each CO^3 is normalized in size to fit its cell.

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

Notation: $\vec{\xi}$ is the PDF of CO^3s, $\mathcal{K}$ is the set of detected connected components in the sample. Scalar vector elements are shown as indexed uppercase capitals. Steps: First, the PDF is initialized to zero. Then, each 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 [30], [32]. The CO^3 vector $\vec{f}_i$ consists of the X and Y values of the normalized contour resampled to 100 points. In the table of prenormalized Kohonen SOFM 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 ($N = |\mathcal{K}|$) to obtain, finally, $p(CO^3)$. This PDF is assumed to be a writer descriptor containing the connected-component shape-emission likelihood for uppercase characters, by a given writer (5).

2.4 Stage 3: 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 $p(CO^3)$ are computed for set A and B, separately, for each writer. Using the χ^2 distance measure (9), for each writer descriptor in set B, the nearest neighbor in set A was searched.

$$\chi_{ij}^2 = \sum_{k=1}^{n} \frac{(\xi_{ki} - \xi_{kj})^2}{\xi_{ki} + \xi_{kj}}, \tag{9}$$

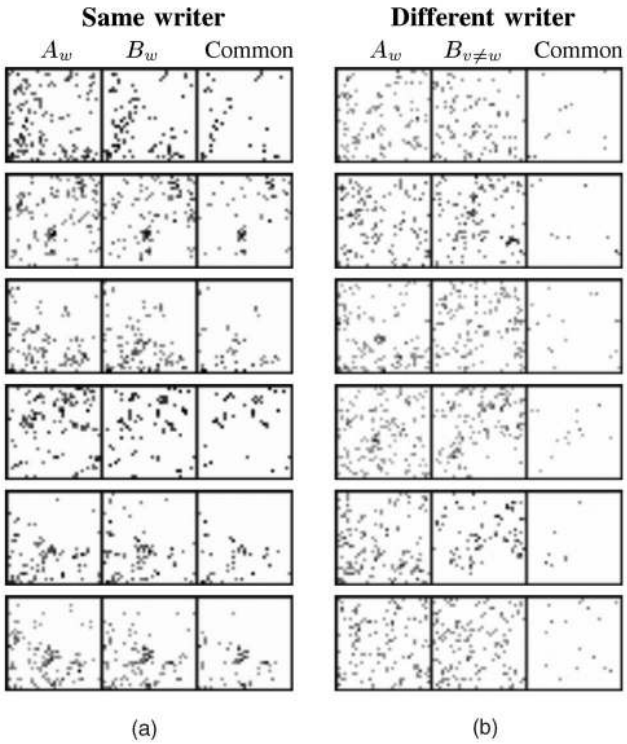

Fig. 6. Density plots of $p(CO^3)$. Each cell presents the probability density of the CO^3s in the 33×33 Kohonen codebook. The maximum probability is depicted in black, a probability of zero is represented as white. (a) shows, for a number of writers w (i.e., the rows), the densities for a set A (left column), a set B (middle column), and the densities of CO^3s which are both present in set A and B (right column, "Common"). (b) shows the densities for the case where A and B are samples from two different writers w and $v \neq w$, yielding a much lower density in the third column ("Common") than is the case in the left panel.

where i and j are sample indices, k is the bin index, n represents the number of bins in the PDF, and ξ represents the probability of a CO^3 codebook entry, as in (5).

The advantage of using the χ^2 distance measure is that differences in low-probability regions in the PDF are weighed more importantly than is the case in simple Euclidean, but also in the Bhattacharya distance measure for PDFs. Fig. 6 shows density plots for sample combinations originating from the same writer (Fig. 6a) or originating from two different writers (Fig. 6b). Samples were selected for this figure on the basis of visual clarity.

3 Results

Using an independent test set of N = 150 writers, a number of performance comparisons were performed. Tests will be organized as follows: For each writer from the test set, a paragraph labeled A and a distinct paragraph labeled B will be entered into the test. The purpose of testing is to find a corresponding paragraph B from a query A, and vice versa, for each writer. A single test on 150 writers was typically performed in 7s on a personal computer with a 600 MHz CPU (gcc, Linux). This corresponds to 328ms per sample. The computational complexity is $O[(N_{samples} - 1) * N_{cells} * N_{(X,Y)}]$ for a single-sample query.

The tests named "AB" refer to a leave-one out approach, where all A and B samples are lumped together in one set, taking a query sample out, one by one. This means that for an "A" query, the pair "B" sample written by the same subject will be the target, the distractors being the remaining 149 "A" samples and the 149 "B" samples of the other writers. Consequently, the "AB" sets constitute a reasonably-sized problem with 300-1 patterns to be searched in the set. The a priori hit probability thus equals $1/299$.

TABLE 2
An Overview of the Features Used in the Tests and Their Dimensionalities

Feature	*Name*	*PDF*	*Ndim*
f0	Edge directions	$p(\phi)$	16
f1	CO^3	$p(CO^3)$	1089
f2	Edge-hinge angles	$p(\phi_1, \phi_2)$	464
f1 $\cup$ f2	Combined feature vector	-	1553

The tests named "A versus B" are based on traditional disjoint sets, where the target set only contains a single sample from each writer. Consequently, the number of distractors for a query is much lower: 150-1, and the a priori probability of a hit equals $1/150$. As a consequence, the disjoint "A vs B" tests will yield better results than the more realistic leave-one out "AB" tests.

