Pattern Recognition 43 (2010) 387–396

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

Reducing forgeries in writer-independent off-line signature verification through ensemble of classifiers

D. Bertolini[a], L.S. Oliveira[b,*], E. Justino[a], R. Sabourin[c]

[a]Pontifical Catholic University of Parana (PUCPR), R. Imaculada Conceição, 1155, Curitiba 80215-901, PR, Brazil
[b]Federal University of Parana (UFPR), Department of Informatics, Rua Cel. Francisco Heráclito dos Santos, 100, Curitiba, PR, Brazil
[c]Ecole de Technologie Superieure 1100 rue Notre Dame Ouest, Montreal, Quebec, Canada

ARTICLE INFO

Article history:
Received 13 June 2008
Received in revised form 9 February 2009
Accepted 7 May 2009

Keywords:
Signature verification
Graphometrics
Forgeries

ABSTRACT

In this work we address two important issues of off-line signature verification. The first one regards feature extraction. We introduce a new graphometric feature set that considers the curvature of the most important segments, perceptually speaking, of the signature. The idea is to simulate the shape of the signature by using Bezier curves and then extract features from these curves. The second important aspect is the use of an ensemble of classifiers based on graphometric features to improve the reliability of the classification, hence reducing the false acceptance. The ensemble was built using a standard genetic algorithm and different fitness functions were assessed to drive the search. Two different scenarios were considered in our experiments. In the former, we assume that only genuine signatures and random forgeries are available to guide the search. In the latter, on the other hand, we assume that simple and simulated forgeries also are available during the optimization of the ensemble. The pool of base classifiers is trained using only genuine signatures and random forgeries. Thorough experiments were conduct on a database composed of 100 writers and the results compare favorably.

1. Introduction

The main objective of a signature verification system is to exploit the singular and personal character of writing [15,28]. This kind of system should verify that what has been signed corresponds to the unique characteristics of an individual. A failure in this case is referred as type I error (false rejection), i.e., rejecting a genuine signature. Besides, the system should cope with a more challenging problem, i.e., avoiding the acceptance of forgeries as being authentic. The second error is referred as type II error (false acceptance).

The signature verification problem can be categorized into on-line and off-line. In general, on-line systems achieve better performance since they can count on dynamic features such as, time, pressure, and speed, which can be easily obtained from the on-line mediums [23]. Off-line systems are difficult to design as many desirable characteristics such as the order of strokes, velocity, and other dynamic information are not available during off-line image acquisition. The verification process has to rely only on features that can be extracted from the trace of the static signature image [11].

To deal with the problem of off-line signature verification, researchers have investigated two different approaches: writer-dependent and writer-independent [33]. The former is the standard approach for signature verification, where a specific model is built for each writer. In this context, hidden Markov models have been successfully applied [29,6,17]. Still in the same vein, different machine learning models have been tried out, such as neural networks [1], distance classifier [7,14] and support vector machines (SVM) [26]. In these cases, some samples of a given writer are used to model the genuine class and some samples of other writers, chosen randomly, are used to model the forgery class.

The forgeries usually are divided into three different subsets (random, simple, and simulated forgeries). The random forgery is usually a genuine signature sample belonging to a different writer, one who is not necessarily enroled in the signature verification system. The simple forgery occurs when the forger knows the writer's name, but has no access to a sample of the signature. Thus, the forger reproduces the signature in his own style. Finally, the simulated forgery is a reasonable imitation of the genuine signature model. Fig. 1 depicts some examples of these forgeries.

The main drawbacks of the writer-dependent approach are the need of learning the model each time a new writer should be included in the system and the great number of genuine samples necessary to build a reliable model. In real applications, usually a limited

* Corresponding author.
E-mail address: lesoliveira@inf.ufpr.br (L.S. Oliveira).

0031-3203/$ - see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.patcog.2009.05.009

Fig. 1. Examples of the signature: (a) genuine, (b) simple forgery, and (c) simulated forgery.

number of signatures per writer is available to train a classifier for signature verification, which leads the class statistics estimation errors to be significant, hence, resulting in unsatisfactory verification performance. To surpass this problem, some writers generate more data through transformations of the genuine signatures [13].

An alternative to the writer-dependent approach is the writer-independent, which models the probability distributions of within-class and between-class similarities. These distributions are used to determine the likelihood of whether a questioned signature is authentic or forgery. The concept of similarity/dissimilarity representation for pattern recognition was introduced by Pekalska and Duin [27] and the seminal work using this concept in the field of author identification was presented by Cha and Srihari [5]. Later, Santos et al. [32] use the idea of dissimilarity representation for signature verification. The main benefit provided by this approach is the possibility of reducing an n-class pattern recognition problem to a 2-class problem, in the case of signature verification, genuine and forgery.

This work takes into account the framework initially proposed by Santos et al. [32]. It is based on a forensic document examination approach and can be defined as writer-independent approach as the number of models does not depend on the number of writers. In this vein, it is a global model by nature, which reduces the pattern recognition problem to a 2-class problem, hence, makes it possible to build robust signature verification systems even when few signatures per writer are available. It also applies the ideas of dissimilarity representation introduced by Pekalska and Duin [27] and support vector machines as classifiers.

An important aspect in signature verification that is very often neglected is the class distribution. A tacit assumption in the use of recognition rate as an evaluation metric is that the class distribution among examples is constant and relatively balanced. In signature verification this is rarely the case. Usually one has few genuine signatures and a bunch of random forgeries (signatures from other writers) to train a model. In this context, ROC (receiver operating characteristic) curves are attractive due to its property of being insensitive to changes in class distribution. If the proportion of positive to negative instances changes in a test set, the ROC curves will not change [8]. It is worth of emphasis that simple and simulated forgeries cannot be used neither for training nor for the estimation of decision thresholds.

The contribution of this paper is twofold. First, we introduce a new graphometric feature set that takes into account the curvature of the most important segments, perceptually speaking, of the signature. It simulates the shape of the signature by using Bezier curves. The features are then extracted from these curves.

Second, we propose an ensemble of classifiers to improve the resistance of the signature verification system against forgeries. The pool of base classifiers contains classifiers trained with four different graphometric feature sets, which were trained using just genuine samples and random forgeries. The ensemble was built using a standard genetic algorithm and two different validation sets were used to drive the search. In the first case, the validation set contains only genuine samples and random forgeries while in the second case simple and simulated forgeries are also available. The writers used to build the validation set are not the same used for training. We show that if forgeries become available even for writers who did not contributed to the training set, then the system can be fine-tuned and the best classifiers selected to compose the ensemble. Three different fitness functions were assessed in our experiments. The first is the minimization of the overall error rate of the ensemble. The other two objective functions are derived from the ROC, namely, the maximization of the AUC (area under the curve) and the maximization of TPR (true positive rate) for a FPR (false positive rate) for a fixed threshold.

Besides, we evince the usefulness of the dissimilarity representation for signatures verification, which enables us to convert a n-class problem to a more general 2-class one. Hence, SVMs were used as base classifiers since they are suitable to deal with binary classification problems. In this context, an analysis on the size of the reference set is also presented. The results show that in some cases, after a certain number of references, the false acceptance cannot be further reduced. Through a set of comprehensive experiments on a database composed of 100 writers, we demonstrate that the proposed approach can reduce considerably the false acceptance rate while keeping the false rejection at acceptable levels. Moreover, it compares favorably to other combination strategies reported in the literature.

