I. Siddiqi, N. Vincent. Writer Identification in Handwritten Documents, 9e International Conference on Document Analysis and Recognition, ICDAR 2007, Brasil.

Writer Identification in Handwritten Documents

Imran Ahmed SIDDIQI, Nicole VINCENT
Laboratoire Systèmes Intelligents de Perception – SIP
Université René Descartes – Paris 5
45, rue des Saints-Pères, 75270 Paris, Cedex 06
{imran.siddiqi, nicole.vincent}@math-info.univ-paris5.fr

Abstract

This work presents an effective method for writer identification in handwritten documents. We have developed a local approach, based on the extraction of characteristics that are specific to a writer. To exploit the existence of redundant patterns within a handwriting, the writing is divided into a large number of small sub-images, and the sub-images that are morphologically similar are grouped together in the same classes. The patterns, which occur frequently for a writer, are thus extracted. The author of the unknown document is then identified by a Bayesian classifier. The system trained and tested on 50 documents of the same number of authors, reported an identification rate of 94%.

1. Introduction

Despite the development of electronic documents and predictions of a paperless world, the importance of handwritten documents has retained its place and the problems of identification and authentification of the writers have been an active area of research over the past few years. Compared to the electronic or printed text, the handwritten text carries additional information about the personality of the person who has written. There exists a certain degree of stability in the writing style of an individual which makes it possible to identify the author for which one has already seen a written text.

The need to identify the author of a document is a recurrent problem that arises often in the court of justice where the authenticity of a document (e.g. a will) has to be concluded [4]. It is posed in the field of medicine as well where the prescription has to come from an authorized person [3] and in banks for the verification of signatures [11]. It can also be used for the analysis of ancient documents to be able to do their indexing and retrieval. We can employ it for the recognition of handwritten text as well, exploiting the principle of adaptation of the system to the type of writer [9].

In this paper, we present a system for offline identification of handwritings. The objective is to find an index of similarity between a document of an unknown author and a document in the reference base, whose writer is known. This is different from the verification model where given two handwriting samples *s1* and *s2*, we would like to determine whether the two samples were written by the same person or by two different people [5].

This paper is outlined as follows: In the first section we give a brief account of some of the proposed approaches for writer identification. In the next part we describe our system and the intermediate steps. The third section discusses the experimental results and finally we give a conclusion and some end remarks.

2. Background

Research in writer identification has received renewed interest in the recent years. A wide variety of features, local or global and structural or statistical, have been proposed that serve to distinguish the writing of an individual from another. Said et al. [7] presents a global approach and considers each handwriting as a different texture, employing Gabor filters and co-occurence matrices. [5] establishes the individuality of handwriting by extracting a set of twelve macro features (document and paragraph level) and 5120 binary micro features (extracted from ten characters). The same research was continued in [2]. In [6] the difference between handwritings is analysed by extracting a set of twelve structural feautures (average height/widht/slope/spacing etc.) for each line of text. The approaches based on the fractal analysis of handwriting include identification by the fractal dimension [10] and by the fractal compression/decompression [3]. The concept of writer invariants introduced in [9] was followed by Bensefia A. [4] who proposed a writer identification system based on the direct correspondence of the graphemes extracted from the two texts to compare. The edge-based directional probability distributions have been used as features in [1] and the identification performance is compared to a number of non-angular features.

3. Proposed Approach

In the recent years, the research on handwritten text has mainly focussed on the extraction of characteristics which are specific to the writer and our method is based on the same approach. We have developed a local approach, higlighting the frequent details in a handwriting by exploting the redundancy of individual patterns within a handwriting. Traditionally, our system can be divided into two stages, first is the offline training of the system to enroll te authorized authors and second is the identification of the writer of a test document.

3.1. Training

The training of the system is carried out to create a reference base by extracting the inherent features of an author. To extract these features, the handwritten text is divided in to a large number of small windows of fixed size *n* x *n*. This size should be large enough to contain ample information about the style of the author and small enough to ensure a good identification performance [3]. The simplest technique would be to divide the entire image regularly from left to right and top to bottom and eliminating the windows which do not contain any part of text. This not only gives a large number of sub-images (containing text pixels), but the windows are also not well positioned over the text, dividing certain features into different sub-images. Thus, we have employed an improved version of the technique proposed in [3] carrying out a component by component division of the text. For each connected component in the text, we fix the vertical origin and move each window from left to right to find the first text (black) pixel. While [3] shifts the entire grid horizontally, we do it for each individual window, thus achieving a better window positioning. The proposed method has been illustrated in figure 1 while the resulting division of the word '*majority*' into sub-images in figure 2.

