Electronic Letters on Computer Vision and Image Analysis 5(2):72-86, 2005

Handwritten Document Analysis for Automatic Writer Recognition

Ameur BENSEFIA, Thierry PAQUET, Laurent HEUTTE

Laboratoire PSI – FRE CNRS 2645,
UFR des Sciences, Université de Rouen,
76821 Mont-Saint-Aignan Cedex, France

{Ameur.Bensefia,Thierry.Paquet,Laurent.Heutte}@univ-rouen.fr

Received 16 July 2004; accepted 23 May 2005

Abstract

In this paper, we show that both the writer identification and the writer verification tasks can be carried out using local features such as graphemes extracted from the segmentation of cursive handwriting. We thus enlarge the scope of the possible use of these two tasks which have been, up to now, mainly evaluated on script handwritings. A textual based Information Retrieval model is used for the writer identification stage. This allows the use of a particular feature space based on feature frequencies. Image queries are handwritten documents projected in this feature space. The approach achieves 95% correct identification on the *PSI_DataBase* and 86% on the *IAM_DataBase*. Then writer hypothesis retrieved are analysed during a verification phase. We call upon a mutual information criterion to verify that two documents may have been produced by the same writer or not. Hypothesis testing is used for this purpose. The proposed method is first scaled on the *PSI_DataBase* then evaluated on the *IAM_DataBase*. On both databases, similar performance of nearly 96% correct verification is reported, thus making the approach general and very promising for large scale applications in the domain of handwritten document querying and writer verification.

Keywords : Handwritten documents, writer identification, writer verification, information retrieval, mutual information, hypothesis testing, graphemes.

1 Introduction

This article is focused on handwritten documents analysis for experts. The method that we propose allows the retrieval of digitized handwritten document images according to their graphical contents i.e. according to the handwritings and authors on whom the user has expressed an information need by means of a specific query. For these documents, one can broadly exhibit two kinds of use which correspond to two different kinds of requests:

- Handwritten documents can be analysed for their textual content. In this case querying a handwritten document database would require to resort to a transcription phase of the handwritten texts prior to the indexing of their textual content using standard techniques dedicated to information retrieval. Unfortunately, the state of the art in the field of handwriting recognition does not allow to apply such

Correspondence to: <Thierry.Paquet@univ-rouen.fr>

Recommended for acceptance by <J.M. Ogier, G. Sanchez>
ELCVIA ISSN: 1577-5097
Published by Computer Vision Center / Universitat Autonoma de Barcelona, Barcelona, Spain

an approach. The recognition of the handwriting remains indeed badly controlled on omni-writer applications when calling upon large size lexicons [7].

- Handwritten documents can also be considered for their graphical contents. In this case querying handwritten document databases can be carried out using graphical requests. One seeks for example to retrieve the documents of the database that contain certain calligraphy corresponding to specific writers. Other possible applications can concern the detection of the various writings present on a document, or the dating of the documents compared to the chronology of the work of the author.

- Automatic handwriting analysis techniques allow to consider from now some applications as the one we propose in this paper.

One can consider that these two applications fall into the problem of information retrieval either textual or graphical. These two tasks have been largely studied either in the electronic document retrieval field or in the image processing field. In the field of automatic handwriting analysis, the task falls into the writer identification paradigm.

Each writer can be characterized by his own handwriting, by the reproduction of details and unconscious practices. This is why in certain cases of expertise, handwriting samples have the same value as that of fingerprints. The problem of writer identification arises frequently in the court of justice where one must come to a conclusion about the authenticity of a document (e.g. a will). It also arises in banks for signature verification [6], or in some institutes which analyse texts of former authors, and are interested in the genetics of these texts, as for example the identification of the various writers who took part in the drafting of a manuscript or who made corrections...

As for any biometric-based identification applications (fingerprints, faces, voices, signatures…), forensic analysis of handwriting requires to query large databases of handwritten samples of known writers due to the large number of individuals to be considered. Therefore in order to come to a conclusion about the identity of an unknown writer, two tasks must be considered:

- the writer identification task concerns the retrieval of handwritten samples from a database using the handwritten sample under study as a graphical query. It provides a subset of relevant candidate documents, on which complementary analysis will be developped by the expert.

- the writer verification task, on its own, must come to a conclusion about two samples of handwriting and determines whether they are written by the same writer or not.

When dealing with large databases, the writer identification task can be viewed as a filtering step prior to the verification task. In this case, the verification task consists in matching the unknown writer with each of those in the selected subset. Therefore the verification task can sometimes be adapted to each known reference writer based on the individual description of their handwriting. On the contrary, when the number of potential writers is too large even unknown or infinite, an individual description of each handwriting cannot be used. In this case one can for instance derive a specific set of feature differences to model the overall within and between writer distance (intra and inter writer variability) on a set of examples [12].

In the first part of this paper we investigate the use of one of the most popular schemes used in IR [8] and apply it to the task of writer identification. For this task, a particular set of features is derived using a segmentation procedure of handwritten components followed by a cluster analysis.