The remaining of the paper is organized as follows: Section 2 describes how the writer-independent approach works. Section 3 presents the database used in this work. Section 4 introduces all the graphometric feature sets used in this works and Section 5 discusses important issues of ensemble of classifiers for signature verification. Finally, Section 6 reports the experiments we have made and Section 7 concludes this work.

2. Writer-independent and dissimilarity

The idea of the writer-independent approach is to classify a handwriting sample into genuine or forgery. The approach used in this work is the one employed by forensic experts, who compare the questioned samples with some references to assert whether a piece of handwriting is genuine or forgery. During this comparison, the experts extract different features to compute the level of similarity between the samples being compared.

The concepts of similarity, dissimilarity, and proximity have been discussed in the literature from different perspectives [31,10,22,27]. Pekalska and Duin [27] introduce the idea of representing the relations between objects through dissimilarity, which they call dissimilarity representation. This concept describes each object by its dissimilarities to a set of prototype objects, called the representation set R. Each object x is represented by a vector of dissimilarities $D(x,R)=[d(x,p_1),d(x,p_2),\ldots,d(x,p_n)]$ to the objects $P_i \in R$.

Let R be a representation set composed of n objects. A training set T of m objects is represented as the $m \times n$ dissimilarity matrix $D(T,R)$. In this context, the usual way of classifying a new object x represented by $D(x,R)$ is by using the nearest neighbor rule. The object x is classified into the class of its nearest neighbor, that is the class of the representation object p_i given by $d(x,p_i)=\min_{p\in R}D(x,R)$. In another approach, each dimension corresponds to a dissimilarity $D(\cdot,p_i)$ to an object p_i. Hence, the dimensions convey a homogeneous type of information. The key here is that the dissimilarities should be small for similar objects (belonging to the same class) and large for distinct objects. In this way, $D(\cdot,p_i)$ can be interpreted as an attribute.

The concept of dissimilarity turns out to be very interesting when a feasible feature-based description of objects might be difficult to obtain or inefficient for learning purposes, e.g., when experts cannot define features in a straightforward way, when data are high

dimensional, or when features consist of both continuous and categorical variables [27]. In the case of signature verification, however, several different features have been proposed so that intra- and extra-class variation can be modeled.

In light of this, in this work we propose to combine feature-based description with the concept of dissimilarity. The idea is to extract the feature vectors from both questioned and reference signatures and then compute what we call the dissimilarity feature vector. If both samples come from the same writer (genuine), then all the components of such a vector should be close to 0, otherwise (forgery), the components should be far from 0.

To implement this, we use a reference set of n genuine signature samples Sk_i $(i=1,2,3,\ldots,n)$ and then compare each Sk with a questioned sample Sq. Let V_i be the graphometric features extracted from the reference signatures and Q the graphometric features extracted from the questioned signatures. Then, the dissimilarity feature vectors $Z_i=|V_i-Q|$ are computed to train the classifiers C_i, which provide a partial decision. The final decision D depends on the fusion of these partial decisions, which are usually obtained through the majority vote rule. Fig. 2 depicts the global approach.

3. Database

The signature database used in this work is composed of 100 writers and it has been divided into 40, 20, and 40 for training, validation, and testing, respectively.

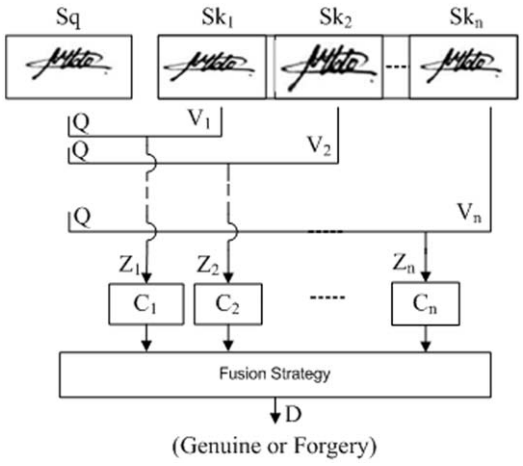

Fig. 2. Architecture of the global approach.

The signatures were provided by undergraduate students in four different sessions, once a week during one month. The signatures were collected on an A4 white sheet of paper with no overlap and then scanned in gray level with 600 dpi. Ten genuine signatures were collected from each writer in each session. Regarding the forgeries, 10 people with no experience in making forgeries were selected as forgers. Simple forgeries were produced informing to the forger only the name of the writer. Simulated forgeries were produced showing to the forger four genuine signatures of the writer. Taking into account that we have forgeries for 60 writers and 20 forgeries per writer (10 simple and 10 simulated), each forger produced 120 forgeries (60 simple and 60 simulated) in average.

Considering the dissimilarity-based approach adopted in this work, the classifiers should be trained to discriminate between genuine (positive) and forgeries (negative). To generate the positive samples, we have computed the dissimilarity vectors among four genuine samples of each writer, which results into six different combinations. Considering 40 writers of the training set, it sums up to 240 positive samples. Fig. 3 exemplifies this process.

To build the negative samples, we consider the first two samples of the first 36 writers of the training set. Then, we compute the dissimilarities among these two signatures and two signatures of four different writers selected randomly. It totalizes 288 negative samples. An example of this process is illustrated in Fig. 4. These 528 samples make our dissimilarity-based training set, which is used to train an SVM classifier. The training is done using k-fold cross validation. Different values for k have been tried out but $k=10$ produced the best results.

In the dissimilarity approach, the number of positive and negative samples in the testing set depends directly on the number of references discussed in Section 2. Consider for example that five genuine signatures from a given writer that did not contribute to the training set are available as references. In this case, the dissimilarity vectors among a given testing sample (questioned signature) and the five references are computed, yielding five dissimilarities feature vectors. In our case, the testing set contains 40 writers with 40 samples per writer, which are divided into 10 genuine signatures, 10 random forgeries (selected randomly), 10 simple forgeries, and 10 simulated forgeries. The number of references is subject of investigation and in our experiments it ranges from 3 to 15.

Two different validation sets devoted to compute the fitness during the search are available. The former is composed of 20 writers and 20 samples per writer, 10 genuine signatures and 10 random forgeries (selected randomly). The latter contains the same 20 writers but with 40 samples per writer, 10 genuine

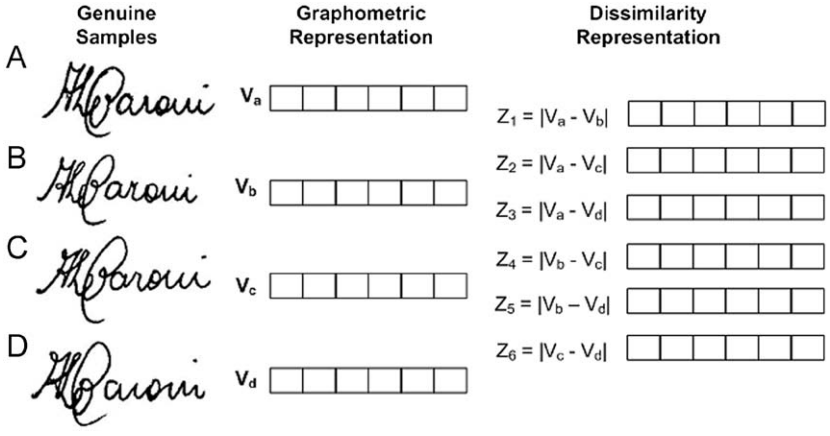

Fig. 3. Dissimilarities among genuine samples of the same writer to generate the positive samples. From four genuine samples, six dissimilarity vectors are created.

