Using off-line handwritten text for writer identification

CARLOS F. ROMERO, CARLOS M. TRAVIESO, JESÚS B. ALONSO, MIGUEL A. FERRER.
Departamento de Señales y Comunicaciones
Universidad de Las Palmas de Gran Canaria
Campus de Tafira, E35017, Las Palmas de Gran Canaria. SPAIN
e-mail: fabian_romero@ciudad.com.ar; {ctravieso, jalonso, mferrer}@dsc.ulpgc.es
URL: http://www.gpds.ulpgc.es

Abstract: - This work has calculated and implemented some methods used by professional person on forensic analysis, for test on dubitative documents. This system obtains different types of characteristics and they are tested with known samples from our database. It has been used writing samples from 30 writers, and we have got a success rate of 94,66%, applying as classifier Neural Network, and after, the technique of "more voted" algorithm, with 10 Neural Networks.

Key-Words: Writer Identification, Document Analysis, Handwriting Individuality, Features Extraction, Neural Networks.

1 Introduction

Nowadays Computer Science advances and the proliferation of computers in the modern society, it is an unquestionable fact. But the great importance, that continues having the handwritten document and the own writing, is true.

For this reason and for this wide use, many handwritten documents are exposed to possible forgeries, deformations or copies, and generally, with illicit use. Therefore, a high percentage of routine work is made by Experts and Professionals in this field, whose task is to certify and to judge the authenticity or falsehood of handwritten documents (for example: testaments) in a judicial procedure.

Writer identification is possible because the writing for each person is different, and everyone has personal characteristics. The scientific bases for this idea are 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 this order are sent by the brain. Generally, this effect is projected toward the writing by two types of forces, they are:

- Conscious or Known: because it can do a control of the own free will.
- Unconscious: because it escapes to control of the own free will. This is divided into: forces of type mechanical and emotional, where are harboured feelings.

Nowadays, the writer identification is a great challenge because these researches are not as studied as the identification based on fingerprints, hands, face or iris (other biometric techniques), due mainly to that the operation of the brain is very difficult of parameterize. On the other hand, the mentioned techniques use widely researched biometric information.

At present, two software tools exist available for the Experts or Professionals, which permit to show and to visualize certain characteristics, but Experts have to investigate and use so much time to extract theirs conclusions about the body of writing. Therefore, these tools save neither time nor a meticulous analysis over the writing. They have to work with graph paper and templates for obtaining parameters (angles, dimensions of the line, directions, parallelisms, curvatures, alignments, etc.). Too, they have to use magnifying glass with graph paper for doing measures of angles and lines.

1.1 Description of system

The objective of the identification writer is to capture the individual characteristics, ignoring the content of the message. This is the difference with the writing recognition, because it removes the individual variability and recognizes the message.

In a judgment, a lot of times, documents are evidences that are only available off-line mode. Therefore, our system works with this mode. We have to scan the document, before its analysis.

1.1.1 Database

For the building of our database, we have used a paragraph of 15 lines. With this size, writers can show theirs personal characteristics and so that they remain reflected the writing habit.

This database has been built with 30 writers, and each one has made 10 times this template (paragraph

of 15 lines). The size of paper was a DIN-A4 format (297 mm. × 210 mm). The sheet was writen and writers used a pen with black ink. Each writer of our database had one week for doing the writing, and therefore, we produce a temporal invariance.

The conditions of creation our database were the normalized, the same type of paper, pen, and similar place of support (for doing the writing).

The samples are scanned 300dpi, obtaining images in grey scale, with 8 bit of quantification.

1.1.2 Framework

As the majority of the works proposed up to now, on biometric recognition, the framework of the system depends on the following basic phases.

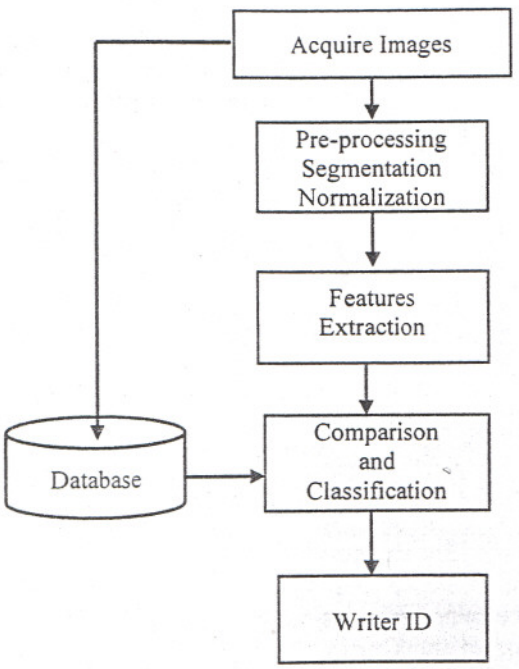

Fig. 1. System of writer identification.