The second part of this paper is devoted to the writer verification task or writer authentication. Indeed if the approach suggested for the writer identification task is relevant to find handwritings known by the system, it is on the other hand unable to detect query made up of unknown handwritings and in this case to reject any writer proposal. The proposed approach exploits the same segmented entities, the graphemes, to build an hypothesis test based on a mutual information criterion between the feature set and the set of the two documents under study.

The two approaches have been evaluated on two different handwritten document databases : the first database has been built at PSI and is made up of 88 writers; the second database is the web available subset of the IAM database [13], which contains around 150 different writers. The PSI_database is written in French while the IAM_database is written in English. In both cases the approach demonstrates the excellent capacities to retrieve and authenticate handwritings. As a consequence the proposed methodology shows that handwriting can be considered, in certain condition, as a possible biometric identifiant.

2 Writer Identification

In the identification approach, the user information need concerns the author identification of a single document, which will be considered to be the query. The possible set of writer candidates is supposed to be finite and made up of the set of *N* writers (N documents) stored in a database. Although the writer identification falls under the same general problems as handwriting recognition, it does not seem however to pose the same kind of difficulties. Indeed, the identification task can exploit the handwriting variability in order to discriminate the handwritings, whereas the recognition task, on the contrary, must eliminate the variability between writers in order to be able to identify the textual message whatever the writer.

2.1 Previous works

Features used for the writer identification task are mainly global features which are based on statistical measurements, extracted from the whole block of text to be identified. These features can be broadly classified in two families:

- *features extracting from texture*: the document image is simply seen in this case as an image and not as a handwriting. For example, application of Gabor filters and co-occurence matrices were considered in [7].

- *structural features*: in this case the extracted features attempt to describe some structural properties of the handwriting. One can quote for example the average height, the average width, the average slope and the average legibility of characters [3].

Note that it is also possible to combine the two families of features [12]. The nature of these statistical features, extracted from a block of text, has allowed to reach interesting performance, which are however always difficult to compare due to a lack of common references.

One can also categorize the previous works according to the number of writers and the nature of training samples used by the system (see table 1). On the one hand the system is required to deal with as much writers as possible while on the other hand, training samples of each handwriting may represent several lines of text or on the contrary a few words. The work suggested in [7], for example, makes it possible to identify 95% of the 40 writers the system can deal with by the analysis of some text lines of handwriting. The work presented in [14] reports a correct writer identification performance of 92,48% among 50 writers by using 45 samples of the same word that the participants were asked to write. It should be noted that the work presented in [12] dealt with the largest database (1000 writers) using the same text written 3 times by each writer.

	# Writers	**Sample size**	**Lexicon dependency**	**Performance (%)**
Said 2000	40	Few lines of handwritten text	Yes	95
Zoïs 2000	50	45 samples of the same word	Yes	92
Marti 2001	20	5 samples of the same text	Yes	90
Srihari 2001	100 900	CEDAR letter / paragraph / word CEDAR letter / paragraph / word	Yes Yes	82 / 49 / 28 59 / 25 / 9
Schomaker 2004	150	One copied text paragraph in uppercase handwriting	No	95
Bensefia 2004	88 150	paragraph / 3-4 words paragraph / 3-4 words	No No	93 / 90 86 / 68

Table 1 : Comparison of performance and test conditions for writer identification in most recent studies.

2.2 Organization of the system

We present in the following subsections the various steps of our writer identification system. Figure 1 gives a brief overview of the data processing sequence. It uses three traditional steps: a pre-processing step where the main objective is to locate information that will be used to perform writer identification, then a feature extraction step is used to obtain a relevant representation for the decision process which represents the final step of the system.

2.2.1 Handwritten document pre-processing

First, the connected components of the document image are extracted and analysed in order to eliminate some charts like erasures, overloaded or underlined zones which one knows, *a priori*, that they don’t characterize the handwriting. Then, the remaining connected components are segmented into graphemes. This denomination does not refer to any specific handwriting description and may be confusing. The graphemes are actually elementary patterns of the handwriting that are produced by a segmentation algorithm based on the analysis of the minima on the upper contour [5]. The concatenation of two (respectively three) adjacent graphemes (grapheme i and grapheme i+1) provides what we call bigrams (respectively trigrams) of graphemes.

We now give full details of each processing step of the writer identification system.

