Writer Identification Based on Forensic Science Approach

Francis L. Baranoski, Luiz S. Oliveira, and Edson J. R. Justino
Pontifícia Universidade Católica do Paraná
Programa de Pós-Graduação em Informática
Rua Imaculada Conceição, 1515
Curitiba, PR, Brazil- 80215-901
francisbaranoski@gmail com
soare@ppgia.pucpr.br
justino@ppgia.pucpr.br

Abstract

Writer identification is an important field of Forensic Science correlated to the Questioned Document Examination area. It is applied to many types of investigation like fraud, homicide, suicide, drug trafficking and clandestine labs, sexual offences, extortion and others. Computer approaches have been developed to help the experts on the handwriting graphical analysis, removing from the analysis the subjectivity imposed by the human measurements.. This paper reports a method to writer identification based on Forensic Science approach, graphometric features and Support Vector Machines (SVM).

Keywords: Forensic, Document, Patter Recognition, SVM, Manuscript.

1 Introduction

The Questioned Document Examination (QDE) is an area of the Forensic Science with the main purpose to answer questions related to questioned document (authenticity, authorship and others). The QDE has a large field of applications and are used for different agencies, like the Federal and Civil Law Enforcement and Justice Area. There are basically two different sub areas in the QDE: the document analysis and the handwriting analysis. The first one approaches the structural analysis of the document to find adulteration, falsification, obliteration and others. The second one approaches the originality or the association between one or more manuscripts to an author [1] and [2].

H. de Gobineau and R. Perron [16] elaborated a basic theory of graphometry, or more exactly a statistical method to measure the similarity between handwriting graphic elements. The graphometry or handwriting analysis is applied to many types of investigation like fraud, homicide, suicide and others.

The graphometry has two basic analysis subjects, manuscripts and signatures. Even with distinct features, both keep a narrow relation having the same root or origin in the writer's learning process, in other words, they carry the experiences acquired by the writer during and after his learning process through the improvement of the handwriting personal style [3].

2 Questioned Document Examiner's Approach

The experts classify a manuscript, in relation to the authorship, like association or dissociation [4]. The association indicates that the questioned manuscript was elaborated by the presumed author. The dissociation indicates that the manuscript wasn't elaborated by the presumed author.

During the analysis the expert uses a set of n manuscripts of known author (reference) K_i *(i=1,2,3...n)* in comparison with the questioned manuscript Q. The expert observes, based on the graphometric features f_i *(i=1,2,3,..,L)*, similarities or dissimilarities between the reference and the questioned manuscript (Figure 1). He or she repeats the process for all references. The expert's report D depends on the sum of the obtained results from the individual comparisons of the pairs (reference/questioned) (Equation 1).

