Writer and Writing-Style Classification in the Recognition of Online Handwriting. *

Lambert Schomaker,
Gerben Abbink,
& Sjoerd Selen

Nijmegen Institute for Cognition and Information (NICI).

Address: NICI/University of Nijmegen, P.O.Box 9104,
6500 HE Nijmegen, The Netherlands.
E-mail: Schomaker@nici.kun.nl,
Tel: +31-80-616029, Fax: +31-80-615938.

1 Introduction

One of the problems in the automatic recognition of cursive and mixed-cursive handwriting is the large variation of handwriting styles in a population. Automatic detection of the generic handwriting style, or identification of the writer could be useful to counteract this problem. The starting point for the writing style analyses is an existing recognition system for online connected-cursive handwriting (Schomaker & Teulings, 1990; Schomaker 1993). The input to this recognizer consists of pen-tip movements produced during the writing of a single word, using equidistant sampling in time. Data are lowpass filtered and normalized on size and slant. In the segmentation stage, strokes are used, which are defined as the pen-tip trajectory between two consecutive minimima in the pen-tip velocity. A neural-network technique, the Kohonen self-organizing map, is used to obtain a finite list of prototypical strokes (PS): A stroke alphabet (PSA). This stroke alphabet approximates the handwriting in the training set with a minimized rms error, and can be shown to generalize well to strokes in the handwriting of unknown writers. Thus, up to this stage of processing, the recognition system is writer independent. At the next level of processing, character classification takes place, in which sequences of stroke codes are classified as letters by a probabilistic stroke transition network. It should be noted, that at this level there is a very strong dependence on writer and writing style. However, simply combining all possible letter shape variants of all writers in the population in a single recognizer may lead to undesirable effects. First, the processing time increases steeply with the number of possible stroke interpretations. Second, the different interpretations of character shapes

*Schomaker, L, Abbink, G. & Selen, S. (1994). *Writer and Writing-Style Classification in the Recognition of Online Handwriting.* Proceedings of the European Workshop on Handwriting Analysis and Recognition: A European Perspective, 12-13 July, 1994, London: The Institution of Electrical Engineers, Digest Number 1994/123, (ISSN 0963-3308).

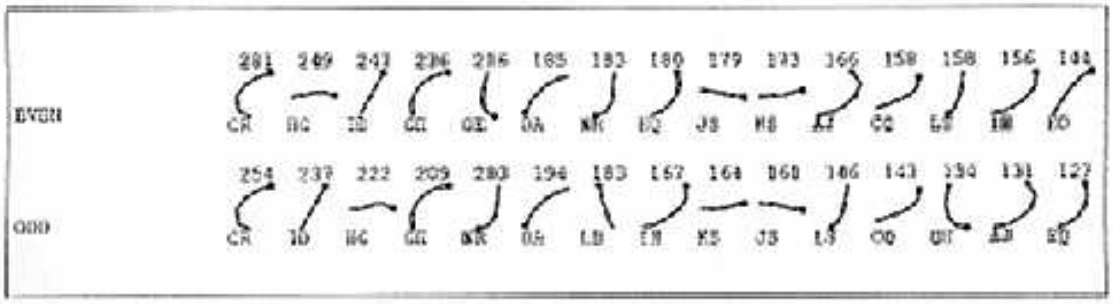

Figure 1: The top-15 of used strokes in two different sets of 20 words from a single writer. The top row denotes the frequency of usage, the middle row represents the size-normalized stroke shapes, and the bottom row represents a two-letter stroke code (figure is scan of original)

from different styles may interfere, leading to unnecessary computation with sometimes inappropriate recognition results. As an example, some younger writers (Irish and Dutch) use a small-written upper case */R/* in place of a lowercase */r/*. Clearly, the lower case interpretation is inappropriate in writers who never use the upper case shape. In other words: a considerable portion of the knowledge stored in a writer-independent ("multi-scriptor") handwriting recognition system may be irrelevant for the writer currently using that system.

