Coarse writing-style clustering based on simple stroke-related features [1]

Louis Vuurpijl & Lambert Schomaker

NICI, Nijmegen Institute for Cognition and Information, University of Nijmegen,
P.O.Box 9104, 6500 HE Nijmegen, The Netherlands, http://www.nici.kun.nl

Abstract

Two methods are presented for the automatic detection of generic writing styles like e.g. *mixed*, *cursive* and *handprint*. Based on a set of handwritten words, three features are determined: a *cursivity index* c, which indicates the tendency of a writer to write cursive, and two distance measures d_c and d_h. The distance measures represents the distance between the stroke feature vectors in the input handwriting and the strokes contained in two style-specific Kohonen Self-Organizing Maps (SOM). One SOM is tuned for the writing style *handprint*, and the other for *cursive*. The first method uses some linear decision criteria based on the feature vector $\{c, d_c, d_h\}$, for the classification of one of the three writing styles. The second method uses non-linear decision boundaries found via agglomerative hierarchical clustering of the three-dimensional feature vectors. Using the second method, several more distinctive writing style classifications are proposed.

1. Introduction

At the NICI, several projects are carried out within the frame work of online handwritten word recognition. Experience with building and analyzing word recognition systems shows that no monolithic system based on a single approach will be capable of handling the large variability and variance in human handwriting. A recognizer will comprise various specialized modules, where each module is dedicated to some peculiarities which exists in the way an individual writer — or a group of writers — writes.

Our current research interests include the specialization of a word recognition systems on specific writing style categories. Also in the area of off-line recognition preliminary attempts are being undertaken to identify style families [1]. In another study, it is shown that a specialized system is able to reach similar or better recognition rates than a system trained for all styles. Furthermore, such a system consumes less memory and computation resources and exhibits less confusion errors. The goal of the work presented in this paper is to arrive at an automated detection

[1] This work was funded by HP Labs, Bristol, UK.

of writing style. Given a writer's handwriting and its derived style classification, the proper specialized recognition module can be activated. Whereas in this paper we use handwriting data recorded in UNIPEN format, this approach can be also be used in a dynamic scenario where a recognition system "gets acquainted" with the way a writer writes. In such a scenario, upon entering the system, the new writer is asked to write a small set of pre-defined words. If all words are validated by the writer, the obtained information is used to determine the writing style. Although insufficient for automatic writer identification *per se*, this method may prove useful in combination with a larger number of more specific features like slant, rotation and velocity parameters.

The feature selection and set up of the style classification process are introduced in section 2. In section 3, a simple decision criterion is used to classify a set of 187 writers in the three writing styles *mixed*, *cursive* and *handprint*. A more advanced classifier is described in section 4, which describes an agglomerative hierarchical clustering technique. The results of this technique are presented in section 5.

2. Feature selection

The data set considered here comprises 187 UNIPEN files, each file containing handwriting produced by a single writer. For each writer from this data set, three features are computed. Based on such a feature vector $\{c, d_c, d_h\}$, the writer's handwriting is classified as belonging to a specific writing style.

2.1. A simple measure of writing style: cursivity index

The rationale behind this measure is that, given the word interpretation of a recorded sample of on-line handwriting, it should be able to derive a measure for the degree in which a writer produces isolated handprint characters or fully-connected cursive script. During an enrollment process, the new user of a pen computer may be asked to write down a particular list of words. From the ink data and the (presumably correct) word labels, for all n words without letter $< i >$ or $< j >$, a cursivity index should then be calculated:

$$\tilde{C_n} = \frac{1}{n} \sum_{w=1}^{n} \left(\frac{N_{l,w} - N_{pd,w} + 1}{N_{l,w}} \right) \tag{1}$$

where $\tilde{C_n}$ is the Cursivity Index, $N_{l,w}(> 0)$ is the number of letters in a non-$< ij >$ word and $N_{pd,w}$ the number of pen-down streams in this word.

