A statistical approach to reliability estimation for fingerprint recognition

Luis Di Martino[*], Alicia Fernández[†], Rafael Grompone von Gioi[‡], Federico Lecumberry[§], Javier Preciozzi[¶]

Abstract

In this work we focus in the reliability estimation of biometric systems output. We explain why this is a very important problem when deploying a biometric system and face it using a statistical approach. In particular, we present a solution based in the *a-contrario* approach widely used in the image processing field. We show how this strategy could be adapted and its key advantages with respect to other state-of-the-art reliability measures. A comprehensive set of experiments is used to validate the approach, using different fingerprints databases, matching systems, and comparing the performance with other state-of-the-art confidence measure strategies.

1 Introduction

The identification process can be done in two different scenarios: closed-set and open-set identification. The former occurs when it is certain that the searched identity is enrolled in the database and therefore the assigned identity is the one corresponding to the gallery sample closest to the query sample. The second corresponds to the case where the searched identity may have been or not previously enrolled in the system. In this case, the distance of the gallery closest sample is validated against some pre-defined threshold before assigning its identity to the input sample. This threshold has to be adjusted considering the performance of the biometric system. Usually this is done using a training dataset and the obtained value its applied globally for all the different system inputs.

[*]Dirección Nacional de Identificación Civil / Instituto de Ingeniería Eléctrica - Facultad de Ingeniería - UdelaR, dimartino@fing.edu.uy

[†]Instituto de Ingeniería Eléctrica - Facultad de Ingeniería - UdelaR, alicia@fing.edu.uy

[‡]CMLA, ENS Cachan, grompone@cmla.ens-cachan.fr

[§]Instituto de Ingeniería Eléctrica - Facultad de Ingeniería - UdelaR, fefo@fing.edu.uy

[¶]Dirección Nacional de Identificación Civil / Instituto de Ingeniería Eléctrica - Facultad de Ingeniería - UdelaR, jpreciozzi@dnic.gub.uy

The use of a validation threshold over the result of the identification being done allows to implement a quality control and therefore estimate the confidence (or reliability) of the system output. Even in the closed-set identification context it is useful to have this control as the gallery closest sample could correspond to a different identity of the one corresponding to the input sample. This outcome could be caused by a bad quality input or gallery enrolled sample or just because some enrolled sample of an incorrect identity is more similar than the one of the correct identity. The obtained confidence measure is also useful when multiple biometrics systems are used, as one needs to know how to assign the relevance of each output in order to combine them and obtain a unique result. When only one biometric system is used, the reliability control could be adjusted to meet the application requirements: for instance, the identification of a person at passport issuance offices needs to be more reliable than the automatic identification on social networks for tagging.

A commonly used strategy to estimate the confidence of a biometric system output is to evaluate the quality of the input biometric sample. In the case of fingerprints, several characteristics can be used to measure it. In [10] a summary of different efforts in this direction are presented, as well as the key ideas used in the definition of the *NFIQ* (NIST Fingerprint Image Quality) index. These strategies have the advantage of being independent of the particular feature extraction and matching techniques later used for processing the biometric sample. Nevertheless, this is also their main disadvantage: if there is little relation between the characteristics used on the quality analysis and the ones used on the matching process, the confidence measure obtained from the former may be inaccurate at the classification level.

Another common approach to solve the problem of reliability estimation is to use margins. These quantize the risk associated to a particular system output distance or score. In [8] a margin based on the *false reject rate* (*FRR*) and *false acceptance rate* (*FAR*) indices is presented. The authors derive a threshold value where these two measures are equal, *equal error rate* (*EER*) operating point, and use the difference between the obtained output and threshold as a confidence measure. The farther the output is from the threshold the more confidence is assigned. Finally, the match is validated or rejected according to the sign of this difference. This margin approach differs from other margin strategies as margin in boosting [4] or Vapnik's *margin slack variable* [11] in that the last two can only be computed once the result corresponding class/label is known. Therefore, these strategies are only useful in a training phase where they could be used to select those examples that are difficult to classify and use them to retrain the classifier. The *EER* based margin only requires labeled data in a development phase to obtain the optimum threshold value and then it could be directly applied in a testing scenario. Despite this, the approach presents a major problem

for its implementation. The used margin function is global in the sense that the same threshold is used for all the different biometric system inputs. A good reliability measure should be adaptable to the particular features of the input sample and its relation to the gallery enrolled samples. As it is well known that, given a biometric trait, some people are more difficult to classify than other (Doddington's zoo [3]).

