Face Recognition Vendor Test 2002

Supplemental Report

NISTIR 7083

Patrick Grother

February 2, 2004

AUTHENTICATED
U.S. GOVERNMENT
INFORMATION
GPO

Highlights

1.	Fusion of four scores from five images per person cuts verification error rates by about half,	[Section 6.3].
2.	Fusion of two leading FRVT systems also reduces verification errors by about 50%,	[Section 6.4].
3.	In combination, fusion of scores from additional images *and* from three systems gives 99.7% true match rate at 1% false match rate, a ten-fold reduction in false non-match rate over the comparable single score result from the best system.	[Section 6.7].
4.	The Chinese sub-population of the FRVT corpus is markedly easier to recognize,	[Section 3.1].
5.	Performance varies little for eye-to-eye distances on the range 40 to 100 pixels, but does degrade with discrepancy between enrolled and test images,	[Section 4.1].
6.	Although performance varies little with JPEG file size, (from 6k to 11k, image size 300 x 252 pixels), there is considerable degradation with compression-ratio computed *on the face*,	[Section 4.2].
7.	The median inter-pupil FRVT image distance, 73 pixels, exceeds the minimum resolution of the *token* image in the draft face recognition interchange format. In addition the on-face compression ratio is 18 which is lower than the allowed maximum of 20,	[Section 4.2].
8.	Performance is insensitive to in-plane head rotation up to ± 12 degrees, but degrades mildly with discrepancy,	[Section 4.3].
9.	Normalization is a legitimate means of boosting verification and open-set identification performance, by using background images. It's action is equivalent to setting user-specific thresholds,	[Section 5.3].
10.	Normalization is efficient, achieving its full effectiveness on background galleries as small 30,	[Section 5.5].
11.	z-Normalization is a simple, effective normalizer proven in speech, fingerprint and gait tests. For face, its performance approaches that of the vendors' normalizers,	[Section 5.2].
12.	The *perfect* impostor is defined and used to model worst case security performance,	[Section 3.4].
13.	Generalized rocs are defined and used as the canonical means of displaying open-set 1:N identification performance, for $N \geq 1$,	[Section 2.2].
14.	In closed-set 1:N identification, the rate at which a user's match is found in the top p % of the enrolled population is largely independent of N,	[Section 2.3].

Note: The identification of any commercial product or trade name does not imply endorsement or recommendation by the National Institute of Standards and Technology, or any other FRVT 2002 sponsor or supporter.

Contents

1 Introduction **4**

2 FRVT 2002, Definitions, Metrics and Datasets **4**

2.1 Definitions . . . 4

2.2 Metrics . . . 5

2.2.1 Autonomous Operation . . . 5

2.2.2 Degenerate Cases . . . 7

2.3 Graphical Statements of Performance . . . 8

2.4 Metrics in Deployed Systems . . . 8

3 Subject-specific Effects **8**

3.1 Ethnicity . . . 8

3.2 Impostor Models . . . 11

3.3 Zero-effort Impostors . . . 11

3.4 Perfect Impostors . . . 11

3.4.1 Relation to Watch List Metric . . . 14

3.4.2 Extreme Non-Match Values . . . 14

4 Image-specific Effects **14**

4.1 Resolution . . . 14

4.2 Compression Ratio . . . 16

4.3 In-plane Rotation . . . 16

4.4 Mixed Effects . . . 16

5 Normalization **16**

5.1 Overview . . . 19

5.2 z-Normalization . . . 19

5.3 Operational Legality . . . 19

5.4 Normalization in FRVT 2002 . . . 20

5.5 Efficiency . . . 20

6 Fusion **22**

6.1 Score Level Fusion . . . 22

6.2 Scaling . . . 23

6.2.1 Learned Scaling . . . 23

6.3 Multi-image Fusion . . . 23

6.3.1 Effect of Elapsed Time . . . 28

6.4 Multi-system Fusion . . . 28

6.5 Weighted Fusion . . . 28

6.6 Three System Fusion . . . 30

6.7 Multi-system and Multi-image Fusion . . . 30

7 Correction **31**

8 Acknowledgements **31**

1 Introduction

This report[1] is intended to supplement the primary FRVT 2002 reports and to be relevant to more specialized audiences, particularly standards makers, policy makers, and developers in face recognition and other biometric fields including fusion. In addition some aspects have been included in response to the many comments received since the primary report was published in March 2003. Although the author acknowledges that the utility of performance numbers themselves is being eroded by technologocial advances through time, the report intends to highlight other relevant issues. For instance, certain elements of best practice evaluation methodology are advocated. In addition the material here is just a small subset of the analyses that could be conducted using the archived FRVT 2002 data, (for example, in the areas of database segmentation, independence, the biometric zoo) but equally there are gaps that can only be filled by further tests (for example, global ethnic variations, persons under the age of 18, image quality). Indeed several results in this paper are included to spur further investigation. This is particularly true on the statistical side, the assumption being that progress on the technology side is adequately active already.

The contents of the report is broadly categorized into four areas: subject-specific effects, image-specific effects, normalization, and fusion.

2 FRVT 2002, Definitions, Metrics and Datasets

This section summarizes the FRVT definitions and performance metrics.

2.1 Definitions

Several terms are used throughout this paper:

1. *FRVT 2002:* The *Face Recognition Vendor Test 2002*, conducted in July and August 2002, presented images to vendors who ran their systems and returned results back to NIST for analysis.
2. *FRVT Report:* Shorthand here for *Face Recognition Vendor Test 2002, Evaluation Report* [11] published in March 2003, available from `http://www.frvt.org`.
3. *Vendor:* Any of the companies participating in FRVT 2002; the terms algorithm and recognition system refer to the company's face recognition system.
4. *High Computational Intensity Test:* Part of the FRVT study that utilized 121589 images from 37437 persons, as extracted from an operational U.S. Department of State Non-immmigrant visa database collected in Mexico.
5. *Legitimate user:* A person previously enrolled in a biometric system.
6. *Gallery:* The imagery from a set of legitimate users who are enrolled in a system. There is always one entry per individual in the gallery.
7. *Impostor:* A person not known to a recognition system.
8. *Verification attempt:* The recognition of a legitimate user or an impostor image with the enrolled image of the person they claim to be.
9. *Identification attempt:* The presentation of an unknown sample for comparison with all enrolled persons.
10. *Probe:* Generic term for the imagery used in a verification or identification attempt.
11. *Match:* The entry in the gallery belonging to the same individual as a probe. As an adjective it implies "from the same individual".
12. *Fixed Threshold:* A value representing the similarity threshold above which users and impostors are accepted. It is conceptually fixed for all users for the life of a deployed system.

[1] **Keywords:** *Face Recognition, Biometrics, Performance Evaluation, Fusion, Identification, Negative identification, Compression, Resolution, Vulnerability Analysis, Demographics*

System	Verification		Identification			
	P_V at $P_{FA} = 0.01$		P_I at Rank 1			
	Small Gallery		Small Gallery		Large Gallery	
	Chinese	Worldwide	Chinese	Worldwide	Chinese	Worldwide
Cognitec	0.93	0.90	0.92	0.86	0.91	0.73
C-vis	0.72	0.65	0.52	0.48	0.44	0.26
Dreammirh	0.39	0.33	0.29	0.25	0.23	0.11
Eyematic	0.91	0.86	0.84	0.79	0.81	0.65
Identix	0.94	0.90	0.91	0.83	0.88	0.70
Imagis	0.66	0.59	0.65	0.57	0.62	0.40
Visage	0.60	0.67	0.47	0.51	0.41	0.31
Visionsphere	0.67	0.53	0.61	0.46	0.58	0.29

Table 1: Comparison of identification and verification performance on the Chinese population vs. the general FRVT population.

2.2 Metrics

The formal description of the metrics used in FRVT 2002 [5], defines *open-set* 1:N identification as the most general biometric task and poses *closed-set* identification and *verification* as special cases of it. Performance on the open-set problem should be quantified over two populations. First the impostors, those persons who are not present in the gallery, are used to compute the *false match rate*, P_{FA}, which is needed to quantify rejection capability:

$$P_{FA}(N,t) = \frac{\text{Number of impostors matching } \textit{any} \text{ of } N \text{ enrollees above threshold } t}{\text{Number of impostors}} \tag{1}$$

Second, for those persons who are "known" (i.e. previously enrolled) to a system, the open-set *identification rate*, P_{TA}, is used to quantify user performance:

$$P_{TA}(N,t) = \frac{\text{Number of known users whose match is above threshold } t}{\text{Number of known users}} \tag{2}$$

A test quantifies performance using operating characteristics computed solely from all similarities, s_{ij}, corresponding to the comparison of images i and j. Formally, the *false match rate* is the fraction of probes, p_j, from an imposter set, $\mathcal{P}_{\mathcal{N}}$, who are more similar than t to *any* (i.e. one or more) gallery entries, g_i from gallery, $\mathcal{G}$:

$$P_{FA}(t) = \frac{|\{p_j : \max_i s_{ij} \geq t\}|}{|\mathcal{P}_{\mathcal{N}}|} \quad \forall p_j \in \mathcal{P}_{\mathcal{N}} \quad \forall g_i \in \mathcal{G} \tag{3}$$

The definition of *identification rate* is the fraction of probe images p_j from the user set, $\mathcal{P}_{\mathcal{G}}$, whose matching gallery image, g_i, from enrolled population $\mathcal{G}$, has similarity score at or above an operating threshold, t:

$$P_{TA}(t) = \frac{|\{p_j : s_{ij} \geq t,\ \text{id}(p_j) = \text{id}(g_i)\}|}{|\mathcal{P}_{\mathcal{G}}|} \tag{4}$$

where the operator $id()$ represents the identity of the one person appearing in the image. Note that i and k subscript gallery elements (rows) of the similarity matrix, and j indexes probes (columns).