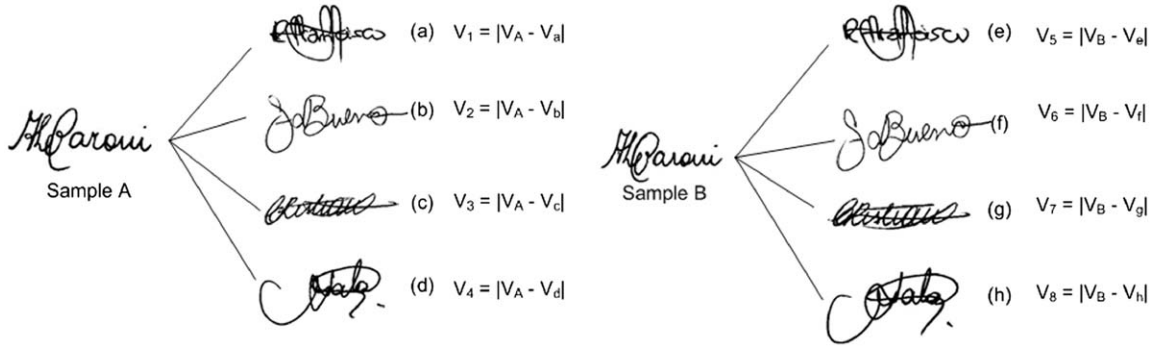

Fig. 4. Dissimilarities among genuine samples from different writers to generate the negative samples.

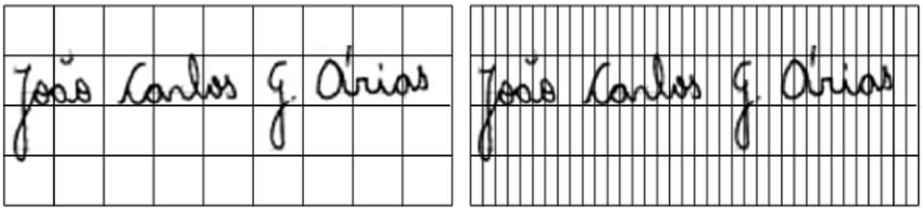

Fig. 5. Example of two different configurations of the grid used for feature extraction.

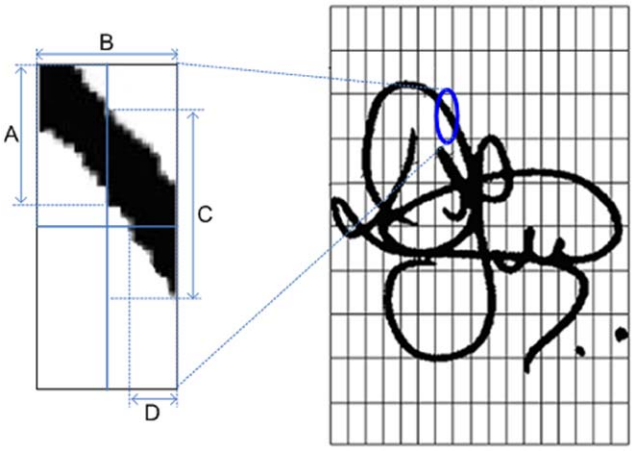

Fig. 6. An example of distribution of pixels. It uses a different number of cells to better illustrate the process.

signatures, 10 random forgeries (selected randomly), 10 simple forgeries and 10 simulated forgeries. Alike the testing set, the number of positive and negative samples depends on the size of the reference set.

Note that these samples are used just to guide the search of the optimization algorithm, i.e., the base classifiers were trained using genuine and random forgeries only. As stated before, our argument is that if some simulated and simple forgeries become available, we can use them to select the best classifiers to build the ensemble, without retraining the base classifiers.

4. Feature sets

In this section we present the feature sets considered to build the pool of base classifiers used to generate the ensembles. All of them are grid-based, i.e., the image of size 400×1000 ($H \times W$) is segmented using a grid and then the features are computed for each cell of the grid. Fig. 5 shows two different grid configurations on the same signature. To make this paper self-contained, we describe the four characteristics used to train the classifiers, namely, density, slant, distribution, and curvature. The first three have been applied to signature verification with relative success [18], while the latter is a new feature set introduced in this work.

The density is what we call apparent pressure, since it describes the width of the strokes. To extract this feature set we put a grid over the image and count the number of black pixels in each cell. To compute the slant we have applied the concept presented by Hunt and Qi [14], which determines the slant in two steps. First, a global slant is computed over the entire image and then the slant for each cell is computed as well. In this way, each cell has a slant value and the final global value is the most frequent value among the cells of the segmentation grid. The distribution of pixels is based on four measures as depicted in Fig. 6. In this case each cell is delimited by two projections, vertical, and horizontal. Then, the height and

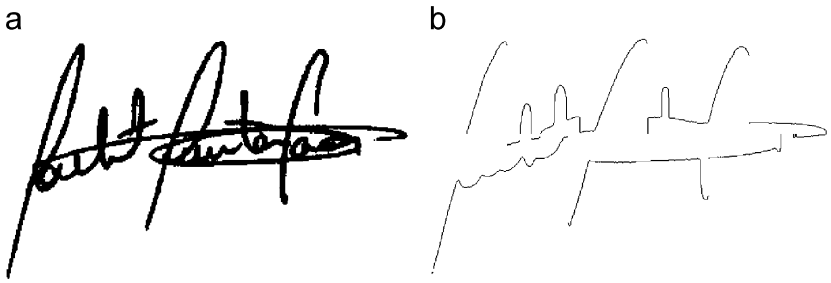

Fig. 7. (a) Original signature, and (b) signature profile.

width of the stroke are computed in four directions limited to these projections. These values are exemplified by the letters A, B, C, and D in Fig. 6, which represent the height of the left part of the stroke, the width of the upper part of the stroke, the height of the right part of the stroke, and the width of the lower part of the stroke, respectively. A more complex approach, but based on the same idea was proposed by Sabourin et al. [30].

Finally, the last feature set tries to capture the information about the curvature of the most important segments of the signature. In order to do that, we try to reproduce these segments using the well-known cubic Bezier curves [24], which are defined by four points: two endpoints (origin and destination) and two control points. To reduce the complexity of this task, first the image of the signature is thinned and then the upper and lower profiles are extracted. Only the longest segment of each cell is considered for feature extraction, which is detected as follows: First, intersection and terminal points are detected in the thinned image. All the paths between two different terminal points, two different intersection points, or the path between a terminal point and an intersection point are considered as independent segments. Then, for each segment three equidistant points (N_i) are defined. Fig. 7 shows a signature and its respective profile.

For each point N_i $\{i=1,2,3\}$, we compute θ and the control points (Pl_i and Ph_i) using Eqs. (1) and (2), respectively.

$$\theta = \arctan \frac{y_{N_{i-1}} - y_{N_{i+1}}}{x_{N_{i-1}} - x_{N_{i+1}}} \tag{1}$$

$$\begin{cases} Pl_i(x) = N_i(x) + \cos(\theta) \times dist(N_i, N_{i-1}) \\ Pl_i(y) = N_i(y) + \sin(\theta) \times dist(N_i, N_{i-1}) \\ Ph_i(x) = N_i(x) + \cos(\theta) \times dist(N_i, N_{i-1}) \\ Ph_i(y) = N_i(y) + \sin(\theta) \times dist(N_i, N_{i-1}) \end{cases} \tag{2}$$

where *dist* stands for the Euclidean distance. Fig. 8a shows an example of the features computed for N_i, where d_{1_i} and d_{2_i} represent the Euclidean distance from N_i to the two control points. As we can notice from Fig. 8a, the bigger the distance, the lower the curvature of the stroke between the points. Summarizing, we extract three features for each point (θ, d_{1_i}, and d_{2_i}), which gives us nine features per cell of the grid. Fig. 8b shows an example of the points detected in a real segment extracted from the signature of the Fig. 7b.

