A novel personal entropy measure confronted to online signature verification systems' performance

Nesma Houmani, Sonia Garcia-Salicetti, Bernadette Dorizzi

A Novel Personal Entropy Measure confronted with Online Signature Verification Systems' Performance

Nesma Houmani, Sonia Garcia-Salicetti, and Bernadette Dorizzi,
Institut TELECOM; TELECOM & Management SudParis;Dept EPH , Evry, France
E-mail: {Nesma.Houmani, Sonia.Salicetti, Bernadette.Dorizzi}@it-sudparis.eu

***Abstract*—In this paper, we study the relationship between a novel personal entropy measure for online signatures and the performance of several state-of-the-art classifiers. The entropy measure is based on local density estimation by a Hidden Markov Model. We show that there is a clear relationship between such entropy measure of a person's signature and the behavior of the classifier. We carry out this study on a Dynamic Time Warping classifier, a Gaussian Mixture Model and a Hidden Markov Model as well. It is worth noticing that the HMM classifier differs from the HMM used for entropy computation. Signatures were split into three categories according to their entropy value. These categories are coherent across four different databases of around 100 persons each: BIOMET, MCYT-100, BioSecure data subsets DS2 and DS3. We studied the impact of such categories on classifier's performance with a larger signature data subset of DS3, of 430 persons.**

I. INTRODUCTION

In general, a biometric system's performance is measured globally on all the available data of a database, in terms of the two types of errors that a biometric system can make, namely False Rejections and False Acceptances. Instead of considering global performance on a database, a first possible way to better tackle performance assessment is to consider the fact that the degree of recognition difficulty is not the same from one individual to another. Therefore, to have a better insight on the behavior of a classifier, it is wise to split the database in subsets, according to a criterion related to the difficulty of recognizing an individual.

Previous works for the speech face and fingerprint modalities have proposed categorizations of users in terms of their behavior with respect to an automatic recognition system [1, 2, 3, 4, 5]. Such categorizations are well-known as the Biometric Menagerie, since animal names were used for different types of users. A first classification of speakers divided the population in four non exclusive categories [1]: Sheeps (easy to recognize speakers), Goats (difficult to recognize speakers), Lambs (easy to imitate speakers) and Wolves (successful at imitating other speakers). Such classification was applied on the face modality later [2] and completed by adding Worms, Chameleons, Phantoms and Doves, according to a user's relationship between genuine and impostor match scores [3]. This last categorization was applied on the fingerprint modality. To our knowledge, no study in this direction concerned the online signature modality.

Nevertheless, such categorizations have the weakness of strongly depending on the classifier that is used. In the same direction, [6] measures the quality of a biometric signal (such as online signature) in function of the scores of a specific classifier. In this paper, we propose another way to tackle this problem for the online signature modality through a novel personal measure of the entropy of the writer (client-entropy), based on a set of genuine signatures [7]. This measure allows to categorize users according to the information content or alternatively to the "degree of disorder" or "uncertainty" of their signatures. We show that there is a strong relationship between such measure and the behavior of the user with respect to different automatic recognition systems. Therefore, it has become possible to quantify the difficulty of recognizing a person according to the value of its entropy measure.

To that end, the signature is seen as a discrete random variable described by its raw Cartesian coordinates *(x,y),* as these are the only available features in all types of databases, whether acquired on fixed platforms (as digitizing tablets) or on mobile platforms (as Personal Digital Assistants). As the online signature is piecewise stationary, it is natural to estimate the probability density locally, namely on portions of the signature. To that end, we propose to use a Hidden Markov Model [8] (HMM) as a local estimator of the probability density. To our knowledge, it is novel to measure the entropy of a single person's signature based on several instances of such signature.

Indeed, the concept of entropy was previously used in the document analysis literature to measure the complexity of a character image database [9], by defining the entropy related to each pixel in the image. In a more security oriented recent work [10], the concept of entropy is used to measure how likely it is that two persons have accidentally the same signature, based on a particular classifier. Some authors have tried to analyze in terms of other measures than entropy, the variability and complexity in signatures. In [11], both complexity and variability criteria were proposed for off-line signature verification by a human expert. Indeed, a human operator labels signatures according to both criteria and their impact on performance is studied.

