SIViP (2013) 7:493–505
DOI 10.1007/s11760-013-0451-4

ORIGINAL PAPER

Entropy-based template analysis in face biometric identification systems

Maria De Marsico · Michele Nappi · Daniel Riccio · Genoveffa Tortora

Received: 19 December 2011 / Revised: 7 June 2012 / Accepted: 10 October 2012 / Published online: 17 March 2013

Abstract The accuracy of a biometric matching algorithm relies on its ability to better separate score distributions for genuine and impostor subjects. However, capture conditions (e.g. illumination or acquisition devices) as well as factors related to the subject at hand (e.g. pose or occlusions) may even take a generally accurate algorithm to provide incorrect answers. Techniques for face classification are still too sensitive to image distortion, and this limit hinders their use in large-scale commercial applications, which are typically run in uncontrolled settings. This paper will join the notion of quality with the further interesting concept of representativeness of a biometric sample, taking into account the case of more samples per subject. Though being of excellent quality, the gallery samples belonging to a certain subject might be very (too much) similar among them, so that even a moderately different sample of the same subject in input will cause an error. This seems to indicate that quality measures alone are not able to guarantee good performances. In practice, a subject gallery should include a sufficient amount of possible variations, in order to allow correct recognition in different situations. We call this gallery feature *representativeness*. A significant feature to consider together with quality is the sufficient representativeness of (each) subject's gallery. A strategy to address this problem is to investigate the role of the entropy, which is computed over a set of samples of a same subject. The paper will present a number of applications of such a measure in handling the galleries of the different users who are registered in a system. The resulting criteria might also guide template updating, to assure gallery representativeness over time.

Keywords Face recognition · Entropy · Pose and illumination distortions · Image Quality Index

M. De Marsico
Sapienza University of Rome, via Salaria 113, 00198 Rome, Italy
e-mail: demarsico@di.uniroma1.it

M. Nappi (✉) · D. Riccio · G. Tortora
University of Salerno, via Ponte don Melillo, 84084 Fisciano, Italy
e-mail: mnappi@unisa.it

D. Riccio
e-mail: driccio@unisa.it

G. Tortora
e-mail: tortora@unisa.it

1 Introduction

Much research by industry, academic world, and government agencies presently aims at investigating quality measures of images used for biometric recognition, as well as other features which may affect the accuracy of obtained results. Specific capture conditions and/or factors related to single subjects often influence the correctness of returned responses. For example, the input samples from which the biometric features are extracted can be affected by a number of distortions (e.g. poor illumination on a face, cuts on a fingerprint, and reflections in an iris). On the one hand, the accuracy of a biometric matching algorithm relies on its ability to better separate score distributions from genuine and impostor subjects. On the other hand, the above-mentioned factors may take an algorithm with "good" performances in optimal conditions, to provide incorrect answers. As a matter of fact, experiments performed by Becker and Ortiz using typical images from Facebook show that many of the most well-known techniques for face classification are still too sensitive to image distortion. This hinders their use in large-scale commercial applications, which are typically run in uncontrolled settings. It is often deemed that quality of input images is the main hindering factor in these cases. Quality measures

generally focus on characteristics of single images, such as the amount of distortion. However, it is often the case that a recognition system can fail even with good quality samples. To investigate this problem, we entail a kind of assessment that uses the entire set (gallery) of images pertaining to a certain subject, to assess if and how it supports a correct recognition in different situations. This property can be defined as representativeness. In practice, the gallery of a biometric system should contain more templates for each subject, acquired in different situations and with different quality, in order to increase the probability to recognize the subject under different trait variations. The idea is that a noisy image, as well as an image captured under unfavorable conditions, can increase the rate of correct recognition when the probe samples are acquired in uncontrolled conditions. For example, pose, illumination, and expression variations (PIE) are relevant to face recognition. Of course, the reverse of the medal is the possible increase in false recognitions, so that a compromise must be found through appropriate operating thresholds. Moreover, while quality measures can be also used to pre-select probe images (when applicable), representativeness is generally computed only on gallery templates. An exception to this latter policy could be during a template updating procedure, when a new probe is temporarily added to the gallery to check it may be worth adding it permanently. In our approach, quality measures should be considered together with measures of the representativeness of a subject's gallery. In this work, we focus on the concept of entropy.

The outline of the paper is as follows. In Sect. 2, we describe the architecture of a recognition system, in particular of a face recognition one, with a brief sketch of how its parts are composed and how they work. We also define the task of template updating. Section 3 focuses on the definition of a quality measure for an input sample and discusses quality in relation to representativeness. The main content of the paper is presented in Sect. 4, which provides a brief introduction to the concept of entropy and contextualizes such theoretical tool within biometric face recognition at different levels. In the first place, entropy is discussed as a quality measure computed on a single template, while afterward its computation is extended to the evaluation and handling of the overall gallery of templates pertaining to a given subject. Section 5 describes the experimental framework, where the proposed approach was tested. In particular, it presents details related to images, algorithms, and adopted performance measures. It then reports the obtained results and discusses the nature and meaning of each performed experiment.

2 Face biometric identification

A biometric system presents a structure which is quite similar to that of general pattern recognition systems [1]. It is composed of a capture system, a pre-processing module, a feature extraction module, and a possible separated module for their classification. In general, a biometric system converts data derived from physical or behavioral features of a subject in a *template*, which is used afterward in all matching operations.

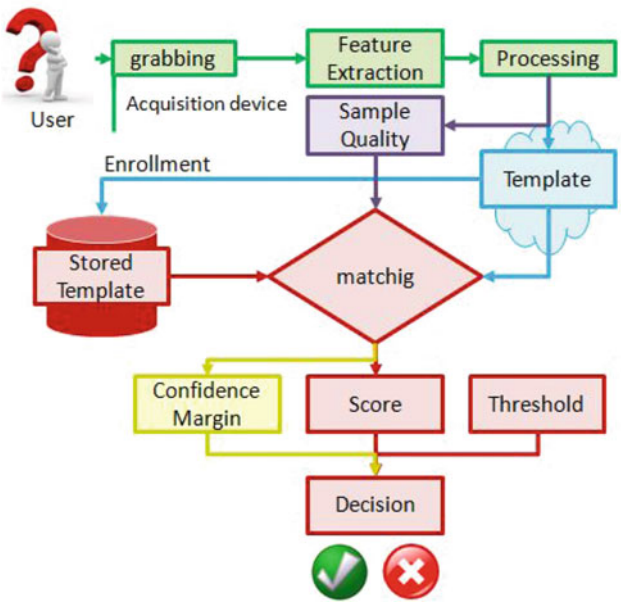

Fig. 1 Architecture of a biometric system

2.1 Architecture of a face recognition system

The general architecture of a biometric system and the articulation of its processing phases are shown in Fig. 1. The *acquisition* represents the operation through which the user provides to the system his/her static or behavioral biometric characteristics through a capture device. *Feature extraction* is the phase during which the characteristic features of a biometrics are located and coded, in order to generate a template. We define as *template* a set of data (vector, matrix, file), which is extracted from the input data, and presents a much lower size though maintaining a high discriminative power. Even if some kind of template is a frequent element of most biometric recognition algorithms, there are cases when matching is performed directly with original input data. Templates can be acquired either during user registration (enrollment) or during recognition. *Registration* is the process through which a new legitimate user is inserted in the system, in order to be later recognized after an access request. Registration, often referred as enrollment, is a basic phase for any recognition system. If it fails, it must be repeated with a better quality input. During the registration phase, more templates may be acquired for the same user. Some biometric systems return a score for the template acquired during enrollment,

