Using Genetic Algorithms and Ensemble Systems in Online Cancellable Signature Recognition

Fernando Pintro and A M Canuto
Informatics and Applied Mathematics Department
Federal University of Rio Grande do Norte (UFRN)
Natal, RN - BRAZIL, 59072-970
pintro@ppgsc.ufrn.br, anne@dimap.ufrn.br

Michael C Fairhurst
School of Engineering and Digital Arts at the
University of Kent, Canterbury, UK.
mcf@kent.ac.uk

Abstract— Biometric-based identification systems can offer several advantages over traditional forms of identity authentication. However, concerns have been raised about the privacy of the personal biometric data, since these systems need to ensure their integrity and public acceptance. In order to address these issues, the notion of cancellable biometrics was introduced. It describes biometric templates that can be cancelled and replaced, in case of being lost or stolen. However, this concept still raises new issues, since they make the authentication problem more complex and difficult to solve. Thus, more effective authentication structures are needed to perform these tasks. In this paper, we investigate the use of ensemble systems in cancellable biometrics, using online signature identification. In order to improve the effectiveness of the ensemble systems, we used genetic algorithms in the choice of an optimized set of weights that are used along with the output of the individual classifiers to define the final output of the system. In addition, we proposed the use of genetic algorithm in the procedure to create the cancellable biometric data, aiming to obtain more efficient cancellable data. The main of this paper is to provide more security in the biometric-based identification process.

I. INTRODUCTION

The development of secure techniques in the identification of users is an important aspect in a hugely interconnected society [1]. The security and reliability of data in information systems are constant concerns, raising numerous challenges in identifying and authenticating users. Generally users of these systems can be authenticated through passwords and tokens, or by human characteristics (biometrics) that are present in all individuals. Biometrics are physical or behavioral human features that are characterized for being unique to each individual and appear as an evolution in authentication and identification of individuals. As they are part of an individual, they cannot be lost, forgotten or stolen. Therefore, biometric-based systems have a high level of security, increasing its reliability, convenience and universality. These characteristics have led to the widespread deployment of biometric-based authentication systems [2,3].

There are several biometric modalities that can be used in the identification or authentication systems, such as: fingerprints, faces, signatures, iris, among others. In this study, we will use the signature data (online signature data) as the main source of analysis, since many issues still can and need to be addressed in this biometric modality. As in all biometric-based systems, aspects such as accuracy, computational time and safety still need improvements.

An important concern in biometric-based systems is related to the security of biometric data, mainly when a biometric template is compromised, either through lost, stolen or unauthorized copies. Biometric information is permanently related to a specific user and they cannot be changed if they are compromised. In this sense, the development of techniques that allow the use of biometric data without compromising the system security is fundamental. The idea of cancellable biometrics has been proposed to address such security issues [4]. Cancellable biometrics consist of applying functions (generally not invertible) in the original biometric data in order to obtain transformed or intentionally-distorted biometric data. Once the biometric data is transformed, an authentication system will use this transformed data for processing and storage. Therefore, in case of being compromised, a biometric template can be replaced through the application of a new transformation function [4].

Although the use of cancellable biometrics bridges the gap between the convenience of biometric authentication and security vulnerabilities, the use of such transformed data generally decreases the performance of the biometric-based system. This is because the level of complexity for the transformed biometric can be much higher than for the original data. For that reason, cancellable biometric systems must support a good trade-off between discrimination capability (accuracy) and non-invertibility (high security), when applied to any biometric modality. In our previous works, [5,6], we applied ensemble systems to cancellable biometrics and we have reached promising results, since we have demonstrated that the use of ensemble systems improves the accuracy of cancellable biometrics. However, we noticed that some cancellable transformations can degrade the accuracy level of ensemble systems.

In order to improve the efficiency of ensembles systems in cancellable biometric recognition, we will apply genetic algorithm (GAs) in the two separate parts of the design of an ensemble system to biometric recognition, aiming to maximize the performance of these systems. The first genetic
algorithm is used in the cancellable transformation (pre-processing phase), and its goal is to improve the efficiency of the discretization process of the cancellable transformation through the use of optimized discrimination intervals. The second GA tries to improve the performance of ensemble systems through the use of an optimized set of weights that are used along with the output of the individual classifiers to define the final output of the ensemble systems (weighted combination methods). In addition, we will also present a methodology to extract features that can be used in online signature datasets.