$$D = \sum_{i=1}^{n} Comparison(K_i, Q) \tag{1}$$

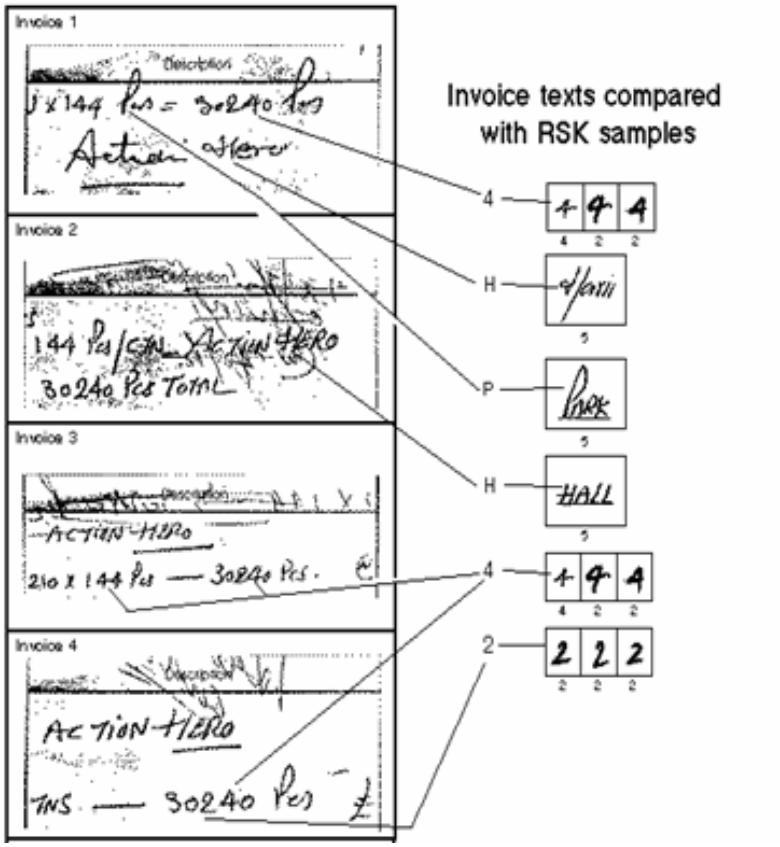


Figure 1. An example of forensic handwriting analysis based on Questioned Document Examination approach [17].

3 Writer Identification and SVM

The writer identification based on computer method is divided on two approaches, global and personal, depending on the model used [5]. The personal approach uses one model by author, while the global approach uses a general model for all authors. The personal model usually needs a large set of genuine samples to generate a robust model. It has the advantage of modeling appropriately the author's intrapersonal variability. The global model usually needs a reduced number of samples to training the model and excuse new training process when a new author is included in the database. But has the disadvantage of the generalization.

In the training process the W_1 class represents the genuine manuscripts class from the authors used for the training. W_2 class represents the manuscripts class belonging to other authors. The global model is then used for the comparison with the unknown or questioned manuscript.

Support Vector Machine [6] is a statistical learning technique based on the Structural Risk Minimization principle (SRM). The SRM induction principle has two objectives. The first one is controlling the empiric risk in the training set. The second one is controlling the capacity of the decision function used for obtaining this risk value. The linear SVM decision function is described by a weight vector $\vec{w}$, a threshold b and an output pattern $\vec{x}$ (Equation 2).

$$f(\vec{x}) = sign(\vec{w}.\vec{x} + b) \tag{2}$$

Given a set of training vectors S_1 (Equation 3) belonging to two separable classes, W_1 ($yi = +1$) and W_2 ($yi = -1$), the SVM finds the hyperplane with the maximum Euclidian distance from the training set. According to the SRM principle there'll be just one hyperplane with the maximum margin δ defined as the sum of the distances from the hyperplane to the closest point of the classes. This linear classifier threshold is the hyperplane optimal separation (Figure 2).

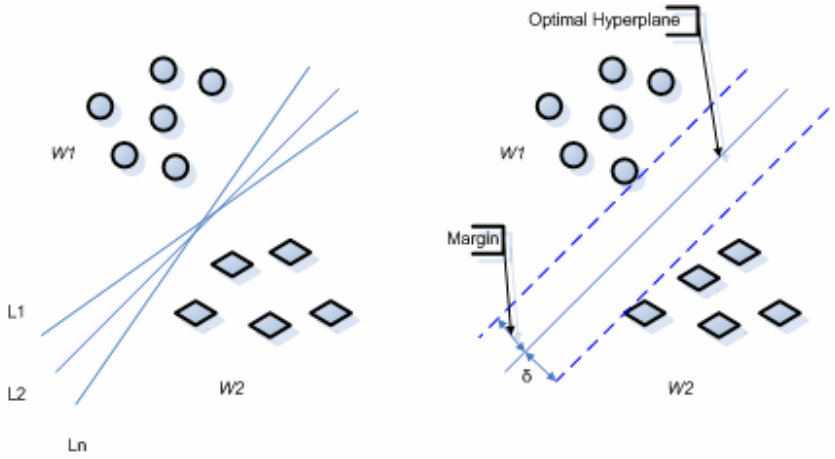
$$S_l = ((\vec{x}_1, y_1),\mathrm{K}\ (\vec{x}_l, y_l)),\ \vec{x}_i \in \Re^n,\ y_i \in \{-1,+1\} \tag{3}$$

In case of non-separable training sets, the ith point has a variable ξ_i, which represents the magnitude of the classification error. The penalty function $f(\xi)$ represents the sum of the classification errors (Equation 4).

$$f(\xi) = \sum_{i=1}^{l} \xi_i \tag{4}$$

The SVM solution can be found through minimization of the error in the training set (Equation 5).

$$\min_{\vec{w},b,\xi} = \frac{1}{2}\vec{w}.\vec{w} + C\sum_{i=1}^{n}\xi_i, \tag{5}$$



(a) (b)

Figure 2. Classification between two classes W_1 and W_2 using hyperplanes: (a) arbitrary hyperplanes li and (b) hyperplane with optimal separation, maximum margin.

