# Information Retrieval Based Writer Identification

A. Bensefia, T. Paquet, L. 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 Writer Identification task. Our previous work has shown the interest of using the graphemes as features for describing the individual properties of Handwriting. We propose here to exploit the same feature set but using an information retrieval paradigm to describe and compare the handwritten query to each sample of handwriting in the database. Using this technique the image processing stage is performed only once and before the retrieval process can take place, thus leading to a significant saving in the computation of each query response, compared to our initial proposition. The method has been tested on two handwritten databases. The first one has been collected from 88 different writers at PSI Lab. while the second one contains 39 writers from the original correspondence of Emile Zola, a famous French novelist of the last 19th century. We also analyze the proposed method when using concatenation of graphemes (bi and tri-gramme) as features.*

## 1.Introduction

In this communication we present a methodology for the identification of the writer of a document. This task has been defined as the one of assigning to an unknown handwritten document its correct writer among a finite set of possible candidates [9]. The implicit hypothesis behind this task is the handwriting individuality. This assumption has proved to be founded since interesting performance have already been obtained in various experiments [9, 10, 11].

Two main approaches have generally been considered in the literature regarding the kind of features used to characterize each handwriting. When a sufficient amount of handwritten material is available in the image, few global robust features can be defined on text blocks and provide the information need. On the contrary, when the input image contains few samples of handwriting, local features are required to characterize the handwritten fragments.

In this communication, like in our previous work [5], the second approach is adopted. This choice leads to represent each handwritten input image in a high dimensional feature space in order to capture the whole variability over the database. As a consequence, the writer identification task consists in finding similar documents represented in a high dimensional feature space. In the field of information retrieval (IR), this problem has been intensively studied and is still motivating a large number of research, especially due to the need for web document retrieval.

In this communication we investigate the use of one of the most popular schemes used in IR [1] and apply it to the task of writer identification. In part 2 of this communication we recall our previous approach. Part 3 is devoted to the presentation of the IR model known as the "vector space model" in the literature. In parts 4 and 5 we evaluate the proposed approach on two different databases. The first one has been constituted in our lab, originally for recognition purposes. It contains 88 different writers. The second one contains 39 writers that have taken part in a correspondence with Emile ZOLA, a French novelist of the late 19th century.

## 2. Writer Identification

In our previous work [6] an original approach for writer identification has been proposed based on local features such as graphemes. This study has also shown that although prone to variability, each handwriting can be characterized by a set of invariant features also called the writer's invariants. Writer identification can be efficiently carried out using the writer's invariants instead of using elementary graphemes, without no significant loss in the identification performance.

Each grapheme is produced by the segmentation module of our recognition system [6,7]. In this system, letter hypothesis are analyzed up to the concatenation of 3 consecutive graphemes.

Each handwritten document $D_j$ is thus described by the set of graphemes $x_i$ it is made of :

$$D_j = \{x_i, i \leq card(D)\} \quad (2.1)$$

A similarity measure between an unknown handwritten document $Q$ and a reference document in the database $D$ can be defined according to the following relation :

Proceedings of the Seventh International Conference on Document Analysis and Recognition (ICDAR 2003)
0-7695-1960-1/03 $17.00 © 2003 IEEE

$$SIM(Q, D) = \frac{1}{card(Q)} \sum_{i=1}^{card(Q)} Max_{y_j \in T}(sim(y_i, x_j)) \quad (2.2)$$

where $y_i$, $x_j$ are graphemes that belong respectively to document $Q$ and $D$, and $sim(y_i, x_j)$ is a similarity measure between two graphemes. Among many others, the correlation measure has been chosen for its average properties. Therefore, two documents will be all the closer as this measure will be close to one. The writer of document $Q$ will be determined as the writer of the closest document in the database.

$$Writer(Q) = Writer(Arg\max_{D_j \in base}(SIM(Q, D_j))) \quad (2.3)$$

The first evaluation of this approach was carried out on a database of 88 writers that has been constituted in our lab. Two experiments were conducted: the first one was design to measure the performance of the approach on large blocks of text; the second one was designed on small handwritten queries.

The results are encouraging, giving rise to a correct identification of nearly 98% when working with large handwritten samples as queries (typically 3 lines of text). When dealing with small queries (typically 50 graphemes) the correct writer was determined in nearly 93% of the cases. These results have shown the interest of using graphemes as local features for writer identification.

