SPARSE-PARAMETRIC WRITER IDENTIFICATION USING HETEROGENEOUS FEATURE GROUPS

L. Schomaker *M. Bulacu*

AI Institute, Groningen University
The Netherlands
(schomaker, bulacu)@ai.rug.nl

M. van Erp

NICI, Nijmegen University
The Netherlands
M.vanErp@nici.kun.nl

ABSTRACT

This paper evaluates the performance of edge-based directional probability distributions as features in writer identification in comparison to a number of non-angular features. It is noted that angular features outperform all other features. However, the non-angular features provide additional valuable information. Rank-combination was used to realize a sparse-parametric combination scheme based on nearest-neighbor search. Limitations of the proposed methods pertain to the amount of handwritten material needed in order to obtain reliable distribution estimates. The global features treated in this study are sensitive to major style variation (upper- vs lower case), slant, and forged styles, which necessitates the use of other features in realistic forensic writer identification procedures.

1. INTRODUCTION

Image-based writer identification has a number of requirements differing from many other applications in pattern recognition. For searching suspects, queries are performed in databases of the order of 10^4 handwritten samples. The process is never fully automatic, due to a wide range of scan-quality and foreground/background separability problems. Three image-related information sources are usually integrated in operational systems: 1) automatic feature extraction from a region of interest (ROI), using general image information, 2) manually measured script features by forensic experts, and 3) character-dependent shape information. This paper will focus on category 1), which is only applicable if good preprocessing has been performed and a crisp image of ink on a homogeneous background is available in gray scale. The target performance in the application domain is to reduce a list of 10^4 suspects to a top-hit list of one hundred candidates, with a probability of near to 1.0 that the correct writer is in this reduced hit list. In this respect writer identification is akin to the application area of Information Retrieval. Given the time-variant contents of the sample databases and the interactive manner in which feature groups are used, there are disadvantages to the use of parameter-greedy methods such as the multi-layer perceptron and the support-vector machine. Additionally, the number of samples per class (i.e., writer) is very limited, limiting the extent to which the statistical within-writer parameters can be estimated. It is important to note that no single feature will be powerful enough for the defined performance target, necessitating the use of classifier-combination schemes. Again, in the combination stage, trained meta classifiers are impractical here. In this paper we will present two new and useful features for writer identification and report on results using nearest-neighbor matching with a sequential rank-combination scheme [1] using several common feature groups.

2. DATA

We evaluated the effectiveness of different features in terms of writer identification using the *Firemaker* data set [2][1]. A number of 251 Dutch subjects, predominantly students, were required to write four different A4 pages. On page 1 they were asked to copy a text presented in the form of machine print characters. On page 2 they were asked to describe the content of a given cartoon in their own words. Pages 3 and 4 are not used here. The recording conditions were standardized: the same kind of paper, pen and support were used for all the subjects. This is an idealized situation compared to the conditions in practice. However, it is also true that the detection of the correct writer on the basis of the character shapes proper is the core function which is to be evaluated with this data set. The response sheets were scanned with an industrial-quality scanner at 300 dpi, 8 bit / pixel, gray-scale. Our experiments are entirely image-

[0] This paper was published as: Lambert Schomaker, Marius Bulacu, Merijn van Erp, *Sparse-parametric writer identification using heterogeneous feature groups*, Proc. of Int. Conf. on Image Processing (ICIP 2003), IEEE Press, 2003, pp. 545-548, vol. I, 14-17 September, Barcelona, Spain

[1] This database has been collected with the help of a grant from the Dutch Forensic Institute NFI, The Hague, The Netherlands

Table 1. Feature groups used for writer identification and the used distance function $\Delta(\vec{u},\vec{v})$ between two samples $\vec{u}$ and $\vec{v}$. Feature group 8 (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.

	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.

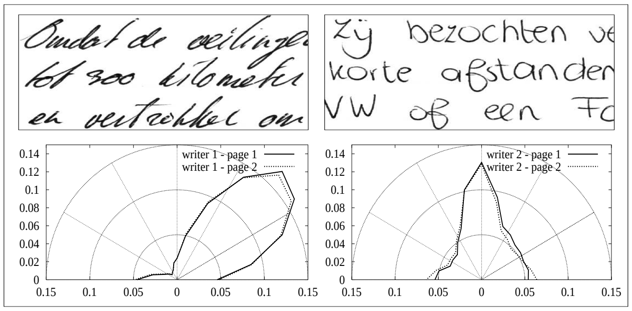

Fig. 1. Two handwriting samples from two different subjects. We superposed the polar diagrams of the edge-direction distribution $p(\phi)$ corresponding to pages 1 and 2 contributed to our data set by each of the two subjects.

based, no on-line information is available (e.g. speed of writing, order of different strokes).

