Exploiting Character Class Information in Forensic Writer Identification

Fernando Alonso-Fernandez, Julian Fierrez, Javier Galbally, Javier Ortega-Garcia

Biometric Recognition Group - ATVS, Escuela Politecnica Superior
Universidad Autonoma de Madrid, Avda. Francisco Tomas y Valiente, 11
Campus de Cantoblanco, 28049 Madrid, Spain
fernando.alonso, julian.fierrez, javier.galbally, javier.ortega@uam.es
http://atvs.ii.uam.es

Abstract. Questioned document examination is extensively used by forensic specialists for criminal identification. This paper presents a writer recognition system based on contour features operating in identification mode (one-to-many) and working at the level of isolated characters. Individual characters of a writer are manually segmented and labeled by an expert as pertaining to one of 62 alphanumeric classes (10 numbers and 52 letters, including lowercase and uppercase letters), being the particular setup used by the forensic laboratory participating in this work. Three different scenarios for identity modeling are proposed, making use to a different degree of the class information provided by the alphanumeric samples. Results obtained on a database of 30 writers from real forensic documents show that the character class information given by the manual analysis provides a valuable source of improvement, justifying the significant amount of time spent in manual segmentation and labeling by the forensic specialist.

1 Introduction

Analysis of handwritten documents with the aim of determining the writer 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 as to the known and the questioned document, 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 present some difficulties due to their 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 into a database of known identities, therefore making easier the subsequent confrontation for the forensic expert [5, 4].

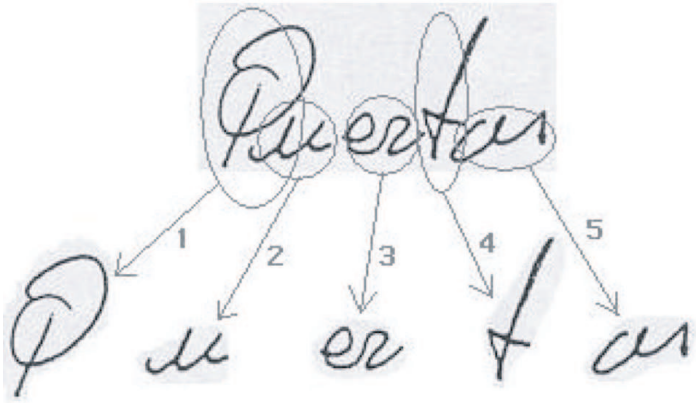

Fig. 1. Connected components from a handwritten sample.

In the last years, several writer recognition algorithms have been described in literature based on different group of features [7]: i) general geometric script features, like word or line spacing; ii) textural features capturing for example slant and curvature of the script; iii) placement features, i.e. writers placement preferences in the process of placing ink elements across the page; iv) micro level features measuring ink deposition characteristics; and v) character-fragment features measuring writer' preferred use of allographic elements.

A machine expert for off-line writer recognition making use of textural features based on contour information has been built in this work. It is focused on discriminating writers by capturing the distinctive visual appearance of the samples. Previous works following this direction used connected-component images or contours [8, 9] using automatic segmentation. Perfect automatic segmentation of individual characters still remains an unsolved problem [5], but connected components encompassing several characters or syllables can be easily segmented, and the elements generated (see Figure 1) also capture shape details of the visual appearance of the samples used by the writer [9]. The system in this paper, however, makes use of individual characters segmented manually by a forensic expert or a trained operator which are also assigned to one of the 62 alphanumeric classes among digits "0"∼"9", lowercase letters "a"∼"z", and uppercase letters "A"∼"Z". This is the setup used by the Spanish forensic group participating in this work. For a particular individual, the authenticated document is scanned and next, a dedicated software tool for character segmentation is used. Segmentation is done manually by a trained operator, who draw a character selection with the computer mouse and label the corresponding sample according to the 62 classes mentioned. We depict in Figure 2 (right) some examples of the manual selection of characters. In this work, we adapt the recognition method based on contour features from [9] to work with this setup. Additionally, the system is evaluated using a database created from real forensic documents (i.e. confiscated to real criminals or authenticated in the presence of a police officer), which is an important point compared with experiments of other works where the writing samples are obtained with the collaboration

