Sparse-parametric writer identification using heterogeneous feature groups

L.R.B. Schomaker[1], M. Bulacu[1] and M. van Erp[2]
[1]AI Institute, Groningen University, [2] NICI, The Netherlands
{schomaker,bulacu}@ai.rug.nl, M.vanErp@nici.kun.nl

Problem

Traditional methods for **forensic writer identification** require considerable manual efforts in individual-character measurements by human experts. However, with current background removal methods, it now becomes feasible to use automatic image-based features on regions of interest which describe the individuality of handwriting style.

Nevertheless, a single feature representation cannot be expected to capture all particularities of writing style, and combination methods are needed. The application domain precludes the use of training on the large datasets such that sparse-parametric combination methods are preferred, excluding MLP or SVM-based combination functions.

Method

Forensic writer search is similar to Information Retrieval yielding a hit list, in this case of suspect documents, given a query in the form of a questioned script sample. Given the requirements, simple nearest-neighbour search is a viable solution. However, a proper distance function has to be identified. For the combination of results, rank combination (Borda) will be tested.

Feature & Distance function Overview

A number of feature groups has been selected for this experiment, on the basis of literature and earlier work on on-line writer identification. Complementarity of extracted information in the feature group was an important design goal.

Table 1: Feature groups used for writer identification and the used distance function $\Delta(\vec{u},\vec{v})$ between two samples $\vec{u}$ and $\vec{v}$. Colors correspond to performance-curve colors in Figure 6.

	Feature	Explanation	Dim.	$\Delta(\vec{u},\vec{v})$
f1	ACF	Autocorrelation in horizontal raster	100	Euclid.
f2	VrunB	PDF of vertical run lengths of ink	100	χ^2
f3	HrunW	PDF of horizontal run length of 'white'	100	χ^2
f4	Brush	Ink-density PDF at stroke endings	225	χ^2
f5	$p(\phi)$	Edge-direction PDF	16	Euclid.
f6	$p(\phi_1,\phi_2)$	Hinge angle combination PDF	464	χ^2
f7	$p(\phi_1,\phi_3)$	Horiz. edge-angle co-occurrence	512	χ^2
f8	WR	Writer: handedness, sex, age, style	16	Euclid.

f1: ACF, autocorrelation function of the horizontal raster, detects the presence of regularity in writing: regular vertical strokes will overlap in the original row and its horizontally shifted copy for offsets equal to integer multiples of the local wavelength. Every row of the image is shifted onto itself by a given offset and then the normalized dot product between the original row and the shifted copy is computed. The maximum offset ('delay') corresponds to 100 pixels. All autocorrelation functions are then accumulated for all rows and the sum is normalized to obtain a zero-lag correlation of 1.

f2: VrunB, PDF of vertical run lengths in ink

f3: HrunW, PDF of horizontal run lengths in background pixels Run lengths are determined on the binarized image taking into consideration either the black pixels corresponding to the ink trace width distribution or the white pixels corresponding to the horizontal stroke and character-placement distribution for the writer. The histogram of run lengths is normalized and interpreted as a probability distribution. We use horizontal run lengths of up to 300 pixels (f3) and vertical run lengths (f2) of up to 100 pixels, i.e., the height of a written line in the data set used (resolution is 300 dpi). This feature is not size invariant. However, size normalization is not an issue in interactive writer search. The run-length PDFs provide orthogonal information to the directional features.

f4: Brush, ink-density PDF at stroke endings. It is known that axial pen force ('pressure') is a highly informative signal in on-line writer identification [1]. In ink traces of ball-point pens, there exist lift-off and landing shapes in the form of blobs or tapering [2] due to the ink-depositing process during take off and landing of the pen. A convolution window of 15x15 pixels was used, only accumulating the local image if the current region obeys to the constraints for a stroke ending. This constraint is determined by a supraliminal ink intensity in the central pixel of the window, co-occurring with a long run of white pixels along minimally 50% of the perimeter of the window, which is interrupted by one ink strip of at least 5 % of the window perimeter (Figure 1).

f4: (continued Brush PDF) After summing all luminances, the accumulator window is normalized to a volume of 1, yielding a PDF for ink presence at stroke endings in any direction. This feature is not size invariant: the window of 15^2 pixels was chosen because it captures 6-7 pixel-wide ink traces (size normalization is assumed). Figure 2 displays the overall shape and subtle writer differences.

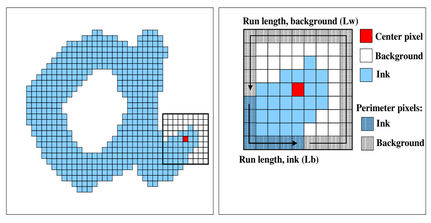

Figure 1. (left): A lower-case letter **a** *with its tail stroke. (right): An example of detecting end strokes on the basis of a central inked pixel and a constrained ink and paper runlength configuration around the window border (actually 15x15 pixels).*

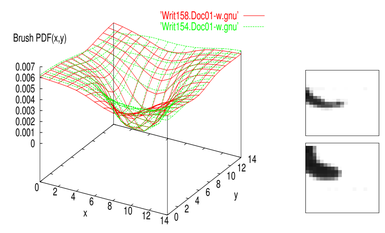

Figure 2. Superimposed brush PDFs for two writers, and examples of an "a" tail for two writers

f5: $p(\phi)$, simple edge-direction PDF, is computed by considering the PDF of quantized directions of the Sobel edges in the image. Sixteen bins were used in the histogram (Figure 3).

