# WRITER VERIFICATION BASED ON GRAPHOMETRIC FEATURES USING FEED-FORWARD NEURAL NETWORK

Carlos F. Romero, Carlos M. Travieso, Jesús B. Alonso and Miguel A. Ferrer
*Signals and Communications Department, Technological Centre for Innovation on Communication (CeTIC)*
*University of Las Palmas de Gran Canaria,Campus of Tafira, Ed. Telecomunicación*
*Pabellón B. 35017, Las Palmas de G.C., Spain*

Keywords: Biometric systems, Writer verification, Handwriting analysis, Neural networks, Pattern recognition.

Abstract: This paper shows a writer verification automatic system based on a set of graphometric characteristics extracted from handwritten words. That dataset has been tested with our off-line handwritten database, which consists of 110 writers with 10 samples per writer, where a sample is a dataset of 34 words. After our experiments, we have got a verification success rate of 95.63% and Equal Error Rate (EER) of 3.90% is achieved. For previous results, we have used as classifiers a Neural Network, for each writer.

## 1 INTRODUCTION

The Biometric Recognition Systems have proliferated greatly in the last decade. Its objective is to substitute or to change the conventional key (passwords, cards, PINs, etc.) for the individual's innate keys. Nowadays, the writing remains of great importance to society because of its wide and extended use in various activities of persons. For this reason, the writer verification on the basis of scanned handwriting images is a very useful biometric modality.

Writer verification involves a one-to-one comparison with a decision as to whether or not the two samples are written by the same person. In this paper, we use Neural Network based handwriting recognition systems for the purpose of writer verification. Thus for N different writers, we obtain N different Neural Networks (NNs) (Bishop, 1995), (Juang and Rabiner, 1992), each NN can be understood as an expert specialized in recognizing the handwriting of one particular person.

The verification is possible because our proposal extracts information biometric of the writing. The scientific bases for this idea come from the brain human. If we try to do writing with the less skilful hand, there will be some parts or forms very similar to the writing with the skilful hand, due to those orders are sent by the brain, and each brain is intrinsic of each person (Romero et al., 2007).

The act to write is a phenomenon governed by the brain and integrated in the psychomotricity of the individual; in contrast to mimic movements, the handwriting movements are fixed toward a plane that allows its study and measurement. The writing, like codified and dynamic message that reflects certain biometric information of the individual in his/her communication with the others, is fundamentally individual, recognizable, univocal and unique; that makes possible the user identification.

Recently, different approaches to writer verification have been proposed in (Cha and Srihari, 2000), they addressed the problem of writer verification, i.e., the problem of determining whether two documents are written by the same person or not. In order to identify the writer of a given document, they model the problem as a classification problem with two classes, *authorship* and *non-authorship*. Given two handwriting samples, one of known and the other of unknown identity, the distance between two documents is computed. Then, the distance value is used to classify the data as positive or negative.

(Leedham and Chachra, 2003) present a set of eleven features which can be extracted and used for the identification and verification of documents containing handwritten digits. These features are represented as vectors and using the Hamming distance measure and determining a threshold value

F. Romero C., Travieso C., Alonso J. and Ferrer M. (2010).
WRITER VERIFICATION BASED ON GRAPHOMETRIC FEATURES USING FEED-FORWARD NEURAL NETWORK .
In *Proceedings of the Third International Conference on Bio-inspired Systems and Signal Processing*, pages 353-358
DOI: 10.5220/0002589003530358

for the intra-author variation a high degree of accuracy in authorship detection is achieved.

(Schalapbach and Bunke, 2004) used HMM based recognizers for the identification and verification, a text line of unknown origin is presented to each of these recognizers and each one returns a transcription that includes the log-likelihood score for the considered input. These scores are sorted and the resulting ranking is used for both identification and verification. The system must decide, based on a verification criterion, whether a text line with a claimed identity is in fact from this writer or if it is an impostor attempt. Using a confidence measures to define the following verification criterion: if the confidence measure is minor than a certain threshold, then it assumes that the text line is from certain writer; otherwise the input is classified as not being of the claimed identity.

(Bensefia et al., 2005) proposed a writer verification building a hypothesis test, based on the mutual information between the grapheme distributions in the two handwritings that are compared. Mutual information allows measuring the independence between the two writers, for low values of mutual information indicate a strong independence showing that the set of features is distributed in the same way on the two documents and reflecting the same identity for the two writers. While the high values denote a strong dependence, showing different identities for the two writers.