2.2.1 Autonomous Operation

Rank based criteria may be included in the true match rate definition. This is particularly appropriate in situations where a system is unattended, that is with automated, not manual, secondary processing.

$$P_{TA}(N,t,k) = \frac{\text{Number of known users whose match is above threshold } t \text{ and at rank } k \text{ or better}}{\text{Number of known users}} \tag{5}$$

The most common application is the case $k = 1$ representing a single forced choice, best guess, decision. In any case the formal definition is:

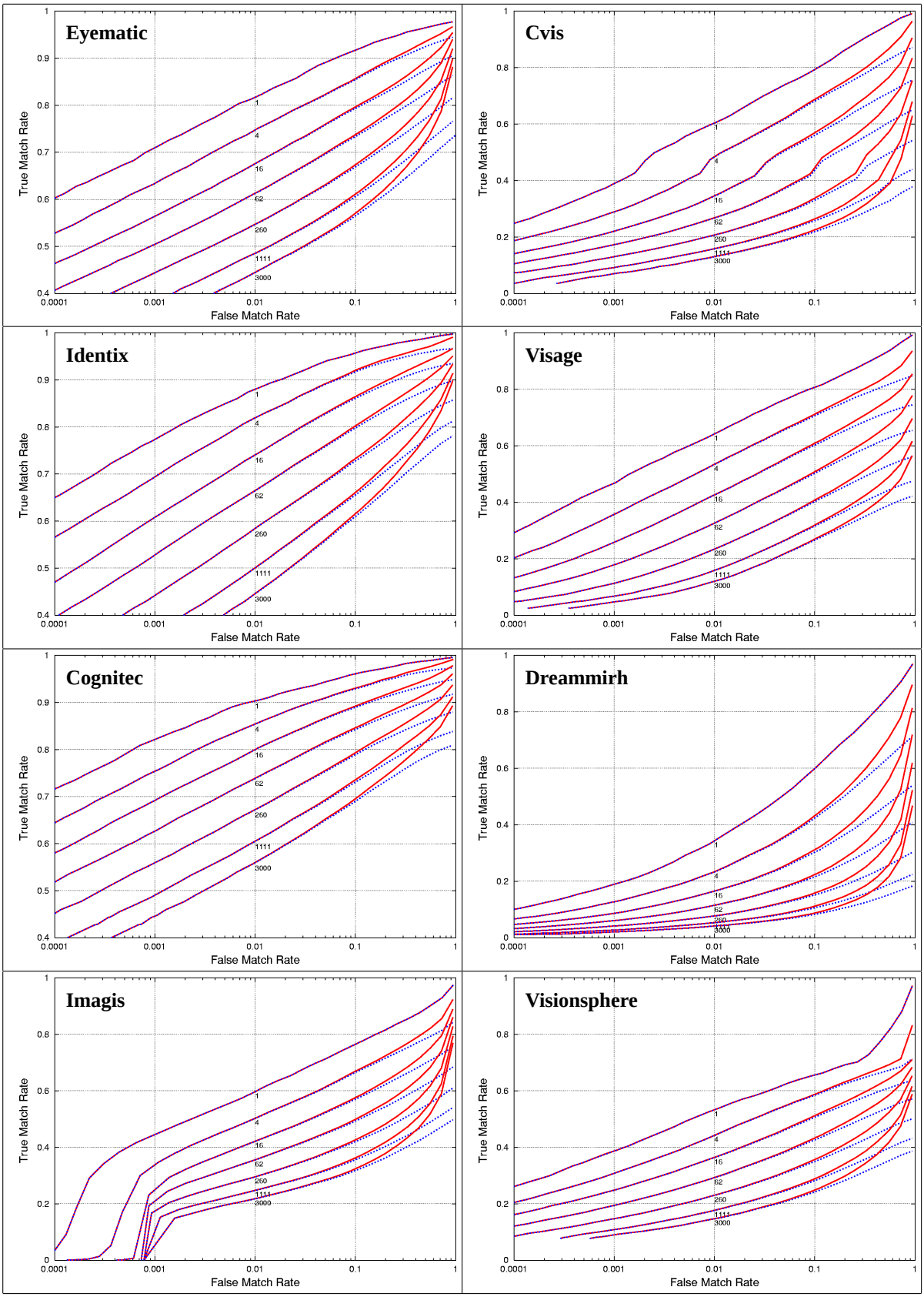


Figure 1: Open-set 1:N identification performance. Each generalized ROCs shows the tradeoff between true and false match rates for various enrolled population sizes, N. The top trace gives the degenerate 1:1 verification case. The upper (red) trace ignores rank (eqs. 1 and 2), and the lower (blue) trace gives true matching at rank 1 (eq. 5). Normalization (sec. 5) was not used. Note the differing y-axis scales.

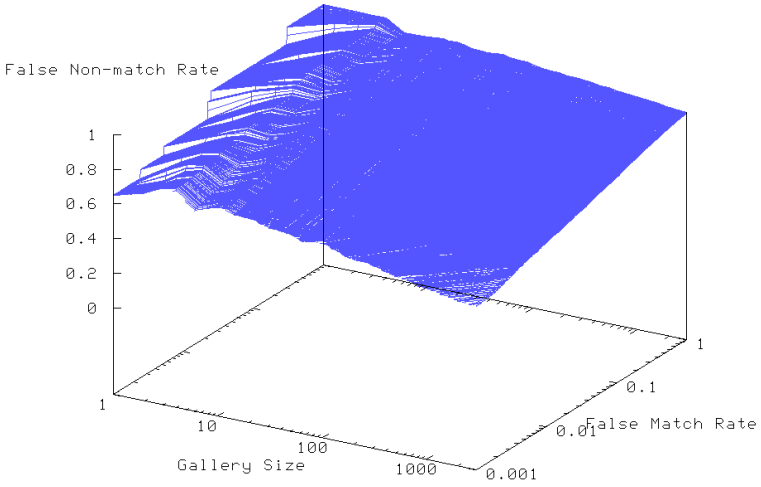


Figure 2: Open-set identification performance plotted as generalized ROC, for the Cognitec system. The axes plot identification rate, versus false match rate, and enrolled population size.

$$P_{TA}(k,t) = \frac{|\{p_j : R(p_j) \leq k,\ s_{ij} \geq t,\ \text{id}(p_j) = \text{id}(g_i)\}|}{|\mathcal{P}_\mathcal{G}|} \tag{6}$$

where the rank R (for non-tied scores) is the number of gallery images, including the match, that have similarity greater than or equal to the match:

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

In the case of $N = 1$ rank is moot. Note that i and k subscript gallery elements (rows) of the similarity matrix, and j indexes probes (columns).

2.2.2 Degenerate Cases

The 1:N open-set task is the general case. In FRVT 2002 it was referred to as the watch-list problem and was presented after the two more commonly quoted special cases:

1:1 Verification This is the situation in which a user claims one particular identity and it corresponds to open-set identification in a gallery size of one [5]. Using $|\mathcal{G}| = 1$ in (6) and (3) causes the rank inequality to disappear and the equations give simply the fractions of matches and non-matches above threshold.

> *Many of the results in this paper quantify performance by quoting the true match rate and false match rate against some variable of interest (resolution, compression, etc.). This is done at* **a fixed threshold** *chosen to give a false match rate of 5% (unless noted otherwise) near the middle of the range of the variable of interest. The use of fixed thresholds is a necessary requirement of operation; particularly it is not sufficient to quote, for example, just P_{TA} at fixed P_{FA} because those true match rates would be realized at different thresholds for different, say, compression ratios.*
> *One-to-one verification is used in this role for two reasons. First it is relevant to access control applications, and it is a well understood measure of biometric strength, that is a measure of a system's capability.*

1:N Closed-set Identification On the other hand, closed set identification is the 1:N comparison of an image whose match is *known* to be in the gallery. This is a special case that renders false match moot, which can be achieved by setting the threshold $t = -\infty$ in (3). Closed-set identification is only occasionally applicable to real-world applications, but is notable as an academic construct because prior probabilities are not subsequently required.

2.3 Graphical Statements of Performance

Open-set Identification The receiver operating characteristic (ROC) is the most common and important statement of biometric performance. It depicts the trade-off between false and true match rates by plotting them parametrically with threshold t. It thereby cleanly devolves recognition performance from policy decisions involving priors and costs.

> *This document advocates the use of the ROC as the canonical means of displaying open-set identification performance for arbitrary gallery sizes. Both 1:1 verification* **and** *1:N identification performance are plotted on the same axes. Figure 1 presents such ROCs for eight FRVT systems on open-set populations of size $N = 1, \ldots, 3000$, for both the rankless (eqs. 3 and 4) and rank one cases (eqs. 3 and 5).*

Figure 1 is a more compact presentation of the data presented in Figure 2, which includes its ROCs as vertical slices of the 3D surface obtained by plotting identification rate as a function of both false match rate and populations size.

Closed-set Identification Figure 1 shows that the inclusion of a rank condition practically has an effect only for larger false match rates, i.e. those that are usually of less operational relevance. In the closed-set limit $P_{FA} = 1$ the rank effect is significant. Without it $P_{TA} = 1$ at $P_{FA} = 1$.