The bibliography presents many kernels possibilities to the SVM in applications involving the pattern recognition, [7], [8] and [9]. In this initial study only the linear kernel was used (Equation 6).

$$K(\vec{x}, \vec{y}) = (\vec{x}.\vec{y}) \tag{6}$$

4 Forensic Letters Database

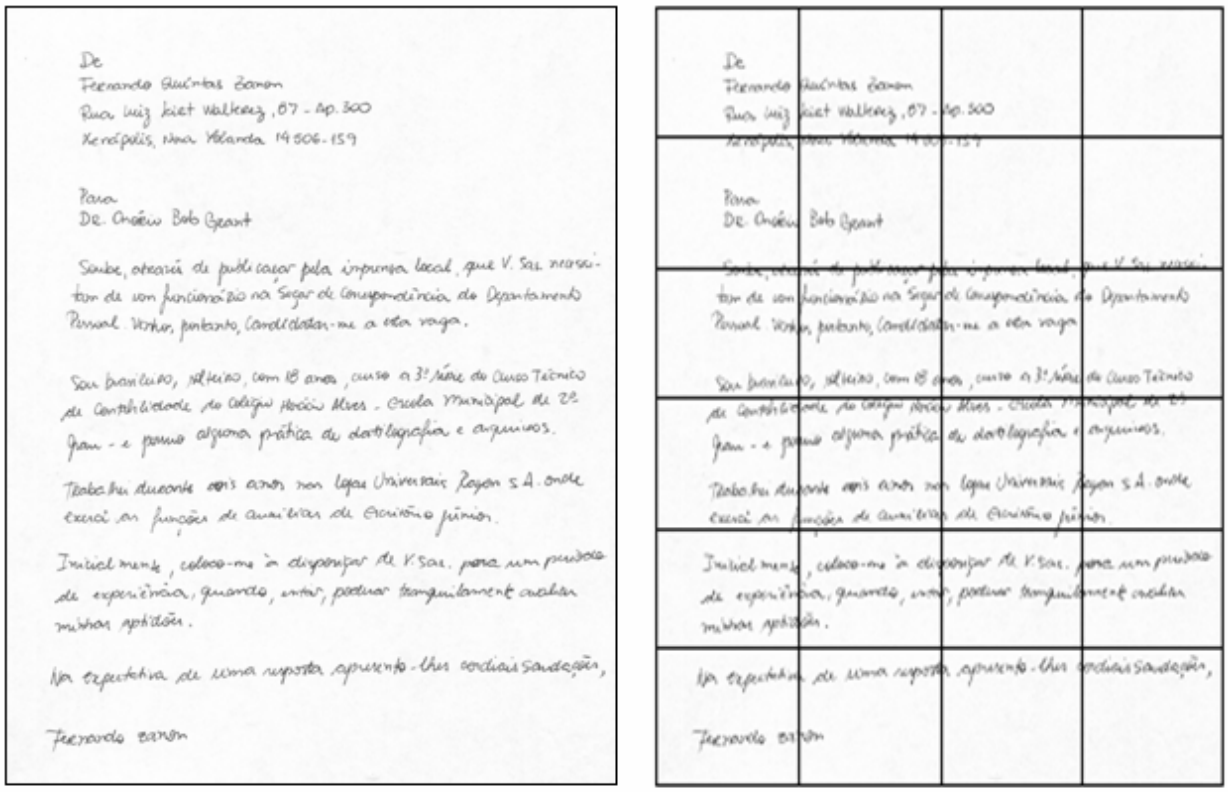
There are several models of forensic letters used by the literature [10]. The firs set of letters are composed of the traditional forensic letters used by experts in a visual analysis (16 class latter, Egypt latter and others) The second one are composed of letters for computer applications (CDAR letter) [10]. But all of then were wrote in English. The PUC forensic letter is a pattern text wrote in Portuguese, instead. The text contains all Latin alphabet letters in uppercases and lowercases, minimal graphics (ã, ç, é, à, etc.) and numbers between 0(zero) and 9 (nine), in different combinations [4].