As a measure of base-line performance, the PDF of edge-orientation angles was used ("feature f0"), which is known to be an informative feature for writer and handwriting style identification [33], [34]. Then, the performance on our newly introduced feature $p(CO^3)$ ("feature f1") will be introduced. Finally, the performance of a recent edge-based orientation and curvature feature ("feature f2") will be presented, in isolation, and in combined use with "f1." Table 2 gives an overview of the features used. The edge-based features f0 and f2 will be explained in the next section.

3.1 Histogram (PDF) of Edge-Directions (Feature f0)

It has long been known from online handwriting research [35], [33] that the distribution of directions in handwritten traces, as a polar plot, yields useful information for writer identification or coarse writing-style classification [34].

We developed an offline and edge-based version of the directional distribution [36], [37], [28]. Computation of this feature starts with conventional edge detection: convolution with two orthogonal differential kernels (Sobel), followed by thresholding. This procedure generates a binary image in which only the edge pixels are "on." We then consider each edge pixel in the middle of a square neighborhood and we check, using the logical AND operator, in all directions emerging from the central pixel and ending on the periphery of the neighborhood for the presence of an entire edge fragment. Fig. 7 shows how the local angles are determined from the character edges. All the verified instances are counted into a histogram that is normalized to a probability distribution $p(\phi)$ which gives the probability of finding in the image an edge fragment oriented at the angle ϕ measured from the horizontal. In order to avoid redundancy, the algorithm only checks the upper two quadrants in the neighborhood because, without online information, we do not know which way the writer "traveled" along the found oriented edge fragment. The orientation is quantized in n directions, n being the number of bins in the histogram and the dimensionality of the feature vector. A number $n = 16$ directions (5-pixel long edge fragments) performed best and will be used in the test.

The distribution of the writing directions is characteristic of a writer's style. Using edges to extract it is a very effective

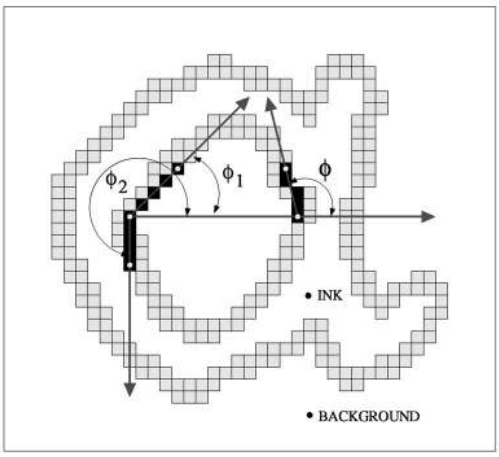

Fig. 7. Schematic description of the determination of edge orientation ϕ for feature f0 and edge-hinge orientations ϕ_1 and ϕ_2 for feature f2 on the edges of a character *a*. More details can be found in [36], [37], [28].

method because edges follow the written trace on both sides and they are thinner, effectively reducing the influence of trace thickness. As can be noticed in Fig. 8, the predominant direction in $p(\phi)$ corresponds, as expected, to the slant of writing. Even if idealized, the example shown can provide an idea about the "within-writer" variability and "between-writer" variability in f0 feature space.

We must mention an important practical detail: our generic edge detection does not generate one-pixel wide edges, but they can usually be one to three pixels wide and this introduces smoothing into the histogram computation because the "probing" edge fragment can fit into the edge strip in a few directions around a central main direction. This smoothing taking place in the pixel space has been found advantageous in our experiments.

Table 3, columns "f0" show the results for the edge feature "f0" or $p(\phi)$, using the χ^2 distance function, for hit lists of size 1 to 10. From a Top-1 performance of 34 percent on the leave-one out test, to a Top-10 performance of 79 percent can be expected for this simple feature (n = 299 samples, 150 writers). Using disjoint sets "A versus B," these performances are 55 percent to 90 percent, respectively (n = 150 samples, i.e., 150 writers). The use of Hamming distance yields comparable results, Euclidean distance yielded worse results.

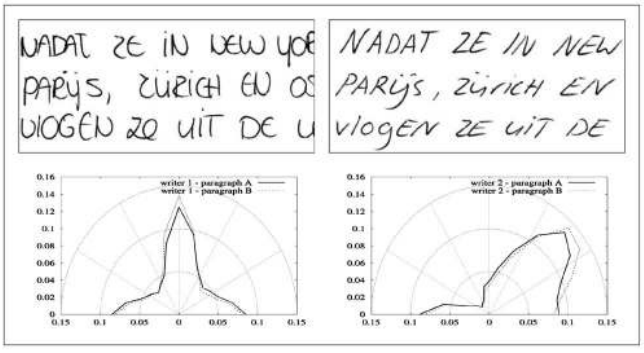

Fig. 8. Two uppercase handwriting samples from two different subjects. We superposed the polar diagrams of the edge-direction distribution $p(\phi)$ (feature f0) corresponding to paragraphs A and B contributed to our data set by each of the two subjects.

TABLE 3
Nearest-Neighbor Writer-Identification Performance in Percent of Correct Writers, as a Function of Hit-List Size (χ^2 Distance), for Basic Feature f0 (Edge Orientation PDF), and the Proposed PDF of Connected-Component Contour Pattern Presence

Hit list size	f0 $p(\phi)$ AB	f0 $p(\phi)$ A-vs-B	f1 $p(CO^3)$ AB	f1 $p(CO^3)$ A-vs-B
1	34	55	72	85
2	45	68	78	93
3	54	76	83	95
4	60	80	85	97
5	66	83	88	97
6	71	85	89	97
7	73	86	91	99
8	75	88	91	99
9	78	89	92	99
10	79	90	93	99

The 95 percent confidence limits are +/- 3.5 percent for N = 150 at a performance of 95 percent. The reader is referred to Table 2 and the text for further details.