(Bucalu and Shomaker, 2007) perform writer verification using a distance measure between the feature vectors to compute the similarity in individual handwriting style, between any two chosen samples. Distances up to a predefined decision threshold are deemed sufficiently low in order to consider that two samples have been written by the same person. Otherwise, samples are considered written by different persons.

The goal of this paper is to introduce a writer verification system from isolated words. The rest of the paper is organized with the following sections. In sections 2 and 3 are shown the methodology of this work and the building of our database, respectively. In section 4 is briefly described the image pre-processing and the segmentation of the words. In the following section, the procedure for the extraction of the characteristics is explained. Section 6 includes the classification system based on Neural Networks. In section 7 is shown verification tasks, and the following section can be observed conclusions. And finally, this present work ends with Acknowledgement and references.

## 2 DESCRIPTION OF THE SYSTEM PROPOSED

As the majority of the works proposed up to now (see Figure 1), on biometric recognition, the framework of the system depends on the following basic steps.

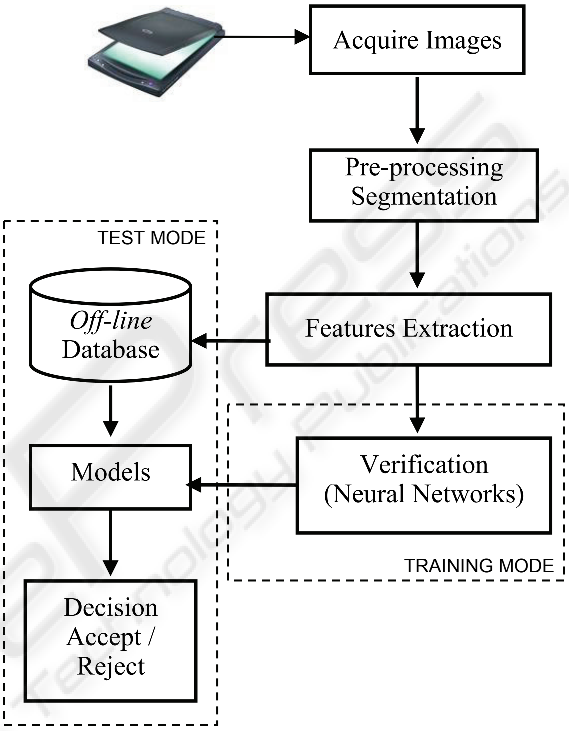

Figure 1: Proposal of our writer verification system.

1. Image pre-processing and segmentation. Preparation and modification of images, so that the module of segmentation produce the results desired. The segmentation separates the zones of interest (lines, words or characters), and it is key for the success or error of the following analysis (Feature Extraction).
2. Feature Extraction. This is the calculation of qualitative and quantitative measures that permit to obtain a significant geometrical characterization of the style of writing, in order to differentiate writers among themselves. Pressure, speed, direction, inclination, cohesion, continuity, opening of ovals constitutes some of the graphologist features. The aspect of the letters, its form and dimension is some of the classical features. The graphologist features help to be more discriminating. In this present work, the most of our graphologist features are

automatically got. In future works, we hope to reach it.

3. Classification: A statistical supervised analysis of the extracted characteristics is carried out, which will permit the comparison with the samples of our database, if the query sample possess similarities is accepted, otherwise is rejected.

## 3 DATABASE

For the building of our database, we have used a paragraph of 15 lines,the language used is Spanish. That text is from "*Don Quijote de la Mancha*" from Miguel de Cervantes, and we have used the same text for each writer. With this size of text, writers can show their personal characteristics, because they keep their writing habits. This database has been built with 110 writers, and each one has made 10 times this template (paragraph of 15 lines). The size of paper was DIN-A4 format (297 mm. x 210 mm). The sheet was written with a pen of black ink. Each writer of our database had one week for doing the writing, and therefore, it is considered like an effect of temporal invariance on this database.

The creation conditions of our database were the normalized with the same type of paper (80 gr/m2), ballpoint pen, and similar place of support (for doing the writing). Of this way, our work is centred on the writing and the efficiency of proposed parameters. In the future work, we are going to change the rest of variables.

The samples are scanned with 200 dpi, obtaining images on grey scale, with 8 bit of quantification (see figure 2).

## 4 IMAGE PRE-PROCESSING AND SEGMENTATION

The first step of the image pre-processing consists of utilizing Otsu's the method, in order to get the binarization of the samples (Otsu, 1979).

As a result of the binarization, in most cases, the line of writing remains with irregular aspect. For that reason, we have implemented other image preprocessing for skew elimination. We developed an algorithm in order to detect the maximum projection of the word, using histogram tool. This permit to smooth out the baseline, so, the baseline remains well defined. Besides, it is eliminated the existing noise in the images after scanning process, by morphological mathematics.