The forensic letters database is composed of 945 samples, three letter samples from 315 distinct authors (Figure 3a). Their images were digitized in 300 dpi and 256 gray levels. Differently of Cha local method [10], where was used words and characters manual segmentations, each letter was automatically subdivided into 24 regular fragments. The fragments were used to extract a global feature to compose the training, reference, and testing databases (Figure 3b). The fragments without text were removed.

The first sample letter was used for training the model. For this purpose a set of 3 fragments per writer were used. The second sample was used for reference, i.e., during the comparison with the questioned fragment (Figure 5). The third letter sample was used for the test. The test database was composed of 5 fragments.

For the experiment two different databases size were used, for training, testing, and references. The first one was composed of 50 authors for training and 256 for testing/references. The second one was composed of 200 authors for training and 115 for testing/references.

The training database is composed of two classes, one of them from the same author W_1 and the other one from different authors W_2.



(a) (b)

Figure 3. (a) Example of the PUC forensic letter; (b) Example of segmentation.

5 Proposed Method

The proposed method is based on the questioned document examiner's approach and the general model (Figure 3). The proposed method was submitted to a set of phases described on the sequence.

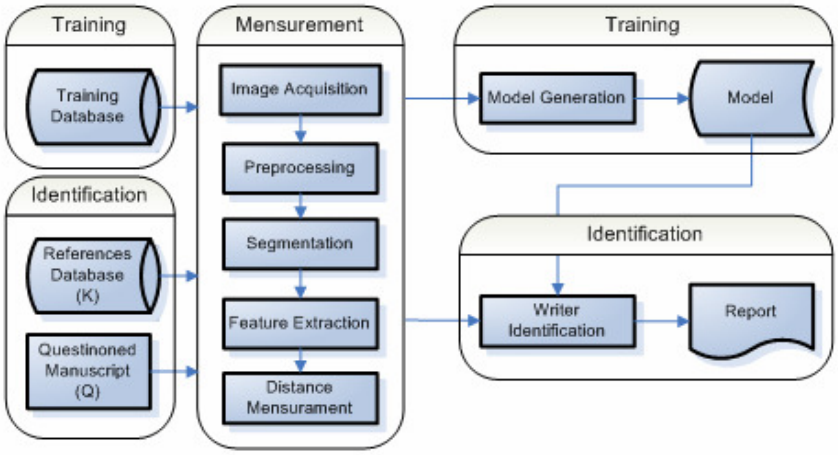
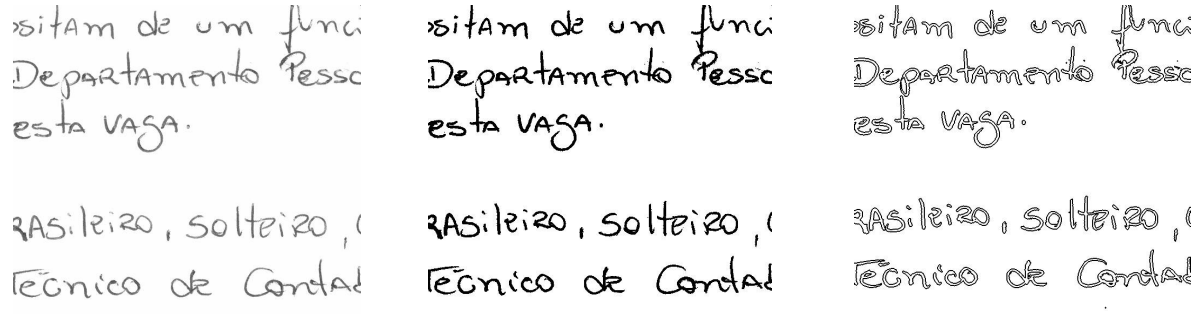


Figure 4. The writer identification framework for the global model.

5.1 Preprocessing

The 256 gray scale images from the manuscript segments were converted in binary images [11] (Figure 5b). The trace contours or borders were obtained through the morphological filters application [12] (Figure 5c).