As mentioned earlier, the size of the grid depends on the feature being used. In this context, we have extracted the feature sets considering different grid sizes and used all of them to build the ensemble. The decision of which classifier should be part of the ensemble will be provided by the search algorithm. We have considered the following 16 (Horizontal × Vertical) different variations for the grid: 4×5, 4×10, 4×20, 4×25, 5×5, 5×10, 5×20, 5×25, 8×5, 8×10, 8×20, 8×25, 10×5, 10×10, 10×20, and 10×25. For each different grid size, a classifier is trained with one of the four feature sets, resulting in 64 different classifiers.

Following the protocol introduced previously, these feature sets are extracted from the questioned (Sq) and reference (Sk) images

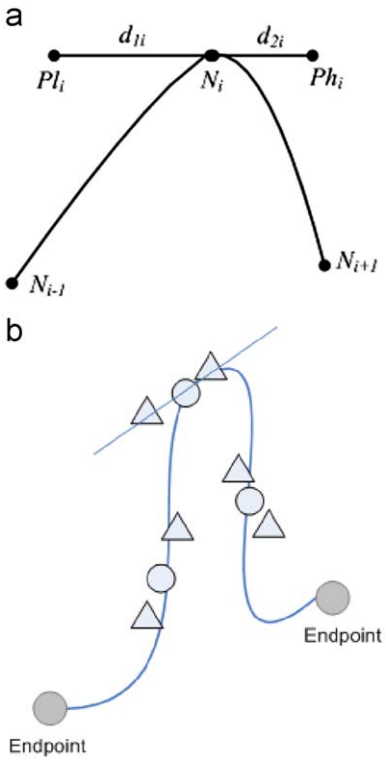

Fig. 8. (a) Example of the features extracted from a segment and (b) example of the points detected in a real segment extracted from the signature of the Fig. 7b

as well, which produce the aforementioned graphometric feature vectors V_i and Q. Once those vectors are generated, the next step consists in computing the dissimilarity feature vector $Z_i = |V_i - Q|$, which will be used to train the SVM classifiers. Considering that the reference set is composed of n images, the questioned image Sq will be compared n times, yielding n partial decisions. Then, the final decision can be based on any fusion rule, such as, majority, voting, max, min, product, etc. What we have observed from previous experiments [25] is that the Max rule achieves the best results for the database considered in this work.

In this work we report the performance of the system in terms of the overall error rate, which is given by Eq. (3). Given the test set discussed in the previous section, the a priori probability of a type I, type II_a (random), type II_b (simple), and type II_c (simulated) is 0.25:

$$\text{Overall Error} = \tfrac{1}{4} \times \text{type I} + \text{type II}_a + \text{type II}_b + \text{type II}_c \tag{3}$$

The overall error rates of the classifiers described above range from 9% to 25% on the testing set, considering five references. It is important to notice that the SVM has been trained with samples

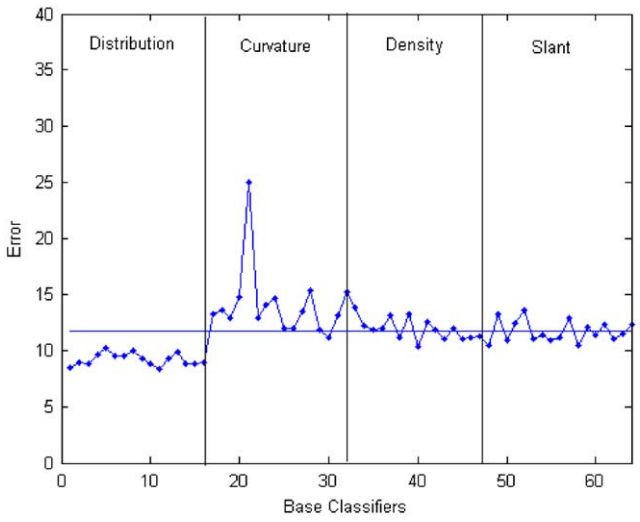

Fig. 9. Performance of the base classifiers.

Table 1
Best classifier of each feature set on the test set.

Features	Overall error	Type I error	Type II error		
			Simulated	Random	Simple
Distribution	8.42	18.83	7.50	3.66	3.66
Curvature	11.19	27.16	9.48	3.32	4.80
Density	10.35	25.32	7.80	3.80	4.48
Slant	10.48	17.48	10.00	6.80	7.64

coming from the 40 writers of the training set. The error rates depicted in Fig. 9 are computed based on 40 writers who did not contribute to the training of the writer independent classifier. As we can observe, the distribution feature set has the best classifiers while those trained with the curvature presents error rates above the average.

Table 1 reports the error rates separately for genuine and the three different classes of forgeries for the best classifier of each feature set, considering five references.

We could notice that the curvature feature set suffers more intensively with the intra-class variability, hence, it produces an higher type I error, as reported in Table 1. This drawback, however, is useful to avoid type II error, specially in the case of the simulated forgeries. By analyzing the results we realized that several forgeries not detected by the other three feature sets were found by the classifiers trained with the curvature feature set.

The main challenge of a signature verification system consists in minimizing as much as possible the type II error while keeping the type I error at acceptable levels. Two hypotheses have been done here: (1) the approach will generalize well for genuine signatures from unknown writer if the intra-class variability of "genuine signatures" is reasonably low and (2) the elimination of random forgeries (e.g., low false positive rate) results in a good detection of simple forgeries according to the nature of this class of forgeries. In the next sections we describe the efforts we have made in this direction using ensemble of classifiers.

5. Ensemble of classifiers for signature verification

Several studies have been published demonstrating the benefits of the combination paradigm over the individual classifier models [20]. During the last years, a considerable amount of research has gone into ensemble of classifiers. According to the literature, the most popular methods for ensembles creation are Bagging [2] and Boosting [9]. The effectiveness of such methods comes primarily from the diversity caused by re-sampling the training set while using the complete set of features to train the component classifiers. In addition, some attempts have been made to incorporate the diversity into ensemble creation methods by over-producing classifiers and then choosing some of them to compose the ensemble. The random subspace method (RMS) proposed by Ho in [12] was one early algorithm that construct an ensemble by varying the subset of features. An alternative to bring diversity to the ensemble is to combine classifiers trained with different feature sets. The efficiency of this strategy has been reported by several authors [34,19,21].

In this paper the underpinning concept adopted was the "overproduce and choose". The over-production of classifiers is achieved by varying the size of the segmentation grid during feature extraction. Based on our previous experience, this is an important parameter of segmentation-based signature verification systems. Therefore, each different configuration of the grid size is used to train a different classifier. As stated in Section 4, 16 different grid configurations were considered for each feature set. The "choose" step is performed by a genetic algorithm, which selects a subset of classifiers and combines them using a fusion rule. In this work, several fusion rules were tried out at this level, and the sum rule produced the best results in average.