Considering the previous statements, in [7] the authors present a list of required properties that a good confidence measure should meet. It should take into consideration the whole gallery and the input individual query sample; it should be well adjusted to the particular features of the biometric system being used, and it should not depend on any *a priori* knowledge of the whole query dataset. As from the operational point of view, it must provide for each input, a unique reliability measure that can be easily interpreted and used. The authors present two *"system response reliability"* (*SRR1* and *SRR2*) measures and apply them in the identification process of a face recognition system. The former use the difference of distances between the two samples closest to the input sample. The latter is a density measure that takes in consideration the relation between the gallery sample whose identity is assigned to the input and all the other enrolled biometric samples. This, as explained by the authors, makes the measure a little harder to compute but more robust with respect to outliers. The proposed reliability measures complies with the stated requirements and improve results with respect to Poh and Bengio margin strategy. Despite this, it still has two drawbacks: first, the *SRR1* and *SRR2* indices depends on thresholds that are obtained in a training phase. If the characteristics of the gallery dataset or input biometric samples drastically change these thresholds should be retrained. Second, as both measures use different criteria to estimate the system output reliability normally they do not perform good both at the same time. Therefore, a choice of which measure to use should be made and the selected one could no be optimal for a particular input.

In this article we present a confidence measure based on a statistical approach that comply with the established requirements and overcomes the problems observed in other state-of-the-art strategies. In particular, we show how the *a-contrario* framework could be adapted to the problem of reliability estimation for biometric identification. Some recent articles on face [2] biometric trait have already used a *a-contrario* framework (see [1] for a general reference) for biometric verification. But, these works do not tackle the problem of reliability estimation and its applicability to the identification operation mode. The article continues as follows: In Section 2 we introduce the proposed reliability measure strategy for identification, which is the main contribution of the present article. In Section 3 we present the experimental setup: the metrics used to analyse the different methods, the databases used to perform the experiments and the different experiments we have done in order to evaluate the proposed technique. In Section 4 we review the main

results obtained and we finally draw some conclusions and perspectives for future work in Section 5.

2 An *a-contrario* response validation strategy

The main idea behind the *a-contrario* framework can be easily explained by means of the *Helmholtz Principle*, that states that perceptually relevant events present large deviations from randomness. That is, given a random or *background* model, an event which is very rare under this model must follow a particular causality. This principle could be easily applied in the particular context of reliability estimation for a biometric identification system. As part of the identification a query sample is compared against all the representatives in the gallery dataset. The comparison against its corresponding sample of the same id would produce, normally, a distance that deviates (is much lower) from the other values obtained. When this occurs, according to the principle, the event is consequence of some particular causality and should be carefully considered. The idea could be formalized as follows. Let be q_i and g_j a query and gallery sample respectively, and let $d(q_i, g_j) = \delta_{i,j}$ be the distance between them. If we consider the event of obtaining distance $\delta_{i,j}$, it can be classified in one of the two following hypothesis:

- $\mathbf{H_0}$ **(null hypothesis)**: $\delta_{i,j}$ is observed only "by chance", indicating a realization of the background model.
- $\mathbf{H_1}$: $\delta_{i,j}$ is obtained by some causality, indicating a relevant event realization under the background model.

In our biometric scenario, we will associate $\mathbf{H_0}$ hypothesis to the comparison of two biometric samples that corresponds to different people, and $\mathbf{H_1}$ hypothesis to the comparison of two samples of the same person. Using these definitions is clear that an observation that corresponds to $\mathbf{H_0}$ is not significant because it is the most probable case and does not follow any causality. As explained in the introduction, a common approach to system reliability is to use a threshold on the distances, let's note it $\bar{\delta}$. A realization of the event that produces a distance bigger than $\bar{\delta}$ is classified as belonging to $\mathbf{H_0}$, and it is classified as belonging to $\mathbf{H_1}$ on the other case. In this scenario, the goal of the *a-contrario* approach is to control the number of false alarms (*NFA*):

$$NFA(q_i, g_j) = N_{test} PFA(q_i, g_j) \tag{1}$$

where $PFA(q_i, g_j)$ is the *"Probability of False Alarm"* associated to the particular match between q_i and g_j:

$$PFA(q_i, g_j) = P\left(d\left(q_i, g_j\right) \leq \bar{\delta} | H_0\right) \tag{2}$$

the term N_{test} accounts for the total number of comparisons realized in the particular experiment. In this way, the NFA index is a direct approximation to the expected value of false alarms that will occur as a result of the experiment. The main idea on the *a-contrario* approach is to bound this expectation, which leads directly to a bound in the distance.