- Pre-processing: 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 is key for the success or error of the following analysis (Feature Extraction).
- Feature Extraction: They are qualitative and quantitative measures that permit to obtain a significant characterization of the style of writing, to differentiate writers among themselves.
- Classification: A statistical analysis of the extracted characteristics is carried out, which will permit the comparison with the samples of our database, seeking the writer, who possesses more similarities.

2 Pre-processing

The first step of the image pre-processing consists of utilizing Otsu's the method, which permits us to determine the necessary grey threshold value to carry out the binarization of the samples [1].

As a result of the binarization, in most cases the line of writing remains with pixel view, that is, little consistent in some parts. For that reason, another pre-processing is carried out that permits to smooth out the line, so that remain well defined. And also, it eliminates the existing noise in the images of the samples after scanned.

As previous step to the separation of words or components connected, the detection and elimination of the punctuation marks (points, accents and comma) is carried out.

Finally, it is segmented words, that 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" [2] 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. To each segmented word, it is applied a correction of the Skew.

3 Feature Extraction

In this step, it is created a list with statisticians of different quantitative measures to analyzed document. After, it will be compared with the obtained samples of the database.

So that the characteristics represent the style of writing, they should comply with the following requirement: the fluctuations in the writing of a person should be as small as be possible, while the fluctuations among different writers should be as large as be possible.

The characteristics extracted in this work are the following:

- long of the words,
- quantity of pixels in black,
- estimation of the wide one of the letters,
- height of the medium body of writing,
- heights of the ascending and descending,
- height relation between of the ascending and medium body
- height relation between descending and medium body

- height relation between descending and, ascending
- height relation between medium body and the wide of writing.

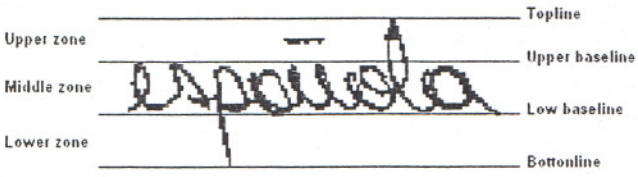

Fig. 2. Zones and baselines.

The quantity of black pixels and the long words, they will give us an estimation of the dimension and thickness of the line, the wide of letters and the height of the medium body. Besides these are distinctive characteristics of the style of writing.

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

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 (see figure 2).

For the obtaining of the coordinates of the maximums and minimums of the contour, in the first step, we have to keep in mind the pen can be separated of the paper, and it produces no-connected among of the same word. For it, a morphological operation was carried out to achieve that the line of writing be continuous along the word. This morphological operation [3] is a closed operation to the image of the word, which consists of a dilatation and after this, the erosion is applied. We have used a structural element of 3x3 because we have found that is a size would without distortion for the form of the word.

In the following expressions, we have shown those operations, where I is the image, and E the structural element. Both of them are considered as an set of pixels in a n-space D^n with elements $i = (i_1,..., i_n)$ y $e = (e_1,..., e_n)$, respectively. The morphological operations are defined with the following expressions:

- Dilatation
  $I \oplus E = \{x \in D^n \mid x = i + e\}$ *for* $i \in I$ and $e \in E$
- Erosion
  $I \ominus E = \{x \in D^n \mid x + e \in I\}$ *for* $e \in E$
- Closed
  $I \bullet E = (I \oplus E) \ominus E$

Subsequently, we are going to work with the external contour of the line of words, we have filled holes of letters (for example, the interior of the letter a, o, etc.), eliminating in this way, the interior contour. After this, a dilatation is carried out and the image obtained is subtracted with the original image, obtaining an image where only appears the exterior contour of the line with connectivity-4.

Once it is obtained the contour, we have passed the image to connectivity-8, eliminating corner pixels. A corner pixel is every pixel of the contour that does not contribute prominent information; therefore, it can be eliminated without loss of continuity in the line. Its detection is carried out sweeping the image with a mask of 3x3 pixels (see figure 3).

