Biometrics system concepts: first report

Luca Stradiotti

1 Introduction

In this report, the focus is on evaluating the performance of two biometric systems both in a verification and identification setting. Two systems are considered: the fingerprint for the left and the one for the right index. To evaluate these systems, the actual predicted fingerprint similarity scores are used; it allows to bypass all steps of preprocessing, feature extraction and matching so as to concentrate only on the score evaluation. The data (biometrics scores set BSSR1) came from the American National Institute of Standards and Technologies (NIST).

2 Validation of verification system

The first setting is the verification. This is used to authenticate a claimed identity, to verify that a person is who he/she claims to be. There are two possible classes, denoted as:

- 0 - Impostor (False)
- 1 - Genuine (True)

Furthermore, the set of genuine scores is represented as G, while the set of impostor ones as I.

2.1 Genuine and impostor score distributions

Firstly, the impostor and genuine score distribution for both systems can be plotted to observe some initial insights in the systems.

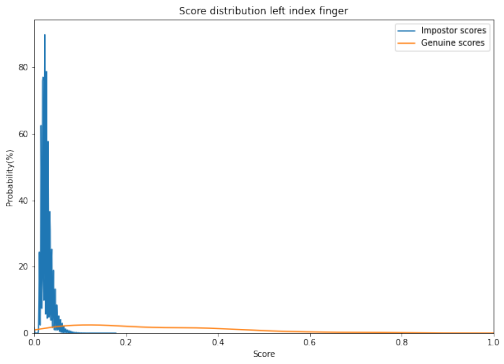

Figure 1: Score distribution left index system

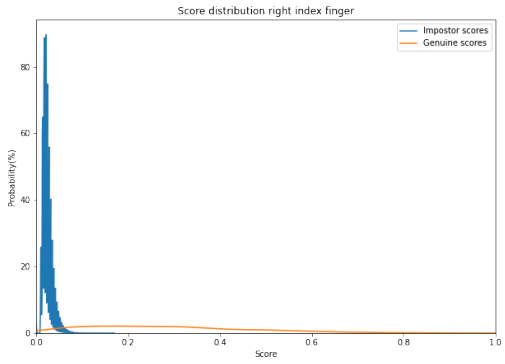

Figure 2: Score distribution right index system

In both systems, the impostor and genuine score distributions overlap only for a small amount and there are no impostor scores higher than 0.2. The impostor score distribution has a lower standard deviation than the genuine one, because genuine scores take value between 0 and 1. Moreover, the class distribution is highly unbalanced (1000 genuine, 999000 impostor).

2.2 FMR, FRR and Receiver Operating Characteristic (ROC) curve

The False Match Rate (FMR) represents the proportion of zero-effort impostor who are falsely recognized as matching a nonself template. It simply consists of an alias for the usual False Positive Rate metric.

The False Non-Match/Rejection Rate (FNMR/FRR) represents the proportion of genuine attempt who are falsely declared not to match their own template. It is simply an alias for the usual False Negative Rate metric. The FMR and FRR can easily be obtained given the previously computed probability distributions and a threshold

value t_0. Practically this corresponds to a counting problem, having $\mathcal{I}$ the indicator function (return 1 if x is true, else 0):

$$FMR(t_0) = p(s \geq \eta | I) \approx \frac{1}{|I|} \sum_{s \in I} \mathcal{I}(s \geq \eta),$$

$$FRR(t_0) = p(s < \eta | G) \approx \frac{1}{|G|} \sum_{s \in G} \mathcal{I}(s < \eta).$$

It is impossible to optimize both FMR and FRR: choosing a threshold is always a trade-off between them.

Considering different thresholds, different values of FAR and FRR can be obtained.

