Biosecure reference systems for on-line signature verification: a study of complementarity

Sonia GARCIA-SALICETTI[1], Julian FIERREZ-AGUILAR[2],
Fernando ALONSO-FERNANDEZ[2], Claus VIELHAUER[3],
Richard GUEST[4], Lorene ALLANO[1], Tung DOAN TRUNG[1],
Tobias SCHEIDAT[3], Bao LY VAN[1], Jana DITTMANN[3], Bernadette DORIZZI[1],
Javier ORTEGA-GARCIA[2], Joaquin GONZALEZ-RODRIGUEZ[2],
Martino BACILE DI CASTIGLIONE[4], Michael FAIRHURST[4]

Abstract

In this paper, we present an integrated research study in On-line Signature Verification undertaken by several teams that participate in the BioSecure Network of Excellence. This integrated work, started during the First BioSecure Residential Workshop, has as main objective the development of an On-line Signature Verification evaluation platform. As a first step, four On-line Signature Verification Systems based on different approaches are evaluated and compared following the same experimental protocol on MCYT signature database, which is the largest existing on-line western signature database publicly available with 16 500 signatures from 330 clients. A particular focus of work documented in this paper is multi-algorithmic fusion in order to study the complementarity of the approaches involved. To this end, a simple fusion method based on the Mean Rule is used after a normalization phase.

Key words: Biometrics, Handwriting, Signature, On-line processing, Authentication, Comparative study, Statistical method, Hidden Markov model, Reference model, Mathematical distance, Experimental study, Research program.

SYSTÈMES DE RÉFÉRENCE DE BIOSECURE POUR LA VÉRIFICATION DE SIGNATURE EN LIGNE : UNE ÉTUDE DE LA COMPLÉMENTARITÉ

Résumé

Dans cet article, nous présentons un travail commun sur la vérification de signature en-ligne, réalisé par 4 équipes qui participent au Réseau d'Excellence BioSecure. Ce travail commun, débuté durant le premier « Workshop » résidentiel, a pour principal objectif le développement d'une plateforme d'évaluation pour la vérification de la signature en-ligne. Tout d'abord, quatre systèmes de vérification de signature en-ligne basés sur différentes approches sont évalués et comparés en utilisant le même protocole expérimental sur la base de signa-

1. GET/INT (Institut National des Télécommunications), Dept. EPH – 9 rue Charles Fourier, 91011 EVRY France.
2. ATVS/Biometrics Research Lab., Escuela Politecnica Superior, Universidad Autonoma de Madrid, Spain
3. Otto-Von-Guericke University of Magdeburg, School of Computer Science, Dept. ITI – Universitaetsplatz 2, 39016 Magdeburg, Germany
4. Department of Electronics, University of Kent – Canterbury, CT2 7NT, United Kingdom.

tures MCYT, *la plus grande base existante de signatures en-ligne disponible, avec 16500 signatures de 330 personnes. Ensuite, l'accent est mis sur la fusion multi-algorithmique afin d'étudier la complémentarité des approches impliquées. Pour cela, une méthode de fusion simple est utilisée, basée sur une moyenne des scores après une phase de normalisation.*

Mots clés : Biométrie, Écriture manuscrite, Signature, Traitement en ligne, Authentification, Étude comparative, Méthode statistique, Modèle Markov caché, Modèle référence, Distance mathématique, Étude expérimentale, Programme recherche.

Contents

I. INTRODUCTION

The Network of Excellence (NoE) BioSecure started in June 2004 grouping the critical mass of expertise required to promote Europe as a leading force in the field of Biometrics. The main objective of this network is to strengthen and to integrate multidisciplinary research efforts in order to investigate biometrics-based identity authentication methods, for the purpose of meeting the trust and security requirements in the progressing digital information society. This goal will be attained through various integrated efforts. Among them, a common evaluation framework including Reference Systems, assessment protocols and databases is at the centre of the objectives of the Network. Indeed, this framework permits, for the first time, the creation of standard evaluation conditions at the European level, that is to evaluate the existing systems at the international level with regard to Reference Systems developed by some partners of the network.

This paper presents the particular work in the On-line Signature modality. For the first time, three BioSecure Reference Systems for On-line Signature Verification, from GET/INT (Institut National des Télécommunications) in France, University of Magdeburg (AMSL) and University of Kent, are presented and evaluated jointly with another state-of-the-art system from Universidad Politécnica de Madrid, (UPM) – an additional member of the BioSecure Network. The systems encompassed both attempts to develop high performance optimised verification algorithms and simpler benchmark structures to broaden the base of comparisons which could be made. These four systems were evaluated in the context of integrated work undertaken by the four institutions participating to the first BioSecure Residential Workshop in August 2005. Among the four systems here presented, the approaches taken are very different: two are based on a statistical approach, Hidden Markov Models [1], and two on reference-based methods: Levenshtein distance [2], and distance measures in general. Systems are evaluated with the same protocol using the largest existing on-line western signature database, the signature section of the MCYT database [3], containing 330 clients. The work is particularly focused on the combination of such systems to exploit the complementarity of the approaches involved.

Combining multiple systems has already been the subject of intensive research [4-8]. We chose in this work a simple fusion method, the Mean Rule, that has proven to be efficient after a score normalization phase [4,6]. Indeed, the aim of normalization is to obtain comparable scores in order to attain good results through simple fusion rules. This fusion method avoids a time-consuming learning phase (as found in other schemes such as Support Vector Machine for example), but still requires a dedicated development set of scores to compute normalization factors. To that end, a fusion protocol based on Cross-Validation [9] is proposed on the MCYT database.

This paper is organized as follows: first, the four systems here presented and the normalization techniques used for their scores' combination are described in Section II, then the experimental setup is detailed in Section III (MCYT database description, individual systems' protocol and fusion protocol). Finally, Section IV presents the analysis of results and Section V draws our conclusions.

II. DESCRIPTION OF SYSTEMS AND FUSION TECHNIQUES

II.1. HMM-based approaches

Two approaches studied in this framework are based on Hidden Markov Models (HMM). The first is based on the fusion of two sorts of information derived from a writer's HMM and is designated Reference System 1 (Ref1). This system was developed by GET/INT [10]. The second is based on the standard Log-likelihood information and developed by UPM [11]; it is called System 4 (Sys4) in the remainder of this paper.

II.1.1. Reference System 1

Signatures are modeled by a continuous left-to-right HMM. In each state a continuous density multivariate mixture of 4 Gaussians is used. A complete HMM description can be found in [1]. 25 dynamic features are extracted at each point of the signature. Features are given in Table I and described in more detail in [10]. They are divided into two sub-categories, namely dynamic features and local shape related features.

The topology of our signature HMM only authorizes transitions from each state to itself and to its immediate right-hand neighbor. Also, the covariance matrix of each multivariate Gaussian in each state is considered diagonal.

The number of states in the HMM modeling the signatures of a given person is determined individually, according to the total number (T_{total}) of all the sampled points available when summing all the genuine signatures that are used to train the corresponding HMM. We consider it necessary to have on average at least 30 sampled points per Gaussian for a good re-estimation process. Then, the number of states N is computed as:

$$N = \left\lfloor \frac{T_{total}}{4*30} \right\rfloor \text{ where brackets denote the integer part.}$$

TABLE I. – The 25 dynamic features extracted from the on-line signature.

Les 25 caractéristiques dynamiques extraites de la signature en-ligne.