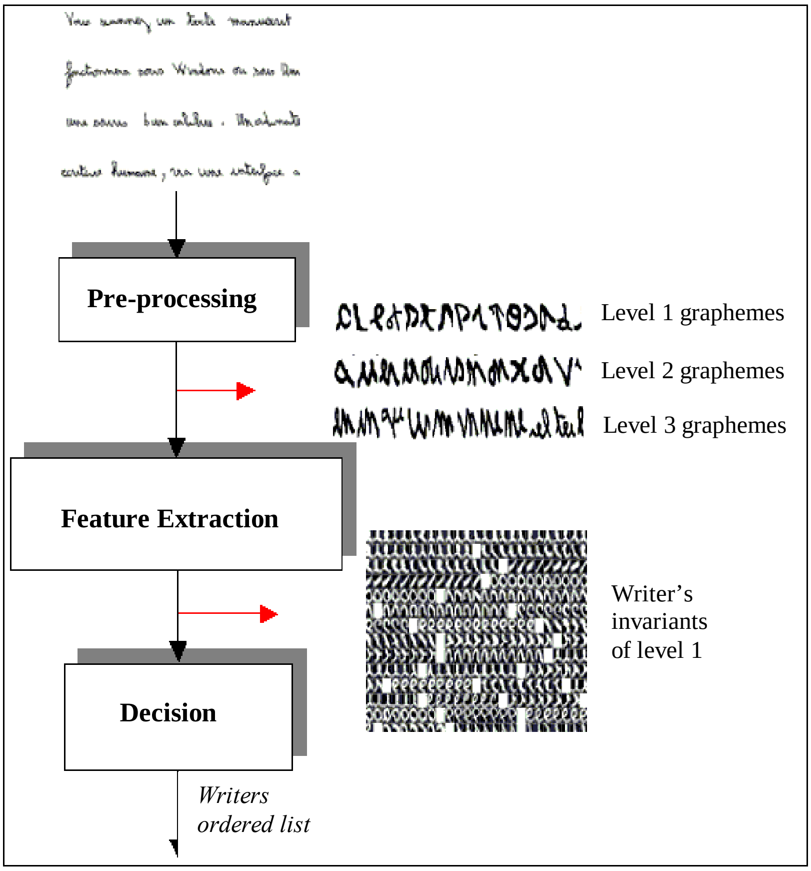


Figure 1. Writer Identification Organisation System

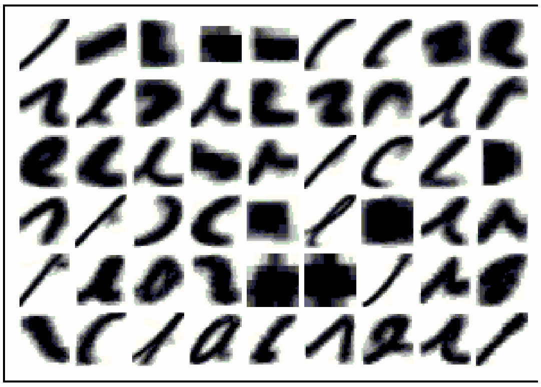


Figure 2. Some invariants of level 1

2.2.2 Feature extraction

The writer identification task lies in the definition of a feature space common to all the handwritten documents. Previous study has shown that graphemes can characterize each handwriting [5]. In this study we have extended this principle to the whole document database. Following the segmentation of the handwritten document, a set of binary features is defined thanks to a clustering procedure. In this manner the feature set is adapted to the handwritings of the database under study, and not *a priori* defined. We briefly recall the main characteristics of the clustering procedure. Several sequential clustering phases are iterated with random selection of the elements. Each of them provides a variable number of clusters. The invariant clusters are defined as the groups of patterns that are always clustered together during each sequential clustering phase. The main step of our clustering procedure are the following ones:

```
Begin
   Define a proximity threshold T
   Repeat sequential clustering 3 times
        Step 1  (initialisation)
             Select randomly a first candidate in the database and define it as the first
             cluster centroïd
        Step 2
             Select randomly a next candidate in the database
             Compute its nearest cluster centroïd
             If the distance to the nearest centroïd is under threshold T
                   Then  Add the candidate to its nearest cluster
             Else
                   create a new cluster with the current candidate as cluster centroïd
        Step 3 (end test)
             go to step2 until the entire database has been visited
   EndRepeat
   Compute intersection of clusters obtained by the three sequential procedures
   Put candidates out of the intersection clusters into additionnal clusters
End
```

The number of clusters depends only on the chosen threshold *T* (which is the same in our experiments for all the handwritings), on the variability of the handwriting under study and on the initial number of graphemes. For example for 600 graphemes of level 1 we can obtain from 150 to 400 clusters. Figure 2 gives some of the most frequent clusters obtained on the *PSI_DataBase*. As one can see on this figure, some of these features look like black boxes: these ones are noise and should be eliminated for the identification process (see section 2.2.3). On the other hand, some features look like letters or parts of letter: these ones should be kept in general. Finally, all these clusters constitute a set of binary features which will be the basis of our writer identification method.

2.2.3 Information Retrieval Model

In this study we formulate the writer identification task within the framework of Information Retrieval. Information Retrieval is the process of finding relevant documents for a user need (expressed by a query) in a large database [11]. For this purpose the query and the documents of the database are generally described in the same feature space. The choice of the feature space is therefore of primary importance. As any kind of request may be used to retrieve the documents, one generally seeks to describe the documents by preserving the whole set of extracted features which leads to a description of the documents in a high dimensional feature space.

We can formulate our problem of writer identification as a process of finding graphical contents (set of graphemes extracted from the document to identify) in a large database of documents (set of reference documents). The retrieved documents will be classified according to their similarity with the query, from the nearest to the farest document. There are several types of Information Retrieval models [11]: the Boolean model, the probabilistic model and the vector space model (VSM) are the most popular models. Among them the Vector Space Model proposed by Salton [8] although very simple and of rather old design still remains very effective [2].