The primary FRVT report plotted closed-set identification rates in two ways, first as a function of rank for a specific gallery size giving *cumulative match characteristics*, and second as a function of gallery size for rank one. In this report we modify the latter by plotting the identification rate versus gallery size for various *percentage ranks*: Figure 3 shows the expected fraction of users that are recognized within the top $k = pN/100$ ranks, for various percentages, p, as a function of N. That the identification rate is approximately constant for large N is an indication that the 1:N closed-set model [6] is reasonable, at least for large gallery sizes. The model is

$$P_I(N, k) = 1 - M\left(F^{-1}\left(1 - \frac{k}{N}\right)\right) \tag{7}$$

where M is the cumulative distribution function (CDF) of the match scores and F^{-1} is the inverse CDF of the non-match scores. The CDFs are rarely known and so must be estimated over some (large) population.

2.4 Metrics in Deployed Systems

Operational installations of biometric identification systems usually report just the closest few gallery members. Indeed it is common in AFIS systems to return *candidate lists* that contain zero or more entries. The logic used in the construction of a candidate list is generally proprietary and unknown, and may not even be fixed. Most often the list is a function of the scores returned in relation to a decision *threshold* and on some *rank* condition. For example a system might return no more than 10 gallery elements whose score to the user image is above 900. Nevertheless, performance in this paper is reported using equations (1) and (2), which are *attempt*-based rather than score-based, eqs. (3) and (4).

3 Subject-specific Effects

3.1 Ethnicity

Ethnic variation is clearly a potential source of performance variability for biometric technologies, particularly for visible face recognition. The effect of ethnicity on face recognition engines has not been documented publicly in any systematic and large scale way, so the Chinese result is included below to give one instance of an apparently significant population effect on recognition performance, and thereby to imply a need for testing on other populations.

The State Department imagery used in FRVT 2002 is annotated with demographic information, and while the primary FRVT report detailed the effects of sex and age on performance, it did not address dependence on ethnicity. Indeed

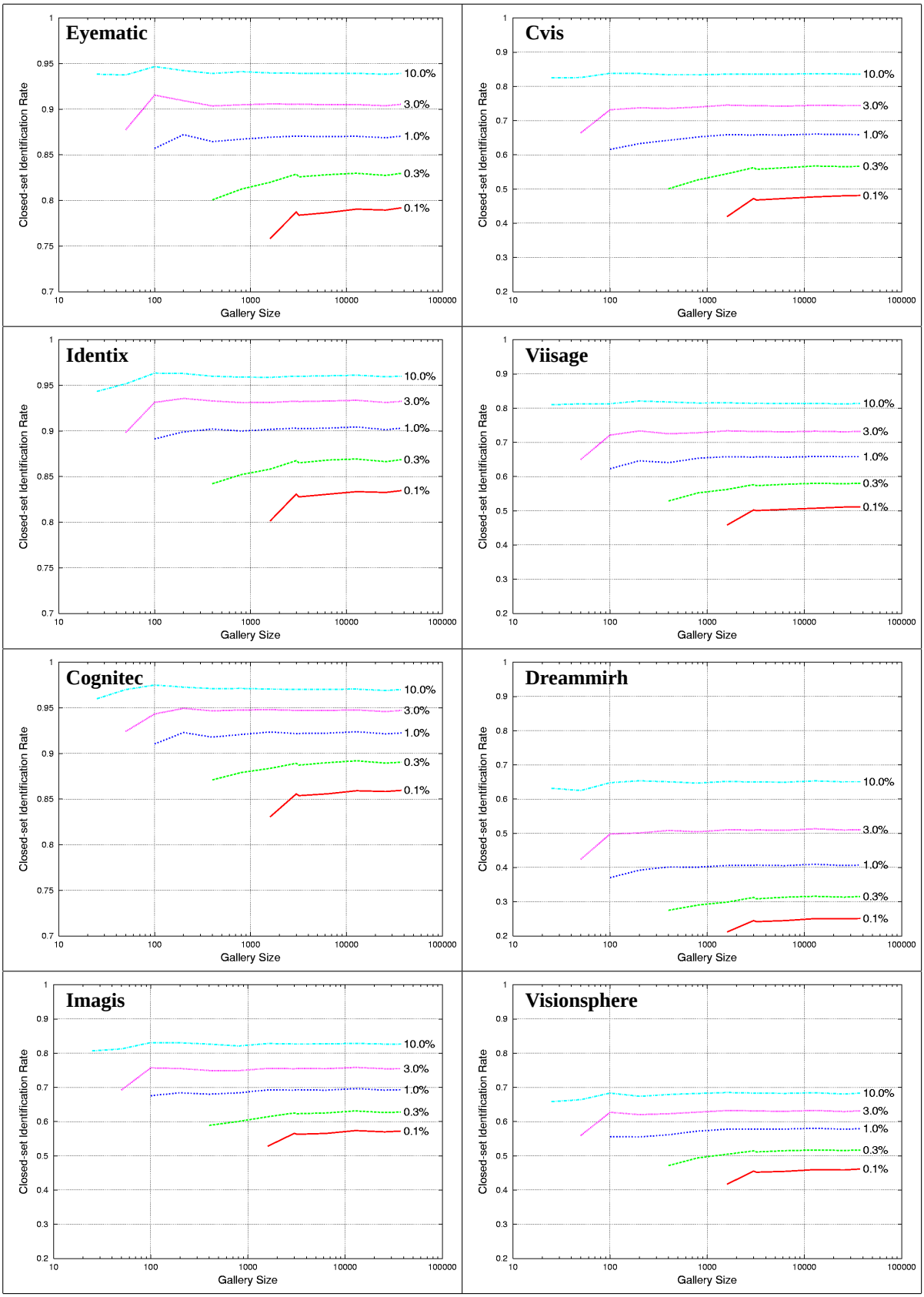


Figure 3: Closed-set identification performance versus gallery size, for various ranks expressed as a percentage of the gallery size.

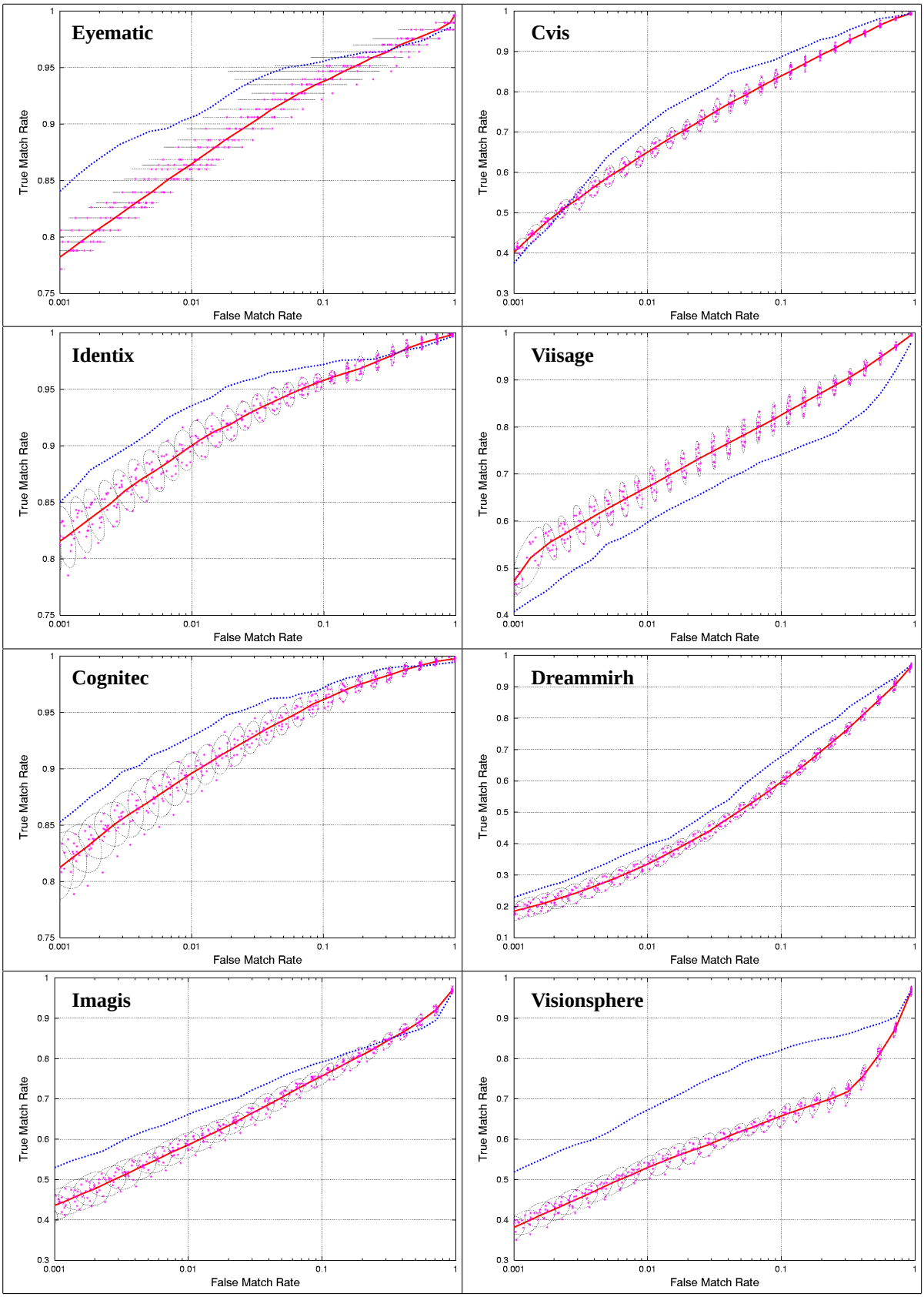


Figure 4: ROC curves for the general population (with ellipses) and for the Chinese sub-population. In all but one case (Viisage) the general population is more difficult to recognize than the Chinese sub-population. Normalization (sec. 5.4) was used.

because 98.4% of the imagery used is of persons whose birthplace is in Mexico a formal study of this effect was beyond the scope of the FRVT 2002 design. However we are able to report on the next largest ethnic component, namely the 639 Chinese individuals who sought visas in Mexico, and who were included in the FRVT image set.