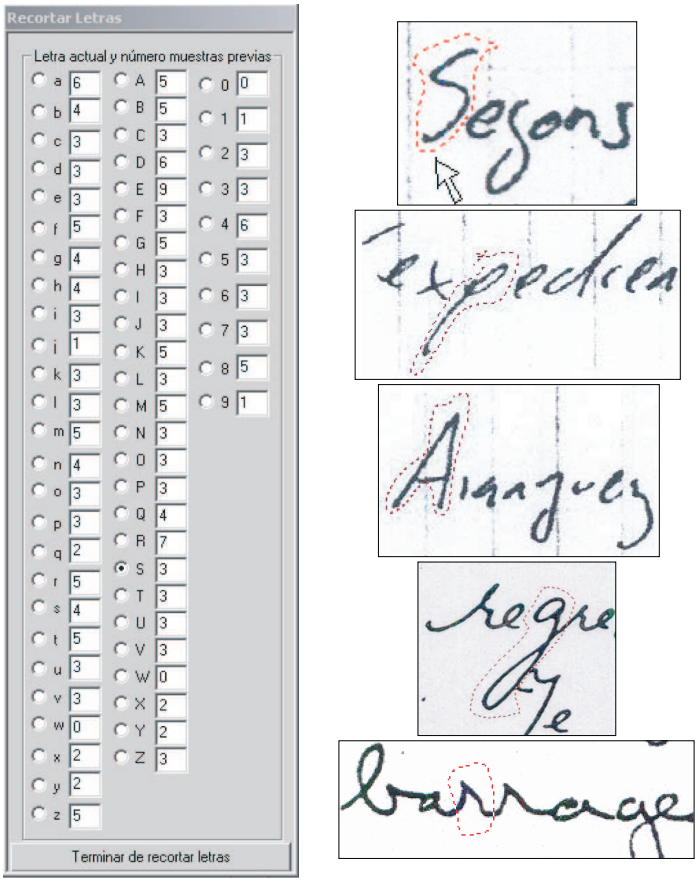

Fig. 2. Left: the 62 classes of alphanumeric characters used in this paper (digits "0"∼"9", lowercase letters "a"∼"z", and uppercase letters "A"∼"Z"). Right: manual selection of individual characters with the computer mouse using a dedicated software tool. Images from [6].

of volunteers under controlled conditions [10]. We evaluate in this paper three different scenarios for identity modeling, exploiting to a different degree the class information provided by the manual segmentation of alphanumeric samples: modeling *per individual sample*, modeling *per alphanumeric channel*, and modeling *per writer*. Results show that the class information provides a considerable improvement, justifying the writer identification approach used in our forensic system, where a significant amount of time is spent every time a new writer is included.

The system is evaluated in identification mode, in which an individual is recognized by searching the reference models of all the subjects in the database for a match (one-to-many). As a result, the system returns a ranked list of candidates. Ideally, the first ranked candidate (Top 1) should correspond with the correct identity of the individual, but one can choose to consider a longer list (e.g. Top 10) to increase the chances of

	Feature	Explanation	Dimensions	Source
f1	$p(\phi)$	Contour-direction PDF	12	contours
f2	$p(\phi_1, \phi_2)$	Contour-hinge PDF	300	contours
f3h	$p(\phi_1, \phi_3)_h$	Direction co-occurrence PDF, horizontal run	144	contours
f3v	$p(\phi_1, \phi_3)_v$	Direction co-occurrence PDF, vertical run	144	contours
f5h	$p(rl)_h$	Run-length on background PDF, horizontal run	60	binary image
f5v	$p(rl)_v$	Run-length on background PDF, vertical run	60	binary image

Table 1. Features used in this work.

finding the correct identity. Identification is a critical component in negative recognition applications (or watchlists) where the aim is checking if the person is who he/she (implicitly or explicitly) denies to be, which the typical situation in forensic/criminal cases [11].