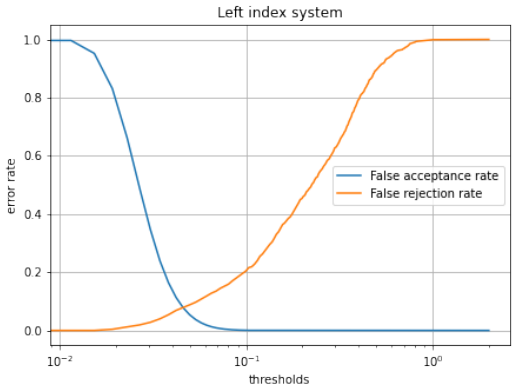

Figure 3: FAR and FRR left index finger

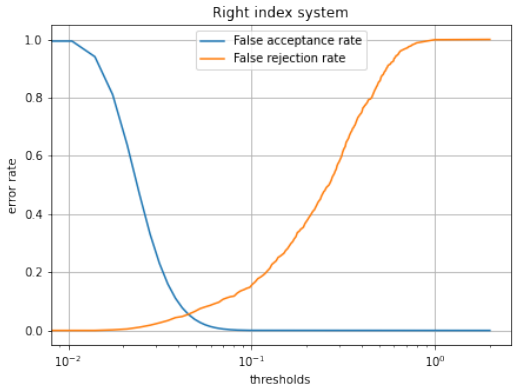

Figure 4: FAR and FRR right index finger

The curves show that FAR values for both systems are almost identical for all the considered thresholds, while the right index finger has generally lower FRR. The intersection point has lower FAR and FRR for the right index system, so this latter should perform better.

Another way to observe the impact of the threshold can be to plot the ROC curve, which plots FPR and TPR for different threshold values, or the DET curve, which plots FPR and FNR instead. These two curves are plotted using a logarithmic scale to obtain quite straight lines which are easier to be compared.

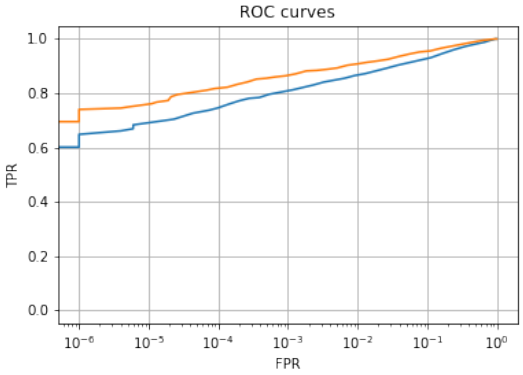

Figure 5: ROC curve for left and right index system

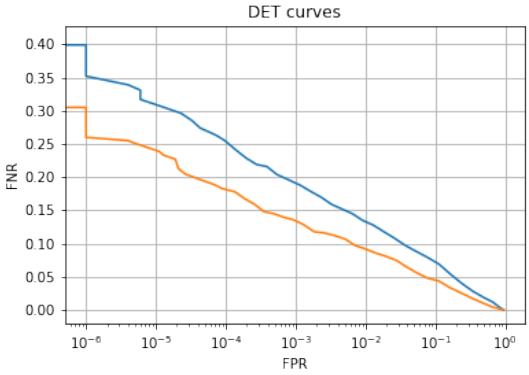

Figure 6: DET curve for left and right index system

Again, the right index finger performs better than the left one. Both curves represent the same information, since FNR can be computed as 1-TPR.

2.3 F1 and accuracy as metrics

2.3.1 Accuracy

Biometrics system are usually evaluated using FMR and FRR or ROC/DET curve because classes are usually highly unbalanced, but accuracy can also be considered to evaluate the performances of these systems. Considering different threshold values, different accuracy measures can be obtained and plotted.