When using search algorithms to build ensemble of classifiers, the most common fitness function is the minimization of the overall error rate or the maximization of some diversity measure. In the context of signature verification, however, minimizing the overall error rate may not be the ideal since the hypothesis that the class distribution among examples is constant and relatively balanced does not hold. In this context, ROC curves are attractive due to its property of being insensitive to changes in class distribution.

Two other fitness functions can be derived from the ROC computed on the validation set. The first one is the area under the ROC (AUC), which reduces the ROC performance to a single scalar value representing the expected performance. The AUC has an important statistical property: the AUC of a classifier is equivalent to the probability that the classifier will rank a randomly chosen positive instance higher than a randomly chosen negative instance [8]. The second function is the maximization of the true positive rate for a given false positive rate. The FPR usually is a constraint imposed by the application.

6. Experiments

In our experiments, the genetic algorithm is based on bit representation, one-point crossover, bit-flip mutation, and roulette wheel selection (with elitism). The following parameter setting were employed: population = 100, number of generations = 300, probability of crossover = 0.7, and probability of mutation = 0.03. In order to define the probabilities of crossover and mutation, we have used the one-max problem, which is probably the most frequently used test function in research on genetic algorithms because of its simplicity [4]. The population size and the number of generations were defined empirically.

Let $A = \{C_1, C_2, \ldots, C_p\}$ be the pool of p base classifiers and B a chromosome of size p of the population. The relationship between A and B is straightforward, i.e., the gene i of the chromosome B is represented by the classifier C_i from A. Thus, if a chromosome has all bits selected, all classifiers of A will be included in the ensemble. As presented earlier, in this work we have build a pool of 64 base classifiers ($p = 64$) by varying the size of the segmentation grid.

In order to find an ensemble of classifiers that brings a good generalization, we have compared three different fitness functions. The first one is the minimization of the overall error rate of the ensemble on the validation sets. In this case, the decision threshold in the ROC curve is the one that minimizes the overall error rate. The other two objective functions are derived from the ROC, namely,

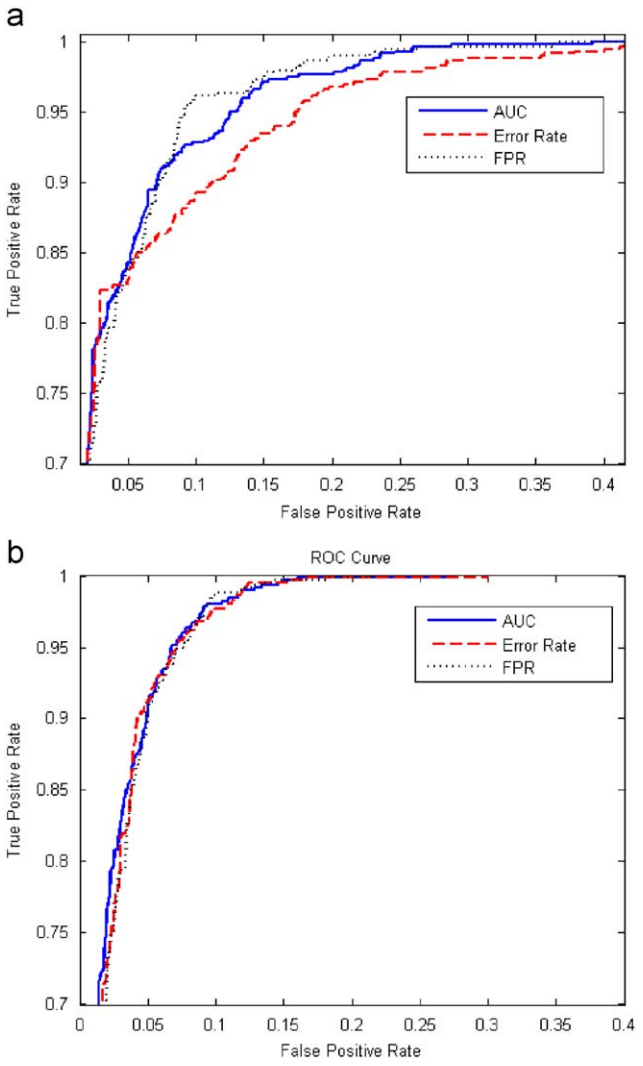

Fig. 10. Comparison among the three objective functions considered in this works. (a) $Sk = 5$ and (b) $Sk = 15$. The validation set II was considered here.

the maximization of the AUC and the maximization of TPR for a FPR fixed at 10%. As stated before, two different databases were used to assess the fitness. In the first one only genuine samples and random forgeries are available, while the second set contains genuine samples and all types of forgeries. In this way, we will be able to verify the impacts of having simple and simulated forgeries during the optimization process. It is worth of remark, though, that those forgeries were not used to train the classifiers.

As mentioned in Section 3, one aspect worth of investigation is the size of the reference set *Sk* during the search. In theory, the bigger the reference set, the bigger the intra-class variation the writer-independent model can absorb. In real applications, however, usually a limited number of genuine signatures per writer is available.

In light of this, seven experiments were performed using $Sk = \{3, 5, 7, 9, 11, 13, 15\}$. Taking into account that three different objective functions were assessed, it sums up to 21 different experiments. Each experiment was replicated 10 times to verify the reproducibility. Therefore, all the results presented here consider the average of these 10 replications.

6.1. Objective functions

Our first concern when analyzing the results of the experiments was the impact of the objective functions. As one can observe from Fig. 10, the smaller the size of the reference set, the bigger the impact of the objective function during optimization. From Fig. 10a, we can observe that the most homogeneous curve was produced by the ensemble that maximizes the AUC. It is also clear that the ROC produced by the maximization of the TPR for FPR fixed at 10% really improved the results for that operational point. The worst performance, on the other hand, was produced by the minimization of the overall error rate. This can be explained by the fact that the error rate is sensitive to changes in class distribution.

What we have observed is that as the size of the reference size increases the impacts of the objective function is minimized. Fig. 10b compares the three objective functions for $Sk = 15$, while the numerical results are reported in Table 2 It is important to mention that rarely such a number of references is available. Therefore, we believe that an objective function derived from the ROC is more suitable in the context of signature verification.

Table 2
Results on the testing set for ensemble tuned on the validation set II.

Objective function	*Sk*	Overall error	Type I error	Type II error		
				Simulated	Random	Simple
Error rate	3	8.06	14.32	8.64	4.48	4.80
	5	7.09	21.08	**3.80**	**2.00**	**1.48**
	7	6.65	12.48	6.48	3.64	4.00
	9	6.46	**7.32**	8.32	5.32	4.88
	11	6.74	9.16	8.32	5.00	4.48
	13	6.36	11.00	7.00	3.64	3.80
	15	5.90	8.32	7.16	3.80	4.32
AUC	3	7.12	16.32	6.16	3.00	3.00
	5	6.61	17.64	**4.64**	**2.00**	**2.16**
	7	6.03	11.16	6.48	3.16	3.32
	9	6.11	14.00	5.00	2.80	2.64
	11	6.36	14.00	6.00	2.80	2.64
	13	6.20	**9.16**	7.48	4.16	4.00
	15	5.65	10.16	6.48	3.16	2.80
FPR fixed (10%)	3	8.02	13.80	8.16	5.32	4.80
	5	7.57	18.00	**6.16**	**2.80**	**3.32**
	7	7.02	14.32	6.48	3.48	3.80
	9	6.81	13.80	6.00	3.64	3.80
	11	7.12	10.16	9.16	4.00	5.16
	13	7.03	10.16	9.32	4.16	4.48
	15	5.99	**9.00**	7.48	3.48	4.00

6.2. Size of the reference set (Sk)

One of the important issues of the approach adopted in this work is the size of the reference set and its impact on the reliability of the signature verification system, which can have different meanings depending on the application. It is clear that a reliable system is the one where both types I and II errors are reduced simultaneously. However, in general there is a trade-off between these types of errors and for this reason the concept of reliability can change depending on the application requirements. For example, for some applications, reducing type II error is much more important than reducing type I, and vice versa.