Four databases are used in our study of performance assessment across personal entropy-based categories : the most two widely used in the Online Signature Verification

literature, BIOMET Signature subset [12] and the freely available subset of 100 users MCYT-100 of MCYT database [13], and two data sets containing the same 104 users but differing by the type of acquisition platform (digitizing tablet vs. mobile platform), the DS2 and DS3 subsets acquired in the framework of BioSecure Network of Excellence [14].

This paper is organized as follows: in Section 2, we present our novel client-entropy measure and the resulting categories across the four databases: BIOMET, MCYT-100, DS2 and DS3. In Section 3, we first present the three classifiers used for performance assessment across the client-entropy-based categories, namely a Dynamic Time Warping classifier [8], a Hidden Markov Model [8], and a Gaussian Mixture Model [15]. Then, we analyze results on a larger signature data subset of DS3 containing 430 persons. Finally, conclusions are stated in Section 4.

II. GENERATING SIGNATURE CATEGORIES BASED ON CLIENT-ENTROPY

A. Measuring "Client-Entropy" with a Hidden Markov Model

As mentioned above, signatures are described by their raw coordinates *(x,y)*. We consider each signature as a succession of portions, generated by its segmentation via the client-Hidden Markov Model (HMM) [8]. Therefore, we obtain as many portions in each signature as there are states in the client-HMM. Then we consider each point *(x,y)* in a given portion as the outcome of one random variable that follows a given probability mass function.

Therefore, the entropy associated to a given portion of a signature is represented by the entropy of an ensemble of outcomes of a random variable Z. Such random variable is discrete since its alphabet A has a finite number of values, as many as found in the Cartesian product XxY of all possible values of ordered pairs *(x,y)*. The cardinal of A is of course related to the resolution of the acquisition surface that may be a digitizing tablet or a Touch Screen in the case of a mobile acquisition platform.

Each outcome of $Z=(x,y)$ has a probability value attached to it; if Z has as probability mass function $p(z) = Pr(Z = z)$ where z belongs to A, its entropy is defined as:

$$P(Z) = -\sum_{z \in A} p(z).\log(p(z)) \qquad (1)$$

where log denotes the logarithm to the base 2.

Although the random variable $Z=(x,y)$ is discrete, we take advantage of the continuous emission probability law estimated on each portion by the client-HMM. Such density is a mixture of Gaussian components. This choice is motivated by the fact that the discrete version of entropy and the continuous one (called Differential Entropy) are directly related when the density is Riemann integrable [7], which is the case as we have a linear combination of Gaussian densities. Also, a more orthodox version of $H(Z)$ could have been computed by quantizing the domain of Z in bins and using as probability mass function, the density value obtained by the Mean Value Theorem in each bidimensional bin. But as the quantization is actually very fine because of the fine resolution of acquisition devices (in general 1 million pixels for the acquisition surface), we can assume that the continuous density value is close to the quantized one (the density value obtained by the Mean Value Theorem). To compute client-entropy, we consider several genuine signatures of a given user or client, namely 10, and a personalized number of states, computed as:

$$N = \frac{T_{Total}}{M * 30} \qquad (2)$$

where T_{total} is the total number of sampled points available in the training signatures, and $M=4$ is the number of Gaussian components per state.

We ensure this way that the number of sample points per state is at least 120 in order to obtain a good estimation of the Gaussian Mixture in each state (four Gaussian components). Then we compute, following (1), the entropy per portion first (after segmentation which is performed by the client-HMM), by using all the sample points belonging to each portion across the 10 instances of the person's signatures. We then average the entropy over all the portions of a signature and normalize the result by the average length of the ten signatures considered in order to generate the client-entropy measure.

B. Databases description

We used five databases in this work: the freely available MCYT subset of 100 persons [13], the BIOMET signature sub-corpus of 84 persons [12], and subsets of the online signature databases acquired in the framework of the BioSecure Network of Excellence [14]: DS2 (for Second Data Set of the whole data collection), acquired on a digitizer, and DS3 (for Third Data Set of the whole data collection), acquired on a mobile platform, a Personal Digital Assistant (PDA) [16]. Two subsets of DS3 and DS2 that we use in this work both contain data from the same 104 persons. Also, a larger subset of DS3 of 430 persons is used in our study of classifiers' performance assessment per client-entropy category. Indeed, the whole BioSecure Signature Subcorpus DS3 and DS2 are not yet publicly available but, acquired on several sites in Europe, they are the first online signature multi-session databases acquired in a mobile scenario (on a PDA) for DS3 and on a digitizer for DS2. DS3 contains the signatures of 713 persons, acquired on the PDA HP iPAQ hx2790, at the frequency of 100Hz and a touch screen resolution of 1280*960 pixels. Three time functions are captured from the PDA: x and y coordinates and the time elapsed between the acquisition of two successive points. The user signs while standing and has to keep the PDA in his or her hand. Two sessions were acquired spaced of around 5 months, each containing 15 genuine signatures. The donor was asked to perform, alternatively, three times five genuine signatures and twice five forgeries. Indeed, for skilled forgeries, at each

