# Biografo: An Integrated Tool for Forensic Writer Identification

Javier Galbally, Santiago Gonzalez-Dominguez, Julian Fierrez, and Javier Ortega-Garcia

Biometric Recognition Group-ATVS, EPS, Universidad Autonoma de Madrid
C/ Francisco Tomas y Valiente 11, 28049 Madrid, Spain

**Abstract.** The design and performance of a practical integrated tool for writer identification in forensic scenarios is presented. The tool has been designed to help forensic examiners along the complete identification process: from the data acquisition to the recognition itself, as well as with the management of large writer-related databases. The application has been implemented using JavaScript running over a relational database which provides the whole system with some very desirable and unique characteristics such as the possibility to perform all type of queries (e.g., find individuals with some very discriminative character, find a specific document, display all the samples corresponding to one writer, etc.), or a complete control over the set of parameters we want to use in a specific recognition task (e.g., users in the database to be used as control set, set of characters to be used in the identification, size of the ranked list we want as final result, etc.). The identification performance of the tool is evaluated on a real-case forensic database showing some very promising results.

**Keywords:** Forensics, writer identification, data acquisition, database management

## 1 Introduction

Analysis of handwritten documents with the aim of determining the writer identity is an important application area in forensic casework, with numerous cases in courts over the years that have dealt with evidence provided by these documents [1]. Handwriting is considered individual, as shown by the wide social and legal acceptance of signatures as a mean of identity validation, which is also supported by experimental studies [2]. The goal of writer recognition is to determine whether two handwritten documents, referred to as the *control* document (i.e., generated by a known writer) and the *questioned* document (i.e., generated by an unknown writer), were written by the same person or not. For this purpose, computer vision and pattern recognition techniques have been applied to this problem to support forensic experts [3, 4].

The forensic scenario presents some difficulties due to its particular characteristics in terms of [5]: frequently reduced number of handwriting samples, variability of writing style, pencil or type of paper, the presence of noise patterns, etc. or the unavailability of online information. As a result, this application domain still heavily relies on human-expert interaction. The use of semi-automatic recognition systems is very useful

to, given a questioned handwriting sample, narrow down a list of possible candidates which are comprised in a database of known identities, therefore making easier the subsequent confrontation for the forensic expert [4, 5].

However, before reaching the recognition phase itself, forensic examiners have to manually go through a number of steps which include: labeling the data, segmenting the characters of the new handwriting samples or manually handling all the data of large databases. Although some efforts have been made in the automation of several of these steps [6, 7], usually, for each of the stages, different independent tools are used or, in the worst cases, no practical applications are available. This fact hinders and slows down the already difficult task of the forensic specialists and increases the chances of human errors.

In this context, we have developed Biografo, a tool that integrates over a relational database the different steps involved in the forensic identification of unknown writers, automating all the tasks related to the management of data and presenting a number of functionalities thought to make more efficient the work of forensic examiners. The application intends to give practical solutions to problems encountered by examiners in real-world case scenarios and has been designed based on a previous very schematic and simple software [8], according to the advice and suggestions received from the experts of the Spanish forensic laboratory of the National Police Force (*Dirección General de la Guardia Civil, DGGC*).

The present contribution also includes some preliminary results on the performance of the recognition module included in Biografo, based on the extraction of gradient-related features of individual characters. The evaluation has been carried out on a subset of a real forensic database comprising original confiscated/authenticated documents, which has been captured by trained operators using the acquisition tool integrated in Biografo.

The rest of the paper is structured as follows. The general tool is introduced in Sect. 2. Each of the two specific modules comprised within Biografo are described in Sects. 3 and 4 respectively. The preliminary performance results are presented in Sect. 5. Finally conclusions are drawn in Sect. 6.