	N°	Feature name
Dynamic features	1-2	Normalized coordinates $(x(t)-x_g, y(t)-y_g)$ relatively to the gravity center (x_g, y_g) of the signature
	3	Speed in x
	4	Speed in y
	5	Absolute speed
	6	Ratio of the minimum over the maximum speed on a window of 5 points
	7	Acceleration in x
	8	Acceleration in y
	9	Absolute acceleration
	10	Tangential acceleration
	11	Pen pressure (raw data)
	12	Variation of pen pressure
	13-14	Pen-inclination measured by two angles
	15-16	Variation of the two pen-inclination angles
Local shape related features	17	Angle α between the absolute speed vector and the x axis
	18	Sine(α)
	19	Cosine(α)
	20	Variation of the α angle: φ
	21	Sine(φ)
	22	Cosine(φ)
	23	Curvature radius of the signature at the present point
	24	Length to width ratio on windows of size 5
	25	Length to width ratio on windows of size 7

To improve the quality of the modeling, we also normalized separately for each person each of the 25 features described in [10], in order to give an equivalent standard deviation to each of them. This guarantees that each parameter contributes with the same importance to the emission probability computation performed by each state on a given feature vector. This also permits a better training of the HMM, since each Gaussian marginal density is neither too flat nor too sharp. Indeed, if it is too sharp, for example, it will not tolerate variations of a given parameter in genuine signatures or, in other words, the probability value will be quite different on different genuine signatures. For more details, the reader should refer to [10].

The Baum-Welch algorithm [1] is used for parameter reestimation. In the verification phase, the Viterbi algorithm [1] permits the computation of an approximation of the log-likelihood of the input signature given the model, as well as the sequence of visited states (called "most likely path" or "Viterbi path").

On a particular test signature, we compute the distance d_l (Likelihood distance) between its log-likelihood and the average log-likelihood on the training database. This distance is then shifted to a similarity value s_l (Likelihood Score) between 0 and 1, by the use of an exponential function:
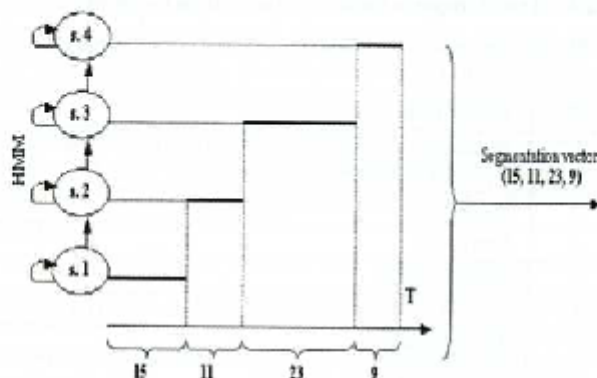
FIG 1. – Computation of the Segmentation Vector.
Obtention du vecteur de segmentation.

$$s_l = \exp\left(\frac{-d_l}{n_{par}}\right),$$ where n_{par} denotes the number of parameters describing the signature.

Given a signature's most likely path, we consider a N-components *Segmentation Vector*, N being the number of states in the claimed identity's HMM. This vector has in the ith position, the number of feature vectors that were associated to state i by the Viterbi path (see Figure 1). We then characterize each of the training signatures by a *Segmentation Vector*. In the verification phase, as shown in Figure 2, for each test signature, we compute the City Block distance d_v between its associated *Segmentation Vector* and all the *Segmentation Vectors* of the training database, and we average such distances. This average distance is then shifted to a similarity measure s_v between 0 and 1 (Viterbi Score) by an exponential function, as follows:

$$s_v = \exp\left(\frac{-d_v}{T_{avg}}\right),$$ where T_{avg} denotes the average length of a client's enrollment signatures.

This score normalization is intrinsic to the system and only exploits information from the client's enrollment signatures.
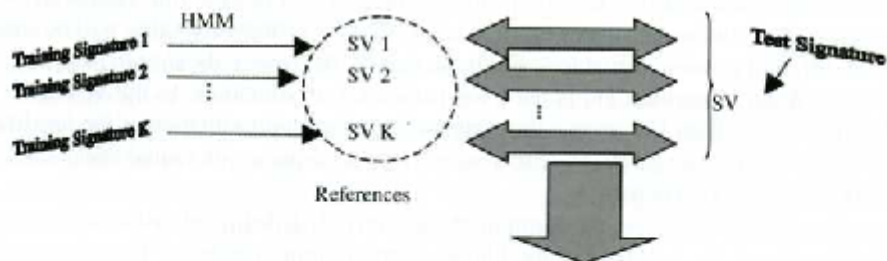

FIG 2. – Exploitation of the Viterbi Path information. sv stands for Segmentation Vector.
Exploitation de l'information du chemin de Viterbi. SV signifie Vecteur de Segmentation.

Finally, on a given test signature, these two similarity measures (s_t and s_v), are fused by a simple arithmetic mean.

This system was evaluated during svc'2004, the First International Signature Verification Competition, in the two tasks of the competition [12]; it gave very good results in Task 2, but its best result, obtained in Task 1, unfortunately could not be reported in [12] because of "incomplete results" (the system could not process 10 signatures from the evaluation set). Indeed, if we consider those 10 samples as classification errors of the system, our system is then classified in position 2, which is an excellent result.

II.1.2. System 4

On-line Signature Verification System 4 is based on functional feature extraction and Hidden Markov Models (HMMs) [1]. This system was used by UPM in the First International Signature Verification Competition (SVC 2004) with excellent results [12]: in Task 2 of the competition, where both trajectory and pressure signals were available, System 4 was ranked first when testing against random forgeries. In case of testing with skilled forgeries, System 4 was only outperformed by the winner of the competition, which was based on Dynamic Time Warping [13].

Below we provide a brief sketch of System 4, for more details we refer the reader to [11].

Feature extraction is performed as follows. Coordinate trajectories (x_n, y_n) and pressure signal p_n are the components of the unprocessed feature vectors, where $n = 1, ..., N_s$ and N_s is the duration of the signature in time samples. In order to retrieve relative information from coordinate trajectories (x_n, y_n) and not being dependent on the starting point of the signature, signature trajectories are preprocessed by subtracting the center of mass. Then, a rotation alignment based on the average path tangent angle is performed. An extended set of discrete-time functions is derived from the preprocessed trajectories. The resulting functional signature description consists of the following feature vectors (x_n, y_n, p_n, θ_n, v_n, ρ_n, a_n, $\dot{x}_n$, $\dot{y}_n$, $\dot{p}_n$, $\dot{\theta}_n$, $\dot{v}_n$, $\dot{\rho}_n$, $\dot{a}_n$) with $n = 1, ..., N_s$, where the upper dot notation represents an approximation to the first order time derivative and θ, v, ρ, and a stand respectively for path tangent angle, path velocity magnitude, log curvature radius and total acceleration magnitude. A whitening linear transformation is finally applied to each discrete-time function so as to obtain zero mean and unit standard deviation function values.

Given the parameterized enrollment set of signatures of a user, a continuous left-to-right HMM was chosen to model each signer's characteristics: this means that each person's signature is modeled through a double stochastic process, characterized by a given number of states with an associated set of transition probabilities, and, in each of such states, a continuous density multivariate Gaussian mixture. No transition skips between states are permitted. The Hidden Markov Model (HMM) λ is estimated by using the Baum-Welch iterative algorithm. Given a test signature parameterized as $\mathbf{O}$ (with a duration of N time samples) and the claimed identity previously enrolled as λ, the similarity matching score s

$$s = \frac{1}{N^s} \log p(\mathbf{O} \mid \lambda)$$

is computed by using the Viterbi algorithm.