We expected to show that we can improve even further the performance of cancellable signature through the use of more efficient ensemble systems. With this, it is expected to contribute to a better performance (accuracy and safety) in the identification of the individual, something that is increasingly important in current days.

II. RELATED STUDIES

Many studies can be found in the literature showing efficient ways of using ensemble systems or genetic algorithm in cancellable biometric, such as in [7,8,9,10,11,12] among others. We would like to highlight the works described in [7,8], since they are more related to the subjects of this paper.

In [7], for instance, an on-line signature authentication system based on an ensemble of local, regional, and global matchers is presented. The following matching approaches are taken into account: the fusion of two local methods employing Dynamic Time Warping, a Hidden Markov Model based approach where each signature is described by means of its regional properties, and a Linear Programming Descriptor classifier trained by global features. Moreover, the original biometric characteristics are transformed using two transformation functions, BioHashing and BioConvolving that presented a good performance (Equal Error Rate=4.51%) even when an impostor had access to the data security key.

In [8], two different structures of genetic algorithms (GAs) are used in a feature selection problem in online signature verification. First, a structure of a binary encoding pattern of GA is used to find a suboptimal subset of features that minimizes the verification error rate of the system. Second, the curse of dimensionality phenomenon is further investigated using a GA with integer coding. In the used database, signatures are represented by a set of features divided into four different groups according to the signature information they contain, namely: time, speed and acceleration, direction, and geometry. The results provided by GAs show that features from subsets time and geometry are the most discriminative when dealing with random forgeries. On the other hand the parameters of a sub-assembly velocity and acceleration, and geometry are superior in ratings with skilled forgeries.

Unlike most of the aforementioned works, in this paper we propose to combine ensemble and genetic algorithm to cancellable online signature recognition. In this paper, we use an optimization technique to improve the efficiency of ensemble systems and the cancellable transformation, aiming to improve even further the recognition of cancellable biometric data.

III. BIOMETRIC DATA

One of the biggest concerns is related to the security of biometric data, due to the fact of biometrics are unique and are permanently related to a user, making it impossible for its replacement or cancellation if compromised. In traditional biometric systems, when the original sample is somehow compromised, the biometric characteristics used to generate such a sample may not be used again (reversible). One of the solutions is related to the explicit storage of biometric models in these systems, eliminating any possibility of leakage of the original biometric trait [2]. In order to fulfill the needs involving security issues, the idea of cancelable biometrics has emerged as an important tool. The main reason for this is because cancellable biometrics allow the reuse of a compromised biometric characteristic without affecting the security of the system [4]. However, a common problem found in these procedures is the decrease of the performance on the identification systems, motivated by the complexity found in the transformed data to be much larger than the original data. Therefore, it is important for a biometric-based system to have a good trade-off between robustness (accuracy) and non-invertibility (security) when using feature transformations in any biometric modality.

There are many methods of biometric template protection proposed that, according to [4], can be divided into two main categories, which are: Cryptosystem and Feature Transformation.

- Cryptosystems: In this case, some public information about the biometric template is stored. This public information is usually referred to as helper data, and hence biometric cryptosystems are also known as helper data based methods.

- Feature Transformation: A transformation function \( f \) is applied to the biometric model \( T \) and only the transformed model \( f(T) \) is stored in the database. The parameters of the transformation function are provided by password or by a random key \( K \).

According to the characteristics of the transformation function \( f \), the techniques based on Feature Transformation can be further classified into: salting or non-invertible transformations. In the salting, the function \( f \) is invertible and its security is based on a secret key or password. Thus, if an intruder obtains access to the security key, the biometric data will be at risk. On the other hand, non-invertible transformations use one-way functions, making it computationally difficult or impossible the inversion of the model transformed, even having access to the security key. In this work, an adaptation of an-invertible function, called BioHashing, will be used. This transformation function will be described in the next subsection.
A. The BioHashing Cancellable Transformation

The original BioHashing algorithm is characterized by transforming the original biometric characteristics of an individual in a non-invertible binary sequence (discretization).