## 2 Biografo

Biografo is a forensic tool formed by a client application programmed in JavaScript running over a relational database implemented in the platform MySQL Server 5.5. As shown in Fig. 1, Biografo presents two different operating modes: $i$) *local*, in which both the client application and the database run on the same machine, and $ii$) *remote*, in which several copies of the client application installed in different local machines communicate with one single copy of the database installed in a remote server. In the second case different forensic experts (that may be located at any point in the globe) can use the tool at the same time (e.g., launching queries or introducing new data) without compromising the consistency of the data.

Biografo has two main functionalities which are intended to help the forensic experts over the whole examination process, from the acquisition of the data to the identification of the individuals: $i$) acquisition of handwritten characters of individuals from

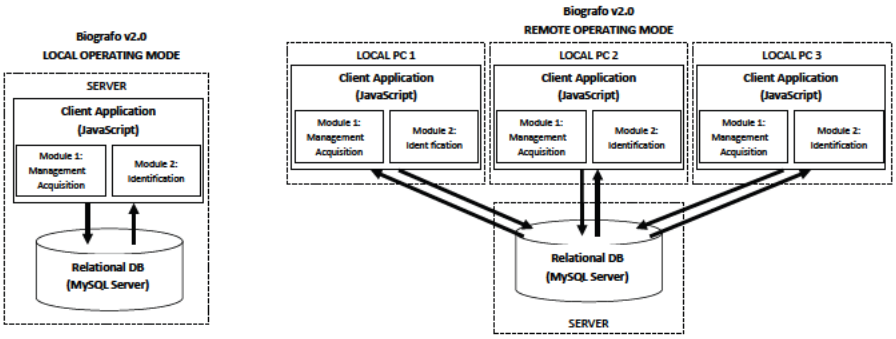


**Fig. 1.** General diagram of Biografo local (left) and remote (right) operating modes.

control or questioned documents, and management within a relational database of the acquired data; $ii$) run automatic identifications of individuals based on the samples acquired and handled within the database. Each of these functionalities is implemented in two separate modules that form the core of the client application as shown in Fig. 1.

Before describing in the next sections the specific functionalities of both modules, and in order to better understand the design of Biografo, it is important to clarify at this point that the identification module of Biografo works at the character level, that is: identification is performed comparing the samples of each character of the questioned individual to those of the known individuals, and performing a majority voting. The questioned individual will be identified with the known individual which has given a highest similarity score for the most number of characters.

Therefore, it is important to understand the difference that will be made throughout the document between: *character*, referring to each of the elements in the occidental written alphabet (i.e., we will consider 62 characters corresponding to the uppercase letters "A-Z", lowercase letters "a-z", and the ten digits "0-9"); *sample*, referring to each of the particular executions of a character carried out by a writer.

In general, for each individual several samples of one same character will be acquired. Each of these samples will have been captured from a digitalized document which, in turn, will be associated with a given individual. This link individual-document-sample is the basis of the tool operation.

In this scenario, a typical use-case for Biografo would be as follows (see Fig. 2):

- A. Data acquisition from documents of *known origin*. We will assume that a certain forensics laboratory has a number of handwritten documents coming from $N$ known individuals. These documents are digitalized and, using the acquisition and management module in Biografo, several samples for each character are acquired and stored in the relational database.
- B. Data acquisition from *questioned* documents. Now lets assume that a handwritten document found in a crime scene is sent to the forensics laboratory for identification. Again, the experts will digitalize the document and store the acquired samples into the database.