session, a donor is asked to imitate five times the signature of two other persons. In order to imitate the dynamics of the signature, the forger visualized on the PDA screen the writing sequence of the signature he/she had to forge and could sign on the image of such signature in order to obtain a better quality forgery both from the point of view of the dynamics and of the shape of the signature.

On the other hand, DS2 contains data from 667 persons acquired in a PC-based offline supervised scenario and the digitizing tablet WACOM INTUOS 3 A6. The pen tablet resolution is 5080 lines per inch and the precision is 0.25 mm. The maximum detection height is 13 mm and the capture area is 270mm (width) x 216mm (height). Signatures are captured on paper using an inking pen. At each sampled point of the signature, the digitizer captures at 100 Hz sampling rate the pen coordinates, pen pressure (1024 pressure levels) and pen inclination angles (azimuth and altitude angles of the pen with respect to the tablet). This database contains two sessions, acquired two weeks apart. The acquisition protocol is the same as the one used for DS3.

C. *Client categories with Entropy measure*

We performed on the four databases containing around 100 persons, a *K*-Means on the client-entropy values for different values of *K* and reached a good separation of signatures with *K=3* on all databases as shown in Figures 1 to 4, respectively on MCYT-100, BIOMET, DS2 and DS3.

We notice visually that on the four databases, the first category of signatures, those having the highest client-entropy, contains paraph-like signatures, the shortest. At the opposite, signatures in the third category, those of lowest client-entropy, are the longest and their appearance is rather that of handwriting, some are even readable. In between, we notice that signatures with medium client-entropy (second category) have sometimes more the aspect of those of highest client-entropy and sometimes the aspect of those of lowest client-entropy.

III. CLASSIFIERS FOR INTRACLASS PERFORMANCE ASSESSMENT

A. *Score computation by the three Classifiers*

Three classifiers are used in this study considering only the raw coordinates description of signatures as input data: a Dynamic Time Warping classifier [8], a Hidden Markov Model [8], and finally a Gaussian Mixture Model [15]. We consider in this study only the largest database at our disposal, namely a subset of DS3 containing 430 persons, denoted as DS3-430.

For performance assessment, we considered only random forgeries in this study, in order to fit to the evaluation conditions in other biometric modalities where in general no skilled forgeries are available [1,2,3,4,5]. Five random samplings are carried out on genuine and impostor signatures

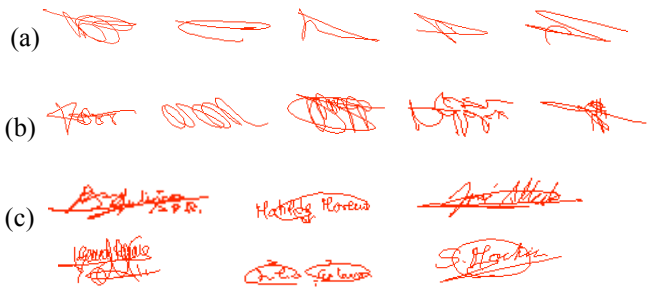

Figure 1: Examples of signatures from MCYT-100 of (a) highest, (b) medium and (c) lowest client-entropy

Figure 2: Examples of signatures from BIOMET of (a) highest, (b) medium and (c) lowest client-entropy

Figure 3: Examples of signatures from DS2 of (a) highest, (b) medium and (c) lowest client-entropy

Figure 4: Examples of signatures from DS3 of (a) highest, (b) medium and (c) lowest client-entropy

in the following way: each sampling contains five genuine signatures used as reference signatures for Dynamic Time Warping and as the training set for statistical approaches. For test purposes, the remaining 25 genuine signatures (belonging to two sessions) and 30 impostor signatures randomly

sampled in equal number in each client-entropy category (10 random forgeries per category) are used. The False Acceptance and False Rejection Rates are computed relying on the total number of False Rejections and False Acceptances obtained on the whole five random samplings.