Although birthplace information is only a proxy for ethnicity, manual inspection of the images confirmed that in this case birthplace does indeed imply east asian ethnicity. Such an association is generally suspect, especially in developed nations, but is reasonable here because of the low levels of immigration to China in recent decades. Nationality information was also available but, for obvious reasons, was ignored as a less reliable indicator of ethnicity.

To assess the recognizability of this group we compare a standard performance measure for the Chinese group with the population as a whole. The FRVT Evaluation Report gave, in Figure 8, verification performance for twelve disjoint populations of size 3000, using *error ellipses* to show the joint variation in the true-, and false-match rates, (P_{FA}, P_{TA}). Figure 4 uses the same approach to quantify population performance variation. It shows for each FRVT system the receiver operating characteristic for the Chinese population and also for the general population, the latter being computed over 12 disjoint populations of size 639 (i.e. the size of the available Chinese population) and plotted as their mean with their error ellipses. The overall result is that the Chinese population is substantially easier to recognize; this cannot be attributed to age nor sex[2] because the population is only three months older with a 56.8% males rather than the overall 50.5% used in FRVT.

The single point corresponding to 1% false match rate from the ROCs is repeated in columns 2 and 3 of Table 1. In addition the table gives results for 1:N closed-set identification for two galleries $N = 639$ (i.e. small) and $N = 37437$ (large). Note that the identification rate for the Chinese population declines very little (0.91 vs. 0.92 for Cognitec) compared to the general population (0.86 vs. 0.73). This is due to a *database segmentation* effect: when an individual from one class is used in a 1:N search of a gallery made up mainly of individuals of a different class, the recognition process is much easier because the gallery is *effectively* much smaller than its nominal size, in this case 37437. Thus columns 4-5, and 2-3, more appropriately contrast the two populations.

Database segmentation effects have been observed for sex and for age, and although these variables were addressed in the FRVT 2002 primary report, a more formal analysis of segmentation is warranted.

Performance of face recognition systems is dependent on the ethnicities of the human subjects. By conducting a larger study across many ethnicities, performance of a deployed face recognition system can be simulated if a specification of the local ethnic composition is available. The method should not be considered a replacement for good experimental design [9] however.

3.2 Impostor Models

One to one verification is arguably the archetypal biometric application, and is ubiquitously deployed as an access control mechanism. It is incumbent on recognition systems to correctly accept legitimate users, and without reconfiguration, and to reject impostors. This section discusses two means of measuring reject performance.

3.3 Zero-effort Impostors

Although false acceptance is a natural hazard in human populations classified with imperfect algorithms, the existence of an *active* impostor is a primary security concern. Active here implies that the impostor has taken steps to mimic a legitimate user's biometric, the obvious examples in face recognition being cosmetic, subdermal or surgical disguises. However performance of biometric systems faced with active impostors is difficult to test on a large scale. Instead most research efforts use randomly selected members of the available data as *zero-effort impostors*. This essentially quantifies the natural inter-personal variation in the human population, and is representative of real world performance only in as much as an algorithm capable of differentiating zero-effort impostors from legitimate users can repeat the feat for persons actively making an effort. The primart FRVT 2002 report considered only zero-effort impostors.

3.4 Perfect Impostors

Although the FRVT 2002 test was not designed to address active impostors, a worst case one-to-one false match performance can be computed by emulating the *perfect* impostor - i.e that impostor who most closely matches the claimed identity. Verification performance, stated as an ROC, is obtained by computing the non-match distribution

[2]The primary FRVT report shows evidence that older people and males are easier to recognize.

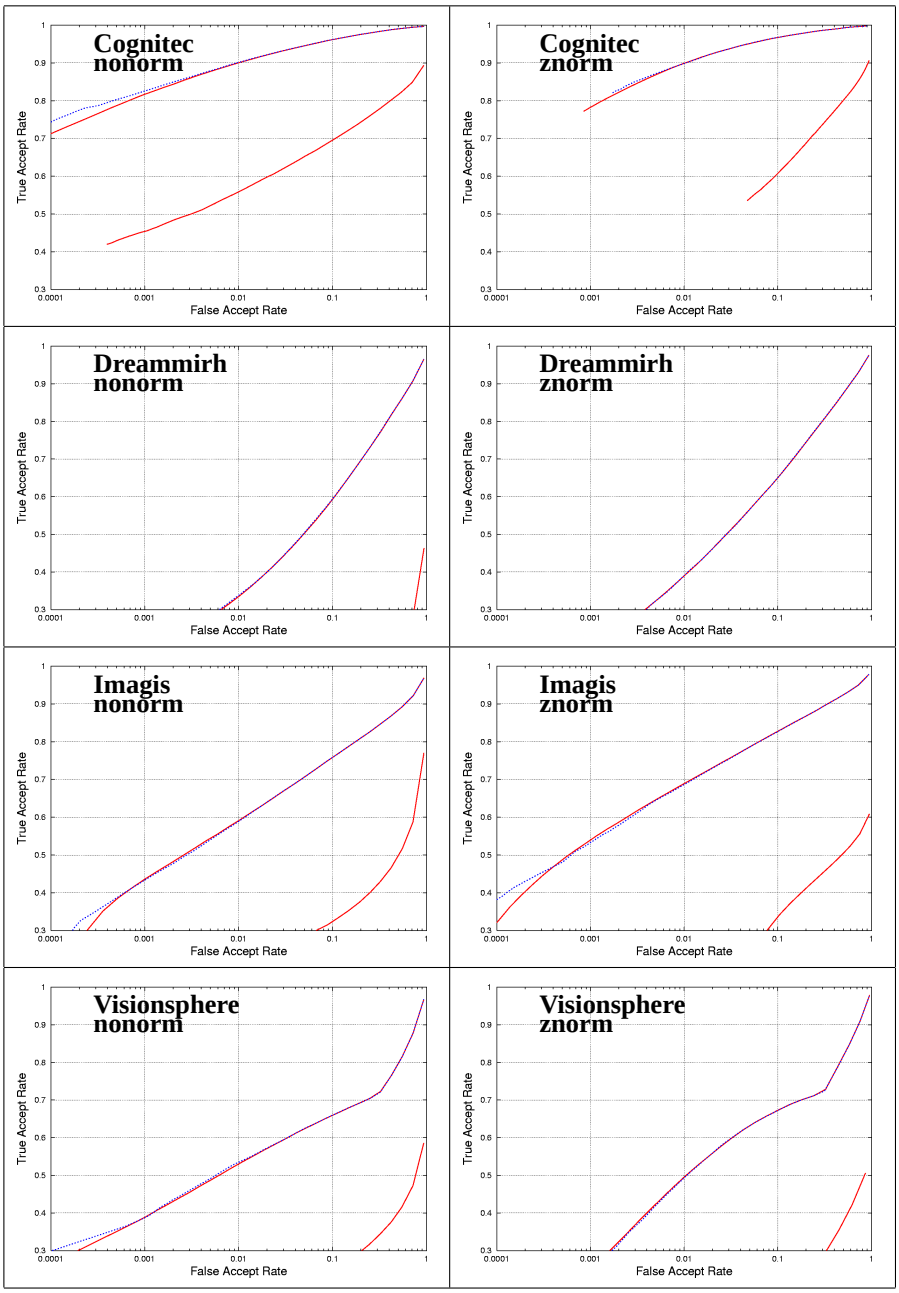


Figure 5: The effect on ROC of different non-match sampling schemes and the effect of normalization. Three plots are shown on each graph. The lower curve results from the worst case impostor, the other two, which largely overlap use all impostors and single randomly selected impostors. The systems shown provided no native normalization functionality.

