Generating copybooks from consistent handwriting styles

Ralph Niels and Louis Vuurpijl
Nijmegen Institute for Cognition and Information (Radboud University Nijmegen)
{*r.niels,vuurpijl*}*@nici.ru.nl*

Abstract

The automatic extraction of handwriting styles is an important process that can be used for various applications in the processing of handwriting. We propose a novel method that employs hierarchical clustering to explore prominent clusters of handwriting. So-called membership vectors are introduced to describe the handwriting of a writer. Each membership vector reveals the frequency of occurrence of prototypical characters in a writer's handwriting. By clustering these vectors, consistent handwriting styles can be extracted, similar to the exemplar handwritings documented in copybooks. The results presented here are challenging. The most prominent handwriting styles detected correspond to the broad style categories cursive, mixed, and print.

1. Introduction

It is well known that each handwriting is individual [20]. Characteristics that distinguish different handwritings from each other are (i) global holistic features like slant and spacing between characters, words or lines, (ii) local features that exhibit the occurrence of prototypical character shapes (allographs), and (iii) sub-allographic features like ligatures, descenders, ascenders, crossings and loops [3]. Groups of writers that have a significant amount of characteristics in common, share the same *handwriting style* [2, 5, 10]. Determining the handwriting style of a writer is an important process, serving three broad areas of application:

(i) Handwriting recognition, where knowledge about a handwriting style enables the development of handwriting recognition systems that are targeted on particularities in the handwriting belonging to a specific style. Rather than having one monolithic system that deals with the required different preprocessing, segmentation, feature extraction, and character shape variants, specialization in handwriting style categories like cursive, mixed and handprint is known to boost recognition performance while decreasing computational complexity of the system [7, 23].

(ii) Handwriting synthesis, in particular when the production of personalized texts in a certain handwriting style are concerned [11]. The successful selection of appropriate styles rely heavily on the determination of coherent collections of exemplar character shapes: "the handwriting fonts".

(iii) Forensic writer identification, where forensic experts use the notion of writing style to describe the handwriting of groups of writers. Similar coarse style categories like cursive and handprint may be employed to classify a writer's handwriting. Alternatively, by comparing the handwriting to template shapes listed in so-called copybooks, an attempt can be made to indicate the country of origin of the writer.

This paper presents a novel approach to the automatic determination of handwriting styles. Our method may be used for any of the applications sketched above, but our focus is on using knowledge about handwriting styles for forensic document examination. Traditionally, the distinguishing characteristics listed above are used to analyze the handwriting from a so- called "questioned document", for, e.g., writer identification or verification purposes [10]. The use of automated technology to support this laborious process has received much interest [22]. However, most research pursues the development of writer identification techniques [13, 17, 18] or writer identification systems [8, 15, 21] rather than the determination of handwriting style.

As explained in [4, 7, 24], assigning handwriting to a handwriting style relies on the availability of a set of template characters that are representative for the style. Such a set of representative allographs may be represented in a "top-down" manner, based on knowledge and experiences from forensic document examiners, or based on exemplar character shapes from copybooks [4]. Copybooks describe handwriting styles that are used for the acquisition of handwriting skills, using material that is taught to children. Typically, such styles differ between, and in many cases also within, countries. The work described in [4, 12], presents preliminary, yet promising, results on the automated comparison of copybook styles for the determination of the country of origin. In [6], it is described how native Arabic

writers can be distinguished from non-natives. Both methods employ directional features (respectively by convolving Sobel edge detectors and directional Gabor filters) for this task. However, whereas the latter paper uses these features to yield a two-class distinction by means of support vector machines, the former work performs a one-by-one comparison between the characters segmented from a questioned document to the corresponding characters from each copybook style. This approach, where the similarity between handwriting and a writing style is expressed as a combination of similarities between mutual allographic character variants, is very similar to the work presented in this paper.

