Face Recognition Vendor Test 2002
Performance Metrics

Patrick Grother, Ross Micheals
and P. Jonathon Phillips

31 March 2003

The paper is published as **NIST IR 6982** and will appear in the Proceedings of the **Fourth International Conference on Audio-Visual Based Person Authentication**, in June 2003.

The paper details the metrics used to quantify the performance of the face recognition systems tested in FRVT 2002. The methods are suited to any recognition evaluation, online or offline, technology or scenario, for which complete similarity scores are archived.

The paper shows that the open-set identification problem known as the watch list task is the general case: it requires systems to perform 1:N recognition with concurrent possibilities of false acceptance and rejection. Two special cases are demonstrated: 1:1 Verification is simply the watch list task with N=1; and closed-set identification is that with no false accept rate.

The paper also presents the computation of standard error ellipses used to show the effect of population variation on false accept and false reject rates.

Face Recognition Vendor Test 2002 Performance Metrics

Patrick Grother, Ross J. Micheals, and P. Jonathon Phillips

National Institute of Standards and Technology, Gaithersburg MD 20899, USA,
pgrother@nist.gov

Abstract. We present the methodology and recognition performance characteristics used in the Face Recognition Vendor Test 2002. We refine the notion of a biometric imposter, and show that the traditional measures of identification and verification performance, are limiting cases of the open-universe watch list task. The watch list problem generalizes the tradeoff of detection and identification of persons of interest against a false alarm rate. In addition, we use performance scores on disjoint populations to establish a means of computing and displaying distribution-free estimates of the variation of verification vs. false alarm performance. Finally we formalize gallery normalization, which is an extension of previous evaluation methodologies; we define a pair of gallery dependent mappings that can be applied as a post recognition step to vectors of distance or similarity scores. All the methods are biometric non-specific, and applicable to large populations.

1 Introduction

The evaluation protocol [2] used in FRVT 2002 [4] is a framework for the quantitative determination of the performance of recognition technologies using arbitrary biometrics. Specifically, it applies to either online (live human subjects) or offline (stored imagery) testing as long as it produces output recognition data that is available for subsequent analysis. This has the advantage that tests can be conducted uniformly and fairly across participants, the results are repeatable, and very large image sets can be tested expeditiously. We first present normalization in section 1.1 and then introduce the terminology of the FRVT 2002 protocol and define our performance metrics in section 2.

In the FRVT 2002 protocol a system is required to compare two biometric signatures and report a scalar similarity score. A biometric signature can be an arbitrary ensemble of pieces of imagery from an individual, for example a face and a fingerprint, or face and voice audio sequence. In FRVT 2002, both single face stills and video sequences were used. We use the term image to include such modalities, unless stated otherwise. A similarity score is a measure of the sameness of identity of the individuals appearing in the images. Without loss of generality the protocol requires that images of the same person have a large similarity score. Systems reporting distance measures, where a small value indicates sameness of identity, have their values negated before any processing.

The FRVT 2002 tests are structured around sets of images. An algorithm compares all images in a *query* set, $\mathcal{Q}$, with all images in a *target* set $\mathcal{T}$. The result is a similarity matrix whose ij-th element is the similarity between the i-th element of $\mathcal{T}$ and the j-th element of $\mathcal{Q}$. The matrix is computed and stored in column order; each column corresponds to an unknown query image being compared with all the known, enrolled, target images. In the general case the matrix is rectangular, but for FRVT 2002 we used the special case $\mathcal{T} = \mathcal{Q}$.

This framework allows for great flexibility in arriving at quantitative results. The sets $\mathcal{T}$ and $\mathcal{Q}$ are not normally used directly. Instead we consider *virtual* image sets, a conceptual innovation first defined in the FERET protocol[3]. Here a *gallery*, $\mathcal{G}$, a subset of $\mathcal{T}$, contains identically one signature per subject and represents the set of images that have been enrolled in a biometric system. Likewise a *probe* set, $\mathcal{P}_\mathcal{G}$, is a subset of $\mathcal{Q}$. Each of its images have a match in the gallery, and represent a legitimate user. The images of a third set, the *imposter* set $\mathcal{P}_\mathcal{N}$, also a subset of $\mathcal{Q}$, do not have a match in the gallery. This set represents persons attempting to defeat a system. A *match* describes the comparison of probe and gallery images of the same individual. A *non-match* likewise arises from images of different persons.