The idea is that: 1) real *Cursive* writing style yields "one word/one ink blob" ($C = 1$), 2) real *Print* yields "one ink blob per letter" ($C = 0$), 3) the *Mixed* style will yield intermediate values, and 4) extreme cases of *Block Print* and Chinese characters will yield "more than one ink blob per symbol" ($C < 0$). The calculation of the cursivity index is thus based on the correct word ASCII string and the detected number of pen-down streams.

2.2. Kohonen SOM distance.

In our approach, the basic building block of handwriting is the stroke. We define a stroke as the sequence of recorded samples between two subsequent minima in the pen-tip velocity. A live `Java` demonstration of stroke segmentation points can be seen via WWW[2]. As explained in [3], each stroke can be described by a feature vector, which can be trained by a Kohonen Self-Organizing Map neural network. Several other feature schemes have been studied. The angular information seems to be more stable than absolute or relative size-based features. For the work presented here, we extract 14 features from the stroke samples: the vertical start and stop level, 5 consecutive angles along the stroke, 4 angles of the previous and following stroke, loop area, pen pressure and total length of the stroke.

Two 20x20 Kohonen SOMs were trained with respectively stroke feature vectors of *cursive* and *handprint* writers. For each of the 187 writers considered here, two distance measures d_c and d_h are computed as the average Euclidean distance between all stroke feature vectors produced by the writer and the respective weight vectors of the "winning" neurons in the cursive and handprint SOMs:

$$d = \frac{1}{nstrokes} \cdot \sum_{i=1}^{nstrokes} \| \mathbf{W}(i) - \mathbf{V}_i \| \qquad (2)$$

In (2), $nstrokes$ is the number of strokes produced by a writer, $\mathbf{W}(i)$ is the Kohonen weight vector of the neuron closest to stroke i, and $\mathbf{V}_i$ is the feature vector of stroke i. By specializing the Kohonen SOMs, d_c and d_h indicate how much the strokes produced by a writer resemble the strokes produced by cursive or handprint writers. Let d_c represent the distance to the cursive SOM, and d_h the distance to the handprint SOM. For d_c small and d_h large, it is safe to assume that a writer has a cursive handwriting. For both d_c and d_h large, the writer has some mixed writing style.

3. Writing style distinction

A simple decision criterion $S(c, d_c, d_h)$ was developed to determine a writer's style:

```
S(c,d_c,d_h) = if (s<T1) then handprint
               else if (s<T2) then mixed
               else cursive,
```

where $s = c * d_h / d_c$ and $T1, T2$ are threshold values (linear decision boundaries).

As all of the 187 data files contain a UNIPEN `.STYLE` definition, the accuracy of this method could be determined. It appeared that for 117 of the 187 writers a correct classification was made (see confusion matrix below).

[2] `http://www.nici.kun.nl/unipen/hwr-tutor/velocity.html`

Table 1
Guessed styles compared to corresponding .STYLE annotations.

	Actual style			#writers
Guessed style	print	mixed	cursive	
print	30	4	0	34
mixed	18	52	9	79
cursive	0	39	35	74

However, examining each of the files which were misclassified, it appeared that although the annotated style definition indicated otherwise, the decision made by our method was justifiable in most of the cases (see figure 1). Furthermore, no misclassifications were made between the writing styles cursive and handprint, only between mixed and cursive, or mixed and print.