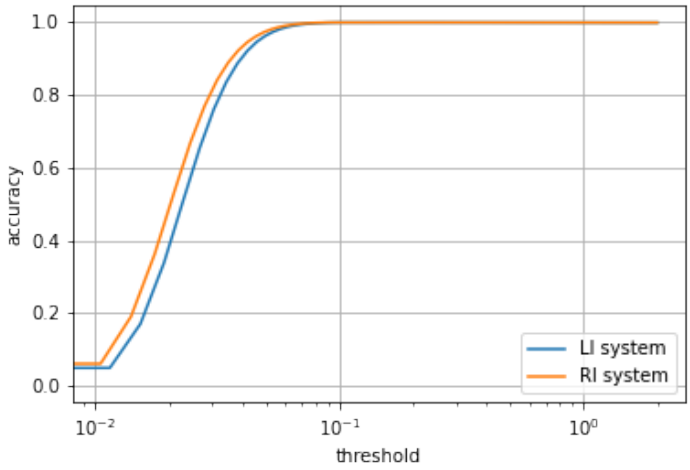

Figure 7: Accuracy for left and right index system

Both systems perform quite the same: their accuracy increases fastly close to 1, since there are no impostor scores higher than 0.2 and so with thresholds higher than this value all the impostor are well-classified. The maximum values for accuracy are reached when:

	t_{max}	Accuracy
Left index system	0.126	0.9997
Right index system	0.122	0.9998

Table 1: Maximum accuracy left and right index system

Accuracy is not a suitable classification metrics for these two systems because impostor and genuine classes are highly unbalanced.

2.3.2 F1 score

Also F1 score can be considered as evaluation metrics for biometrics system. It corresponds to the harmonic mean of precision, which is class prior dependent, and recall, which is instead class prior independent.

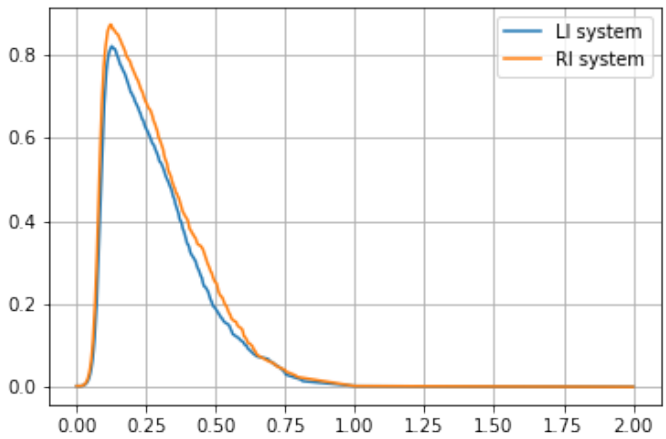

Figure 8: F1 score for left and right index system

The right index system performs again slightly better then the left one. The maximum values are:

	t_{max}	F1 score
Left index system	0.126	0.821
Right index system	0.122	0.873

Table 2: Maximum F1 score left and right index system

Maximum values for F1 score are reached for the same thresholds as for accuracy, nevertheless the performance difference between the two systems is higher.

The threshold with maximum F1 score could be an interesting operating point for a forensic application since, considering figures 3 and 4, it has a very low FAR value; for other applications, this point is not a good choice since the FRR value is quite high.

2.4 AUC and EER as summary measures

Comparing curves is not so easy, that is why metrics that return a single value are often considered. The overall performance considering all threshold settings is typically squeezed using the Area Under the Curve (AUC) or the Equal Error Rate (EER).

2.4.1 Area Under the ROC curve (AUROC)

The Area Under the ROC curve (AUROC) value can be used to obtain a summary measure about a system. This value reveals the amount of area under the ROC curve: so the larger this number, the better.

	AUROC
Left index system	0.971
Right index system	0.983

Table 3: AUROC left and right index system

The two values are very close to 1 and the right index system performs again slightly better than the left one.

However, AUROC is a summary measure and, for example, it is possible for a lower AUROC classifier to outperform a higher AUROC one in a specific region: so, it is convenient to always inspect the full ROC curve to make decisions.

2.4.2 Equal Error Rate (EER)

Another way to obtain an overall performance for a system is considering the Equal Error Rate, which represents the operation point where FAR and FRR have the same value.