3.2 Histogram (PDF) of CO^3s (Feature f1)

Subsequently, the identification performance on the CO^3 PDF was measured. The computation of this feature vector has been described in the Methods section. Table 3, columns "f1" show the results for the $p(CO^3)$ feature. Performances vary from Top-1 72 percent to a Top-10 rate of 93 percent for the leave-one out "AB" test. Again, disjoint sets yield a higher performance (Top-1 of 85 percent to Top-10 of 99 percent). Also, here, the use of Hamming distance yields comparable results, Euclidean distance yielded worse results. These results clearly outperform the simple edge-based feature "f0" and appear to be very promising. However, for use in the application domain, such results are not sufficient. The target performance indicated by forensic experts would be "99 percent probability of finding the correct writer in the Top-100 hit list, on a database of 20,000 samples." Therefore, the use of other orthogonal feature groups is necessary. Therefore, we will combine the "f1" feature with another edge-based feature that we recently developed [36]. This feature (f2) captures both writing slant and curvature, by estimating "hinge" angles along edges of script. Consequently, this complementary information will be expected to boost performances. Fig. 9 shows an example of a good hit list, with the target sample on the first position, and a homogeneous impression of script style. Fig. 10 shows an example of a hit list which does not contain the target sample, while the samples reveal a heterogeneity of style.

3.3 Histogram (PDF) of "Edge-Hinge" Angles (Feature f2)

In order to capture both slant and the curvature of the ink trace, which are known to be discriminatory between different writers, we have designed another feature [36], using local angles along the edges. The computation of this feature is similar to the computation of "f0," but it has added complexity. The central idea is to consider in the neighborhood, not one, but two edge fragments emerging from the central pixel and, subsequently, compute the joint probability distribution of the orientations of the two edge fragments constituting the legs of an imaginary hinge. All the instances found in the image are counted and the final normalized histogram gives the joint probability distribution $p(\phi_1, \phi_2)$

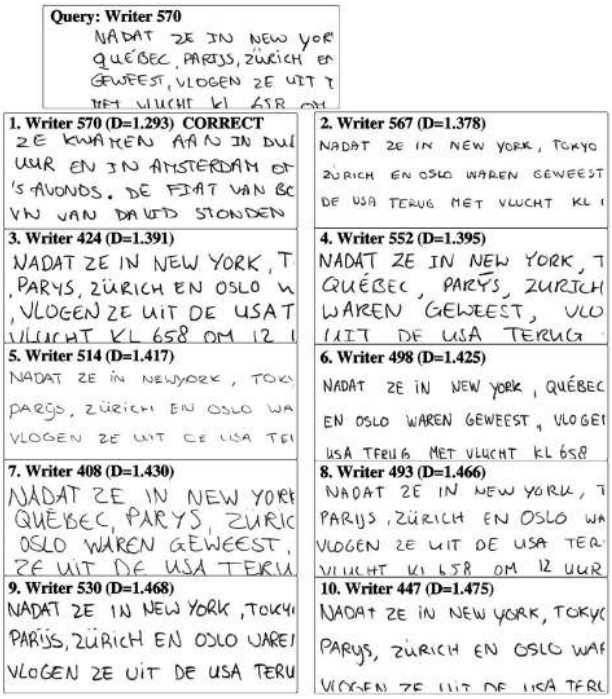

Fig. 9. An example of a successful hit list. The query sample is at the top. The nearest neighbor is the sample directly below it, which is correctly from the same writer. The distance value increases with left-to-right reading order down the hit list.

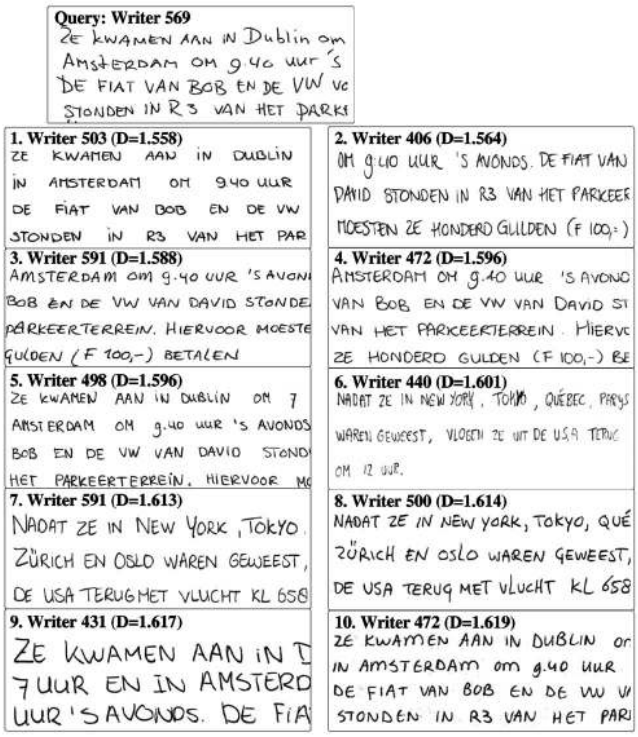

Fig. 10. An example of an unsuccessful hit list. The query sample is at the top. None of the nearest neighbors are from the correct writer of the query sample. The distance value increases with left-to-right reading order down the hit list. As can be seen, this query attracts samples from many styles. The probability of such an undesirable case is less than 7 percent for a hit list of 10 samples, assuming a 1 versus 299 test (cf. Table 3, column "f1").

quantifying the chance of finding in the image two "hinged" edge fragments oriented at the angles ϕ_1 and ϕ_2, respectively (see Fig. 7).