In our work, however, the lists of prototypical characters describing handwriting styles are obtained in a "bottom-up", data driven, approach. We have shown that by hierarchical clustering of a large collection of characters, a set of allographs can be obtained that represents the most prominent character shapes from the handwriting of hundreds of writers [16, 24]. The research described in this paper pursues the question how such a list of allographs can be used to distinguish the handwriting from different writers in a number of coherent handwriting styles. Similar to the method described in [4], our method matches the characters written by a writer to a hierarchically structured set of allographs and records the best matching allograph for each character. The resulting *membership vector* is an array containing the frequency of occurrence of each allograph. Our assumption is that if the handwritings from different writers are alike, their membership vectors are similar and thus, that clustering of handwritings represented by such vectors reveals handwriting styles. In other words, writers with similar handwritings have the same allographic prototypes in common and are member of the same handwriting style.

The procedure outlined below is explained in detail in the remainder of this paper. To illustrate the feasibility of our work, lowercase characters were used. However, our methods can handle other characters and alphabets as well [14]. Section 2 describes how a collection of handwriting styles can be generated. Hierarchical clustering is used to generate a relatively large set of allographs from characters selected from the UNIPEN v07_r01-trainset. These allographs are subsequently clustered to yield a hierarchical structure of prototype clusters. Together with a matching process to compute the similarity of a prototype cluster and a character, this structure implements a *membership function*, which can be used to compute membership vectors. We used exhaustive clustering of the membership vectors from handwritings from 41 different writers to yield many instances of handwriting styles. If two of these resulting handwriting styles have exactly the same members, or contain the same allographs, they can be considered similar as well. In Section 3, these two measures are explored for assessing the consistency of the generated handwriting styles.

2. Extracting handwriting styles

Three datasets were used to develop, train, and test our methods. These are the UNIPEN [9] v07_r01-trainset (referred to as *Trainset*), the UNIPEN devtest_r01_v02 (referred to as *Devset*), and the *Plucoll* [24] dataset. Only online characters were used. Below, in Figure 1, the process of generating handwriting styles based on these sets is depicted. In this section, this process is described in detail.

