Writer identification – clustering letters with unknown authors

Joanna Putz-Leszczynska

Writer identification – clustering letters with unknown authors

Joanna Putz-Leszczynska
Warsaw University of Technology,
Faculty of Electronics and Information Technology
Nowowiejska 15/1900-665 Warsaw, Poland,
jputz@elka.pw.edu.pl

Abstract. This paper provides a simple algorithm for writer identification of historical letters. The collected database is an original historical database, of 100 pages belonging to 25 people selected from a 500 letters database. In the article there is presented an article for a cauterization of the letters, because the system doesn't have the templates of the classes and doesn't know how many classes is in the database. The obtained result shows that automatic identification can help historical experts to segregate the documents, before they would analyze the text information.

1. Introduction

Historical archives are an extensive collections of handwritten documents. A significant portion of these archives are being scanned and stored in electronic form in order to facilitate research. Many of the documents are not signed or associated with any author, whereas such association could be of benefit for researchers like historians or genealogists.

The aim of this work was to verify the effectiveness of automatic separation/ grouping by author of handwritten historical documents. In the literature, one can find a number of items related to verification of identity based on text (R. Messerli, H. Bunke) rather than a signature, but most of them work in controlled conditions - same ink color, guides etc . The present study, using the results so far published are a step further and examine whether these algorithms can to work on real pieces of writing, created under varying conditions, where the authors were in different positions, different places and different times of writing.

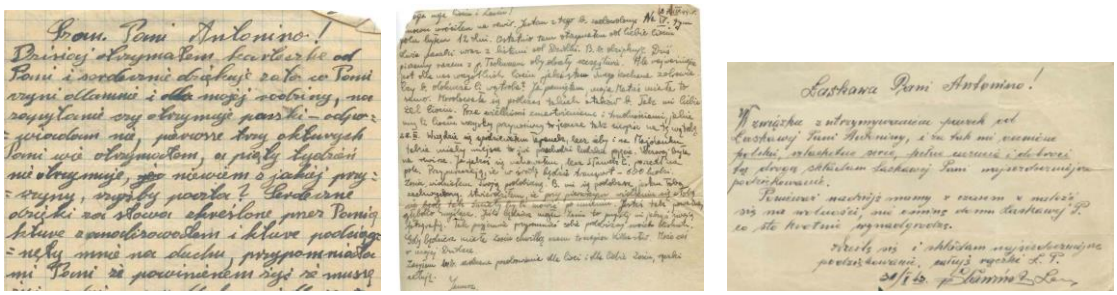

Figure 1 Examples of letters

In this paper, the research used a collection of approx. 500 letters - secret letters written from the Nazi concentration camp at Majdanek . This is one of many collections, which would facilitate an automatic segregation analysis and work with others. Some of the letters are quite clear and organized , written on a piece of lined paper (Figure 1). Others are more disorganized, where the disorder stems from lack of guides, or additional text which should be exempt from the characteristics extraction.

2. Database

As part of the work, 416 pages of secret messages from Majdanek have been scanned in 600 dpi and 300 dpi. In the second step, 25 classes have been selected, with 4 scans representing each scan. Only a part of a database was used, because only for this letters the 'clusering' by the human expert was done. The rest of the letters were postponed for further study as difficulties in identifying the class arose. An extended study would require assistance from handwriting experts. Finally the 300 dpi scans were used for study. The calculations were faster and 600 dpi did impact the results of the verification in a significant way.

3. Identification algorithm

The present algorithm consists of the following steps:

1. Pre- processing, where a color image is converted to a number of glyphs represented as binary image
2. Feature extraction, where using morphological operations are used to obtain 32 characteristics for each glyph.
3. Comparison based on clusters similarity distance

3.1. Pre-processing

The result of the preprocessing are glyphs, which are later used for feature extraction. An image in the RGB space is converted to a grayscale image. Next, **binarization** is performed using a dynamic threshold, which is determined for each image based on the mean value determined for this picture based on the gray-scale image. Next, **correction of image orientation** is performed.