As already mentioned in the description of "f0," in our case, edges are usually wider than 1-pixel and, therefore, we have to impose an extra constraint: We require that the ends of the hinge legs should be separated by at least one "nonedge" pixel. This makes certain that the hinge is not positioned completely inside the same piece of the edge strip. This is an important detail, as we want to make sure that our feature properly describes the shapes of edges (and, implicitly, the shapes of handwriting at a low level) and avoids the senseless cases.

In contrast with feature "f0" for which spanning the upper two quadrants ($180°$) was sufficient, we now have to span all the four quadrants ($360°$) around the central junction pixel when assessing the angles of the two fragments. The orientation is now quantized in $2n$ directions for every leg of the "edge-hinge." From the total number of combinations of two angles ($4n^2$), we will consider only the nonredundant ones ($\phi_2 > \phi_1$) and we will also eliminate the cases when the ending pixels have a common side. The final number of combinations is $C_{2n}^2 - n = n(2n-3)$. For $n = 16$, the edge-hinge feature vector will have 464 dimensions.

In Table 4, columns "f2" display the writer-identification performance for the hinge feature vector. Clearly, this is a powerful feature. Its virtue resides in the local computation on the image and, as such, it can be directly applied also to cursive (lowercase) handwriting when character segmentation is very difficult. However, a weakness is the strong dependence on natural slant. The performance ranges from Top-1: 83 percent to Top-10 97 percent using the χ^2 distance measure on the leave-one out set "AB." Again, the disjoint-set test "A versus B" yields a higher performance (Top-1: 91 percent to Top-10: 98 percent). The Hamming distance delivers comparable results (Table 5). There seems to be a complementary behavior for these distance functions. Choosing the optimum for the Top-1 performance will yield a lower performance at the 10th position in the list or vice versa, depending on the choice of Hamming or χ^2 distance and the particular data set.

TABLE 4
Nearest-Neighbor Writer-Identification Performance in Percent of Correct Writers, as a Function of Hit-List Size (χ^2 Distance), for the Edge-Hinge Feature f1, and for a Combined Feature Vector of f1 and f2

Hit list size	f2 $p(\phi_1,\phi_2)$ AB	f2 $p(\phi_1,\phi_2)$ A-vs-B	f1 ∪ f2 $p(CO^3)$∪ $p(\phi_1,\phi_2)$ AB	f1 ∪ f2 $p(CO^3)$∪ $p(\phi_1,\phi_2)$ A-vs-B
1	83	91	81	94
2	89	95	88	97
3	93	97	92	99
4	95	97	93	99
5	96	97	95	99
6	96	97	96	99
7	97	97	97	99
8	97	97	97	99
9	97	97	98	100
10	97	98	98	100

The 95 percent confidence limits are +/- 3.5 percnet for N = 150 at a performance of 95 percent. The reader is referred to Table 2 and the text for further details.

TABLE 5
Nearest-Neighbor Writer-Identification Performance in Percent of Correct Writers, as a Function of Hit-List Size (Hamming Distance), for the Edge-Hinge Feature f1, and for a Combined Feature Vector of f1 and f2

Hit list size	f2 $p(\phi_1,\phi_2)$ AB	f2 $p(\phi_1,\phi_2)$ A-vs-B	f1 ∪ f2 $p(CO^3)\cup p(\phi_1,\phi_2)$ AB	f1 ∪ f2 $p(CO^3)\cup p(\phi_1,\phi_2)$ A-vs-B
1	80	87	87	95
2	86	92	92	98
3	91	96	95	98
4	93	97	96	99
5	93	97	97	99
6	95	97	97	99
7	96	98	97	99
8	96	98	97	99
9	97	98	98	99
10	97	98	98	99

The 95 percent confidence limits are +/- 3.5 percent for N = 150 at a performance of 95 percent. The reader is referred to Table 2 and the text for further details.

Finally, the effect of combining the main feature of this paper, the CO^3 PDF, with the hinge feature "f2" is tested and displayed in Table 4, columns $f1 \cup f2$, for the χ^2 distance measure, and similarly for the Hamming distance, Table 5. The combined feature vector used consists of an adjoined f1 and f2, yielding a 1,553-dimensional vector. No feature-group weighing has been performed. Extensive optimization tests yielded only marginal improvements. All axes are thus scaled as probabilities. The CO^3 dimensions outnumber the hinge dimensions with a ratio of roughly 2:1. For the χ^2 distance, the range of Top-1 to Top-10 performance is 81 percent to 98 percent for the leave-one out test "AB," and 94 percent to 100 percent for the disjoint test "A versus B." For the Hamming distance, the Top-1 results for leave-one out are better than the results for χ^2, i.e., 87 percent compared to 81 percent, with little difference at Top-10.

4 Discussion

Results indicate that the use of connected-component contour shapes in writer identification on the basis of uppercase script yields valuable results. We think that the reason for this resides in the fact that, ultimately, writing style is 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 online handwriting recognition show evidence that the amount of shape information at the level of the characters is increasing monotonously as a function of the number of writers (Fig. 1 in [38]). Other image properties which are determined by slant, curvature, and character placement yield additional information to the overall character-shape elements of allographs, but these features require a thorough normalization and they are sensitive to simple forging attempts (slant, size). However, as we have shown, the combination of character-shape elements and image properties such as the edge-hinge angular joint probability distribution function will yield usable classification rates. We can anticipate on a number of objections which could potentially be raised with respect to the proposed approach and the experiments which have been performed. We will try to refute these potential objections or put them into a different perspective in the next paragraphs.

