Writer Identification By Writer's Invariants

Ameur BENSEFIA, Ali NOSARY, Thierry PAQUET, Laurent HEUTTE
Laboratoire Perception Systèmes Information,
UFR des Sciences, Université de Rouen,
F-76821 Mont-Saint-Aignan Cedex, France.
ameur.bensefia@univ-rouen.fr

Abstract

This communication deals with the problem of writer identification. If the assumption of writing individuality is true then graphical fragments that constitute it should be individual too. Therefore we propose a morphological grapheme based analysis to make writer identification. Template Matching is the core of the approach. The redundancy of the individual patterns in a writing, defined as the writer's invariants, allows to compress the handwritten texts while maintaining good identification performance. Two series of tests are reported. The first series is designed to evaluate the relevance of our approach of identification on a basis of 88 writers by evaluating the influence of the text representation (with or without invariants) on the quality of the method. The method gives about 97,7% of correct identification when using large compressed samples of handwriting. The second series of tests is designed to evaluate the influence of the sample size of the writing to be identified on the quality of the method. It is shown that writer identification can reach a correct identification rate of 92,9% using only samples of 50 graphemes of each writing.

1. Introduction

Writing is a personal act: each writer is characterized by his writing, by the reproduction of details and unconscious practices. This is why in certain cases of expertise, the analysis of writing samples has the same value as the analysis of fingerprints.
The problem of writer identification frequently arises in the court of justice where one must come to a conclusion about the authenticity of a document (e.g. a will). It is also posed in the banks for signature verification [7]. It is also posed in some institutes which analyze texts of former authors, and are interested in the genetics of these texts, the identification of the various writers who took part in the drafting of a manuscript or who made corrections. The significant results of these last years in the field of handwriting recognition make it possible to bring today first significant answers to this particular problem.
Recently, some works in this field were proposed [1, 4, 8, 9, 10], in which the developed methods rely on three traditional steps:

- pre-processing: the image is cleaned by noise reduction, then lines and words are extracted.
- feature extraction: features which are quantitative measurements are used to discriminate the writers as well as possible; they can be global or local and structural or statistical.
- classification: the search of the nearest writer is guided by the extracted features using an adapted metric.

In this communication, we present a system for the off-line identification of cursive handwritings. We briefly recall in the first part works proposed recently to carry out writer identification. We then present the approach we developed to solve this problem. To avoid the delicate design of suitable features, we chose to use the elementary patterns the writing is made up of (graphemes). The assumption of the individuality of each writing justifies this choice a priori. Indeed, if each writing is individual, the elements which make it up are also individual. Therefore they should bring the suitable information to make discrimination possible. The results we present in the third part of this communication seem to validate this assumption. For this purpose, two series of tests have been carried out. The first one evaluates the performance of the proposed method when the available information is a set of some lines of writing (a short text). The second series of tests evaluates the method according to the size of the sample considered: some characters or some words. This point has been rarely considered in the literature yet it seems to be fundamental in real applications.

2. State of the art

The problem of writer identification can be tackled according to two main approaches [8]: the verification and the identification of the writer. In the writer

Proceedings of the Eighth International Workshop on Frontiers in Handwriting Recognition (IWFHR'02)
0-7695-1692-0/02 $17.00 © 2002 IEEE

verification approach, one must come to a conclusion about two documents read in input and determine whether the two documents are written by the same writer or by two different writers. The problem of writer identification consists in identifying a writer among a set of N candidates. The system must therefore be learnt from a set of handwriting samples of each individual candidate.
Whereas the verification approach can be "simply" formulated as a two-class discrimination problem, the identification approach requires the use of a nearest neighbor based decision, due to the potentially large number of candidates.
Until now, the features used in these two approaches have been global features which are statistical measurements extracted from the whole block of text to identify. They can be:

- *features from texture*: the document image is seen in this case as a simple image and not as a writing. For example, the application of Gabor filters and co-occurence matrices was considered in [10].
- *structural features*: in this case the extracted features attempt to describe the writing properties. We can quote for example, features such as the average height, the average width, the average slope and the average legibility of the characters [4].