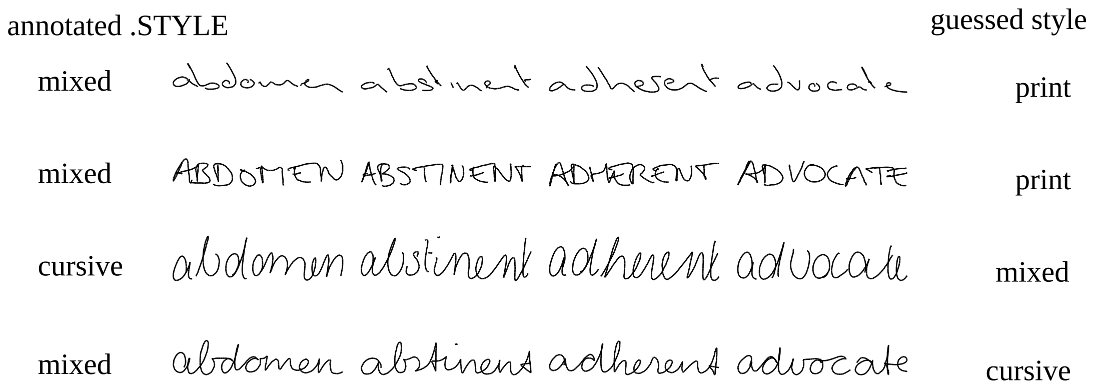
annotated .STYLE
guessed style
mixed abdomen abstinent adhesent advocate print
mixed ABDOMEN ABSTINENT ADHERENT ADVOCATE print
cursive abdomen abstinent adherent advocate mixed
mixed abdomen abstinent adherent advocate cursive

Figure 1: Situations were the .STYLE annotations and guessed style differ.

4. A hierarchical clustering technique

Agglomerative hierarchical clustering techniques are often used to find clusters in data. Many different paradigms exist, but they all feature the same principle. For a computed or given nxn distance matrix Δ, the agglomerative hierarchical clustering methods all operate following the following algorithm:

1. Start with n clusters each consisting of exactly one entity. Let the clusters be labeled with the number 1 through n.

2. Search the distance matrix for the most similar pair of clusters. Let the chosen clusters be labeled p and q and let their associated distance be $\delta(p, q)$, with $p < q$. Reduce the number of clusters by 1 through merger of clusters p and q. Label the product of the merger as p and update the distance matrix entries in order to reflect the revised distances between cluster p and all other existing clusters. Mark the column and row $\Delta[q]$ as unused.

3. Perform step 2 a total of $n - 1$ times.

For the results presented here, the merger operation and initialization of the distance matrix are performed using Ward's method. The implementation of this method following the heuristics given above was based on [2]. After $n-1$ iterative steps, the clustering process is finished. Validation and examination of the clusters found is described below.

5. A more distinctive writing style classification

Of the three coarse styles most commonly used, the mixed category seems to be the most ambiguous. Probably, a number of sub-style classes exist in this category. To search such classes, we used agglomerative hierarchical clustering techniques. In these techniques, based on a distance measure, iteratively the two most similar samples from a population are merged into one class. The 187 feature vectors were normalized using the z-transform and clustered using Ward's clustering method. Six clusters were found, which was decided based on 1) considering the resulting dendrogram, 2) visualizing the three-dimensional data points, 3) counting the known writing styles of writers belonging to each cluster, 4) considering the means and standard deviations of the features and 5) observing the handwriting of writers contained in each cluster. Examination of the clustered writer population showed that some ordering, ranging from cursive handwriting, via mixed to handprint, exists. Table 2 depicts that ordering, where the writing styles indicated by the UNIPEN ".STYLE" definition and the guessed styles described in the previous section were counted and used as order criterion.

Table 2
Clusters ordered by the number of writers belonging to a style as determined by the ".STYLE" definition or the method described in section 3 (indicated by the boxed numbers). Also given are mean and standard deviations of features per cluster.