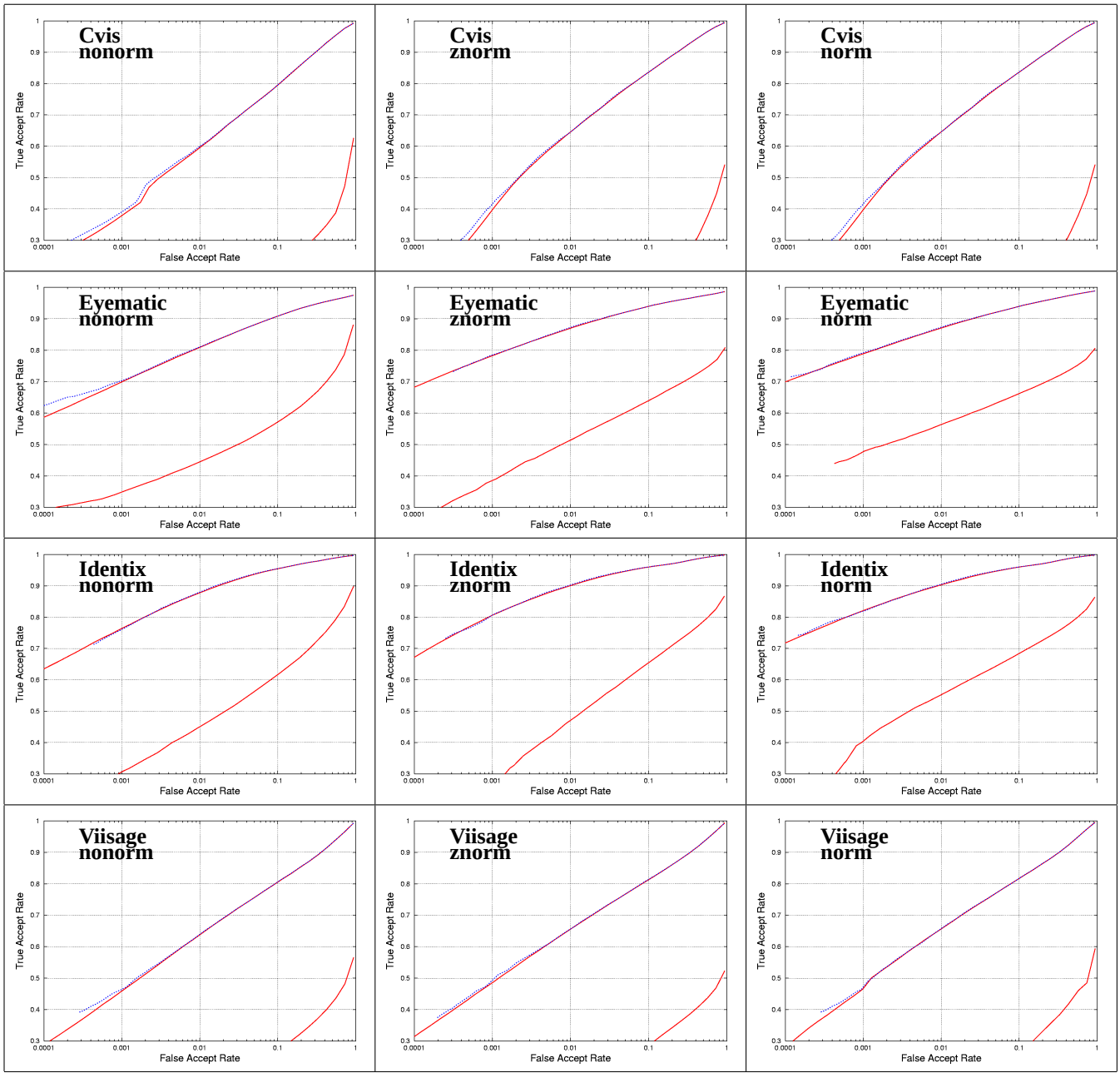


Figure 6: The effect on ROC of different non-match sampling schemes and the effect of normalization. Three plots are shown on each graph. The lower curve results from the worst case impostor, the other two, which largely overlap use all impostors and randomly selected impostors.

from the *top* non-match score. This represents a worst case scenario in which all impostors make optimal decisions about whom to pretend to be, or equivalently, who benefit from having someone with a resemblance in the database. This latter aspect is predicated on the algorithm in question, and it should be emphasized that the perfect impostor for one system may not be for another. Thus the perfect impostor ROCs that appear in Figures 5 and 6 are computed from non-match distributions that are themselves computed from generally not identical image sets. The match distributions are the same in all cases, but the impostors differ. The figures include the effect of normalization which is discussed in section 5.

Not shown in those figures is the effect of *incompetent* impostors who consistently claim to be precisely the wrong person. The probabilty of verification in such cases is essentially 1 for all false match rates.

3.4.1 Relation to Watch List Metric

Unfortunately the perfect impostor in 1:1 verification trials is not just a theoretical construct. The primary FRVT 2002 report included a lengthy treatment of the watch list application. This pertains to the situation in which a general population, passing through an airport concourse perhaps, are compared with a set of enrolled individuals, the persons on the watch list. This is perhaps the classic example of open-set negative identification, where the word negative is used to indicate that subjects claim implicitly *not* to be known to the system. The difficulty of the watch list task for large populations is that if a passer-by, a person not on the watch list, matches *any* of the watch list members above the system threshold an alarm is thrown. The computation of the false match rate is then precisely the same calculation (equation 3) as that for the perfect impostor shown above.

3.4.2 Extreme Non-Match Values

The non-match scores that define false match rates for the watch list are the extreme values of the non-match distribution. In practice if this extreme value distribution is estimated from similarity data, the sample mean will generally increase with the number of samples, i.e. the enrolled population size, considered. This arises because the expectation value of the maximum non-match score is a function of the number of samples drawn from the distribution. From order statistics the dependency on the population size is obtained by considering the expected value of the largest of G samples drawn from the non-match distribution.

$$E(x_{max}) = \int_{-\infty}^{\infty} GsN(s)^{G-1}n(s)ds \tag{8}$$

where $n(s)$ and $N(s)$ are, respectively, the density and the CDF of the non-match distribution.

4 Image-specific Effects

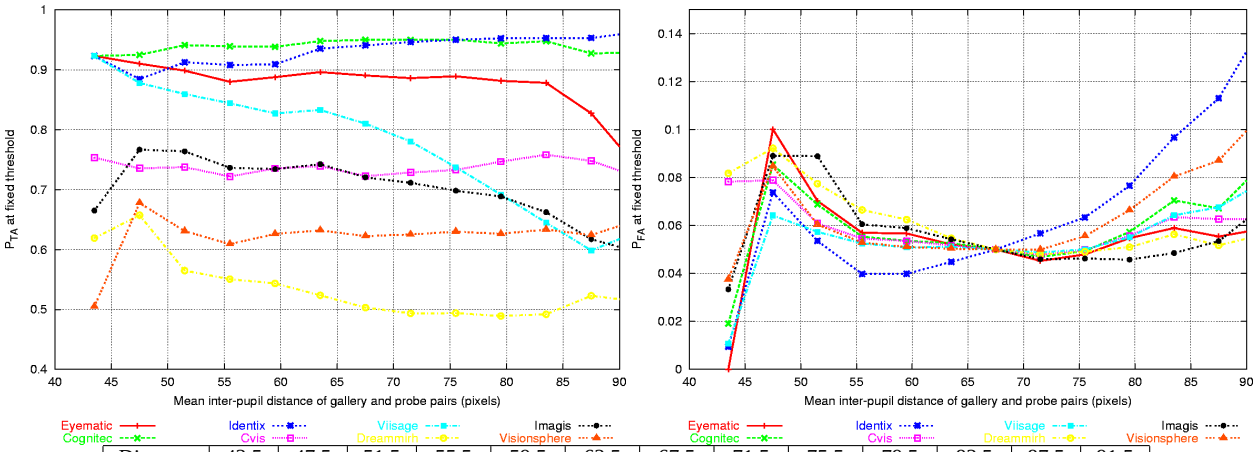
This section addresses the variation in face recognition performance with the image-specific quantities, compression ratio, resolution and head rotation. Because the recognition process involves both the user's probe image and the enrolled gallery image, performance is a function of whatever quantity is under investigation. The most general way of presenting the result is to plot performance as a function of both the gallery and probe's variable values on a 3D scatter plot. However in the sections that follow alternative, the two values are combined to make plots more readable.

4.1 Resolution

The amount of information available to a recognition engine is related to the size of the face in the image. This is usually quantified by the number of pixels between the pupils.

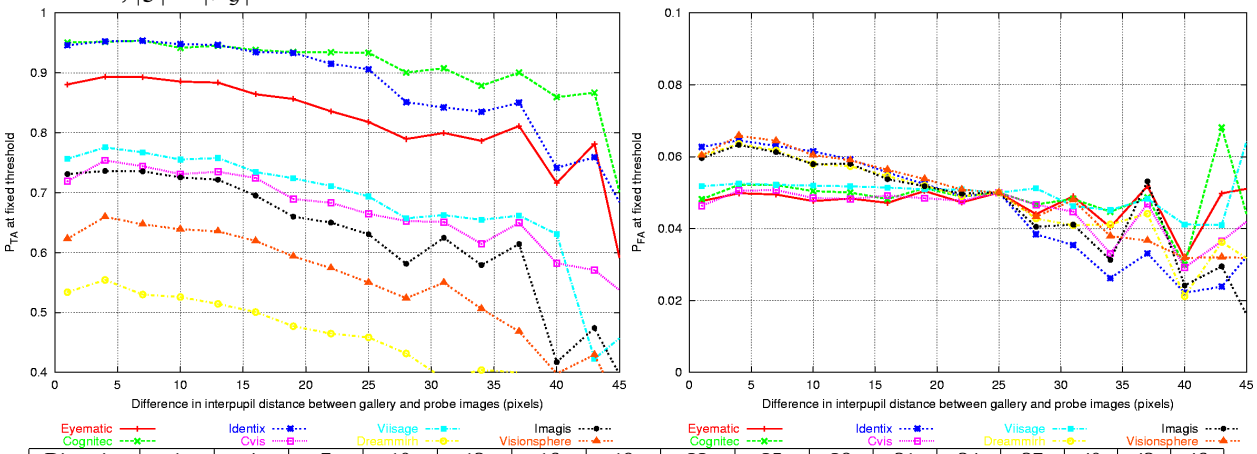
The FRVT images are loosely conformant to the specification of the canonical *token image* set forth in the draft international standard (SC37 - [4]), particularly: the size of the image, 252 x 300, is close to the specification of 240 x 320; the median inter-pupil distance of 72 exceeds the standard's mandated minimum value of 60. However given a standard deviation of 9.4, some 7.3% of the images have an inter-pupil distance less than the standard's minimum.

Figures 7 and 8 show verification performance as functions of the average of gallery and probe inter-pupil distance, and of the discrepancy between the gallery and probe's inter-pupil distance. The distances were obtained either by



Distance	43.5	47.5	51.5	55.5	59.5	63.5	67.5	71.5	75.5	79.5	83.5	87.5	91.5
Count	16	119	458	1146	2509	4629	6697	7971	7001	4243	1896	578	184

Figure 7: Performance as a function of mean inter-pupil distance. At left the true match rate at a fixed threshold and, at right, the false match rate at that same threshold. Normalization was not used. The table shows the number of images in each bin, $|\mathcal{G}| = |\mathcal{P}_\mathcal{G}|$. Note the small numbers at left.