The rest of the paper is structured as follows. In Section 2 we describe the main stages of our recognition system. Section 3 describes the database, the scenarios for identity modeling and the experimental results. Finally, conclusions are drawn in Section 4.

2 System Description

The writer recognition system of this paper makes use of the contour features presented in [9], which are adapted to the particular setup of this paper. It includes three main stages: i) preprocessing of the individual characters, ii) feature extraction, and iii) feature matching. These stages are described next.

2.1 Pre-processing Stage

The writer identification method used by the forensic group participating in this work is based on manually reviewing the handwritten material, as mentioned in Section 1. After manual segmentation and labeling of alphanumeric characters from a given document, they are binarized using the Otsu algorithm [12], 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 [13]. Next, a connected component detection, using 8-connectivity, is done. In the last step, internal and external contours of the connected components are extracted using the Moore's algorithm [13]. Beginning from a contour pixel of a connected component, which is set as the starting pixel, this algorithm seeks a pixel boundary around it following the meaning clockwise, and repeats this process until the starting pixel is reached for the same position from which it was agreed to begin the algorithm. The result is a sequence with the pixels coordinates of the boundary of the component. This vectorial representation is very effective because it allows a rapid extraction of many of the features used later.

2.2 Feature Extraction Stage

Features are calculated from two representations of the handwritten samples extracted during the preprocessing stage: the binary image without noise and the contours of the connected components. The features used in this work are summarized in Table 1, including the image representation used by each one. A handwritten sample is shaped like a texture that is described with probability distribution functions (PDFs). Probability distribution functions used here are grouped in two different categories: direction PDFs (features f1, f2, f3h, f3v) and length PDFs (features f5h, f5v). A graphical description of the extraction of these features is depicted in Figure 3. To be consistent with the work in which these features where proposed [9], we follow the same nomenclature used in it.

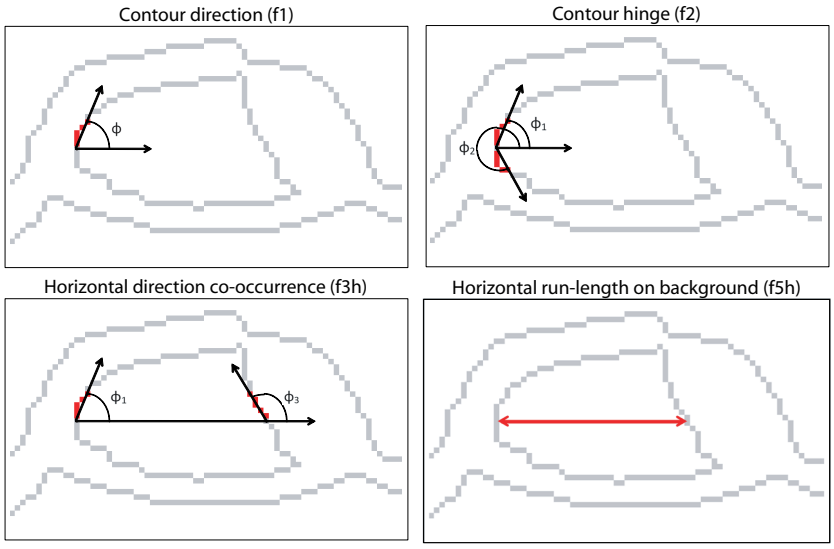

Fig. 3. Graphical description of the feature extraction: contour direction (f1), contour hinge (f2), horizontal direction co-occurrence (f3h) and horizontal run-length (f5h).

