Forensic Writer Identification Using Allographic Features

Ruben Fernandez-de-Sevilla, Fernando Alonso-Fernandez, Julian Fierrez, 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
ruben.fernandezdesevilla, fernando.alonso, julian.fierrez, javier.ortega@uam.es

Abstract

Questioned document examination is extensively used by forensic specialists for criminal identification. This paper presents a writer recognition system based on allographic features operating in identification mode (one-to-many). It works at the level of isolated characters, considering that each writer uses a reduced number of shapes for each one. 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. A codebook of shapes is then generated by clustering and the probability distribution function of allograph usage is the discriminative feature used for recognition. 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 proposed approach. We also evaluate the selection of different alphanumeric channels, showing a dependence between the size of the hit list and the number of channels needed for optimal performance.

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].

In the last years, several writer recognition algorithms have been described in literature based on different group of features [6]. This paper presents an offline writer recognition system making use of features at the allographic level, which focuses on discriminating writers encoding their preferred or more used allographic elements by capturing their occurrence probability. Previous works following this direction used connected-component images [7] 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 also capture shape details of the allographs used by the writer [10]. The system in this paper, however, makes use of individual characters segmented manually by a forensic expert 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 charac-

978-0-7695-4221-8/10 $26.00 © 2010 IEEE
DOI 10.1109/ICFHR.2010.54

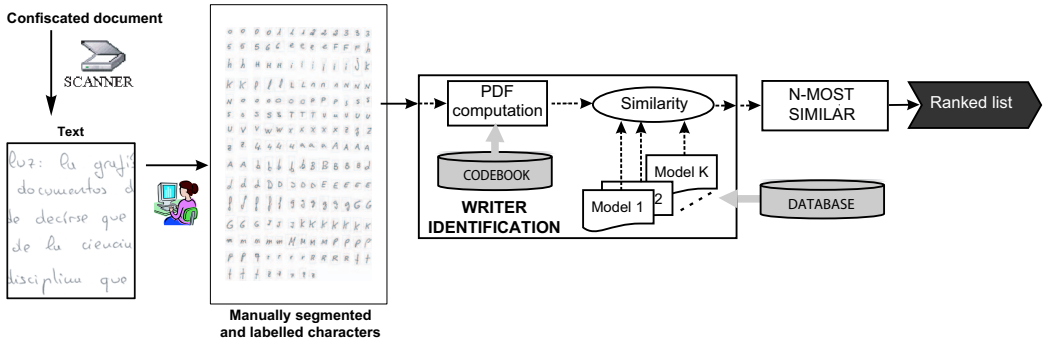

Figure 1. System model for forensic writer identification based on allographic features.

ter segmentation is used. Segmentation is done manually by a trained operator (a forensic expert), who draw a character selection with the computer mouse and label the corresponding sample according to the 62 classes mentioned. In this work, we adapt the off-line recognition method based on allographic features from [10] to work with this setup. Other similar approaches making use of on-line information are also found in the literature [11]. 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 of volunteers under controlled conditions [12].

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 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 [13].

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 and the experimental protocol used. Experimental results are presented in Section 4. Finally, conclusions are drawn in Section 5.

2. System Description

The writer recognition system used in this paper is an implementation of the system presented in [10], which is adapted to the particular setup of this paper. It considers the writer as a stochastic pattern generator of handwritten shapes. The probability distribution function (PDF) of these shapes in a given handwritten sample is used to characterize the writer, which is computed using a common codebook of shapes obtained by means of clustering techniques. This way, the codebook provides a common shape space and the PDF captures the individual shape usage preference of the writer. This writer identification system includes three main stages: i) handwriting preprocessing, ii) shape codebook generation, and iii) computation of the writer-specific PDF. In Figure 1, the overall model of the identification system used in this work is depicted.

Handwriting Preprocessing

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 [14], followed by a margin drop and a size normalization to 32×32 pixels, preserving the aspect ratio.