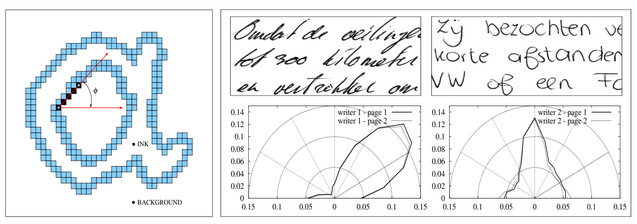

Figure 3. (left) Two handwriting samples from two different subjects. (right): We superimposed the polar diagrams of the edge-direction distribution $p(\phi)$ corresponding to pages 1 and 2 contributed to our data set by each of the two subjects.

f6: $p(\phi_1,\phi_2)$, hinge-angle combination PDF. In order to capture the curvature of the ink trace, which is very typical for different writers, another feature is needed, using local angles along the edges [3]. The computation of this feature is similar to the one previously described, but it has added complexity. The central idea is to consider the **two** edge fragments emerging from a central pixel and, subsequently, compute the joint probability distribution of the orientations of the two fragments of this 'hinge'. The final normalized histogram gives the joint probability distribution $p(\phi_1,\phi_2)$ quantifying the chance of finding in the image two "hinged" edge fragments oriented at the angles ϕ_1 and ϕ_2 respectively. The orientation is quantized in 16 directions for a single angle. We will consider only the non-redundant angles ($\phi_2 > \phi_1$) and we will also eliminate the cases when the ending pixels have a common side. Therefore the final number of combinations is $C(2n,2) - 2n = n(2n-3)$ (464 dimensions). See Figure 4 for more details.

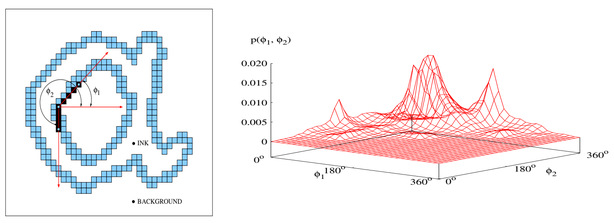

Figure 4. (left): The computation of the angular 'hinge' feature and (right): An example of a single-writer hinge PDF.

f7: $p(\phi_1,\phi_3)$, horizontal edge-angle co-occurrence. This feature is an variant of the edge-hinge feature, in that the combination of angles is computed along the rows of the image. For the angle of a found edge fragment i, the co-occurrence probability is computed with the angles of fragments j which are horizontally displaced from i (Figure 5).

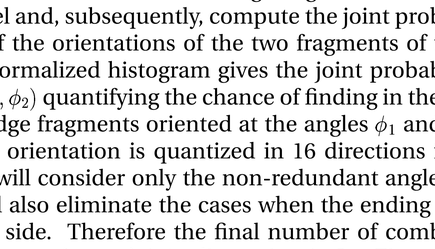

Figure 5. Computation of the horizontal (or vertical) edge-angle co-occurrence

f8: Writer characteristics (WR) is a 'pseudo' feature vector, containing writer parameters which are often known in the application context: Style may be one of Handprint, Cursive or Mixed. The parameters are represented as a bit vector. This feature is added to underscore the possibility of using heterogeneous sources of information in a rank-combination scheme.

Borda Rank-Combination Schemes

Given a sample of unknown identity u and a universe of samples of known writer identity W, each uniquely labeled $[1:m=|W|]$, and assuming there exist N feature groups describing a sample, we can construct a Borda rank-combination scheme. Assume a set of N distance vectors $\vec{\delta}_i(u,W)$ between the unknown sample u and the reference set W for each feature group $i=[1,...,N]$, such that each dimension of the distance vectors corresponds to one and the same sample index. Furthermore, given that a vector of ranks will be denoted by $\dot{}$, assume the availability of a rank operator $\dot{\rho}(\vec{x}), \vec{x}\in\mathbb{R}^m$ which returns for each dimension in $\vec{x}$ (i.e, handwritten sample) its unique rank in the set W with respect to u according to values in $\vec{x}$, in ascending order. The dimension $d(r)$ of a rank value uniquely refers to a sample in W. Thus $\dot{\rho}(\vec{\delta})$ guarantees that $(\Delta_{u,d(r_1)} < \Delta_{u,d(r_2)} < ... < \Delta_{u,d(r_m)})$, where $\Delta_{u,d(r_j)}$ is the distance between an unknown sample u and a known sample of j^{th} rank, indexed $d(r_j)\in W$. Then a Borda rank combination scheme can be considered as a rank-combination function $\vec{\beta}()$ operating on a tensor:

$$\dot{L}_{Borda}(u,W) = \dot{\rho}\left(\vec{\beta}\begin{pmatrix}\dot{\rho}(\vec{\delta}_1(u,W))\\ \dot{\rho}(\vec{\delta}_2(u,W))\\ \dot{\rho}(\vec{\delta}_3(u,W))\\ .\\ .\\ .\\ \dot{\rho}(\vec{\delta}_N(u,W))\end{pmatrix}\right) \quad (1)$$

where $\vec{\beta}()$ returns a vector in $\mathbb{R}^m$ which has a monotonous relation to the combined rank vector. The output hit list contains the samples in the final rank order $\dot{L}$. In the regular Borda vote, $\dot{L}_{Borda}(u,W) = \dot{\rho}(\sum_{i=1}^{N}\dot{\rho}(\vec{\delta}_i(u,W)))$, i.e., $\vec{\beta}()$ is the Sum function: ranks are summed per dimension before being resorted by $\dot{\rho}()$. However, many Borda-operator variants are known $\beta()$: Sum, Max, Median, Min, Majority, Plurality etc. In this study, we tested the use of the Sum operator. The problem of the Sum operator is that all votes are treated equally. Since the Median did not improve on this, we applied the Sum rule in a sequential and cumulative fashion from worst to best feature group. This is comparable to taking a weighted sum with rank weights $\omega_q = 2^{-q}$ where $q=[1:N]$ is the quality index of the feature group (1=best, $q=N$ is worst) after optimal group reordering.

Data & Evaluation

We evaluated the effectiveness of different features for writer identification using the *Firemaker* data set [4] A number of 251 Dutch subjects wrote four different A4 pages. On page 1 they were asked to copy a text presented as machine-printed characters. On page 2 they were asked to describe a given cartoon in their own words. The same kind of paper, pen and support were used for all subjects. The A4 sheets were scanned at 300 dpi, 8 bit / pixel gray-scale. Performance was tested using leave-one out. For a query sample, the set W will contain one matching sample of the same writer and 500 distractor samples by 250 other writers.

Results

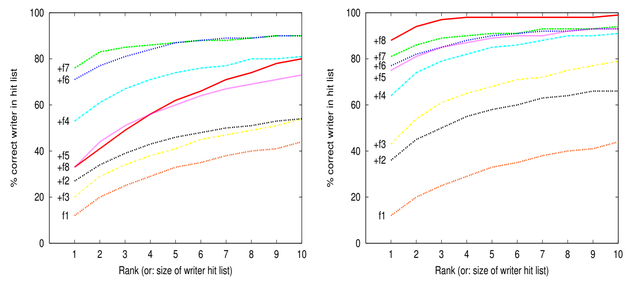

Figure 6. (left): Results for individual feature groups, (right): Results for sorted feature groups, using sequential Borda rank combination

Recent tests with the Min operator, not reported here, have given indications that this rule may be preferable to sequential Borda. Ongoing studies have revealed still better identification performances if (a) feature vectors are computed separately from upper and lower parts of lines of text [5], and additional improvement if (b) local component-shape features are used (Fig. 7).

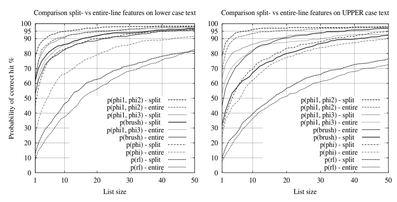

Figure 7. Recent results on refinement of angular features [5]

Actual forensic systems: System A: 34%,(90%) and System B: 65%,(90%) for Top1,(Top10) using only $N_{writers}=100$ from the same data are largely outperformed by our method: 79%,(95%).

Conclusions

- Localized, angular (co)occurences based on edges are very good features for writer identification.
- For feature vectors which are PDFs, the χ^2 distance measure is mostly the natural choice.
- In multiple feature groups where trained parametric combination cannot be applied, a sequential Borda approach which overweighs the better feature groups can be useful

References

[1] 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.

[2] D.S. Doermann and A. Rosenfeld, "Recovery of temporal information from static images of handwriting," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, 1992, pp. 162–168.

[3] M. Bulacu, L. Schomaker, and L. Vuurpijl, "Writer identification using edge-based directional features," in *Proc. of ICDAR 2003*, 2003, pp. 937–941.

[4] L.R.B. Schomaker and L.G. Vuurpijl, "Forensic writer identification [internal report for the Netherlands Forensic Institute]," Tech. Rep., Nijmegen: NICI, 2000.

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

Acknowledgements: Thanks to the Dutch Forensic Institute (NFI) who enabled us to set up a large data collection, to Katrin Franke of the Fraunhofer Institut (IPK) with her expertise in background removal, who revived the problem of writer identification in our lab and Louis Vuurpijl (NICI) for the data collection and preparatory research. This poster was presented at ICIP'2003, Barcelona.