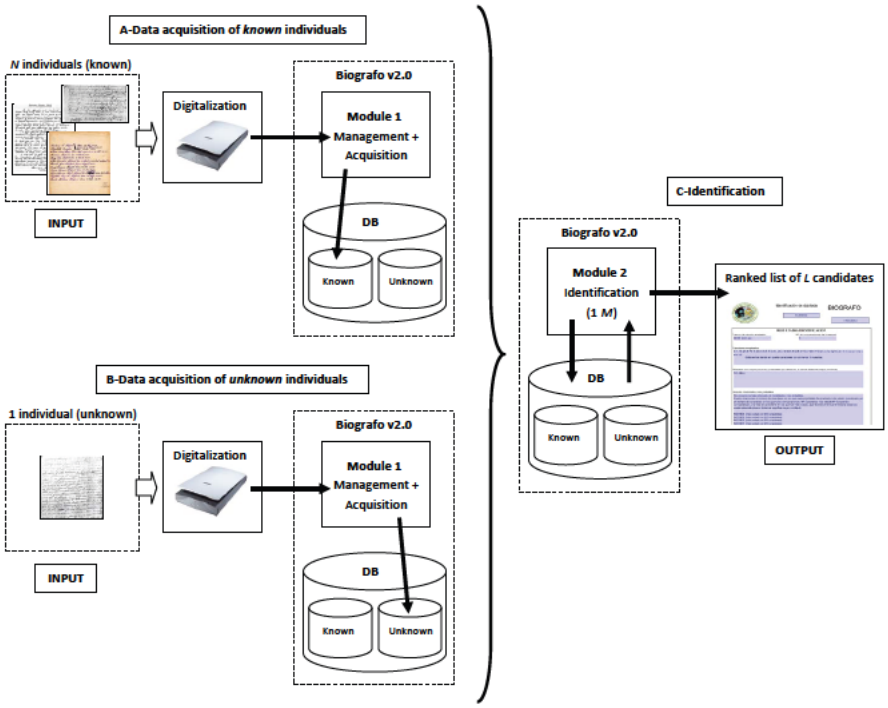


**Fig. 2.** General diagram of a typical use-case for Biografo.

– C. Identification. The forensic examiners will run the identification module in Biografo comparing the data corresponding to the questioned document to that of $M$ known subjects, being $M$ a subset of $N$, i.e., $M \leq N$. The identification module will give as output a ranked list of the $L$ most probable candidates, where $L \leq M$.

## 3 Module 1: Management and Acquisition

This module of the client application is responsible for the acquisition of the data (samples) from documents written by known or unknown individuals, and managing all these data maintaining at all times the consistency individual-document-sample. For this purpose the tool has implemented three main menus labeled as Individuals, Documents, and Samples (*Individuos*, *Documentos* and *Muestras* respectively in Spanish). Each of these menus presents the next functionalities:

**Individuals**. Different screenshots from this menu are shown in Fig. 3.

– Register new known or unknown individuals in the database (see Fig. 3, left). A unique alphanumeric identifier is assigned to each subject and we may also include other meta data associated to the individual such as name, id-card, date of birth, general comments etc.

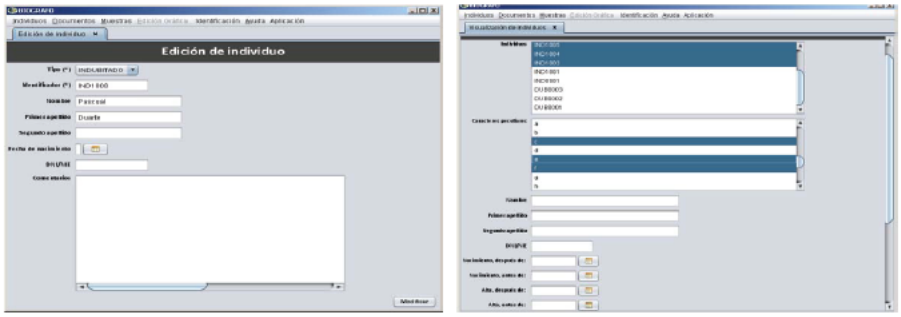


**Fig. 3.** Screenshots from the individuals menu in Biografo corresponding to the options: registration (left) and search (right) of an individual.