which indicates a measure of quality of the received input, and allows discriminating cases when the process must be repeated. *Recognition* can be performed as *verification* or as *identification*. In the first case, the user claims an identity that must be verified, so that matching is 1:1 and only involves the probe template and the template (or templates) of the claimed identity. Identification entails a 1:N matching process, since the user must be searched in the whole database by matching its probe template with the template (or templates) of all registered users. In both cases, the basic operation is matching. Matching measures the similarity between two different templates. The result of this operation is a numeric value, which expresses the similarity degree of the compared templates. This value is also called score, with a meaning that is deeply different from that used for enrollment. It determines if a subject is recognized or not, depending on its comparison with a threshold: a subject is recognized if it achieves a similarity value higher than the threshold and is not recognized otherwise. On one hand, score as similarity degree is typical of biometric systems, since traditional systems such as passwords or badges produce a response that is strictly binary (correct/not correct, 0/1). On the other hand, its comparison with the threshold in the context of a verification system also produces a Boolean response YES/NO.

2.2 Template selection and subject gallery updating

Two issues related to biometric systems are attracting research attention, which are particularly related to the assessment of specific properties of biometric templates: (a) template selection, to create the gallery of a biometric module from scratch [22] and (b) template updating, to renew an existing gallery by substituting old as well as corrupted templates for an identity with more recent or representative ones [27]. Defining an evaluation criterion for the state of the gallery can represent an important added value in defining effective strategies for both these tasks, even if "the state of the art related to template update is still in its infancy" [22].

Biometric template handling, in the sense described above, applies in many settings, either commercial or related to security access control. As an example, a continuous control of the identity of previously authorized subjects may be required to access or stay in a restricted area. In such context, the system acquires a high number of input samples per subject. If these are suitably selected and added to the gallery built during enrollment phase, they can make recognition particularly robust to both short-term variations (illumination, pose, occlusions), and to long-term ones (age, look, permanent signs on face). In some settings, the presence of a high number of input samples is equally usual, as in face-based video indexing (movies, security tapes) [8,26]. In this case, variations between videos, or between scenes of a same video, can be very marked, so that the presence of a high number of input samples to integrate the initial gallery represents a viable strategy for correct recognition and indexing. On the other hand, many out of the new input samples are purely redundant, since they do not provide any additional information to that of the gallery. In this case, it is desirable to devise a criterion to select the minimum number of input samples able to maximize the gallery information content. This would represent a powerful tool to improve system performances in terms of both effectiveness and efficiency.

3 Face sample quality versus representativeness

3.1 Quality measurement

Quality of input samples can be exploited for both selecting gallery templates and possibly updating them and for discarding samples that cannot be automatically safely processed. In most biometrics research, the concept of quality of an input sample, which is frequently an image, is not defined according to the personal judgment of a human operator. It rather relates to specific factors that may affect the performances of an automatic matching algorithm, and to the extent, this may happen. For instance, a fingerprint may represent a very good sample for a human operator, who may be perfectly able to handle it. However, it may be a very bad input for a matching algorithm that exploits minutiae if it substantially lacks a sufficient number of such features. This may happen even if it presents well clear ridges in a particularly sharp image. This also highlights that the measure of the quality of an input sample is also often bound to the exploited matching algorithm. As an example, when the matching algorithm does not use minutiae, it has no sense measuring the quality of a fingerprint according to their number. Even for face images, the definition of quality mainly obeys to conditions that may either support or hinder the accuracy of an automatic recognition system. Therefore, we consider a full frontal, uniformly illuminated face as a "canonical" sample and take this condition as a reference to evaluate face quality. On the contrary, a tilted, shadowed, not symmetric face is considered of lower quality. On the other hand, a skilled end user (e.g. a forensic expert) may be perfectly able to recognize relevant subjects in photos in "difficult" settings: quarter, shadowed, etc. Given this circumstance, one could think that automatic systems are still too limited to be sufficiently reliable. However, when for example the number of photos to compare dramatically increases (as in certain identification tasks), an automatic system can still provide a valuable support. In those cases, the help of an automatic system would release the end user from performing the "easiest" activities (those for which the input image is deemed of sufficient quality to be automatically handled), only leaving "hard" cases to human analysis.

According to these considerations, we note that the definition of a quality measurement algorithm (QMA) is useful in different settings (enrollment, identification, and verification) as well as in different processing phases (pre-processing, matching, and decision) of a biometric system. It further provides the ability to measure the variation of sample quality with respect to time, or to different places (the environment) where the system operates, and to possibly perform corrective actions (e.g., readjusting a light source).

A QMA can be defined as a function which maps an input sample x into a scalar value that represents its quality $q = Q(x)$. The authors of [15] show that it is reasonable to assume that a QMA and a matcher which exploits it are related by an increasing monotone function, such that higher values of the first (in terms of sample quality) correspond to higher score values (in terms of similarity). Actually, in many cases, the function might not be strictly monotonic, due to some occasional fluctuation. However, in general, this function can be bounded by a monotonic one, with a sufficiently small margin of error. An example is given by the functions in Fig. 8 in the experimental section. Even this more relaxed assumption is being reconsidered in recent literature [3]. It is worth noticing that most QMAs only consider single images and not their relations with other images involved in recognition operations. Using a set of high-quality samples (according to a pre-defined measure) may therefore not completely eliminate the amount of error in a biometric system. We will return on this point later.

For the present work, we dealt with five of the available quality indices which consider some major hindering factors for face recognition, such as illumination, pose, or a combination: Sharpness Estimation (SE [18]), Universal Image Quality Index (UIQI [28]), Sample Pose (SP, [10]), Sample Illumination (SI, [10]), and Sample sYmmetry (SY, [11]). It was out of our scope to perform an exhaustive survey/analysis of all the available quality indices.

3.2 Gallery representativeness

One of the limits of the above measures and of similar ones is that they base their "quality" assessment on the information contained in a single biometric sample. However, a deeper consideration of the relation between image quality and system accuracy demonstrates that the former is less decisive in determining the latter than traditionally claimed. As a matter of fact, both past and recent researches underline also different contributions to the overall system performance. A first example refers to the consideration of the features that are intrinsic to the biometric trait of particular classes of individuals, in the way characterized by the Doddington's "biometric menagerie" [12]. Along a line which focuses a greater attention on the relations among subject gallery images, Philips et al. [21] classify enrolled users in "good, bad, and ugly". We aim at going further and at defining measures to assess representativeness of a subject gallery. This concept especially applies to the case when each registered subject is represented by more samples in the system, i.e., to the case of a multi-sample gallery. In such a system, the presence of more templates is usually exploited to increase the ability to recognize the subject under different conditions. However, knowing that each such sample is of high quality is not sufficient to guarantee high system robustness in terms of tolerance to distortions. Though being of excellent quality, the gallery samples belonging to a certain subject might be very (too much) similar among them. As a consequence, an "atypical" sample in input may be very different from those in the gallery and cause a wrong response lowering system accuracy. A viable candidate to assess true representativeness of a subject's gallery is provided by entropy resulting from samples of the same subject. Such measures would be exploited in many ways in handling the galleries of the different users who are registered in a system. First of all, we notice that entropy computed on the single image can be used as a measure of the amount of information in the biometric sample, to assure a sufficient informative content. In this way, it would be used in a way similar to the above-mentioned quality measures, with the same advantages and the same limits. On the other hand, mutual information, expressed in terms of entropy, would allow further assessing the amount of information overlapping among single samples, so that minimizing it would provide a criterion to increase gallery representativeness. This criterion would represent a heuristics to exploit as a guide in firstly populating the gallery of a recognition system. In a system which also implements dynamic and automatic gallery template updating procedures, the value added of a smart and well-structured starting population would be to significantly decrease the workload required by these during normal operation. In addition, this criterion might also become the guideline to follow during template updating, when appropriately adding/substituting samples for a subject would assure to maintain gallery representativeness also after long operating periods have elapsed.