BioHashing combines biometric characteristics and passwords or tokens (physical tokens, smart card, USB token, among others). This forces the presence of both parts (token+biometric) for authentication a user, protecting it against a biometric falsification without the knowledge of both parts. In addition, it makes it impossible to obtain the original data (biometric characteristics) through the transformed model.

The revocability of the data is provided by the token replacement and BioHashing has significant functional advantages when compared with the original biometric data, such as: zero equal error rate (EER) point and it eliminates the occurrence of FAR without overly expose the FRR performance [13].

Most work based on BioHashing found in the literature are aimed to the recognition of fingerprint, palm and face, as in: [13,14,15,16,17,18]. However, its adjustment can be made to any biometric modality, since the size of the features vectors have a fixed dimension after the processing of the biometric data.

A simplified BioHashing procedure can be described as follows.

1. From the vectors containing the original biometric characteristics \( I \in R^n \), with \( n \) being the dimension it is generated a set of pseudo random vectors \( \{ p_i \in R^n \mid i = 1, ..., m \} \), where \( m \) represents the number of vectors of dimension \( n \) and \( m \leq n \).
2. It is then applied the Gram-Schmidt process in \( \{ p_i \in R^n \mid i = 1, ..., m \} \), aiming to obtain a set of orthonormal vectors \( \{ r_i \in R^n \mid i = 1, ..., m \} \).
3. After that, it is calculated the inner product between the vector of biometric features \( I \) and each of the pseudo random orthonormal vectors \( r_i \), \( \langle I, r_i \rangle \) = 1, ..., \( m \) , where \( \langle \cdot \rangle \) indicates the inner product operation.
4. Then, it is created a BioHashing model, which has dimension of \( m \). This is done through the binary discretization of values obtained in the inner products, \( \{ b = b_i \mid i = 1, ..., m \} \), where:

\[
b_i = \begin{cases} 
0 & \text{if } \langle I, r_i \rangle \leq \tau, \\
1 & \text{if } \langle I, r_i \rangle > \tau,
\end{cases}
\]

(1)

With \( \tau \) being a threshold determined empirically.

In the BioHashing function, the system performance is dependent on the variable \( m \), which is the number of pseudo random orthonormal vectors. The number of vectors must be defined empirically and will correspond to the number of attributes for each instance of the transformed base.

The use of BioHashing in the original biometric model has not always resulted in improvement in the accuracy level of the identification systems and its use depends on the characteristics of specific biometric dataset. In order to define an efficient and robust cancellable model, an adaptation of the BioHashing model was proposed and it is basically done in step 4 of the BioHashing procedure (discretization function).

The adaptation of the BioHashing model proposed in this paper consists of two basic concepts, which are: discretization of data through 4 values \([0;0.33;0.66;1]\), instead of only in 2 values \([0;1]\), and the use of GAs to determine which are the optimal intervals for discretization.

The purpose of the use of more discretization values is to enlarge the space of coding, enabling a better representation of the transformed base, without losing the characteristics of the BioHashing function.

For the proposed discretization step, a standard normalization is carried out. Then, the normalized data is discretized according to the following intervals:

\[
b_i = \begin{cases} 
0 & \text{if } 0 \leq \langle I', r_i \rangle < a, \\
0.33 & \text{if } a \leq \langle I', r_i \rangle < b, \\
0.66 & \text{if } b \leq \langle I', r_i \rangle < c, \\
1 & \text{if } c \leq \langle I', r_i \rangle \leq 1.
\end{cases}
\]

(2)

The values for the discretization coefficients \( a, b \) and \( c \) of Eq. 2 are determined by a genetic algorithm (GA).

IV. ENSEMBLE SYSTEMS

Ensemble systems are composed of a set of \( N \) individual classifiers (ICs), structured in a parallel way, which receive the biometric data and provide their output to a combination module, which is responsible for providing the final output of the system. These systems exploit the idea that different classifiers can offer complementary information about the biometric data to be classified, thereby improving the effectiveness of the recognition process [19,20].