Codebook Generation

The objective of this stage is to generate a common codebook of shapes that we can observe on a handwriting sample, for which an external database of segmented alphanumeric characters is used (obtained from an independent set of writers not "participating" in the forensic material). For this purpose, we make use of the CEDAR database [15]. This database[1] contains digitized images of handwritten words and ZIP codes (300 dpi, 8-bit) and binary handwritten isolated digits and alphanumeric characters (300 dpi, 1-bit). Data were scanned from envelopes in a working post office

[1] Available for a fee at http://www.cedar.buffalo.edu/Databases

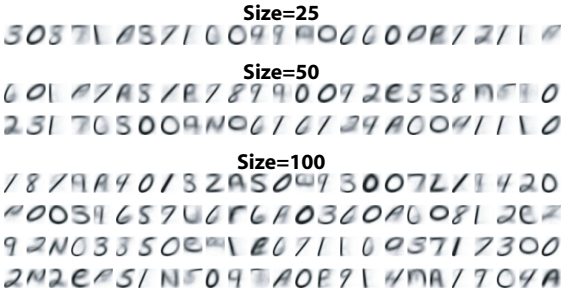

Figure 2. Global codebooks of different sizes.

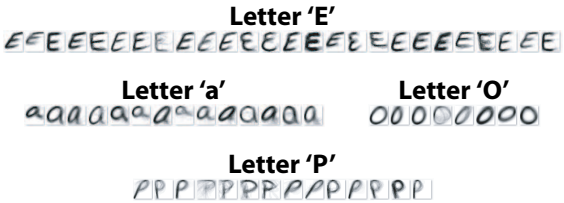

Figure 3. Example of optimal "sub-codebooks" for different characters.

at Buffalo in the US, therefore no constraints were imposed in the writer, style, pencil, etc. In this paper, we use the set of isolated digits and alphanumeric characters, which contains 27,837 mixed alphas and numerics segmented from address blocks and 21,179 digits segmented from ZIP Codes. Since the database was extracted from handwritten text on real postal letters, the distribution of samples is not uniform, having over 1000 samples of some characters, like '1', and less than 10 samples of another ones, like 'j'. For the experiments of this paper, we drop the margins of the binary images by calculating their bounding boxes, followed by a size normalization to 32×32 pixels, preserving the aspect ratio of the handwritten sample.

In this paper, we evaluate the following two scenarios for codebook generation:

1. A *global* codebook that does not use the character class information. We just use as input all the alphanumeric character images of the CEDAR database and generate a unique global codebook.

2. A *local* character-based codebook, composed of 62 "sub-codebooks", one per each character (10 numbers and 52 letters, including lowercase and uppercase letters). In this case, we exploit the class information given by the character segmentation and labeling carried out by the forensic expert.

Clustering is then applied to the CEDAR database in order to obtain the codebooks according to these scenarios. The clustering technique used is k-means [16] because of its simplicity and computational efficiency [17]. We generate codebooks with different sizes in order to obtain the optimal size for each scenario (i.e. yielding the best performance). The maximum size for each sub-codebook in the scenario 2 depends on the number of samples of the corresponding character in the CEDAR database. For example, characters like "q" or "j" allow only codebooks of size 2 or 3, while "0" or "A" allow codebooks of up to 500 clusters. Figure 2 depicts several global codebooks of different sizes obtained according to this protocol, whereas Figure 3 depicts some of the 62 optimal "sub-codebooks" obtained in the experiments of Section 4.

PDF Computation and Matching

In this stage, the main objective is to obtain the discriminative PDF of each writer describing his/her individual shape usage preference. It is computed by building an histogram in which one bin is allocated to every codebook sample. For each alphanumeric sample of a writer, we find the nearest codebook sample using Euclidean distance. Therefore, for each writer we obtain 1 histogram (in the case of global codebook) or 62 histograms (one per character, in the case of local sub-codebooks). Each histogram is finally normalized to a PDF, which will be the discriminative feature used for recognition. To compute the similarity between two PDFs $\mathbf{o}$ and $\boldsymbol{\mu}$ from two different writers, the χ^2 distance is used:

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

where N is the dimensionality of the vectors $\mathbf{o}$ and $\boldsymbol{\mu}$. When using a global codebook, only one distance is obtained. In the case of using 62 character-based sub-codebooks, 62 sub-distances between any two given writers are obtained, one per alphanumeric channel.

