A Set of Novel Features for Writer Identification

Caroline Hertel and Horst Bunke

Department of Computer Science, University of Bern
Neubrückstrasse 10, CH-3012 Bern, Switzerland
`bunke@iam.unibe.ch`

Abstract. A system for writer identification is described in this paper. It first segments a given page of handwritten text into individual lines and then extracts a set of features from each line. These features are subsequently used in a k-nearest-neighbor classifier that compares the feature vector extracted from a given input text to a number of prototype vectors coming from writers with known identity. The proposed method has been tested on a database holding pages of handwritten text produced by 50 writers. On this database a recognition rate of about 90% has been achieved using a single line of handwritten text as input. The recognition rate is increased to almost 100% if a whole page of text is provided to the system.

Keywords: personal identification; handwriting analysis; writer identification; feature extraction; k-nearest neighbor classifier.

1 Introduction

The identification of persons based on biometric measurements has become a very active area of research [1, 2, 3]. Many biometric modalities, including facial images, fingerprints, retina patterns, voice, signature, and others have been investigated. In the present paper we consider the problem of personal identification using samples of handwritten text. The objective is to identify the writer of one or several given lines of handwritten text. In contrast with signature verification [4] where the identity of an individual is established based on a predefined, short sequence of characters, the methods proposed in the present paper are completely text-independent. That is, any text consisting of one or a few lines may be used to establish the identity of the writer. In particular, we don't suppose that the meaning (i.e. the ASCII transcription) of the given handwritten text is known. In contrast with signature verification, which is often performed in the on-line mode (where the writer is connected to the system via an electronic pen or a mouse and the writing is recorded as a time-dependent process) we assume off-line handwritten text as input modality. That is, only an image of the handwriting is available, without any temporal information. Applications of the proposed approach are forensic writer identification, the retrieval of handwritten documents from a database, or authorship determination of historical manuscripts.

For a survey covering work in automatic writer identification and signature verification until the end of the 1980's see [4]. An extension including work until 1993 has been published in [5]. In [6] a system for writer identification using textural features derived from the grey-level co-occurrence matrix and Gabor filters is described. For this method whole pages of handwritten text are needed. Similarly, in [7, 8] a system for writer verification is described. It takes two pages of handwritten text as input and determines if they have been produced by the same writer. The features used to characterize a page of text include writing slant and skew, character height, stroke width, frequency of loops and blobs, and others. Morphological features obtained from transforming the projection of the thinned writing are computed in [9]. In this approach only single words are used to establish the identity of a writer.

In contrast to [4, 5, 6, 7, 8, 9] the method proposed in [10] works on an intermediate level using text lines as basic input units from which features are computed. In the current paper a continuation and extension of this work is presented. The novel contribution of the paper is a significantly extended set of features that are suitable to characterize an individual's handwriting. This new set of features has been tested on a data set that is an extension of the data set described in [10] from 20 to 50 writers. On this extended data set a recognition rate of about 90% has been achieved using a single line of handwritten text as input. The recognition rate is increased to almost 100% if a whole page of text is provided to the system.

The remainder of the paper is organized as follows. In the next section the new set of features is introduced. Then a series of experiments with the new features are described in Sect. 3. Finally, conclusions are drawn in Sect. 4.

2   Features for Writer Identification

In this paper four groups of novel features for writer identification are introduced. They will be described in the following four sub-sections. Additionally some features that have already been used in other systems [7, 8, 10] will be briefly sketched in Sect. 2.5. Throughout this section we assume that a page of handwritten text has been segmented into individual lines. The segmentation methods used in the present paper are the same as those described in [10].

2.1   Connected Components

Some people tend to write a whole word in a single, continuous stroke, while others break up a word into a number of components. The features introduced in this sub-section attempt to model this behavior.