The utility of this framework is that many different recognition experiments are embedded within $\mathcal{T}$ and $\mathcal{Q}$. All the results of FRVT 2002 are obtained from the similarity elements corresponding to the rows defined by the subset $\mathcal{G}$, and the columns defined by $\mathcal{P}_\mathcal{G}$ and $\mathcal{P}_\mathcal{N}$. Together these form a similarity matrix S from which various performance statistics are computed. Disjoint gallery and probe sets allow performance to be estimated on particular recognition tasks. For example to explore the effect of subject ageing, a gallery containing the earliest image of all persons is used with a probe set of more recent images.

This protocol evaluates recognition technologies rather than deployed systems. Particularly it ignores efficiency and performance when databases are partitioned. See Wayman [5] for a treatment of these issues.

1.1 Normalization

The FRVT 2002 protocol allows a *normalization* option. This is a post-processing transform of similarity scores, that may exploit the fact that the gallery, unlike the target set, contains only one image per person by definition. Specifically, a vector, $\mathbf{s}$, contains the column of elements s_{ip} from S corresponding to the single probe p against gallery $\mathcal{G}$. Normalization is defined as a function, $f : R^N \rightarrow R^N$, mapping $\mathbf{s}$ to a new vector, $\mathbf{t}$. For an algorithm that uses normalization, the final performance scores are computed over these transformed values. Notably the normalization option is operationally realistic only if each probe is processed independently of all others. Certain unjustified performance gains may be realized if normalization were allowed across probes.

Two classes of normalization functions were allowed in FRVT 2002. The first is simply $\mathbf{t} = f_1(\mathbf{s})$. The second also takes a matrix of similarities, S_{GG}, between all gallery pairs, whose off-diagonal elements contain information available to an operational system. The second form of normalization, then, is simply

$\mathbf{t} = f_2(\mathbf{s}, S_{GG})$. Notably, the use of this second form is impractical for even moderate gallery sizes because S_{GG} imparts $\mathrm{O}(N^2)$ resource constraints. FRVT 2002 participants were given the option of submitting different f_1 and f_2 functions for each of the recognition tasks: identification, verification and watch list, discussed in the next section. The functions were supplied to, and run by, NIST as functions in an object file.

It should be noted that the use of normalization causes all searches to become one-to-many operations. The verification task, detailed below, is usually considered a one-to-one search and is correspondingly efficient, but if normalization is to be used then a gallery, of some size and content, must be constructed. If the gallery is changed, normalization must be recomputed.

2 Performance Metrics

In FRVT 2002, we evaluate an algorithm on three related tasks: identification, verification, and watch list, and we state separate appropriate statistics for each. As described above, performance on each of these tasks is obtained solely from the similarity values extracted from the similarity matrix and from the subject identities. In a proctored test such as FRVT 2002, identity information is not available to the recognition systems.

The watch list problem is a generalization of both identification and verification. For this task, a probe p is compared to a gallery which we term the watch list. The probe ranks the gallery, so we state performance as an identification rate. However a significant operational constraint is that a system should not attempt identification of individuals not on the watch list. We must also, therefore, measure a false alarm rate. This makes clear that the generalized watch list problem is defined over an open-universe.

In the next three subsections we formalize the watch list problem and show that verification is the special case when the watch list size is 1, and identification conversely is a closed universe watch list task.

2.1 Watch List

We measure the watch list performance using a watch list $\mathcal{G}$ and two probe sets: $\mathcal{P}_{\mathcal{G}}$ with subjects who should be identified and, $\mathcal{P}_{\mathcal{N}}$ with true imposters who should not throw an alarm. The former is used to state the detection and identification rate equal as the fraction of probes in $\mathcal{P}_{\mathcal{G}}$ that are detected at or above threshold t *and* recognized at rank r or better:

$$P_{DI}(t, r) = \frac{|\{p_j : \mathrm{rank}(p_j) \leq r,\ s_{ij} \geq t, \mathrm{id}(p_j) = \mathrm{id}(g_i)\}|}{|\mathcal{P}_{\mathcal{G}}|} \quad \forall p_j \in \mathcal{P}_{\mathcal{G}} \qquad (1)$$

where the rank is defined as the number of watch list entries which have greater than or equal similarity to the probe than the matching entry:

$$\text{rank}(p_j) = |\{g_k : s_{kj} \geq s_{ij}, \;\; \text{id}(g_i) = \text{id}(p_j)\}| \quad \forall g_k \in \mathcal{G}. \tag{2}$$