4 Entropy-based template analysis

4.1 Background

In recent literature, we can find many examples of classical problems that are addressed through models inherited from Information Theory. Biometrics is not an exception, as works like [4,25] demonstrate. In this perspective, modules performing capture, pre-processing, and feature extraction can be modeled as a signal noisy source. In a symmetrical way, a matcher can be considered as a decoder on a noisy channel. From this point of view, one of the main problems,

as well as one of the most common concerning biometric systems, is to measure the amount of information contained in a biometric template which can be used for recognition. In other words, it is extremely interesting to estimate how much a biometrics is able to guarantee the univocal recognition of a subject. This property takes to consider the concept of probability of random correspondence (PRC) that is the probability that an impostor template contains an amount that is sufficient for recognition, of the biometric information that characterizes an enrolled user. In [5], the authors claim that such measure is a more reliable indicator of the accuracy of a system than standard performance measures such as Equal Error Rate (EER) and Receiver Operating Curve (ROC), because, differently from them, it does not depend on the size of the dataset against which the system is evaluated. Schmid et al. [25] analyze the recognition capacity of a system by modeling database templates as instances of a random linear process and applying the Chernoff capacity in this context. In [4], the authors model the genuine/impostor hypotheses as a signal to which a suitable noise pattern is summed. The application of the channel coding theorem to this model allows them to define the constrained capacity as a limit for the performances of the biometric system at hand.

The main critical aspect in defining useful tools for such kind of estimates is that it depends not only on single template features, but also on the used matching algorithm. Adler et al. [2] deeply studied the problem and proposed *relative entropy* as a solution. It represents a measure of the uncertainty that characterizes a random variable. In particular, it is usually adopted to measure the distance between two different distributions, is used in this case to measure the amount of information that distinguishes a subject from a given population. Sabah et al. [17] demonstrate a strict correlation between the relative entropy resulting from a given feature extraction technique (FET) and the performances in terms of accuracy of a recognition technique relying on such features. In addition, relative entropy is influenced by the quality of samples exploited, since the former increases as the latter does.

Information theory provides valuable support also in evaluating security of a biometric template [16]. In [14], Shannon's entropy is used to demonstrate the optimality of the generated code, in terms of template protection, in a system for the generation of multiple and revocable biometric keys, where matching is based on Euclidean metric.

More recently, concepts from information theory have been related to the problem of information fusion in multimodal systems [5,6], in particular in systems that are defined as Multiple Input Multiple Output (MIMO) [13]. In many application fields, a number of encoding algorithms employ multiple copies of a given signal to provide redundancy to address channel errors; a simple example is the classical repetition code. Likewise, employing signal diversity in the form of multiple signals transmitted from different antennas substantially improves the error performance in a MIMO (multi-transmit-multi-receive) information model. However, it is still difficult to devise how to precisely bridge a MIMO information model with a multimodal biometric system [24]. In this case, different subsystems should be considered as copies of the same one, to generate redundancy and correct errors during decoding (template matching).

4.2 The basics of the proposed analysis

Shannon entropy measures the uncertainty of a random variable [9], and in the specific case of a biometric system has the potential advantage to quantify the difference of a single subject from a whole population on the base of features extracted by a FET. This latter aspect strongly depends from the way it is contextualized within the biometric recognition process. The easiest way to insert entropy into sample quality evaluation of biometric systems is to use it as an estimate of the degree of randomness of image pixels. In this case, each pixel x in an image I is considered as a symbol in the alphabet emitted by a source S. In particular, in the case of a grayscale image, the alphabet is represented by the set of 8-bit integers giving 256 possible different shades of gray from black to white $(0, \ldots, 255)$. The image histogram represents the frequency table of all symbols (graylevels), being computed as $h(x) = |\{(i, j) : I(i, j) = x, 0 \leq i \leq \text{height}(I), 0 \leq j \leq \text{width}(I)\}|$, where $|S|$ represents the cardinality of the set S. Once the values $h(x_k), k = 1, \ldots, 255$, in the histogram have been normalized in the range [0,1], and according to the total number of pixels in the image, each of them represents the probability of occurrence $p(x_k)$ of symbol x_k in $I, k = 1, \ldots, 255$. Entropy $H(I)$ can be therefore defined as follows:

$$H(I) = -\sum_{k=0}^{255} p(x_k) \log_2(p(x_k)), \tag{1}$$

This formulation of image entropy in (1) can be exploited as a generic quality measure, with all related advantages and limitations. It expresses the information content of the bin distribution of the color of the image. However, it does not take into account spatial correlation; moreover, an image could present irrelevant details that would increase the entropy. As a matter of fact, entropy cannot be proposed as sufficient by itself to solve the problem of quality. Each quality measure evaluates an image with respect to a particular issue, and it is unlikely to find in the literature measures that are able by themselves to perform a full and complete assessment of the quality of a biometric sample. They are rather regarded as individual components of a pool of measures and can contribute, each with a relative weight, to a final estimate of the quality. In this context, we want to evaluate in the paper which is the potential contribution from image entropy,

although aware that the noise, seen as detail, could undermine its performance. A similar case is represented by the Sharpness Estimation, commonly used as a measure of quality and that measures the goodness of a sample as a function of the detail in the image. In Sect. 5.1, we underline the relation between such two measures as an interesting result. As a further consideration, we have to take into account that, like the commonly used quality measures (e.g., sharpness estimation, sample pose, or illumination), entropy evaluates the single sample uniquely according to its individual information content. According to some researches, this is the reason why, though using a set of high-quality samples (according to a pre-defined measure), a biometric system reduces, but not eliminates its error rate (see for example [3]). Entropy-based template analysis assures real advantages, when it is not used to evaluate the randomness amount for a single sample, but rather to relate the discriminant power of the templates of a single subject with those of other subjects. This idea is the base of most techniques that exploit relative entropy to estimate the degree of uniqueness which is assured by a biometric trait when processed by a FET. In this paper, entropy is introduced as a tool to evaluate the contribution of each sample in guaranteeing a suitable diversification of the templates in a subject gallery. To this aim, we modify (1) such that we do not consider the color of a single pixel as the value of a random variable, but the relation (similarity) of a probe with the elements of subject's gallery. This concept is further extended to define a matching scheme where the similarity score is defined as a function of the entropy variation produced by introducing the probe in the gallery of each enrolled subject.