Contour-Direction PDF (f1)

This directional distribution is computed very fastly using the contour representation, with the additional advantage that the influence of the ink-trace width is eliminated. The contour-direction distribution f1 is extracted by considering the orientation of local contour fragments. A fragment is determined by two contour pixels (x_k,y_k) and ($x_{k+\epsilon}$,$y_{k+\epsilon}$) taken a certain distance ϵ apart. The angle that the fragment makes with the horizontal is computed using

$$\phi = \arctan\left(\frac{y_{k+\epsilon} - y_k}{x_{k+\epsilon} - x_k}\right) \tag{1}$$

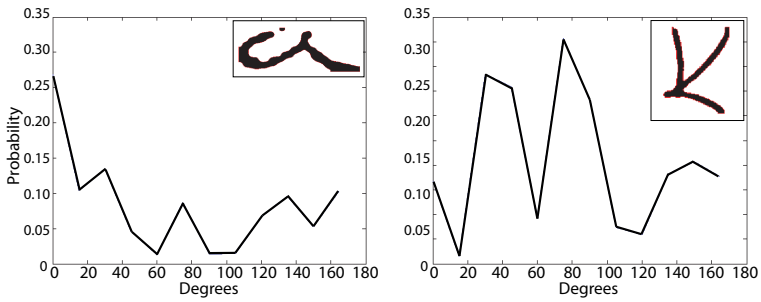

Fig. 4. Example of the countour direction feature (f1) for two different handwritten characters.

As the algorithm runs over the contour, the histogram of angles is built. This angle histogram is then normalized to a probability distribution f1 which gives the probability of finding in the handwritten sample a contour fragment oriented with each ϕ. The angle ϕ resides in the first two quadrants because, without online information, we do not know which inclination the writer signed with. The histogram is spanned in the interval 0°-180°, and is divided in $n = 12$ sections (bins). Therefore, each section spans 15°, which is a sufficiently detailed and robust description [9]. The parameter ϵ controls the length of the analyzing contour fragment, which is set to $\epsilon = 5$. These settings will be used for all of the directional features presented in this paper. An example of extraction of this feature for two handwritten characters is depicted in Figure 4.

Contour-Hinge PDF (f2)

In order to capture the curvature of the contour, as well as its orientation, the "hinge" feature f2 is used. The main idea is to consider two contour fragments attached at a common end pixel and compute the joint probability distribution of the orientations ϕ_1 and ϕ_2 of the two sides. A joint density function is obtained, which quantifies the chance of finding two "hinged" contour fragments with angles ϕ_1 and ϕ_2, respectively. It is spanned in the four quadrants (360°) and there are $2n$ sections for every side of the "contour-hinge", but only non-redundant combinations are considered (i.e. $\phi_2 \geq \phi_1$). For $n = 12$, the resulting contour-hinge feature vector has 300 dimensions [9].

Direction Co-Occurrence PDFs (f3h, f3v)

Based on the same idea of combining oriented contour fragments, the directional co-occurrence is used. For this feature, the combination of contour-angles occurring at the ends of run-lengths on the background are used, see Figure 3. Horizontal runs along the rows of the image generate f3h and vertical runs along the columns generate f3v. They are also joint density functions, spanned in the two first quadrants, and divided into n^2 sections. These features give a measure of a roundness of the written characters and/or strokes.

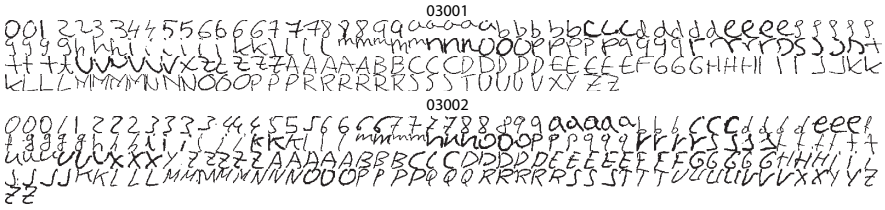

Fig. 5. Training samples of two different writers of the forensic database used in this paper.