As previous step to the separation of words or components connected (by labelling), the detection and elimination of the punctuation marks (points, accents and comma) is carried out, by a size threshold, In particular, we have removed components connected if they were minor than 60 pixels.

Finally, it is segmented words, which compose the lines of writings (baselines) and for it, it is must establish limits of each one of the words. For this estimation, the method of the "Enclosed Boxes" (Jaekyu and Haralick, 1995) was used, which provides us the coordinates that will permit segment the words. The enclosed boxes are defined as the most minimum rectangle that contains to the component connected.

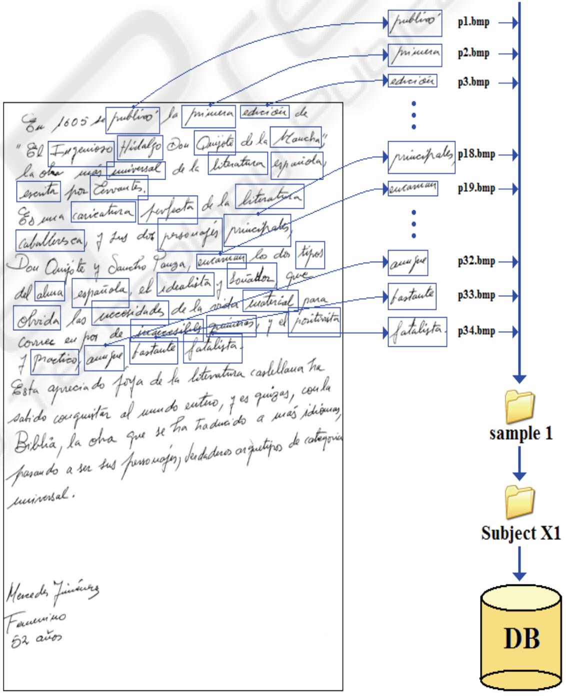

Figure 2: Example of the words extraction for a sample.

## 5 FEATURE EXTRACTION

Writer verification and identification methods fall into two broad categories: text-dependent versus text-independent methods. If any text may be used to establish the identity of the writer the system is text independent. Otherwise, if a writer has to write a particular predefined text to identity the verification task is text dependent.

This present work has used a selection of the most discriminative features from (Romero et al. 2009), (Hertel and Bunke. 2003), (Srihari et al., 2001), which are depending on a set of 34 specific words. These features were separated into two groups, one set of dependent text feature and other set of independent text feature.

Handwriting cohesion is called to the percentage of unions that appear between the letters of the same one; when saying unions, we are talking about the final strokes of the letters, continuing with the initials of the following letters without ballpoint pen is risen of the paper.

In order to make an estimation of the cohesion in the handwriting, the images of the 34 words of each sample are selected and binarized and their components connected with connectivity-8 are labelled to them. As soon as the quantity of components connected of each word is obtained, it is proceeded to calculate the average and the variance of the components connected. Those words have been selected by their size, the largest.

The estimation of the width of letters is carried out, seeking the row with greater quantity of black to white transition (0 to 1). It is counted the number of white pixels between each transition, this result is averaged.

In order to measure the height of the medium body of the words, the goal is to determine the upper and lower baseline through maximums and minimum values and to measure the distance among them.

In order to approach baselines of each word, it was decided to use the adjustment of minimum mean square error that is based on find the equation (see expression 1) that better be adjusted to an set of points "n" (Morita et al., 2001). The equation is the following:

$$y = ax + b \tag{1}$$

where the coefficients “a” and “b” determine the lineal polynomial regression by means of the following expressions:

$$a = \frac{n\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} \tag{2}$$

$$b = \frac{\sum_{i=1}^{n} y_i - a\sum_{i=1}^{n} x_i}{n} \tag{3}$$

Those values of “a” and “b”, based on the coordinates of minimums or maximums detected in the contour of the word, are different baselines. Minimums are to approach the lower baseline and the maximums for the superior baseline (see figure 3).

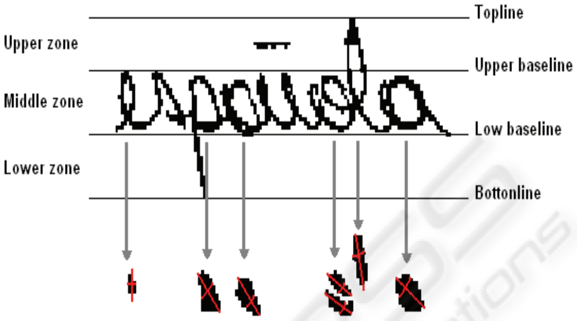

Figure 3: Ovals and loops with its respective major and minor axes.