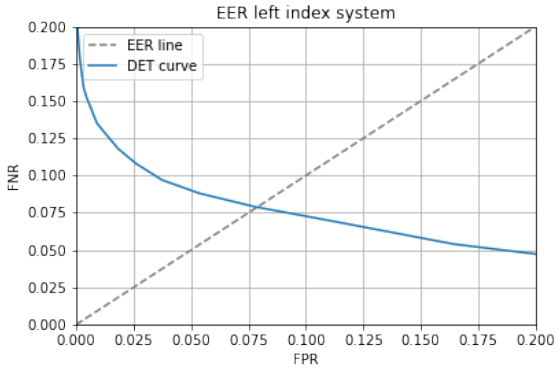

Figure 9: EER left index system

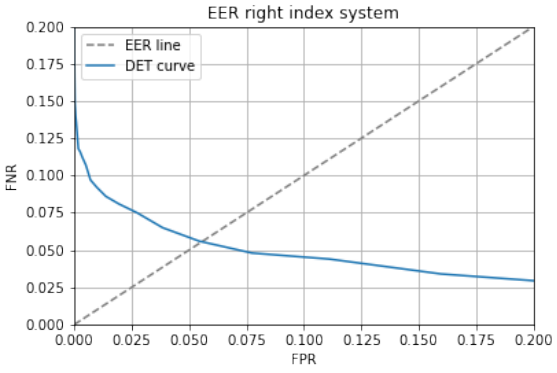

Figure 10: EER right index system

For the left index system, the EER is reached when $FAR = FRR \approx 0.075$; these values correspond to a threshold value of approximately 0.05. While, for the right index system the EER is reached when $FAR = FRR \approx 0.05$, and it is obtained considering approximately the same threshold of the left system.

2.4.3 Other strategies for "optimal" performance measure

Another possible approach could be to minimize the sum of FAR and FRR.

	Threshold for min(FAR+FRR)
Left index system	0.057
Right index system	0.059

Table 4: Threshold for min(FAR+FRR) left and right index system

Actually this measure assumes a balanced cost for misclassification error (the cost for false positive is equal to the cost for false negative) and a balanced class prior distribution. So, this operating point could be a suitable choice for system with equal cost for False Positive and False Negative and with balanced class prior.

If the misclassification costs and the class prior differ, it could be a better idea to minimize the overall classification cost, expressed as

$$\mathrm{C} = \mathrm{C_{FP}} * \mathrm{FPR} * \mathrm{P}(impostor) + \mathrm{C_{FN}} * \mathrm{FNR} * \mathrm{P}(genuine)$$

where $\mathrm{C_{FP}}$ and $\mathrm{C_{FN}}$ represent respectively the cost for false positive and the cost for false negative, FPR represents the false positive rate, FNR the false negative rate and $\mathrm{P}(genuine)$ and $\mathrm{P}(impostor)$ represent class prior probability. This measure could be a good trade-off also for a system with highly unbalanced prior since it takes into account also class prior distribution.

This approach could be useful if an overall performance measure is required, but often a rank classifier (ROC, DET curve) should be better since it allows the comparison between the performances of different systems in different regions.

2.5 Precision-Recall curves and related summary measures

Since classes are highly unbalanced in both systems, Precision-Recall curves represent a better performance metrics than ROC curves because the latter give an overly optimistic view of an algorithm's performance in this scenario. Precision-Recall curves summarize the trade-off between the true positive rate (recall) and the positive predictive value for a predictive model using different probability thresholds. They are better than ROC curves for high unbalanced classes since they take into account also a class prior dependent measure (precision).

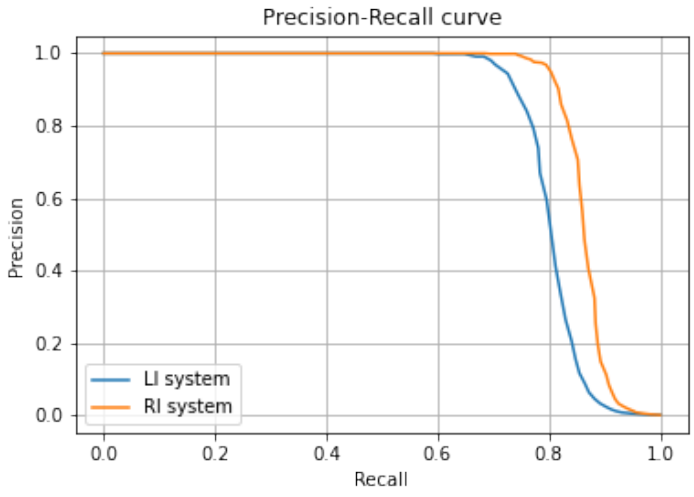

Figure 11: Precision-Recall curve for left and right index system

These curves can be summarized using two different summary measures: the Area Under the Precision-Recall curve (AUPR) or the Average-Precision score.