Objection 1. The data are academic and clean, written on the same paper type, and using a single brand of ball-point pen, by 250 Dutch subjects. The test set contains 150 writers (300 samples): This is hardly representative of conditions seen in the application domain.

Reply. It is true that the data are uniform in many senses. Additionally, the same texts have been written by all subjects. However, it is also an advantage of the current experiment that the results can be attributed to the differences in writing style, proper. For any robust writer-identification algorithm, one would hope that an unknown sample can be identified if it has been written on similar material by the same writer. For the connected-component contours, only a clear black/white image has to be realized. The ink trace must be thick enough to avoid singularities in the behavior of the Moore contour follower. Many heuristics have been developed over the years to improve on this. The edge-based approaches are fairly size insensitive. However, a size normalization and optionally, a normalization of slant, may be an option in a realistic application context. Most scans have to be processed anyway, to remove background textures. However, it is granted that an experiment on nonacademic data, including the required preprocessing, needs to be performed in future studies. Examples of difficult material include: a photograph of a threat written with a lipstick on a mirror, faint traces on carbon copies of bills, smeared and textured hotel registration forms containing a combination of handwritten material, and machine-print text.

In regard to the limited amount of data, it is clear that more research is needed with sets in the order of 10^4 samples. It should be noted that, with search sets of this magnitude, the presence of administrative and labeling errors which are accumulated during the enrollment process starts to play a problem, even in the case of the powerful biometric methods which are based on (bio)physical traits [39]. The resulting "adventitious matches"—in forensic jargon—will pull the achievable performance asymptote below the exact 100 percent. In order to explore the consequences of using large search sets, simulated data were generated, with random values from a Gaussian distribution with the same means and standard deviations as in the original feature vector $f1 \cup f2$ (Ndim = 1,553). At 6,000 samples, a drop of 7 percent in Top1 performance was observed. However, experiments with real data are needed to provide reliable estimates of the performance as a function of data set size. It should be noted that, in a practical case, the search set size can often be reduced on the basis of nonhandwriting evidence. About 6.6 bits of information are needed to reduce an existing data base of 30,000 samples to a search set of 300 samples, the size used in the current experiments. Restrictive constraints concern known information on writer handedness, age category, sex, nationality, etc.

Objection 2. Although the approach presented appears to be probabilistic through the use of probability density functions, no use is made of Bayesian statistics, which seem to be perfectly suitable to this problem.

Reply. A naive Bayesian approach, discounting the probabilities of evidence conjunctions in the joint PDF could be used here, indeed. Given a Kohonen SOFM codebook C of connected-component contours ($n = 33 * 33$),

$$P(w|C_1, C_2, \ldots, C_n) = \\ = P(w) * \frac{P(C_1|w)}{P(C_1)} * \frac{P(C_2|w)}{P(C_2)} * \ldots * \frac{P(C_n|w)}{P(C_n)}, \quad (10)$$

the posterior probability of finding a writer w, given the set of found connected-component contours C_i, equals the product of the prior probability of finding a writer with all conditional probabilities of finding a connected-component contour C_i given this writer, divided by the prior probabilities of finding evidential shapes C_i in the ensemble. As regards $P(w)$, the probability of finding a writer in the set was the same for all writers, in the current experiment. In the application domain, however, one may argue that it is not a desirable property if, e.g., the probability of deciding for a writer A equals the fivefold of the probability of deciding for a writer B, because writer A has five samples and writer B has only one sample in the database: One would like to take the identification decision on the basis of shape evidence, alone. Other sources of identity evidence should be incorporated in procedures which are outside the realm of writer identification on the basis of script shapes, proper. In regard to normalization by $P(C_i)$ and taking the normalized product of conditional probabilities, results in any case indicated a dramatic reduction in writer-identification performance.

Objection 3. In how far is the Kohonen self-organized map of connected-component contours representative of all possible writing styles? Only 100 writers were used here: This can hardly be called representative for the ensemble of possible uppercase allographs.

Reply. The size of script sample collections in forensic practice may be 20k-70k. Indeed, it would be better to use more training data, and the size of the Kohonen network may have to be enlarged. However, inspection of Fig. 5 will reveal that if one searches for recognizable fragments of variations on letters in the alphabet, not all of them seem to be present. Rather than representing an exhaustive list of all possible shapes, the Kohonen network spans up a shape space. The CO^3s of an unseen allograph will find their attractor shapes in the map: It is the overall shape of the resulting probability density function that will characterize the writer.

Objection 4. Currently, powerful methods for class separation exist, such as the multilayer perceptron (MLP) and the support-vector machine (SVM). One would expect that the use of these methods will yield higher performances than reported on the simple distance measures and nearest-neighbor search.

Reply. The use of a technique like the SVM is not trivial in the *writer-identification* problem. The amount of writers in a realistic problem may exceed the number of 20,000. Training writer-specific SVMs, using, e.g., a one-versus-others training scheme becomes prohibitive. A more realistic solution would entail the use of a trained distance function between two given sample feature vector. Although the idea of trained distance functions as such is appealing, preliminary experiments revealed that the results where not much better than those obtained by nearest-neighbor search. The number of contrasting classes (writers) is large, and it is difficult to find a distance function which suits all local sample configurations with a smooth margin separating "near" (same-identity) from "far" (different-identity) samples. At this moment, the combination of a comparable or lower performance with the additional cost of training efforts and additional parameters seems unattractive. However, more research is needed here, indeed.