Recall that this is an *a-contrario* decision, because the acceptance of $\mathbf{H_1}$ is done as soon as the null hypothesis ($\mathbf{H_0}$) is not likely to be valid for a particular realization of the event. Finally, a threshold ε is applied over *NFA*:

$$NFA(q_i, g_j) \leq \varepsilon \tag{3}$$

A particular realization of the event is considered relevant, and called $\varepsilon - meaningful$, if this restriction is met. The statistical test being done in the *a-contrario* framework could be easily related with the classical *Fisher's hypothesis test*. In fact, if a single test is made ($N_{test} = 1$) the threshold ε accounts for the significance level of the test and the test would be rejected whenever its *p-value* is less than ε. But, the *a-contrario* framework is applied in a scenario of multiple hypothesis testing or multiple comparisons [9]. In this context, the *NFA* measure equals to N_{test} times the *p-value* of the test and corresponds to what is called the Bonferroni Correction. Nevertheless, even if a formal analogy exist between the two approaches, the underlying aim of them is different. This is discussed in [5] and more completely in [1].

It is worth noting that, finally, the classification is performed by means of detecting the events that do not comply the null hypothesis. This is one of the reasons that explains the popularity of *a-contrario* methods with respect to classical hypothesis testing: the model we test against is not the one that describes the rare events but the *a-contrario* one that in general can be obtained easily and with more precision as there are many more representatives of this hypothesis.

2.1 Application to reliability estimation

The previously introduced framework could be easily adapted to the problem of reliability estimation. The key idea in this adaptation is that the outputs of the system that should be reliable are those in which the system achieves a correct identification. This corresponds to the case in which the query sample q_i is correctly associated with the gallery sample g_i. This match is very rare to occur when performing the identification in a large database and therefore could be thought as a realization of the hypothesis $\mathbf{H_1}$ defined above. Therefore, the presented *a-contrario* framework could be used to estimate the likelihood that a particular match belongs to the null hypothesis $\mathbf{H_0}$. If this is very likely to occur, the output should be considered not reliable.

In order to compute $PFA(q_i, g_j)$, the hypothesis $\mathbf{H_0}$ should be characterized. This is done by computing the probability density function (*pdf*)

$p_{q_i,\mathbf{H_0}}$ of distances against the query sample q_i under this hypothesis. It is worth noting that this calculation requires various samples of distances belonging to the hypothesis $\mathbf{H_0}$. Fortunately, in a closed-set identification scenario, these distances are already available as a result of the comparison between the input sample and those of the identities enrolled in the gallery. The final *PFA* is:

$$PFA(q_i, g_j) = \int_0^{\delta_{i,j}} p_{q_i,\mathbf{H_0}}(x)dx \tag{4}$$

In a confidence measure estimation scenario only the matches corresponding to the identifications are to be evaluated. With this in mind, if a particular query sample q_i has been associated the identity of the gallery sample g_j (the one that produces the lower distance), the *NFA* corresponding to this match is computed as follows:

$$NFA(q_i) = |N_Q|\, PFA\,(q_i, g_j) \tag{5}$$

where $|N_Q|$ is the size of the query dataset Q. It is worth noting that, in this reformulation, the constant multiplying the *PFA* is adjusted to the size of the hypothesis test being done. Finally, this measure is used to perform the validation according to the value of ε being used. If the match is considered $\varepsilon - meaningful$, it is labeled as reliable. On the other case, we consider it a realization of the null hypothesis $\mathbf{H_0}$ and labeled as not reliable.

3 Experimental setup

In order to compare the different strategies presented above, we perform three experiments using different databases and matching systems. The *a-contrario* reliability measure is compared, in each case, with *SRR2* (as presented in [7]) and quality based confidence measure strategies.

3.1 Databases

The experiments are performed using two different datasets: *FVC2004* [6] and *DNIC*.