This model involves two different phases: a preliminary indexing phase is intended to describe each document through a high dimensional feature vector; the retrieval phase then makes it possible to evaluate the relevance of each document D_j of the database with respect to a specific query Q. According to the vector space model, relevance of each document is evaluated by the scalar product between the vector describing the query Q and the one describing a document D_j of the database. We now present each of the two phases of the model.

Indexing phase

Assume a binary feature set has been chosen. Denote φ_i , $0 \leq i \leq m\text{-}1$ the ith binary feature. For IR purposes each feature is all the more relevant to describe a document as it is relatively frequent in this document compared to any other document in the database. Using this principle, each document D_j as well as the query Q, can be described as follows:

$$\vec{D}_j = (a_{o,j}, a_{1,j}, \ldots\ a_{m-1,j})^T \quad \text{and} \quad \vec{Q} = (b_o,\ b_1,\ \ldots\ b_{m-1})^T$$

where : $a_{i,j}$ and b_i are weights assigned to each feature φ_i, and are defined by:

$$a_{i,j} = FF(\varphi_i, D_j)\ IDF(\varphi_i) \quad \text{and} \quad b_{i,} = FF(\varphi_i, Q)\ IDF(\varphi_i)$$

$FF(\varphi_i, D_j)$ is the *Feature Frequency* in document D_j. $IDF(\varphi_i)$ is the *Inverse Document Frequency* that is the inverse of the number of documents that contain this feature φ_i and is exactly defined by :

$$IDF(\varphi_i\ ) = log\ (\frac{1+n}{1+DF(\varphi_i\ )})$$

where n denotes the total number of documents in the database and $DF(\varphi_i)$ is the *Document Frequency,* i.e. the number of documents that contain this feature.

Note that $IDF(\varphi_i) = 0$ when φ_i occurs in each document. Such features will therefore be given a null score and should indeed be eliminated from the feature set.

Retrieval phase

Each document as well as the query being described in the same high dimensional feature space, a similarity measure between a document and the query is required to provide an ordered list of pertinent documents. Many similarity measures have been proposed in the literature. Most of them are defined on binary feature vectors such as Dice, Jaccard, Okapi measures. When dealing with real valued feature vectors, a similarity measure can be defined by the normalized inner product of the two vectors e.g. by the cosine of the angle between the two vectors. Therefore the similarity measure between a document D_j and the query Q is defined by:

$$\cos(Q, D_j) = \frac{\sum \mathrm{a}_{i,j}\ \mathrm{b}_i}{\sqrt{\sum_{\varphi_i} a_{i,j}^2 \sum_{\varphi_i} b_i^2}}$$

where the two terms in the denominator are the lengths of the document and the query respectively. The retrieval process has thus a complexity of *O(mN)*, where *m* is the size of the feature vector and *N* the number of documents in the database.

2.3 Application

In this section we discuss the implementation of the Vector Space Model of IR for the writer identification task. The central point lies in the definition of a common feature space over the entire database. Then indexing and retrieval phase can be implemented following the definitions given in section 2.2.3. Thus the central point in the evaluation of the IR model concerns the feature choice. Here we have chosen the invariant clusters obtained on the whole document database as described in section 2.2.1.

For the evaluation, two different databases have been used. The first one has been constituted at PSI and contains 88 writers who have been asked to copy a letter (in French) that contains 107 words. The scanned images have been divided into two parts: two thirds for the learning base and one third of each page for the test base. The second database that we have used is part of the IAM database [13]. The fraction of this database that we have used contains texts written in English by 150 writers. Textual content varies from one writer to another. As graphemes can be grouped together to produce either bi or tri-gram (a larger window could eventually be used), the writer identification has been carried out on these three levels. Indeed it is unclear whether concatenations of these features can better characterize a writing or not.

2.4 Performance

Figure 3 gives the results of the approach on the PSI and the IAM DataBase. This table exhibits similar good results on the two databases. Results on the PSI database shows higher performance (+10%) in the top 1 writer candidates that can be explained by the lower number of candidate writers in the PSI database compared to the IAM database. Another difference between the two databases is the lower performance of the bi-gram level on the IAM database that can be explained by the fact that this database contains smaller text samples than the PSI database. In both cases tri-grams show the same significant decrease in identification performance: this can be explained by the fact that tri-gram features may be more dependent on the textual content. Therefore, while some of the tri-grams may constitute pertinent features for the writer, their frequency may be so low (due to the low frequency of textual passage) that the size of our database does not allow to measure it. As a preliminary conclusion theses results show that the vector space model of IR is pertinent for the task of writer identification when using local features.

A second experiment was conducted on the same databases in order to evaluate the influence of the size of the query on the identification performance. Results are given in Figure 4. Correct identification is achieved using 50 graphemes in nearly 90% of the cases on the PSI database (compared to 93 % using text blocks) while these performance decrease to nearly 68% for the IAM database in the same conditions (compared to 88% using text blocks). We can conclude that short queries do not degrade significantly the identification performance. In addition we can also highlight that in both databases tri-grams have a significantly lower discriminative power than graphemes or bi-grams as was already observed using large queries (see Figure 3).