For the estimation of pressure by means of the width of the strokes, the run lengths are determined on the binary image taking into consideration black pixels corresponding to the ink trace. There two basic scanning methods: horizontal along the rows of the image and vertical along the columns of the image.

The area of thinning is defined as the area remaining in the resulting image of the word after it undergoes a process of three morphological stages: thinning, dilation and thinning again.

The following features are dependent text: length of the words, quantity of pixels in black, heights of the ascending and descending, height relation between descending and ascending, height relation between descending/ascending and medium body. As for the analysis of the ovals and loops of the words, is done the measurement of the minor and major axes of each ovals and loops from 34 words. Axes are calculated by maximum projection using histograms. Finally, it is calculated the average size of the above mentioned axes of the handwriting sample in analysis.

## 6 NEURAL NETWORKS

As classifier, we have used a Feed-Forward Neural Network (NN) with a Back-propagation algorithm for training (Bishop, 1995) (Juang *et al.*, 1992), where the number of input neurons is given by the dimension of the vector of features with 374 parameters (11 parameter x 34 words). And the number of output neurons is given by the number of writers to identify.

Too, we have researched and checked this system varying the number of neurons in the hidden layer in order to improve the success rate.

Finally, we have used the method of the 'more voted' algorithm, where we have built a schedule with 10 neural networks (see figure 4), and we have reached a improvement between 3 and 4 points on success rates

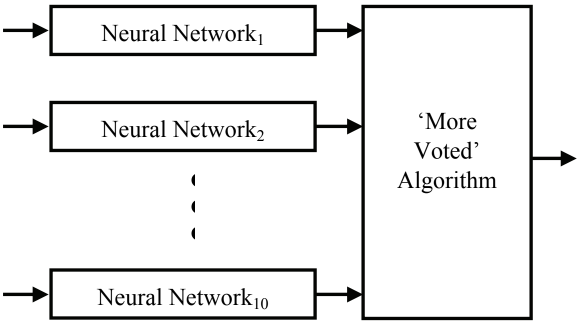

Figure 4: Classification System with 'more voted' algorithm, based on NN.

## 7 WRITER VERIFICATION EXPERIMENTS

The verification can be seen as a problem of classification of two classes. When are compared the samples: the variation of a writer among their own samples should be smaller than the variation among samples of two different writers.

In order to give solution to this problem, the methodology of the used verification was independent supervised classification. Therefore, we have a system with two modes, training and test modes (see Figure 1).

For the training, we have used the 50% of our database, and the remainder to carry out the test mode. That is, five samples have been chosen to training and other five for the test, since we have 10 samples for each writer. Besides, a total of 34 words have been extracted from paragraphs, and there will be 34 words by sample.

Like our parameters depend on words used and its writer, we have used the same word for this process; in particular, we have used a set of 34 words from the paragraph of 15 lines. But, the samples set for training and test mode of these 34 words are different, being obtained from 10 different samples of each writer. Therefore, this system works with a close set of words (34 words).

In order to calculate the characteristics, we have used 170 words (34 x 5 samples/writer) on the training process. The criterion of selection to choose the previous 34 words was their length, upper than 5 letters, because with this length, they offer information more general than a word with a shorter size.

For writer verification, we employed a threshold to decide if the samples query is accept or reject. By varying the threshold, a Receiver Operating Characteristics (ROC) curve (see Figure 5) is obtained that illustrates the inevitable trade-off between the False Acceptance Rate (FAR) and False Rejection Rate (FRR). The Equal Error Rate (EER) corresponds to the point on the ROC curve where FAR=FRR and it quantifies in a single number the writer verification performance.

To analyze the performance of our system, we used dependent text feature, independent text feature and combination of both. The results of the verification experiments are given in Table 1.

The best ROC curve is produced using combination features with a threshold of 0.027 yielding the following results: nearly 96% of correct verification with an EER of about 3.90%, FAR of 4.23% and FRR of 4.37%.

Table 1: Writer verification performances.

| % | Independent Text | Dependent Text | Combination |
|---|---|---|---|
| Threshold | 0.030 | 0.027 | 0.027 |
| EER | 8.10 % | 6.10 % | 3.90 % |
| FAR | 8.35 % | 6.17 % | 4.23 % |
| FRR | 8.19 % | 6.19 % | 4.37 % |
| Success Rate | 92.23 % | 93.82 % | 95.70 % |

## 8 CONCLUSIONS

In this paper, we have presented a system that uses NNs for the task of off-line writer verification. The basic input units presented to the system are handwritten words. We have tested our system on the verification task using independent and dependent text features are extracted from each word.