In this section we address the impacts of the size of the reference set used during the optimization. Here, we assume that a reliable system should reduce as much as possible type II error, i.e., be resistant against forgeries. However, the results reported in Tables 2 and 3 enable us to verify in which conditions the system is more resistant against type I error as well.

Regarding the resistance against forgeries, what is clear from these experiments is the importance of using simple and simulated forgeries during the ensemble optimization. The beauty of having a universal classifier based on dissimilarity is that if simulated and

simple forgeries become available, they can be used to find more reliable ensembles without retraining the classifiers. Table 2 shows that when the validation set II is used, the most reliable configuration against forgeries were reached when $S_k = 5$. On the other hand, if only random forgeries are available during the optimization (validation set I), the number of references should be considerably increased to achieve comparable rates. See Table 3.

As mentioned before, one could expect that increasing the number of references would consequently reduce the error rates. However, what we could observe is that the behavior reported in Table 3 is also related to the acquisition of the database. The signatures were collected during four sessions, where each writer provided 10 samples per section. In some cases, besides the normal intra-class variability, we could notice that the last signatures feature extra variability caused by the fatigue of the writers. Fig. 11 shows an example of this variability, where (a), (b), and (c) show the first genuine signature collected, the superposition of the first three signatures, and the superposition of the last three signatures, respectively. It is easy to observe the bigger intra-class variability in the end of the process. Consequently, adding more references did not help if they have such a large intra-class variability.

Very often, pattern classification systems impose some constraints in terms of FPR. In such cases, as depicted in Fig. 10a, could be interesting to optimize the ensemble taking into account such restrictions, even though the overall performance is poorer. If no constraint is presented, then AUC seems a good option due to its property of being insensitive to changes in class distribution. What happens in this case, is that the ensemble is composed of more restrictive classifiers, which are more sensitive to intra-class variations. That is to say, small variations on the signature are considered forgeries. On the other hand, the price to pay in this case is the higher type I error (false rejection).

Table 3
Results on the testing set for ensemble tuned on the validation set I.

Objective function	*Sk*	Overall error	Type I error	Type II error		
				Simulated	Random	Simple
Error rate	3	8.85	20.32	6.80	3.80	4.48
	5	8.81	16.16	7.64	4.80	6.64
	7	7.82	12.32	9.00	4.64	5.32
	9	7.15	10.80	7.64	5.16	5.00
	11	7.19	**9.64**	9.48	4.48	5.16
	13	7.54	17.00	7.00	**3.00**	3.16
	15	6.28	11.32	**6.48**	4.32	**3.00**
AUC	3	7.86	15.32	7.48	4.16	4.48
	5	7.32	11.32	8.00	4.48	5.48
	7	6.32	11.32	5.00	4.16	4.80
	9	7.04	10.00	7.00	5.00	6.16
	11	7.19	17.64	**4.32**	**3.32**	**3.48**
	13	6.73	**7.32**	7.80	5.32	6.48
	15	6.48	9.16	6.64	4.80	5.32
FPR fixed (10%)	3	7.99	7.64	10.00	7.16	7.16
	5	7.78	10.80	9.00	5.32	6.00
	7	7.28	13.00	7.16	4.64	4.32
	9	6.80	8.00	8.16	5.32	5.64
	11	6.88	**5.16**	9.48	6.64	6.16
	13	6.98	15.16	5.64	3.32	3.80
	15	6.34	13.64	**5.16**	**3.16**	**3.32**

If we compare the results produced by the ensembles with the best classifiers trained with the four feature sets presented in Table 1, we can assert that the ensembles are quite effective in mitigating all kind of forgeries, even though when only genuine signatures and random forgeries are used during the optimization. One aspect that is clear from our experiments is that increasing the size of the reference set does not necessarily reduces type I error, but it generally reduces the overall error rate.

The results reported here compare favorably to other combination strategies [25], where the error rates reported for simulated, random, and simple forgeries were 8.16%, 5.32%, and 4.48%, respectively, for five references.

6.3. Selected classifiers

Another important facet when discussing ensemble of classifiers consists in analyzing the selected classifiers. According to the theory, a good ensemble contains good classifiers (not necessarily excellent) that disagree as much as possible on difficult cases [20]. This behavior could be observed in our experiments as several classifiers with good performance were not selected and others below average were chosen to be part of the ensemble. Figs. 12 and 13 show the classifiers selected for the ensembles using validation sets I and II, respectively. In both figures, each cell represents one classifier trained with a different grid size. The order of the cells is as follows: 4×5, 4×10, 4×20, 4×25, 5×5, 5×10, 5×20, 5×25, 8×5, 8×10, 8×20, 8×25, 10×5, 10×10, 10×20, and 10×25. The three different objective functions are addressed in these figures.

According to Figs. 12 and 13, the classifiers trained with the slant are selected more often. This does not mean necessarily that these classifiers have more discriminative power, but rather that they provide more complementary information to the other feature sets used in this work. We can observe in Table 1 that the overall error produced by classifiers based on distribution are smaller than those based on slant.

The opposite occurs for those classifiers based on density. In spite of the fact that this feature set shows a good performance against forgeries (see Table 1), the classifiers based on this feature set are not selected for the ensemble very often. In some cases, they are not even used (see Fig. 12a and c). About the curvature feature set, it can be observed that it plays an important role in the ensemble. From Figs. 12 and 13, we can notice that even the weaker classifier of the pool (the fifth classifier of the curvature feature set, with error rate = 25%) has been selected very often. This corroborates to our argument that reproducing the signature using Bezier curves carry complementary information to the other feature sets considered in this work. The classifiers based on curvature are the second most selected for the ensembles.

From the theoretical point of view, the ensemble that generalizes well on unknown data is the one composed of base classifiers that maximize the diversity of opinions on hard cases (ambiguous cases). In this way, very performing classifiers might be discarded during the selection process because they are highly correlated together

Fig. 11. Example of variability: (a) genuine signature, (b) first signatures superposed, and (c) last signatures superposed.