Disparity	1	4	7	10	13	16	19	22	25	28	31	34	37	40	43	46
Count	1621	5322	6840	6485	5519	3931	2892	1934	1166	801	441	255	139	66	40	17

Figure 8: Performance as a function of the discrepancy between inter-pupil distances. At left the true match rate at a fixed threshold, and, at right, the false match rate at that same threshold. Normalization was not used. The table shows the number of images in each bin, $|\mathcal{G}| = |\mathcal{P}_\mathcal{G}|$. Note the small numbers at right.

manual location of the eyes, or from an implementation of the Schneiderman algorithm [14]. The result is a stronger dependency on the discrepancy between images than on the absolute value. This is somewhat counter-intuitive since recognition systems will inevitably degrade at low enough inter-pupil distances, and because they *do* take steps to normalize out geometrical variations.

4.2 Compression Ratio

The Department of State images used in FRVT are JPEG compressed. The images, having size 300 x 252 pixels with three color channels and a mean file size 9467 bytes, depart from the draft standard specification [4] in that the compression ratio of 24:1 exceeds the mandated maximum of 20:1. High compression ratios, like low resolution values, will ultimately lead to a degradation in recognition performance. This has previously been quantified [4] on a smaller data set, and on a set of originally *uncompressed* images [7]. Although the FRVT imagery was not designed to allow a similar controlled test of the effect of compression, the results that follow are included because of the large number, and operational nature, of the images used. Also by computing the compression ratio calculation over just the area that approximates the *Inner Region*, (i.e. the face plus minimal border; see Figure 13) as specified in the draft standard, the mean compression ratio, at 18:1, is within specification. The on-face ratio is computed by first determining the JPEG quality applied to the original image, and then using the JPEG compressor on the cropped image at that quality.

Figure 9 shows, at left, a decline in true match rate as filesize is decreased, and, at right, a larger increase in false match rates. The computation is suspect because the filesize incorrectly includes the easy-to-compress uniform background, and should be considered a worse indicator of the effect of compression than that appearing in Figure 10 which shows that both true and false match rates suffer at compression ratios above about 21. This is congruent with the effect noted in the draft standard [4]. The Identix system is apparently alone in its successful suppression of false accepts at low compression ratios. Whether the other systems are tuned for moderate compression ratios is not inferable from these figures.

The figures do not show performance as a function of the discrepancy between compression in gallery and probe images.

4.3 In-plane Rotation

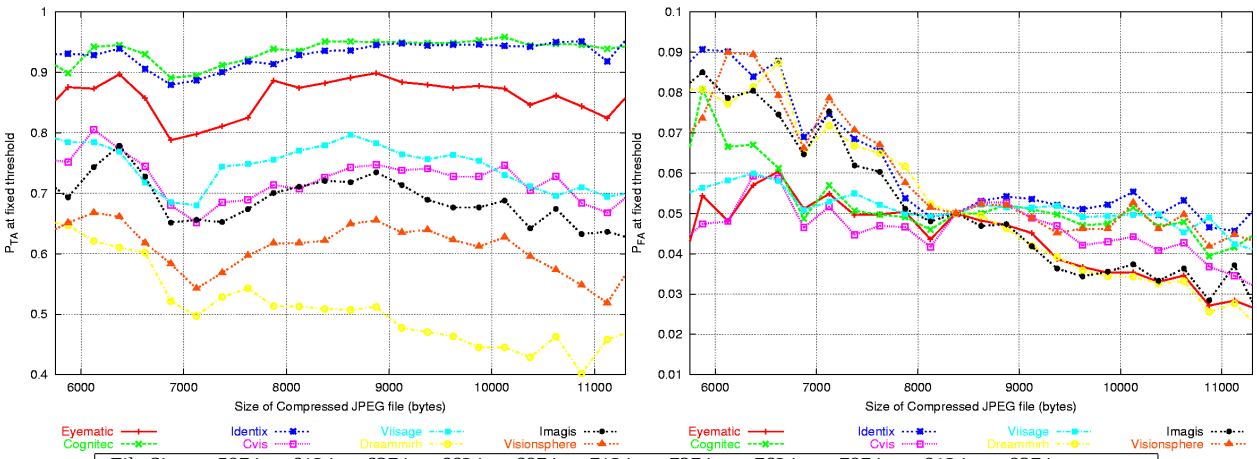
Although left to right head tilt (about optical axis) is a far less pernicious problem than out-of-plane rotations (facing down, or facing left, for example) it is generally removed by in-plane rotation prior to formal recognition. We therefore compute the effect of head tilt using the available FRVT data. A minority of the FRVT images depart from the proposed face recognition interchange format [4] in that the images exhibit some rotation. We therefore bin these by the tilt angle, as computed from the eye coordinates, and, as before, plot performance as functions of the mean in-plane head rotation angle, and the discrepancy. Figures 11 and 12 show an insensitivity to rotation up to 8 degrees. Beyond that there is generally an increase in error rates. However the overall variation over the available range of angles is small in comparison to the resolution and compression studies above. One exception, when there is a discrepancy between gallery and probe tilts, is that a drop in verification rate is partially offset by a *drop* in false match rate. This anomaly doesn't apply for the leading systems however.

4.4 Mixed Effects

It must be noted that the single-effect analyses given in this section will miss correlated covariates. For example the on-face compression ratio is positively correlated with the distance between the eyes. In that case formal statistical methods for failure analysis are warranted. These should generally include both subject and image specific covariates. Recent work [3, 2] along these lines uses generalized linear models in the analysis.

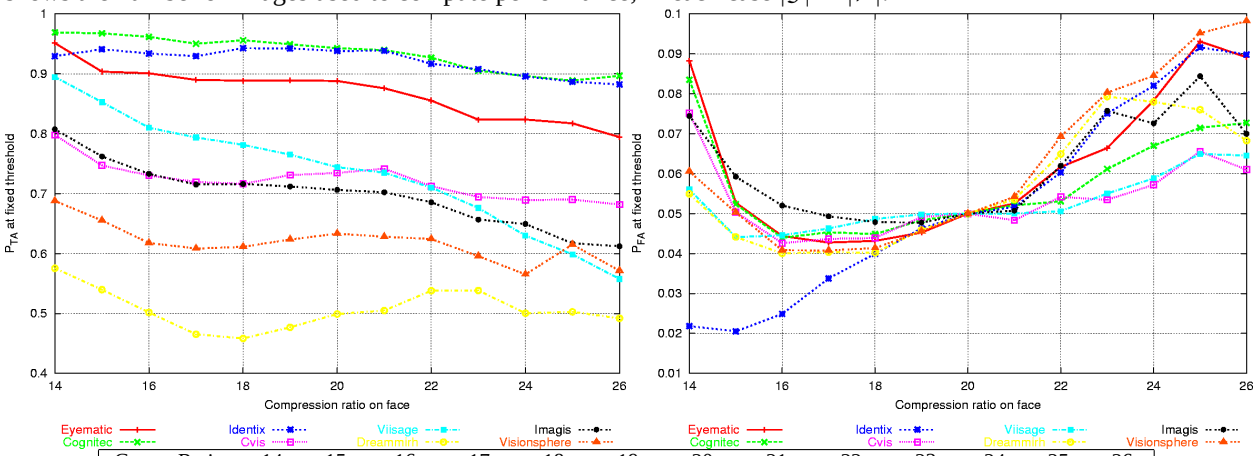
5 Normalization

The use of the normalization in FRVT 2002 has generated considerable discussion since the primary report. This section defines the operation, restates its use in FVRT, addresses its realizability and operational use, asseses it's utility



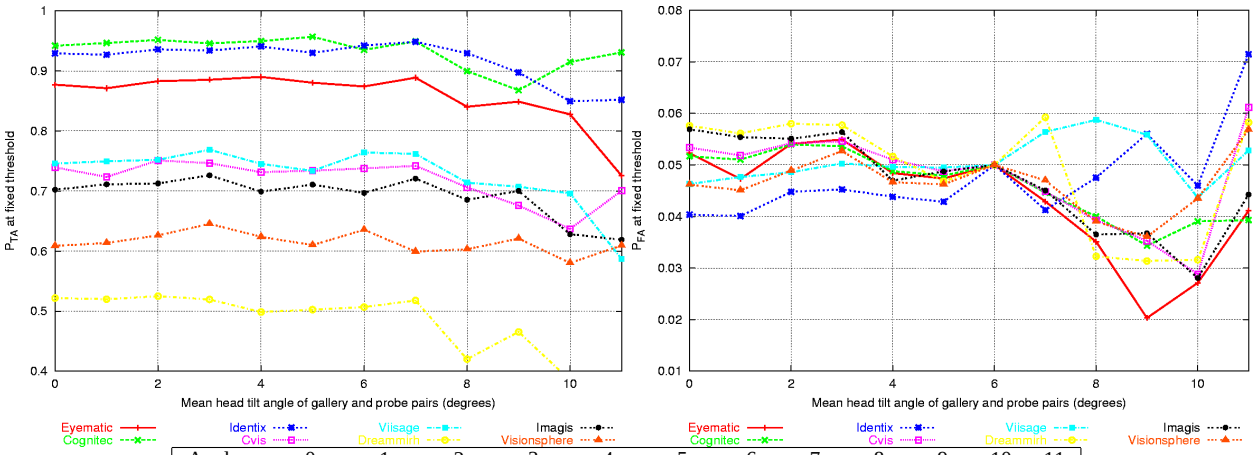
File Size	5874	6124	6374	6624	6874	7124	7374	7624	7874	8124	8374	
Number	167	242	332	486	582	792	995	1190	1498	2128	2871	
File Size	8624	8874	9124	9374	9624	9874	10124	10374	10624	10874	11124	11374
Number	3743	4186	4339	3736	3131	2290	1708	1111	764	437	271	174

Figure 9: Verification performance as a function of the JPEG file size in bytes as it resides on disk. The left figure gives the true accept rate at a fixed threshold, and at right is the false match rate at that same threshold. The table shows the number of images used to compute performance; in each case $|\mathcal{G}| = |\mathcal{P}|$.