In order to clarify the notation, we detail the context at hand. We consider a recognition system T, which is characterized by a gallery G of templates, a feature extraction technique A, a similarity measure between templates d. Gallery G is the union of galleries G_k of the single users ($G = \cup_k G_k, G_k \cap G_h = \emptyset\ \forall k \neq h$). Each G_k contains all templates $g_{i,k} \in G_k$ pertaining to the subject k. The FET A is defined such that it takes a sample image I as input and produces a template v as output, which contains features extracted by I, i.e., $v = A(I)$. Similarity measure d associates a real scalar value to a pair of templates; if we compare a probe template v with a gallery template $g_{i,v}$, we get $s_{i,v} = d(v, g_{i,k})$. In particular, $s_{i,v}$ is a value in the real interval [0,1], otherwise, i.e., if the matching function returns values outside such interval, a score normalization technique could be applied to remap the distribution of values produced by d in that interval. In the case that d is a distance measure instead of a similarity, it is always possible to consider the value $1 - s_{i,v}$ in place of $s_{i,v}$. More in general, the following proposed definitions are independent from the used similarity (distance) measure d, provided that returns real scalar values for all templates in a gallery. Of course, on the other hand, the final system performance is affected by the accuracy of the FET and of the associated recognizer, as it would happen anyway.

It is appropriate to underline that, in this preliminary step, entropy is not introduced to assign a new template v to its correct gallery G_k. We are rather interested in measuring the representativeness of G_k and how v alters it. In this context, we assume that someone else (an oracle, such as a matching algorithm) has assigned the template v to its corresponding identity k, so that the score $s_{i,v}$ can be interpreted as the probability that template v conforms to $g_{i,k}$ that is

$$s_{i,v} = p\left(v \approx g_{i,k}\right), \tag{2}$$

In order for $s_{i,v}$ to represent such a probability, it must range in the interval [0,1], and the sum over all templates in G_k must be 1 since we do not question about v correct assignment to k; therefore, each $s_{i,v}$ is normalized with respect to $\Sigma_i(s_{i,v})$.

The concept of entropy defined in (1) can be applied also to the case of the probability distribution obtained from applying (2) to a whole gallery G_k with respect to a probe v as follows:

$$H(G_k, v) = -\frac{1}{\log_2(|G_k|)} \sum_{i=1}^{|G_k|} s_{i,v} \log_2\left(s_{i,v}\right), \tag{3}$$

where $1/\log_2(|G_k|)$ is a normalization factor. It corresponds to the maximum entropy, which is obtained when (2) has the same value for all the templates in the gallery G_k. Therefore, irrespective of the size of the gallery, we always obtain a value in the range [0,1] so that it is also possible to consistently compare the values obtained for different galleries.

We can now introduce a measure of entropy for the gallery G_k, which is computed by considering each gallery template $g_{j,k}$ in turn as a probe v. Given Q the set of pairs $q_{i,j} = (g_{i,k}, g_{j,k})$ of elements in G_k such that $s_{i,j} > 0$, the entropy for the gallery is defined as the following:

$$H(G_k) = -\frac{1}{\log(|Q|)} \sum_{q_{i,j} \in Q} s_{i,j} \log_2\left(s_{i,j}\right), \tag{4}$$

The proposed formulation allows to obtain values in the range [0,1] irrespective of the size of the gallery.

The value of $H(G_k)$ represents a measure of heterogeneity for G_k. The ways to use it within a biometric schema are manifold, e.g., reliability estimation and template selection. In Sect. 4.4, the use of entropy is further extended, since it is just used to identify a new template v, by exploiting (2). In practice, we implement a matching algorithm that tries to assign v to each gallery G_k and selects the most appropriate value for k by analyzing the obtained $H(G_k)$ values. To the best of our knowledge, the overall entropy-based approach proposed in this paper is novel in the literature.

4.3 Gallery entropy for template selection

In many applications related to biometric systems, a very large number of input samples is continuously collected and added to the gallery. Handling such samples may weigh down system operation, till to compromise its usefulness. An example is the handling of digital books of photos which are cataloged with respect to faces. Another example is face processing from videos, where tens of input samples are captured for each subject, according to the length of the clip at hand. Of course, not all such samples are really representative, due to a high redundancy [23]. *Sample selection* criteria are adopted, so that some samples are discarded a priori, in order to lower the computational weight for the system.

To this aim, $H(G_k)$ can represent a valid tool to select a subset of representative samples out of a much larger set, given that a strategy is defined for the progressive selection. The proposed procedure takes a gallery G_k as input, and starting from it computes a similarity matrix M_k and the value for $H(G_k)$. M_k is computed by applying the similarity measure d to all pairs of templates in G_k, i.e., $M_k(i, j) = d(g_{i,k}, g_{j,k}), \forall g_{i,k}$ and $g_{j,k} \in G_k$. For each $g_{i,k} \in G_k$, the matrix M_k is used to compute the value of $H(G_k \setminus \{g_{i,k}\})$ that would be obtained by considering $g_{i,k}$ as a new sample v, not already contained in G_k. The sample $g_{i,k}$ achieving the minimum difference $f(G_k, g_{i,k}) = H(G_k) - H(G_k \setminus \{g_{i,k}\})$ is selected; the matrix M_k is updated by deleting the i-th row and column, and the process is repeated, until all elements of G_k have been selected. In practice, we first select the most representative samples, i.e., those causing the lower entropy (representativeness) decrease. In this way, the minimum entropy difference tends to increase, as expected. However, from a certain point, it tends to decrease again due to the much lower number of samples which are involved in the computation. We empirically identified the parabola as the simplest curve to approximate this behavior with sufficient accuracy. The pseudo-code of the ParseGallery algorithm and of its associated functions is shown in Fig. 2.

This approach can be considered as a mechanism to progressively reduce the inhomogeneity of a set of samples. The entropy-based function represents a possible choice in this sense, together with other criteria such as the standard deviation. In Sect. 5.2, we compared the performance of the latter with the entropy-based function $f(G_k, g_{i,k})$ defined herein. However, we experimentally observed that the use of standard deviation causes apparently good samples to be interleaved with samples with much worse capture conditions. In other words, we lose the characteristic condition of sample ordering for ascending perceived quality of acquisition.

Figure 3 shows an example, where samples $g_{i,k} \in G_k$ have been ordered according to their selection during the ParseGallery procedure. We did not carry out a systematic study to quantitatively measure the correlation between the

```
PIVOTS ← ParseGallery(Gk)

  p ← []
  h ← 0
  PIVOTS ← []
  Mk ← ComputeDistanceMatrix(Gk)

  while(p)
    p ← FindPivot(Mk)
    Mk (p,:) ← 0
    Mk (:,p) ← 0
    PIVOTS[h] ← p
    h ← h + 1
  end
```

```
Mk ← ComputeDistanceMatrix(Gk)

  for i=1 → |Gk|
     for j=1 → |Gk|
       Mk (i, j) = si,j
     end
end
```

```
p ← FindPivot(Mk)

    if ∃ (i,j) ∍ Mk (i,j)≠0
       Et ← entropy(Mk)

       for h=1 → | Mk |
         Mtmp ← Mk
         Mtmp(h,:) ← 0
         Mtmp(:,h) ← 0
         E(h) ← Et - entropy(Mtmp)
       end

       p ← argmin(E)
    else
       p ← []
    end
```

```
E ← ComputeEntropy(Mk)

    Q ←{(i,j) ∍ Mk(i,j)≠0}
    E ← 0

    ∀ qi,j=(i,j) ∈ Q
      E ← E - Mk(i,j)log2(Mk(i,j))/log2(|Q|)
```