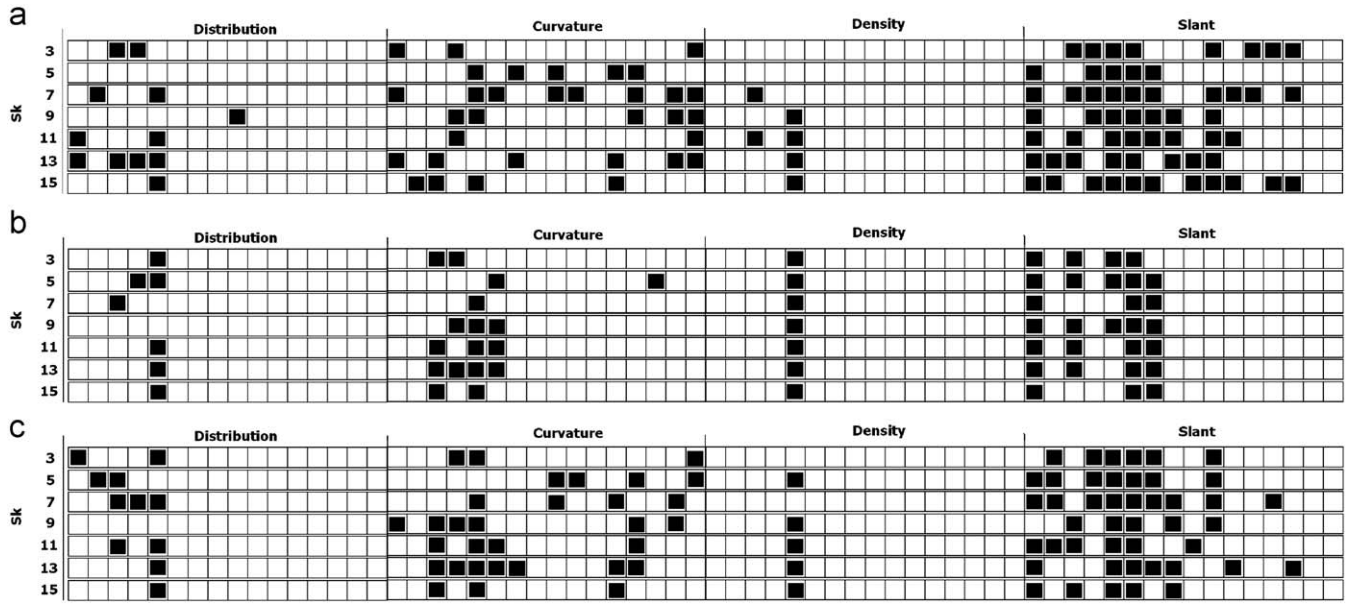

Fig. 12. Classifiers selected during the search using validation set I and fitness: (a) overall error, (b) AUC, and (c) FPR fixed at 10%.

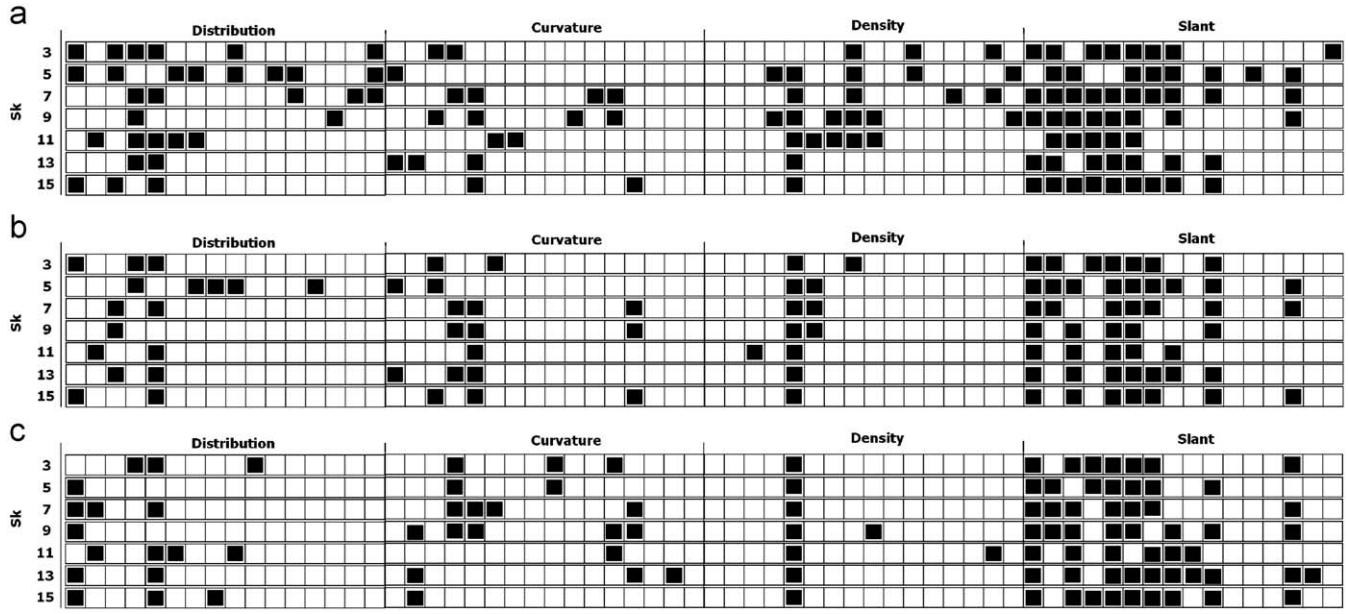

Fig. 13. Classifiers selected during the search using validation set II and fitness: (a) overall error, (b) AUC, and (c) FPR fixed at 10%.

on ambiguous samples, i.e., they could have the same level of competence in their respective representation space for the classification of the unknown sample [3,16]. This is supported by our experiments as some writers are well handled by some specific classifiers, such as density-based and some curvature-based. This explains why some weak classifiers are always selected, in spite of their weak performance.

7. Conclusion

In this paper we have discussed ensemble of classifiers as a strategy to improve the reliability of off-line signature verification systems. It is important to highlight the usefulness of the dissimilarity-based scheme, which allows the design of a universal classifier where new users can be added without retraining the classifiers. As we have demonstrated, if forgeries are available for some writers who did not participate in the training, these samples can be used to fine tune the system and select the best ensemble of classifiers.

Comprehensive experiments taking into account two different scenarios (simple and simulated forgeries available and not available) demonstrated that ensembles based on graphometric features are quite efficient and can reduce considerably the type II error (acceptance of forgeries). What we could observe is that if forgeries are available, a good performance on detecting forgeries can be achieved even using few signatures in the reference set. The efficiency of the proposed methodology was proved on a database composed of 100 writers.

Besides, a new graphometric feature set was introduced. The idea was to simulate the most important segments of the signature by using Bezier curves and then extracting features from them. In spite of the fact that this feature set can be improved in some ways, it has been shown that it can be useful for signature verification. As future works, we plan to investigate different ways to define the singular points used for feature extraction. We believe that, the better the reconstruction, the more reliable will be the feature set.

Acknowledgments

This research has been supported by The National Council for Scientific and Technological Development (CNPq) Grants 471496/2007-3 and 306358/2008-5.