3. FEATURES

Table 1 shows the feature groups used in this study and the distance function used in plain nearest-neighbor matching. Experiments have been performed with Hamming distance, Minkowski up to 5th order, Hausdorff and for the probability distribution functions, χ^2 and Bhattacharyya. Only best-performing distance functions will be used here. Features f1-f3 are known from literature, feature f4 is by the first author, f5 is known from literature but adapted to edges, features f6 and f7 are new features by the second author. We believe that the use of empirical probability-density functions (PDFs) is a more sensitive way of representing a writer’s uniqueness than enforcing a parametric distribution model. Furthermore, the use of PDFs in general allows for a homogeneous feature vector for which excellent distance functions exist.

3.1. Edge-direction distribution (f5)

The distribution of the writing directions is characteristic of a writer’s style. The polar probability density function was used in an on-line study of handwriting [3] to describe differences between upward and downward strokes. It was also used off-line [4] as a preliminary step to handwriting recognition that allows a partition of the writers by unsupervised fuzzy clustering in different groups. The edge-direction distribution is strongly influenced by slant, which may have been forged since it is a geometric property which is under voluntary control. However, slant normalization by shear is an easy step to perform in cases of doubt. Here, we will test the performance of this feature, assuming that a reference sample and a questioned sample were written with a comparable slant (non-forged conditions). 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. 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 on-line 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 of 16 direction performed best and will be used in the sequel. A more detailed description of the method can be found elsewhere [5].

As can be noticed in fig. 1, 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 the feature space.

3.2. Edge-hinge distribution (f6)

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. 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, as before. From the total number of combinations of two angles we will consider only the non-redundant ones ($\phi_2 > \phi_1$) and we will also eliminate the cases when the ending pixels have a common side. Therefore the final number of combinations is

$C(2n, 2) - 2n = n(2n - 3)$. The edge-hinge feature vector will have 464 dimensions (16 directions considered). A more detailed description of the method can be found elsewhere [5].

3.3. Horizontal co-occurrence of edge angles (f7)

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.

3.4. Run-length distributions (f2 and f3)

Run lengths, first proposed for writer identification by Arazi [6], are determined on the binarized image taking into consideration either the black pixels corresponding to the ink trace or, more beneficially, the white pixels corresponding to the background. Whereas the statistical properties of the black runs mainly pertain to the ink width and some limited trace shape characteristics, the properties of the white runs are indicative of character placement statistics for a writer. There are two basic scanning methods: horizontal along the rows of the image and vertical along the columns of the image. Similar to the edge-based directional features presented above, the histogram of run lengths is normalized and interpreted as a probability distribution. Our particular implementation considers only horizontal run lengths of up to 300 pixels (f3) and vertical run lengths (f2) of up to 100 pixels (the height of a written line). This feature is not size invariant. However, size normalization is not an issue in interactive writer search. An advantage of the use of run lengths is that these features provide orthogonal information to the directional features.

3.5. Autocorrelation (f1)

The autocorrelation function 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. This results in a large dot product contribution to the final histogram for periodic signal components: 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.

3.6. Brush function: Ink density at stroke endings (f4)

It is known that axial pen force ('pressure') is a highly informative signal in on-line writer identification [7]. In ink traces of ball-point pens, there exist lift-off and landing shapes in the form of blobs or tapering [8]. These shape phenomena are due to the ink-depositing process by a ball-point pen during take off from and landing on paper. Force variations will be reflected in the saturation and width of an ink trace. In order to capture the statistics of this process, 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 requiring a supraliminal ink intensity in the central window, co-occurring with a long run of white pixels along minimally 50% of the perimeter of the window, which is interrupted by an ink strip of at least 5 % of the window perimeter. 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 clearly not size invariant: the window of 15^2 pixels was chosen because it captures 6-7 pixel-wide ink traces. This feature thus aims at writer identification on the basis of stroke-brush patterns, assuming that a suitable normalization of size with a comparable ink-trace thickness between known and questioned sample has been realized in preprocessing.

4. SINGLE-FEATURE GROUP RESULTS

Table 2 shows the identification performance of the feature groups using hit list sizes of one to ten. It is very clear that the angular features (f6, f7) outperform the traditional features (f1, f2, f3), whereas the proposed Brush feature (f4) performs moderately well (53% Top1). The dimensionality of the feature vectors is high indeed, and PCA analysis confirmed that a reduction to 10% of the original dimensions may still yield reasonable results. Such computations, however, are in conflict with the sparse-parametric philosophy which is proposed here. Since the results on the single feature groups are well below the application target, and since the feature groups are theoretically reasonably independent, it can be expected that classifier combination may produce improved results. However, for aforementioned reasons, a non-parametric or sparse- parametric procedure is needed here.

