# Analysis of texture and connected-component contours for the automatic identification of writers

Marius Bulacu  Lambert Schomaker

AI Institute, Groningen University, The Netherlands

We developed new techniques for offline writer identification that use probability distribution functions (PDFs) extracted from scanned images of handwriting to characterize writer individuality. Our methods operate at two levels of analysis: the texture level and the character-shape (allograph) level.

At the *texture level*, a generic descriptor that can be used to characterize individual handwriting style is the probability distribution of *edge-angles* $p(\phi)$. While this classical texture feature proves to be effective for writer identification, we obtained significant further improvements in performance by designing more complex features that use the edge orientation as a building block. These new features are bivariate edge-angle probability distributions ($p(\phi_1, \phi_2)$, $p(\phi_1, \phi_3)$) computed separately on the upper and lower halves of text lines and then adjoined (see fig. 1a). They encode, besides orientation, also curvature and location specific information, giving an intimate characterization of the individual handwriting style (see fig. 1b).

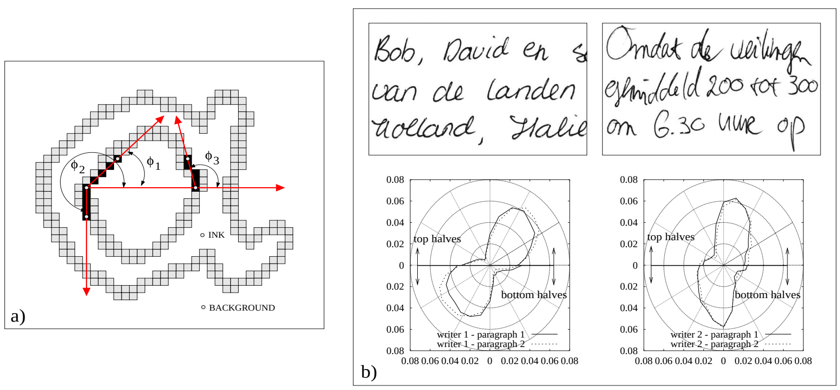
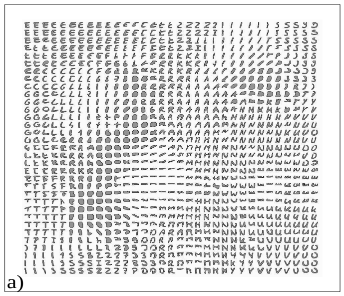



Figure 1: a) Extraction of the edge-based texture features on letter "a", b) Examples of lowercase handwriting from two different subjects and the corresponding polar diagrams of the "split-line" edge-direction distribution $p(\phi)$.

In our analysis at the *allograph level*, the writer is considered to be characterized by a stochastic pattern generator, producing a family of *connected components*. A codebook of connected-component-contours (COCOCOs or $CO^3$s) is generated from an independent

[0]This paper was published as: Marius Bulacu, Lambert Schomaker, *Analysis of texture and connected-component contours for the automatic identification of writers*, Proc. of 16th Belgium-Netherlands Conference on Artificial Intelligence (BNAIC 2004), 2004, pp. 371-372, 21-22 October, Groningen, The Netherlands



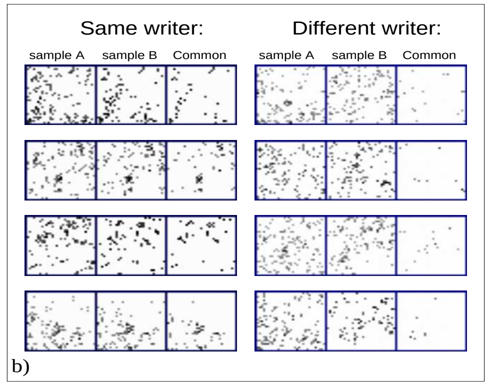



Figure 2: a) A Kohonen self-organized map of 33x33 $CO^3$s, b) Density plots of $p(CO^3)$. If A and B are samples from two different writers, the overlap between their PDFs is much lower as can be seen in the third column ('Common').

training set of handwritten samples using a Kohonen self-organizing map (see fig. 2a). The PDF of $CO^3$s is then computed for an independent test set containing unseen writers. Results revealed a high-sensitivity of $p(CO^3)$ for identifying individual writers (see fig. 2b). This method can be applied directly on samples of uppercase handwriting (isolated handprint). For lowercase (cursive) handwriting a segmentation stage is needed and we use a procedure based on finding the minima of the lower contour. The PDF of fragmented $CO^3$s $p(FCO^3)$ is used as a writer characteristic. This fragmentation approach is applicable to general free-style handwriting, when it is not a priori known whether the handwritten sample contains lowercase, uppercase, cursive or handprint script.

The proposed automatic techniques bridge the gap between image-statistics approaches on one end and manually measured allograph features of individual characters on the other end, covering both the angular and Cartesian domains. Our methods outperformed two systems (X and Y) used in current forensic practice (see table 1).

Table 1: Writer identification accuracy. The dataset contains two samples per writer. One selected sample is matched against the remaining $2N-1$ samples that contain only one target sample (the pair) and $2N-2$ distractors. The performance percentages express how often the correct writer is in the top 1 respectively top 10 entries in the sorted list. The performance of $p(FCO^3)$ depends on the category of samples used in the evaluation (whether copied or self-generated handwriting).

| Method / Feature | N writers | lowercase | | UPPERCASE | | Ref. |
|---|---|---|---|---|---|---|
| | | Top 1 (%) | Top 10 (%) | Top 1 (%) | Top 10 (%) | |
| $p(\phi)$ | 150 | 53 | 88 | 34 | 79 | [1] |
| $p(\phi_1, \phi_2)$ | 150 | 84 | 97 | 84 | 97 | [1] |
| $p(\phi_1, \phi_3)$ | 150 | 70 | 94 | 68 | 91 | [1] |
| $p(CO^3)$ | 150 | – | – | 72 | 93 | [2] |
| $p(FCO^3)$ | 150 | 71 - 97 | 90 - 100 | 73 | 91 | [3] |
| system X | 100 | 34 | 90 | – | – | [2] |
| system Y | 100 | 65 | 90 | – | – | [2] |

## References


[1] Marius Bulacu and Lambert Schomaker. Writer style from oriented edge fragments. In *Proc. of CAIP 2003: LNCS 2756*, pages 460–469, 2003.

[2] Lambert Schomaker and Marius Bulacu. Automatic writer identification using connected-component contours and edge-based features of uppercase western script. *IEEE Trans. on PAMI*, 26(6):787–798, 2004.

[3] Lambert Schomaker, Marius Bulacu, and Katrin Franke. Automatic writer identification using fragmented connected-component contours. In *Proc. of IWFHR 2004*, pages 185–190, 2004.