Dynamic Time Warping determines the dissimilarity between two time sequences with different lengths [8]. This method, with polynomial complexity, computes a matching distance by recovering optimal alignments between the two time series. The alignment is optimal in the sense it minimizes a cumulative distance measure consisting of "local" distances between aligned samples. In this system, the DTW-distance between two time series $x_1 \ldots x_M$ and $y_1 \ldots y_N$ is *D(M,N)* computed as:

$$D(i,j) = \min \begin{Bmatrix} D(i,j-1) + w_p \\ D(i-1,j) + w_p \\ D(i-1,j-1) + w_p \end{Bmatrix} + d(i,j) \quad (3)$$

where the "local" distance function *d(i, j)* is the Euclidian Distance between i^{th} reference point and j^{th} testing point, with $D(0,0) = d(0,0) = 0$, and equal weights w_p are given to insertions, deletions and substitutions. We use the *Sakoe-Chiba* band constraint [8] to ensure that the warping path stays close to the diagonal of the matrix which contains the local distances *D(i,j)*. The DTW-based classifier aligns by Dynamic Time Warping (DTW) a test signature with each reference signature and the average value of the resulting five distances is used to classify the test signature as being genuine or a forgery. The dissimilarity matching score for the DTW classifier is:

$$Score = \frac{1}{N} \sum_{reference=1}^{N} D(test, reference) / L \quad (4)$$

where *N* is the number of reference signatures and *L* is the length of the test signature.

If the final distance is lower than the value of the decision threshold the claimed identity is accepted, otherwise it is rejected.

Concerning statistical models, we used a GMM and a left-to-right HMM of the same complexity in terms of Gaussian components. It is worth noticing that the HMM classifier differs from the HMM used for client-entropy computation. Indeed, the former is devoted to classification, while the latter only performs local density estimation. We considered for the HMM classifier a six states and four Gaussian components per state, as a tradeoff in complexity between the signatures of the two extreme categories. For the GMM, accordingly, we considered 24 Gaussians to model a person's signatures. The dissimilarity matching score for both statistical models is:

$$Score = \left| LL - LL_{BA} \right| \quad (5)$$

where *LL* is the Log-Likelihood of the test signature (normalized by the length of the test signature) and LL_{BA} is the corresponding average Log-Likelihood of the training signatures.

B. Results and analysis

For each classifier, we show in Figure 5 the client and impostor scores distributions per client-entropy category (high, medium and low client-entropy) and, in Figure 6, the associated DET-Curves on DS3-430 database.

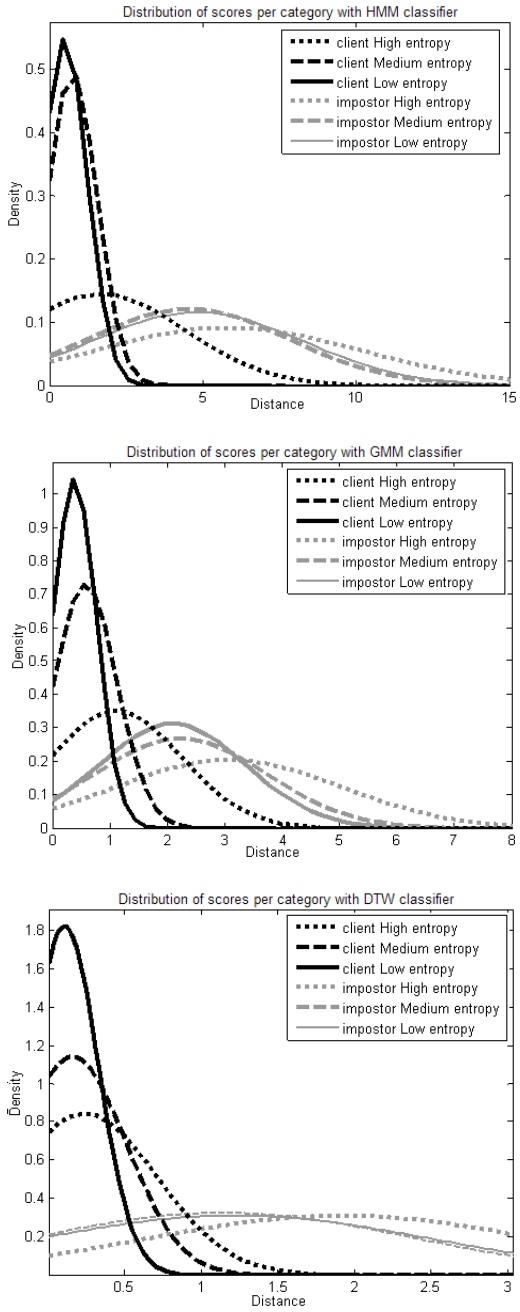

Figure 5: Client and impostor scores distributions considering random forgeries on categories of highest, medium and lowest client-entropy on DS3-430 database with HMM, GMM and DTW classifiers.