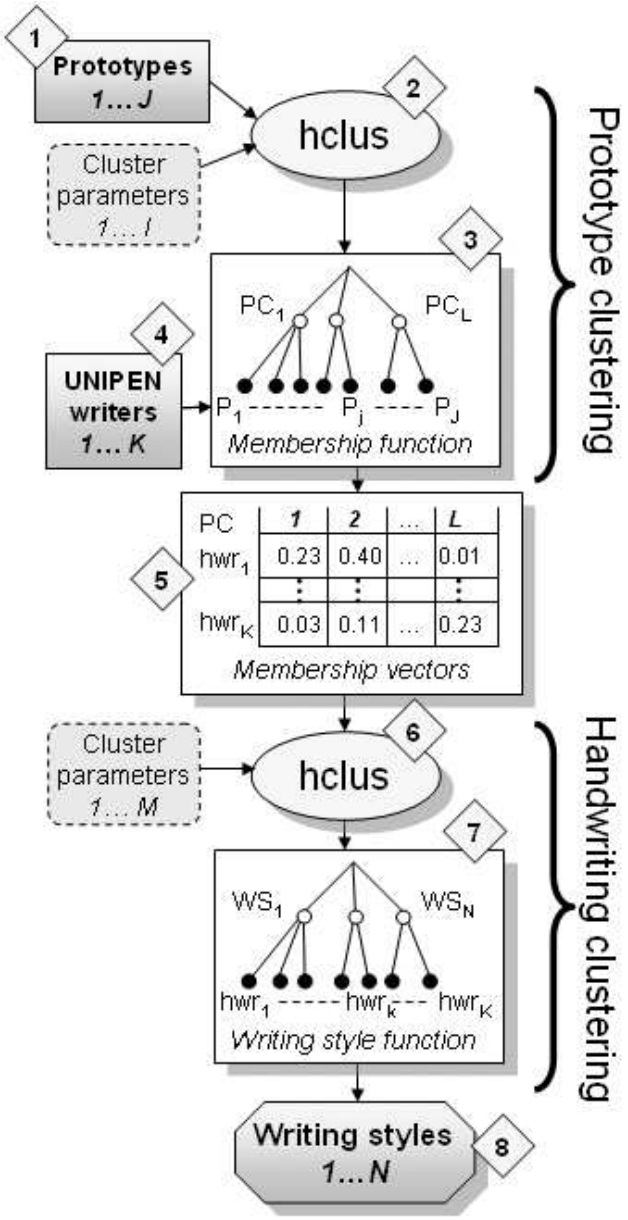

Figure 1. Graphical summary of the writing style creation process.

2.1. Allograph prototype generation

About one third (14.448) of the lowercase characters in the *Trainset* were randomly selected for step(1). Hierarchical clustering was performed for each letter, resulting in 26 cluster dendrograms. This process is described in detail in [16]. For matching two characters, Dynamic Time Warping (DTW) was employed. Using DTW, the distance between two coordinate trajectories can be computed as the average Euclidean distance between each pair of most suitable coordinates. Allographs were manually selected by human experts. The result of this processing step(1) is a list of 1583 allographs.

2.2. Allograph prototype clustering

The number of allograph prototypes generated in the previous step determines the length of the membership vectors used in step(3). Because using a large number of allographs would result in sparse vectors (since only a small proportion of these allographs would occur in a writer's handwriting), the vector length was reduced by allograph clustering (Fig. 1-(2)). Hierarchical clustering (*hclus* [24]) was used for this purpose. The result is a hierarchical organization of *prototype clusters* (Fig. 1-(3)). As described in [24], several parameters can be controlled to rule the outcomes of the clustering process. Typically, the number of resulting clusters, the size and variance of clusters and the arity of nodes from the dendrograms, ruled the different outcomes. Furthermore, as is well-known from cluster analysis, the selection of clusters from a dendrogram can be performed in various ways [1]. This involves that, depending on parameter settings and cluster selection criteria, the resulting prototype clusters can vary in both number and content. However, all leaves P_i of each dendrogram remain the same for each clustering.

2.3. Computing membership vectors

Both the *Trainset* and the *Devset* datasets were used to select $K = 43$ writers who wrote at least 5 instances per lowercase letter from the alphabet. For each writer, a membership vector was computed through a *membership function*. This membership function assigned each character in a persons writing to a prototype cluster.

Note that this was not done by finding the matching cluster centroid, as is done in [24]. In stead, this was done by matching each character to all leave allographs P_i (using DTW as matching function). Each character was then assigned to the cluster of which the best matching leave allograph was a member. Figure 2, in which level 0 represents the leave level, shows that prototype clusters could be selected at different levels, resulting in membership functions of different length.

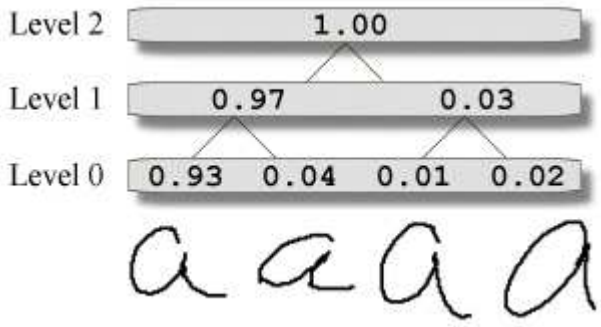

Figure 2. Example of an allograph cluster (at Level(2)), containing multiple allographs. The level (0) allographs are used for matching and reflect the relative frequency of prototypical *a* in a persons handwriting.