II.2. Approaches based on Distance measures

This section deals with two approaches based on distance measure functions. One is based on the comparison of a test signature to a reference by an Adapted Levenshtein Distance of two character strings [14], called Reference System 2 (Ref 2) and developed by the University of Magdeburg. The other is a distance measure between two feature vectors representing the test and the reference signatures respectively by a series of statistical features, called Reference System3 (Ref 3).

II.2.1. Reference System 2

The basis for this algorithm is the transformation of the dynamic handwriting signals (position, pressure and velocity of the pen) into a character string and the comparison of two character strings according to the Levenshtein distance method by V. I. Levenshtein [2]. This distance measure determines a value for the similarity of two character strings. To get this character strings, the online signature sample data must be transferred into a sequence of characters as described by Schimke & al. [14]. From the raw data of the writing (pen position and pressure) the pen movement can be interpolated and other signals can be determined such as the velocity. In order to transfer a signature into a string, we use the local extrema (minimum, maximum) of the function curves of the pen movement. We name the occurrence of such an extreme value as an *event*. Another type of event is the gap after each segment of the signature. A segment is the signal between the pen-down and the subsequently following pen-up. Short segments are another type of events. It is not possible for short segments to determine extreme points because too few data are available. These events can be subdivided into single points and segments from which the stroke direction (e.g. from left to right) can be determined. We analyze the pen movement signals, extract the feature events f and arrange them in temporal order of their occurrences in order to achieve a string-like representation of the signature. An overview of the described events ε is represented in Table II.

TABLE II. – The possible event types.

Les différents types d'événements possibles.

E-Code	S-Code	Description
$\varepsilon_1 \cdots \varepsilon_6$	$x\,X\,y\,Y\,p\,P$	x-min, x-max, y-min, y-max, p-min, p-max
$\varepsilon_7 \cdots \varepsilon_{12}$	$vx\,Vx\,vy\,Vy\,vV$	v_x-min, v_x-max, v_y-min, v_y-max, v-min, v-max
$\varepsilon_{13} \cdots \varepsilon_{14}$	$g\,d$	gap, point
$\varepsilon_{15} \cdots \varepsilon_{22}$		short events; directions: , $\uparrow$, $\nearrow$, $\rightarrow$, $\searrow$, $\downarrow$, $\swarrow$, $\leftarrow$, $\nwarrow$

At the transformation of the signature signals the events are encoded with the characters of the column 'S-Code' resulting in a event string: positions are marked with x and y, pressure with p, velocities with v, v_x and v_y, gaps with g and points with d. Points are temporally very short strokes for which no velocity or direction can be determined because they are overlapped or the writing duration lies below a fixed value. Maximum values are encoded by

capital letters and minimum values by lower case letters. A difficulty at the transformation is the simultaneous appearance of extreme values of the signals because here no temporal order can be determined. This problem of simultaneous events can be treated by the creation of a combination event, requiring the definition of scores for edit operations on those combination events. In addition, a normalization of the string lengths is also required because the lengths of the strings can be different at the biometric authentication by fluctuations of the biometric input. In this approach, an additional normalization of the distance is performed due to the possibility of different lengths of the two string sequences.

The signals of the pen movement are represented by a sequence of characters now. Starting out from the assumption that similar strokes have also similar string representations, a biometric authentication based on signatures can be carried out by using the Levenshtein Distance.

The Levenshtein Distance determines the similarity of two character strings by the transformation of one string into the other one by using operations on the individual characters. For this transformation a sequence of the operations *insert*, *delete* and *replace* is applied to every single character of the first string to transfer this into the second string. The distance between the two strings is the minimal number of edit operations at the transformation. Another possibility is the use of weights for each edit operation. The weights depend on the assessment of the individual operations. For example, it is possible to weight the deletion of a character higher then replacing with another character. A weighting with respect to the individual characters is also possible. A formal description of the algorithm is given by the following recursion:

$$
\left.
\begin{aligned}
D(i, j) &:= \min[D(i, -1, j) + w_d, \\
&\qquad D(i, j - 1) + w_i, \\
&\qquad D(i - 1, j - 1) + w_r] \\
D(i, 0) &:= D(i - 1, 0) + w_d \\
D(0, j) &:= D(0, j - 1) + w_i \\
D(0, 0) &:= 0
\end{aligned}
\right\} \forall i, j > 0
$$

In this description, i and j are lengths of the strings S_1 and S_2 respectively. The weights of the operations *insert*, *delete* and *replace* are w_i, w_d and w_r. The weight w_r is 0 if characters $S_1[i] = S_2[j]$. A smaller distance D between S_1 and S_2 denotes greater similarity than a larger distance.

II.2.2. Reference System 3

Reference System 3 (Ref3) is based on a series of statistical features shown below, and is used principally to provide a simple non-optimised baseline for performance comparison against the more highly tuned, high performance systems. The features represent the proposed set for a developing standard within the biometrics community; a process which requires performance characteristics to be established.

The features extracted from each signature were:
a) The standard deviation of all sampled X values – prior to rotation
b) The standard deviation of all sampled Y values – prior to rotation

c) 1000 x {1 + The correlation coefficient of X and Y values to three significant digits}
d) The total signature/sign time in milliseconds
e) The total in-contact signature/sign time (Force (F) > 0) in milliseconds
f) The mean of all sampled Force (F) values which are greater than 0
g) The standard deviation of all sampled F values which are greater than 0
h) The mean of all sampled Azimuth (Az) angles in degrees
i) The standard deviation of all sampled Az angles in degrees
j) The mean of all sampled Elevation (El) angles in degrees
k) The standard deviation of all sampled El angles in degrees
l) 1000 x {1 + The correlation coefficient of Az and El angles to three significant digits}
m) Total number of pen down pen up sequences
n) Pen distance in X-axis divided by the pen distance in the Y-axis

The Canberra distance measure was used to evaluate the similarity between a reference and a sample signature:

$$d_{xy} = \frac{\sum_{i=1}^{n} \frac{|x_i - y_i|}{|x_i| + |y_i|}}{n}$$

where:

    x = reference statistics
    y = sample statistics
    i = feature number
    n = total number of features

In this metric, the numerator represents the dissimilarity and the denominator normalizes this dissimilarity. Therefore the result will never exceed one and there will be no scaling effect.

Calculations are made between pairs of statistical features extracted from a reference signature (known a priori to be genuine) and a test signature (whose authenticity will be found out a posteriori). The metric provides a final distance result, which is in the range 0 to 1. If such distance is lower than the value of the decision threshold the claimed identity is accepted, otherwise it is rejected.

II.3. Fusion Techniques

The individual system scores are combined using a simple Arithmetic Mean Rule (AMR) after performing a normalization of these scores. Indeed, the aim of normalization is to obtain comparable scores in order to attain good results through simple fusion rules. Two types of normalization are studied: the first one is based on the Min-Max normalization [7] and the second one, called Bayes normalization, uses a posteriori class probabilities [6].

The *"Min-Max" normalization* of score s of one unimodal expert is defined as $n = (s-m)/(M-m)$ where M is the maximum and m is the minimum. We consider the mean (μ) and standard deviations (σ) of both the client and impostor distributions in the training database, and set: $m = \mu_{imp} - 2\sigma_{imp}$ and $M = \mu_{cl} + 2\sigma_{cl}$. Indeed, assuming that genuine and impostor scores follow Gaussian distributions, 95% of the values lie in the $[\mu - 2\sigma, \mu + 2\sigma]$ interval; following this model, our choice of m and M permits to cover most of the scores. Values higher than M or lower than m are thresholded. This normalization maps the score in the [0,1] interval.