In the design process of an ensemble system, two main issues must be considered which are: the individual classifiers and the combination method that will be used. In relation to the first issue, we need to define the size of the ensemble system (number of individual classifiers) as well as the classification algorithms to be used as the individual classifiers. Depending on its particular structure, an ensemble can be designed in two main approaches: heterogeneous and homogeneous. In the first case, different types of classification algorithms are chosen as individual classifiers, while only one type of classification algorithm is chosen in the second case (ensemble of k-NN, for instance).

After the definition of the set of individual classifiers, we need to choose an effective way of combining their outputs. According to [19], in an ensemble system, the possible ways of combining the outputs of the \( N \) individual classifiers (ICs) depend on what information we obtain from these classifiers. They can vary from the simplest approach using class labels or rank values to the utilization of more elaborate information, such as support degree D. In this paper, we will
use the support degree of the ICs as input information for the combination module.

The choice of the best combination method for an ensemble system usually needs an exhaustive testing procedure. In other words, the choice of the combination technique of an ensemble system is very important and generally difficult to optimize. According to their functioning, the combination module typically adopts one of three main strategies, as follows:

- **Fusion or combination based methods**: This combination method uses the output of all individual classifiers in order to reach the final output.
- **Selection-based methods**: This combination method selects one individual classifier to use as a guide to reach the final output. Hence, once an individual classifier is chosen, the output of the selected classifier becomes the final output of the ensemble system;
- **Hybrid methods**: This combination method can use a mixture of selection or fusion. Usually, selection is the first choice but selection is chosen if and only if the best classifier is good enough to classify the test pattern. Otherwise, fusion is chosen.

In our investigation, fusion-based combination methods are considered. Additionally, we will use the support degree as input for the combination methods and these methods will be treated as classification methods, receiving real numbers as input and providing as output the class to which the biometric data belongs.

V. **GENETIC ALGORITHMS**

Genetic algorithms (GAs) were first developed by Holland [21] and popularized by [22]. The basic idea of the GAs operation is to treat the possible solutions of a problem as individuals of a population, which will develop at each iteration or generation. For this, it is necessary to build a model of development where individuals are solutions of a problem. The evolution usually starts from a set of randomly created solutions (initial population). At each generation, the adaptation of each solution in the population is evaluated using a fitness function and some individuals are selected to be used as genetic operators in order to form a new population. The new population is then used as input to the next iteration of the algorithm.

The structure of a GA can be described by evaluating six principles, as follows:

- **Chromosome**: every chromosome (individual) describes a possible solution for a problem. They are usually represented by a sequence of codes grouped together.
- **Population**: set of chromosomes or solutions to the problem.
- **Fitness evaluation**: checks the degree of fitness of the chromosomes of the population.
- **Selection**: It is a stochastic function that selects chromosomes. The main purpose of this function is to select the best solutions for future generations more often than the worst ones [23].
- **Genetic Operators**: they are extremely necessary for the population to diversify and to keep the adaptation characteristics acquired by previous generations. Genetic operators most frequently used are the Crossover and mutation.
- **Stopping Criterion**: it aims to terminate the algorithm when an optimal solution or the initial established parameters are found.

In the context of ensemble systems, GAs have been widely used to optimize the structure of ensembles, including the features used by the base classifiers. This is due to the great capacity of adaptation (flexibility) of GAs that makes it an important tool in the development of ever more efficient recognition systems.

VI. **EXPERIMENTAL SETTING UP**

In order to demonstrate the efficiency of using genetic algorithms to optimize ensemble systems in the development of more efficient systems, as well as to demonstrate the efficiency of adaptation proposed in the BioHashing transformation function, an empirical analysis will be performed. In this analysis, three datasets composed of digital online signature features (original and transformed) will be applied to individual classifiers, as well as to the ensemble systems (homogeneous and heterogeneous structures). These three datasets were obtained through the use of different feature acquisition process and will be describe in Section VI.B.