Two major drawbacks of this approach can however be pointed out. The first one is that it is especially computationally expensive due to the pattern matching technic employed. Assume T is the average size of a document, then the complexity of the retrieval process is $O(T^2N)$, where N is the number of documents in the database. The second one arises when using invariant graphemes as features. In this case, when calculating the similarity between two documents, each feature is assigned the same weight, no matter its effective frequency in the document.

## 3. Information Retrieval Model

Information Retrieval techniques have been designed in order to query textual documents described in a high dimensional feature space such as terms. Therefore, the problem of binary feature encoding and document querying has been particularly studied in this field. An Information Retrieval system is characterized by [2]:

- The set of documents that constitute the database.
- An Information Retrieval model that orders documents in the database according to their respective similarity with the query.
- Document processing: documents are processed in order to gather statistical information.

One of the most popular model in IR was proposed by Salton [1]. Its first advantage is to propose a retrieval model that integrates the description of the documents and the query in a single high dimensional feature space. High dimensionality ensures a minimum loss of information when describing each document in the database as well as the query. The second advantage is that, once the feature space has been defined, each document can be described independently from the query, thus avoiding any other access to the document content when responding to a query. This last point is of particular interest regarding our problem of writer identification which requires intensive image matching. Although very simple, this model is still popular in the IR community [3, 4].

Various kinds of features can be used to describe an electronic document. They can be words, n-grammes, letters, html tags... In the feature space a similarity measure will then be defined between the query and each document, thus giving an ordered list of relevant documents regarding the query content. Two distinct steps are required: the indexing phase concerns the processing of each document in order to obtain a high dimensional vector that describes the document; the retrieval phase concerns the calculation of the relevance score of each document for a particular query.

### 3.1. Indexing phase

Assume a binary feature set has been chosen. Denote $\varphi_i$ , $1 \le i \le m$ the $i^{th}$ 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:

$$\vec{D_j} = (a_{o,j}, a_{1,j}, \ldots . \quad a_{m-1,j})^T \quad (3.1)$$

$$\vec{Q} = (b_o, b_1, \ldots . \quad b_{m-1})^T \quad (3.2)$$

where : $a_{i,j}$ and $b_i$ are weight assigned to each characteristic $\varphi_i$, and are defined by:

$$a_{i,j} = FF(\varphi_i, D_j)\, IDF(\varphi_i) \quad (3.3)$$

$$b_i = FF(\varphi_i, Q)\, IDF(\varphi_i) \quad (3.4)$$

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

$IDF(\varphi_i)$ is the *Inverse Document Frequency* and is the inverse of the number of documents that contain this characteristic $\varphi_i$, it is exactly defined by :

$$IDF(\varphi_i) = log\left(\frac{1+n}{1+DF(\varphi_i)}\right) \quad (3.5)$$

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 characteristic. Notice that $IDF(\varphi_i) = 0$ when $\varphi_i$, occurs in every document. Such characteristics will therefore be given a null score and should indeed be eliminated from the feature set.

### 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. 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 of the two vectors. Therefore the similarity measure between document $D$ and the query $Q$ is defined by:

$$\cos(Q, D_j) = \frac{\sum TFIDF_{\varphi_i, D_j} TFIDF_{\varphi_i Q}}{\sqrt{\sum_{\varphi_i} TFIDF^2_{\varphi_i, D_j} \sum_{\varphi_i} TFIDF^2_{\varphi_i, Q}}} \qquad (3.6)$$

where the two terms in the denominator are the lengths of the document and the query respectively. Compared to the direct pattern matching method, the retrieval process has a complexity of O(TN), where T is the size of the feature vector and N the number of documents in the database.

## 4 IR applied to writer identification

In this section we discuss the implementation of the IR model 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 3.

Let us recall that our initial works have implemented writer identification based on local features such as graphemes (see section 2). Besides, we have shown that writer identification can be efficiently carried out using invariant clusters within the set of graphemes of each writer.

Therefore, the writer's invariants can be viewed as binary features defined within the writer's set of graphemes. In order to define a set of binary features common to all the handwritten documents it is required to cluster all the graphemes of the database. For this purpose, the procedure described in [6] is used. We briefly recall its main characteristics. Many sequential clustering phases are iterated with random selection. Each of them provides a variable number of clusters. The invariant clusters are defined as the groups of patterns that have been clustered together after each sequential clustering phase.