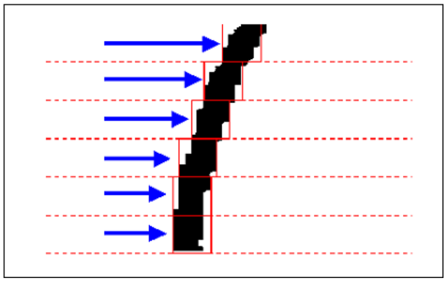

Figure 1. Window positioning on text

The next step is the clustering of the sub-images. The objective is to group similar patterns in the same classes. We have used a simple clustering algorithm that does not need to know a priori the number of clusters to retain. Two sub-images are compared by a correlation similarity measure [4]. We choose a similarity threshold and start with the first sub-image as the centroid of the first class. For each of the subsequent patterns, we calculate the measure of similarity between the current element and the mean of each class. The element is then attributed to the nearest cluster. In case, it is not close to any of the clusters (with respect to the similarity threshold), a new cluster is created.

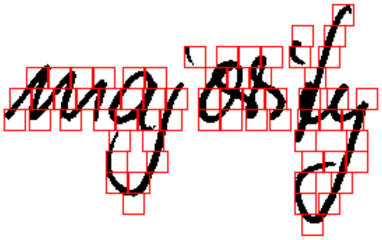

Figure 2. Division of text into sub-images

Once the sub-images have been clustered, we filter the classes to keep only the clusters which have at least Φ elements, Φ being fixed to 5 in our experiments.

$$C = \{C_i \mid card(C_i) \geq \phi\}$$

For each class, we only keep a single representative sub-image, the one which is close to all the other elements of its class with respect to the similarity measure, thus representing a document D as:

$$D_r = \{\overline{x}_i \mid \overline{x}_i = Rpr(C_i)\}$$

Figure 3 shows the features (the representative patterns of each class) that have been extracted from a document image.

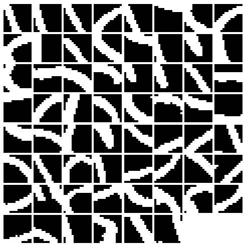

Figure 3. Representative features of a writer

For each class, we also calculate the probability of its occurrence:

$$P(C_i) = \frac{card(C_i)}{\sum_{i=1}^{card(C)} card(C_i)}$$

And the matrix of covariance:

$$Cov_i = (X_i - Mean(C_i))(X_i - Mean(C_i))^T$$

With X_i being the matrix whose columns are the elements of class *i* and *Mean(C$_i$)* being the arithmetic mean of class *i*. Thus, we construct a vector for each class and represent the document by the set of these vectors.

$$F_i = \{Rpr(C_i), P(C_i), Cov_i\}$$

$$D = \{F_i, i \leq card(C)\}$$

These vectors are determiend for each of the writers and thus a reference base is created.

3.2. Identification

In the identification phase, given a handwriting sample *s* whose writer is unknown and samples of handwriting of *N* known writers, the objective is to identify the writer of *s* among the *N* writers [5]. The first step towards identification is the extraction of features from the document whose writer is to be identified. We start with a binarization of the test image followed by the division of text into small sub-images and then their clustering. Once the document is represented by its features we proceed to the classification for which we have employed the Baye's decision theory.

The Bayesian Classifier is based on the assumption that decision problem can be specified in probalistic terms and that all of the relevant probability terms are known. With the posterior probability *P(C$_i$|x),* class conditional probability (likelihood) *p(x|C$_i$)* and the prior probability *P(C$_i$),* the Baye's decision rule can be written as :

$$i_{Bayes} = \arg\max_i(P(C_i|x) = \arg\max_i\, p(x|C_i)P(C_i)$$

The objective is to find the class *i* that maximises the probability of pattern *x* belonging to class *C$_i$*. If the class conditional probability *p(x|C$_i$)* is assumed to have a Gaussian distribution for each class *C$_i$*, we have:

$$p(x/C_i) = \frac{1}{(2\pi)^{d/2}|Cov_i|^{1/2}} \exp(-\frac{1}{2}(X-\mu_i)^T Cov_i^{-1}(X-\mu_i))$$

With μ_i being the mean of class *i* . Thus, we can have:

$$i_{Bayes} = \arg\max_i[\log p(x|C_i) + \log P(C_i)]$$

Taking into account the law of normal distribution and dropping the constants, we have the following expression to maximize:

$$-\frac{1}{2}\log|Cov_i| - \frac{1}{2}(X-\mu_i)^T Cov_i^{-1}(X-\mu_i) + \log P(C_i)$$

To comapre an unspecified handwritten document T (represnted by the frequent patterns of its writer), with a document D in the reference base, we define the following index of similarity:

$$SIM(T,D) = \frac{1}{card(T)} \sum_{j=1}^{card(T)} \underset{C_i \in D}{Max}(P(C_i|x_j))$$