3. Database and Protocol

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). As described in Section 2, alphanumeric characters of the handwritten samples are segmented and labeled by a forensic expert of the DGGC. The whole database contains 9,297 character samples of real forensic cases from 30 different writers, with around 300 samples on

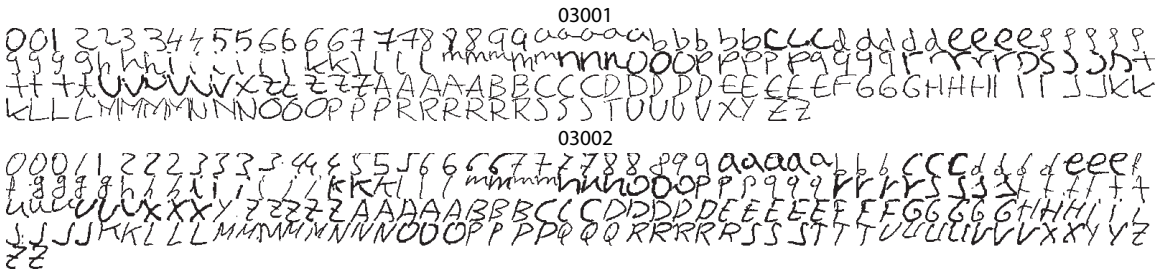

Figure 4. Training samples of two different writers of the forensic database used in this paper.

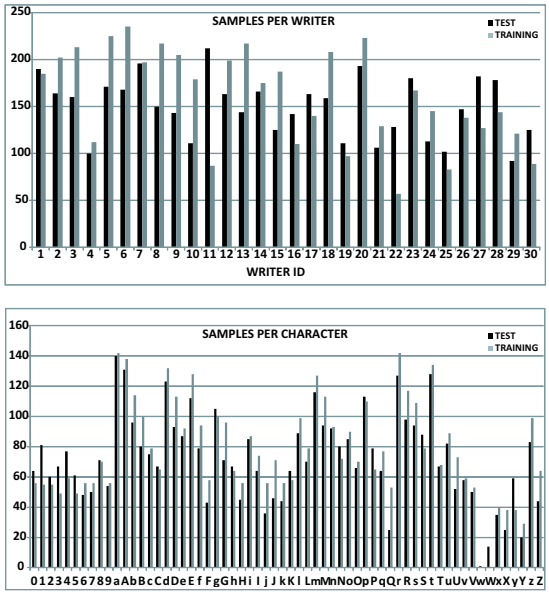

Figure 5. Distribution of samples per writer (top) and per character (bottom) of the forensic database used in this paper.

average per writer distributed between a training and a testing data set. In Figure 4 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. As in the CEDAR database, and given the nature of the database, it does not contain uniformly distributed samples of every character. Figure 5 shows the distribution of samples per writer and per character of our database.

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. When using a global codebook, only one distance between two writers is obtained, which is used for identification. This results in 30×30=900 computed distances. When using 62 sub-codebooks, we compute the closest identity to each alphanumeric character based on the sub-distance of each channel. A decision is then made 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 $62\times30\times30$=55,800 computed distances. In case of writers having the same number of winning channels, they are subsequently ranked using the next 4 criteria, listed in descending order of weight: 1) average of winning sub-distances, 2) minimum winning sub-distance, 3) average of the 62 sub-distances between the test and the training writers and 4) minimum of the 62 sub-distances between the test and the training writers.

4. Results

The first step is to obtain the optimal size of the codebooks. We plot in Figure 6 results of the identification experiments depending on the size of the global codebook for a hit list size of N=1 (Top 1). We observe that the identification rate oscillates for small codebook sizes and it tends to increase with codebooks of size above 400 clusters, reaching a plateau around a size of 750.

Similarly, we vary the size of each of the 62 sub-codebooks separately in the corresponding scenario of Section 2, obtaining the identification rates of each alphanumeric channel. The optimal size of each sub-codebook is set at that resulting in the highest identification rate for a hit list size of 1. In Figure 7 we plot the best identification rate obtained for each channel, together with the optimal size of each sub-codebook. We observe that the characters with best rates are "d", "r", "s" and "N". For some characters, like "j", "q", "Q", "w" and "W", the identification rates are null. As explained in Section 2, for characters "q", "Q" and "j" we

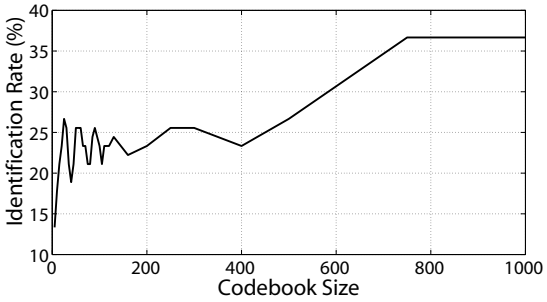

Figure 6. Writer identification rates depending on the size of the codebook (global codebook, hit list size=1).