These statistical features extracted from a block of text make it possible to reach interesting results, which are however always delicate to compare due to a lack of common references.
Finally, the various works suggested can be categorized on the one hand according to the number of writers to be discriminated, and on the other hand, according to the size of the sample available to carry out the identification of the writer (several lines of text or some words). For example, the approach proposed in [9] is able to identify 95% of the 40 writers known by the system, using a sample of several lines of each writing. The work presented in [10] reaches 92,5% of correct identification among 50 writers using 45 samples of the same word the participants were asked to write. It is worth noticing that the work proposed in [8] has dealt with a problem of writer identification / verification using the largest database (1000 writers) made of 3 samples of a same text written by each participant.

3. Proposed approach

Thanks to the experience of the last decades in the field of handwriting recognition, the problem of writer identification does not seem to give rise to the same difficulties. Indeed, it is now well established that one of the major difficulties for the automatic recognition is to manage the large variability of writing styles. This difficulty has been tackled by using suitable features or by defining multiple classes of letter allographs, or lastly by introducing the lowest level of description: the grapheme level. That is to say that the task of modeling the writing is difficult from the recognition point of view, and is still calling research efforts in spite of the results obtained these last years.

3.1. Writer's properties

The fundamental property of handwriting, which makes written communication possible, is that there exist *inter-writer invariants* since the morphological differences between patterns representing distinct letters are more important than those between different allographes of a same letter (inter-writer variability). We postulate now that each writer draws the same letters using the same patterns (his own handwriting references): we call these references the "*writer's invariants*". The writer's invariants, reflecting the morphological redundancy of his handwriting, can be defined as the set of similar patterns or graphemes extracted from the segmentation of his handwriting using a particular segmentation technique. We can expect to obtain the writer's particularities only when large samples of his handwriting can be collected.
The detection of the morphological invariants is performed using an automatic classification of graphemes issued from the segmentation of the handwritten text. The method has been presented previously in [5]. We just recall the main points of the algorithm.
Among the different clustering methods, a simple and fast sequential clustering algorithm that does not need to know a priori the number of clusters has been retained [3]. One characteristic of this sequential clustering method is that the choice of the templates depends on the order of presentation of the elements in the grapheme set. To cope with this problem, multiple sequential clusterings are iterated with a random selection order of the elements. Finally, only the elements that are always grouped together, when iterating the clustering procedure, are retained to constitute a cluster, while elements that are not always assigned to the same group constitute singular elements e.g. clusters with a single element.
This algorithm requires, as any clustering method, an index of proximity and an adequate choice of regrouping threshold. The number of generated invariant clusters is a function of the selected threshold. Among the several types of proximity measures tested, only the correlation similarity measure has been retained. Figure 1 gives an example of the clusters obtained after four iterations of a sequential clustering.

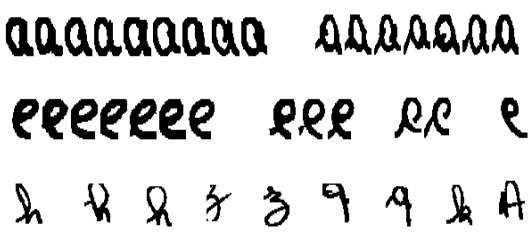


Figure 1. Samples of invariant clusters extracted from a handwritten page.

As one can see in figure 1, the proposed method succeeds in determining regular patterns that occur in the text. Of course, this primary result does not allow to evaluate the robustness of the method over a large set of handwritten texts.

The first results presented in [5] have shown however that handwriting variability can be measured through the writer's invariants. This experiment made it possible to show the existence of a certain level of stability in each handwriting and has led us to develop a system for handwritten text recognition based on writer adaptation [6]. These experiments demonstrate that inter-writer variability is more significant than intra-writer variability. Therefore, the writer identification problem should find a natural solution by taking benefit of the individual patterns of each handwriting.

For this reason we propose a system for writer identification without resorting to the traditional features used in pattern recognition but rather by exploiting the rough information brought by each individual pattern of each handwriting. Each handwriting sample is thus represented by the set of graphemes produced by our segmentation module which is part of our recognition system [6].