Figure 2 gives some of the most frequent clusters obtained on the PSI_BASE (see next section for details). These features can occur for different writers. A feature is all the more pertinent as it belongs to a low number of writers. TF-IDF scores will thus be calculated for each feature and each document during the indexing phase.

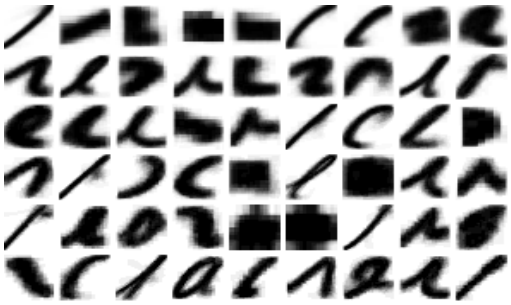


**Figure 2.** Some invariant clusters of the PSI_BASE.

## 5 Experiment

### 5.1. Description of the databases

Two databases have been used for this experiment. The first one (PSI_BASE) contains 88 writers who have been asked to copy a letter 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 base (ZOLA_BASE) contains 39 writers that have taken part to a correspondence with Emile ZOLA, a famous novelist of the late 19th century (1840-1902). These images have been scanned from a microfilm with a resolution of 300 dpi. They present a higher degree of difficulty than those of the PSI_BASE for various reasons: presence of noise, overlapping lines, slant, type of nib or quill used at the end of the 19th century. Finally this database contains completely free writing. The original microfilm contains nearly 700 documents. This database was first inspected and manually annotated in order to discard from the analysis, irrelevant areas such as printed zones, marks, etc... Although it contains a relatively large number of documents, they are far from being equally distributed among writers. The number of words can vary dramatically from one document to another. For these reasons, the ZOLA_BASE was designed with text blocks having a sufficient amount of information and at the same time with a sufficient number of writers. The result was

thus a compromise of these two criteria. The learning base contains 39 documents each containing between 5 and 7 handwritten lines, while the test base contains text blocks between 3 and 5 lines long. Figure 3 gives some samples of the ZOLA_BASE.

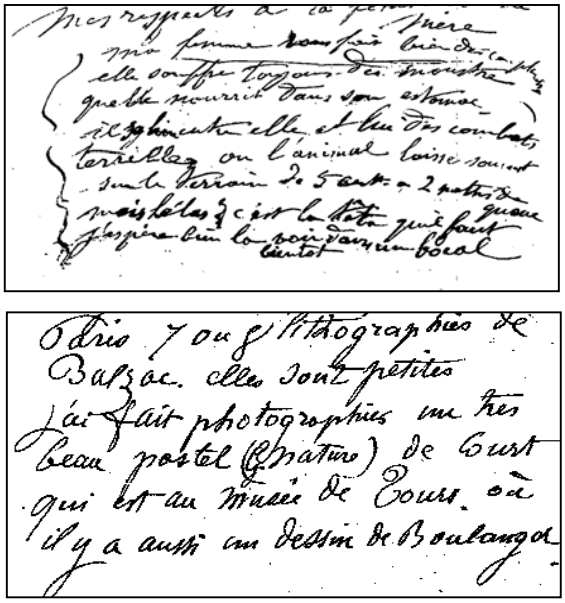


**Figure 3.** Some samples of the ZOLA_BASE.

Due to the variability of the ZOLA_BASE it was necessary to modify our segmentation algorithm in order to operate on slanted connected components and without any knowledge of the reference lines. Therefore, the grapheme produced by this segmentation step can vary from those produced on the PSI_BASE. Figure 4 gives the segmentation result on one example.

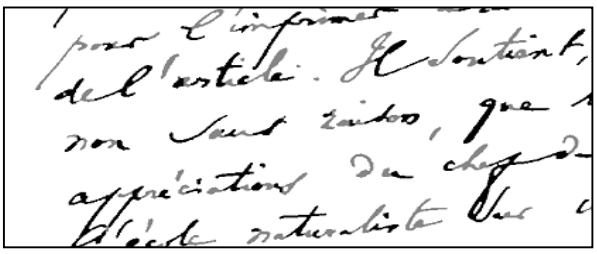


**Figure 4.** Segmentation produced on the ZOLA_BASE.

As connected 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, if our previous study has shown that graphemes are good local features, it is however unclear whether concatenations of these features can better characterize a writing or not. Table 1 summarizes the properties of the two databases on the three levels of analysis.