Fig. 2 Pseudo-code for ParseGallery

results produced by the algorithm and the similar product from a human operator. However, an experiment was conducted in which the results produced by the algorithm was assessed by a human operator. For each experiment run, we acquired a sequence of 16 samples, in which the subject was free to move and produce all kinds of variations in the face (expression, pose, and occlusion). The algorithm was run on the images captured, and the result was a reordering from the most distorted ones (e.g., pose far from the frontal one, mouth covered, …) to the least affected by changes (e.g., frontal pose, neutral expression). This result was submitted to the human operator, in order to estimate the adherence with his expectations. In a sample of 50 subjects, the overall

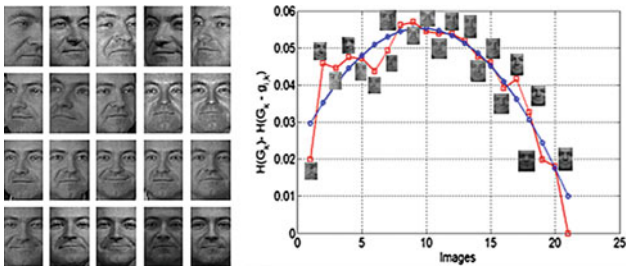

Fig. 3 Example of using $H(G_k)$ to evaluate the heterogeneity of the gallery G_k. On the *left*, samples $g_{i,k}$ have been ordered according to their selection during ParseGallery procedure. On the *right*, the graph of computed values of $f(G_k, g_{i,k}) = H(G_k) - H(G_k \setminus \{g_{i,k}\})$ is shown, together with an approximating polynomial function

response was 98 % match with the expected result over a set of 50 experiment runs.

It is worth making some comment. Experiments using standard deviation instead of entropy, confirmed that the specific shape of the function $f(G_k, g_{i,k})$ is typical only of entropy. The steps-like trend of the function f is due to the inherent nature of the computed measure that, in some way, represents how much a template is different from a given set of templates. More in detail, given a set of templates, the function f selects the one that minimizes the difference between the entropy of the original set of templates and that of the same set, having excluded the selected item. Clearly, if a template very similar to the one just selected is still in the set, it will be, with high probability, the next to be selected. In this way, the function f undergoes a small variation in the form (generally remaining quite constant or slightly decreasing). As soon as similar samples end up, the algorithm selects a sufficiently different one, which causes a peak in the function f. Actually, being different from the others, deleting it from the set generates a significant change in the entropy of the set. Afterward, the function f assumes again its constant/descending trend, up to the selection of a substantially different sample. The parabolic course is instead typical of the various functions and was also expressed by the standard deviation. This is dictated by the fact that the differential calculated by the function f undergoes an initial greater variation, due to the greater amount of template in the selection process, before stabilizing and then reversing its trend as the number of templates remained in the set becomes smaller.

After performing sample ordering according to the ParseGallery procedure, it is necessary to define a criterion allowing selecting those that guarantee a suitable representativeness. As it can be noticed from Fig. 3(left), samples with higher distortion achieve the first positions, where the graph (right) takes an increasing trend; the latter decreases with those samples which present a "canonical" aspect in terms of pose and illumination. The graph can be approximated by a second-order polynomial function, assuming a parabolic shape.

Figure 3 (right) shows that the local maxima of function $f(G_k, g_{i,k})$ correspond to significant variations of the features of the corresponding face with respect to the preceding one. A valid criterion for sample selection is therefore to choose only samples that correspond to a local maximum in the curve of $f(G_k, g_{i,k})$. Such criterion has been adopted during the experiment phase, for both entropy-based and standard deviation-based homogeneity.

It is to notice that variations affecting gallery entropy also implicitly include quality variations, e.g., the entropy computed over image pixels.

4.4 Gallery entropy for people identification

In Sect. 4.3, we observed that samples with a high distortion are located in the left ascending part of the graph, while those acquired in "canonical" conditions tend to lay on the right descending part. This observation has to be extended also to those samples that, though presenting suitable capture conditions, pertain to a different subject. In other terms, suppose that a sample v is temporarily introduced in the gallery G_k even without belonging to the subject k. Even in this case, it will be generally located in the left part of the graph of the ordering produced by ParseGallery. According to this, it is possible to define a new matching schema to recognize a new sample v submitted as query (probe) to the system.

Recognition obeys the following protocol. For each registered subject k, the algorithm builds the new gallery $G_{k,v} = G_k \cup \{v\}$ and applies the procedure described in Sect. 4.3 to sort samples $g_{i,k}$ and compute the parabola which approximates the function $f(G_{k,v}, g_{i,k}) = H(G_{k,v}) - H(G_{k,v} \setminus \{g_{i,k}\})$; an example is shown in Fig. 4. The number of samples $|G_{k,v}|$ can be considered as the difference between the positions of the first and of the last sample on the graph. Let us set h equal to the distance between the vertex $U = (u_x, u_y)$ of the parabola and the line r parallel to the x-axis and passing through the minimum value a of function $f(G_{k,v}, g_{i,k})$. The value R represents the distance between the sample v and the axis of the parabola $(v_x – u_x)$, while the value E corresponds to the distance between sample v and line $r(v_y - a)$. The similarity function for sample v with respect to gallery G_k is expressed by the following formula:

$$s_{v,k} = \frac{1}{2}\left[\frac{(v_x - u_x)}{\left(|G_{v,k}| - u_x\right)} + \frac{\left(v_y - a\right)}{h}\right]. \tag{5}$$

where the two terms represent the relative distance from the most "typical" templates and the relative representativeness of sample v. Notice that the function in (5) is computed after any similarity measure d, according to the conditions mentioned in Sect. 4.2, i.e., provided that it returns real scalar

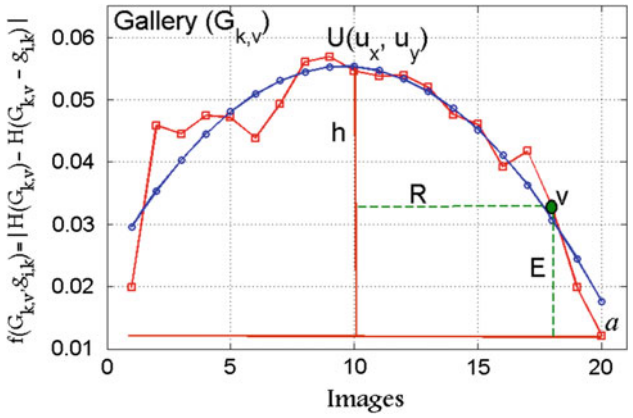

Fig. 4 Example of a new sample located in the graph of the function defined in Sect. 4.3 and of how this can be used to define a similarity measure

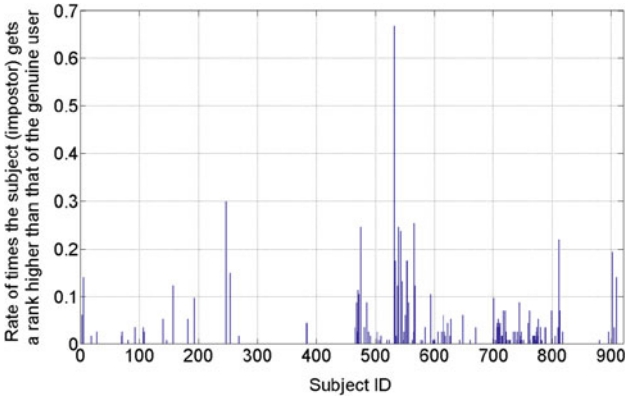

Fig. 5 Histogram of the rate of occurrences when an (impostor) user is returned before the genuine user in the ordered list of ranks returned by the identification system

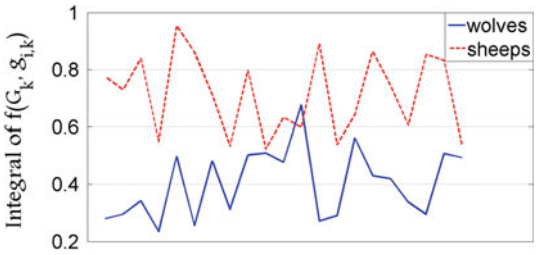

Fig. 6 Value of the integral of function $f(G_k, g_{i,k})$ for the two different categories of users known as wolves and sheep

values for all templates in a gallery. In our implementation, the distance d used is that defined for the comparison between the templates that are generated by FACE and which is described in [10]. However, other viable examples are given by all the most commonly used metrics, such as the city block distance, Euclidean, or cosine.