2.4. Generating copybooks containing handwriting styles

To generate *handwriting styles*, the membership vectors yielded by the previous step were clustered, again using the *hclus* algorithm (Fig. 1-(6)). The match between membership functions was computed using Euclidean distances. Each node of the resulting dendrograms represents a handwriting style. Note that the clusters selected from each clustering can be considered as a *copybook* containing handwriting styles. As described in the next section, many different clusterings, performed on different sets of membership functions and different cluster parameter settings, were assessed to yield consistent handwriting styles.

2.5. Assessing consistent handwriting styles

For clustering, we employed a modified hierarchical clustering algorithm which has shown to yield a hierarchical organization that reflects the true structure occurring in the data [24]. As described above, different parameter settings yield different clusterings. We performed exhaustive clustering of the membership functions resulted in copybooks differing in size and contents of handwriting styles. From 240 random clusterings 240 copybooks were selected, each containing a number of handwriting styles. However, overlap was present between some of the styles in the different copybooks. This overlap can be expressed by (i) the number of writers that are shared between handwriting styles and (ii) by considering the frequency of occurrence of allographs contained in handwriting styles.

From the *Plucoll* set, the handwritings of 41 writers were selected to generate membership vectors as described in

Section 2.3. So, the handwriting of each writer resulted in a membership vector which subsequently can be compared to handwriting styles (as depicted in (Fig. 1-(7)). Classifying membership vectors into handwriting styles was performed by computing the Euclidean distance between the leave membership vectors from each WS_i. For any new writer, this method can be used to assign handwriting styles to the handwriting of that writer.

Stated otherwise, each handwriting style can be described by the writers it contains. This classification provides us with a description of each writing style from the copybooks, in terms of assigned Plucoll-writers. Two writing styles were considered identical, if the same group of writers was assigned to them both. The number of identical writing styles was counted, and styles with the highest frequency were considered as being the most consistent ones. Fig. 3 illustrates the 3 most consistent handwriting styles extracted from the handwriting of 43 writers. Each style shares exactly the same writers and occurs in multiple copybooks. As depicted in Fig. 3, we are attempting to provide meaningful names to each style. It is apparent that the most prominent styles emerging from our method correspond to the well-known broad categories cursive, mixed, and print.

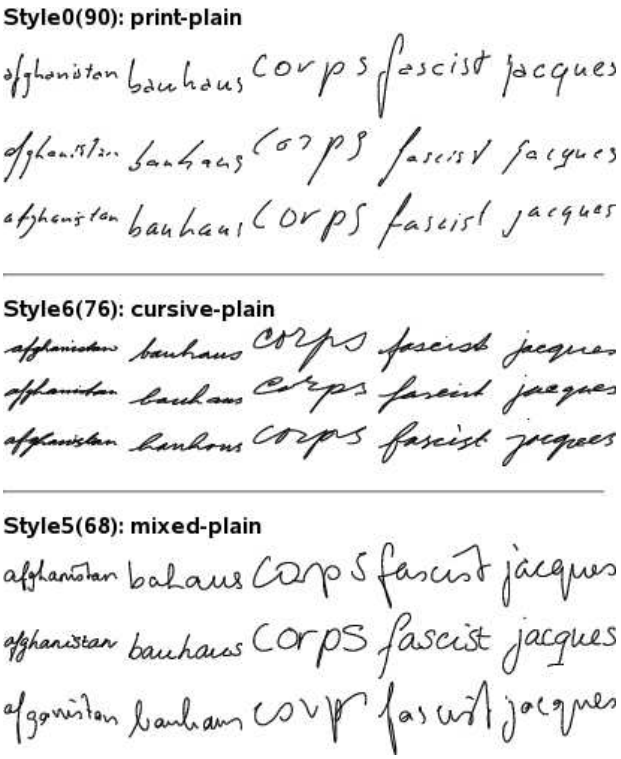

Figure 3. Examples of handwritings contained in the three most consistent handwriting styles.