Throughout we use i and k to subscript gallery elements, corresponding to rows of the similarity matrix, and j to subscript the probe sets corresponding to columns of the matrix. In practice, this needs to be modified to handle tied similarity values: we elected to use the mean of the lower and upper ranks of the run of tied values:

$$\begin{aligned} 2\,\text{rank}(p_j) = \;& |\{k : s_{kj} \geq s_{ij}, \;\; \text{id}(g_i) = \text{id}(p_j)\}| + \\ & |\{k : s_{kj} > s_{ij}, \;\; \text{id}(g_i) = \text{id}(p_j)\}| + 1 \end{aligned} \tag{3}$$

The imposter set is used to compute the false alarm rate as the fraction of probes from $\mathcal{P}_\mathcal{N}$ whose similarity to *any* gallery image is at or above threshold:

$$P_{FA}(t) = \frac{|\{p_j : \max_i s_{ij} \geq t\}|}{|\mathcal{P}_\mathcal{N}|} \quad \forall p_j \in \mathcal{P}_\mathcal{N} \quad \forall g_i \in \mathcal{G} \tag{4}$$

2.2 Identification

Identification is a special case of the watch list task: If we mandate a closed universe, then the false alarm rate is undefined and a pure identification rate specifies performance. Formally for each probe p from $\mathcal{P}_\mathcal{G}$ we sort the similarity scores against gallery $\mathcal{G}$, and obtain the rank of the match. Identification performance is then stated as the fraction of probes whose gallery match is at rank r or lower. If the set of probes with a close match is

$$C(r) = \{p_j : \text{rank}(p_j) \leq r\} \quad \forall p_j \in \mathcal{P}_\mathcal{G} \tag{5}$$

where the rank is defined as before. We now define the Cumulative Match Characteristic (CMC) to be the identification rate as a function of r:

$$P_I(r) = \frac{|C(r)|}{|\mathcal{P}_\mathcal{G}|} \tag{6}$$

which we plot as the primary measure of identification performance. It gives an estimate of the rate at which probe images will be classified at rank r or better. It is a non-decreasing function of r. Although the CMC is most often summarized with rank one performance, other points and the steepness of the curve are also relevant operationally. One drawback of the characteristic is its dependence on gallery size, $|\mathcal{G}|$. For this reason we also plot identification performance at fixed rank as a function of the gallery size.

2.3 Verification

The use of biometric systems for the verification task is perhaps more common than identification. The operational model assumes that a probe p_j is compared with its matching gallery image and that the single similarity score is compared

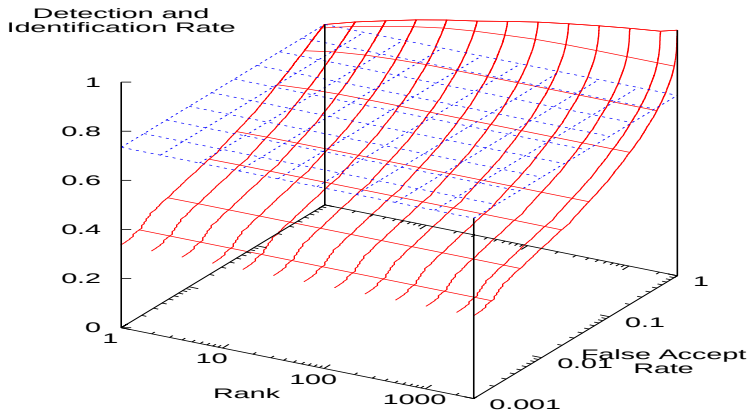

Fig. 1. Identification rate as a function of rank and false alarm rate for a watch list size of 3000. Note the weak dependence on rank, except at high false accept rates. The horizontal plane gives the rank one closed-universe identification rate.

against a threshold to verify the individual or otherwise. Two types of error can occur in this process: first a false accept in which an imposter claims an identity and is matched by the system above threshold; and secondly a false reject in which the system incorrectly matches the individual below threshold. This method for accepting or rejecting a claim models the Neyman-Pearson observer. A Neyman-Pearson model maximizes the verification rate for a constant false accept rate[1].