	guessed			.STYLE			cursivity		d_c		d_h	
cluster	P	M	C	P	M	C	μ_c	σ_c	μ_{d_c}	σ_{d_c}	μ_{d_h}	σ_{d_h}
cluster 1	0	4	$\boxed{44}$	0	16	$\boxed{32}$	0.89	0.11	10.79	0.99	11.96	0.63
cluster 4	0	3	$\boxed{31}$	1	10	$\boxed{23}$	0.83	0.11	8.82	0.61	10.16	0.37
cluster 0	0	2	$\boxed{19}$	1	$\boxed{10}$	$\boxed{10}$	0.83	0.09	13.38	0.79	14.06	0.93
cluster 5	0	$\boxed{26}$	7	5	$\boxed{23}$	5	0.48	0.21	10.55	0.58	10.83	0.28
cluster 3	5	$\boxed{10}$	0	$\boxed{7}$	$\boxed{6}$	2	0.33	0.28	15.81	0.87	14.59	1.06
cluster 2	$\boxed{24}$	12	0	$\boxed{20}$	14	2	0.19	0.27	12.66	0.84	12.13	0.84

In this table its is shown that clusters 1 and 4 represent the cursive writing style, as μ_c is high and μ_{d_c} is low compared to μ_{d_h}. This can also be concluded from table 2. Cluster 0 also contains cursive writers, but the distances to the Kohonen SOMs are high, indicating that these writers produce strokes not well known to the networks. Cluster 5 represents the "pure" mixed writing style, and clusters 3 and 2 represent two forms of handprint.

Figure 2 below depicts some exemplar handwriting (the word *agreement*) for the clusters. A gradual change from cursive to handprint writing style can be observed.

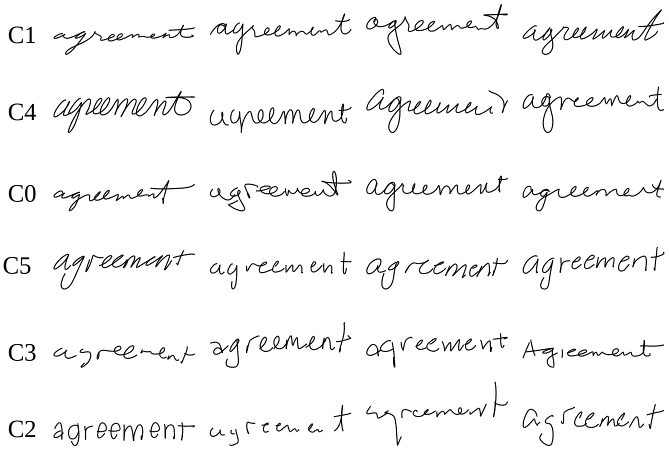

Figure 2: Typical handwriting for each of the six clusters.

6. Conclusions

Two methods of automatically determining the writing style for a given writers handwriting are introduced. Both methods use a three-dimensional stroke-based feature vector which can easily be constructed. The major conclusions of this work are that: 1) The .STYLE definition in UNIPEN files often is ambiguous. Especially in the mixed writing style, a number of sub-classes must be distinguished. 2) Using the linear decision method presented here, a clear distinction can be made between the writing styles handprint and cursive. 3) Using the clustering method described in section 5, a more subtle classification of writing styles is possible.

These results indicate that it is possible to develop a pre-processing system that determines to which writing style out of a number of writing styles a writer belongs. It was shown in other work that specialized word recognizers are smaller, faster and achieve similar or better recognition results. In future work we will use the pre-processing systems discussed here as a front end for a set of such specialized recognizers.

References

[1] Jean-Pierre Crettez. A set of handwriting families: style recognition. In *Third International Conference on Document Analysis and Recognition*, pages 489–494, Montreal, August 1995. IEEE Computer Society Press.

[2] M.R. Anderberg Air Force Systems Command United States Air Force. *Cluster Analysis for Applications*. Probability and Mathematical Statistics. Academic Press, New York San Franc, 1973.

[3] L.R.B. Schomaker and H.-L. Teulings. A handwriting recognition system based on the properties and architectures of the human motor system. In *Proceedings of the International Workshop on Frontiers in Handwriting Recognition (IWFHR)*, pages 195–211, Montreal: CENPARMI Concordia, 1990.