2.5 Discussion

These first results show that the vector space model of IR is pertinent for the task of writer identification when using local features. Furthermore bi-gram features may be even better features for the task. The writer identification is based on the principle of similarity between the query (document to be identified) and all the documents of the reference database. The output of this process is an ordered list of all the documents of the database. This principle raises however the problem where the writer to be identified is unknown in the reference database. In this case, a possible solution lies in the authentication (verification) of the writer proposed by the approach of identification. Indeed with an approach of verification, we can accept or reject the writer proposed with a given error rate. We now propose to investigate the use of graphemes in the writer verification task.

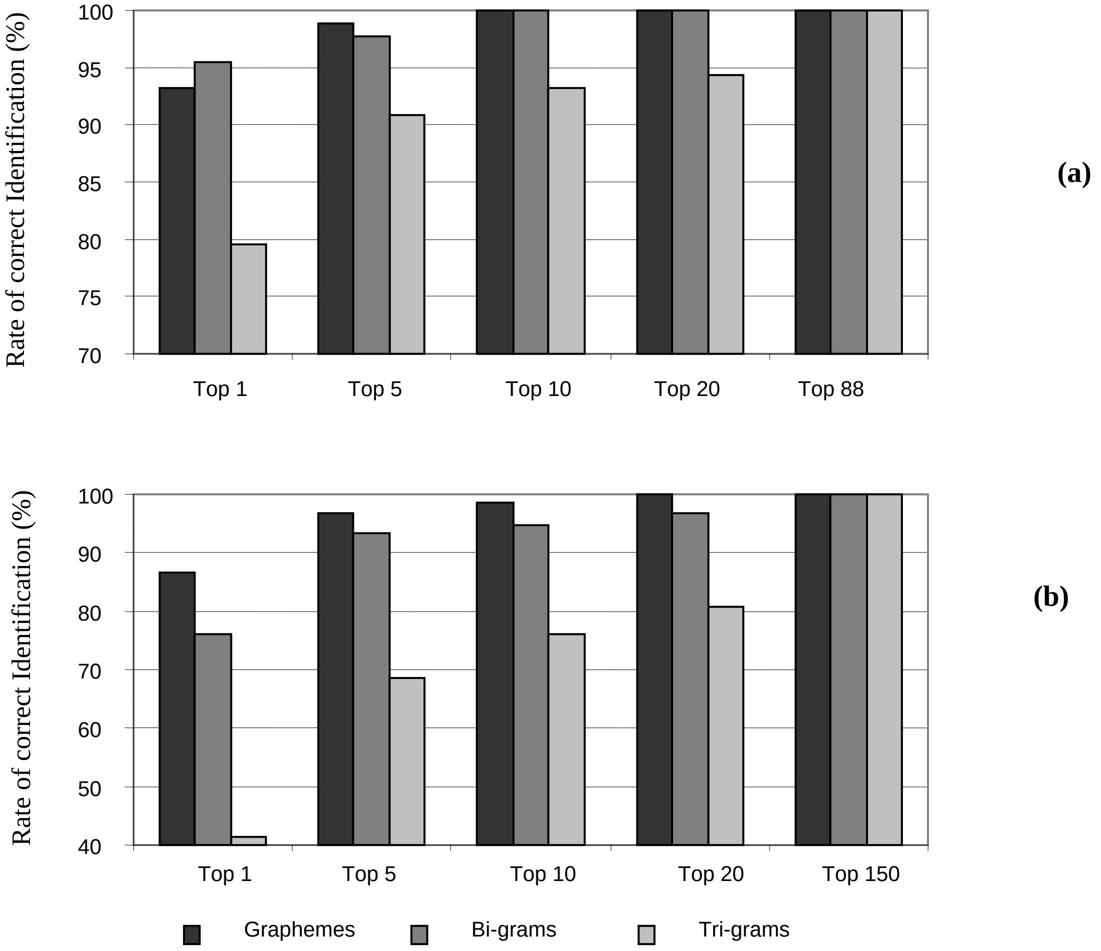


Figure 3. Writer identification performance on the PSI database (a) and on the IAM database (b), where Top N stands for the presence of the correct writer among the N first solutions.