3.2. Identification

Each handwritten document D is represented by the whole set of graphemes x_i it is made up, that is to say:

$$D=\{x_i, i \le card(D)\}$$

We define a similarity measure between the handwritten document D and an unspecified handwritten document T by the following relation:

$$SIM(D,T)=\frac{1}{card(D)}\sum_{i=1}^{card(D)} Max_{y_j \in T}(sim(x_i,y_j))$$

where: x_i, y_j are graphemes of documents D and T respectively, and $sim(x_i,y_j)$ is a similarity measure between two unspecified graphemes. Several similarity measures have been already defined in the literature [2], we quite simply retained the correlation measure, defined by the following relation:

$$sim(x,y)=\frac{n_{11}n_{00}-n_{10}n_{01}}{[(n_{11}+n_{10})(n_{01}+n_{00})(n_{11}+n_{01})(n_{10}+n_{00})]^{1/2}}$$

Where n_{ij} is the number of pixels for which the two normalized images x, y have the corresponding value: x(k)=i, y(k)=j. Notice that according to this measure, two handwritten documents will be all the more close as the similarity measure will be close to 1. In the extreme case where it would be equal to 1, this would indicate that all the graphemes of the unknown document D have an exact correspondent in the considered document T. By construction this measure is not symmetrical because we want to take into account the size of the unknown document, which can be a document comprising several lines of writing or on the contrary only some graphemes. Each reference document has, as for it, a standard size since it corresponds to the writing of the same known text.

The writer of the unknown document D will be finally classified as the writer of the document of the reference set that is the most similar to the unknown document (according to the meaning of the measure we have just defined) that is to say:

$$Writer(D) = Writer(Arg\max_{T_i \in Reference Set}(SIM(D, T_i)))$$

3.3. Identification with writer's invariants

The documents D and T are represented by their respective set of graphemes, the number of which varies according to whether the document D is a whole text or just a word. Using all the graphemes of the two documents D and T is computationally expensive especially when the reference set of writers becomes large. In order to accelerate the procedure of writer identification, we propose to represent the handwritten texts by their invariant graphemes. We thus hope to operate some compression of the handwritten information without degrading to a significant degree the proposed method of identification.

When the sample to be identified is only made up of some words, the use of writer invariants before identification is not necessarily justified in terms of computing time neither in terms of compression rate.

In the first series of tests carried out, we have evaluated each of the four possible combinations of the textual representations of the 2 documents (with and without

compression by invariants). The results show the interest of the representation by invariants since we reach the same level of performance in terms of identification rate with however a notable saving in computation time thanks to the compression of the representation by an average factor of 4.

4. Experiments

We carried out two series of different tests in this first study. The first series is designed to evaluate the relevance of our approach of identification on a basis of 88 writers by evaluating the influence of the text representation (with or without invariants) on the quality of the method. The second series of tests is designed to evaluate the performance of the method as a function of the sample size of writing to be identified.

4.1. Database

For these two series of tests we have built a database of 88 writers who have been asked to copy out one letter chosen among two suggested, each one being made up of 107 and 98 words respectively. Some samples are presented in figure 2. The majority of these handwritings are of cursive nature.
The texts obtained were cut in two nonequal parts: two thirds, one third. The first two thirds were used as the reference set of writers, and the remaining third was used for testing.
In the first series of tests the last third of each text is used to identify the writer. In the second series of tests, we extract from each last third of each text 5 examples of grapheme sequences of 5 various lengths: for each writer, 5 examples for each of the 5 lengths selected (10, 20, 30, 40 and 50 graphemes). We thus have for this second series of tests 2200 sequences we try to identify the writer with.

Figure 2. Some samples of the data base.