In a pilot study, it was found that the histogram of usage of prototypical strokes (PS) was different for different writers, but at the same time it was comparable for different samples from the same writer. It is the purpose of this study to exploit the differential usage of prototypical strokes by different writers.

2 Writer and Style Classification

The database consists of isolated online handwritten words collected earlier from 30 writers of different nationalities, at various occasions. Words were obtained, during page, sentence and word copying tasks. The content and style of the words was miscellaneous Dutch, English, and Italian. A consecutive list of 40 words was randomly selected. From this database, two sets of data are created, i.e., the even-numbered (N=20) and odd-numbered words (N=20) per writer. The pen-tip displacement data are segmented into velocity-based strokes, and for each stroke a 14-dimensional feature vector is calculated (Schomaker & Teulings, 1990). The Kohonen network, or prototypical stroke alphabet (PSA) was trained on the handwriting of these writers. The network is organized as a 20x20 layer of cells, thus there is a stroke alphabet with a total of 400 PSs. For each input stroke, the PSA is searched for the closest matching prototypical stroke (PS) on the basis of Hamming distance. For each PS, the frequency of usage was counted, separately for each writer and for the even and odd sets. The resulting histograms were considered as the writer feature vector (WFV) and were entered into subsequent analyses. Figure 1. shows the top-15 strokes, separately for the even and odd numbered set of one writer.

Analysis 1: Clustering.

With the SPSS program CLUSTER, groupings of writer feature vectors were analyzed. It appeared that at the lowest level, the even and odd set of a given writer were always in

the same isolated cluster. At higher levels, the clustering was meaningful, i.e., groupings of female handwriting style and of nationality could be found. Figure 2. shows the cluster dendrogram.

Analysis 2: Discriminant Analysis.

With the SPSS program DISCRIMINANT, separability of writer feature vectors was analyzed. The even-numbered set was used for determining the discriminant functions, the odd-numbered set was used for testing. Since the dimensionality of the WFV was too high for SPSS, a random subset of 40 PSs was used as a reduced-dimensionality WFV. Already with the reduced WFV, a 100% correct writer classification could be obtained, as was to be expected on the basis of the cluster analysis. Some remarks have to be made at this point. Note that the texts were different among nationalities, which may have biased the clustering. However, Dutch and Irish male handwriting samples are in the same cluster as well as the Dutch and Irish female samples.

On the basis of these preliminary results, an analysis of the shape of the PSs is currently being executed, as well as their relation to letter shape variants (allographs). The described analyses will be repeated on new sets, within a single nationality, and with all writers producing the same text. The data collected within the international UNIPEN project* will be very useful in this respect. Future work will be focused on the relationship between the stroke shape alphabet and the character shape alphabet for a given writer or style group.

3 Conclusion

The results indicate that a reliable classification writers is possible, and interpretable style groupings can be formed. As opposed to the broad distinction between cursive, mixed and handprint, we propose to use a more subtle classification based on the characteristical set of stroke shapes produced by a given writer. Utilization of this type of knowledge will ultimately allow for recognition systems to adapt to a given writer with less user intervention, such as training the system at the character or word-level.

4 References

Schomaker, L.R.B. (1993). Using Stroke- or Character-based Self-organizing Maps in the Recognition of On-line, Connected Cursive Script. **Pattern Recognition, 26(3)**, 443-450.

Schomaker, L.R.B., & Teulings, H.-L. (1990). A Handwriting Recognition System based on the Properties and Architectures of the Human Motor System. **Proceedings of the International Workshop on Frontiers in Handwriting Recognition (IWFHR)** (pp. 195-211). Montreal: CENPARMI Concordia.