For line segmentation, the author has decided to use a signal of the number of black pixels in each row of the image - ***lp*** as a function of ***r*** rows of the image (feature used in signature verification and proposed in [4]). This function has also been used to correct the orientation. To this end, for each image, a set of two graphs ***lp*** were calculated :

- Right hand side of the scan : **:** $lp_p(r)$ - black area on the scan Figure 5
- Left of the scan : $lp_l(r)$ - blue area on the scan Figure 5

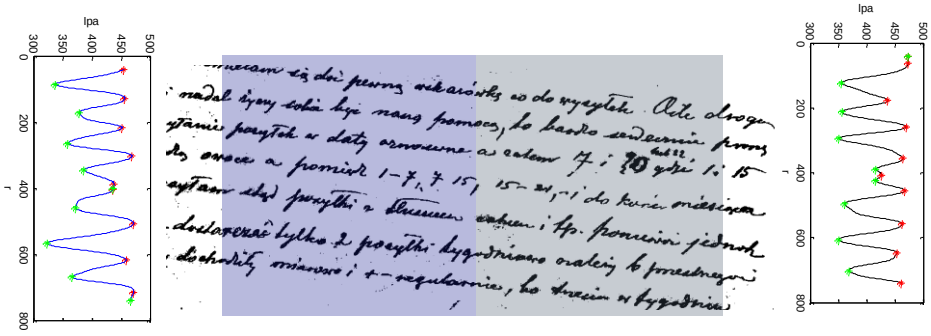

Figure 2 $lp_p(r)$ - black, $lp_l(r)$– blue.

Each signal was smoothed using moving average over the signal. This algorithm has also been successfully used in studies of gait biometrics . This step simplifies the extremes detection in the signal. Equation (1) describes a moving average algorithm:

$$lpa(r) = \sum_{i=-k}^{k} w_i lp(r+i) \qquad (1)$$

where:

2k + 1- the width of the time window
wi – samples weight
lp(r) - original data value at time t
lpa(r) - smoothed data value at time t

Each ***lpa*** signal is converted into an extremes vector (signal). The correction algorithm consists of choosing such a rotation angle where distance calculated using Dynamic Time Warping (DTW) between the signal extrema positions for the left and right side will be the lowest :

$$\vartheta = \arg\min_{\vartheta} \boldsymbol{D}(ex_l, ex_p) \qquad (2)$$

where ***D*** is the dissimilarity calculated using DTW .

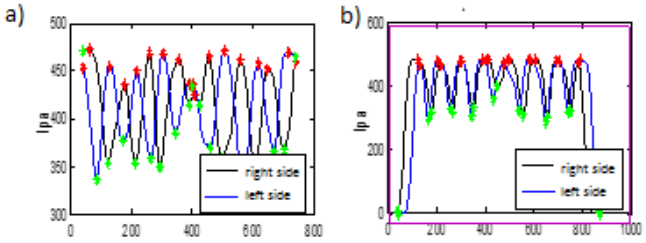

Figure 3 a) lpa signals for 0 degree b) lpa signals for 4 degree- the optimal correction

The result of this minimization can be seen in Figure 4 - the lowest value of ***D*** = 1734 is reached for 4 degree rotation. As a result of experiments, a -5 to 5 degrees range was determined.

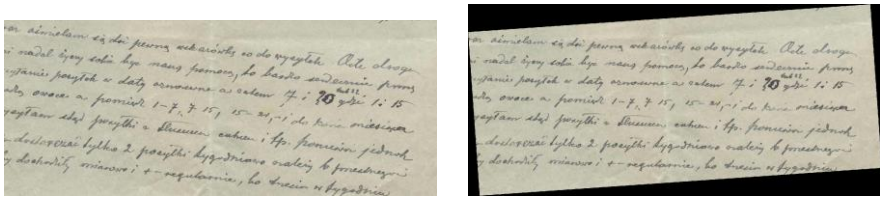

Figure 4 left: before correction, right: after correction

Then, the ***lp*** function calculated for the whole image is used for row segmentation using an experimentally set threshold (Fig 5).