Finally, *Bayes normalization* uses the a-posteriori client class probability $P(C \mid s)$ given score s, as a normalized score. A-posteriori probabilities are obtained using Bayes' rule as follows:

$$P(C \mid s) = \frac{P(s \mid C)\, P(C)}{P(s \mid C)\, P(C) + p(s/I)\, P(I)}$$

where $P(C)$ and $P(I)$ are the client and impostor priors, and $P(s \mid C)$ and $p(s \mid I)$ are the client and impostor likelihoods. Conditional probability densities are computed from Gaussian score distributions whose parameters are estimated on the training database. Assuming independence between the two scores s_1 and s_2, and following [6], we compute the arithmetic mean of $P(C \mid s_1)$ and $P(C \mid s_2)$.

III. EXPERIMENTAL SETUP

III.1. MCYT brief description and the associated protocol for each system

III.1.1. MCYT Signature Database

The number of existing large public databases oriented to the performance evaluation of recognition systems in On-line Signature is quite limited. The MCYT project, however, is a large bimodal corpus consisting of fingerprints and on-line signatures from more than 300 subjects [3]. A subset of 100 subjects of the MCYT database was made freely available at the end of 2003[1]. In this Section, we give a brief description of the signature corpus of MCYT. Since this corpus is the largest existing On-line western Signature database compared to other existing available corpus [15], it was chosen as the reference benchmark data within BioSecure On-line Signature Verification Evaluation platform.

In order to acquire the dynamic signature sequences, a WACOM pen tablet, model INTUOS A6 USB was employed. The pen tablet resolution is 2 540 lines per inch (100 lines/mm), and the precision is 0.25 mm. The maximum detection height is 10 mm (pen-up movements are also considered), and the capture area is 127 mm (width) $\times$ 97 mm (height). This tablet provides the following discrete-time dynamic sequences: position x_n in x-axis, position y_n in

---

1. This subset can be obtained following the instructions at http://atvs.ii.uam.es

y-axis, pressure p_a applied by the pen, azimuth angle γ_a of the pen with respect to the tablet, and altitude angle ϕ_a of the pen with respect to the tablet. The sampling frequency is set to 100 Hz. The capture area is further divided into 37.5 mm (width) × 17.5 mm (height) blocks which are used as frames for acquisition.
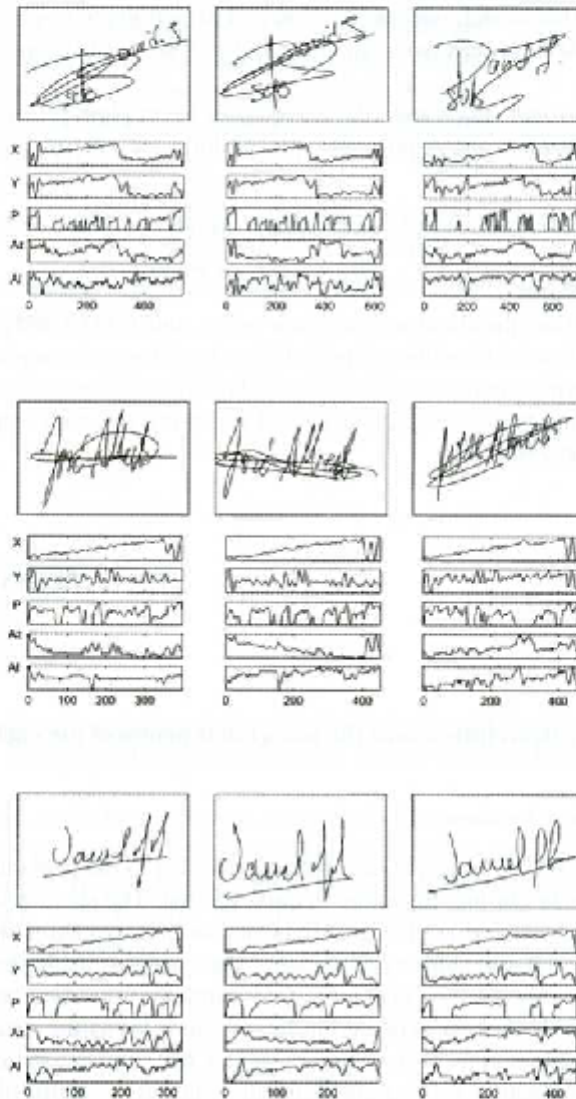

FIG 3. – Signatures from MCYT database corresponding to three different subjects. For each subject, the two left signatures are genuine and the one on the right is a skilled forgery. Plots below each signature correspond to the available information, namely: position trajectories, pressure, and pen azimuth and altitude angles.

Signatures de la base MCYT correspondant à trois personnes. Pour chaque personne, les deux signatures à gauche correspondent aux signatures authentiques et celle à droite à une vraie imitation. Les figures situées en dessous correspondent aux informations disponibles: coordonnées sur la trajectoire, pression et angles d'inclinaison du stylo.

The signature corpus comprises genuine and shape-based highly skilled forgeries with natural dynamics. In order to obtain the forgeries, each contributor is requested to imitate other signers by writing naturally, without artifacts such as breaks or slowdowns. The acquisition procedure is as follows. User n writes a set of 5 genuine signatures, and then 5 skilled forgeries of client n-1. This procedure is repeated 4 more times imitating previous users n-2, n-3, n-4 and n-5. Taking into account that the signer is concentrated in a different writing task between genuine signature sets, the variability between client signatures from different acquisition sets is expected to be higher than the variability of signatures within the same set. As a result, each signer contributes with 25 genuine signatures in 5 groups of 5 signatures each, and is forged 25 times by 5 different imitators. The total number of contributors in MCYT is 330. Therefore the total number of signatures present in the signature database is $330 \times 50 = 16\,500$, half of them genuine signatures and the rest forgeries.

Some signatures samples from MCYT database are shown in Figure 3.

III.1.2. Individual systems' Evaluation Protocol

The first fifty (50) writers of the MCYT database are considered as the Development Set. This set was used by the different teams involved to tune their systems, for example for choosing the best combination of parameters, or the best topology of the system in the case of HMM-based approaches. The remaining 280 persons are considered as the Evaluation Set.

In the Evaluation Set, the first five genuine signatures of each writer in the MCYT database are used to build each client's model or references. The remaining 20 genuine signatures, together with the 25 skilled forgeries and 279 random forgeries (6th signature of each other client of the Evaluation Set) are used for test purposes.

III.2. Description of the Fusion Protocol and the different experiments

III.2.1. Protocol

As mentioned in Section III.1.2, the first 50 writers of the MCYT database are considered as the Development Set, called in the following "MCYT-50", and the remaining 280 users as Evaluation Set.

The Evaluation Set is split into two parts of equal size (140 persons) designated respectively the Fusion Learning Set (FLS) and Fusion Test Set (FTS). The first is used to compute normalization factors that are applied to the Fusion Test Set, in order to test the fusion system.

We have chosen a Cross-Validation (CV) procedure because it permits us to obtain results on the entire database instead of only on a predefined test subset. We consider a 2-fold Cross-Validation (CV) protocol [9]. It consists in splitting the database in 2 subsets $S1$ and $S2$ and in first using $S1$ as Learning Set (FLS) and $S2$ as Test Set (FTS) and then in interchanging their roles, that is ($FLS = S2$, $FTS = S1$). Several splits must be considered to reduce the bias effect related to one particular split.

For each split, instead of computing error rates (FAR, FRR) in each of the two steps above described (Step 1: ($FLS = S1$, $FTS = S2$) and Step 2: ($FLS = S2$, $FTS = S1$)), we compute global error rates (FAR, FRR) on the whole Evaluation Set. This way, for a given value of the decision

threshold, a single couple of error values (FAR, FRR) is associated to each split of the database. Then, for each split, by varying the value of the decision threshold, we obtain a DET curve [16].