(a) (b) (c)

Figure 5. Manuscript fragment example: (a) 256 gray scale image; (b) Binary image. (c) Image contour's.

5.2 Feature Extraction

For this first experiment only one dynamic feature was used, the axial slant. This graphometric feature is very usable by the questioned document examiners and has been extensively used automated writer identification [13] and [14]. The process consists of going through the image considering the border pixel in the center of a square structural element. After that, all the directional gradients L (angles Θ) are verified, starting from this central pixel and checking the following pixels finishing in the structural element extremities, only if there's the presence of an entire border fragment. That is, if all the neighboring pixels are black, the border is considered calculating the fragment position in a position vector for the histogram construction. This directional gradient vector is finally normalized by the probability distribution $P(\theta)$, which is the probability of finding in the image an orientated border fragment in an angle θ in relation to the horizontal axis.

For the proposed method different structural elements with k = 3, 4 and 5 pixels were tested. Each one generated respectively L = 9, 13 and 17 directional gradient angles Θ. The best result was obtained using $\{k = 5, L = 17\}$ (Figure 6). Figure 7 shows examples of different axial slants and the probability distribution $P(\theta)$.

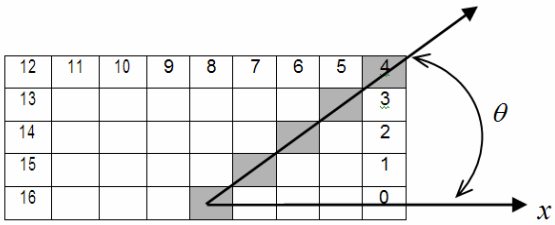
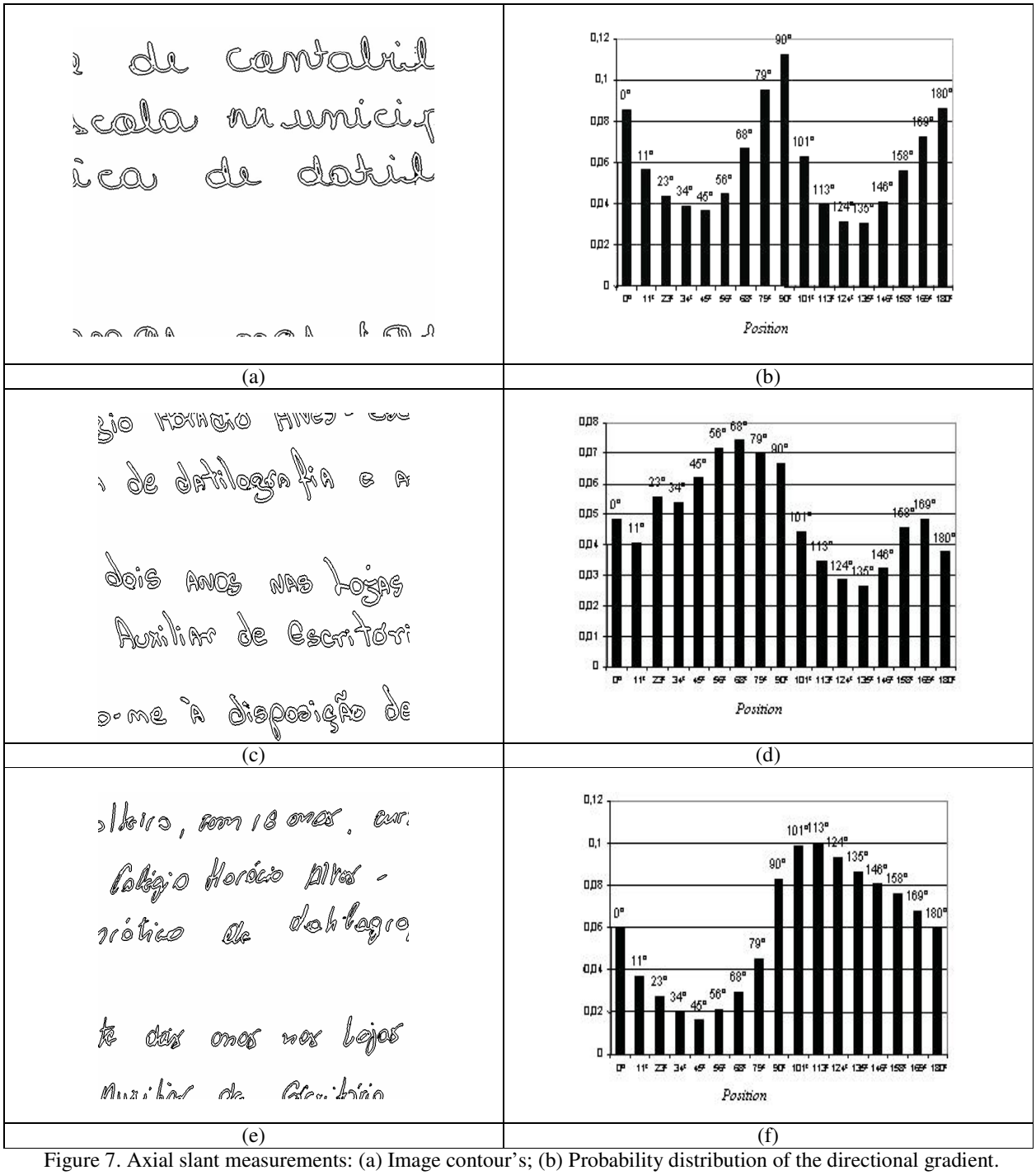