| | | level 1 | level 2 | level 3 |
|---|---|---|---|---|
| *PSI_BASE* | # graphemes | 43178 | 25088 | 15953 |
| | # binary features | 7230 | 13876 | 12722 |
| *ZOLA_BASE* | # graphemes | 25907 | 15647 | 10670 |
| | # binary features | 3567 | 5489 | 6266 |

**Table 1.** Properties of the two databases.

### 5.2. Results

Figure 5 gives the performance of the approach on the PSI_BASE. It shows that the correct writer is determined in 93% (83/88) of the cases using first level graphemes. Identification rate rises up to 95.45% (84/88) using bi-grams as features, while tri-grams give only 80% (70/88) of correct identification. Let us recall that in our initial work [5] a correct identification rate of 97% was obtained on the first level graphemes but intensive pattern matching was required in this case.

This first result shows 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.

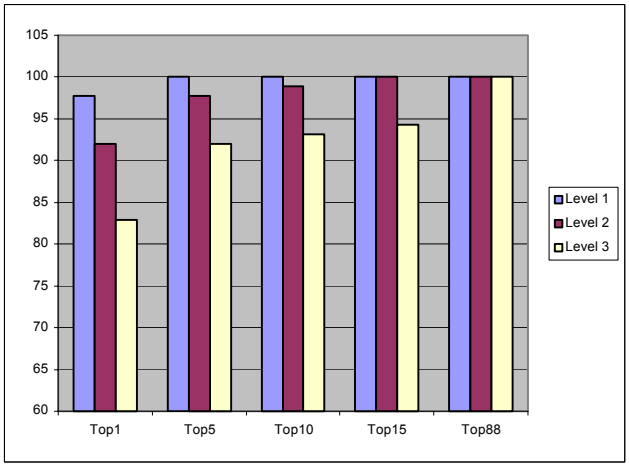
Two reasons can explain the lower performance obtained on tri-gram. The first one is due to the fact that tri-gram features being more numerous, each one of them is thus less frequent and therefore cannot be as representative of a particular writer as lower level features (bi-gram or graphemes). The second reason is that tri-gram features may be more dependent on the textual content. Therefore, while it may be a pertinent feature for the writer, its 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.

Results obtained on the ZOLA_BASE are significantly lower than those obtained on the PSI_BASE. Particularities of this database are given in section 5.1 and can explain these results. Nevertheless, the method allows a correct identification of 93,3% (36/39) in the top 5 propositions. However bi-gram features are not as informative as on the PSI_BASE.

**Figure 5.** Writer identification on the PSI_BASE.

**Figure 6.** Writer identification on the ZOLA_BASE.

## 6. Conclusion




In this communication we have presented an information retrieval based writer identification method. The results obtained are comparable to those presented in our previous work, but the information retrieval model has a linear complexity which is one order less than our initial method.

The method has been tested on two different databases. On a clean database the method performs very well, furthermore it is shown that bi or tri-gram features can also bring interesting information about the writer. On a more noisy database, performance decrease but the method still provides an interesting means to query handwritten documents.

**Acknowledgement**

This study was sponsored by the French program CNRS STIC-SHS.

The authors are grateful to DPCI for scanning the microfilm of Zola's letters (www.dpci.com)

## References

[1] Salton, Wrong "A vector Space Model for Automatic Indexing", Information retrieval and language processing, pp 613-620, 1975.

[2] P.Schaüble, « Multimedia Information Retrieval : Content–Based Information Retrieval from Large Text and Audio Databases », Kluwer Academic publishers, 1997.

[3] B. Pouliquen, D.Delamane, P.Lebeux, *« indexation des textes médicaux par extraction de concepts et ses utilisations »* JADT, 6éme journées d'analyse statistique des données textuelles, 2002.

[4] D. Memmi, « Le modèle vectoriel pour le traitement des documents ». Les cahiers du laboratoire Leibniz – IMAG-grenoble, France n°14, 2000.

[5] A.Bensefia, A.Nosary, T.Paquet, L.Heutte, "*Writer Identification by Writer's Invariants* », International Workshop on Frontiers in Handwriting Recognition, IWFHR'01, pp 274-279, 2002.

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

[7] Nosary A., Reconnaissance automatique destextes manuscrits par adaptation au scripteur, Thèse de Doctorat, Université de Rouen, 2002.

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

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

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

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