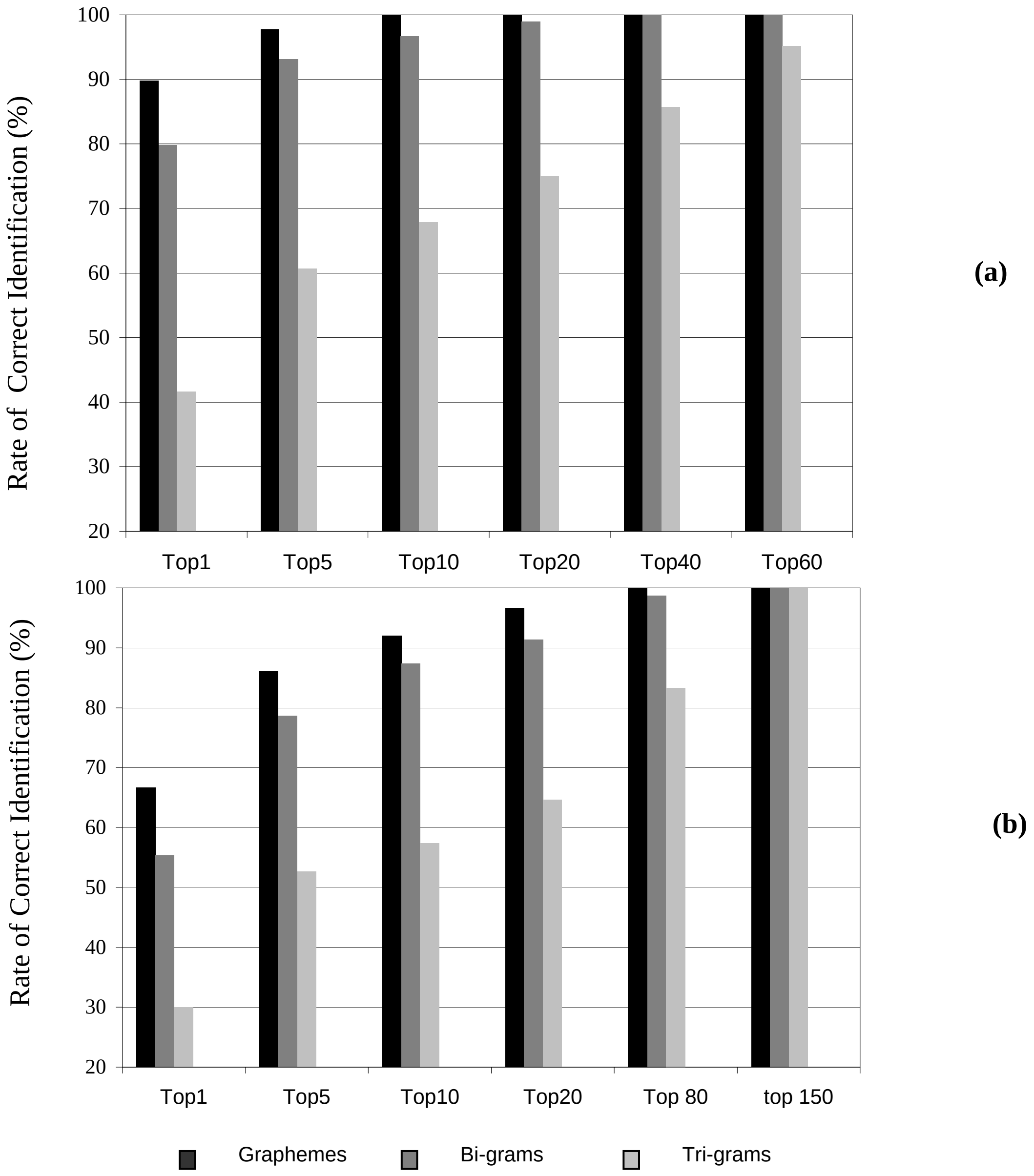


Figure 4. Writer identification performance using queries of 50 graphemes on the PSI database (a) and the IAM database (b), where Top N stands for the presence of the correct writer among the N first solutions.

3 Writer Verification

The writer verification task is the task of authenticating the writer of a document. Most of the time this task is carried out by an expert and is prone to an important subjectivity [4]. In any case, the confidence which can be associated to a decision of this type is not scientifically proven. Recent works have however proposed a scientific methodology of handwriting analysis for the task of writer verification [1]. It should be noted that this task of verification has been less studied than the task of identification. This is undoubtedly due to the fact that verification implies a local process of decision-making which generally depends on the textual contents. Indeed, one generally has to compare the possible shapes of a character or a specific word that occur on the document under study. This is why the complete automation of this task does not seem to be very realistic because it would depend on the automatic recognition task.

In this section we propose a writer verification approach which is independent of the textual contents. Note that this is possible only when the amount of information available for the analysis is sufficiently important i.e. under the same conditions as for the task of writer identification (a block of text). If this orientation may seem, *a priori*, very restrictive for the expertise, it appears nevertheless to be completely complementary to the writer identification task we have introduced in section 2. Indeed, by construction, the writer identification approach does not make it possible to detect an unknown handwriting in the database. The proposed approach of verification allows to validate or to reject the handwritten documents output by the identification phase. The approach can benefit from the set of local features already exploited at the time of the identification phase (graphemes) in order to assess that two handwritten documents can come from the same writer or not. For this purpose, we build an hypothesis test [9], based on a mutual information criterion between the two handwritten documents.