We notice in Figure 5 that for the statistical classifiers the distributions' shapes behave in the same way for client and impostor scores as well: they are the sharpest in the lowest entropy category (long, complex, stable and readable signatures) and become more flat when we change to a higher entropy category. As a consequence, we observe that in the highest entropy category, the overlap between client and impostor scores distributions is the most important, and such area is considerably higher than the area of overlap obtained in the lowest entropy category (around a factor 3).

For the DTW classifier, we note that this phenomenon of increasing sharpness of the distributions with decreasing client-entropy is only observed for client scores distributions. Indeed, for the three client-entropy categories, the impostor scores distributions are similarly flat. Nevertheless, as the DTW client distributions shapes are still different per category, the area of overlap of client and impostor scores distributions in the lowest entropy category is still lower than on the other categories. This result leads to different behaviors in terms of performance according to the category of client-entropy that we consider for all the three classifiers.

	High entropy		Medium entropy		Low entropy	
	EER	CI95%	EER	CI95%	EER	CI95%
HMM	13.93%	±0.78	8.5%	±0.74	4.79%	±0.85
GMM	19.49%	±1.23	13.39%	±1.04	7.49%	±0.87
DTW	5.76%	±0.59	2.95%	±0.49	2.03%	±0.19

Table1: Equal Error Rate and Confidence Interval on each client-entropy category on DS3-430 database with HMM, GMM and DTW classifiers.

As shown in Table 1, for the three classifiers, at the Equal Error Rate functioning point, performance is roughly improved by a factor three when switching from the highest entropy category to the lowest one. Confidence Intervals at 95% are given to show the significance of results. At other functioning points, this behavior is maintained for the statistical classifiers; nevertheless, for the DTW classifier, the difference between the high client-entropy category and the two others remains the same at other functioning points, but there is no longer a difference between the two categories of lowest client-entropy.

We notice that DTW outperforms the statistical approaches independently of the client-entropy categories. This can be explained by the fact that the description of signatures by the raw coordinates is not sufficient to allow a good adaptation of the statistical models. Moreover, the HMM topology and GMM complexity were not optimized but rather used in a "baseline" manner. Indeed, our goal in this paper was not to compare the classifiers but to show the relationship between the behavior of classifiers in terms of performance and the client-entropy categories.

In order to study the link between our client-entropy-based categories and the version of the Biometric Menagerie in [3], we plotted in Figure 7 for each classifier, the Average Impostor Score (averaged on all the available impostor accesses) vs. the Average Genuine Score. First, for statistical models, we notice that the highest client-entropy category shows low average client scores, and a pronounced scattering of persons along the average impostor score axis. This situation, in which individuals have variable impostor scores and low client scores, corresponds to "Phantoms" or "Goats" [3]: Phantoms when both types of scores are low at the same time, Goats when the user is easy to imitate.

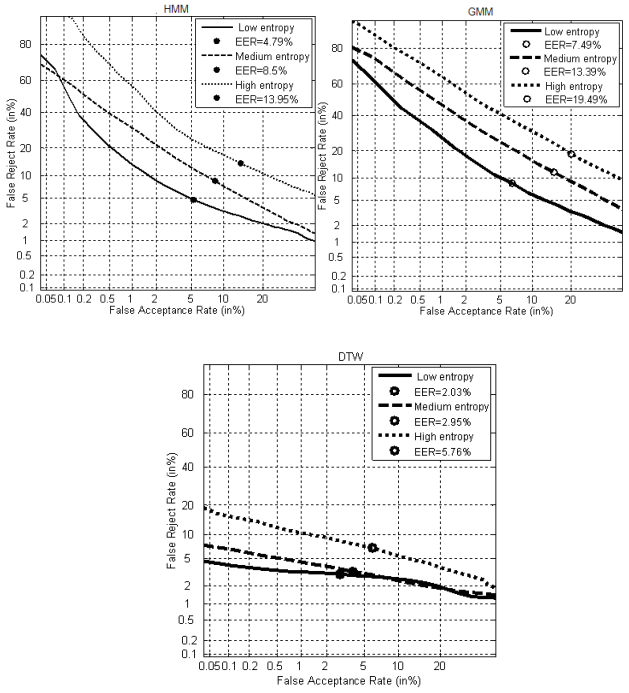

Figure 6: DET-Curves considering random forgeries on categories of highest, medium and lowest client-entropy on DS3-430 database with HMM, GMM and DTW classifiers.