In order to add more diversity in the ensemble systems, random feature selection process was applied, generating different subsets of attributes for each of the individual classifiers of the ensemble systems. In an initial analysis, we tested several possibilities of random reductions and we observed that 80% of reduction did not cause a strong reduction in the accuracy of the ensemble systems. Therefore, each resulting base was composed by 80% of the total existing attributes and was created exclusively for one individual classifier (different subsets of attributes for each individual classifier).

The classification algorithms used as individual classifiers applied in this analysis were $k$-NN (Nearest Neighbor), Naive Bayes and Multi-layer Perceptron Neural Networks (MLP). The choice of these methods as individual classifiers is based on the diversity in the classification criteria used by these methods. In addition, the decisions of individual classifiers were combined by fusion-based methods functions, which are also responsible for assigning the final output of the ensemble systems. In this analysis, we used three combination methods, which are: Sum, majority voting and $k$-NN.

The evaluation of the classification systems was made through two different steps. In the first step, the classification of legitimate users was done using a standard k-fold cross validation, with $k=15$. In the second step, an analysis of
signatures made by falsifiers is performed, where a single test is performed using genuine instances as training set and qualified falsifications (10 for each user) were tested. Therefore, the results presented in this analysis are related to the performance of the tests sets of the legitimate signatures (first step), summing up with the analysis of the falsifications (second step). We believe that this technique provides a better estimate of system performance, evaluating all instances of the dataset, but in two different steps.

In addition, k-NN combination method is considered trainable method. In this case, we generated a validation dataset to train these methods. The k-NN used as combination function has been configured with k = 1. In addition, all classification and optimization methods used in this study were written in Matlab.

In order to compare the accuracy of the classification systems, a statistical test will be applied, which is called the hypothesis test (t-test). For this test, the confidence level is 95% ($\alpha = 0.05$).

A. Genetic Algorithms

As mentioned previously, we used genetic algorithm (GA) to improve the performance of the recognition process of cancellable biometrics. In order to do that, two genetic algorithms were used. The first one was applied in the BioHashing method and the second one was used to define the optimal weight set for the individual classifiers of the ensemble systems.

Firstly, we used a standard genetic algorithm to define the best intervals for discretization in the BioHashing cancellable transformation. The chromosomes were represented by real numbers that lie in the [0,1] interval of size 3 (one for each coefficient of Equation 2). The initial population was randomly chosen. The population size was defined as 250 and standard genetic operators such mutation and crossover were used. This GA used as a fitness function the accuracy of the resulting transformed data through the average of performance of two classifiers, a single $k$-NN with $k=1$ and a standard Naive Bayes. When using a single classifier as a fitness function, the GA tries to find the best solution for the specific classifier and may lose the ability of generalization, since the solutions are biased to that particular classifier. With the use of two classifiers, with different classification criteria, it is expected to smooth out this problem by maintaining the ability to generalize its performance (performance improvement) for more than one classifier.

In ensemble systems, weights can be associated with every sample to denote the confidence (influence) of the individual classifiers in classifying the input pattern to a particular class. Different ways of calculating the weights (confidence) of each class for each individual classifier can be used in determining the relative contribution of each classifier within the system. In order to illustrate the use of weights, it is important to define the output of an ensemble, which can be described as:

$$\text{Ensemble} = \text{Max}(\text{Out}_{i=1,..,C} = \sum_{n=1}^{NC}(O_{nc} \times W_{nc}))$$

In equation 3, the output of individual classifier $(i)$ to class $(c)$, $O_{nc}$, is used along with the weights of the corresponding classifier to the corresponding class ($W_{nc}$) to define the final output of the ensemble system.

Although weighted combination methods appear to provide some flexibility, obtaining the optimal weights is not an easy task. In this paper, we use a second genetic algorithm to define the best set of weights to be used in an ensemble system.

For this second GA, we used a population of size 200, with each individual being represented by a vector of real number representing the weights for each class of each individual classifier. The population was initialized assigning to all weights of all individual classifiers. We applied mutation and crossover as genetic operator and the stopping criterion was defined by the maximum number of iterations (varying from 400 and 1000, depending on the used dataset). As fitness function, we used the accuracy level of the ensemble systems using the chosen combination method ($k$-NN, Sum or Voting).

