Available online at www.sciencedirect.com
SCIENCE DIRECT®
Pattern Recognition Letters 26 (2005) 2080–2092
Pattern Recognition Letters
www.elsevier.com/locate/patrec

A writer identification and verification system

Ameur Bensefia, Thierry Paquet, Laurent Heutte *

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

Received 27 February 2004; received in revised form 8 December 2004
Available online 23 May 2005

Communicated by E. Backer

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 document; Writer identification; Writer verification; Information retrieval; Mutual information; Hypothesis testing; Graphemes

1. Introduction

Research studies concerning handwriting analysis have mainly investigated the automatic recognition of handwritten words in various particular contexts using either temporal information (on-line recognition) or scanned images (off-line recognition) (Plamondon and Srihari, 2000). This paper investigates the retrieval process of scanned handwritten document images in archive databases. Foreseen applications concern user assistance in querying large databases of images as for example forensic applications or digital libraries. When

* Corresponding author. Fax: +33 2 3514 6618.
E-mail addresses: ameur.bensefia@univ-rouen.fr (A. Bensefia), thierry.paquet@univ-rouen.fr (T. Paquet), laurent.heutte@univ-rouen.fr (L. Heutte).

0167-8655/$ - see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.patrec.2005.03.024

faced to handwritten document image databases, one can exhibit two kinds of use which correspond to two different kinds of query:

- Handwritten documents can be analysed for their textual content. In this case querying a handwritten document database needs 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 (Baeza-Yates and Ribeiro-Neto, 1999). Unfortunately, considering the state of the art in the field of handwriting recognition, such an approach does appear to be realistic yet. Indeed, the recognition of handwriting remains badly controlled on omni-writer applications when calling upon large size lexicons (Koerich et al., 2003).
- 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. Another possible application can deal with the detection of the various writings that may occur on a document, or the dating of the documents compared to the chronology of the work of the author.

It can be considered that these two applications fall into the problem of Information Retrieval either textual or graphical. These two tasks have been widely studied either in the electronic document retrieval field (Baeza-Yates and Ribeiro-Neto, 1999) or in the image processing field (Lew, 2001). 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 (Plamondon and Lorette, 1989), or in some institutes which analyze ancient manuscripts of 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.

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. As a rule, one strives for a near 100% recall of the correct writer in a hit list of 100 writers, computed from a database of up to 10,000 samples, which is the size of the search sets in current European forensic databases (Schomaker and Bulacu, 2004).

Due to the large number of classes, the identification cannot be considered as a simple classification task. Therefore, a two-stage strategy must be used to come to a conclusion concerning the authenticity of an individual. The first stage is the writer identification task while the second one is defined as the writer verification task:

1. 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 achieved by the expert.
2. 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 produced by the verification stage. 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-writer and between-writer distances (intra and inter writer variability) on a set of examples (Srihari et al., 2001).

The writer identification approach that we propose in the second section of this paper is based on a characterization of the writers derived from the graphemes obtained after a segmentation stage of the handwriting components. Therefore the method is not restricted to any particular style of writing (e.g. script or upper case, etc...). Furthermore, thanks to these local descriptors, the approach is completely independent of the textual content of the documents under analysis. The last originality of the method lies in the use of the well-known Vector Space Model originally proposed in the field of Information Retrieval (IR) in textual electronic document database (Salton and Wrong, 1975). This technique has proved to be particularly efficient on large databases for both its low complexity and its ability to deal with high dimensional feature vectors. The proposed method therefore exhibits similar properties for the writer identification problem on large databases.

Section 3 of this paper is devoted to the writer verification task or writer authentication. Indeed, if the approach proposed for the writer identification task is relevant to retrieve writers known by the system, it is not intended to reject samples of unknown writers. The verification approach uses the same segmented entities, the graphemes, to build a hypothesis test based on a mutual information criterion between the feature set and the set of the two documents under study, thus making the approach independent from the textual content by using local features. The main originality of this second stage lies in the self-adaptation of the feature set to each couple of handwritten samples under analysis, thanks to the use of an unsupervised clustering stage.