We consider 50 different splits leading to 50 DET curves. For each value of the decision threshold, we also compute an average error rate over the 50 splits (error rates are directly averaged).

III.2.2. Experiments

In our experiments we study combinations of pairs of systems and compare them to the four individual systems here presented in order to have some insight on systems' complementarity. Then the best combinations of two systems are compared to the combination of all the four systems, and also to the combination of the three Reference Systems. Two schemes are studied in the following: one considering both skilled and random forgeries simultaneously, the other considering only skilled forgeries.

III.3. Systems' optimization on the Development Set

This Section is devoted to the optimization of systems on the Development Set, namely MCYT-50.

In fact, Reference System 3 (the baseline system) has no structural parameters to tune before evaluation. Tuning of the system would indeed constitute the selection of a subset of features; this alternative was not chosen because the aim of the system was precisely to test a proposed set of features for a developing standard within the biometrics community. On the other hand, The Adapted Levenshtein Distance algorithm proposed by Magdeburg University as a Reference System, is based on rules which convert an online handwriting signal into a string. It is very difficult to improve or to enlarge these rules to a given set. For this reason, it was chosen not to tune the algorithm to MCYT-50.

Concerning INT's HMM-based approach, no optimization of the system was performed in order to measure the system's performance on totally unseen data.

In this Section, we report configuration experiments of On-line Signature System 4 on the Development Set. The two experiments described in the following are related to the configuration of the functional feature set and the modeling complexity, respectively.

III.3.1. System 4 Optimization on MCYT-50: Functional Feature Set

The configuration for the functional feature extraction experiment is based on the previous work reported in [17]. The modeling complexity is fixed to 4 HMM states and 8 Gaussian mixtures per state. Training data of each user consist of 5 training signatures, each one from a different acquisition set. Results are given as the average individual EER for all the 50 signers in the Development Set, considering all the skilled forgeries available for each user. Each individual EER is computed following the operational procedure introduced in [18]. Results for different functional sets are shown in Figure 4.

We first observe that although pen inclination signals have shown discriminative capabilities in other works [19], the inclusion of these two functions worsens the verification performance in the Signature System 4. In particular, average EER decreases from 10.37% to 4.54% when pressure signal is included to the basic position trajectory information, but increases to 6.28% and 4.84% when azimuth and altitude signals are respectively considered.

In Figure 4 we also show the verification performance for an increasing number of extended functions, computed over the basic set $\{x, y, p\}$. In particular, when path tangent angle θ, path velocity magnitude v, log curvature radius ρ, and total acceleration magnitude a are progressively included, 2.57%, 1.99%, 1.44%, and 0.68% average EERs are obtained. The set composed by these 7 functions $\{x, y, p, \theta, v, \rho, a\}$ will be referred to as w. The verification performance result for the final functional configuration of System 4 consisting of the 7 functions in w and their first order time derivatives Δw is also plotted in Figure 4.

III.3.2. System 4 Optimization on MCYT-50: Modeling Complexity

The functional feature set is fixed to $\{w, \Delta w\}$. In order to configure the system to have good generalization capabilities, we use here 5 training signatures from the first acquisition set, and test with the remaining ones. Results are given as the average individual EER for the 50 users in the signature development corpus considering all the skilled forgeries of each user, as in the previous functional set experiment. The topology of the HMM is fixed to left-to-right transitions without skipping states. The modeling complexity is related then to the number of states N and number of Gaussian mixtures per state M. Verification performance results are given in Table III.
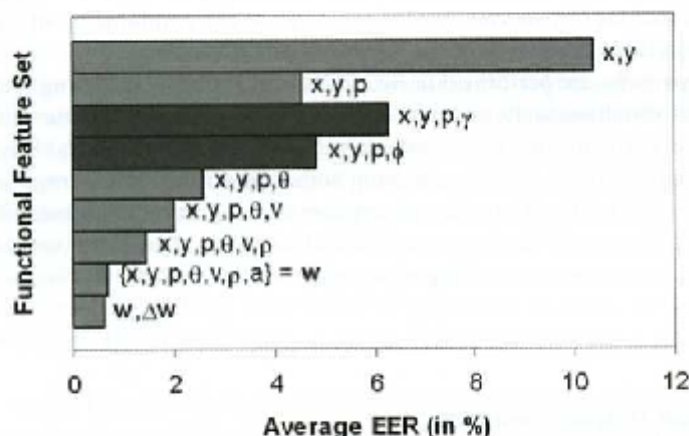

FIG 4. – Functional Feature Extraction experiments. Verification performance results for skilled forgeries are given for various function sets including: position trajectories x and y, pressure p, azimuth γ, altitude ϕ, path tangent angle θ, path velocity magnitude v, log curvature radius ρ, total acceleration magnitude a and their first order time derivatives Δ.

Expériences d'extraction de caractéristiques fonctionnelles. Les performances de vérification sont données pour différents ensembles de caractéristiques fonctionnelles et en considérant de vraies imitations.

From Table III, we first observe that, for a fixed number of states N, the greater the number of mixtures per state M, the lower the error until the minimum error is reached. Similarly, for a fixed number of mixtures per state M, the greater the number of states N the lower the error until the minimum error is reached. We also observe that the best configuration of System 4 from all the tested instances is $N = 2$, $M = 32$.

TABLE III. – Average EER for skilled forgeries (in %) for different HMM complexity configurations. N = number of states, M = number of Gaussian mixtures per state.

EER moyen pour les vraies imitations (en %) et différentes configurations de complexité du MMC (Modèle de Markov Caché). N = Nombre d'états, M = Nombre de Gaussiennes par état.

N＼M	2	4	8	16	32	64
2	-	-	1.51	0.74	0.30	0.44
4	-	1.64	0.87	0.52	0.48	-
8	1.81	0.79	0.76	0.35	-	-
16	1.20	0.96	0.74	-	-	-
32	0.97	-	-	-	-	-

IV. ANALYSIS OF RESULTS

In this Section we present the results from analyzing pairs of systems and compare them to the four individual systems in order to provide an insight on systems' complementarity. The best combinations of two systems are compared to the combination of all the four systems, and also to the combination of the three Reference Systems.

These comparisons are performed in two frameworks: first, considering both random and skilled forgeries simultaneously, and then considering only skilled forgeries. In the first case, only a Min-Max normalization is performed; in the second, Min-Max and Bayes normalizations are compared. Indeed, when considering both random and skilled forgeries, Bayes normalization associated with a Gaussian assumption of the client and impostor distributions is not well suited, because of the impostor distribution shape. Therefore, we consider in this case a Min-Max to perform a rescaling of scores.

IV.1. Individual systems' comparison

In Table IV, the four individual systems' (Ref1 for Reference System 1, Ref2 for Reference System2, Ref3 for Reference System 3 and Sys4 for System 4) performance are presented at the Equal Error Rate (EER) point. For more insight, we also report the performance of the two modules of Ref1 system, denoted by Ref1-Lik for the system giving as output the Likelihood Score, and Ref1-Vit for the system giving as output the Viterbi Score.

TABLE IV. — Individual systems' performance at the EER point considering both random and skilled forgeries.

Performances des systèmes au point EER en considérant à la fois les vraies imitations et les imitations aléatoires.

System	Ref1	Sys4	Ref1-Vit	Ref1-Lik	Ref2	Ref3
EER	2.91 %	4.3 %	4.6 %	6.61 %	9.18 %	11.5 %

We first notice that the best system is the statistical system based on the fusion of the two scores above mentioned, the Likelihood Score and the Viterbi Score (both descended from the same Hidden Markov Model), that is the Reference System 1. It is worth noting first that this score normalization is intrinsic to the system as explained in Section II.1.1; also, it is a personalized score normalization because it only exploits information from the client's enrollment signatures.