4.2. Writer identification from texts

We are first of all interested in the relevance of the basic method we propose which is based on template matching using the correlation measure. The results obtained on the test database are represented in figure 2 and are referred as *Method 1*. The correct writer identification rate is 97,7% in first proposal; moreover, if one retains the two texts which are the most similar to the unknown example, the correct writer is present in the two first solutions in 100% of the cases.
This first result is quite interesting because beyond the quality of the correct writer identification rate, it accredits the strong assumptions which led us to develop this approach: handwriting is individual and this individuality can be detected thanks to the graphemes the writing is made up of. As a consequence, resorting to another representation space does not seem necessary for the identification task. The question remains open for the writer verification task.
In the following tests we are interested in the influence of the text representations on the identification performance. More precisely we want to know if the writing identification remains possible if one represents the texts by their respective groups of invariants. Indeed, since any clustering method is accompanied by an error, we want to know the incidence of this error on the identification method qualities.

In *Method 2* (figure 3), only the unknown text is represented by its invariant clusters, the texts of the reference set are, as for them, represented by the whole set of graphemes they are made up.
In *Method 3* (figure 3), we use the opposite representation of that used in *Method 2*. The texts of the reference test are represented in a compressed manner by their invariants, while the unknown text to classify remains represented by its whole set of graphemes.
Lastly, in *Method 4*, we systematically use a representation compressed by the invariants for the texts of the reference set as for the texts to be tested.
The most interesting results of this series of tests can be emphasized by examining the results of method 3. In this particular case, the results obtained are close to those obtained with the direct method (method 1) but using the compressed representation by invariants. It follows from there a saving of computation time of a factor 16 on average, compared to the direct method, without notable loss on the quality of the writer identification. The two other methods (methods 2 and 4) although less powerful are nevertheless interesting but they seem to point out the limit of the method when we choose to work on a compressed representation of the unknown text.

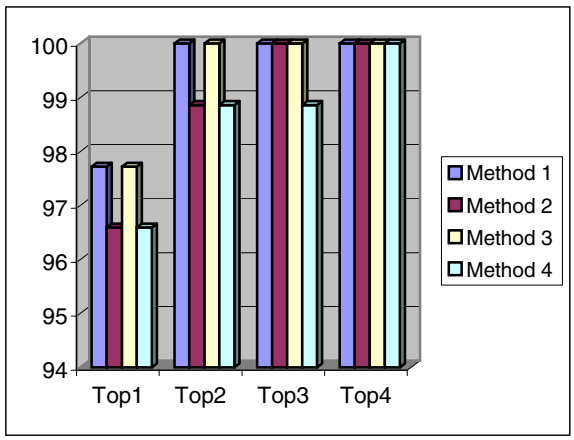
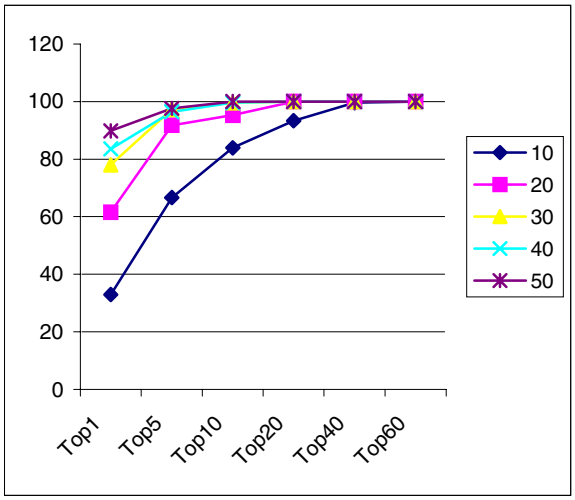


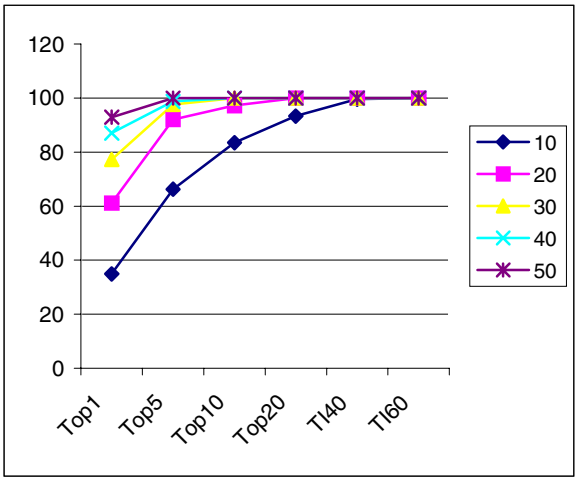
Figure 3. Identification results on texts using various compressed representations.