From the binary image of a line of text, connected components are extracted first. Each connected component C is described by its bounding box $(x_1(C), y_1(C), x_2(C), y_2(C))$, where $(x_1(C), y_1(C))$ and $(x_2(C), y_2(C))$ are the coordinates of the left-lower and right-upper corner of the bounding box of C,

Fig. 1. Five sample text lines from the data set used in the experiments; these lines have been produced by different writers

respectively. Given all connected components of a line of text, the average distance between two successive bounding boxes is computed first. For this purpose we order all connected components according to their x_1-value. Given the ordered list $(C_1, C_2, \cdots, C_n)$ we calculate the average value of $(x_1(C_{i+1}) - x_2(C_i))$. This quantity is used as a feature that is potentially useful for writer discrimination. The next two features are the average distance of two consecutive words and the average within-word distance of connected components. In order to compute these two features, a clustering procedure is applied that groups connected components together if they are likely to belong to the same word. This clustering procedure uses a threshold t on the distance of two consecutive connected components, C_i and C_{i+1}. If $(x_1(C_{i+1}) - x_2(C_i)) < t$ then it is assumed that C_i and C_{i+1} belong to the same word. Otherwise, C_i is considered to be the last component of a word w_j and C_{i+1} the first component of the following word w_{j+1}.

Other features derived from connected components are the average, median, and standard deviation of the length $(x_2(C) - x_1(C))$ of connected components C in a line of text, and the average number of black-to-white transitions within each connected component.

2.2   Enclosed Regions

If we analyze the closed loops occurring in handwritten text we observe certain properties that are specific to individual writers. For example, the loops of some writers are of circular shape while the loops of other writers tend to be more elliptical. To simplify our computational procedures, we don't analyze the loops directly, but the blobs that are enclosed by a loop. These blobs can be easily computed by standard region growing algorithms. For a graphical illustration see Figs. 1 and 2. In Fig. 1 five lines of text from the database used in the experiments described in Sect. 3 are shown. They have been produced by different writers,
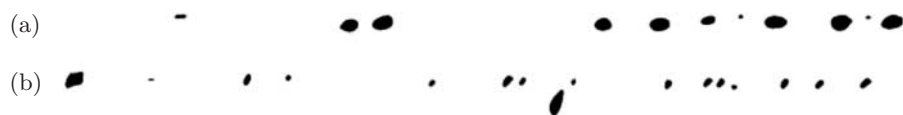
(a)

(b)

Fig. 2. Blobs enclosed by loops: (a) extracted from the first line in Fig. 1; (b) extracted from the second line in Fig. 1

as can be readily seen. In Fig. 2 the blobs enclosed by the loops corresponding to the first two text lines are displayed. One notices a clear difference in the shape of these blobs. In the next paragraph we describe features derived from such blobs.

The first feature is the average of the form factor $f = 4A\pi/l^2$ taken over all blobs of one text line, where A is the area of the blob under consideration and l the length of its boundary. The second feature is similar. It measures the roundness $r = l^2/A$ of an object. Again the average over all blobs in a line of text is taken. The last feature is the average size of the blobs in a text line.

2.3   Lower and Upper Contour

The lower (upper) contour of a line of text is defined as the sequence of pixels obtained if only the lower(upper)-most pixel in each column of the text image is considered. Obviously, if there are gaps between words or parts of a word in the text, these gaps will be present in the lower (upper) contour as well. Gaps of this kind are eliminated by simply shifting the following pixels of the lower (upper) contour by the amount of the gap to the left. After this operation there is exactly one black pixel at each x-coordinate in the lower (upper) contour. However, there are usually discontinuities in the y-coordinates of two consecutive points. These discontinuities are eliminated by shifting the following elements along the y-axis by an appropriate amount. A graphical illustration of this procedure is shown in Fig. 3.

The sequence of pixels resulting from the operations described in the previous paragraph is called the *characteristic lower (upper) contour*. A visual analysis reveals that these characteristic contours are quite different from one writer to another. An example is shown in Fig. 4. Comparing Fig. 4 with Fig. 3 we notice a clear difference between the two contours.