We want to point out, nevertheless, that an SVM or MLP trained distance function offers a very effective solution to the *writer-verification* problem when the question is: Are these two given samples written by the same person? The SVM seems to be the ideal classifier to give the yes/no answer to this question of authentication in a one-to-one comparison.

Objection 5. The proposed edge-based features, here and in other studies [36], [28], [37] may perform well, but the first attempt at disguising identity by a forging writer is to alter the habitual slant angle.

Reply. As stated in the introduction, the goal of the proposed method is to correctly identify a writer on the basis of handwritten samples produced under natural conditions. The introduction of the connected-component contour PDF is in fact inspired by the goal to complement the information derived from exact edge orientations with allographic style information. Connected components are usually small, and the contour feature, which consists of normalized x, y coordinates, is quite robust to naturally occurring slant variations. For structural slant deviation in a suspected sample, the shear transform can be applied in order to align the average slant of an unknown sample with a standard slant value. Such a normalization can be realized automatically on the basis of the modal edge orientation if a handwritten sample contains a sufficient amount of characters. Such methods are widely applied in automatic handwriting recognition. After such a slant normalization, residual writer-specific information may be expected to be present in an edge-orientation (polar) PDF.

Objection 6. The proposed approach is, in the end, hybrid: The CO^3 PDF is apparently not powerful enough and an additional edge-orientation feature has to be called in to achieve performances which become interesting.

Reply. As discussed in the Introduction, the identification of writers is not as easy as is DNA-based, fingerprint-based, or iris-based identification of individuals. This is mainly due to the fact that properties of the working brain are involved, as contrasted with the low-level biochemical or biomechanical information that can be used in these other techniques. Under these conditions, a pragmatic use of all the available shape information seems to be preferable. In order to put the performances in perspective, for the *Firemaker* set, the following additional findings may be presented:

- Using the edge-hinge (f2) feature, but computed separately for the upper and lower halves of written text-lines and subsequently concatenated [37], performances of 79 percent on Top-1 and 96 percent on Top-10 are reported on the same *uppercase* handwriting samples used for the present study. The reported performance was obtained under difficult testing conditions: All 250 writers were included in the test set (since no training was needed) and searches were performed in the leave-one-out "AB" manner. This shows that a combination of this method ("f2-split") with $p(CO^3)$ may yield even better figures than reported here for a homogeneous f2 computation over the sample as a whole.
- For the predominantly *lowercase* text samples from the Firemaker set and under similarly difficult testing conditions, the following results have been reported [37]: a Top-1 and Top-10 performance of 78 percent and 95 percent, respectively, for feature "f2-split."

TABLE 6
Nearest-Neighbor Performance of Other Features on Set "ab:" Leave One Out (1 versus 299 Samples), N = 150 Writers, as Before

Feature	Description	$Ndim$	$Top1$ (%)	$Top10$ (%)
e	normalized entropy	1	2	19
w1	wavelets,Haar	99	5	14
w2	wavelets,Odegard	99	14	28
w3	wavelets,Adelson	99	14	29
w4	wavelets,Antonini	99	14	29
w5	wavelets,Brislawn	99	14	29
w6	wavelets,Daubechies 14	99	15	29
w7	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, 15x15	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 $Top1$ and $Top10$ percentages of the correct writer found in a sorted hit list of size 1 and 10, respectively.

- Using an existing system for forensic writer identification (X) and a subset of predominantly lower-case text by 100 writers, Top-1 and Top-10 performances of 34 percent and 90 percent were realized.
- Another existing practical system for forensic writer-identification (Y), showed a Top-1 and Top-10 performance of 65 percent and 90 percent, respectively, on image-based statistics and 45 percent and 81 percent on features measured by script experts. It is important to note that the number of writers (distractors) in these experiments (100) is two-third of the number of writers used for testing the system presented in the current study (150), which makes the identification easier for the systems X and Y.

Objection 7. It is unclear how the proposed approach performs on these data in comparison to other known image features.

Reply. Table 6 shows performances of a number of features on this same data set (AB), in leave-one-out mode. Feature e represents a one-dimensional feature, i.e., the number of 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 (Top10: 19 percent). The wavelet-based features *(w1-w7)* are computed on the basis of Davis' Wavelet package [40], 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 6, *w6*). 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 ([36], [28], [37]). The "brush" feature [28] shows an interesting performance (Top1: 69 percent). 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. Performances for features *f0, f1, f2* as described elsewhere in this paper are duplicated here for ease of comparison.

Taking all of these points in consideration, we are quite confident that the results presented here on the combined use of connected component contours and edge-hinge angles will be robust and replicatable.

5 Conclusion

This paper presents a theoretically founded approach for the use of a connected-component contour codebook for the characterization of a writer of uppercase Western letters. The use of the connected-component contour (CO^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 uppercase samples, but this is an infrequent processing stage. Writer-identification performance on this new feature is promising, and could be improved using better distance measures. However, as we have illustrated, the combination with another strong feature concerning the edge-orientation distribution has proven to be highly effective. In this manner, we have used two major and complementary features in image processing: edges and the shapes of connected components, covering the angular and Cartesian domains, respectively. Current experiments concern the use of the CO^3 codebook approach for writer identification on the basis of regular mixed-style scripts, obtaining promising results.