Compared to the distance functions in the literature, the one defined in (5) has the advantage of relating the probe template with the trends of similarity which exist among the elements in the gallery, given by the kind of parabola in Fig. 4, instead of just calculating a global distance, such as a sum of distances from the gallery samples or from the gallery centroid. Existing measures calculate a distance matrix and then apply a fusion rule. With the rule in (5), the two processes occur simultaneously and implicitly. A further advantage is the ability to detect wolves (people who could replace one or more persons registered in the system, see [12]). Other measures require a specific analysis for this task. From tests performed on a real database, extracted from FERET [20] (see Sect. 5), we observed that, though genuine and impostor distributions are sufficiently different to guarantee acceptable values of EER (or equivalently of ROC), the graph of Cumulative Match Curve (CMC) is not satisfying, because it starts quite low. However, it reaches very quickly a satisfying height for a certain small value of the rank. This behavior lets us suppose the presence of some few subjects which tend to be often accepted (returned at the first positions in the list) though being impostors. In other words, impostor subjects identified for a genuine one (which lower CMC curve) are always the same.

This hypothesis is supported by an experimental verification on the same dataset. As a matter of fact, Fig. 5 reports an histogram of the percentage of times that each subject has been returned (as an impostor) before the genuine one in the ordered list of ranks returned by the system as response for an identification operation.

From Fig. 5, we can see that a number of subjects tend to be returned by the system before the subject which is actually searched (genuine) in more than 20 % cases, with peeks of 70 %. These are called "wolves" by Doddington [12]. A deeper analysis of gallery composition and of the corresponding function $f(G_k, g_{i,k})$ for these subjects has shown the existence of common traits and has demonstrated that $f(G_k, g_{i,k})$ is a valid tool to preventively identify such subjects and signal the most potential wolves (see [12] for more details on the topic). In particular, we experimentally observed that wolves present a much lower value for the area below the function $f(G_k, g_{i,k})$. Figure 6 reports an example of the value of the integral of the function for wolves and sheep users.

Notice that identification performed using the above technique can represent the oracle mentioned in Sect. 4.2 and trigger a new template selection procedure as described in Sect. 4.3.

5 Experimental setup

A variety of aspects, which are bound to the use of entropy in face sample analysis, have been introduced in the preceding section. Each of them needs an experimental assessment.

For this reason, we chose a face database which is sufficiently versatile and useful to this aim, namely FERET [20]. Images were acquired in 15 sessions, in analogous environmental settings, except for small variations in capture devices. For each subject, 2 frontal images are present at least, which are organized in subsets *fa* and *fb* (expression variations), while smaller subsets include pose variations which are more (*ql, qr, hr, hl*) or less (*fc*) marked. There are 24 datasets, with a total of 14,126 images captured from 1,199 subjects. Face pre-processing, after location, was performed according to the approach in FERET protocol [20]. The center of eyes is used to center the face and crop the image, which is then resized to guarantee a fixed value for inter-ocular distance.

The matching algorithm used to extract features and compute scores is based on a localized version of correlation index. Details are provided in [10], which describes Face Authentication for Commercial Entities (FACE). It is to underline that, out of the whole FACE approach, we only applied the matching function and not the pre-processing phase (pose and illumination correction to achieve a "canonical" setting), since we were interested in working with a gallery containing as much differentiated samples as possible. This choice partly penalizes FACE, but on the other hand, it highlights the robustness on entropy-based schemes. For the same reason, the distance measures (scores) exploited for the entropy-based approach are the same exploited by FACE.

Different performance measures were used to assess the accuracy of the system. In the first place, EER, i.e., the common value taken at the intersection by the curves representing false acceptance (FA) and false rejection (FR) rates versus variations of the acceptance threshold (the lower EER, the better). A higher amount of information is provided by the ROC, which relates False Acceptance Rate (FAR) with Genuine Acceptance Rate (GAR). This measure is typically used for verification systems rather than for identification ones. Cumulative Match Score (CMS) at rank n is specifically defined for identification systems instead. It represents the probability that the correct identity is among the first n retrieved ones. Setting rank n on x-axis and the corresponding values on the y-axis, we obtain a graph named CMC. CMS at rank 1 is also defined as Recognition Rate (RR). Bolle et al. [7] demonstrated that, when scores are derived from one-to-one comparisons, as in our case, there is a very tight relation between EER and CMS. Therefore, FAR, FRR, EER, and derived measures can be used for identification systems too, given the described condition.

When quality measures are introduced in an identification schema, some samples are discarded by the system, since they are deemed not reliable enough. These responses are neither correct not wrong, and therefore, they are not considered in the classical performance evaluation. On the contrary, a further measure of system robustness is the Rate of Adequate Images (RAI), i.e., the percentage of images that were deemed useful by the system.

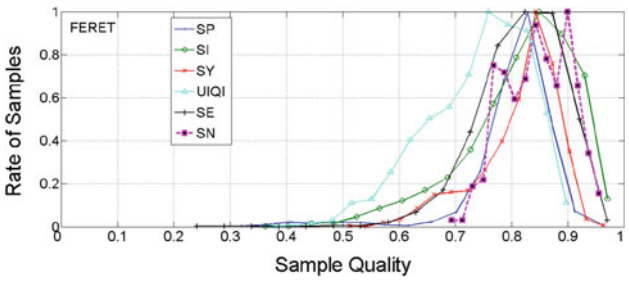

Fig. 7 Behavior of the distributions of the values from the different quality measures for the *fa* sample subset of FERET

5.1 Entropy as an image quality measure

In the first experiment, we tested the behavior of entropy when it is computed over image pixels and used as a quality measure for the input samples. Its performances are compared with those obtained from five other quality measures, following a protocol similar to [11]. In practice, the first 250 images from subset *fa* were selected, which correspond to 116 subjects. Through an extended version of the approach based on Active Shape Model described in [19], face was located in images, and the six quality measures were computed over it, namely Sample Pose index (SP) and Sample Illumination index (SI) [10], Sample sYmmetry index (SY) [11], Universal Image Quality Index (UIQI) [28], Sharpness Estimation (SE) [18], Sample eNtropy (SN).

Figure 7 reports the distributions of quality values for the different quality measures over the FERET *fa* subset. Figure 7 shows that SP values are particularly concentrated around the mean value 0.8 (scarce pose variability), while SI spans a wider interval, which underlines the presence of a variable illumination. SY presents an average behavior, which accounts for both kinds of variation. Differently from the other measures, SN plot is more irregular, though the general trend is similar to the other plots. This underlines a certain discriminating power with respect to the conditions of the dataset over which it is computed.