The two approaches have been evaluated on two different handwritten document databases: the first database was built in our lab (PSI) and is made up of 88 writers; the second database contains nearly 150 handwritten texts from 150 writers derived from the IAM database (Zimmermann and Bunke, 2002). The *PSI_DataBase* is written in French while the *IAM_DataBase* is written in English (see Section 2.3). In both cases, the approach demonstrates excellent capacities to retrieve and to verify handwritten samples. As a consequence, the proposed methodology shows that handwriting can be considered, in certain conditions, as a possible biometric identifiant.

The rest of the paper is organized as follows: Section 2 is dedicated to the writer identification process including a review of the major studies done in the field recently (Section 2.1) and a full description and evaluation of the approach (Sections 2.2–2.5). Section 3 is dedicated to the presentation of the writer verification approach including the construction of the hypothesis test (Section 3.1) and experimental results (Section 3.2). Finally, some conclusions and future works are drawn in Section 4.

2. Writer identification

In the identification approach, the user information need concerns the author identification of a single document. In this paper, this document will be considered to be the query, in reference to the Information Retrieval paradigm. The possible set of writer candidates is supposed to be finite and made up of the set of N writers (i.e. documents) stored in the database. While handwriting recognition has been largely studied during the last 20 years and still requires research efforts to reach acceptable performance for general purpose applications, the writer identification task does not seem to exhibit the same kind of difficulties. Indeed, whereas the recognition task must eliminate the variability between writers in order to be able to identify the textual content of the image for any writer, the writer identification task, on the contrary, has to use the variability between writers in order to discriminate them. From this point of view, the two tasks appear rather different. In this study each writer is supposed to use one single and stable writing, therefore detection of forgeries or manuscript dating are beyond the scope of this paper.

2.1. Previous works

One of the main difficulties associated with the writer identification task is to define a set of

features able to reflect the large variability between handwriting. In the literature, the features proposed for the writer identification task are most often global features (or macro features) which are based on statistical measurements, extracted from the whole block of text to be identified. Generally, these features can be broadly classified in two families:

- *Features from texture*: In this case the document image is simply seen as an image and not as a handwriting. For example, application of Gabor filters and co-occurrence matrices has been considered in (Said et al., 2000).
- *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 (Marti et al., 2001).

Note that it is also possible to combine these two families of features (Srihari et al., 2001). The nature of these statistical features, extracted from a block of text, has allowed to reach interesting performance that we have summarized in Table 1. Also notice that these results are however always difficult to compare due to a lack of common reference database and but also due to the large variability in the experiments carried out.

For example, the previous works can be categorized according to the number of writers and the nature of the training samples used by the system. On the one hand, the system is required to deal with as many writers as possible, while on the other hand training samples of each handwriting may sometimes represent several lines of text or on the contrary a few words. The work presented in (Said et al., 2000), for example, makes it possible to identify 95% of the 40 writers the system can deal with by the analysis of some lines of handwriting. The work presented in (Zois and Anastassopoulos, 2000) 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. Another parameter which is also associated with the size of the sample used for the identification is relative to its textual content, i.e. does the identification process require a particular textual content to operate or not? These experimental settings may largely restrict the generality of the approach when the experience is limited to one particular textual content that is not always reproductible from one writer to another, i.e. image data captured from real documents do not generally obey a particular textual content known in advance.