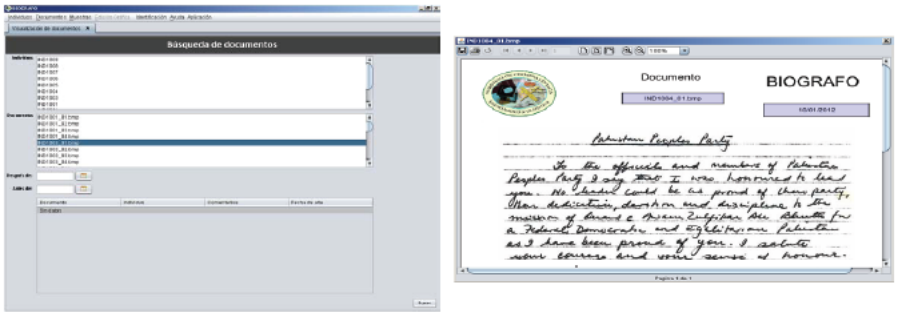


**Fig. 4.** Screenshots from the document menu in Biografo corresponding to the options: search (left) and visualization (right) of a document.

- Remove an individual from the database.
- Search for a certain individual (see Fig. 3, right). The tool permits to launch queries attending to different parameters such as the date in which the subject was incorporated to the database, name, surname, date of birth, those who write a certain character in a very particular manner, etc.
- Retrieve and print information of a given individual. Once an individual has been found with the search option, the tool can generate a document with all the data comprised in the database related to that subject: number of acquired samples of each character, number of handwritten documents, meta data, etc.

**Documents**. Different screenshots from this menu are shown in Fig. 4.

- The main purpose of this menu is to import a previously digitalized document into the database and assign it to a given individual. Documents accepted by Biografo are greyscale images in bmp format.

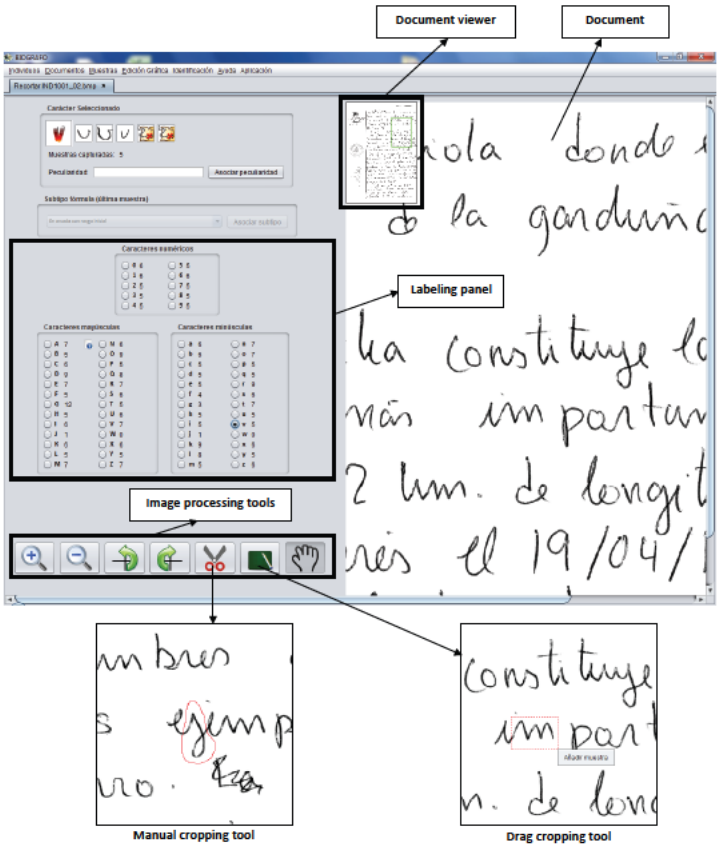


**Fig. 5.** Screenshots from the samples menu in Biografo corresponding to the acquisition options.

– This menu also has the option to search for documents within the database for their visualization (see Fig. 4 right). The search can be performed in terms of: the date in which the document was imported to the database, the individual to whom they are assigned, or directly with the name of the document (see Fig. 4 left).