Another way of assessing the consistency of extracted handwriting styles is by determining the most distinctive allograph prototypes, i.e., the most important cells from the membership vectors contained in a handwriting style. This notion of characteristic shapes that distinguish handwriting styles is very common in forensic writer identification. To illustrate how handwriting styles can be assessed in this manner, we selected three letters that are known to be discriminative in Western handwriting: forensic experts [19] as well as recent findings described in [4], indicate that the letters 'k', 'r' and 't' are known to make this distinction. For the three handwriting styles depicted in Figure 3, the frequency of occurrence of a prototype cluster for the letter l was determined as n_{Pl}/n_l, where n_{Pl} is the number of occurrences of prototype cluster P_l and n_l the number of occurrences of the letter l in the handwritings belonging to a handwriting style. Figure 4 depicts the prototype occurrence for each of the three handwriting styles.

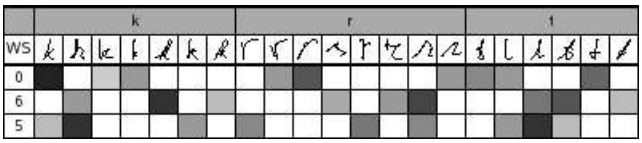

Figure 4. Prototype occurrence for the letters 'k', 'r' and 't' of the three handwriting styles depicted in 3. Black cells indicate a fraction of 1, white cells a fraction of 0.

We consider such a visualization as a proper tool to assess the properties of our extracted handwriting styles. First, the similarity of styles is shown by columns with similar prototype occurrence, marked by cells with similar grey values. Second, the discriminative power of prototypes is marked by the intensity of each cell. And third, this tool can be used to highlight characteristic prototypes occurring in the handwriting of an unknown writer. The latter option can be valuable in forensic writer search applications [4].

3. Discussion

We have presented a novel procedure for extracting handwriting styles from the handwritings of different writers. We have argued that handwriting can be described by the occurrence of prototypical characters and that by clustering different handwritings, consistent handwriting styles can be obtained. Our method is data driven, employing hierarchical clustering and a character matching function (DTW) to: (i) determine a set of allographs, (ii) cluster these allographs to build prototypical allographs, (iii) compute so-called *membership vectors* indicating the frequency of

occurrence of prototypical allographs in new, unseen handwriting, and (iv) derive copybooks comprising handwriting styles by clustering these membership vectors.

The domain of our research is forensic writer identification, but the determination of handwriting styles can be used for applications like handwriting recognition and synthesis as well. We consider the results presented here as promising, but there many challenging opportunities for further research.

For example, the attempt to adorn clusters of handwriting styles with symbolic, meaningful names is a process that has our ongoing attention. We are discussing these results with forensic experts. Second, we have presented a tool to assess the discriminative power of allograph prototypes and use this as a similarity measure for comparing different handwritings. Forensic handwriting experts traditionally use such lists of discriminative characters and the accuracy of this method for writer search remains to be explored.

In our tests, we used pre-segmented on-line data, which is often not available in the forensic practice. Techniques exist, however, to generate this data automatically from offline data [16] and often interactive sessions with human experts can be performed. Furthermore, the underlying idea of style clustering can also be applied directly to off-line data, using matching techniques for off-line data.

The proposed technique is developed within the Trigraph project [15], which aims at improving the reliability of automatic writer identification programs for the forensic practice. We consider the techniques described in this paper as a new and promising step in the right direction.

References

[1] V. Barnett. *Interpreting multivariate data*. John Wiley & Sons, New York, 1981.

[2] A. Bharath, V. Deepu, and S. Madhvanath. An approach to identify unique styles in online handwriting recognition. In *Proc. ICDAR 2005*, pages 775–778, Seoul, Korea, 2005.

[3] V. Blankers and R. Niels. Writer identification by means of loop and lead-in features. In *Proc. NSVKI*, Nijmegen, 2007.

[4] S.-H. Cha, S. Yoon, and C. C. Tappert. Handwriting copybook style identification for questioned document examination. *Journal of Forensic Doc. Examination*, 17:1–14, 2006.