2.5.1 Area Under the Precision-Recall curve

The Area Under the Precision-Recall curve (AUPR) can be obtained computing the integral over the Precision-Recall curve. So the larger this number, the better.

	AUPR
Left index system	0.803
Right index system	0.863

Table 5: AUPR left and right index system

The right index system has a higher AUPR value, as it can also be noticed through figure 11.

2.5.2 Average Precision score

The Average Precision score summarizes a precision-recall curve as the weighted mean of precisions achieved at each threshold; the weights correspond to the increase in recall from the previous threshold.

$$\mathrm{AP} = \sum_n (R_n - R_{n-1})\, P_n$$

For the two systems:

	Average precision
Left index system	0.799
Right index system	0.860

Table 6: Average precision left and right index system

The Average Precision score is an approximation of the Area Under the Precision-Recall curve. These values, even if they are similar, are actually different to AUPR.

3 Validation of identification system

The second setting is the identification. This is used to determine the identity of a given person among all the possible identities that are enrolled in the system.

3.1 Evaluation using CMC curves

An identification scenario is usually evaluated using the Cumulative Match Characteristic (CMC) curve which plots the probability that a correct identification is returned among the top-k ranked matching scores for $k = 1, 2, ..., N$, where N is the number of enrolled users.

CMC curves can easily be computed through the similarity matrix of the scores since these can give us the ranked matching scores for every test sample.

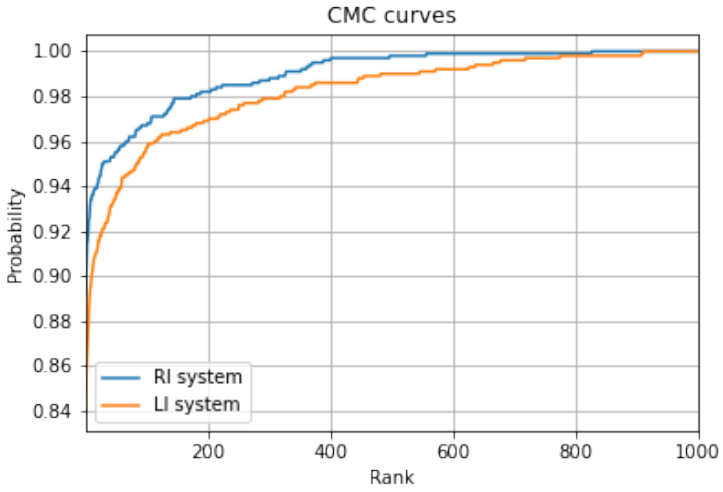

Figure 12: CMC curve for left and right index system

Also for the identification setting, the right index system performs better than the left one.

Considering the Rank-1 identification rate, the probability that a correct identification is returned as the highest matching score:

	Rank-1
Left index system	0.839
Right index system	0.892

Table 7: Rank-1 left and right index system

The right index system return a correct identification as highest matching score for 53 sample more than the left index one (this can be easily computed since the total amount of samples is 1000).

4 Conclusions

The right index finger performs slightly better than the left one with all the performance metrics analyzed in this report.

Moreover, for these two systems a good evaluation metric should take into account class prior distribution, since classes are highly unbalanced. For this reason Precision-Recall curve should be preferred to ROC or DET curve.

In conclusion, rank classifiers (e.g. ROC curve, DET curve, precision-recall curve)allow the comparison between the two systems in different regions, so they are more informative than a single performance measure (e.g. AUROC, AUPR, F1 score).