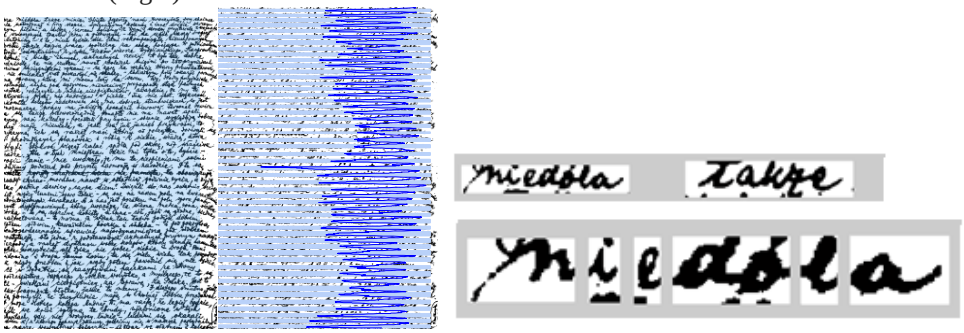

Figure 5 left: image segmentation into rows, right: rows segmentation into symbols

The rows segmentation into glyphs is realized by plotting the number of pixels, but this time as a function of the column number . The threshold, which is the parameter of the method determines whether the glyphs are whole words or individual characters/ groups of characters (Figure 5). After testing, the author decided to use the one that separates into individual characters – on average 215 symbols per page.

3.2. Feature extraction

A large number of feature calculation in off-line handwriting and signature verification has been proposed in the literature. Some are based on global features such as height or width of the symbol , others on the characteristics of texture. Some approaches try recreating the time of the formation of individual pen strokes and thus go to the field of signal processing (S. Chen and S. Srihari, P.S. Deng, H.-Y. Liao, B. Fang, C.H. Leung).
The aim of the study was to verify whether or not user grouping is possible in real data . For this reason, the author chose the features proposed in the paper (J. Fierrez-Aguilar at al.), which was further elaborated by the author in (Putz et al.). The proposed approach uses morphological operations. For each glyphs, features are determined by the steps of :

a) Dilation- feature is the number of pixels lit after morphological dilation. Dilation of that element is performed five times and each time the number of pixels is recorded. As a result, a single structural element is used to designate exactly five features. Thus, as a result of operations using 4 structural elements, we get the 20 features.
b) Erosion - feature is calculated as the number of pixels lit after morphological erosion. One structural element is used exactly once per original symbol giving only one feature . Hence, for the 16 structural elements we get 16 features.

In summary, the scan is converted to a set of glyphs, each represented by 36 features. This can be regarded as a collection of points in 36 dimensions.

3.3. Comparison

The comparison measure denotes a similarity between sets of clusters. Each scan, or a collection of points in 36 dimensions, is subjected to clustering using K-means. A clustering method with a preset number of clusters was selected deliberately, based on the assumption that the number of clusters, or groups of glyphs for handwriting in general should be constant. The task here is to compare two sets of clusters - one of which belongs to a scan looking for class, the second is a representative of the class to which it is compared. In the paper (M Hayvanovych et al.) has proposed a method of comparing symmetric clusters. In the presented solution it was decided to propose the asymmetric form of the formula. The reasoning was that in the case being considered, the first set of clusters suspected of belonging to a class $C_q = \{S_1, S_2, \dots, S_K\}$ is compared to the second set of clusters (class representative) $C_j = \{S'_1, S'_2, \dots, S'_K\}$. The value of dissimilarity between to clusters is determined as the sum of distance between centroids of clusters assigned to each of the two sets. In other words, for each cluster

from a set of scan verified $S_i \in C_q$, the distance is determined in 36 -dimensional space between the centroid and centroid S_i nearest cluster from a set of clusters being compared C_n class . Finally, the dissimilarity value is:

$$d(C_q, C_n) = \sum_{i=1}^{K} \min_{j=1,\ldots,K} |S_i, S'_j| \qquad (3)$$

3.4. Results

The following tests were carried out using two indicators :

a) EER (Equal Error Rate) – equal error rate of false acceptance and false rejection

b) ANE (Accepted No Errror) - indicating the effectiveness of the correct assignment element (glyph) to the group in the absence of misallocation the group

Tests were conducted to select the best parameters. Here I present one that shows the relation of number of clusters to identification efficiency. As it is presented here, the best results were obtained for 2 clusters.