In the second experiment, for each measure, we associated the corresponding quality value to each face sample. If the quality value is below a predetermined measure-specific threshold th, the sample is discarded from the set used to compute the performances of the recognition system. On the set of accepted face samples, for each exploited quality measure and identified threshold, EER was computed by an all-against-all comparison: each probe was used claiming in turn either the correct or one out of all the other identities. Threshold th (for each measure) was varied in the range [0,1]. Figure 8 shows that, since SY accounts for both pose

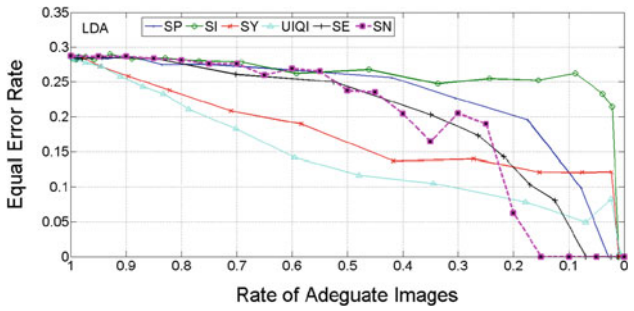

Fig. 8 Graphs of the variations of system performance in terms of EER (y-axis) when only samples achieving a quality value above the corresponding threshold are considered; the Rate of Adequate Images is on x-axis

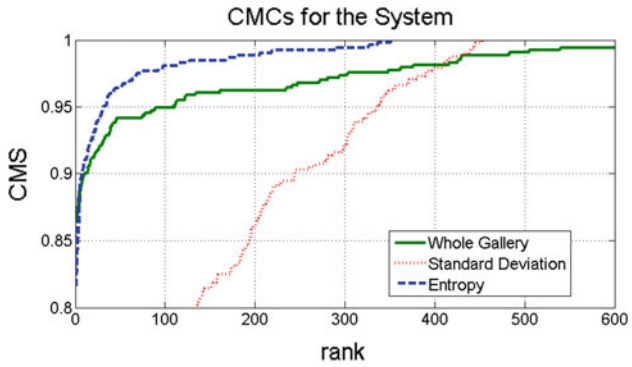

Fig. 9 Variation of CMC curve after gallery pruning using function $f(G_k, g_{i,k})$

and illumination variations, it can be considered as the most indicative among the first three ones (SP, SI, SY). Since pose and illumination variations are limited in this dataset, SP and SI achieve a less marked improvement (see graphs) when considered separately. It is interesting to notice the similarity between the trends of SE and SN. Since one measures the sharpness of an image and the other one measures the entropy, this seems to underline their relationship.

5.2 Gallery entropy for template updating and matching

Some of the concepts discussed in Sect. 4 are based on the assumption that each subject who is enrolled in the system is represented by a sufficient number of samples (e.g., about 5). In order to perform an appropriate experimentation with significant results, we extracted a different subset from FERET database, with 177 subjects for whom the location procedure in [19] provided eight correctly segmented images at least.

In the third experiment, given the gallery G_k for a subject k, we tested the ability of the function $f(G_k, g_{i,k})$ of selecting an appropriate subset $\hat{G}_k \subseteq G_k$ of such gallery, which is sufficiently adequate to represent the subject k. In practice, we have a recognition system A with K registered subjects, and a gallery G where each subject has a number of sample images $|G_k| \geq 1$ ($G = \cup_k G_k, G_k \cap G_h = \emptyset \; \forall k \neq h$). We measure the performances of such system, in terms of CMC, when a certain set of probe images P, which belong to the same K subjects but are not included in the gallery, are submitted as queries. Afterward, the function $f(G_k, g_{i,k})$ is used as described in Sect. 4.3 to prune gallery G_k and substitutes it with $\hat{G}_k$. Finally, performances of the system with the same probe set P are measured again in terms of CMC. If the performance variation is negligible, this means that the discarded samples were redundant. In this experiment, the probe set included 531 images, 3 for each of the 177 subjects. In the first case (whole gallery), no pruning was performed and the gallery includes 2,062 images. The pruning operation on the gallery was performed by adopting both standard deviation as homogeneity criterion, which selected a subset of 456 samples, and entropy, which selected 451 samples. The score for comparing two samples was computed with the method in [10], as previously mentioned.

Figure 9 shows that the most significant variation pertains to the first ranks. In particular, the most significant difference regards rank 1 (RR), that is 0.86, 0.59, and 0.82, respectively. However, even in this case, we achieve a probability of correct recognition of above 0.82, which is a very good result in any case, if compared with the lowered computational workload which might be very appealing in intensive, low-security applications. We can further observe that the curve obtained by using entropy grows better than the one obtained on the whole gallery. This can be explained by the removal of many samples which were redundant or poorly representative. Such removal reduces the probability that the system returns the wrong gallery image during a probe identification. In other words, removed templates do not constitute "useful" information. As a matter of fact, it may seem that less information producing best performance is a contradiction. However, even some of these templates, which, in an absolute sense, contain information, in terms relating to the recognition rather constitute noise. Because of the excessive distortion they suffer for, they obtain the contrary effect of increasing the interclass similarity and reduce intraclass similarity. That is why the process of selection after ordering must be suitably performed, so to exclude both "too similar" and "too different" samples for a subject. As anticipated in Sect. 4.3, our sample selection is a winning strategy achieving a twofold goal: better performance with less computational power.

The standard deviation provides poor performance, since it reduces the number of samples by considering only how much the scores computed between the probe and the gallery images diverge from the average value, without considering the actual representativeness of the selected samples.

Figure 9 highlights how entropy allows to select a lower number of samples, with an higher representativeness.

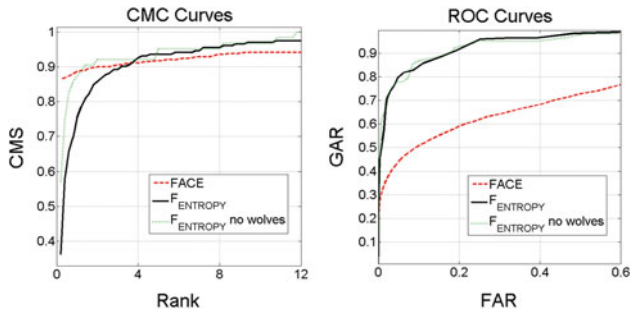

Fig. 10 Performance comparison of FACE and F_{ENTROPY}. CMC curves are on the *left*, ROC ones on the *right*

In the fourth experiment, function $s_{v,k}$ (5) defined in Sect. 4.4 was considered as a similarity score in a recognition system. Performances of such system, labeled F_{ENTROPY} in the graphs in Fig. 10, were measured in terms of CMC and ROC and compared with those achieved by FACE.

Figure 10 shows that FACE produces a higher CMC curve and therefore provides better identification results, while F_{ENTROPY} provides far better performances in terms of ROC, therefore resulting more appropriate for verification. In this specific case, verification entails the comparison of the probe samples with all samples of the claimed identity, using F_{ENTROPY} approach. CMCs in Fig. 10 further show that, though F_{ENTROPY} starts with a lower CMS, it quickly arrives to similar performances as FACE.

Finally, Fig. 10 also shows the curves corresponding to an experiment "no wolves". These plot performances of F_{ENTROPY}, when subjects considered as "*wolves*", are discarded from the gallery G according to the arguments in Sect. 4.3. In this experiment, identification performances are improved, so demonstrating the usefulness of a well-structured gallery.