The goal remains to realize sparse-parametric solutions [28] for writer identification since there is limited room for extensive training and retraining, and the use of an abundance of weights entails a risk of biased system solutions. Again, it should be stressed that it is not the goal of this paper to introduce a single and ultimate solution. However, the use of automatic and computation-intensive approaches in this application domain will allow for massive search in large databases, with less human intervention than is current practice. By reducing the size of a target set of writers, detailed manual and microscopic forensic analysis becomes feasible. It is important to note also the recent advances [1], [41] that have been made at the detailed allographic level, when character segmentation or retracing is performed by hand, followed by human classification. In the foreseeable future, the toolbox of the forensic expert will have been thoroughly modernized and extended. Besides their forensic applicability, the methods described in this paper may have interesting potential applications in the field of historic document analysis. Examples are the identification of scribes on medieval manuscripts or identification of the printing house on historic prints.

References

[1] S. Srihari, S. Cha, H. Arora, and S. Lee, "Individuality of Handwriting," *J. Forensic Sciences,* vol. 47, no. 4, pp. 1-17, July 2002.
[2] K. Franke and M. Köppen, "A Computer-Based System to Support Forensic Studies on Handwritten Documents," *Int'l J. Document Analysis and Recognition,* vol. 3, no. 4, pp. 218-231, 2001.
[3] H. Said, T. Tan, and K. Baker, "Writer Identification Based on Handwriting," *Pattern Recognition,* vol. 33, no. 1, pp. 133-148, 2000.