This result is followed by the performance of System 4, also based on a Hidden Markov Model. It is interesting to notice that System 4, whose output score is the Log-likelihood of the test signature given the model, does better than Ref1's Likelihood Score alone (Ref1-Lik in Table IV) and also than Ref1's Viterbi Score alone (Ref1-Vit in Table IV).

IV.2. Individual systems' combination

In Table V, we show results of combinations of systems reporting the Equal Error Rate (EER) over 50 splits of the database, since a normalization (Min-Max) of scores is performed before fusing the scores of different systems. We also report the standard deviation over such 50 splits at the EER point to evaluate the confidence of results.

Table V shows that the best result seems to be the combination of all systems, closely followed by the combination of the two systems based on HMMs (Ref1+Sys4). In fact, the standard deviation of errors over 50 splits of the database shows that the difference between the two is not significant: indeed, the difference between the mean EER over 50 trials for both combinations of systems is of the same order of magnitude than the standard deviations reported in both cases. Therefore, we may conclude that the two combinations of systems are equivalent.

For more insight, Figure 5 shows the relative performance of the individual systems and the two best combinations: all the four systems and the two HMM-based approaches. We confirm that the fusion of both HMM-based approaches is equivalent to that of the four systems for any value of the threshold.

TABLE V. – Performance of Systems' combination at the EER point after Min-Max normalization.

Performances des combinaisons de systèmes au point EER après une normalisation des scores par la méthode du Min-Max.

Systems Combination	Min-Max normalization	
	EER %	Std EER %
All: Ref1+Ref2+Ref3+Sys4	1.22	0.03
Ref1+Sys4	1.28	0.04
Ref1-Vit + Sys4	1.55	0.04
Ref1-Lik + Sys4	2.11	0.04
Ref1+Ref2+Ref3	2.12	0.05
Ref1+Ref2	2.35	0.06
Sys4+Ref2	2.62	0.06
Ref1 +Ref3	2.88	0.04
Ref1-Lik + Ref1-Vit	2.89	0.04
Ref1-Vit + Ref2	3.07	0.07
Ref3 +Sys4	3.55	0.07
Ref1-Lik + Ref2	3.79	0.10
Ref2 +Ref3	5.14	0.11

For more insight, Figure 5 shows the relative performance of the individual systems and the two best combinations: all the four systems and the two HMM-based approaches. We confirm that the fusion of both HMM-based approaches is equivalent to that of the four systems for any value of the threshold.
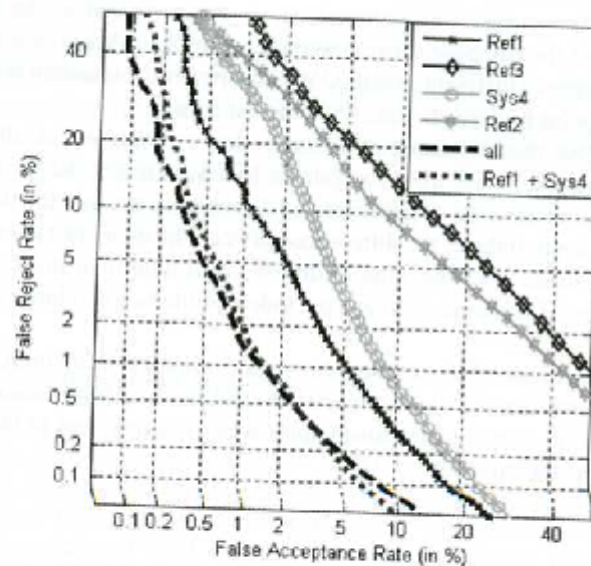

FIG 5. – DET curves of the four Individual systems and their 2 best combinations.
Courbes DET pour les 4 systèmes seuls et leurs 2 meilleures combinaisons.

IV.2.1. HMM-based systems and their possible associated combinations

When Ref1's scores are combined separately to System 4, we notice that one of the two combinations is significantly better than the other: (Ref1-Vit + Sys4) reaches 1.55% of EER while (Ref1-Lik + Sys4) shows a higher EER (2.11%). It seems indeed that the two Likelihood information from both HMM-based systems are less complementary than Sys4's Likelihood and Ref1's Viterbi Score. This can be explained by the fact that the Viterbi Score corresponds to another level of description of signatures than the one that characterizes the Likelihood. Indeed, the Viterbi Score corresponds to an intermediate level of description, that of portions of the signature that are the outcome of the segmentation performed by the target model on the signature. The Likelihood is an average over the whole signature of very local "scores", namely emission probabilities. This averaging effect introduces, in fact, a certain loss of local information; it is why the Viterbi Score that keeps local information at the level of segments is so complementary. It helps particularly the system to discriminate impostors when the length of the impostor signature is different from the average length of the client's signatures: indeed, the more the respective lengths of client's and impostors' signatures are different, the more their respective segmentations performed by the client's model will be different, and thus the more the discrimination capabilities of the system are enhanced.

Also, for comparison purposes, we studied the combination of the two scores from Ref1 (Ref1-Lik + Ref1-Vit) after a Min-Max normalization as the other combinations considered. We see in Table V that any combination of Sys4's Likelihood score to one of the two scores from the other HMM, Ref1-Lik or Ref1-Vit, gives a better result than combining Ref1-Lik and Ref1-Vit, as done in Ref1. This can be explained by the fact that the two HMM-based systems extract different information from a signature, and model such information in a different way, although both HMMs have the same type of topology (Left-to-Right with no skips), and the same model for the emission probability (Gaussian mixture). The two HMMs differ in some of the features extracted from the signature (Ref1 uses more local shape-related features whereas Sys4 uses mainly dynamic information), the number of states (Ref1 uses a variable number of states according to the client's enrollment signatures whereas Sys4 uses only 2 states), and the number of Gaussian components per state (4 in the case of Ref1 and 32 in the case of Sys4). Figure 6 shows the corresponding DET curves that confirm these remarks for any value of the threshold: there exists in fact a tangible complementarity of the two HMM-based systems.

To conclude about the relative quality of the two HMM-based approaches, certainly Sys4's Likelihood score is a better system than Ref1's Likelihood Score alone, but Ref1's Viterbi Score is very effective when combined to the Likelihood information, particularly in increasing the discrimination capabilities of the system to impostors.

Finally, Figure 7 shows the performance of the fusion of the 3 Reference Systems (Ref1+Ref2+Ref3) compared to the fusion of both HMM-based approaches (Ref1+Sys4). We notice that, close to the EER point, combining the two HMM-based approaches (Ref1+Sys4) lowers the error rate by a factor 2 compared to combining the 3 Reference Systems.

In the following, we compare the best combination of HMM-based systems (Ref1+Sys4) to the combination of distance-based systems.