0	0	X	X	1	0
0	1	1	0	1	1
X	1	0	0	0	X
X	0	0	0	1	X
1	1	0	1	1	0
0	1	X	X	0	0

Fig. 3. Masks for detecting the corner pixel. (X= any value).

If the region analyzed of the image coincides with some of these masks, the central pixel is eliminated (that is, it takes the value of the background colour).

The following step will be to carry out a sweeping of the contour, keeping all the coordinates of the pixels of the contour in a vector; but before a starting point should be found, for it, we used the coordinates of the centre of mass on the binary image $I(i, j)$ (the image of the word). These expressions come given by:

$$C_X = \frac{1}{N}\sum_{(i,j)\in R} i \qquad C_Y = \frac{1}{N}\sum_{(i,j)\in R} j \qquad (1)$$

Where N is the quantity of pixels (area) that occupies the word, R is the pertinent region to the line of the word, and the indexes $(i\ j)$ correspond to the coordinates x, and of the pixels of the line.

Therefore, as the component on "y" of the starting point are used C_Y, and the same for the component on "x" (C_X).

Once the starting point is obtained, we have worked with Chain Code method, on connectivity-8 [4], which supposes that in a binary image, the edge of the word is represented by a of "1"; that is, that connected line consists on a segment of pixels connected, therefore, to carry out the sweeping of

contour, we have worked the opposite of clockwise with the following matrix:

4	3	2
5	P	1
6	7	8

Fig. 4. Code matrix.

The code matrix is utilized to show which is the way that continues the line. It is a matrix of 3x3 where the central element is the pixel selected, and each one of the eight bordering elements, they indicate the possible directions of the line, since pixels of the contour are connected by a single neighbour.

The method starts with the starting point on the centre of the code matrix. After, to sweep the neighbouring pixels in the opposite of clockwise. When a following pixel is detected for a element from the code matrix, this value is saved by its coordinates (x,y) of the present pixel P and the system follows this process until the last pixel of the contour (starting point).

Finally, the vector, where all the coordinates of the pixels of the contour of the word have been kept, is taken and all points outside of the two bands are filtered (a filter for maximums and another for minimums).

Once obtained the two bands, it is carried out a filter, keeping in a vector, only the coordinates of the pixels of the contour, those are inside of the specific bands. Then the two resultant vectors are taken (Vb_1 and Vb_2), and it is calculated the derivate with respect to axis of coordinates "y", from the discrete function represented by the values of each vector.

$$\frac{\partial Vb_1}{\partial y} = 0 \quad \Rightarrow Maximum \tag{2}$$

$$\frac{\partial Vb_2}{\partial y} = 0 \quad \Rightarrow Minimum \tag{3}$$

The pixels, where the derivate is zero, will be maximums and minimums looked for.

To approach the 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 4) that better be adjusted to an set of points "n" [5]. The equation is the following:

$$y = ax + b \tag{4}$$

Where the coefficients a and b are determined by regression lineal with the following expressions:

$$a = \frac{n\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} \tag{5}$$

$$b = \frac{\sum_{i=1}^{n} y_i - a\sum_{i=1}^{n} x_i}{n} \tag{6}$$

The values of a and b are the coordinates of minimums or maximums detected in the contour of the word. Minimums are to approach the lower baseline and the maximums for the superior baseline.

Due to the resultant baselines can have different value of inclination in some cases, that is, they are not always parallel, then; three measurements of distance among the lines base are carried out (an in the right side, another in the left side and the last measure, in the centre of the word). And according to the three resultant values, the average distance is calculated. That will be equal the average height of the body of writing of the word.

4 Classification and results

The identification can be seen as a problem of classification of N classes, in our case N writers. There are two variations of interest when are compared the samples: about the writing of a same writer and between the writings of two different writers. The variation of a writer among their own samples should be smaller that the variation among samples of two different writers.

For giving solution to this problem, the methodology of the used identification was supervised classification. Therefore, we have a system with two modes, training and test mode.

For the training, we have used the 50% of our database, and the remainder to carry out the test mode. That is, five words have been choosen to training and other five for the test, because we have 10 samples for each writer.

A total of 34 words have been extracted from the paragraph. Therefore, we have used 170 samples (34 × 5 words) on the process of training. The criterion for choosing the previous 34 words was theirs length, more than 5 letters, because with this length, they offers more information than a short size.