The ***FVC2004*** database was created for its use in the third international *Fingerprint Verification Competition* carried on in Italy in 2004. This database includes four subsets (*DB1, DB2, DB3, DB4*) of fingerprints of forefinger and middle finger collected using different sensors as well as a synthetic fingerprint generator. In this work, the *DB1* subset from this database was chosen for the evaluation. In this database an additional division is realized, resulting in subsets *DB1_A* and *DB1_B*, these datasets were used in the competition for testing and training respectively. Only the first subset is considered in this paper, it corresponds to a test dataset including 800 fingerprints (100 identities with 8 samples each one).

The problem with this dataset in a identification scenario is that it includes multiple biometric samples per identity, this is commonly used in a identity verification evaluation. As in this case the identification is being evaluated, an adaptation of the database should be done in order to define a gallery and query subsets as usual in this type of test. The following procedure was followed: from each identity only two fingerprints were selected randomly from the eight available fingerprints. One of them was included in the gallery and the other was added to the query dataset. In this way, the experiments with the *FVC2004* database consist in the identification of 100 identities having only one sample per person in the gallery dataset. This database was chosen in order to report the performance of the proposed technique in a public and standard database.

The ***DNIC*** database corresponds to a test subset taken from a on-production environment where fingerprints are acquired as part of the enrollment process of a subject in a citizens IDs issuance office. The fingerprints in the gallery subset were obtained from scans at 500 dpi, of historical records of thumbs rolled fingerprints. The images have a size of, at least, 700×700 pixels. The corresponding query samples where obtained using a fingerprint optical sensor that scans fingerprints at 500 dpi and produces images of size 416×416 pixels. The selected subset includes 1000 identities, each one having a unique sample in the gallery dataset. This database was chosen because it is a good representative of a real, on-production environment that presents variations that could not arise in a controlled lab-environment.

3.2 Matching

In order to achieve a performance evaluation of the proposed strategy as complete as possible, we also used two different fingerprint matching systems, we called them "*M1*" and "*M2*".

M1, is open and public, this system is obtained from the NBIS (Nist Biometric Image Software) software distribution developed by the National Institute of Standards and Technology (NIST). The MINDTCT utility automatically locates and records minutiae points, these feature are used with the BOZORTH3 matching software to obtain a similarity score between fingerprints. Additionally, the NFIQ software is used for the estimation of fingerprint quality. This utility uses various features and returns a quality map for each fingerprint and a total score in the range [1,5] where 1 and 5 indicates highest and lowest quality respectively. More information of the software and the particular used modules could be find in [12]. *M2* is a closed system acquired by the DNIC agency for the automatic matching of two fingerprints. This system receives as input two *WSQ*[6] compressed fingerprints and apply private algorithms for the minutiae extraction. The

[6] *WSQ* stands for "Wavelet Scalar Quantization", a compression algorithm that is the standard for the exchange and storage of fingerprint images.

extracted features are compared resulting in a score that indicates how similar two fingerprints are. The system returns also quality indices for each of the compared fingerprints, this index is in the range [0,255], the algorithms used for the estimation of the fingerprint quality are also private and, therefore, unknown to us.

While both systems produce a similarity score between fingerprints, in this article we are more comfortable using distances between samples. These could be easily obtained by inverting the obtained scores.

3.3 Experiments

We follow the standard procedure used to evaluate biometric systems working in a closed-set identification mode in the three experiments described below.

Experiment 1

In the first experiment we use the *FVC2004* database and *M1* matching system. This experiment shows the results obtained with both a public database as well as public matching system. We believe that presenting this experiment is of great importance for the reproducibility of the obtained results and the evaluation of the proposed technique in conditions well known by the fingerprint recognition community.

Experiment 2

The second experiment is done using the same dataset but changing the fingerprint matching system. By using *M2*, the robustness of the presented confidence measure strategies with respect to changes in the biometric system could be evaluated.

Experiment 3

In the last experiment we present the results obtained in database *DNIC* with the *M2* system. This experiment is of great value because it shows the performance in a bigger database, obtained in a production environment and using a private biometric system. These conditions are difficult to reproduce by other researchers. On the other hand, these are the ones more similar to a real case scenario and therefore of great importance in the evaluation of the applicability of the presented technique.