Then, we notice that for signatures of lowest client-entropy, the average impostor score per user is bounded and the average client score per user is variable. Thus, such signatures resist well to impostures and span three classes of the Biometric Menagerie: Phantoms, when client scores are low, Sheep when client scores are intermediate, and Doves when client scores are high. Concerning the medium client-entropy category, it scatters in all the previously mentioned classes as it is a transition category in terms of client-entropy. Nevertheless, individuals in this category are more concentrated in the Sheep class when client scores are intermediate and impostor scores are low, and in the Phantoms class.

For the DTW classifier, the highest client-entropy category remains concentrated in the Phantoms class showing the same behavior of scattering along the vertical axis (impostor scores) and low average client score per user. With this classifier, at the opposite of statistical models, impostor scores are very variable for all the remaining client-entropy categories (medium and lowest entropy); users in such categories are concentrated in the Sheep class, thus explaining that system performance is improved with the DTW classifier with regard to statistical approaches.

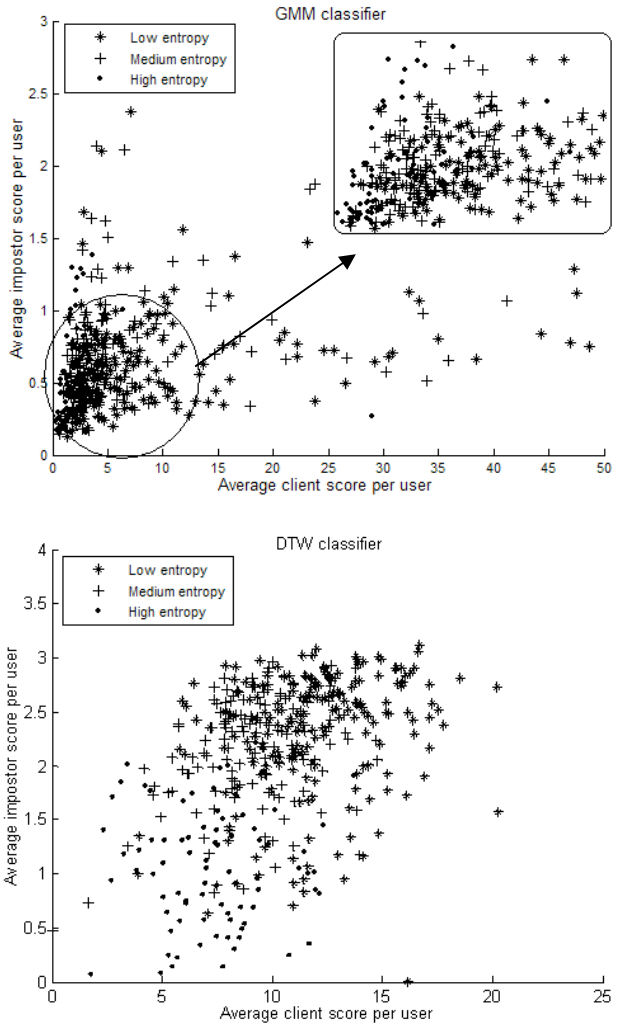

Figure 7: Average impostor score per user vs. average client score per user with the three classifiers (HMM, GMM and DTW) on DS3-430 database. Rectangular area on top right is the zoom of the circled area in bottom left.

IV. Conclusion

We have proposed in this work a novel measure of client-entropy, based on a Hidden Markov Model that performs local density estimation. Such measure allows obtaining three categories of signatures, coherent across several databases, spanning from highly variable, short and low information content signatures (high client-entropy) to stable, longer and complex signatures with the aspect of handwriting (low client-entropy). Then, we analyzed the relationship between this entropy measure and system performance by considering three classifiers: an HMM, a GMM and a DTW. We showed that for all classifiers, the three entropy categories have a different behavior in terms of performance: the lowest entropy category performs three times better than the highest entropy category. This result was explained in terms of the overlap of client and impostor scores distributions that vary across client-entropy categories. Finally, we related our entropy categories to the Biometric Menagerie [1,2,3,4,5]. In spite of the fact that our entropy-based categories do not correspond to identified classes of the Biometric Menagerie, we have observed that the lowest entropy category is more concentrated in the best configurations of the Menagerie (Doves and Sheep). Future work will focus on the impact of skilled forgeries on performance assessment and on the use of this categorization of databases for developing better and more robust identification strategies.