**Samples**. Different screenshots from this menu are shown in Fig. 5.

– The main functionality of this menu is to manually acquire the samples of the different 62 handwritten characters (i.e., uppercase and lowercase letters and the ten digits) of a certain individual. These samples may be captured from any of the documents associated to that subject. Prior to the acquisition of a given sample, the operator selects on the labeling panel (see Fig. 5) the character to which it has to be assigned. On this panel the expert can see the number of samples already captured from each of the characters. On top of the labeling panel the last 5 acquired samples of the selected character are shown.

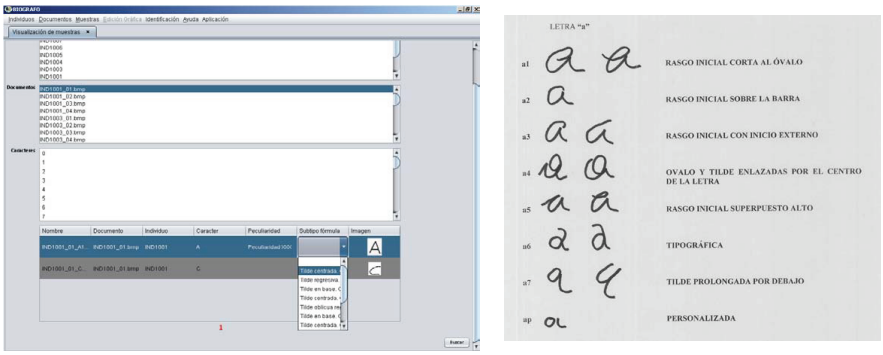


**Fig. 6.** Left, sample search engine implemented within Biografo. Right, different types of the character 'a' that may be selected by the operator to classify the different samples.

- Graphical processing tools. To assist the forensic expert in the acquisition process, Biografo has implemented different graphical tools such as: a hand tool to move the document, zoom in and out, a document viewer with your current location, rotation tools, drag cropping tool (i.e., acquires what is inside a rectangle), manual cropping tool (i.e., acquires what is inside a contour drawn with the mouse). Screenshots of both cropping tools are shown in Fig. 5 (bottom).
- Sample characterization options. Biografo gives the option to assign the samples not just to a given character (e.g., *a*) but also to a specific type within that character (e.g., *calligraphic a* or *typographic a*) following the classification used by the Spanish Forensic Laboratory (see the right panel in Fig. 6 for the complete classification of the handwritten character *a*). In addition, Biografo also permits to identify those characters that are executed in a very particular way and that can be very discriminative of a certain individual. This way each sample may be perfectly characterized so that the identification process can later be performed in a more precise manner (for instance using samples corresponding only to a certain type).
- This menu also offers the option to search for samples within the database (see Fig. 6 left). The search can be performed in terms of: the individual to whom they are assigned, the documents from which they were acquired, and the character they represent. Once the samples are retrieved from the database, the tool gives the possibility to visualize them, print them, or update them (e.g., changing the character they represent in case of an acquisition error.)

## 4 Module 2: Identification

This module of the client application is responsible for the identification of unknown individuals within the database. Biografo permits to fix a number of parameters before running the identification in order to restrict the search options or to discard *a priori*

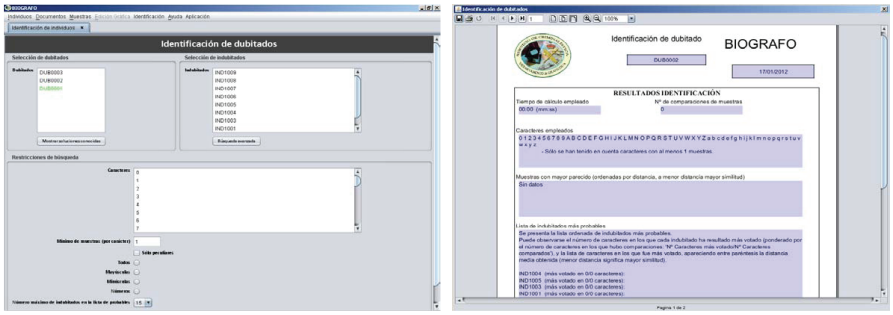


**Fig. 7.** Left, identification module implemented within Biografo. Right, output document with a summary of the results obtained in an identification test.