3.4 Performance measure

Using the distances between samples, three different reliability measure approaches are evaluated. First, the function φ_2 is used for the computation

of index *SRR2*, the threshold γ is applied over this index. Second, the presented *a-contrario* strategy is applied by thresholding the *NFA* associated to each identification. In order to use the same range for the parameters of both the *SRR2* and the *a-contrario* strategy we define and use a threshold ε' as follows:

$$\varepsilon' = \frac{NFA(q_i, g_j)}{N_{test}} \tag{6}$$

In this way the threshold ε' varies in the range $[0, 1]$. Last, the quality indices retrieved from *M1* and *M2* are used as reliability measures of the input fingerprint samples. For each match, the minimum quality of both fingerprints is considered and a threshold applied over it in order to discard or accept the match. Its important to remark that, in this last case, the operating points are limited to the range of the quality values (5 and 255 when *M1* and *M2* are used respectively).

In order to compare the different reliability strategies, we use the following measures: the number of reliable responses (NRR) and the recognition rate (RR). The NRR is defined as follows:

$$NRR(\gamma) = \frac{|N_{rr}(\gamma)|}{|N|} \tag{7}$$

where subset $N_{rr}(\gamma)$ represents the subjects q_i of the query dataset Q in where the reliability measure $r(q_i)$ complies with a minimum required confidence threshold γ: $N_{rr}(\gamma) = \{q_i \in Q | r(q_i) > \gamma\}$.

The recognition rate RR is computed over the reliable responses:

$$RR(\gamma) = \frac{|N_{match}(\gamma)|}{|N_{rr}(\gamma)|} \tag{8}$$

where $N_{match}(\gamma)$ is the subset of reliable responses that corresponds to correct answers of the system:

$$N_{match}(\gamma) = \{q_i \in N_{rr}(\gamma) | d(q_i, g_i) \leq d(q_i, g_j) \forall g_j \neq g_i\}.$$

The number of reliable responses accounts for the amount of outputs of the system that complies with the reliability threshold being used. The recognition rate is simply the ratio of people correctly identified by the system, those in which the correct identity appears in first position in a candidate list, over the identifications marked as reliable. A good confidence measure should achieve an improvement in the RR as NRR becomes lower. This represents the situation in which, as the number of discarded matches increases, the outputs of the system left as reliable are those in which it really performs best. It would not make any sense to use a confidence measure technique that, discarding matches, does not improve the obtained RR.

In the case of the *a-contrario* formulation, the set N_{rr} is redefined as follows:

$$N_{rr}(\varepsilon) = \{q_i \in Q | NFA\,(q_i) < \varepsilon\} \tag{9}$$

Note that the only difference is that in the *a-contrario* framework, the threshold is an upper bound. It is important to remark how the response reliability threshold is applied in each case. When *SRR2* is used, high values in this index indicate great confidence in the response. Therefore, the responses are filtered by using γ as a lower threshold and the *NRR* index decreases as the threshold over *SRR2* increases. When the *a-contrario* approach is used, the use of the reliability threshold is inverted because higher values of the ε threshold relaxes the restriction allowing a higher number of false alarms (*NFA*). Higher values of the ε threshold results in higher *NRR* and lower *RR*.

For both the *SRR2* and *a-contrario* approach the evolution of *NRR* and *RR* is plotted as the thresholds γ and ε are varied. This allows to review the different working points of both techniques but makes very difficult the direct comparison of them. For this reason, the relation *RR* vs. *NRR* is also plotted, in this scenario a better reliability measure would be the one that, for a same *NRR* of reference, produces a higher *RR*. This implies that the system is correctly assigning more confidence to the outputs in which the identification is accurate and, on the other hand, rejecting incorrect matches.

4 Results

The results of the first experiment are shown in Figure 1. In particular, in Figures 1a and 1b the evolution of indices *RR* and *NRR* are shown against the thresholds over *SRR2* and ε respectively. It is important to note that the *M1* matching system does not present, at first, a very good performance. If a reliability control is not used and all outputs are accepted, this is represented by the working point were $NRR = 1$, the recognition rate obtained would be $RR = 0.65$. This could be considered a low performance for a fingerprint recognition system working in a database of only 100 identities. Despite this, both the *a-contrario* and the *SRR2* strategies are shown useful when implementing a reliability control in these conditions. Since both achieve an increase in the *RR* rate while samples are being rejected, represented by a decrease in the *NRR* rate.