[4] U.-V. Marti, R. Messerli, and H. Bunke, "Writer Identification Using Text Line Based Features," *Proc. Sixth Int'l Conf. Document Analysis and Recognition (ICDAR '01),* pp. 101-105, 2001.
[5] Y. Zhu, T. Tan, and Y. Wang, "Biometric Personal Identification Based on Handwriting," *Proc. 15th Int'l Conf. Pattern Recognition,* pp. 801-804, 2000.
[6] M. Benecke, "DNA Typing in Forensic Medicine and in Criminal Investigations: A Current Survey," *Naturwissenschaften,* vol. 84, no. 5, pp. 181-188, 1997.
[7] B. Devlin, N. Risch, and K. Roeder, "Forensic Inference from DNA Fingerprints," *J. Am. Statistical Assoc.,* vol. 87, no. 418, pp. 337-350, 1992.
[8] A. Jain, L. Hong, and R. Bolle, "On-Line Fingerprint Verification," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 19, no. 4, pp. 302-314, 1997.
[9] M. E, V. Ballarin, F. Pessana, S. Torres, and D. Olmo, "Fingerprint Identification Using Image Enhancement Techniques," *J. Forensic Sciences,* vol. 43, no. 3, pp. 689-692, 1998.
[10] J. Daugman, "The Importance of Being Random: Statistical Principles of Iris Recognition," *Pattern Recognition,* vol. 36, no. 2, pp. 279-291, 2003.
[11] L. Schomaker, "From Handwriting Analysis to Pen-Computer Applications," *IEE Electronics Comm. Eng. J.,* vol. 10, no. 3, pp. 93-102, 1998.
[12] E. Dooijes, "Analysis of Handwriting Movements," *Acta Psychologica,* vol. 54, pp. 99-114, 1983.
[13] C. Francks, L. DeLisi, S. Fisher, S. Laval, J. Rue, J. Stein, and A. Monaco, "Confirmatory Evidence for Linkage of Relative Hand Skill to 2p12-q11," *Am. J. Human Genetics,* vol. 72, pp. 499-502, 2003.
[14] J. Gulcher, P. Jonsson, A. Kong et al., "Mapping of a Familial Essential Tremor Gene, Fet1, to Chromosome 3q13," *Nature Genetics,* vol. 17, no. 1, pp. 84-87, 1997.
[15] G.P. Van Galen, J. Portier, B.C.M. Smits-Engelsman, and L. Schomaker, "Neuromotor Noise and Poor Handwriting in Children," *Acta Psychologica,* vol. 82, pp. 161-178, 1993.
[16] E. Moritz, "Replicator-Based Knowledge Representation and Spread Dynamics," *Proc. IEEE Int'l Conf. Systems, Man, and Cybernetics,* pp. 256-259, 1990.
[17] G. Jean, *Writing: The Story of Alphabets and Scripts.* Thames and Hudson Ltd., 1997.
[18] L.R.B. Schomaker and R. Plamondon, "The Relation between Pen Force and Pen-Point Kinematics in Handwriting," *Biological Cybernetics,* vol. 63, pp. 277-289, 1990.
[19] L. Schomaker, "Simulation and Recognition of Handwriting Movements: A Vertical Approach to Modeling Human Motor Behavior," PhD dissertation, Univ. of Nijmegen, NICI, The Netherlands, 1991.
[20] R. Schmidt, "A Schema Theory of Discrete Motor Skill Learning," *Psychological Rev.,* vol. 82, pp. 225-260, 1975.
[21] L. Schomaker, A. Thomassen, and H.-L. Teulings, "A Computational Model of Cursive Handwriting," *Computer Recognition and Human Production of Handwriting,* M.S.R. Plamondon and C.Y. Suen, eds., World Scientific, pp. 153-177, 1989.
[22] R. Plamondon and F. Maarse, "An Evaluation of Motor Models of Handwriting," *IEEE Trans. Systems, Man, and Cybernetics,* vol. 19, pp. 1060-1072, 1989.
[23] R. Plamondon and W. Guerfali, "The Generation of Handwriting with Delta-Lognormal Synergies," *Biological Cybernetics,* vol. 78, pp. 119-132, 1998.
[24] D. Doermann and A. Rosenfeld, "Recovery of Temporal Information from Static Images of Handwriting," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* pp. 162-168, 1992.
[25] K. Franke and G. Grube, "The Automatic Extraction of Pseudo-Dynamic Information from Static Images of Handwriting Based on Marked Gray Value Segmentation," *J. Forensic Document Examination,* vol. 11, pp. 17-38, 1998.
[26] S. Kondo and B. Attachoo, "Model of Handwriting Process and Its Analysis," *Proc. Eighth Int'l Conf. Pattern Recognition,* pp. 562-565, 1986.
[27] L. Vuurpijl and L. Schomaker, "Finding Structure in Diversity: A Hierarchical Clustering Method for the Categorization of Allographs in Handwriting," *Proc. Int'l Conf. Document Analysis and Recognition,* pp. 387-393, Aug. 1997.
[28] L. Schomaker, M. Bulacu, and M. van Erp, "Sparse-Parametric Writer Identification Using Heterogeneous Feature Groups," *Proc. IEEE Int'l Conf. Image Processing,* vol. 1, pp. 545-548, 2003.
[29] T. Kohonen, *Self-Organization and Associative Memory,* second ed., Berlin: Springer Verlag, 1988.
[30] L.R.B. Schomaker, "Using Stroke- or Character-Based Self-Organizing Maps in the Recognition of On-Line, Connected Cursive Script," *Pattern Recognition,* vol. 26, no. 3, pp. 443-450, 1993.
[31] L. Schomaker, G. Abbink, and S. Selen, "Writer and Writing-Style Classification in the Recognition of Online Handwriting," *Proc. European Workshop Handwriting Analysis and Recognition: A European Perspective. Digest Number 1994/123,* p. 4, July 1994.
[32] L. Schomaker, E. de Leau, and L. Vuurpijl, "Using Pen-Based Outlines for Object-Based Annotation and Image-Based Queries," *Visual Information and Information Systems,* D. Huijsmans and A. Smeulders, eds., New York: Springer, pp. 585-592, 1999.
[33] F. Maarse, L. Schomaker, and H.-L. Teulings, "Automatic Identification of Writers," *Human-Computer Interaction: Psychonomic Aspects,* G. van der Veer and G. Mulder, eds., New York: Springer, pp. 353-360, 1988.
[34] J.-P. Crettez, "A Set of Handwriting Families: Style Recognition," *Proc. Third Int'l Conf. Document Analysis and Recognition,* pp. 489-494, Aug. 1995.
[35] F. Maarse and A. Thomassen, "Produced and Perceived Writing Slant: Differences between Up and Down Strokes," *Acta Psychologica,* vol. 54, nos. 1-3, pp. 131-147, 1983.
[36] M. Bulacu, L. Schomaker, and L. Vuurpijl, "Writer Identification Using Edge-Based Directional Features," *Proc. ICDAR'2003: Int'l Conf. Document Analysis and Recognition,* pp. 937-941, 2003.
[37] M. Bulacu and L. Schomaker, "Writer Style from Oriented Edge Fragments," *Proc. 10th Int'l Conf. Computer Analysis of Images and Patterns,* pp. 460-469, 2003.
[38] L. Vuurpijl, L. Schomaker, and V. Erp, "Architecture for Detecting and Solving Conflicts: Two-Stage Classification and Support Vector Classifiers," *Int'l J. Document Analysis and Recognition,* vol. 5, no. 4, pp. 213-233, 2003.
[39] A. Broeders, "In Search of the Source: On the Foundations of Criminalistics and the Assessment of Forensic Evidence," PhD thesis, Leiden Univ., with abstract in English, ISBN 90-130-0964-6, Netherlands: Kluwer, p. 349, 2003.
[40] G. Davis and A. Nosratinia, "Wavelet-Based Image Coding: An Overview," *Applied and Computational Control, Signals, and Circuits,* vol. 1, no. 1, 1998.
[41] M. van Erp, L. Vuurpijl, K. Franke, and L. Schomaker, "The WANDA Measurement Tool for Forensic Document Examination," *Proc. 11th Conf. Int'l Graphonomics Soc.,* pp. 282-285, 2003.

Lambert Schomaker received the MSc degree cum laude in psychophysiological psychology in 1983, and the PhD degree on the simulation and recognition of pen movement in handwriting in 1991 at Nijmegen University, The Netherlands. Since 1988, he has been working in several European research projects concerning the recognition of online, connected cursive script and multimodal multimedia interfaces. Current projects are in the area of cognitive robotics, image-based retrieval, historical handwritten document analysis, and forensic handwriting analysis systems. Professor Schomaker is a member of the IEEE, the IEEE Computer Society, and the IAPR. He is chairman of IAPR/TC-11 (Reading Systems). He has contributed to more than 60 reviewed publications in journals and books. In 2001, he accepted the position of full professor and director of the AI Institute at Groningen University, The Netherlands.

Marius Bulacu received the BSc and MSc degrees in physics from the University of Bucharest, Romania in 1997 and 1998, respectively. He did teaching and research in the Biophysics Department, Faculty of Physics, University of Bucharest from 1999 to 2002. Since March 2002, he has been with the Artificial Intelligence Institute of the University of Groningen, The Netherlands, pursuing the PhD degree. He is currently working on developing the vision system for a mobile robot capable to detect and read the text encountered in its environment. His scientific interests include computer vision, statistical pattern recognition, and intelligent autonomous robots. He is a student member of the IEEE.