It should be noted that the work presented in (Srihari et al., 2001) dealt with the largest database (1000 writers), but they use the same text written three times by each writer. In the work reported in (Schomaker and Bulacu, 2004), the authors define a set of features from a first set of 100 writers using self-organizing feature maps. Writer identification is then carried out on another set of 150 other writers using the same feature set. Reported performance using the same copied text in uppercase letters reach 95% of correct writer identification. Table 1 below summarizes the performance achieved in these various studies. As a

Table 1
Comparison of performance and conditions of test for the writer identification task in most recent studies

	# Writers	Sample size	Lexicon dependency	Performances (%)
Said et al. (2000)	40	Few lines of handwritten text	Yes	95
Zois and Anastassopoulos (2000)	50	Forty five samples of the same word	Yes	92.48
Marti et al. (2001)	20	Five samples of the same text	Yes	90
Srihari et al. (2001)	100	CEDAR letter/paragraph/word	Yes	82/49/28
	900	CEDAR letter/paragraph/word	Yes	59/25/9
Schomaker and Bulacu (2004)	150	One copied text paragraph in uppercase handwriting	No	95
Our approach	88	Paragraph/3–4 words	No	93/90
	150	Paragraph/3–4 words	No	86/68

comparison, the results and conditions of experiment of the work proposed in this paper are reported in the last row of the table.

2.2. Organization of the system

We present in the following subsections the various steps of our writer identification system. Fig. 1 gives a brief overview of the data processing sequence. It uses three steps: a segmentation step first aims at locating information that will be used to perform writer identification, then a binary feature step is used where the goal is to obtain a relevant representation for the retrieval process which represents the final step of the system.

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

2.2.1. Segmentation

Our method is able to cope with unconstrained handwriting thanks to this segmentation step. Up to now, very few studies have dealt with unconstrained handwriting. Yet, while perfect segmentation into characters is impossible and therefore implies sophisticated recognition procedures, the writer identification task is not so restrictive with respect to the segmentation. On the contrary, the variability in segmenting characters will bring to the method more information to characterize each writer.

First, the connected components of the document image are extracted and analyzed in order to eliminate some charts like erasures, overloaded or underlined zones which as one knows, do not obviously 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 of the upper contour (Nosary et al., 1999). Fig. 2 gives one example of the segmentation obtained on the french word "manuscript".

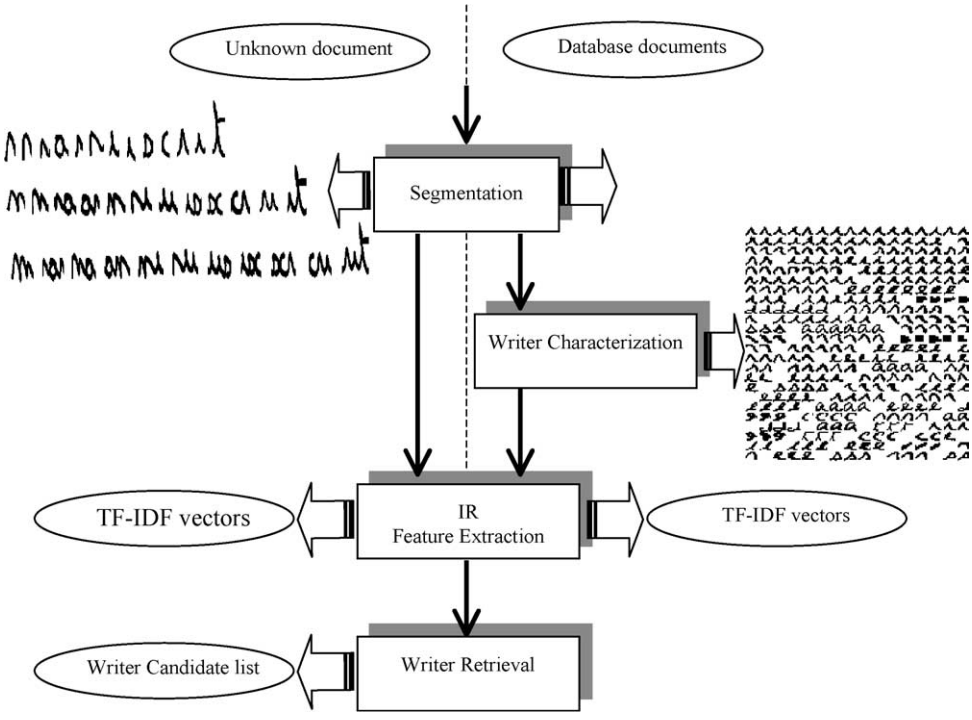

Fig. 1. Writer identification system organisation.