In Figure 1c, both confidence measures are compared with the control established by means of thresholding the fingerprints quality. As stated before, the problem with such approach is that only a few operating points of the system are available, according to the range of the quality index being used. In this particular case, this becomes very problematic as the *NFIQ* quality index being used with matcher *M1* only return values in the $[1, 5]$ range. In this case, it is clear that both techniques obtains better results than the simple use of a quality index obtained only by considering the input sample features. The *a-contrario* technique obtains a significant

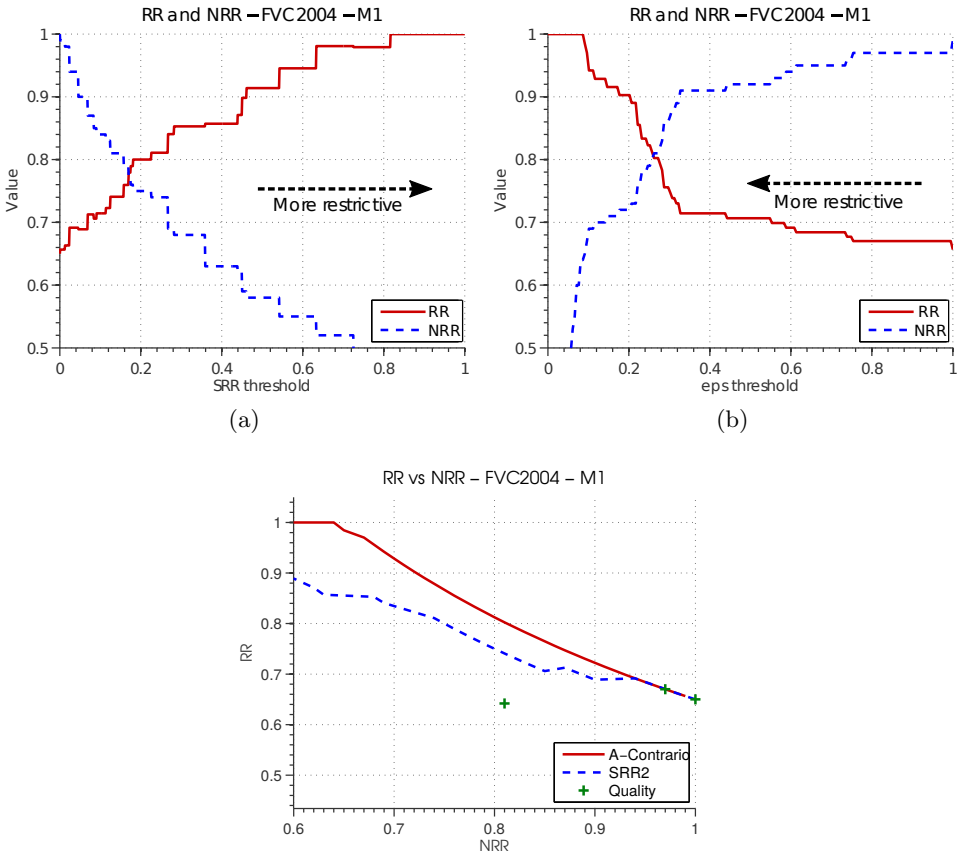

Figure 1: Results in *FVC2004* database using *M1* matcher.

improvement over *SRR2*, represented by higher values of the *RR* measure for any same reference value of the *NRR*. This difference could be attributed to the very nature of both confidence measures and the information they use. The *SRR2* index measures the significance of the identification test by evaluating if each sample in the gallery produced a distance against the input sample greater than a threshold. But the particular distribution of these distances is not considered, with the *a-contrario* approach all this information is taken into account.

The results of the second experiment are shown in Figure 2, in this case only the *RR vs. NRR* relation is shown, several observations can be made. First, the *M2* matcher system present a significant improvement in performance using the same dataset. This is clear when comparing the working point where $NRR = 1$. The *M2* system obtains a recognition rate of $RR = 0.94$, much higher than the $RR = 0.65$ obtained with *M1*. Second, the quality index associated to *M2* gives place to more configuration

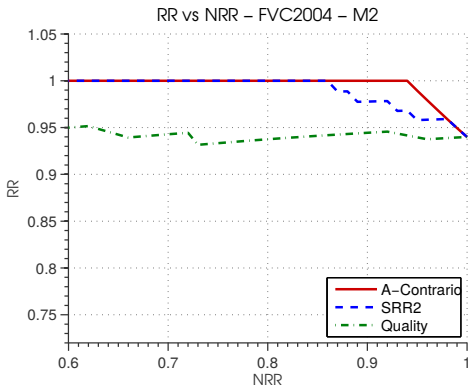

Figure 2: Results in *FVC2004* database using *M2* matcher.