Figure 6. Example of structural element with k = 5 and consequently L = 17.



(a) (b) (c) (d) (e) (f)

Figure 7. Axial slant measurements: (a) Image contour's; (b) Probability distribution of the directional gradient.

5.3 Distances Measure Between Features

The fragments databases were converted in feature vectors, [10]. The F feature vectors of size L (L represents the number of directional gradient angles) are extracted from the manuscript fragments form testing, training and references (Equations 6).

$$F = (f_1 , f_2 , \Lambda \ f_L) \tag{6}$$

The Euclidian distances vector *D* between two fragments *A* and *B* is calculated for the training and testing procedures (Equation 7).

$$D\{A,B\}=\sqrt{(F_A-F_B)^2} \tag{7}$$

5.4 Comparison

The comparison process is composed of two phases, the training and the verification ones. In the training stage, the distance measures between the features *D* are calculated between fragment pairs. When two fragments belong to the same author the feature vector is indicated with 1 (association). When two fragments belong to different authors the feature vector is indicated with -1 (dissociation). The distance between two manuscript fragments is considered small when the samples belong to the same author. The SVM is then trained to separate small distances between features (association) and big distances between features (dissociation).

In the verification stage occur a set of comparisons, between each reference fragment sample and the questioned manuscript fragment sample. The SVM uses the general model obtained in the training phase to separate small distances between features (association) and big distances between features (dissociation).

5.5 Decision Comparisons

Usually, in an exam, the expert uses a set of manuscript sample with reference or models from known author. Each known sample, belonging to the reference set (4 to 10 samples), is compared with the questioned or unknown authorship sample. In this experiment a set of 5 reference samples was used for each author.

With the objective of producing the final decision, the proposed method classifies the solutions in an amount process M. This last stage represents the expert's final decision or report.

6 Experimental Results

Table 1 shows the obtained results using linear kernel. For the first experiment 50 authors were selected for the training and 215 for the tests. For the second one, 200 authors were selected for the training and 115 for the tests. The results demonstrated the discriminative capacity of the graphometric feature (axial slant) even being used in a global approach, in both cases. When the number of training set samples increases, the false acceptance (type II error rate) also increases and the false rejection goes down (type I error rate). This phenomenon occurs because the model is absorbing interpersonal variability [15] and makes the model more flexible. This effect increases the acceptance error rate. A possible solution for this phenomenon is to add other graphometric features.

The segmentation simplicity of the proposed method demonstrates the usability in a real forensic application. Like fingerprint, in writer identification large databases are commonly used, which makes it very important to rely on and simple fast algorithms. The global model is strongly compared to the personal model because, no additional training is necessary. Other important point that differs our method from the others in the literature [10], [13] and [14] is the forensic science framework. This framework is accepted by the Law Enforcement and Justice Area.