3.1 Construction of an hypothesis test

3.1.1 Mutual information criterion

Assume that two handwritten documents D_1 and D_2 have been written by writers S_1 and S_2 respectively. Let us denote S the set of these two writers :

$$S = \{S_1, S_2\}$$

As for the identification task, we can assume that the two handwritten documents have been segmented into graphemes during the preprocessing step. Then an unsupervised classification step (as discussed in section 2.2.2) allows to define a set of features G common to the two analyzed documents:

$$G = \{g_1, g_2, g_3 \ldots\ldots g_N\}$$

Some of these features can be present on the two documents, while others can appear specifically on only one document. Mutual information then allows to measure the independence between the two writers S and the set of features G. Low values of mutual information indicate a strong independence between the two random variables while the high values denote a strong dependence between S and G. Independence between S and G should indicate that the set of features G is distributed in the same way on the two documents and should reflect the same identity for the two writers S_1 and S_2. On the contrary, the mutual information criterion should allow to detect cases that exhibit a strong dependence between S and G and to reveal different identities for the two writers. We point out the expression of mutual information between G and S:

$$I_M(G,S) = H(G) - H(G/S)$$

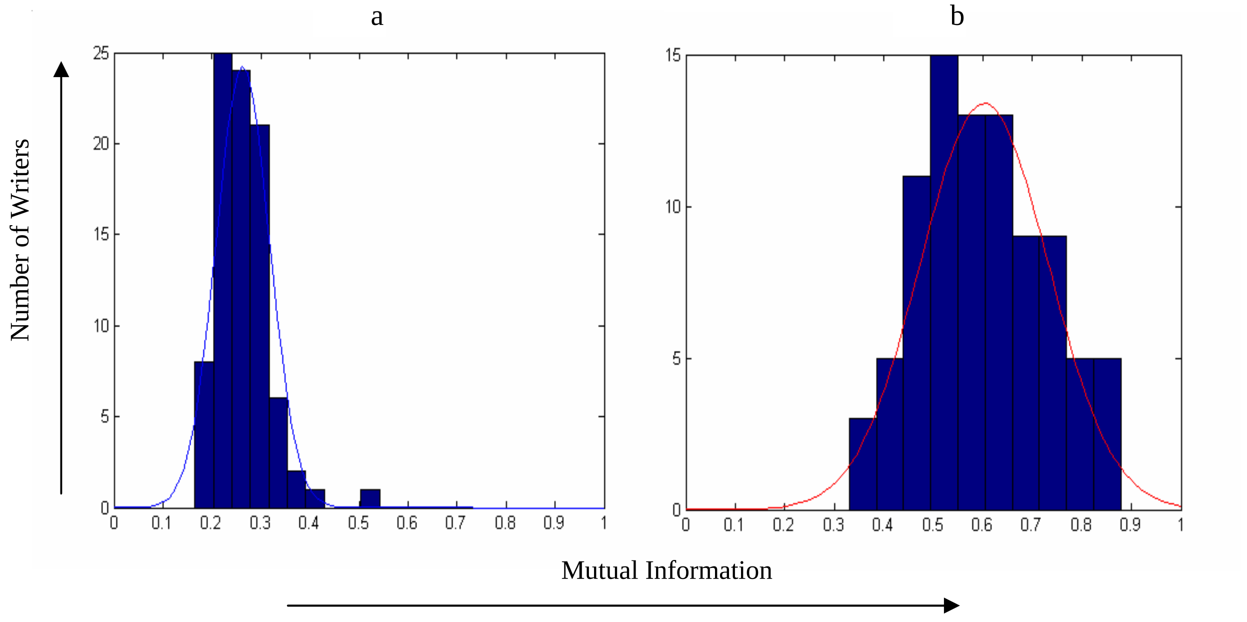
Where $H(G)$ indicates the Shannon entropy [10]:

$$\mathrm{H(G)}=-\sum_{\mathrm{i=1}}^{\mathrm{card(G)}} \mathrm{P(g_i)\log[P(G=g_i)]}$$

and $H(G/S)$ indicates the conditional entropy defined by:

$$H(G|\mathrm{S})=\sum P(\mathrm{Sj})H(G|\mathrm{S=Sj})=-\sum_{i=1}^{card(G)}\sum_{\mathrm{j=1}}^{\mathrm{card(S)}}\mathrm{P(S_j)\,P(g_i|S_j)}\;\;\log_2\left[\mathrm{P(g_i|S_j)}\right]$$

For assessing the interest of this criterion, an experiment was carried out on the PSI database. Figure 5 gives the distribution of the mutual information criterion in the two following cases : figure 5.a gives the distribution of the criterion in the case where the two writers are identical, while figure 5.b gives the distribution in the case where the two writers are different. From the observation of these two distributions it seems clear that mutual information should provide a quantitative criterion for the writer verification task. Furthermore, this figure shows that these two distributions can be approximated with a normal distribution.



(a) (b)

Figure 5. Mutual information criterion distributions in the intra-writer (a) and inter-writer (b) cases on the PSI database

3.1.2 Hypothesis test

We now seek to build a decision criterion between the two following hypotheses:

$$H_0 : S_1 = S_2 \text{ and } H_1 : S_1 \neq S_2$$

This can be accomplished using classical hypothesis testing [9]. H_0 will serve as the null hypothesis or the default hypothesis. Each of the two possible decisions is associated to a probability of correct decision and a probability of false decision or error probability. Probability of error on the null hypothesis is the first kind of error and is denoted α , while probability of error on H_1 is the second kind of error and is denoted β. Table 1 summarises the possible situations.