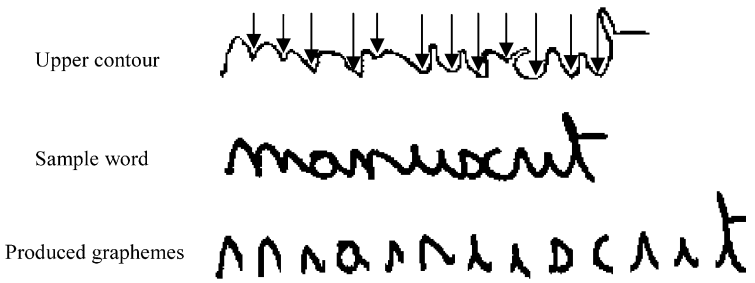

Fig. 2. Potential segmentation points and final segmentation into graphemes.

The concatenation of two (respectively three) adjacent graphemes provides what we call bigrams (respectively trigrams) of graphemes.

2.2.2. Writer characterization

The writer identification task lies in the definition of a feature space common to all the handwritten documents. In a previous study, we have shown that graphemes can characterize each handwriting (Nosary et al., 1999). In this study we have extended this principle to the whole document database. Following the segmentation of the handwritten documents into graphemes, a set of binary features is defined thanks to a clustering procedure. Thus, the feature set is adapted to the database under study, and not defined in advance.

We briefly recall the main characteristics of our clustering procedure which is based on sequential clustering (Friedman and Kandel, 1999). Unlike most of the clustering procedures such as *k*-means or self-organizing maps, the sequential clustering has some advantages that are suitable for our purpose. First it is a simple, fast and effective procedure to cluster a set of data points. Second, it does not require any fixed number of clusters: the total number of clusters can be unknown and is therefore dynamically determined. Finally, it does not resort to multiple iterations to converge towards the final location of cluster centroïds. Despite these advantages that make this clustering procedure well adapted for our problem, it is very sensitive to the order the data points to cluster are being visited. It requires therefore several clustering phases with random selection of the data points in order to be less sensitive to the initial conditions. Each of the clustering phases provides thus 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. These clusters constitute a set of binary features that will be used in the writer identification process. Fig. 3 gives some clusters obtained on the database where gray level shows intra-cluster variability.

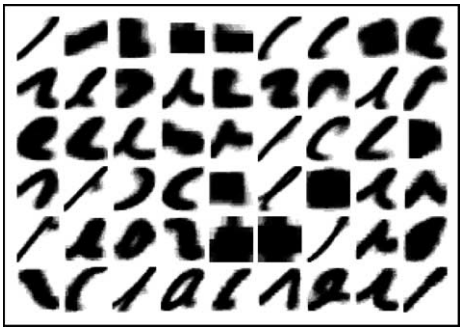

Fig. 3. Some invariant clusters of level 1 obtained on the *PSI_DataBase*.

2.2.3. Information retrieval based feature computation

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 in a large database. The user need is expressed by a query. For this purpose the query and the documents of the database are described in the same feature space. The choice of the feature space is therefore of primary importance. As the documents must be described so as to cope with any kind of query, one cannot resort to any specific

feature selection procedure that could reduce the dimensionality of the feature space. Therefore, one generally seeks to describe documents by preserving the whole set of extracted features. This leads to a description of documents in a high dimensional feature space.

The problem of writer identification can be defined 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 ranked according to their similarity with the query. There are several types of Information Retrieval models (Song and Bruce Croft, 1999): the boolean model, the probabilistic model and the Vector Space Model are the most popular models. Among them the Vector Space Model (VSM) proposed by Salton (Salton and Wrong, 1975) still remains very effective (Feng et al., 2003), even though it is very simple and of a rather old design.