points of a sample quality based reliability control (256 against 5). Despite this, once again the quality measure does not represent a useful confidence measure strategy, as the obtained *RR* does not improve over variations in the *NRR*. Last, it can be seen that both the *SRR2* and *a-contrario* approach represent good solutions to the reliability estimation problem. Both strategies take advantage of the improvements in performance of the matcher *M2*, that produces as a consequence a better separation between the distances corresponding to each hypothesis $\mathbf{H_0}$ and $\mathbf{H_1}$. The *a-contrario* technique continues to perform equal or better than *SRR2*.

The results of the last experiment are shown in Figure 3. The performance of *M2* in this database is worse than the obtained previously, this is to be expected considering the increase in database size as well as the difficulties that may be introduced by using an on-production database. Despite this, both the *a-contrario* and *SRR2* strategies remain good solutions to the reliability measure problem. From the comparison of Figures 3a and 3b, it can be seen that the increase rate in the *RR* as the system becomes more restrictive is higher when the *a-contrario* approach is used instead of *SRR2*. This difference in performance also is evident in Figure 3c that allows to compare both solutions quantitatively. In particular, if a *NRR* of 0.9 is used (a value that represents a coherent tradeoff), the *SRR2* technique achieved approximately a *RR* of 0.95 while the *a-contrario* strategy attains 0.98. While this is a difference of only a 3%, its very significant considering that we are considering *RR* in the 90% range. This difference in performance could be attributed to the fact that, in this experiment, the characterization of the null-hypothesis done in the *a-contrario* solution is more accurate as more representatives of it are available. The quality based criteria, once again, shows a poor performance as a reliability control measure as the *RR* remains practically unchanged as the number of reliable responses is decreased.

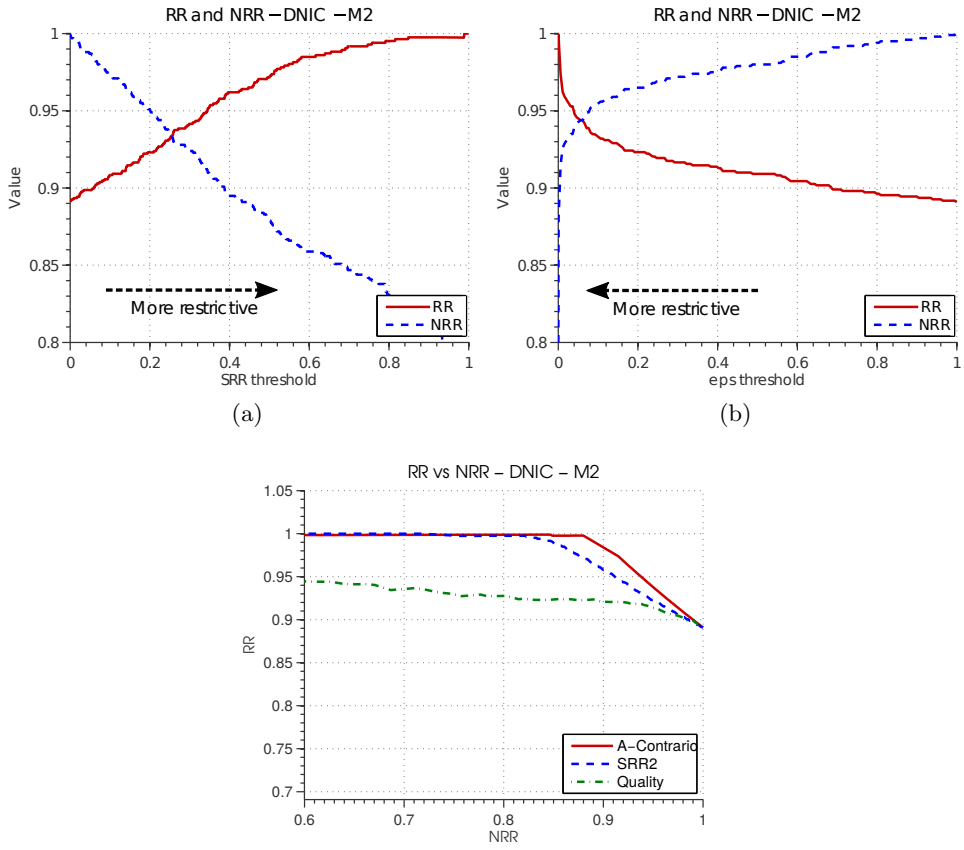

Figure 3: Results in *DNIC* database using *M2* matcher.