The author of the test document T is then identified as the author of the document D of the reference base, which maximises the similarity index[4].

$$Writer(T) = Writer(\underset{D_i \in R}{Arg\max}(SIM(T, D_i)))$$

4. Experimental Results

For our system, we have chosen the IAM database [8] which contains samples of unconstrained handwritten text. The text has been scanned at a resolution of 300 dpi and digitized to 256 levels of gray. For our experiments the text was binarized by a global thresholding of the document image. The experiements were carried out by fixing the value of Φ and varying the window size n. The system was trained on 50 writers with one document image per writer. The evaluation of the system was carried out by using a different text image of each author and achieved an identification rate of 94%. Figure 4 summarizes the results obtained at different window sizes. It can be noticed that a window size of 13x13 produced the best results in our series of experiments.

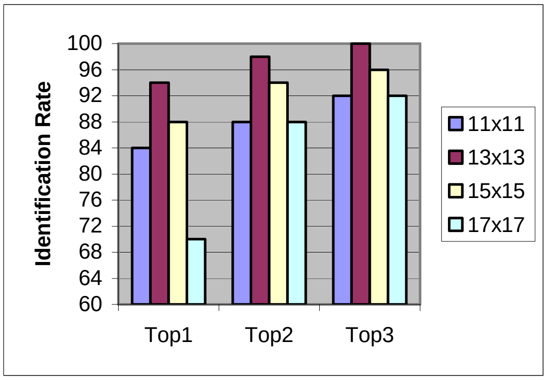

Figure 4. Identification at different window sizes

5. Conclusion

We have presented an effective method for writer identification in handwritten documents. The technique is based on the extraction of forms that a specific writer would use frequently as he draws the characters. The achieved identification rates are very promising and validate the argument of the existance of redundant pattenrs in a handwritten text. The identification can be further improved by using a more adaptive division of text following the natural direction of hand. Varying the window size n, this method can also be applied to other languages like arabic and urdu etc. The system can be made more robust making it capable of automatically adjusting the window size depending upon the writing details. Moreover, the system can be extended beyond identification to perform the verification of the author as well.

6. References

[1] M. Bulacu, L. Schomaker, and L. Vuurpijl, “Writer identification using edge-based directional features”, In *Proc. of 7th International Conference on Document Analysis and Recognition*, volume II, Edinburgh, Scotland, 3-6 August 2003, pp. 937–941.

[2] B. Zhang, S. N. Srihari, S. Lee, “Individuality of handwritten characters”, In *Proc. of 7th International Conference on Document Analysis and Recognition*, volume II, Edinburgh, Scotland, 3-6 August 2003, pp. 1086–1090.

[3] A. Seropian, “Analyse de Document et Identification de Scripteurs”, PhD Dissertation, University of Toulon, France, 2003.

[4] A. Bensefia, A. Nosary, T. Paquet, L. Heutte, “Writer identification by writer's invariants”, In *Proc. of the Eight International Workshop on Frontiers in Handwriting Recognition*, Niagara-on-the-Lake, Canada, August 2002, pp. 274-279.

[5] S. Srihari, S. Cha, H. Arora, and S. Lee, “Individuality of handwriting”, *J. of Forensic Sciences*, 47(4):1.17, July 2002.

[6] U.V. Marti, R. Messerli, H. Bunke, “Writer Identification Using Text Line Based Features”, In *Proc. of the 6th International Conference on Document Analysis and Recognition*, Seattle, USA, 2001, pp. 101-105.

[7] H.E.S. Said, T.N Tan, K.D. Baker, “Personal Identification Based on Handwritting”, *Pattern Recognition*, vol. 33, 2000, pp. 149-160.

[8] U. Marti and H. Bunke, “A full English sentence database for off-line handwriting recognition”, In *Proc. of 5th International Conference on Document Analysis and Recognition*, Bangalore, India, 1999, pp. 705-708.

[9] A. Nosary, L. Heutte, T. Paquet, Y. Lecourtier, "Defining writer's invariants to adapt the recognition task", In *Proc. of 5th International Conference on Document Analysis and Recognition*, Bangalore, India, 1999, pp. 765-768.

[10] V. Bouletreau, N. Vincent, R. Sabourin, H. Emptoz, "Handwriting and signature : one or two personality identifiers?", In *Proc.of Fourteenth International Conference on Pattern Recognition*, Los Alamitos, CA, vol.2, 1998, pp. 1758-1760.

[11] R. Plamondon and G. Lorette, “Automatic signature verification and writer identification – the state of the art”, *Pattern Recognition*, vol. 22, n°2, 1989, pp. 107-131.