This model involves two different phases: a preliminary indexing phase is intended to describe each document with 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, the relevance of each document is evaluated by the scalar product between the vector describing the query Q and the vector describing a document D_j of the database. We now present each of the two phases of the model.

2.2.3.1. Indexing phase: feature extraction. Assume a binary feature set has been chosen. Denote $\varphi_{i,1\leqslant i\leqslant m}$ 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:

$$\vec{D}_j = (a_{0,j}, a_{1,j}, \ldots, a_{m-1,j})^{\mathrm{T}} \quad \text{and}$$
$$\vec{Q} = (b_0, b_1, \ldots, b_{m-1})^{\mathrm{T}}$$

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

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

$\mathrm{FF}(\varphi_i,D_j)$ is the *Feature Frequency* in document D_j. $\mathrm{IDF}(\varphi_i)$ 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

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

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

Note that $\mathrm{IDF}(\varphi_i) = 0$ 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.

2.2.3.2. 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

$$\cos(Q, D_j) = \frac{\sum a_{i,j}b_j}{\sqrt{\sum_{\varphi_i} a_{i,j}^2 \sum_{\varphi_i} b_j^2}}$$

where the two terms in the denominator are the lengths of the document and of the query respectively. The retrieval process has thus a complexity of $\mathrm{O}(TN)$, where T 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 the indexing and retrieval phases 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 2.2.2.

For the evaluation, two different databases have been used. The first one has been constituted in our lab (PSI) and contains 88 writers who have been asked to copy a letter (in french) that contains 107 words (Fig. 4a). The scanned images have been divided into two parts: the first two thirds are used for learning, the last third of each page is used for testing. The second database that we have used is the world wide web accessible part of the IAM database (Zimmermann and Bunke, 2002), which represents nearly 20% of the whole database. The fraction of the database that we have used contains texts written in English by 150 different writers. The textual content varies from one writer to another (Fig. 4b).

As graphemes can be merged to produce either bigrams or trigrams (a larger window could possibly 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 an handwriting or not.

2.4. Performance

Fig. 5 gives the results of the approach on the *PSI_DataBase* and the *IAM_DataBase*. This figure exhibits similar good results on the two databases. Results on the *PSI_DataBase* show higher performance (10%) in the top 1 writer candidate. This can be explained by the lower number of writer candidates in the *PSI_DataBase* compared to the *IAM_DataBase*. Another difference between the two databases is the lower performance of the bigram level on the *IAM_DataBase* which can be explained by the fact that this database contains smaller text samples than the *PSI_DataBase*. In both cases trigrams show the same significant decrease in identification performance and this can be explained by the fact that trigram features may be more dependent on the textual content. Therefore, while some of the trigrams 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, these 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

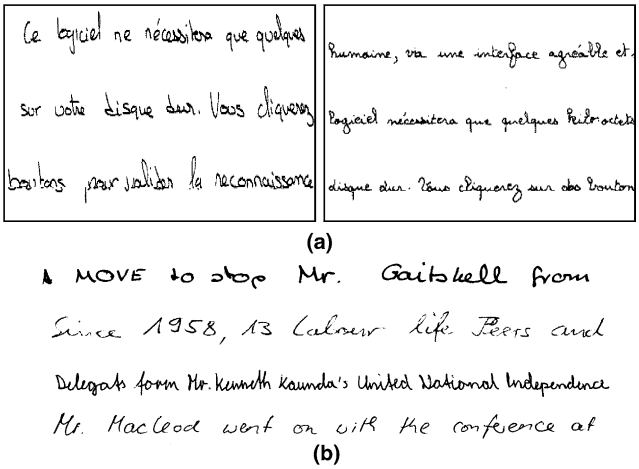

Fig. 4. Some handwriting samples of the *PSI_DataBase* (a) and the *IAM_DataBase* (b).