From both the characteristic lower and upper contour of a line of text a number of features are extracted. The first feature is the slant of the characteristic contour. It is obtained through linear regression analysis. The second feature is the mean squared error between the regression line and the original curve. The next two features measure the frequency of the local maxima and minima on the characteristic contour. A local maximum (minimum) is defined as a point on the characteristic contour such that there is no other point within a neighborhood of given size that has a larger (smaller) y-value. Let m be the number of local maxima and l be the length of the contour. Then the frequency of the local
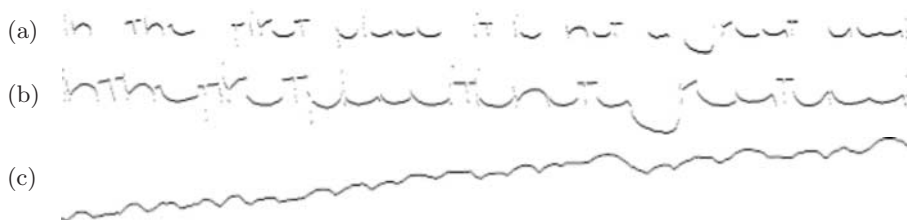
(a)

(b)

(c)

Fig. 3. Illustration of characteristic contour extraction, using the first line in Fig. 1: (a) lower contour; (b) lower contour after gap elimination; (c) lower contour after elimination of discontinuities in y-direction (characteristic contour)

Fig. 4. Characteristic contour of second line in Fig. 1

maxima is simply the ratio m/l. The frequency of the local minima is defined analogously. Moreover, the local slope of the characteristic contour to the left of a local maximum within a given distance is computed, and the average value, taken over the whole characteristic contour, is used as a feature. The same operation is applied for the local slope to the right of a local maximum. Finally, similar features are computed for local minima.

2.4   Fractal Features

In [11, 12] it was shown that methods based on fractal geometry are useful to derive features that characterize certain handwriting styles. While the purpose in those papers was to distinguish between legible and poorly formed handwritings, we take a broader view in this sub-section and aim at features that are useful for writer identification.

The basic idea behind the features proposed in [11, 12] is to measure how the area A (i.e. the number of pixels) of a handwritten text grows when we apply a dilation operation [13] on the binary image. In order to make the features used in this paper invariant with respect to the writing instrument (i.e. stroke width), a thinning operation is applied first. Then the writing is dilated using a disk-shaped kernel of increasing radius $d = 1, 2, \cdots$ and the quantity $\ln(A(d)) - \ln(d)$ is recorded as a function of $\ln(d)$. This function is also called evolution graph [11, 12]. Typically, the evolution graph can be segmented into three parts, each of which behaves more or less linearly. As an example, the evolution graphs derived from the first two lines in Fig. 1 are shown in Fig. 5. The endpoints of the straight line segments of each evolution graph are computed by means
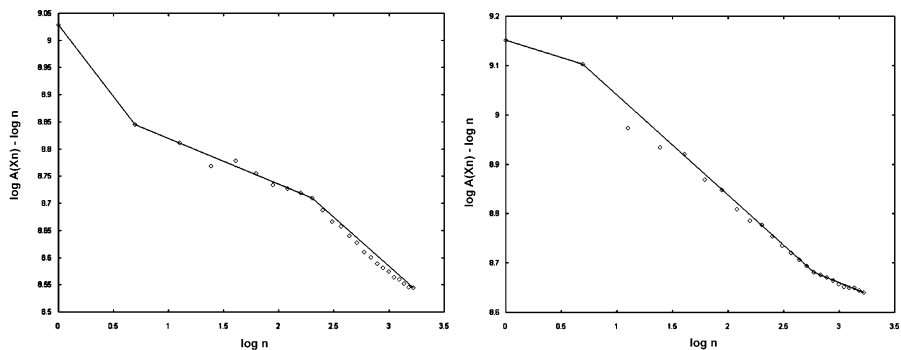
Fig. 5. Evolution graphs and their approximation by three straight line segments, derived from the first line (left) and second line (rigth) in Fig. 1

of an exhaustive search over all possible points on the x-axis. The objective of this search procedure is to minimize the mean squared error between the original points of the evolution graph and the straight line segments used for approximation. Eventually, the slope of each of the three straight line segments is used as a feature to characterize the given handwritten text line. The differences between the two handwriting styles shown in the first two lines of Fig. 1 clearly manifest themselves in the two evolution graphs, and the slopes of the three straight line segments.