The Receiver Operating Characteristic (ROC) is computed to quantify verification performance. It shows the tradeoff between the two types of error by plotting estimates of the verification rate, P_V (i.e. the true accept rate) against the false accept rate, P_{FA} as a parametric function of the prior operating threshold, t. The verification rate is the fraction of probes whose gallery match has similarity greater than or equal to the threshold value, t:

$$P_V(t) = \frac{|\{p_j : s_{ij} \geq t, \ \ \mathrm{id}(g_i) = \mathrm{id}(p_j)\}|}{|P_G|} \quad \forall p_j \in \mathcal{P}_\mathcal{G} \tag{7}$$

The false accept rate is computed over the set of imposters, $\mathcal{P}_\mathcal{N}$, containing those individuals who are not present in the gallery:

$$P_{FA}(t) = \frac{|\{s_{ij} : s_{ij} \geq t\}|}{|\mathcal{P}_\mathcal{N}|\,|\mathcal{G}|} \quad \forall p_j \in \mathcal{P}_\mathcal{N}, \tag{8}$$

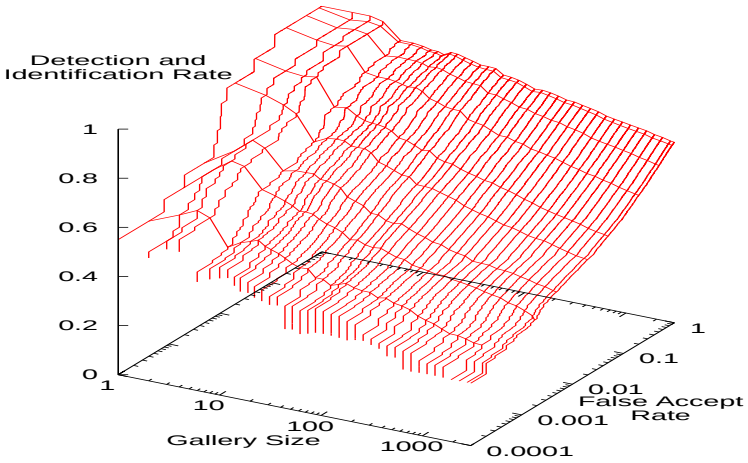

Fig. 2. Watch list performance. The rank one detection and identification rate is plotted as a function of gallery size and false alarm rate.

where the denominator shows our use of all the non-matching similarities in order to improve the fidelity of our P_{FA} estimate. In reality an imposter would usually claim just one identity, but our method of using all scores is realistic unless an imposter has a known prior resemblance to a person.

Note that this definition differs from that for verification false accept rate. Here we determine if there exists at least one gallery image more similar to the imposter than t. This occurs more frequently than in verification where the number of scoes above t between an imposter and *all* galleries are counted.

The use of a true imposter set $\mathcal{P}_{\mathcal{N}}$ contrasts to "round-robin" evaluations, in which $\mathcal{P}_{\mathcal{G}}$ and $\mathcal{P}_{\mathcal{N}}$ are the same set. From an operational standpoint, this models the case where a subject, already with legitimate access to the system (they are in $\mathcal{P}_{\mathcal{G}}$), attempts to gain access to the very same system, under a different identity. There may be some specialized scenarios where this is a valid model. However, we prefer to model the situation in which a person who does not already have access to the system attempts verification. In this model, the persons (not just the images) in $\mathcal{P}_{\mathcal{N}}$ are different from those in the gallery. The rationale for using true imposters is that the non-match distributions estimated from S_{GP_N} and S_{GP_G} may be different.

An empirical ROC is not a curve but a sequence of steps corresponding to a set of operating points where P_V and/or P_{FA} changes. Because systems are operated on the ROC's convex hull which corresponding to P_V changes, we use the $|\mathcal{P}_{\mathcal{G}}|$ match values as thresholds. We sort these values, keep their unique subset as t_i, $i = 1 \ldots |\mathcal{P}_{\mathcal{G}}|$ and insert the artificial threshold $t_0 = -\infty$. This

avoids the use of a much large number of non-matches $|\mathcal{G}||\mathcal{P}_\mathcal{N}|$ in the ROC calculation. The computation of P_V proceeds by binning the number of match values from S_{GP_G} that are in the range $[t_i, t_{i+1})$, and finally by accumulating all of them. The same thresholds and procedure are used for P_{FA} from S_{GP_N}.

2.4 Limiting Cases

Figures 1 and 2 show watch list performance as functions of watch list sizes, false alarm rates and rank. They were generated from the similarity scores of a leading FRVT 2002 participant. For galleries of size 1 the watch list measures reduce to verification; the intersection of the surface and the (1,y,z) plane is the ROC curve. The CMC curve is the plane (x,1,z) of figure 1 and similarly identification performance as a function of gallery size is given by the intersection of the surface and the (x,1,z) plane, as shown in figure 2. These two cases correspond to the relaxation of the threshold, $t \to -\infty$ whence $P_{FA} \to 1$ then the detection and identification rate becomes the identification rate, i.e. $P_{DI}(1, r) = P_I(r)$.