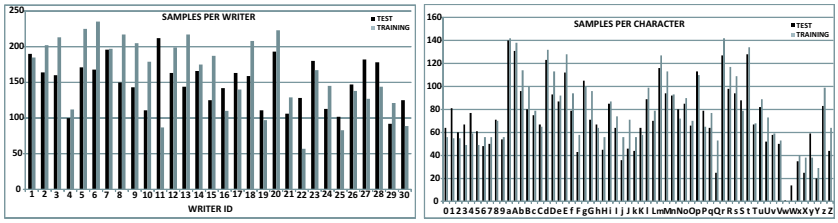

Fig. 6. Distribution of samples per writer (left) and per character (right) of the forensic database used in this paper.

Run-Length PDFs (f5h, f5v)

These features are computed from the binary image of the handwritten sample taking into consideration the pixels corresponding to the background. They capture the regions enclosed inside the letters and strokes and also the empty spaces between them. The probability distributions of horizontal and vertical lengths are used. These features gives the probability of finding in the handwritten sample an enclosed region with each length.

2.3 Feature Matching Stage

Each writer is represented in the system by a PDF or set of PDFs (depending on the experiment at hand, see 3). To compute the similarity between two PDFs $\mathbf{o}$ and $\boldsymbol{\mu}$ from two different writers, the χ^2 distance is used:

$$\chi^2_{\mathbf{o}\boldsymbol{\mu}} = \sum_{i=1}^{N} \frac{(o_i - \mu_i)^2}{o_i + \mu_i} \tag{2}$$

where N is the dimensionality of the vectors $\mathbf{o}$ and $\boldsymbol{\mu}$.

3 Experimental Framework

3.1 Database

To evaluate the system, we use a real forensic database from original confiscated/authenticated documents provided by the Spanish forensic laboratory of the Dirección General de la Guardia Civil (DGGC). Alphanumeric characters of the handwritten samples are segmented and labeled by a trained operator of the DGGC. The whole database contains 9,297 character samples of real forensic cases from 30 different writers, with around 300 samples on average per writer distributed between a training and a testing data set. In Figure 5 we plot the training samples of two different writers of the database. For each writer, training and testing data are extracted from different confiscated documents, meaning that they were "acquired" at different moments. Given the nature of the database, it does not contain uniformly distributed samples of every character, nor time span between training and testing data. Figure 6 shows the distribution of samples per writer and per character of our database.

3.2 Writer Identity Modeling

Given a writer of the test set, *identification experiments* are done by outputting the N closest identities of the training set. An identification is considered successful if the correct identity is among the N outputted ones.

For a particular writer, several samples of individual characters pertaining to one of the 62 alphanumeric classes among digits "0"∼"9", lowercase letters "a"∼"z", and uppercase letters "A"∼"Z" are available thanks to the manual segmentation and labeling. For each feature described in Section 2.2, we evaluate the following three scenarios for writer identity modeling:

1. Modeling *per individual sample* (channel dependent). For example, if a writer has x samples of the digit "0", features for each of the x samples are computed. This process is repeated with all the 62 alphanumeric channels. This modeling captures particular variations in each alphanumeric character (e.g. if the writer has different "a", "b", etc.) Due to the nature of the database, it will not be a uniform number of features among the different channels. It could also be the case that a writer many not have samples in a particular channel, in whose case no features will be extracted. For each individual sample, we find the closest identity by comparing with all the training samples pertaining to the same channel. We then compute the closest identity to each alphanumeric channel based on the majority rule: the winning identity will be the writer having the maximum number of winning samples. In case of writers having the same number of winning samples, they are subsequently ranked using the next 2 criteria, listed in descending order of weight: 1) average of winning sub-distances, and 2) minimum winning sub-distance. Finally, identification is based again on the majority rule, applied in this case to the alphanumeric channels: the winning output identity will be the writer having the maximum number of winning alphanumeric channels, the second winning identity will be the next writer, and so on. In case of writers having the same number of winning channels,

we apply the same above criteria.