6 Conclusions

In this work, we discuss how entropy can be contextualized within a face recognition system. Many works in the literature exploit entropy to evaluate the uniqueness of a biometric trait or the appropriateness of the code that a system assigns to submitted samples. The presented work considers entropy under two very different aspects. In the first place, we directly apply entropy to image pixels, according to the most common definition, to compare its potential as a quality measure to that of other measures used in the context of face recognition. On the other hand, we introduce a different use of entropy as measure of representativeness of the gallery (intended as the set of face images) pertaining to a subject who is registered in a biometric system. In this sense, it is used for template selection, i.e., to prune the gallery by reducing the computational cost for the system, without excessively penalizing its accuracy. By extending these concepts, entropy has been reconsidered as a measure of the similarity between subjects and therefore as a real classificatory. Experiments related to such latter aspect gave extremely encouraging results and suggested interesting possibilities for further investigations related to improving efficiency for the discussed procedures, besides their combination with existing techniques. In third experiment in Sect. 5.2, we used FACE similarity measure as a basis to compute $s_{i,v}$. We performed some preliminary experiment with other techniques (e.g., LDA) but the obtained differences were not significant enough to be included in the paper. However, a more accurate set of comparisons will be a topic for future work. In the same way, mutual information is a further topic of our ongoing study. Results and comparisons will be the subject of a future work, in which the current results obtained with the entropy will be a basis for comparison. Finally, one of the interesting aspects to be analyzed in future work is certainly the comparison with further popular homogeneity measures, as f-divergences, from which KL distance, or relative entropy.

References

1. Abate, A.F., Nappi, D., Riccio, M., Sabatino, G.: 2D and 3D face recognition: a survey. Pattern Recognit. Lett. **28**(14), 1885–1906 (2007)
2. Adler, A., Youmaran, R., Loyka, S.: Towards a measure of biometric information. In: Proceedings of the Canadian Conference on Electrical and, Computer Engineering, pp. 210–213 (2006)
3. Beveridge, J.R., Phillips, P.J., Givens, G.H., Draper, B.A., Teli, M.N., Bolme, D.S.: When high-quality face images match poorly. In: IEEE International Conference on Automatic Face and Gesture Recognition and Workshops (FG 2011), pp. 572–578, 21–25 March 2011
4. Bhatnagar, J., Kumar, A.: On estimating some performance indices for biometric identification. Pattern Recognit. **42**(5), 1805–1818 (2009)
5. Bhatnagar, J., Lall, B., Patney, R.K.: Performance issues in biometric authentication based on information theoretic concepts: a review. IETE Tech. Rev. **27**, 273–285 (2010)
6. Bhatnagar, J., Kumar, A., Saggar, N.: A novel approach to improve biometric recognition using rank level fusion. In: Proceedings of the IEEE Conference on Computer Vision and, Pattern Recognition, pp. 1–6 (2007)
7. Bolle, R.M., Connell, J.H., Pananti, S., Ratha, N.K., Senior, A.W.: The relation between the ROC curve and the CMC. In: Proceedings of the 4th IEEE Workshop on Automatic Identification Advanced Technologies, pp. 15–20 (2005)
8. Choi, J.Y., De Neve, W., Ro, Y.M.: Towards an automatic face indexing system for actor-based video services in an IPTV environment. IEEE Trans. Consumer Electron. **56**(1), 147–155 (2010)
9. Cover, T.M., Thomas, J.A.: Elements of Information Theory. Wiley, New York (1991)
10. De Marsico, M., Nappi, M., Riccio, D.: FACE: Face Analysis for Commercial Entities. In: Proceedings of the International Conference on Image Processing, pp. 1597–1600 (2010)
11. De Marsico, M., Nappi, M., Riccio, D.: Measuring measures for face sample quality. In: Proceedings of the International ACM

Workshop on Multimedia in Forensics and, Intelligence (MiFor'11) (2011)
12. Doddington, G., Liggett, W., Martin, A., Przybocki, M., Reynolds, D.: Sheep, goats, lambs and wolves: a statistical analysis of speaker performance in the NIST 1998 speaker recognition evaluation. In: Proceedings of International Conference on Spoken Language Processing (ICSLP), vol. 4, pp. 1351–1354 (1998)
13. Gallager, R.G.: Information Theory and Reliable Communication. Wiley, New York (1968)
14. Golić, J.D., Baltatu, M.: Entropy analysis and new constructions of biometric key generation systems. IEEE Trans. Inf. Theory **54**(5), 2026–2040 (2008)
15. Grother, P., Tabassi, E.: Performance of biometric quality measures. IEEE Trans. Pattern Anal. Mach. Intell. **29**(4), 531–543 (2007)
16. Jain, A.K., Nandakumar, K., Nagar, A.: Biometric template security. EURASIP J. Adv. Signal Process. Special issue on Pattern Recognition Methods for Biometrics (2008)
17. Jassim, A.J., Al-Assam, H., Abbound, A.J., Sellahewa, H.: Analysis of relative entropy, accuracy, and quality of face biometric. In: Proceedings of the Workshop on Pattern Recognition for IT (2010)
18. Kryszczuk, K., Richiardi, J., Prodanov, P., Drygajlo, A.: Reliability-based decision fusion in multimodal biometric verification. EURASIP J. Adv. Signal Proc. **2007**(1), 74–83 (2007)
19. Milborrow, S., Nicolls, F.: Locating facial features with an extended active shape model. In: European Conference Computer Vision, pp. 504–513 (2008)
20. Phillips, P.J., Wechsler, H., Huang, J., Rauss, P.: The FERET database and evaluation procedure for face recognition algorithms. Image Vis. Comput. J. **16**(5), 295–306 (1998)
21. Phillips, P.J., Beveridge, J.R., Draper, B.A., Givens, G., O'Toole, A.J., Bolme, D.S., Dunlop, J., Lui Y.M., Sahibzada, H., Weimer, S.: An introduction to the good, the bad, & the ugly face recognition challenge problem. In: 2011 IEEE International Conference on Automatic Face and Gesture Recognition and Workshops (FG 2011), pp. 346–353 (2011)
22. Rattani, A., Freni, B., Marcialis, G.L., Roli, F.: Template update methods in adaptive biometric systems: a critical review. In: ICB 2009, pp. 847–856 (2009)
23. Roli, F., Didaci, L., Marcialis, G.: Adaptive biometric systems that can improve with use. In: Ratha, N.K., Govindaraju, V. (eds.) Advances in Biometrics—Sensors, Algorithms and Systems. Springer, Berlin (2008)
24. Ross, A., Nandkumar, K., Jain, A.K.: Handbook of Multibiometrics. Springer, Berlin (2006)
25. Schmid, N.A., O′Sullivan, J.A.: Performance prediction methodology for biometric system using large deviations approach. IEEE Trans. Signal Process. Supplement on secure media **52**(10), 3036–3045 (2004)
26. Torres, L., Vilà, J.: Automatic face recognition for video indexing applications. Pattern Recognit. **35**(3), 615–625 (2002)
27. Uludag, U., Ross, A., Jain, A.: Biometric template selection and update: a case study in fingerprints. Pattern Recognit. **37**, 1533–1542 (2004)
28. Wang, Z., Bovik, A.C.: A universal image quality index. IEEE Signal Process. Lett. **9**(3), 81–84 (2002)