2.5 Comparing Verification Performances

Frequently we seek to compare verification results, either between systems, or across different data sets processed by the same system. Between systems this may be accomplished by looking at the verification rate at fixed false accept rates. This is appropriate because it is not possible, nor operationally feasible, to set a uniform threshold across different systems that report on different ranges and scales. However, for studies using just one system and many gallery and probe sets, fixed P_{FA} values correspond to different thresholds, which are not operationally realizable. The correct approach is to set a single threshold t and acknowledge that the $P_V(t)$ and $P_{FA}(t)$ are random variables across experiments.

Thus, to be able to compute variation in verification and false accept rates across multiple galleries and probe sets, we must compute the ROC using a fixed global set of thresholds. Formally, if R experiments use R different $\mathcal{G}$, $\mathcal{P}_\mathcal{G}$ and $\mathcal{P}_\mathcal{N}$ image sets, we extract the sorted union of R sets of match scores. These thresholds will generally "oversample" each individual ROC. Thus we have R (P_V, P_{FA}) pairs at each threshold, which we plot with an error ellipse that traces two standard errors in the P_V and P_{FA} dimensions. The principal axes of this ellipse are the eigenvectors of the covariance matrix of the R pairs. This is shown in figure 3.

As a summary statistic we usually report the P_V and P_{FA} at a threshold that gives $P_{FA} \approx 0.01$. We favor this over equal error rate (when $P_{FA}(t) = 1 - P_V(t)$) because $P_{FA} = 0.01$ is an operationally realistic number, while equal error rate not only varies, but is usually higher.

3 Conclusion

The paper details the evaluation framework and scoring metrics used in the Face Recognition Vendor Test 2002. We have outlined the concept of operational re-

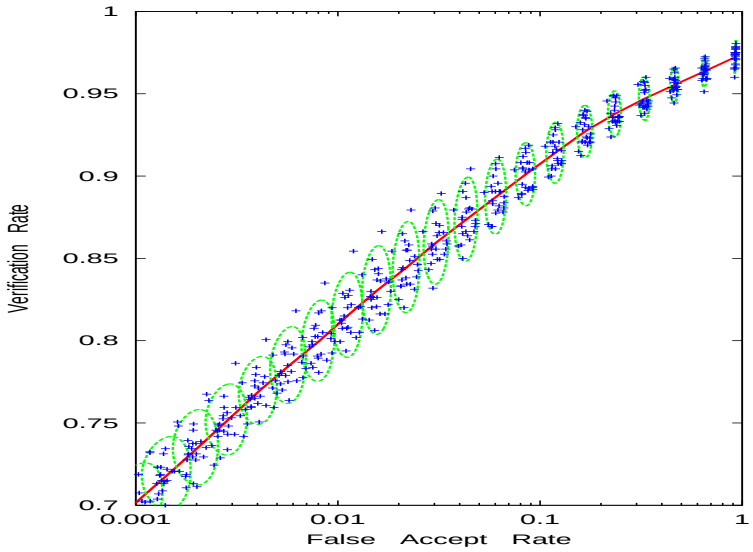

Fig. 3. Verification variation. The figure shows the results of computing ROC curves from 23 disjoint populations of size 800.

alizability based on the real-world usage of systems with fixed operating thresholds. We show how variation in verification performance must be computed at the same threshold. Motivated by operational relevance we have defined true imposters and shown via the non-match distribution that their use is necessary in evaluations. We define the watch-list scenario and show that verification and identification are special cases of it.

References

1. K. Fukunaga. *Statistical Pattern Recognition*, chapter 3. Academic Press, second edition, 1990.
2. R. J. Micheals, P. Grother, and P.J. Phillips. The NIST Human ID Evaluation Framework. In *Proceedings of the Fourth International Conference on Audio- and Video-based Biometric Person Authentication*, June 2003.
3. P. J. Phillips, H. Moon, S. A. Rizvi, and P. J. Rauss. The FERET Evaluation Methodology for Face Recognition Algorithms. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 22:1090–1104, 2000.
4. P.J. Phillips, P. Grother, R. J. Micheals, D. M. Blackburn, E. Tabassi, and M. Bone. Face Recognition Vendor Test 2002. Evaluation Report IR 6965, National Institute of Standards and Technology, www.itl.nist.gov/iad/894.03/face/face.html, March 2003.
5. J. L. Wayman. Error-Rate Equations for the General Biometric System. *IEEE Robotics and Automation Magazine*, pages 35–48, 3 1999.