As we use an initial population composed of identical chromosomes, we applied mutation operator with a variable rate. For the first iteration, mutation rate was 20% and a reduction of 1% was done every five iteration until it reaches 1% and it stabilizes.

B. Datasets

Online signatures are representations of the handwritten signatures of the individuals, but obtained through a graphics digitizer table that captures spatial information and dynamics characteristics of writing. Dynamic information extends considerably the number of possible features of a signature, making it much safer to the identification process. Some of the biometric characteristics that can be used in an identification system of online signature are the number of times that the pen is removed from the table, the speed of the signature, the inclination of the pen regarding to the digitizer table, the spatial coordinates, among others.

The dataset of digital signatures used in this experimental investigation was collected in the school of digital arts at the University of Kent, in the United Kingdom in a controlled environment, and it belongs to a European Project conducted by the EU BioSecure Network of Excellence [24]. The dataset is composed of 79 users divided into two sessions. Together the two sessions contain 50 samples of signatures for each user, in which 30 are examples of real signatures and 20 are attempts to imitate the signature of another user (qualified falsifications). In this investigation, we used only 25 samples from each user, being 15 genuine samples and 10 qualified falsifications. The data were collected using an A4-sized graphics tablet with a density of 500 lines per inch.
The feature extraction process used in this work consists of 6 simple methods which were extracted based on spatial coordinates and the propagation speed of the signatures. We then combined these six methods in a two-by-two basis to generate three different datasets. The choice of these characteristics is made because they are extracted by simple digitizer tables and consequently can be more accessible in real applications.

An online signature can be described by coordinated points representing the orientation of the signature printed in constant intervals of time t. Based on this information, for each signature that compose the datasets, six characteristics vectors (V1, V2, V3, V4, V5 and V6) will be created and casted.

The two first vectors of characteristics (V1 and V2), represent angles in function of time t. The first one is uses the angle of a reference coordinate of a signature, c_t, its subsequent point c_{t+1} and the center of mass cM of the signature. The second one represents the angle between c_t, its third subsequent point c_{t+3} and the center mass cM. These equations provide characteristics of the physical structure of the signatures.

The three following vectors (V3, V4 and V5) correspond to the distance between the coordinates of a signature sample as a function of time t. V3 represents the distance between the reference coordinate c_t and the center mass cM of the signature, V4 is the distance between the coordinate reference c_t and its subsequent coordinate c_{t+1} and V5 is the distance between c_t and c_{t+3}. These vectors describe characteristics of propagation speed of the signatures.

The last vector (V6) aims to extract structural characteristics of the signatures and corresponds to the area formed by triangles between the reference coordinate, c_t, and its subsequent points c_{t+2} and c_{t+3}.

After the creation of the vectors, their size is standardized in 750 attributes (l = 750), where characteristic vectors with more than 750 attributes are cut and the one with smaller dimensions are filled with zeros. After the standardization step these vectors were combined, in a two-by-two basis, by multiplying its corresponding vectors and thus forming three separate datasets (B1, B2 and B3), in accordance with the following equations.

\[ B1_{jxt} = V1_t \times V6_t \]
\[ B2_{jxt} = V2_t \times V5_t \]
\[ B3_{jxt} = V3_t \times V4_t \]

Where,

l is the number of attributes and j is the number of instances of the dataset.

The computational time is also an important issue in the performance of a biometric system. As the number of attributes of an instance has a direct impact on computational time, it is important to design a dataset with the smallest possible size, and still keeping its main characteristics. Techniques such as PCAs (Principal Component Analysis) are characterized by being able to extract the most relevant parts of a signal, thereby reducing their dimension without losing the main characteristics, bringing great benefits to the system. In this paper, we applied a PCA method in the original dataset, selecting the 50 main components. In this case, we reduced the original datasets from 750 to 50 attributes. Although we reduced in 15 times (93.33%) the dimension of the datasets, the accuracy level using individual classifiers increased an average of 1.37% points.