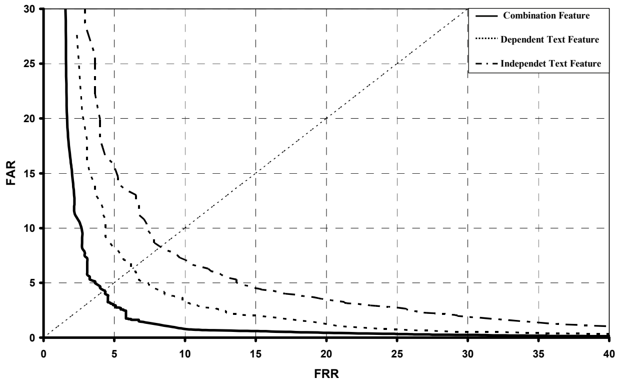

Figure 5: ROC curve of the verification experiment.

Currently, our better performances have been combining independent and dependent text features. The system performs very well on both tasks of accept and reject. An Equal Error Rate (EER) of about 3.90% is achieved.

In the future work, we plan to address at improving the performance independent text feature. We also plan to expand our database and test our system with other databases such as the IAM database.

## ACKNOWLEDGMENTS

This present work has been supported by private funds from Spanish Company, Telefónica. This research project is called "Cátedra Teléfonica-ULPGC 2009"; and by funds from Research Action from Excellent Networks on Biomedicine and Environment belong to ULPGC.

## REFERENCES

Bishop, C.B., 1995. *Neural Networks for Pattern Recognition*, Oxford University Press.

Juang, B.H., Rabiner, L.R., 1992. Spectral representations for speech recognition by neural networks-a tutorial. In *Proceedings of the Workshop Neural Networks for Signal Processing*, pp. 214 – 222.

Romero, C.F., Travieso, C. M., Alonso, J. B., Ferrer, M. A., 2007. Using Off-line Handwritten Text for Writer Identification. In *WSEAS Transactions on Signal Processing*. Issue 1, Vol. 3, pp. 56-61.

Cha, S. H., Srihari, S., 2000. Multiple feature integration for writer verification. In *Proceedings Seventh Inernational. Workshop on Frontiers in Handwriting Recognition*, pages 333-342.

Leedham, G., Chachra, S., 2003. Writer Identification using Innovate Binarised Features of Handwritten Numerals. In *Proceeding of the 7th International Conference on Document Analysis and Recognition*. Vol. 1, pp. 413-416.

Schlapbach, A., Bunke, H., 2004. Using HMM Based Recognizers for Writer Identification and Verification. In *Proceeding of the 9th Int'l Workshop on Frontiers in Handwriting Recognition*

Bensefia, A., Paquet, T., Heutte, L., 2005. Handwritten Document Analysis for Automatic Writer Recognition. *Electronic Letter on Computer Vision and Image Analysis*, pp. 72-86.

Bucalu, M., Shomaker, L., 2007. Text-Independent Writer Identification and Verification Using Textual and Allographic Features. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 29, No. 4.

Otsu, N., 1979. A threshold selection method from graylevel histograms. In *IEEE Transaction on Systems, Man and Cybernetics*, Vol. 9, Is. 1, pp 62-66.

Jaekyu Ha, R. M., Haralick, I. T., 1995. Document page decomposition by the bounding-box project, In *Proceedings International Conference on Document Analysis and Recognition*, Vol. 02, No. 2, p. 1119.

Romero, C.F., Travieso, C. M., Alonso, J. B., Ferrer, M. A., 2009. Hybrid Parameterization System for Writer Identification. In *Second International Conference on Bio-inspired Systems and Signal Processing*, pp. 449-454.

Hertel, C., Bunke, H., 2003. A Set of Novel Features for Writer Identification. In *Proceedings of the Audio and Video Based Biometric Person Authentication*. pp. 679-687.

13. Srihari, S., Cha, S.H., Arora, H., Lee, S., 2001. Individuality of Handwriting: A Validity Study. In *Proceedings of International Conference on Document Analysis and Recognition* pp 106-109.

14. Morita, M.E. Facon, J. Bortolozzi, F. Garnes, S.J.A. Sabourin, R., 1999. Mathematical Morphology and Weighted Least Squares to Correcthandwriting Baseline Skew. In *Proccedings of the 5th International Conference on Document Analysis and Recognition*, pp. 430-433.