The experiments have been carried out in five occasions, for which the results are shown by theirs average rate and theirs standard deviation.

As classifier, we have used two different classifiers, Neural Network (NN) [11][12] and Support Vector Machines (SVM) [13][14][15].

For the first classifier, we have used a Feed-Forward Neural Network (NN) with a Back-propagation algorithm for training, where the number of input units is given by the dimension of the vector of features. And the number of output units is given by the number of writers to identify.

Too, we have researched with different number of neurons in the hidden layer, and finally, 42 neurons were used, because they have presented the better results.

The average success rate of recognition is 89,73%, with a standard deviation of 1,64. But this result was improved using the method of the 'more voted', where we have built a schedule with 10 neural networks (see figure 5), and we have found a recognition rate of 94,66%, with a standard deviation of 0.

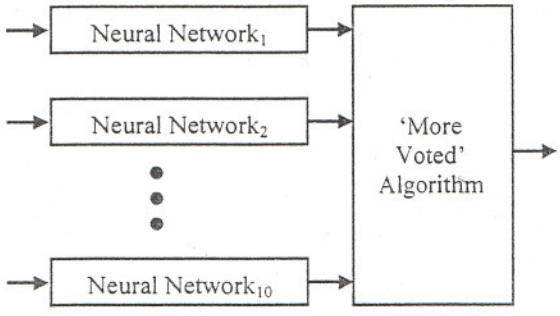

Fig. 5. Classification System with NN and 'more voted' algorithm.

The other classifier is SVM. This model is based on geometric properties to calculate a hyperplane that separate the different classes of data. The basic idea is shown in the figure 5, training this system, we would have two sets of vectors (in two dimensions) that would be the classes to classify. This system would calculate the border hyperplane, H (in two dimensions would be a straight line) among these two sets,

$$\mathbf{w}\cdot\mathbf{x} + b = 0, \tag{4}$$

where:

- w is normal to hyperplane
- $\frac{|b|}{\|\mathbf{w}\|}$ is the distance perpendicular from the hyperplane to origin
- $\|\mathbf{w}\|$ is the Euclidean norma of $\overline{w}$
- b is the independent element
- x is a point contained in the plane

At the same time we would have other two hyperplaness, H1: $\mathbf{x}_i\cdot\mathbf{w} + b = 1$ and H2: $\mathbf{x}_i\cdot\mathbf{w} + b = -1$ that contain the vectors (points) more nearby. Those vectors are known like support vectors. The distance between both planes is knows like "margin". The objective of the algorithm will be to maximize that margin.

Once it is had border hyperplane, due the system has been training, it will decide the side of the limit of decision that belongs (the hyperplane situated between H1 and H2 and parallel to them) a class, from a test given x and the system will assign the corresponding label of class, therefore we will take the class of x to be $\text{sgn}(\mathbf{w}\cdot\mathbf{x} + b)$. In the next figure, we can observed this classifier.

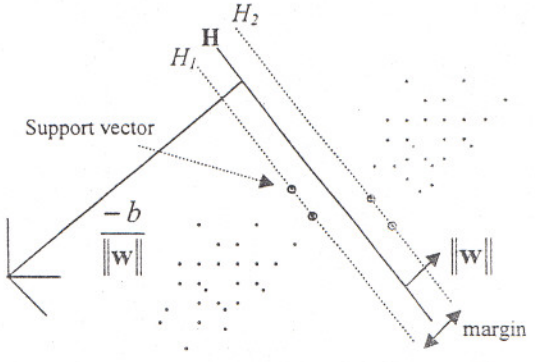

Fig. 6. Lineal border hyperplane (lineal and separable case).

There are different types of kernel functions, the most known and that better results give are; Kernel polynomial of degree P, Kernel of radial base Gaussiana (RBF – "radial basis function") and Kernel sigmoidal. We have used lineal (see figure 6) and RBF kernel. The obtained results are showed in the table 1.

Finally, we have the not separable and not lineal case. Now it is suggested as to be able to generalize the previous method for the case in which the decision function is not a lineal function of the data. It is mapped the data to some another Euclidean space H (possibly of infinite dimension), using a map that we will call Φ, where the data be separable lineally:

$$\Phi : R^d \rightarrow \text{H} \tag{5}$$

For the present work, we have using multiclass SVM with one-versus-all strategy, due to we have 30 different classes (30 writers).