unfeasible candidates: $i$) unknown individual that we want to identify; $ii$) subset of $M$ known individuals (from the total $N$ available in the database) among which we want to run the identification (e.g., all of them, only those registered in the database prior/after a certain date); $iii$) characters that we want to use in the identification (e.g., all of them, only the lowercase letters, only the uppercase letters); $iv$) number of ranked candidates $L$ that we want to obtain in the output list.

Once the parameters above mentioned have been selected, the identification of writers is performed at the character level. Lets assume that we are using for identification the character subset composed of the 26 lowercase letters. All the samples of each of the 26 characters of the unknown individual are compared according to a certain matching function (described below) with all the samples of each of the 26 characters of the $M$ known individuals selected for the search. The closest identity for each character is computed based on the majority rule: the winning identity for a certain character will be the writer having the maximum number of winning samples (i.e., highest similarity score given by the matching function). In case of writers having the same number of winning samples, they are ranked according to the average of the winning scores. Finally, identification is based again on the majority rule, applied in this case to the characters: the winning output identity will be the writer having the maximum number of winning characters. In case of writers having the same number of winning characters, the same above criterion is applied.

A screenshot of the identification tool with the different parameter options to be selected is shown in Fig. 7 left, while on the right appears a document given as output by Biografo with a summary of the results of an identification test.

The current matching function implemented in Biografo is based on gradient-related features [2]. After the manual segmentation and labeling of the samples from a given document, they are binarized using the Otsu algorithm [9], followed by a margin drop and a height normalization to 120 pixels, preserving the aspect ratio. Elimination of noise of the binary image is then carried out through a morphological opening plus a closing operation [10]. After these preprocessing steps the feature vectors are computed as follows (see Fig. 8):

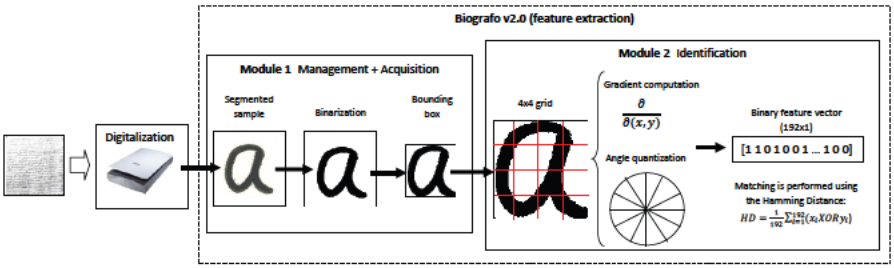


**Fig. 8.** Diagram of the feature extraction process followed by Biografo.

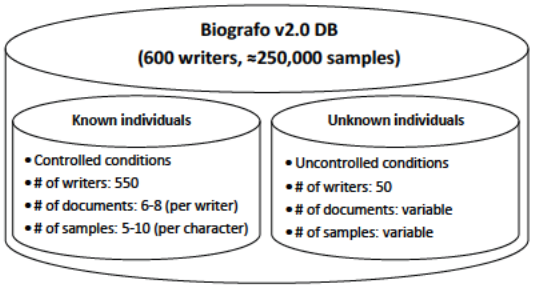


**Fig. 9.** Diagram showing the distribution of the Biografo DB.

- The processed samples are divided into a grid of $4 \times 4$ cells.
- The gradient is computed in each cell using $3 \times 3$ vertical and horizontal Sobel filters [10]. The direction of the gradient vector is quantized to 12 values (i.e., multiples of $\pi/6$).
- A histogram is computed for each cell, showing the number of times a certain direction appears in that given cell.
- All the 16 histograms are binarized according to the Otsu algorithm [9] so that for each cell we have a binary vector of length 12 showing if each of the possible directions are present (1) or not (0) in that cell.
- The final feature vector representing the sample is the binary vector of length $12 \times 4 \times 4 = 192$ that results from concatenating the individual binary vectors from each of the 16 cells.