[5] K. Chellapilla, P. Simard, and A. Abdulkader. Allograph based writer adaptation for handwritten character recognition. In *Proc. 10th IWFHR*, pages 423–428, La Baule, France, 2006.

[6] F. Farooq, L. Lorigo, and V. Govindaraju. On the accent in handwriting of individuals. In *Proc. 10th IWFHR*, La Baule, France, 2006.

[7] G. Fink and T. Plotz. Unsupervised estimation of writing style models for improved unconstrained off-line handwriting recognition. In *Proc. 10th IWFHR*, pages 429–434, La Baule, France, 2006.

[8] K. Franke, L. Schomaker, C. Veenhuis, L. Vuurpijl, M. van Erp, and I. Guyon. Wanda: A common ground for forensic handwriting examination and writer identification. In *ENFHEX news*, volume 1, pages 23–47, 2004.

[9] I. Guyon, L. Schomaker, R. Plamondon, M. Liberman, and S. Janet. UNIPEN project of on-line data exchange and recognizer benchmarks. In *Proc. ICPR'94*, pages 29–33, Jerusalem, Israel, October 1994.

[10] R. Huber and A. Headrick. *Handwriting identification: facts and fundamentals*. CRC Press, Boca Raton, Florida, 1999.

[11] Z. Lin and L. Wan. Style-preserving English handwriting synthesis. *Pattern Recognition*, 40(7):2097–2109, 2007.

[12] M. Manfredi, S. Cha, S. Yoon, and C. Tappert. Handwriting copybook style analysis of pseudo-online data. In *Proc. IGS2005*, pages 217–221, Salerno, Italy, June 2005.

[13] A. Namroobi and S. Gupta. Text independent writer identification from online handwriting. In *Proc. 10th IWFHR*, pages 287–292, La Baule, France, 2006.

[14] R. Niels and L. Vuurpijl. Dynamic Time Warping applied to Tamil character recognition. In *Proc. ICDAR2005*, pages 730–734, Seoul, Korea, August-September 2005.

[15] R. Niels, L. Vuurpijl, and L. Schomaker. Introducing Trigraph - trimodal writer identification. In *Proc. European Netw. of Forensic Handwr. Experts*, Budapest, 2005.

[16] R. Niels, L. Vuurpijl, and L. Schomaker. Automatic allograph matching in forensic writer identification. *Int. J. on Patt. Recognition and AI*, 21(1):61–81, 2007.

[17] A. Schlapbach, V. Kilchherr, and H. Bunke. Improving writer identification by means of feature selection and extraction. In *Proc. ICDAR 2005*, pages 131–135, 2005.

[18] L. Schomaker and M. Bulacu. Automatic writer identification using connected-component contours and edge-based features of upper-case western script. *PAMI*, 26(6):787–798, 2004.

[19] L. Schomaker and L. Vuurpijl. Forensic writer identification: A benchmark data set and a comparison of two systems. Technical report, NICI, 2000.

[20] S. Srihari, S. Cha, and S. Lee. Establishing handwriting individuality using pattern recognition techniques. In *Proc. 6th ICDAR*, pages 1195–1204, Seattle, USA, 2001.

[21] S. Srihari, C. Huang, and H. Srinivasan. A search engine for handwritten documents. *Document Recognition and Retrieval*, XII:66–75, 2005.

[22] S. Srihari and G. Leedham. A survey of computer methods in forensic document examination. In *Proc. IGS2003*, pages 278–282, Phoenix, USA, November 2003.

[23] L. Vuurpijl and L. Schomaker. Coarse writing-style clustering based on simple stroke related features. In *Proc. 5th IWFHR*, pages 29–34, Colchester, UK, 1996.

[24] L. Vuurpijl and L. Schomaker. Finding structure in diversity: A hierarchical clustering method for the categorization of allographs in handwriting. In *Proc. ICDAR4*, pages 387–393. IEEE Computer Society, Aug. 1997.