(UNIPEN original call for data in 1994: deleted in this .ps version)

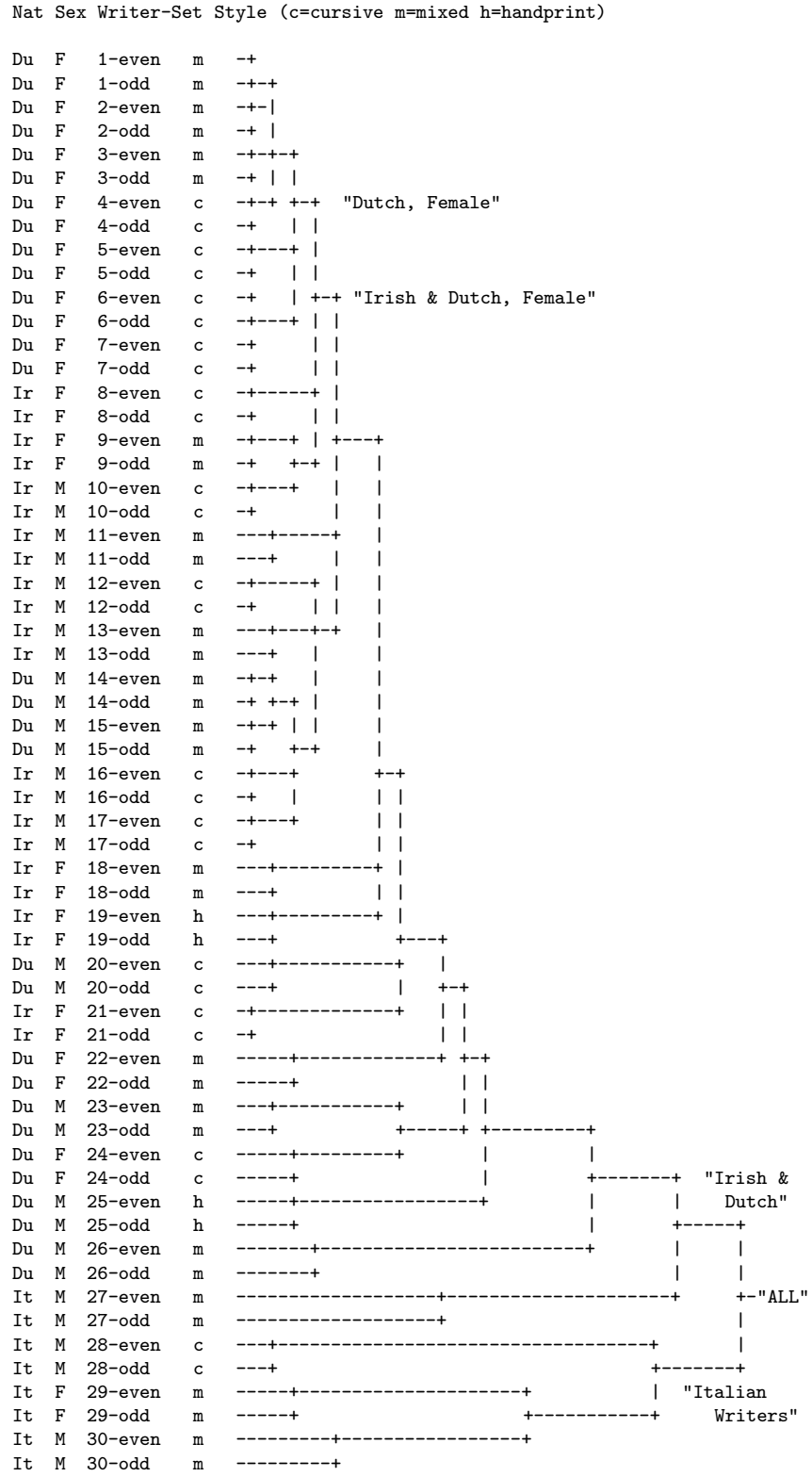

Figure 2. Writer clustering dendrogram on the basis of the histogram of prototypical stroke usage. Note that the smallest clusters (n=2, odd & even set) represent a single writer. Clustering was based on Average Linkage. Note: comments between quotes are only qualitative interpretations.