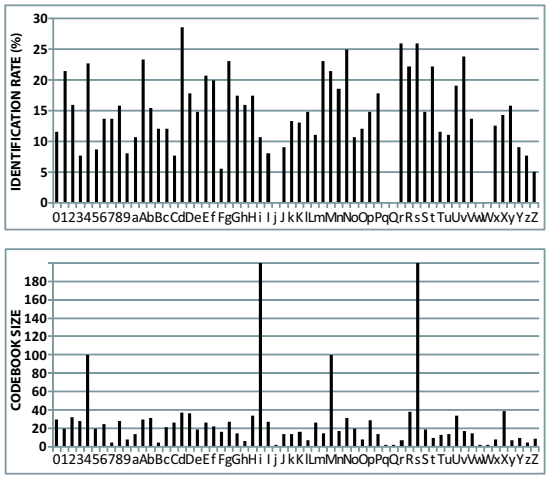

Figure 7. Best identification rates (top) and optimal size of the sub-codebook (bottom) for each individual alphanumeric channel (hit list size=1).

could only generate very small codebooks (up to 2 or 3 clusters) so their PDFs are not very discriminative. For the characters 'w' and 'W', codebooks of enough size can be generated, but in the forensic database there are not samples of them for most users, as long as this characters are not often used in the Spanish language (see Figure 5). We also observe in Figure 7 that for each character, we achieve the best identification rate with a codebook of different size. These optimal sizes are obtained for our forensic database in Spanish language, but it is expected that depending on the size and the language of the database, the size of the optimal sub-codebooks could vary.

Once we have obtained the optimal size of the codebook for each single channel, we evaluate the combination of the 62 alphanumeric channels. We plot in Figure 8 results of the identification experiments depending on the number of channels combined for a hit list size of N=1 (Top 1). Individual channels are ranked in descending order and selected according to its identification rate depicted top in Figure 7 (e.g. the channel with the highest identification rate, the two channels with the highest identification rates, etc.) We observe that the identification rate is increased with the number of channels, reaching its maximum at around 40 channels combined, and then it remains more or less constant.

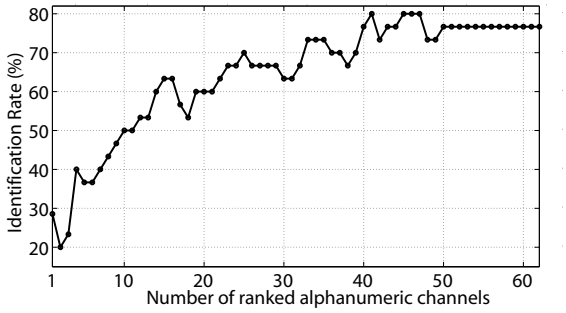

Figure 8. Writer identification rates depending on the number of alphanumeric channels combined (local sub-codebooks, hit list size=1).

We also plot in Figure 9 the identification rates varying the size of the hit list when combining 5, 10, 20, 30, 40 and all the 62 alphanumeric channels. Results are also shown for the global codebook with a size of 750 clusters (according to Figure 6). We observe that working with local sub-codebooks results in much better performance that using a unique single codebook, meaning that the class information given by the character segmentation and labeling carried out by the forensic expert 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.

Concerning the system working with local sub-codebooks, we observe in Figure 9 that there are only slightly differences in performance between combining 40 and all the 62 alphanumeric channels, as mirrored previously in Figure 8. Interestingly enough, if we allow a hit list of size 8-10 (Top 8-10), the combination of only the best 10 alphanumeric channels works as well as other combinations involving more channels. On the other hand, if we want the target identity to be in the first positions of the list (Top 1-2), more alphanumeric channels are needed.

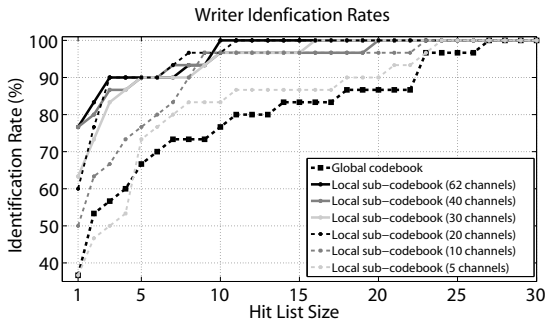

Figure 9. Writer identification rates depending on the size of the hit list size.