References

[1] H. Baltzakis, N. Papamarkos, A new signature verification technique based on a twostage neural network classifier, Engineering Applications of Artificial Intelligence 14 (2001) 95–103.
[2] L. Breiman, Bagging predictors, Machine Learning 24 (1996) 123–140.
[3] G. Brown, J. Wyatt, R. Harris, X. Yao, Diversity creation methods: a survey and categorisation, Information Fusion 6 (2005) 5–20.
[4] E. Cantu-Paz, Efficient and Accurate Parallel Genetic Algorithms, Kluwer Academic Publishers, Dordrecht, 2000.
[5] S.-H. Cha, S.N. Srihari, On measuring the distance between histograms, Pattern Recognition 35 (2002) 1355–1370.
[6] J. Coetzer, B. Herbst, J. du Preez, Off-line signature verification using the discrete random transform and a hidden Markov model, IEEE Transactions on Image Processing 4 (1995) 870–874.
[7] B. Fang, Y. Tang, Improved class statistics estimation for sparse data problems in offline signature verification, IEEE Transactions on Systems, Man and Cybernetics 35 (3) (2005) 276–286.
[8] T. Fawcett, An introduction to ROC analysis, Pattern Recognition Letters 27 (8) (2006) 861–874.
[9] Y. Freund, R.E. Schapire, A decision-theoretic generalization of on-line learning and an application to boosting, Journal of Computer and System Sciences 55 (1997) 119–139.
[10] L. Goldfarb, What is distance and why do we need the metric model for pattern learning, Pattern Recognition 25 (1992) 431–438.
[11] M. Hanmandlua, M. Yusofb, V.K. Madasuc, Off-line signature verification and forgery detection using fuzzy modeling, Pattern Recognition 38 (3) (2005) 341–356.
[12] T.K. Ho, The random subspace method for constructing decision forests, IEEE Transactions on Pattern Analysis and Machine Intelligence 20 (1998) 832–844.
[13] K. Huang, H. Yan, Off-line signature verification based on geometric feature extraction and neural network classification, Pattern Recognition 30 (1) (1997) 9–17.
[14] R. Hunt, Y. Qi, A multi-resolution approach to computer verification of handwritten system, IEEE Transactions on Image Processing 4 (1995) 870–874.
[15] D. Impedovo, G. Pirlo, Automatic signature verification: the state of the art, IEEE Transactions on Systems, Man, and Cybernetics—Part C 38 (5) (2008) 609–635.
[16] K. Jackowski, M. Wozniak, Algorithm of designing compound recognition system on the basis of combining classifiers with simultaneous splitting feature space into competence areas, Pattern Analysis and Applications, 2008.
[17] E. Justino, F. Bortolozzi, R. Sabourin, Off-line signature verification using HMM for random, simple and skilled forgeries, in: 6th International Conference on Document Analysis and Recognition, 2001, pp. 1031–1034.
[18] E. Justino, F. Bortolozzi, R. Sabourin, A comparison of SVM and HMM classifiers in the off-line signature verification, Pattern Recognition Letters 26 (2005) 1377–1385.
[19] J. Kittler, M. Hatef, R.P.W. Duin, J. Matas, On combining classifiers, IEEE Transactions on Pattern Analysis and Machine Intelligence 20 (3) (1998) 226–239.
[20] L. Kuncheva, Combining pattern classifiers, Methods and Algorithms, Wiley, New York, 2004.
[21] C-L. Liu, Classifier combination based on confidence transformation, Pattern Recognition 38 (1) (2005) 11–28.
[22] V. Mottl, O. Seredin, S. Dvoenko, C. Kulikowski, I. Muchnik, Featureless pattern recognition in an imaginary Hilbert space, in: 16th International Conference on Pattern Recognition, 2002, pp. 88–91.
[23] V.S. Nalwa, Automatic on-line signature verification, Proceedings of the IEEE 85 (2) (1997) 215–239.
[24] W.M. Newman, R. Sproull, Principles of Interactive Computer Graphics, McGraw-Hill, New York, 1981.
[25] L. S. Oliveira, E. Justino, R. Sabourin, Off-line signature verification using writer-independent approach, in: 2007 International Joint Conference on Neural Networks, 2007, pp. 2539–2544.
[26] E. Ozgunduz, T. Senturk, E. Karsligil, Off-line signature verification and recognition by support vector machine, in: 2005 European Signal Processing Conference, 2005.
[27] E. Pekalska, R.P.W. Duin, Dissimilarity representations allow for building good classifiers, Pattern Recognition 23 (2002) 943–956.
[28] R. Plamondon, S.N. Srihari, On-line and off-line handwriting recognition: a comprehensive survey, IEEE Transactions on Pattern Analysis and Machine Intelligence 22 (1) (2000) 63–84.
[29] G. Rigoll, A. Kosmala, A systematic comparison between on-line and off-line methods for signature verification with hidden Markov models, in: 14th International Conference on Pattern Recognition, 1998, pp. 1755–1757.
[30] R. Sabourin, G. Genest, An extended-shadow-code based approach for off-line signature verification: part I evaluation of the bar mask definition, in: 12th International Conference on Pattern Recognition, 1994, pp. 450–453.
[31] S. Santini, R. Jain, Similarity measures, IEEE Transactions on Pattern Analysis and Machine Intelligence 21 (1999) 871–883.
[32] C. Santos, E. Justino, F. Bortolozzi, R. Sabourin, An off-line signature verification method based on document questioned experts approach and a neural network classifier, in: 9th International Workshop on Frontiers in Handwriting Recognition, 2004, pp. 498–502.
[33] S. Srihari, A. Xu, M. Kalera, Learning strategies and classification methods for offline signature verification, in: Proceedings of the 9th International Workshop on Frontiers in Handwriting Recognition, 2004, pp. 161–166.
[34] L. Xu, A. Krzyzak, C.Y. Suen, Methods of combining multiple classifiers and their applications to handwriting recognition, IEEE Transactions on Systems, Man, and Cybernetics 22 (3) (1992) 418–435.

About the Author—DIEGO BERTOLINI GONÇALVES received the B.S. degree in Informatics from UNIPAR and M.Sc. degree in computer science in 2008 from Pontifical Catholic University of Paraná (PUCPR) in 2005 and 2008, respectively. Currently he is professor at Maringa State University (UEM). His interests include pattern recognition and evolutionary computation.

About the Author—LUIZ S. OLIVEIRA received the B.S. degree in Computer Science from UnicenP, Curitiba, PR, Brazil, the M.Sc. degree in electrical engineering and industrial informatics from the Centro Federal de Educacao Tecnologica do Parana (CEFET-PR), Curitiba, PR, Brazil, and Ph.D. degree in Computer Science from Ecole de Technologie Superieure, Universite du Quebec in 1995, 1998, and 2003, respectively. From 2004 to 2009 we was professor of the Computer Science Department at Pontifical Catholic University of Parana, Curitiba, PR, Brazil. In 2009 he joined the Federal University of Parana, Curitiba, PR, Brazil, where he is professor of the Department of Informatics. His current interests include Pattern Recognition, Neural Networks, Image Analysis, and Evolutionary Computation.

About the Author—EDSON JUSTINO received B.S., M.Sc. degrees in Electrical Engineering from UTFPR and Ph.D. degree from Pontifical Catholic University of Paraná, in 1985, 1991, and 2001, respectively. Currently he is full professor at Pontifical Catholic University of Paraná and his interests include signature verification and forensics.

About the Author—ROBERT SABOURIN received B.Ing., M.Sc.A., Ph.D. degrees in Electrical Engineering from the Ecole Polytechnique de Montreal in 1977, 1980 and 1991, respectively. In 1977, he joined the physics department of the Universite de Montreal where he was responsible for the design and development of scientific instrumentation for the Observatoire du Mont Megantic. In 1983, he joined the staff of the Ecole de Technologie Superieure, Universite du Quebec, Montreal, P.Q. Canada, where he is currently a professeur titulaire in the Departement de Genie de la Production Automatisee. In 1995, he joined also the Computer Science Department of the Pontificia Universidade Catolica do Parana (PUC-PR, Curitiba, Brazil) where he was coresponsible since 1998 for the implementation of a Ph.D. program in Applied Informatics. Since 1996, he is a senior member of the Centre for Pattern Recognition and Machine Intelligence (CENPARMI). His research interests are in the areas of handwriting recognition and signature verification for banking and postal applications.