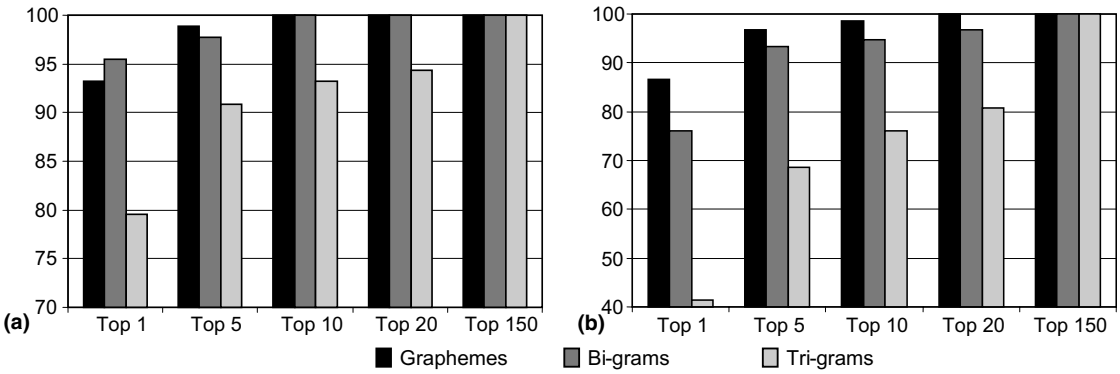

Fig. 5. Writer identification performance on the *PSI_DataBase* (a) and on the *IAM_DataBase* (b).

of the size of the query on the identification performance. Correct identification is achieved using 50 graphemes in nearly 90% of the cases on the *PSI_DataBase* while this performance decreases to nearly 68% for the *IAM_DataBase* in the same condition (see Fig. 6). In both cases we can note that trigrams have a discriminative power significantly lower than graphemes or bigrams as it was also measured using large queries (Fig. 5).

One can thus assert that these results are relevant compared to the other methods proposed in the literature (see Table 1). Moreover, by using segmented graphemes as local features, the method is able to perform correctly (with a slight decrease in performance) using only 3 or 4 words (50 graphemes) on a basis of 88 writers in 90% of the cases.

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 bigram 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. However, this principle raises the problem where the writer to be identified does not belong to the reference database. In this case, a possible solution lies in the authentication (verification) of the writer proposed by the identification stage.

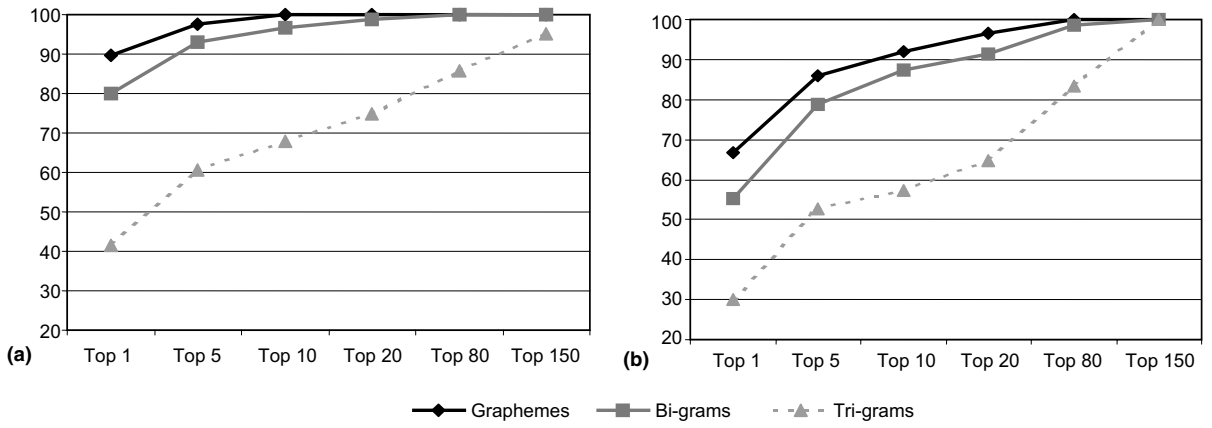

Fig. 6. Writer identification performance using short queries of 50 graphemes on the *PSI_DataBase* (a) and the *IAM_DataBase* (b).