As a result of these processing, three datasets (B1, B2 and B3) were created, using different features of the 1975 used signature samples, where each signature was composed of 50 attributes. In the empirical analysis, we also investigate the combination of these three datasets (B1, B2 and B3) in an ensemble system. The idea was to use each dataset in 1/3 of the individual classifiers of the ensembles for one dataset. We then performed all possible assignments of these three different datasets (B1, B2 and B3) to the individual classifiers, making a total of 6 assignments. The results presented in this paper represent the average of these 6 possible assignments.

In order to apply the cancellable transformations, we used m equal to 50 and this choice was done based on an initial empirical analysis. In this case, the transformed BioHashing data also consists of 50 attributes, the same number in original datasets.

VII. RESULTS AND DISCUSSION

This section presents the results and discussions of the experimental analysis described in the previous section. It is important to highlight that we have addressed a user identification task in this paper. In this case, once a biometric is provided, the classification systems will provide a class label which corresponds to a specific individual user. An alternative would be to implement a verification task. However, verification is, in effect, a two-class classification problem, whereas identification is a C-class problem (where C is the number of enrolled users). Therefore, from a classification point of view, identification is a complex and time consuming task. Although verification results would be directly relevant to authentication, as we want to analyze the benefits of well-established classification structures, we have chosen to focus on identification. Nonetheless, the proposed methodology can be easily adapted for verification tasks. In our analysis, the PIN numbers to be used in the cancellable transformation will be held secretly.

In order to understand the results obtained by the ensemble systems optimized by GA, it is important to first analyze the accuracy level of the individual classifiers used in the set of the experiments. In addition, we can analyze the benefits provided by the proposed adaptation of the BioHashing function in the individual classifiers. Table I shows the accuracy rate and standard deviation of the individual classifier for each dataset in their original and transformed spaces (original and proposed BioHashing).
this table, the shaded cells indicate the best results reached by an individual classifier of the corresponding dataset, among the BioHashing methods. When applying the Wilcoxon signed-ranks test, the bold numbers represents the statistically significant improvements, when comparing with the original BioHashing.

Table I. Accuracy rate and standard deviation of individual classifiers (average)

<table>
<thead>
<tr>
<th>Datasets</th>
<th>Original</th>
<th>BioHashing Original</th>
<th>Proposed BioHashing</th>
</tr>
</thead>
<tbody>
<tr>
<td>B1</td>
<td>76.87±4.23</td>
<td>72.64±5.89</td>
<td>77.77±6.37</td>
</tr>
<tr>
<td>B2</td>
<td>77.36±5.76</td>
<td>68.86±5.13</td>
<td>77.43±4.95</td>
</tr>
<tr>
<td>B3</td>
<td>78.96±5.29</td>
<td>72.13±6.16</td>
<td>79.37±5.46</td>
</tr>
</tbody>
</table>

As noted from Table I, the transformation function BioHashing adapted brought about a significant improvement to the performance of individual classifiers, especially when compared to the original BioHashing transformation function. This improvement is caused by the expansion of the number of intervals of discretization and the efficiency of the GA when determining the discretization intervals that maximize the performance of the systems. In relation to the original dataset the adaptation of BioHashing function creates encoding from the original attributes, extending the extra diversity class and correcting with this some structural failures submitted by original datasets. We then had individual classifiers with accuracy rate similar to the original dataset.

The results from Table I show that the use of GA in the cancellable transformation can be a very important tool in the creation of safety and efficient biometric systems. The use of the proposed BioHashing method was able to smooth out one of the biggest problems founded in cancellable biometric systems, which is the decrease of the accuracy level when using cancellable transformations. The results show that it is possible to create cancellable transformations that can provide a good trade-off between robustness and security.

In table II, the results using the ensembles systems are presented and these systems are applied to the original BioHashing cancellable dataset (O-B) and the proposed methodology which applied GA in BioHashing and in the ensemble systems (P-B). In addition, we also present the results of combining all three datasets generated (B1, B2 and B3) in the ensemble systems, as explained in previous section (called combined scenario, or simply Cb). For all datasets, we analyze them with ensemble systems using three different numbers of individual classifiers (3, 6 and 12). Once again, the shaded cells indicate the best results reached by an ensemble system of the corresponding dataset, from a statistical point of view.