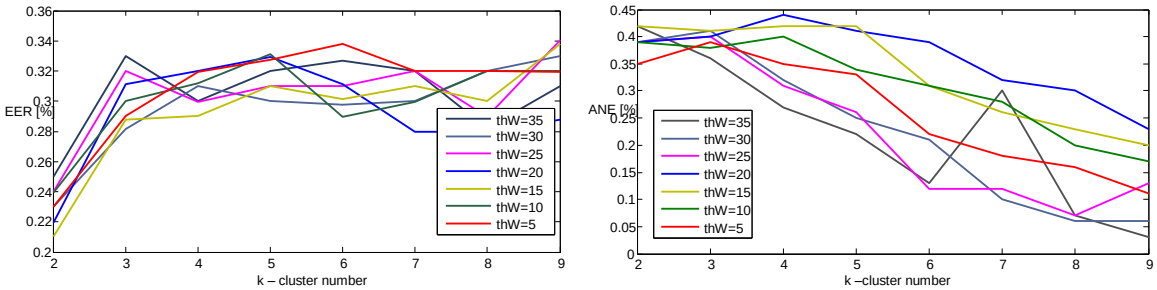

Figure 6: The EER and ANE results for different thW and cluster number.

Additionally, plots for different thW are presented – it is visible that the low thW gives the best results – the individual letters/ groups of letters. The best results obtained are EER ~ 20 % and ~ 45% of the ANE. Both results are very good. In particular, the EER result demonstrates a correct implementation- the result is similar to the ones reported in literature for handwritten signature verification are at this level. The ANE 45% success rate means that almost half of the scans were assigned properly without committing an error.

4. Summary

An algorithm was proposed, implemented in a computer program used to categorize handwritten documents. From the collection of 500 letters, secret messages from the Nazi concentration camp, 100 were selected, belonging to 25 people (4 for each person). The proposed algorithm was applied on the scanned letters, leading to the transformation of a letter in a set of glyphs, then used one of the many well-known approaches for determining the characteristics of the handwriting, features based on morphological transformations. The calculated features were used in a comparison algorithm based on the grouping of clusters. The results achieved error-free or 50% of the group assignments are a good prelude to broader studies involving forensic experts involved in writing, who would do a handmade categorization of the current base, making it possible to use the other 400 letters.

References

H. Baltzakis and N. Papamarkos. A new signature verifcation technique based on a two-stage neural network classifier. Engineering Applications of Artifcial Intelligence, 14:95–103, 2001.

S. Chen and S. Srihari. Use of exterior contours and shape features in off-line signature verification. Document Analysis and Recognition, 2005, Proceedings. Eighth International Conference on, pages 1280–1284, 2005.

P.S. Deng, H.-Y. Liao, C.W. Ho, and H.-R. Tyan. Wavelet-based off-line signature verification. Computer Vision and Image Understanding, 76(3):173–190, 1999.

B. Fang, C.H. Leung, Y.Y. Tang, K.W. Tseb, P.C.K. Kwokd, and Y.K. Wonge. Offline signature verifcation by the tracking of feature and stroke positions. Pattern Recognition, 36:91–101, 2003.

J. Fierrez-Aguilar, N. Alonso-Hermira, G. Moreno-Marquez, and J. Ortega-Garcia. An off-line signature verification system based on fusion of local and global information. Workshop on Biometric Authentication, Springer LNCS-3087, pages 295–306, 2004.

M Hayvanovych, M. Magdon-Ismail, Measuring Similarity between Sets of Overlapping Clusters , Social Computing (SocialCom), 2010 IEEE Second International Conference on, pages 303 – 308, 2010

R. Messerli, H. Bunke, Writer identification using text line based features Document Analysis and Recognition. Proceedings. Sixth International Conference on, pages 101 – 105, 2001

J. Putz-Leszczynska, M. Chochowski, L. Stasiak, R. Wardzinski, and A. Pacut, Two-stage classifier for off-line signature verification, 13th Biennial Conference of the International Graphonomics Society, Melbourne, Australia, pages 138–141, 2007.