2. Modeling *per alphanumeric channel* (channel dependent). For example, if a writer has x samples of the digit "0", histograms of the feature are combined (added) to obtain a unique probability distribution. This process is repeated for all the 62 alphanumeric classes. This modeling averages the different variations of a given alphanumeric character. Therefore, we obtain 62 sub-distances between two writers, one per channel. We then compute the closest identity to each alphanumeric channel based on its distance. Identification is based on the majority rule: the winning output identity will be the writer having the maximum number of winning alphanumeric channels, the second winning identity will be the next writer, and so on. This results in $62 \times 30 \times 30$=55,800 computed distances. In case of writers having the same number of winning channels, we apply the same above criteria.

3. Modeling *per writer* (channel independent). This case computes a unique probability distribution per writer by combining all the available samples of all the alphanumeric characters. In this case, we do not use the character class information, obtaining a unique writing identity model that averages information from the 62 channels. Only one distance between two writers is obtained, which is used for identification. This results in 30×30=900 computed distances.

3.3 Results

We plot in Figure 7 results of the identification experiments varying the size of the hit list from N=1 (Top 1) to N=30 (Top 30). Results are shown for the different features described in Section 2.2 and for the three identity modeling scenarios considered.

We observe that, in general, working with the class information provided by the alphanumeric channels (top and medium plot in Figure 7) results in considerable better performance with respect to using a unique single identity model that does not exploit this information (bottom plot in Figure 7). Thus, the class information given by the character segmentation and labeling carried out by the trained operator provides a considerable improvement. This justifies the writer identification approach used in our forensic system, in which a considerable amount of time is spent every time a new writer is included in the database. It can be seen in Figure 7, for example, that a success rate of 80% is already achieved with some features for a hit list size of N=5 or less when using channel information. However, when using a channel independent identity modeling, it is not achieved until a hit list size of N=13 is considered.

It is worth noting that directional features (f1, f2, f3h, f3v) work consistently better that features based on length properties (f5h, f5v). This suggests that the length of the regions enclosed inside the letters and strokes is not a good distinctive feature in the setup presented in this paper, where we are using a database of isolated alphanumeric handwritten characters. Better results are obtained in other studies making use of complete lines or pages of handwritten material [9].

Finally, by comparing the two scenarios for writer identity modeling that make use of channel information (top and medium plot in Figure 7), it can be seen that the best results are obtained when using identity models per alphanumeric channel. In this case,

for a hit list size of N=5, all the directional features achieve a success rate of $\tilde{8}$0%. On the other hand, when using identity models per individual sample and a hit list size of N=5, the success rate exhibited by the directional features are between 60% and 80%. Thus, averaging all the samples of a given channel provides more robustness than using the samples separately.

4 Conclusions and Future Work

A machine expert for off-line writer identification based on contour features has been evaluated. It encodes several directional properties of contour fragments as well as the length of the regions enclosed inside letters. The system presented in this work is based on manual review of the handwritten material, in which segmentation and labeling of characters is made using a dedicated software tool according to 62 alphanumeric classes (10 numbers and 52 letters, including lowercase and uppercase letters). This particular setup is used by the Spanish forensic group participating in this work, which has also provided us with a database of real forensic documents from 30 different writers, an important point in comparison with other works where data is obtained from collaborative writers under controlled conditions. Experiments are done in identification mode (one-to-many), which the typical situation in forensic/criminal cases.