4.3. Writer identification from sequences of graphemes

Given the excellent results of the identification method we have just proposed, we are interested now in the evaluation of the sufficient quantity of information required to identify the writer correctly. The answer to this question has two major interests. On the one hand, when one wants to identify a writer one does not have necessarily a large sample of his writing; we will thus bring on this point a precise answer. On the other hand, from a strictly computation point of view, one could be tempted to use the only necessary and sufficient information to identify a writing rather than carry out the complete analysis of the page when the latter is available. To answer this question we have thus evaluated our identification method on small fragments of texts made up of the grapheme sequences built as indicated in section 4.1. For this evaluation we proceed using *method 1* and *method 3*.
Results of this experiment are reported in figure 4. Figure 4.a gives identification results using compression of the reference set (*method 3*), while figure 4.b gives identification using direct pattern matching of *method 1* without compression. First of all, we can notice that the uncompressed representation of the reference writings gives better identification performance, essentially for the two larger samples of 40 and 50 graphemes.
The second general interesting conclusion is that samples of 40 to 50 graphemes of the writing allow significant identification rates of nearly 90% on average. A more detailed analysis shows that using uncompressed reference sets allows to identify the correct writer in 92,9% of the cases using a sample of 50 graphemes, while using this same sample size allows the correct writer to be present in a list of 5 candidates in 100% of the cases. A sample size of 30 graphemes can also give significant results to provide candidate lists.



(a)



(b)

Figure 4. Identification results on grapheme sequences with (a) and without compression (b).

5. Conclusion

In this paper we have proposed a new approach for writer identification. Identification is based on the pattern matching of individual components of the handwritings, so called graphemes in the literature. The identification performance obtained on a sample of 88 writers are very promising and show that individual written patterns are representative of the writer's personal style. As a consequence, the use of common features sets as those used for the recognition of handwriting does not seem to be necessary at all for this particular task of writer identification.

Furthermore, the experiments carried out in this paper show that this method is able to give interesting identification performance even using small samples of handwriting. The correct writer can be selected within a list of 5 candidates in almost all the cases, using only a sample of 50 graphemes i.e. few words. These results however are insufficient with respect to forensic applications. In such case, it seems that the proposed approach should bring a good mean to select relevant samples from large databases. These samples should be further analysed for writer verification. This point is at present under developpement and we expect the grapheme representation to be as relevant as for the writer idenfication task.

References

[1] S.H. Cha, S. Srihari, "Multiple Feature Integration for Writer Verification", 7th International Workshop on Frontiers in Handwriting Recognition; IWFHR VII, Amsterdam, The Netherlands, pp 333-342, 2000.

[2] R. Duda, D. Stork, P. Hart, Pattern Classification and Scene Analysis, Wiley & Sons, 2nd Edition, 2000.

[3] P. Gader, B. Forster, M. Ganzberger, A. Gillies, M. Wahlen and T. Yocum, "Recognition of handwritten digits using template and model matching", Pattern Recognition, vol. 24, n°5, pp. 421-431, 1991.

[4] U.V. Marti, R. Messerli, H. Bunke, "Writer Identification Using Text Line Based Features", Proc. ICDAR'01, Seattle (USA), pp 101-105, 2001.

[5] A. Nosary , L. Heutte, T. Paquet, Y. Lecourtier, "Defining writer's invariants to adapt the recognition task", IAPR-ICDAR'99, Bengalore, India, pp 765-768, 1999.

[6] A. Nosary, ''Automatic recognition of handwritten texts through writer adaptation'', PhD Dissertation (in french), University of Rouen, France, 2002.

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

[8] A. Srihari, S. Cha, H. Arora, S. Lee, "Individuality of Handwritig : A Validity Study", Proc. ICDAR'01, Seattle (USA), pp 106-109, 2001.

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

[10] E.N. Zois, V. Anastassopoulos, "Morphological Waveform Coding for Writer Identification", Pattern Recognition, vol. 33, n°3, pp 385-398, 2000.