IV.2.2. Comparison to the combination of distance-based systems

When combining distance-based systems (Ref2+Ref3), results are significantly improved (roughly of around 50% at the EER point) compared to both individual systems' results
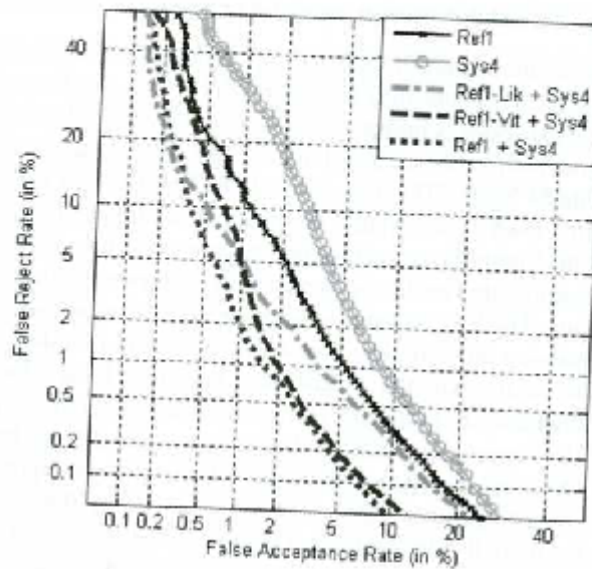
FIG 6. – DET curves of HMM-based systems and their possible associated combinations.
Courbes DET des systèmes à base de MMCS et les combinaisons associées.

(5.15% relatively to 9.18% for Ref2 and 11.1% for Ref3). This can also be seen in Figure 7 showing the corresponding DET curves. We conclude that these two systems that are simple and less computationally expensive compared to HMM-based approaches, when put together, give state-of-the-art performance. This is an interesting result. Of course, HMM-based approaches are certainly computationally more demanding, but they are more fine: this is apparent when we recall that the combination of the two HMMs (Ref1+Sys4) give a result that is, at the EER point, roughly 4 times better than the result obtained by the combination of the two distance-based approaches (Ref2+Ref3), as shown in Table V. This can be confirmed in Figure 7 when comparing the DET curves of (Ref2+Ref3) system to (Ref1+Sys4) system.

IV.2.3. Combination of HMM-based and distance-based systems

We now analyze results after performing fusion of HMM-based approaches to distance-based approaches. The best result is obtained by fusing Ref1 with Ref2 (Ref1+Ref2), although the improvement obtained relatively to Ref1 alone is rather low (2.35% compared to 2.93% at the EER point). Figure 8 shows indeed results obtained by combining Ref2 to the two HMM-based systems separately, (Ref1+Ref2) and (Ref2+Sys4), compared to the combination of the two HMM-based systems (Ref1+Sys4) and the individual systems involved.

On the other hand, when fusing System 4 with Ref2 (Ref2+Sys4), the relative improvement with respect to System 4 alone is significant (2.62% compared to 4.16% at the EER point). This result is confirmed in Figure 8. It is interesting because it shows the power of fusion : certainly, fusion always improves results since it adds an extra dimension to our
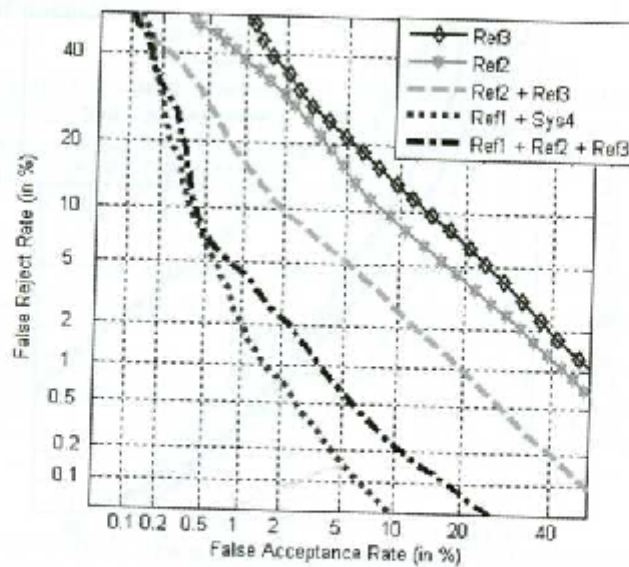
FIG 7. – DET curves of the distance-based systems and their combination compared to the combination of HMM-based systems and the 3 Reference Systems.

Courbes DET pour les systèmes à base de distances et leurs combinaisons, comparés aux combinaisons des systèmes à base de MMCs et des systèmes de référence.

discrimination problem, separating clients from impostors; but, in some cases, in which complementarity exists between systems, a system that is less discriminant (Ref2 in this case) improves the other (Sys4 in this case) of about 50%. This is not observed in the (Ref1+Ref2) combination probably because few extra information is brought by the edit-distance approach after the intrinsic fusion of Ref1, Likelihood Score with Viterbi Score.

Concerning the fusion of HMM-based approaches to the other distance-based approach (Ref3), the combination of Ref3 with Ref1 gives better results than with System 4 (2.88% compared to 3.55% at the EER point). Figure 9 shows the comparison of these combinations between them and to the individual systems, and in particular to the combination of the two HMM-based approaches. On the other hand, the resulting system (Ref1+Ref3) is less discriminant than Ref1 combination with the other distance-based approach (Ref1+Ref2). Indeed, Ref2 is an elastic distance applied on strings encoding the signatures (test and reference), while Ref3 is simpler: a distance measure normalized to the [0,1] interval applied on features extracted on the signatures.

In the following section, we study the case in which only skilled forgeries are considered. In this case, in order to perform Bayes normalization, previously described in Section II.3, we assume that the client and impostor score distributions are Gaussian.
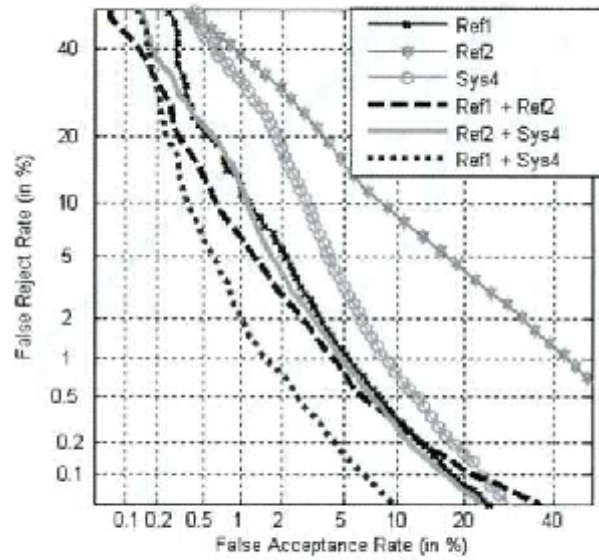
FIG 8. – Fusion of Ref2 system with HMM-based approaches.

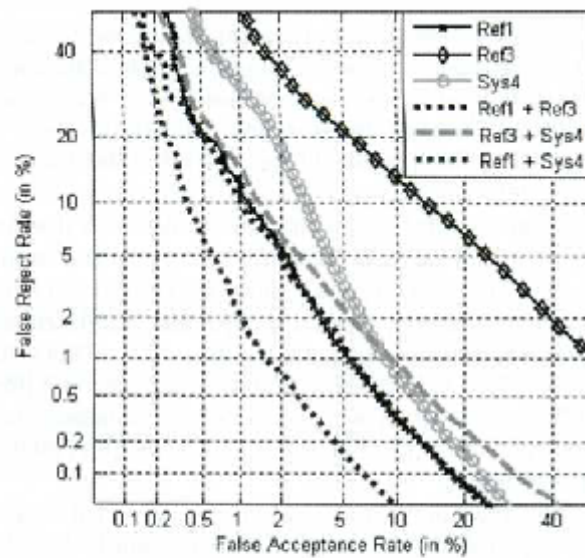*Figure 8. Fusion du système Ref2 avec ceux basés sur des approches à base de MMCs.*

FIG 9. – Fusion of Ref3 system with HMM-based approaches.

Fusion du système Ref3 avec ceux basés sur des approches à base de MMCs.

IV.3. Effects of normalization

In the following, the same approach will be kept to present results: combinations of couples of systems are first compared to the four individual systems; then, as before, the best combinations of two systems are compared to the combination of all the four systems, and also to the combination of the three Reference Systems.