The system of this paper is evaluated in three different scenarios for identity modeling which exploit to a different degree the class information provided by the manual segmentation of alphanumeric samples: i) modeling *per individual sample*, ii) modeling *per alphanumeric channel*, and iii) modeling *per writer*. The two first scenarios make use of the class information given by the manual labeling, whereas the third one is channel independent (i.e. does not use the character class information). Results show that much better performance is obtained by using channel information, justifying the considerable amount of time spent by the trained operator in the segmentation and labeling process. The best scenario is based on identity modeling *per alphanumeric channel*, meaning that averaging all the samples of a given channel provides more robustness than using the samples separately. The latter approach may work better if enough samples representative of writer' particular variations are included in the database, or for specific channels commonly used in the language of the database (as can be seen in Figure 6, characters like 'w' and 'W' are not often used in the Spanish language, while "a", "A", "r" or "t" are quite common).

A drawback found in our experiments is that a success rate of 100% is never achieved with some features and/or identity modeling scenarios. It means that there are some writers in the database whose identity are never found, and test samples from this writer are assigned as pertaining to someone else. It could be due to the majority rule used for identification, as well as the decision criteria when several writers have the same number of winning samples (see Section 3.2).

The analysis of these results with a limited database suggest that the proposed approach can be used for forensic writer identification, pointing out the advantages of manual segmentation and labeling by a trained operator. Future work includes evaluating of our system with a bigger forensic database and improving the performance by applying advanced alphanumeric channel combination methods [14]. Another source of

future work is the use of advanced approaches for user-dependent selection and combination of alphanumeric channels [15, 16], so that the most discriminative channels for each user are used in the fusion.

5 Acknowledgements

This work has been partially supported by projects Bio-Challenge (TEC2009-11186), BBfor2 (FP7 ITN-2009-238803) and "Cátedra UAM-Telefónica". Postdoctoral work of author F. A.-F. is supported by a Juan de la Cierva Fellowship from the Spanish MICINN. The authors would like to thank to the forensic "Laboratorio de Grafística" of the 'Dirección General de la Guardia Civil' for its valuable support.

References

1. Srihari, S., Huang, C., Srinivasan, H., Shah, V.: 17. Biometric and Forensic Aspects of Digital Document Processing. In: Digital Document Processing. Springer (2007) 379–406
2. Srihari, S.N., Cha, S.H., Arora, H., Lee, S.: Individuality of handwriting. Journal of Forensic Sciences **47**(4) (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**(1) (2000) 63–84
4. Srihari, S., Leedham, G.: A survey of computer methods in forensic document examination. Proc. 11th International Graphonomics Society Conference, IGS (November 2003) 278–281
5. Schomaker, L.: Writer identification and verification. In: Sensors, Systems and Algorithms, Advances in Biometrics. Springer Verlag (2008)
6. Tapiador, M.: Análisis de las Características de Identificación Biométrica de la Escritura Manuscrita y Mecanográfica. PhD thesis, Escuela Politécnica Superior, Universidad Autónoma de Madrid (2006)
7. Schomaker, L.: Advances in writer identification and verification. Proc. Intl. Conference on Document Analysis and Recognition, ICDAR **2** (2007) 1268–1273
8. Schomaker, L., Bulacu, M.: Automatic writer identification using connected-component contours and edge-based features of upper-case western script. IEEE Trans. on Pattern Analysis and Machine Intelligence **26**(6) (2004) 787–798
9. Bulacu, M., Schomaker, L.: Text-independent writer identification and verification using textural and allographic features. IEEE Trans. on Pattern Analysis and Machine Inteligence **29**(4) (April 2007) 701–717
10. Tapiador, M., Sigenza, J.: Writer identification method based on forensic knowledge. Proc. International Conference on Biometric Authentication, ICBA, **Springer LNCS-3072** (2004) 555–560
11. Jain, A., Flynn, P., Ross, A., eds.: Handbook of Biometrics. Springer (2008)
12. Otsu, N.: A threshold selection method for gray-level histograms. IEEE Trans. on Systems, Man and Cybernetics **9** (December 1979) 62–66
13. Gonzalez, R., Woods, R.: Digital Image Processing. Addison-Wesley (2002)
14. Jain, A., Nandakumar, K., Ross, A.: Score Normalization in Multimodal Biometric Systems. Pattern Recognition **38**(12) (December 2005) 2270–2285
15. Fierrez-Aguilar, J., Garcia-Romero, D., Ortega-Garcia, J., Gonzalez-Rodriguez, J.: Adapted user-dependent multimodal biometric authentication exploiting general information. Pattern Recognition Letters **26** (2005) 2628–2639
16. Galbally, J., Fierrez, J., Freire, M.R., Ortega-Garcia, J.: Feature selection based on genetic algorithms for on-line signature verification. Proc. IEEE Workshop on Automatic Identification Advanced Technologies, AutoID (2007) 198–203