5. RANKED VOTING METHODS

In ranked voting methods the voters are asked for a preference ranking of the candidates. A monotonous transform from likelihoods or distances to the ranking is assumed to exist. Although conversion to rank constitutes a loss of information, it also evades problems in scaling the voters confidences. The *Borda* count method needs a complete preference ranking from all voters over all candidates. It then computes the mean rank of each candidate over all voters. The classes are reranked by their mean rank and the top ranked class wins the election. Note that the Borda count is the ranked variant of the sum rule in classifier combination. Many variants exist but the use of the average ranks often produces good results [1]. A problem in the current context

Table 2. Top-N recognition rate (%), for individual feature groups. The numbers between parentheses in the last column show the results for a pseudo feature vector containing some general writer parameters (handedness, age, sex, coarse writing style) which are known. There are 501 handwritten samples in each individual query: 1 target and 500 distractors. Confidence interval: ± 5 % at 80 % recognition.

TopN	f1	f2	f3	f4	f5	f6	f7	(f8)
1	12	27	20	53	33	71	76	(33)
2	20	34	29	61	44	77	83	(41)
3	25	39	34	67	51	81	85	(49)
4	29	43	38	71	56	84	86	(56)
5	33	46	41	74	60	87	87	(62)
6	35	48	45	76	64	88	88	(66)
7	38	50	47	77	67	89	88	(71)
8	40	51	49	80	69	89	89	(74)
9	41	53	51	80	71	90	90	(78)
10	44	54	54	81	73	90	90	(80)

is that the individual feature groups yield different performances, such that rankings cannot be easily aligned. The following crude scheme proved quite effective. A cascade of classifiers is used, at each step merging the new ranking with the existing ranking: $\tilde{r}_{t+1} = \alpha r_{t+1} + (1-\alpha)\tilde{r}_t$ and $\alpha = 0.5$. Feature groups are sorted from low Top1 performance to high Top1 performance and the Borda cascade is executed in this order. The results are evaluated and classifiers are reshuffled until a monotonously increasing Top1 performance is result. This usually requires only a limited number of bubble-sort moves. Table 3 shows the results of this procedure, with a stable improvement in performance as feature groups are being added, from left to right. The first and weakest feature may contribute to the final performance in the order of 1%, using eight features. Strong features in the late stages in this Borda cascade dominate the final performance. In order to assess the influence of edge-based features alone, a test was performed using a Borda combination of f5, f6 and f7. Results indicated a Top-1 performance of 72% and Top-10 performance was 90%.

6. CONCLUSION

Although results are below the requirements in the application domain, they are quite robust. Results on weighted combination schemes not reported here showed that with extensive efforts, performances above 94% Top-10 are quite unlikely for these data and features. A kernel-based distance function will be our next effort, but it is with hesitation, since the sparse-parameter methods have distinct advantages. It is evident that global descriptors of ROIs will never suffice in writer identification. Detailed character-shape knowledge is needed, as well, especially in case of forged writing styles. Although edge-based orientation features are powerful, slant may be forged easily: Only a combination of approaches will yield reliable results in practice. Elsewhere [5], results on the stability of the features with respect to the amount of text or ink will be reported.

Table 3. Top-N recognition rate (%), cumulative over feature groups using cascaded Borda rank combination. The numbers between parentheses in the last column show the results in case some general writer parameters are known, as well. Please refer to Tables 1 and 2 for more details.

TopN	f1	+f2	+f3	+f4	+f5	+f6	+f7	(+f8)
1	12	36	43	64	75	77	81	(88)
2	20	45	54	74	81	82	86	(94)
3	25	50	61	79	85	85	89	(97)
4	29	55	65	82	87	88	90	(98)
5	33	58	68	85	89	90	91	(98)
6	35	60	71	86	90	91	91	(98)
7	38	63	72	88	90	92	93	(98)
8	40	64	75	90	92	92	93	(98)
9	41	66	77	90	93	93	93	(98)
10	44	66	79	91	93	93	94	(99)

7. REFERENCES

[1] M. van Erp, L. Vuurpijl, and L. Schomaker, "An overview and comparison of voting methods for pattern recognition," in *Proc. of the 8th IWFHR*. 2002, pp. 195–200, IEEE.

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

[3] F.J. Maarse and A.J.W.M. Thomassen, "Produced and perceived writing slant: differences between up and down strokes," *Acta Psychologica*, vol. 54, no. 1-3, pp. 131–147, 1983.

[4] J.-P. Crettez, "A set of handwriting families: style recognition," in *Proc. of the Third International Conference on Document Analysis and Recognition*, Montreal, August 1995, pp. 489–494, IEEE Computer Society.

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

[6] B. Arazi, "Handwriting identification by means of run-length measurements," *IEEE Trans. Syst., Man and Cybernetics*, vol. SMC-7, no. 12, pp. 878–881, 1977.

[7] 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.

[8] 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.