The first analysis of Table II is related to the effects of using ensemble systems, when compared with the performance of individual classifiers (Table I). We can see that the use of ensemble systems caused a significant increase in the accuracy level of the biometric systems. For instance, the highest result obtained by the individual classifier was 79.37% (the proposed BioHashing for B3). The best combination method for this analyzed case (the proposed BioHashing for B3) in Table II (ensemble systems) was 87.26% (ensemble size 12 combined with Sum method). It means an increase of almost 8% points in the accuracy level when using ensemble systems. These results show that the use of ensemble can be a useful tool in the development of more accurate systems. Furthermore, it shows that the number of individual classifiers can have an important effect in the accuracy level of the ensemble systems. We noticed that using more classifiers provide better results.

In second analysis of Table II, it is investigated the use of a genetic algorithm to define the optimal set of weights to be used in the ensemble systems, comparing O-B and P-B columns of Table II. As we can observe, the use of the proposed approach improved the accuracy level of the ensemble systems in all analyzed cases, showing statistical significance in almost 67% of the cases (24 out of 36). This shows that the use of genetic algorithms may have an important role in the performance of the ensemble systems, ensuring even more successful development of ensemble-based techniques in cancellable biometric data.

In summarizing, we can make three main assumptions in this empirical analysis, which are:

- Weighting the output of classifiers in an ensemble is useful for this application (user identification).
- The genetic algorithm finds the (near-)optimal weighting for combining classifiers in an ensemble.
- Filtering the input dataset does not change the assumption that the data is representative, but makes the process effectively computable.

VIII. CONCLUSION

This work presented a new approach for the identification of online cancellable signatures, where genetic algorithms were applied in two different stages of the recognition process. In addition, we presented a methodology to extract features in online signature datasets. Six techniques were extracted based on information contained in an online signature sample. These techniques were combined to form three different datasets and a PCA method was applied to reduce the set of attributes to 50. The resulting three datasets were used as basis to the empirical analysis of this paper.

Through this analysis, an important conclusion observed is that the use of GAs in the search for more efficient and secure biometric systems had a positive effect, since it increased the accuracy rate of the ensemble systems. In addition, it showed that the flexibility of genetic algorithms makes it possible to apply in different parts of a biometric recognition system, whether it is to optimize the set of
weights in an ensemble system or to refine the processing of a cancellable transformation. Like this, it is ensuring a secure system and, at the same time, with the highest accuracy level.

We believed that the main objective of this study was achieved, showing efficient tools in the implementation of biometric-based systems in its two main areas of study, which determine the feasibility of a system: security and accuracy.

REFERENCES


Table II – The accuracy level and standard deviation of ensemble systems using the different BioHashing approaches (original and proposed).

<table>
<thead>
<tr>
<th>Combination</th>
<th>Cancellable Transformation (Datasets)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>O-B (B1)</td>
</tr>
<tr>
<td>Size 3</td>
<td></td>
</tr>
<tr>
<td>Sum</td>
<td>81.71±3.59</td>
</tr>
<tr>
<td>Vote</td>
<td>81.38±4.09</td>
</tr>
<tr>
<td>Knn</td>
<td>80.46±3.92</td>
</tr>
<tr>
<td>Size 6</td>
<td></td>
</tr>
<tr>
<td>Sum</td>
<td>82.51±3.5</td>
</tr>
<tr>
<td>Vote</td>
<td>82.17±3.92</td>
</tr>
<tr>
<td>Knn</td>
<td>82.24±3.88</td>
</tr>
<tr>
<td>Size 12</td>
<td></td>
</tr>
<tr>
<td>Sum</td>
<td>82.64±3.95</td>
</tr>
<tr>
<td>Vote</td>
<td>83.48±3.34</td>
</tr>
<tr>
<td>Knn</td>
<td>82.07±4.71</td>
</tr>
</tbody>
</table>