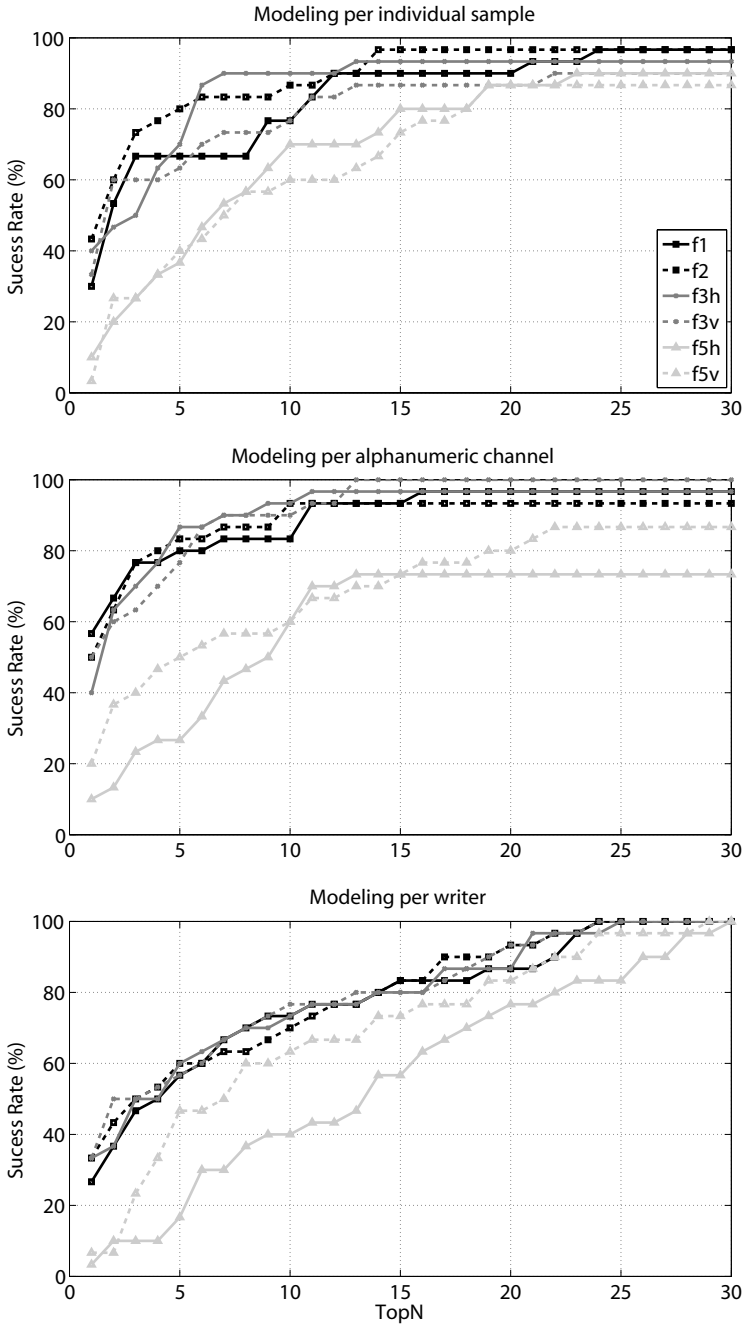

Fig. 7. Writer identification rates for the three scenarios of identity modeling considered.