LinearKernel	**False Rejection (Type I Error)(%)**	**False Acceptance (Type II Error)(%)**	**Total Error Rate (%)**
50 Tr / 265 Ts	2.45	7.92	10.37
200 Tr / 115 Ts	1.73	10.87	12,60
Tr – training		Ts – testing	

Table 1. Comparative results using different number of training and testing set.

7 Conclusion and Future Works

This paper presented a method to writer identification base on Forensic Science approach. For this purpose only an axial slant feature are used and just two classes (association and dissociation). The adopted global model has demonstrated itself as being promising in the sample number reduction by used author in the model training and in the elimination of a new training need when including new authors.

For future works there's the inclusion of other graphometric features allowing the model to absorb more adequately the intrapersonal and interpersonal variability and provide the error rate reduction of the false acceptance.

References

[1] Morris, N. “Forensic Handwriting Identification Fundamental Concepts and Principles”, Academic Press, (2000), p. 238.

[2] Dines, J. E., “Document Examiner Textbook”, Pantex Intl Ltd, (1998), p. 566.

[3] Santos, C. R., Justino, E. J. R., Bortolozzi, F. Sabourin, R. “ An Off-Line Signature Verification Method based on the Questioned Document Expert’s Approach and a Neural Network Classifier”, In: The Ninth International Workshop on Frontiers in Handwriting Recognition, Tokyo, (2004), 10-14p.

[4] Justino, E. J. R. “A Análise de Documentos Questionados”, Monografia para concurso de professor titular, Pontifícia Universidade Católica do Paraná, (2002), 74p.

[5] Justino, E. J. R., Bortolozzi, F., Sabourin R. “An Off-line Signature Verification Method Based on SVM Classifier and Graphometric Features”, The 5th International Conference on Advances in Pattern Recognition, Calcutta , India, (2003), 200-204p.

[6] Vapnik, V. “Statistical Learning Theory”, Wiley, N. Y, (1998).

[7] Burges, C. J. C., “A Tutorial on Support Vector Machines for Pattern Recognition”, Data Mining and Knowledge Discovery 2, (1998), 121-167p.

[8] Müller, K., Mika, S., Rätsch, G., Tsuda, K. and Schölkopf, B. “An Introduction to Kernel-Based Learning Algoritms”, IEEE Transactions on Neural Networks, Vol. 12, No. 2, (March, 2001), 181-202p.

[9] Joachims, T., “Optimizing Search Engines Using Clickthrough Data”, ACM Conference on Knowledge Discovery and Mining (KDD), (2002), 1-10p.

[10] Cha, S. H., “Use of the Distance Measures in Handwriting Analysis. Doctor Theses. State University of New York at Buffalo, EUA, (2001), p. 208.

[11] Abutaleb, A. S., “Automatic Thresholding of Gray-level Pictures using Two Dimensional Entropy”, Computers Graphics & Image Processing, no. 47, (1989), 22-32p.

[12] Gonzalez, R. C., Woods, R. E., “Digital Image Processing”, Addison-Wesley Publishing Company,(1992).

[13] Crettez, J. P. “A set of handwriting families: style recognition”, In Proc. of the 3th. International Conf. on Document Analysis and Recognition, pages 489-494, Montreal, (August 1995). IEEE Computer Society Press.

[14] Bulacu, M. , Shomaker, L. “Writer Identification Using Edge-Based Directional Features”, Proc. Of 7th Int. Conf. on Document Analysis and Recognition (ICDAR 2003), IEEE Computer Society, (1993), pp. 937-941, vol. II, 3-6 August, Edinburgh, Scotland.

[15] Justino, E. J. R., Bortolozzi, F., Saborin, R. The Interpersonal and Intrapersonal Variability Influences on Off-Line Signature Verification In: XV Brazilian Symposium on Computer Graphics and Image Processing, 2002, Fortaleza, Proc. of 15th Brazilian Symposium on Computer Graphics and Image Processing. Los Alamitos: IEEE Computer Society, (2002). v.1. p 197-201.

[16] Gobineau H. and Perron R. “G´en´etique de l´ecriture et ´etude de la personnalit´e: Essais de graphometrie”, Delachaux & Niestl´e, (1954).

[17] James E. Doyle, State of Wisconsin Department of Justice, www.doj.state.wi.us/index.asp, (2006).