Cost(c)	Lineal kernel		RBF kernel		
	mean	std	mean	std	gamma
1	72.27%	3.35	64.67%	3.09	2.10^{-3}
10	77.73%	2.03	71.87%	5.06	5.10^{-4}
50	78.40%	2.77	78.13%	4.33	2.10^{-4}
100	78.40%	2.77	79.07%	5.18	10^{-4}
200	72.67%	3.35	81.87%	7.05	2.10^{-4}
300	73.07%	5.97	83.07%	2.77	10^{-4}
500	74.19%	4.65	81.01%	3.58	10^{-4}

Table 1. Comparison of results among different kernels for SVM.

We can compare our results with other authors, but this comparison depends on databases, like all databases are different, this comparison is relative.

Autor	Number of writers	Success Rates
Said [6]	40	95 %
Zois [7]	50	92,5 %
Marti [8]	20	90,7 %
Hertel [9]	50	90,7 %
Bensefia [10]	150	86 %
This work	**30**	**94,66%**

Table 2. Comparison of results among different published methods vs. our work.

5 Conclusion

In this present work, we have proposed new parameters for identification writer. We have used a back-propagation NN and SVM for the classification. The best results have been found for NN, and for improving results, we are implemented a 'more voted' algorithm. The success rate is 94,66% for our database.

References:

[1] N. Otsu. A threshold selection method from gray-level histograms. IEEE *Transaction on Systems, Man and Cybernetics*, 9(1):62-66, Jan. 1979.

[2] Jaekyu Ha, R.M. Haralick, I.T. Phillips. Document page decomposition by the bounding-box project, *ICDAR*, Vol. 02, No. 2,1995, p. 1119.

[3] R.C. González y R.e. Word. Digital Image Processing. Addison-Wesley Publishing Company, Inc., 1993.

[4] S. Madhvanath, G. Kim and V. Govindaraju. Chaincode Contour Processing for Handwritten Word Recognition. IEEE *Transactions on Pattern Analisys and Machine Intelligence*, 1999, Vol. 21, N° 9.

[5] W. Chin, M. Harvey y A. Jennings. Skew Detection in Handwritten Scripts". IEEE Region 10 Annual Conference. *Speech and Image Technologies for Computing and Telecommunications*, 1997, Vol. 1, p. 319-322.

[6] H. E. S. Said, G. S. Peake, T. N. Tan y K. D. Baker. Writer Identification from Non-uniformly Skewed Handwriting Images. *Proc. of the 9th British Machine Vision Conference*, 1998 pp. 478-487.

[7] E.N Zois, V. Anastassopoulus. Morphological Waveform Coding for Writer Identification. *Pattern Recognition*, Vol. 33, N°3, 2000, pp. 385-398.

[8] C. Hertel and H. Bunke. A Set of Novel Features for Writer Identification. In J. Kittler and M. Nixon, editors, *Audio and Video Based Biometric Person Authentication*, 2003, pp. 679-687.

[9] U. V. Marti, R. Messerli y H. Bunke. Writer Identification Using Text Line Based Features. *Sixth International Conference on Document Analysis and Recognition*, 2001, pp. 101-105.

[10] A. Bensefia, T. Pasquet, L. Heutte. Handwritten Document Analysis for Automatic Writer Recognition. *Electronic Letters on Computer Vision and Image Analysis*, 2005 pp. 72-86.

[11] C. M. Bishop, Neural Networks for Pattern Recognition, *Editorial. Oxford University* Press, 1995.

[12] Hirose, Y., Yamashita, K., Hijiya, S., "Back-propagation algotithm which varies the number of hidden units", *Neural Networks*, vol 4, 1991, pp 61-66.

[13] Christopher J. C. Burgues: A tutorial on Support Vector Machine for Pattern Recognition. *Data Mining and Knowledge Discovery*, Vol. 2, 1998, 121-167.

[14] Jaakkola, T., Diekhans, M., and Haussler, D. (1998). "A discriminative framework for detecting remote protein homologies". Unpublished, available from http://www.cse.ucsc.edu/research/compbio/research.html (visited on 24 September, 2006)

[15] N. Cristianini, J. Shawe-Taylor, Support Vector Machines, *Ed. Cambridge University Press*, 2000.