Truth Decision	H_0 is true	H_1 is true
Accept H_0	1- α	β
Accept H_1	α	1-β

Table 2: Associated probabilities to the different decisions

Assuming normal distribution of the mutual information criterion for the two hypotheses, it is very simple to quantify the errors of first and second order. Using these distributions and by asserting a value of the first order error, we can define the two regions of rejection and acceptance of the null hypothesis and deduce the experimental value of β.

The area of rejection of H_0, noted W_0, is defined by the first order error. The limit of this area allows to define the rejection area of H_1, noted W_1 and to deduce the second order error by the following relations:

$$P(W_0 \mid H_0) = \alpha \qquad \text{and} \qquad P(W_1 \mid H_1) = \beta$$

In the same way, one determines the acceptance regions of the two hypothesis, $\overline{W}_0$ for H_0 and $\overline{W}_1$ for H_1. We have:

$$P(\overline{W_0} \mid H_0) = 1\text{-}\alpha \qquad \text{and} \qquad P(\overline{W_1} \mid H_1) = 1 - \beta$$

3.2 Experimentation

We have evaluated this writer verification test on the IAM Database. Initially the PSI database was used to determine the acceptation regions of the two hypotheses by fixing a value to the first order error. This allows to evaluate the power of the test *(1- β)* to accept the hypothesis H_1 (table 3).

First Order Error (α)	5.0%	2.5%
Power of Test (1-β)	97.4%	96.4%

Table 3: First Order Error and Power of Test on the PSI database.

The test was then applied to couples of writers randomly chosen in the IAM database. Let us recall that this database includes Swiss writers and has been written in English language [13]. It is thus, *a priori*, very different from the PSI database. The writer verification test on this second database allows to obtain the results presented in table 4.

	Correct acceptation		Correct Rejection	
First Order Error (α)	5.0%	2.5%	5.0%	2.5%
Power of Test (1-β)	94.0%	97.3%	97.3%	94.0%

Table 4: Writer verification performance on the IAM database

3.3 Discussion

Concerning the approach proposed in this section for the writer verification task, the results seem particularly promising for several reasons. First of all the choice of a local representation founded on the segmented graphemes seems very relevant since it allows a level of description which is close to characters without however requiring a recognition stage. In addition, it is remarkable to obtain similar performance on the IAM database than those obtained on the PSI database on which the training test of hypothesis was carried out. We are thus able to bring relevant quantitative elements for the handwriting individuality assumption. We show moreover here that it is possible to build a robust statistical test on several databases of handwritings. It will naturally be necessary to validate the approach on more consequent databases of documents.

4 Conclusion

In this paper we have presented two complementary approaches for the writer recognition task. On the one hand we have adapted and applied to handwritten documents an information retrieval approach which is traditionally used on electronic documents. The proposed approach brings an original answer to the problem of writer identification of a document and offers an important potential of extension on large databases of patrimonial documents for example. In addition to its specific use on handwritten documents, this technique could easily be extended to other problems involving the characterization of textual documents by their graphical contents. Let us quote for example the problems of identification of typographies on old printed documents. Also let us notice that the approach is by construction compatible with some compression techniques using dictionaries such as JPEG or DjVu. For all of these reasons, the technique seems particularly interesting.

In addition we have proposed an hypothesis test allowing to verify compatibility between the handwritings of two different documents. This writer verification stage is essential to validate the assumptions made by the system of identification suggested previously. The approach shows excellent capacities of verification in both cases of acceptation and rejection and shows promising ability to be generalized on unknown handwritings. The approach considered here in complement with an information retrieval stage could be completely adapted to the context of biometric identification or legal expertise. With respect to these objectives it should be necessary however to evaluate the approach on forgeries. One of the major drawback of the proposed methodology is that it requires to work on a sufficient amount of handwritten material in order to be independent of the textual contents. A specific approach remains to be developed to work on samples of low size.

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] D. Feng, W.C. Siu, H.J. Zhang, “Multimedia Information Retrieval and Management”. Springer Edition, 2003.

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

[4] R.N. Morris, Forensic Handwriting Identification. Academic Press, 2000.

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

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

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

[8] G. Salton, Wrong, “A vector Space Model for Automatic Indexing”, *Information Retrieval and Language Processing*, pp 613-620, 1975.

[9] G. Saporta, Probabilités analyse des données et statistiques. *Edition Technip*. pp 317-330. 1990.

[10] C. Shannon, The Mathematical Theory of Communication. Bell System Technical. Journal, Roberts, J.A. vol 27; pp 379-423. 1984.

[11] F. Song, W. Bruce Croft, “A General Language Model for Information Retrieval”, *Eighth International Conference on Information and Knowledge Management* (ICIKM'99), 1999.

[12] S. Srihari, S.H.Cha, H. Arora, S. Lee, “Individuality of Handwriting : A Validity Study”, Proc. ICDAR’01, Seattle (USA), pp 106-109, 2001.

[13] M. Zimmermann, H. Bunke, “ Automatic Segmentation of the IAM Off-line Handwritten {English} Text Database”. 16th International Conference on Pattern Recognition, Canada, Vol 4, pp. 35-39, 2002.

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