Using a disk-shaped dilation kernel results in an evolution graph that is invariant under rotation of the original image. However, the direction of individual strokes and stroke segments is a very important characteristic of a person's handwriting style. Because of this observation, not only disks, but also ellipsoidal dilation kernels are used. To generate this kind of kernels, three parameters are involved, namely the length of the ellipse's two main axes and the rotation angle. Through systematic variation of these parameters a set of 18 dilation kernels are generated. For each of these kernels an evolution graph similarly to Fig. 5 is derived and the slope of the three characteristic straight line segments is computed. Thus a total of $57 \, (= 3 + 18 \times 3)$ different fractal features are generated.

2.5   Basic Features

In addition to the features described in the previous sections, some features that were used already in [7, 8, 10] were included in the experiments described in Sect. 3. These features correspond to writing skew and slant, the height of the three main writing zones, and the width of the writing. Computational procedures for extracting these features from a line of text can be found in [10].

3   Experimental Results

The experiments described in this section are based on the IAM database [14]. This database comprises about 1'500 pages of handwritten text, produced by over 500 writers. A subset of 250 pages written by 50 writers was selected. Each writer contributed 5 pages of text. One page comprises about 8 lines of text on the average. The total data set consists of 2185 lines of text.

The original data format is a binary image for each page of handwritten text. From these images the text lines were extracted first. The corresponding procedures are described in [14]. From each individual line of text, the features introduced in Sect. 2 were extracted. As the ranges of the individual features are quite different, a feature normalization procedure was applied resulting in features that all have zero mean and a standard deviation equal to one.

Out of a potentially large number of classifiers, a simple, Euclidean-distance based 5-nearest-neighbor (5-NN) classifier was adopted for the experiments. This classifier determines the five nearest neighbors to each input feature vector and decides for the class that is most often represented. In case of a tie, the class with the smallest sum of distances is chosen. The number of nearest neighbors to be taken into account was experimentally determined. The advantage of this classifier is its conceptual simplicity and the fact that no classifier training is needed.

In the experiments the whole set of handwritten text lines was split into five portions of equal size. One portion was used as test set and the other four as prototypes for the 5-NN classifier. This procedure was repeated four times such that each portion was used once as test set, and the average recognition rate obtained from these five runs was recorded.

A summary of our experimental results is provided in Table 1. In order to see how the proposed method for writer identification scales up with a growing number of classes, i.e. writers, the experiments were not only run on the full data set produced by 50 writers, but also on a subset that came from 20 writers. Each of the groups of features introduced in Sect. 2 was tested individually (see the first five rows in Table 1). Additionally the union of all these features was tested (see row six).

If we compare the individual groups of features with each other we notice that the blob features yield the lowest classification rate (see 2nd row). A possible explanation of this fact is the rather small number of these features (only three). Next are the connected component based features with a recognition performance of about 53% (31%) for the 20 (50) class problem. The features derived from the characteristic lines are doing quite well and are comparable in performance to the basic features. The best performance among the individual groups of features is achieved by the fractal features. In row six of Table 1, the performance of the union of all features is recorded. On the small data set a recognition rate of 96% is achieved. The performance decreases to about 90% for the case of 50 writers. The small data set is the same as the one used in the work reported in [10]. On this data set a recognition rate of about 88% was achieved with a nearest-neighbor classifier in [10], using a simpler set of features

Table 1. Correct recognition rate for various sets of features

features	20 writers	50 writers
connected components	53.6	31.8
enclosed regions	36.0	18.4
lower and upper contour	76.0	52.8
fractal features	92.6	84.2
basic features	75.6	57.9
all features	96.4	90.7
combination of all lines	100.0	99.6

than the ones employed in this paper. Hence the new set of features proposed in the present paper lead to a clearly improved recognition rate.

In an additional experiment, reported in the last row in Table 1, it was assumed that a whole page of text is written by one single person. In other words, all individual lines on the same page must come from the same writer. Consequently, the results obtained for the individual lines on a page were combined with each other. Simple majority voting was applied to determine the writer of a page. Ties were broken based on the distances output by the individual 5-NN classifiers. Under this combination strategy, a recognition rate of 100% was obtained for the small data set. On the 50-writer data set, all pages but one were correctly assigned, which is equivalent to a recognition rate of 99.66%.