TABLE VI. – Individual systems' performance at the EER point when only skilled forgeries are considered.

Performances des systèmes au point EER en considérant uniquement les vraies imitations.

System	Ref1	Ref1-Vit	Sys4	Ref1-Lik	Ref2	Ref3
EER %	5.73	6.70	8.39	10.83	15.89	19.55

First, we notice in Table VI that performance decreases with respect to that reported in Table IV, which is a normal phenomenon, since the task is more difficult for systems when considering only skilled forgeries.

In Table VII, the Equal Error rate (EER) reported is the average EER over 50 splits of the database. We also report the standard deviation over such 50 splits at the EER point. Two score normalizations are considered in this case: Min-Max and Bayes normalization, both previously described in Section II.3.

TABLE VII. — Performance of systems' combination with skilled forgeries at the EER point after Min-Max and Bayes normalization.

Performances des combinaisons de systèmes au point EER en ne considérant que les vraies imitations après une normalisation des scores par les méthodes du Min-Max et par les Probabilités a posteriori.

Systems' combinations	Min-Max normalization		Bayes normalization	
	EER %	STD EER %	EER %	STD EER %
REF1+SYS4	3.40	0.06	3.63	0.10
ALL: REF1+REF2+REF3+SYS4	3.50	0.09	3.44	0.10
REF1-VIT+SYS4	3.52	0.08	3.97	0.10
REF1-LIK+SYS4	5.23	0.09	6.27	0.16
REF1+REF2	5.53	0.10	5.43	0.11
REF1+REF3	5.96	0.09	5.64	0.13
REF1-VIT+REF2	6.15	0.12	7.15	0.14
REF2+SYS4	6.48	0.12	6.83	0.19
REF1-LIK+REF2	8.00	0.12	9.83	0.16
REF3+SYS4	8.29	0.11	7.49	0.20
REF2+Ref3	11.16	0.17	12.85	0.25

We notice in Table VII that the same major tendencies on individual systems and fusion systems are confirmed at the EER point. Results reported in Table VI show that the best individual system remains Ref1 (5.73%), followed in this case by Ref1-Vit (6.70%), then by Sys4 (8.39%) and finally by Ref1-Lik (10.83%). Indeed, when only skilled forgeries are used, the discrimination capabilities of Ref1-Vit are enhanced. Systems based on distance measures follow, with an error rate increase of a factor 2 of the best distance-based system (Ref2) with respect to System 4.

Concerning fusion and the effect of the normalization scheme used, the combination of the two HMM-based approaches (Ref1+Sys4) and the combination of all the systems (Ref1+Ref2+Ref3+Sys4) are the best systems, whatever the normalization scheme is. Also, Table VII shows that considering the standard deviation of the errors, the fusion of all systems is equivalent to the combination of the two HMM-based approaches (Ref1+Sys4) in both normalization schemes.

The combination of Ref1's Viterbi Score with Sys4's Likelihood score (Ref1-Vit + Sys4) remains a good system; when using Min-Max normalization it is equivalent to the best systems whereas when using Bayes normalization scheme it is less efficient than the combination of all systems.

All other combinations are far behind, above 5.2% of Equal Error Rate; in particular, rows in Table VII below the combination of Ref1 and Ref2 systems (Ref1+Ref2) give worse results than Ref1 alone.

V. CONCLUSIONS

This integrated work of four institutions of the BioSecure Network of Excellence is a first step towards a European evaluation platform for On-line Signature : three Reference Systems have been presented, with another system of an institution of the Network, Universidad Politécnica de Madrid (UPM) in Spain. Reference Systems are respectively an HMM-based approach exploiting fusion of two complementary scores from Institut National des Télécommunications (INT) in France [10], an edit-distance comparing strings describing the reference and test signatures proposed by University of Magdeburg (AMSL) in Germany [14], and a simple distance-based method coupled with a series of statistical features for baseline comparison with a non-optimised implementation based on standardised features. The fourth system evaluated is also based on HMMs, on a classical Likelihood score [1]. Both HMM-based approaches were evaluated in SVC'2004, the First International Signature Verification Competition and gave very good results [12].

The individual systems and several of their combinations were evaluated on the largest existing On-line western Signature database, the signature part of the MCYT database [3], containing 330 clients. It is the first time to our knowledge that so many systems, covering a large scope of approaches, are evaluated on such a large database. Indeed, SVC'2004 Evaluation Set contained only 60 persons [12]. Also, it is the first time in the literature that several On-line Signature Verification approaches are combined to study systems' complementarity.

Our study of the combination of such systems was carried out by means of a careful statistical protocol, 2-fold Cross-Validation [9], which permitted us to evaluate confidence on

results. Two configurations were chosen for evaluation: one considering skilled and random forgeries simultaneously, and the other considering only skilled forgeries.

The major tendencies observed on individual systems are the following: the best system is Reference System 1, based on the fusion of two scores, the Likelihood Score and the Viterbi Score, both descended from a HMM. It is followed by System 4 that performs better than the system based only on the Likelihood Score of Reference System 1. Systems based on distance measures follow, with an Equal Error Rate increase of a factor 2 of the best distance-based system (Reference System 2) with respect to System 4.

Concerning the combination of approaches, whatever the normalization scheme is, the combination of the two HMM-based approaches and the combination of all the systems are the best systems. They are followed by the combination of Viterbi Score of Reference System 1 with System 4's Likelihood score; moreover, this system becomes equivalent to the best systems when performing Min-Max normalization with skilled forgeries. This result is interesting; indeed, such two scores proved to be very complementary. All other combinations are far behind.

Furthermore, concerning the influence of normalizations, studied with skilled forgeries, Bayes normalization slightly improves the result of the fusion of all the systems, although this system remains equivalent to the best combination (the two HMM-based approaches). In the case of Min-Max normalization with skilled forgeries, the best system is the combination of the two HMM-based approaches, although closely followed by the combination of all systems. Also, this combination is equivalent in this case to Sys4's Likelihood fused to Viterbi Score of Reference System 1.

More generally, our results show a clear complementarity of the two HMM-based approaches: indeed, any combination of Sys4's Likelihood score to one of the two scores from Reference System 1 (Likelihood Score and the Viterbi Score), gives a better result than Reference System 1 itself. We conclude that the two HMMs extract different information from a signature, and model such information in a different way, although they are both Gaussian Mixture Models with the same type of topology (Left-to-Right with no skips). Furthermore, the two HMMs differ in some of the features extracted of the signature (Reference System 1 extracts some rough local spatial information with sliding windows centered on each point); but, at the same time, many similar dynamical information are extracted by both. Also, they differ in the number of states and the number of gaussian components per state. In this context of complementarity, now concerning its functioning, we remark that Sys4's Likelihood score is a better system than only the Likelihood Score of Reference System 1, and that the Viterbi Score is in fact a very good system on the MCYT database; indeed, we may recall that in a previous study of INT's team [10], it was observed that according to the database chosen, this may vary: the Viterbi Score alone may lead to higher error rates compared to Likelihood Score alone, depending on the characteristics of the database. Finally, in any case, as proven in [10] and in the present work, the Viterbi Score is really effective when combined to Likelihood information, particularly by enhancing the discrimination capabilities of the system to impostors. We may state at this point that the fusion of the two HMM approaches here presented may be considered as one of the best state-of-the-art On-line Signature Verification system (1.28% of EER on skilled and random forgeries, 3.40% of EER with only skilled forgeries and Min-Max normalization, 3.63% with only skilled forgeries and Bayes normalization).