We now investigate the use of graphemes in the writer verification task.

3. Writer verification

Let us recall that 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 (Morris, 2000). In any case, the confidence which one can associate to a decision of this type is not scientifically proven. A recent work has however proposed a scientific methodology of handwriting analysis for the task of writer verification (Cha and Srihari, 2000). 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 decision process which generally depends on the textual content of the document. 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 be prone to character recognition errors.

In this section we propose a writer verification approach which is independent from the textual contents. Note that this is only possible when the amount of information available for the analysis is large enough i.e. under the same conditions as for the task of writer identification (a block of text). Although this assumption may seem very restrictive for the expertise, it appears to be 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 verification approach allows to validate or to reject the handwritten documents output by the identification stage. 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 may come from the same writer or not. For this purpose, we build a hypothesis test based on a mutual information criterion between the two handwritten documents.

3.1. Construction of a 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: $G = \{g_1, g_2, g_3, \ldots, g_N\}$. Notice that this feature set is writer dependent and is therefore different from the one used in the identification process.

Some of these features can be present on the two documents, while others can appear specifically on one single document. Mutual information then allows to measure the independence between the set of 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 high values denote a strong dependence between S and G. Independence between S and G should indicate that the set of features G is evenly distributed over 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 thus revealing 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 (Shannon, 1984)

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

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

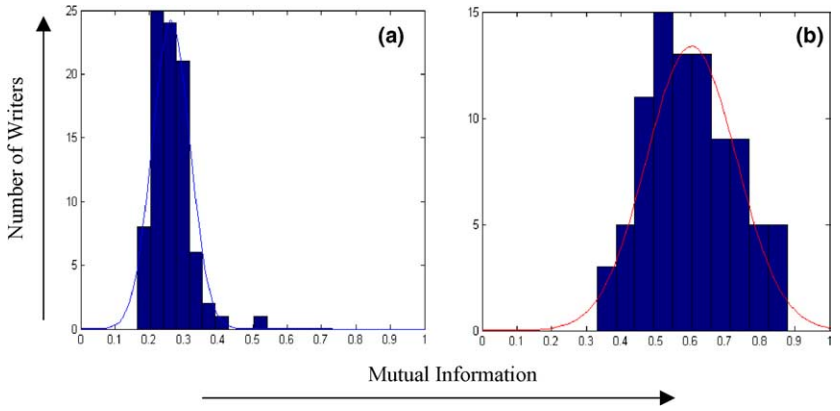

Fig. 7. Distribution of the mutual information criterion in the intra (a) and the inter writer (b) cases.

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

To assess the interest of this criterion, an experiment was carried out on the *PSI_DataBase*. These samples have been split into two documents (first and second half). Fig. 7 gives the distribution of the mutual information criterion in the two following cases: Fig. 7a gives the distribution of the criterion in the case where the two writers are identical (intra-writer distribution), while Fig. 7b gives the distribution in the case where the two writers are different (inter-writer distribution). 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.

3.1.2. Hypothesis test

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

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

This can be done using classical hypothesis testing (Saporta, 1990). 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 the probability of error on H_1 is the second kind of error and is denoted β. Table 2 summarizes the possible situations.

Assuming a normal distribution of the mutual information criterion for the two hypothesis, it is very simple to quantify the errors of first and second kind. Using these distributions and by giving a value to 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:

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

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

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

Table 2
Associated probabilities to the different decisions

Decision	Truth	
	H_0 is true	H_1 is true
Accept H_0	$1 - \alpha$	β
Accept H_1	α	$1 - \beta$

Table 3
First order error and power of the test on the *PSI_DataBase*

First order error (α)	0.05	0.9744
Power of test ($1-\beta$)	0.025	0.9641

Table 4
Writer verification performance on the *IAM_Database*

Correct acceptance (%)		Correct rejection (%)	
$\alpha = 5$	$\alpha = 2.5$	$\alpha = 5$	$\alpha = 2.5$
94	97.33	97.33	94

3.2. Experimentation

We have evaluated this writer verification test on the *IAM_DataBase*. Initially the *PSI_DataBase* was used to determine the acceptance regions of the two hypothesis by fixing a value to the first kind of error. This allows to evaluate the power of the test $(1-\beta)$ to accept the hypothesis H_1 (see Table 3).