4 Conclusions

Handwriting is a modality that can be used for the identification of persons. In the present paper the problem of text-independent writer identification for the case of off-line handwritten text was addressed. The approach proposed in this paper is applicable as soon as at least a single line of text is available from a writer. Thus it is positioned between other approaches proposed in the literature that use either complete pages of text or just single words. In the present paper a number of novel features have been proposed. These features are rather powerful and lead to quite high recognition rates in two experiments involving 20 and 50 writers, respectively.

There are several applications for which handwriting based person identification is important. Examples include forensic science, handwritten text retrieval from databases as well as digital libraries including historical archives. Another application example is personal handwriting recognition systems that automatically adapt themselves to a particular writer in a multi-user environment. In our future work we want to further upgrade the system described in this paper by including more writers in the database and exploring additional characteristic features. Also the application of feature selection algorithms is of potential interest [15].

Acknowledgment

References

[1] A. K. Jain, R. Bolle, S. Pankanti (eds.). Biometrics. *Personal Identification in Networked Society*, Kluwer Academic. 1999. 679

[2] A. K. Jain, L. Hong, S. Pankanti. Biometrics identification. *Comm. ACM 43(2)*, pp. 91 – 98. 2000. 679

[3] J. Bigun, I. Smeraldi (eds.). Audio- and video-based biometric person authentication. *Proc. of the 3rd Int. Conf. AVBPA*, Halmstadt, Sweden. 2001. 679

[4] R. Plamondon and G. Lorette. Automatic signature verification and writer identification - the state of the art. *Pattern Recognition*, 22, pp. 107–131. 1989. 679, 680

[5] F. Leclerc and R. Plamondon. Automatic signature verification: The state of the art 1989-1993. In *Progress in Automatic Signature Verification* edited by R. Plamondon, World Scientific Publ. Co., pp. 13–19. 1994. 680

[6] H. E. S.Said, G. S.Peake, T. N.Tan and K. D.Baker. Personal identification based on handwriting. *Pattern Recognition*, 33, pp. 149–160. 2000. 680

[7] S.-H. Cha and S. Srihari. Writer identification: statistical analysis and dichotomizer. In Ferrie, F. J. et al. (eds.):*SSPR and SPR 2000*, Springer LNCS 1876, pp. 123–132. 2000. 680, 684

[8] S.-H. Cha and S. Srihari. Multiple feature integration for writer verification. In Schomaker, L. R. B., Vuurpijl, L. G. (eds.): *Proc. 7th Int. Workshop Frontiers in Handwriting Recognition*, pp. 333–342. 2000. 680, 684

[9] E. N. Zois and V. Anastassopoulos. Morphological waveform coding for writer indentification. *Pattern Recognition*, 33(3), pp. 385–398. 2000. 680

[10] U. V. Marti, R. Messerli and H. Bunke. Writer identification using text line based features. *Proceedings of 6th ICDAR*. pp. 101–105. 2001. 680, 684, 685

[11] V.Bouletreau, N.Vincent, R.Sabourin and H.Emptoz. Synthetic parameters for handwriting classification. In *Proceedings of the 4th Int. Conf. on Document Analysis and Recognition*, Ulm, Germany, pp. 102–106. 1997. 683

[12] V.Bouletreau, N.Vincent, R.Sabourin and H.Emptoz. Handwriting and signature: One or two personally indentifiers?. In *Proceedings of the 14th Int. Conf. on Pattern Recognition*, Brisbane, Australia, pp. 1758–1760. 1998. 683

[13] P. Soille. *Morphological Image Analysis*, Springer Verlag, Berlin. 1999. 683

[14] U.-V. Marti and H. Bunke. The IAM-database: an English sentence database for off-line handwriting recognition. In *Int. Journal of Document Analysis and Recognition*, vol. 5, pp. 39-46. 2002. 685

[15] J. Kittler, P. Pudil, P. Somol. Advances in statistical feature selection. In Singh, S., Murshed, N., Kropatsch, W. (eds.): *Advances in Pattern Recognition*, Springer, pp. 425 – 434. 2001. 686