The similarity score between two binary feature vectors is finally computed according to the Hamming distance.

## 5 Performance Evaluation

In order to evaluate the recognition performance of the identification module, a real forensic database from original confiscated/authenticated documents provided by the

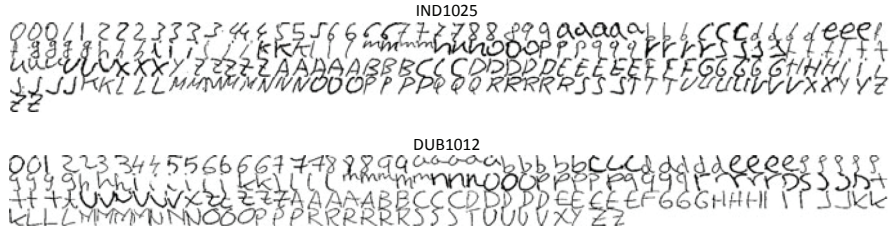


**Fig. 10.** Samples of a known (IND1025) and unknown (DUB1012) writer in Biografo DB.

Spanish forensic laboratory of the Dirección General de la Guardia Civil (DGGC) was captured using the acquisition and management module of Biografo. Samples of the handwritten alphanumeric characters were segmented and labeled by trained operators of the DGGC. The database contains a set of 550 known individuals and a set of 50 unknown writers (see Fig. 9):

- Set of *known* individuals. The documents in this set were written under controlled conditions (type of paper, pen, writing position, etc.) in the police premises after the criminal had been arrested. This way the available data is very large and very consistent for all the writers in this set, with 6-8 documents per subject, and with 5-10 acquired samples per character (except those that are rare in Spanish such as the 'w').
- Set of *unknown* individuals. The documents in this set were retrieved from crime scenes. This way the amount of data in this set is considerably smaller than in the case of the known individuals. Moreover, the variability of the available data among the writers is very big, in terms of amount of samples (some of them do not have samples of all the characters) and in terms of writing conditions (pen or pencil, type of paper, writing direction, etc.)

In Fig. 10 the samples of a known (top) and unknown (bottom) writer in the database are shown.

For the evaluation experiments, 30 out of the 50 unknown writers were manually identified with one of the known subjects in the database by forensic examiners from the DGGC. This correspondence between known and unknown individuals constitutes the ground truth for the performance evaluation of Biografo.

Given a writer of the unknown set, identification experiments are carried out by outputting the $L$ closest identities of the known set. An identification is considered successful if the correct identity is among the $L$ outputted ones. For this preliminary experiments only a subset of $N = 30$ writers from the known set was used (corresponding to those manually identified with the unknown individuals). Identification was performed using: $i$) all the available characters, $ii$) only the lowercase letters, $iii$) only the uppercase letters, $iv$) only the digits. Results are shown in the form of Cumulative Match Curves (CMC) in Fig. 11, from which two main observations may be extracted:

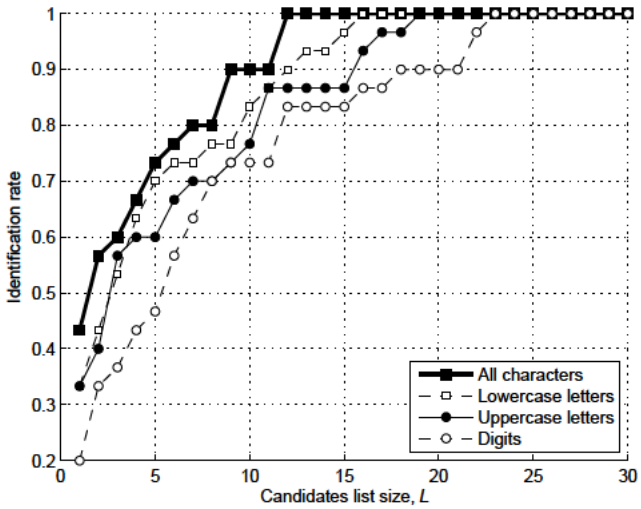


**Fig. 11.** Performance of the identification module in Biografo using: $i$) all the 62 characters, $ii$) only the lowercase letters, $ii$) only the uppercase letters, and $iv$) only the digits.

- Given the challenging nature of the database on which the evaluation was performed, specially for the subset of unknown individuals (with very few available data in some cases), the results show the high potential of the identification module to help forensic examiners narrowing down the search list of potential candidates (100% accuracy for a top 12 list in the best case).
- The multiple options offered by Biografo to define the identification tests have shown that, as could be expected, the lowercase letters are the most discriminant characters, followed by the uppercase letters (which usually have less variability among writers), and the digits. Moreover, it is proven that the best option is to use samples of all the possible characters (lowercase, uppercase and digits), as this increases the chances of a positive recognition.

The design of Biografo also permits to add in the future new recognition approaches to be fused with the current gradient-related matcher in order to further improve the identification rates.

## 6 Conclusions

We have presented the new integrated tool for forensic writer identification Biografo. This software application runs over a relational database and is designed to assist the forensic experts in the whole examination process from the data acquisition to the identification of writers. Preliminary performance results of the tool have also been presented on a real-case database of original forensic documents.

## Acknowledgements

This work has been partially supported by the Spanish *Dirección General de la Guardia Civil*, and projects Contexts (S2009/TIC-1485) from CAM, Bio-Challenge (TEC2009-11186) from Spanish MICINN, BBfor2 (ITN-2008-238803) from the European Commision, and *Cátedra UAM-Telefónica*.

## References

1. Srihari, S., Huang, C., Srinivasan, H., Shah, V.: Biometric and Forensic Aspects of Digital Document Processing. In: Digital Document Processing. Springer (2007) 379–406
2. Srihari, S., Cha, S., Arora, H., Lee, S.: Individuality of handwriting. Journal of Forensic Sciences **47** (2002) 856–872
3. Plamondon, R., Srihari, S.: On-line and off-line handwriting recognition: A comprehensive survey. IEEE Trans. on Pattern Analysis and Machine Intelligence **22** (2000) 63–84
4. Srihari, S., Leedham, G.: A survey of computer methods in forensic document examination. In: Proc. Int. Graphonomics Society Conference (IGS). (2003) 278–281
5. Schomaker, L.: Writer identification and verification. In: Advances in Biometrics: Sensors, Systems and Algorithms. Springer (2007) 247–264
6. Srihari, S., Ganesh, A., Tomai, C.: Information retrieval system for handwritten documents. In: Proc. Int. Workshop on Document Analysis Systems (DAS). (2004) 298–309
7. Franke, K., Schomaker, L., Veenhuis, C., Taubenheim, C., Guyon, I., Vuurpijl, L., Erp, M., Zwarts, G.: WANDA: a generic framework applied in forensic handwriting analysis and writer identification. In: Proc. Int. Conf. on Hybrid Intelligent Systems (HIS). (2003)
8. Tapiador, M.: Anlisis de las Caractersticas de Identificacin Biomtrica de la Escritura Manuscrita y Mecanogrfica. PhD thesis, Escuela Politcnica Superior, Universidad Autnoma de Madrid (2006)
9. Otsu, N.: A threshold selection method for gray-level histograms. IEEE Trans. on Systems, Man and Cybernetics **9** (1979) 62–66
10. Gonzalez, R., Woods, R.: Digital Image Processing. Addison-Wesley (2002)