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 (Zimmermann and Bunke, 2002). It is thus very different from the *PSI_DataBase*. The writer verification test on this second database allows to obtain the results presented in Table 4.

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 based 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* as 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. Moreover, we show 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 larger databases of documents. But up to now and to the best of our knowledge, no other work has shown so general results for the writer verification task.

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 handwritings, 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 based on dictionaries such as those used by the standards JPEG or DjVu. For all these reasons, the technique seems particularly interesting.

In addition, we have proposed a 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 acceptance 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 fully 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 drawbacks 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 lower size samples of handwriting.

References

Baeza-Yates, R., Ribeiro-Neto, B., 1999. Modern Information Retrieval. Addison-Wesley.

Cha, S.H., Srihari, S., 2000. Multiple feature integration for writer verification. In: 7th International Workshop on Frontiers in Handwriting Recognition; IWFHR VII, Amsterdam, The Netherlands, pp. 333–342.

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

Friedman, M., Kandel, A., 1999. Introduction to Pattern Recognition: Statistical, Structural, Neural and Fuzzy Logic Approaches. Imperial College Press, pp. 55–98. (Chapter 3).

Koerich, A.L., Sabourin, R., Suen, C.Y., 2003. Large vocabulary off-line handwriting recognition: A survey. Pattern Anal. Applications vol. 6, 97–121.

Lew, M.S., 2001. Principles of Visual Information Retrieval. Springer Publisher.

Marti, U.V., Messerli, R., Bunke, H., 2001. Writer identification using text line based features. In: Proc. ICDAR'01, Seattle (USA), pp. 101–105.

Morris, R.N., 2000. Forensic Handwritting Identification. Academic Press.

Nosary, A., Heutte, L., Paquet, T., Lecourtier, Y., 1999. Defining writer's invariants to adapt the recognition task. In: Proc. ICDAR'99, Bangalore (India), pp. 765–768.

Plamondon, R., Lorette, G., 1989. Automatic signature verification and writer identification—the tate of the art. Pattern Recognition 22 (2), 107–131.

Plamondon, R., Srihari, S.N., 2000. On-line and off-line handwriting recognition: A comprehensive survey. IEEE Trans. PAMI 22 (1), 63–84.

Said, H.E.S., Tan, T.N., Baker, K.D., 2000. Personal identification based on handwritting. Pattern Recognition 33, 149–160.

Salton, G., Wrong, A., 1975. A Vector Space Model for automatic indexing. Information Retrieval and Language Process., 613–620.

Saporta, G., 1990. Probabilités analyse des données et statistiques. Edition Technip., 317–330.

Schomaker, L., Bulacu, M., 2004. Automatic writer identification using connected-component contours and edge-based features of uppercases western script. IEEE-PAMI 26 (6), 787–798.

Shannon, C., 1984. The mathematical theory of communication. Bell System Tech. J. 27, 379–423.

Song, F., Bruce Croft, W., 1999. A general language model for information retrieval. In: Eighth Internat. Conf. on Information and Knowledge Management (ICIKM'99).

Srihari, S., Cha, S.H., Arora, H., Lee, S., 2001. Individuality of handwriting: A validity study. In: Proc. ICDAR'01, Seattle (USA), pp. 106–109.

Zimmermann, M., Bunke, H., 2002. Automatic segmentation of the IAM off-line handwritten {English} text database. In: 16th Internat. Conf. on Pattern Recognition, Canada, vol. 4, pp. 35–39.

Zois, E.N., Anastassopoulos, V., 2000. Morphological waveform coding for writer identification. Pattern Recognition 33 (3), 385–398.