Acknowledgment

This work was partially funded by BioSecure Network of Excellence. We thank Javier Ortega-Garcia and his colleagues for putting at disposal the subset of the first 100 users of MCYT Signature Subcorpus.

References

[1] G. Doddington, W. Liggett, A. Martin, M. Przybocki, D. Reynolds, “Sheeps, Goats, Lambs and Wolves, A Statistical Analysis of Speaker Performance in the NIST 1998 Speaker Recognition Evaluation”, in Proceeding ICSLP’98, 1998.

[2] M. Wittman, P. Davis, P.J. Flynn, “Empirical Studies of the Existence of the Biometric Menagerie in the FRGC 2.0 Color Image Corpus”. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshop (CVPRW’06), 2006.

[3] N. Yager, T. Dunstone, “Worms, Chameleons, Phantoms and Doves: New Additions to the Biometric Menagerie”. In IEEE Workshop on Automatic Identification Advanced Technologies, 2007.

[4] R. Bolle, S. Pankanti, and N. Ratha. “Evaluation Techniques for Biometrics-based Authentication Systems (FRR)”. In Proceedings of ICPR, 2000.

[5] R. Bolle, J. Connell, S. Pankanti, N. Ratha, and A. Senior. “Guide to Biometrics”. Springer-Verlag, 2003.

[6] J. Richiardi, “Probabilistic models for multi-classifier biometric authentication using quality measures”, PhD Thesis, EPFL, 2007.

[7] Thomas M. Cover, Joy A. Thomas, “Elements of Information Theory”, Second Edition, John Wiley & Sons, 2006. W.-K.

[8] L. Rabiner, B.H. Juang, "Fundamentals of Speech Recognition", Prentice Hall Signal Processing Series, 1993.

[9] Kim Jung-Tae, Bang Sung-Yang, “A Measure of Recognition Difficulty for a Character Image Database”, Proceedings of International Conference on Document Analysis and Recognition (ICDAR’97), pp. 996-1000, 1997.

[10] A. Kholmatov and B. Yanikoglu, “An Individuality Model for Online Signatures”, SPIE Defense & Security: Biometric Technology For Human Identification V, 16-20 March 2008, Orlando FL, USA. H. Poor, An Introduction to Signal Detection and Estimation. New York: Springer-Verlag, 1985, ch. 4.

[11] F. Alonso-Fernandez, M. C. Fairhurst, J. Fierrez and J. Ortega-Garcia, "Impact of Signature Legibility and Signature Type in Off-Line Signature Verification", in Biometrics Symposium, BSYM, IEEE, pp. 1-6, Baltimore, USA, September 2007.

[12] S. Garcia-Salicetti, C. Beumier, G. Chollet, B. Dorizzi, J. Leroux-Les Jardins, J. Lanter, Y. Ni, D. Petrovska-Delacretaz, "BIOMET: a Multimodal Person Authentication Database Including Face, Voice, Fingerprint, Hand and Signature Modalities", Proc. of 4th International Conference on Audio and Video-Based Biometric Person authentication, pp. 845-853,Guildford, UK, 2003.

[13] J. Ortega-Garcia, J. Fierrez-Aguilar, D. Simon, J. Gonzalez, M. Faundez-Zanuy, V. Espinosa, A. Satue, I. Hernaez, J.-J. Igarza, C. Vivaracho, D. Escudero and Q.-I. Moro, “MCYT Baseline Corpus: A Bimodal Biometric Database”, IEE Proceedings Vision, Image and Signal Processing, Special Issue on Biometrics on the Internet, Vol. 150, n. 6, pp. 395-401, December 2003.

[14] http://www.biosecure.info

[15] D. A. Reynolds and R. C. Rose. Robust text-independent speaker identification using Gaussian mixture speaker models. IEEE Transactions on Speech and Audio Processing, 3(1):72–83, Jan 1995.

[16] http://www.int-evry.fr/biometrics/BMEC2007