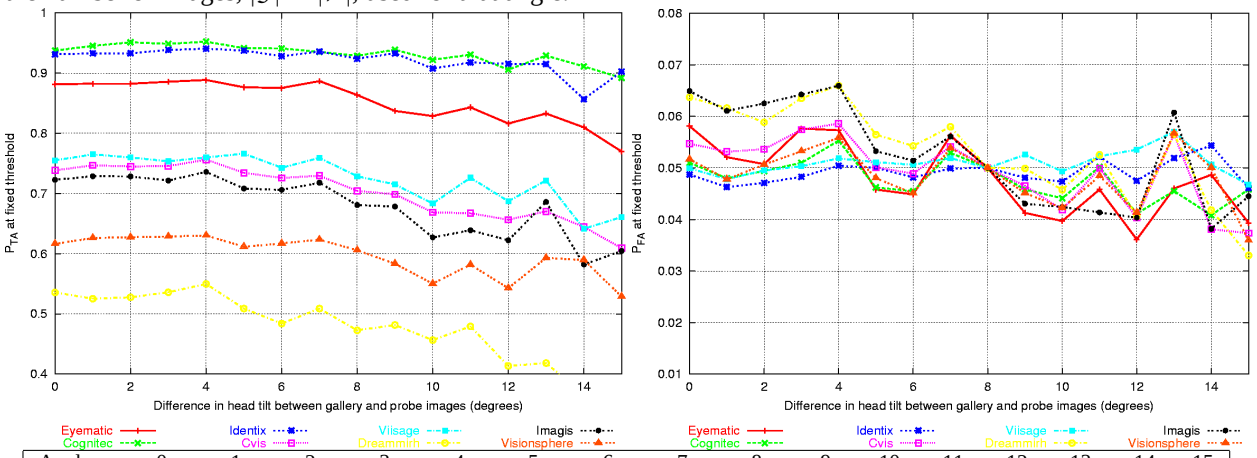
Comp. Ratio	14	15	16	17	18	19	20	21	22	23	24	25	26
Number	248	958	2511	4829	7112	7481	5689	3496	2011	1263	843	478	253

Figure 10: Verification performance as a function of the compression ratio computed over just the rectangle that bounds the face itself. The left figure gives the true accept rate at a fixed threshold, and at right is the false match rate at that same threshold. The table shows the number of images used to compute performance; in each case $|\mathcal{G}| = |\mathcal{P}|$.



Angle	0	1	2	3	4	5	6	7	8	9	10	11
Number	6436	5199	4176	3549	2287	1502	859	483	227	121	57	34

Figure 11: Verification performance vs. mean value of gallery-probe in-plane head rotation in degrees. The left plot shows true match rate at a fixed threshold; the right plot gives false match rate at that same threshold. The table gives the number of images, $|\mathcal{G}| = |\mathcal{P}|$, used for that angle.



Angle	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Number	2357	3504	4843	6624	5809	3702	2990	2815	1759	998	730	516	342	155	127	82

Figure 12: Verification performance vs. discrepancy between gallery and probe in-plane rotation angles. The left plot shows true match rate at a fixed threshold; the right plot gives false match rate at that same threshold. The table gives the number of images, $|\mathcal{G}| = |\mathcal{P}|$, used for that angle.



Figure 13: An image representative of the kind used in FRVT 2002, and a cropped version used to estimate the on-face compression ratio.

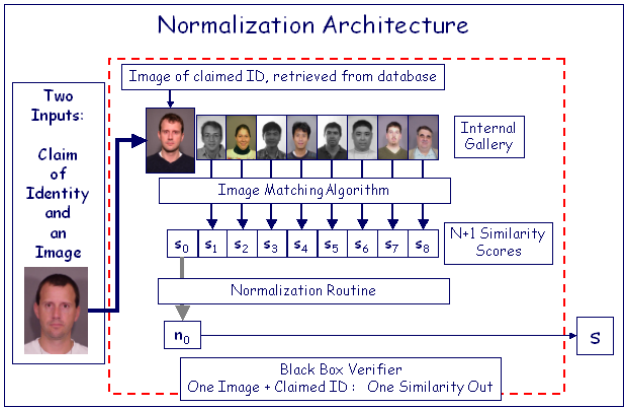


Figure 14: Normalization occurs inside a black box. Additional background images in an internal gallery are used to normalize the score of the image against the claimed identity.

in improving performance, and shows how it renders 1:1 verification a 1:N operation. In addition the vendor-provided *native* normalization functions are compared with a simple and generic yet effective normalizer.

5.1 Overview

Normalization is a technique that aims to improve performance by implicitly setting a sample-specific threshold level. Various methods have been discussed and originate mostly in the speech community [13] where it is used to compensate for environmental variations. Normalization proceeds by comparing the user's image to the claimed match image *and* to other non-matching images, and adjusts the raw score on the basis of the extra scores. Although normalization is used here as a post-processing operation conducted at the similarity score level, it is more generally applicable in higher dimensions in preceeding recognition stages. It's mode of action is to reduce false match rates and it is therefore effective only for open-set tasks.

5.2 z-Normalization

A simple, yet effective normalizer, z-normalization [1], is described here as an example. If a user's image is compared with G images, which may include one from the same user, a vector of G scores is produced. Normalization is a real mapping, $f : R^G \rightarrow R^G$. In the case of z-normalization each element, x_i of a vector of G scores, obtained by comparing a probe image with an internal gallery of G images, is transformed according to:

$$z_i = \frac{x_i - m}{s} \tag{9}$$

where m and s are the sample mean and standard deviation estimated solely from the elements of $\mathbf{x}$:

$$m = \frac{1}{G}\sum_{i=1}^{G} x_i\,, \qquad s^2 = \frac{1}{G-1}\sum_{i=1}^{G}(x_i - m)^2 \tag{10}$$

Thus z-normalization standardizes each user's non-match distributions to approximately zero mean and unit deviation. The effect on performance is given in column two of Figures 5 and 6 the latter of which compares it with vendors' native normalizers. Some systems (Eyematic, Imagis, Dreammirh, Cvis, Identix, and Viisage) show improvement with z-normalization, although this is not always the case for the worst-case ROC. Other systems, Cognitec and Visionsphere, exhibit degraded performance. Particularly the case of Cognitec is interesting in that the variance of the post-normalized scores is large.

5.3 Operational Legality

The essential requirement of any normalization method is that the input is the vector of similarities obtained from *one* probe against some gallery. Simultaneous use of multiple probes' similarity data is invalid because operational realism

dictates that decisions on whether to accept or reject users are inherently sequential and separate [3].

A second qualification is that because normalization involves comparisons with extra images, a throughput performance penalty is incurred. In this sense the use of normalization in the primary FRVT 2002 report was optimistic in that a full gallery (of size 3000, in Figure 8) was used as the background. The addtional benefit of this is covered in the next sections, specifically Table 2. However the full benefit of normalization is legitimately realizable if the application of interest includes a 1:N identification search, in which all N similarity scores are generated and normalization could be computed at "full" size. Thus the relevant metric for quantifying the effect of normalization is the 1:N open-set identification rate (as shown in Figure 1).

5.4 Normalization in FRVT 2002

The use of normalization should be invisible to the operator as it is just another processing step that embeds a vendor's intellectual property. Configuration and use are properly vendor responsibilities. In FRVT 2002, however, normalization functions were provided for post-test execution by NIST. This was done because the FRVT test required *target* and *query* sets of size 121589 to be compared (see primary report), such that scoreable extracts, corresponding to *properly formed* galleries against user and impostor sets, could be normalized and scored separately. The break-out results addressing for example, ethnicity, sex, inter-pupil distance et cetera, are computed from such extracts. In all cases therefore normalization was applied to each column of such a submatrix.

FRVT 2002 allowed two normalization variants termed *F1* and *F2* each in three roles: *ident*, *verif* and *watch*. The *F2* variant, takes the matrix of similarities of the gallery elements against themselves, as additional input. The *ident* function was applied when computing closed-set identification performance; no significant improvement was observed, usually because the functions do not re-order the similarity values, thereby leaving rank-based measures, such as cumulative match characteristic, invariant. The *verif* and *watch* methods showed efficacy, because the open-set problems are sensitive to the distributions of the scores.

5.5 Efficiency

FRVT 2002 quantified the performance gains available from normalization. Specifically the verification results were obtained from vectors of normalized scores of dimension 3000, i.e. normalization was applied to a long vector. Given that normalization on a vector of dimension 1 is ineffective this begs the question: how big a background population is needed to realize normalization's full benefit?

This efficiency issue has considerable implications on the processing time needed for its use. The technique can improve the 1:1 verification but at the expense of making it a 1:N operation. The finding that N can be much smaller typically than an enrolled population is important. Precisely which images should be used in the internal gallery is not addressed here, but clearly could be of considerable importance.