5. Conclusions and Future Work

A writer recognition system based on allographic features has been presented. It 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 considers the writer as a stochastic pattern generator. Using a common codebook of handwritten shapes (also called allographs), the personalized set of shapes that each person uses in writing is obtained by computing their occurrence probability. Experiments are carried out using a *global* codebook (i.e. that does not use the character class information) and a set of *local* character-based subcodebooks (i.e one per alphanumeric character, exploiting the class information given by the manual labeling). Results show that much better performance is obtained with local sub-codebooks, justifying the considerable amount of time spent by the forensic expert in the segmentation and labeling process. For the local case, we also evaluate the use of a different number of alphanumeric channels based on its individual identification rate. We observe that the best identification rate is obtained when using 40 channels, with no additional improvement given by the incorporation of additional ones. It is also worthy to note that in the case of big hit lists, the best performance is already obtained with the use of only 10 alphanumeric channels. However, for small hit lists, more alphanumeric channels are needed.

The analysis of these results with a limited database suggest that the proposed approach can be effectively used for forensic writer identification. Future work includes evaluating of our system with a bigger forensic database and applying advanced feature selection methods [18] to the combination of alphanumeric channels, including used-dependent selection approaches [19].

6. Acknowledgements

This work has been partially supported by projects Bio-Challenge (TEC2009-11186), BBfor2 (FP7 ITN-2009-238803) and "Cátedra UAM-Telefónica". Author F. A.-F. is supported by a Juan de la Cierva Fellowship from the Spanish MICINN. Author J. F. is supported by a Marie Curie Fellowship from the European Commission. 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] S. Srihari, C. Huang, H. Srinivasan, V. Shah. *Digital Document Processing*, ch. 17. Biometric and Forensic Aspects of Digital Document Processing. Springer, 2007.
[2] S. N. Srihari, S.-H. Cha, H. Arora, S. Lee. Individuality of handwriting. *J. Forensic Sc.*, 47(4):856–872, 2002.
[3] R. Plamondon, S. Srihari. On-line and off-line handwriting recognition: A comprehensive survey. *IEEE PAMI*, 22(1):63–84, 2000.
[4] S. Srihari and G. Leedham. A survey of computer methods in forensic document examination. *Proc. IGS*, 2003.
[5] L. Schomaker. *Sensors, Systems and Algorithms, Advances in Biometrics*, chapter Writer identification and verification. Springer Verlag, 2008.
[6] L. Schomaker. Advances in writer identification and verification. *Proc. ICDAR*, 2007.
[7] A. Bensefia, T. Paquet, L. Heutte. Information retrieval-based writer identification. *Proc. ICDAR*, 2003.
[8] L. Schomaker, M. Bulacu. Automatic writer identification using connected-component contours and edge-based features of upper-case western script. *IEEE PAMI*, 26(6):787-798, 2004.
[9] L. Schomaker, M. Bulacu, and K. Franke. Automatic writer identification using fragmented connected-component contours. *Proc. IWFHR*, 2004.
[10] M. Bulacu, L. Schomaker. Text-independent writer identification and verification using textural and allographic features. *IEEE PAMI*, 29(4):701–717, April 2007.
[11] G. X. Tan, C. Viard-Gaudin, and A. C. Kot. Automatic writer identification framework for online handwritten documents using character prototypes. *Pattern Recognition*, 42(12):3313–3323, December 2009.
[12] M. Tapiador, J. Sigüenza. Writer identification method based on forensic knowledge. *Proc. ICBA*, 2004.
[13] A. Jain, P. Flynn, A. Ross, eds. *Handbook of Biometrics*. Springer, 2008.
[14] N. Otsu. A threshold selection method for gray-level histograms. *IEEE SMC*, 9:62–66, December 1979.
[15] J. Hull. A database for handwritten text recognition research. *IEEE PAMI*, 16(5):550–554, May 1994.
[16] R. Duda, P. Hart, D. Stork. *Pattern Classification*, 2004.
[17] M. Bulacu, L. Schomaker. A comparison of clustering methods for writer identification and verification. *Proc. ICDAR*, 2005.
[18] J. Galbally, J. Fierrez, M. R. Freire, J. Ortega-Garcia. Feature selection based on genetic algorithms for on-line signature verification. *Proc. AutoID*, 2007.
[19] J. Fierrez-Aguilar, D. Garcia-Romero, J. Ortega-Garcia, J. Gonzalez-Rodriguez. Adapted user-dependent multimodal biometric authentication exploiting general information. *Pattern Recognition Letters*, 26:2628–2639, 2005.