5 Conclusions and future work

As first conclusion, is worth noting that biometric systems reliability are not commonly measured using a confidence measure complying with the requirements described in this article. This induces a big problem when the biometric system is deployed in a on-production environment because a confidence could be only estimated based in a previous testing of the system. Likely this was done using biometric samples with different characteristics that the ones being used on-production. Secondly, the experiments performed shows that good confidence measures are those that consider each particular output of the system and the whole gallery. By using the output of the complete system and not a estimation in a particular stage (as preprocessing, feature extraction, matching) the confidence measure is able to characterize the biometric system completely and estimate with high precision when a particular output should be considered reliable. In the article

it was shown that a criteria based only in the input sample (as the fingerprint quality) results in very poor reliability estimation control. From the different strategies that meet these restrictions, it is observed that the ones using all the information available in a statistically way results in better confidence measures. They are more complex to compute than the other presented techniques, but present a significant improvement in performance. In particular, the *a-contrario* based technique presented in the article provides an elegant solution to this problem that performs very well and allows to establish a working point of the system in advance. As future work we are planning the application of this same method in other biometric modalities, especially in face recognition. Additionally, we will like to extend the technique to address the fusion of various biometric modalities. We believe that the combination of information from different sources could be used to better characterize the distributions of impostors and genuine distances. This could considerably improve the results.

References

[1] Agnès Desolneux, Lionel Moisan, and Jean-Michel Morel. *From gestalt theory to image analysis: a probabilistic approach*, volume 34. Springer, 2007.

[2] Luis Di Martino, Javier Preciozzi, Federico Lecumberry, and Alicia Fernández. An a-contrario approach for face matching. In *ICPRAM 2014 - International Conference on Pattern Recognition Applications and Methods*, pages 377–384, 2014.

[3] George Doddington, Walter Liggett, Alvin Martin, Mark Przybocki, and Douglas Reynolds. Sheep, goats, lambs and wolves: A statistical analysis of speaker performance in the nist 1998 speaker recognition evaluation. Technical report, DTIC Document, 1998.

[4] Yoav Freund and Robert E. Schapire. A short introduction to boosting. In *In Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence*, pages 1401–1406. Morgan Kaufmann, 1999.

[5] Rafael Grompone von Gioi, Jérémie Jakubowicz, Jean-Michel Morel, and Gregory Randall. On Straight Line Segment Detection. *Journal of Mathematical Imaging and Vision*, 32:313–347, 2008.

[6] Dario Maio, Davide Maltoni, Raffaele Cappelli, JimL. Wayman, and AnilK. Jain. Fvc2004: Third fingerprint verification competition. In David Zhang and AnilK. Jain, editors, *Biometric Authentication*, volume 3072 of *Lecture Notes in Computer Science*, pages 1–7. Springer Berlin Heidelberg, 2004.

[7] Maria De Marsico, Michele Nappi, Daniel Riccio, and Genoveffa Tortora. Nabs: Novel approaches for biometric systems. *IEEE Transactions on Systems, Man, and Cybernetics, Part C*, 41(4):481–493, 2011.

[8] Norman Poh and Samy Bengio. Improving fusion with margin-derived confidence in biometric authentication tasks. In *Fifth Int'l. Conf. Audio- and Video-Based Biometric Person Authentication AVBPA*, 0 2005.

[9] J P Shaffer. Multiple hypothesis testing. *Annual Review of Psychology*, 46(1):561–584, 1995.

[10] Elham Tabassi, Charles L. Wilson, and Craig I. Watson. Fingerprint image quality: Nistir 7151,â tech. rep, 2004.

[11] Vladimir N. Vapnik. *Statistical Learning Theory.* Wiley-Interscience, 1998.

[12] Craig I. Watson, Michael D. Garris, Elham Tabassi, Charles L. Wilson, R. Michael McCabe, Stanley Janet, and Kenneth Ko. User's guide to nist biometric image software (nbis). `http://www.nist.gov/customcf/get_pdf.cfm?pub_id=51097`. Accessed: 2015-01-31.