The procedure below was used to assess the effect of normalization for a given internal gallery size. It is broken into two parts, estimating, respectively, the cumulative distribution functions of the match and non-match scores, $M(x)$ and $N(x)$. The ROC follows simply as a plot of $P_{TA} = 1 - M(x)$ vs. $P_{FA} = 1 - N(x)$.

1. Select a set $\mathcal{P}$ of legitimate users
2. Select a set $\mathcal{I}$ of impostors
   (a) Generate a set $\mathcal{B}$ containing K randomly selected background images
   (b) For each $p \in \mathcal{P}$
      i. Let x_1 be the similarity of p with its actual match
      ii. Let x_k be the similarity of p with the background elements $b_k \in \mathcal{B}$, $k = 2 \ldots K+1$
      iii. Normalize the vector $(x_1, x_2 \ldots x_{K+1})$, store the result as vector $\mathbf{y}$
      iv. Extract the normalized **match** score, y_1 and retain it in S_M.
   (c) For each $p \in \mathcal{I}$
      i. Let x_1 be the similarity of p with some a claimed match (actually a non-match)

[3] Some unusual applications may exist in which a closed population is classified against itself: for example, a roll-call inside a properly access-controlled facility.

G	Eyematic		Cvis		Identix		Viisage	
	native	znorm	native	znorm	native	znorm	native	znorm
0	0.809	-	0.595	-	0.878	-	0.637	-
1	0.713	0.811	0.017	0.595	0.879	0.879	0.523	0.635
2	0.808	0.087	0.062	0.062	0.882	0.071	0.544	0.066
4	0.838	0.535	0.217	0.217	0.880	0.181	0.579	0.313
8	0.839	0.758	0.444	0.444	0.880	0.143	0.587	0.500
16	0.844	0.824	0.551	0.551	0.885	0.488	0.589	0.590
32	0.863	0.853	0.618	0.618	0.885	0.725	0.611	0.630
64	0.864	0.856	0.632	0.632	0.876	0.817	0.603	0.637
96	0.859	0.860	0.649	0.650	0.879	0.863	0.622	0.642
128	0.869	0.863	0.649	0.649	0.896	0.868	0.626	0.646
256	0.868	0.864	0.630	0.630	-1	0.887	0.641	0.656
512	0.873	0.871	0.653	0.653	-1	0.896	0.652	0.660
3000	0.870	0.870	0.645	0.645	0.903	0.901	0.656	0.655

Table 2: Dependence of 1:1 verification performance, at 1% false match rate, on the size of the internal gallery used for normalization. The first line, $G = 0$, gives the un-normalized result; the last line $G = 3000$ is the FRVT 2002 result. The *native* columns refer to the vendor's own normalization algorithm.

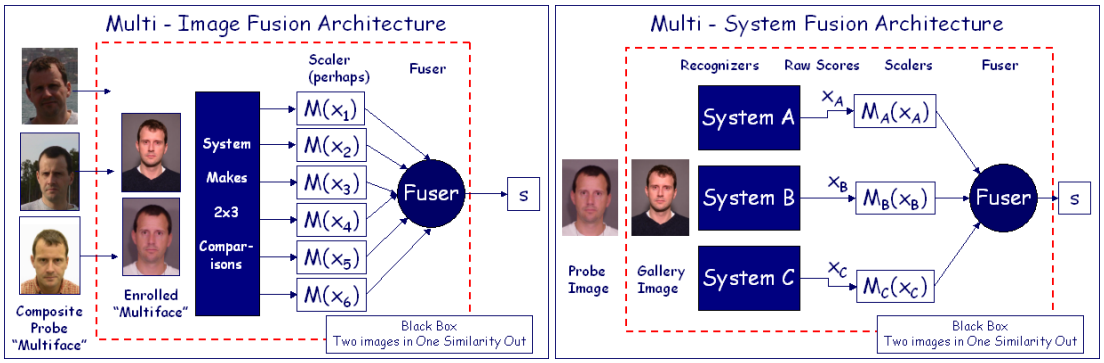


Figure 15: **Left.** Score level multi-image fusion architecture, as used in this paper, top, and the proper architecture in which the testing organization does not contribute technology to the recognition problem.

Figure 16: **Right.** The dashed red box defines the hybrid system formed by internally merging multiple independent recognition systems, via score level fusion. The input-output relationship, and the performance metrics are identical.

ii. Let x_k be the similarity of p with the background elements $b_k \in \mathcal{B}$, $k = 2 \ldots K+1$

iii. Normalize the vector $(x_1, x_2 \ldots x_{K+1})$, store the result as vector $\mathbf{y}$

iv. Extract the normalized **non-match** score, y_1 and retain it in S_{NM}.

(d) increment trial counter, stop if 200, otherwise go to 2a

3. Produce ROC from the retained match and non-match scores, S_M and S_{NM}.

A caveat to this approach is that **it is not known to what extent normalization has already been applied to the scores**. In particular z-normalization actually degrades the Cognitec scores, which may suggest they were normalized at source.

The results are presented in Table 2. The first row, for gallery size zero, is the verification rate (P_{TA} at $P_{FA} = 1\%$) without any external normalization, and the last row, for gallery size 3000, is the result presented in the FRVT report. All the other values are obtained using the sampling method given above. The commercial normalizers attain their benefits within a background gallery of fewer than 100. The z-normalization method underperforms the leading systems' normalizers. Particularly the NIST implementation has no special handling for small background galleries (except for $N = 1, 2$). However for large background galleries z-normalization achieves almost identical performance to all of the supplied native normalizers.

6 Fusion

The benefits available from biometric fusion have been extensively documented for many applications. Indeed with renewed emphasis on very high performance and robustness fusion is attracting considerable attention. The term fusion can be applied in at least three ways:

1. *Samples* Several instances of a given individual are processed by one system.
2. *Systems* Several algorithms processing the same input.
3. *Modes* Multiple biometric modalities are presented to a suitable system. For example: face and fingerprint; iris and hand geometry; and face, iris, and fingerprint.

FRVT 2002 only included face imagery so the third category is beyond study here. But instead system level fusion is reported below for all pairs of FRVT participants, for the three leading systems, and for single systems operating on multiple images. The fusion methods used here were developed by NIST, without input or comment from the vendors.

6.1 Score Level Fusion

Fusion can be conducted at, at least, the image, feature, score, rank or decision stages. It is widely known that early stage fusion, in higher dimensional spaces, offers the most benefit, and for this reason fusion is properly the responsibility of the algorithm developer. Indeed the only role of the tester seeking to demonstrate the utility of fusion is to design it into the test structure.

1. **Multiple Samples:** The tester should arrange to supply a system with multiple *signatures* (i.e. arbitrary collections of biometric samples from a person) [10]. The architecture depicted in Figure 15 shows this for face, but could equally apply to the multi-modal case too.
2. **Multiple Systems:** The tester should arrange to obtain a single similarity value from a common input. This requires no special treatment by the tester because, as Figure 16 shows, multiple systems, working in concert, essentially form a single black box system.

The Human ID Evaluation Framework [10], an XML infrastructure used in the administration of FRVT 2002[4], supports multiple, arbitrary biometric, image sets as input. This capability was only exploited in FRVT 2002 for video sequences and not for the High Computational Intensity Tests considered here. It is therefore not clear to what extent the systems tested in FRVT 2002 were capable of fusion at any level. Thus, although the score-level fusion reported below should be deprecated in standard testing protocols, it is nevertheless reported here because of its policy implications, and because it establishes an empirical lower bound baseline that vendors should, at least, be able to match.

[4] And the Fingerprint Vendor Test (FPVTE): Report due early 2004.

6.2 Scaling

When fusing scores from **multiple systems**, the first step is to map disparately distributed scores from the systems onto a common range. Several methods of doing this are ubiquitous: linear scaling either via maximum and minimum values, or sample mean and variance estimates, and various non-linear scalings [8]. The method used here employs an estimate of each system's match-score distribution, and is not described in the literature surveyed by the author. For fusion of scores from **multiple images** processed by a single system, scaling is optional. Here we compare recognition performance obtained from fusion of scaled and unscaled scores.

Formally the i-th score from the k-th system is transformed using an empirical estimate of the cumulative distribution function of the match scores, M_k,

$$y_{ki} = M_k(x_{ki}) \tag{11}$$

The resulting match scores will be approximately uniform on $[0, 1]$. The non-match scores by contrast have some unknown distribution with a peak at zero corresponding to any score lower than the lowest match score used in the estimation of M_k. That some match scores, too, are at or near zero is not intrinsically deleterious to successful fusion. This may be avoided by using $M_k(x) + N_k(x)$ as the scaling function, with N_k the CDF of the non-match scores. However this, and the linear scaling methods, were found to be consistently inferior.

6.2.1 Learned Scaling

The legitimacy of using $M_k(x)$ merits some discussion. Its use is predicated on the assumption that a reasonable estimate of the match scores can be obtained for a system *before* deployment. If this is possible, for example by using a representative volunteer corpus, or by using early users of the deployed system, then the method is legitimate. Indeed it is not clear to what extent a vendor will know the statistics (even to low order) of the match scores, and be able to exploit them. That is, the same legitimacy question arises for linear scalers too, where low order statistics (minimum and maximum, or mean and standard deviation) are used, because they too must be estimated empirically. Also the usual caveat, that the distributions may change over time because of sensor or illuminations changes, applies.

6.3 Multi-image Fusion

This section considers the utility of fusing one system's scores from several probe-gallery image comparisons. Two methods of fusing are considered. First is the addition of K similarity scores from $K + 1$ images:

$$s_k^\dagger = \sum_{k=1}^{K} s_k \tag{12}$$

Second is the summation of scaled scores:

$$s_k^\dagger = \sum_{k=1}^{K} M(s_k) \tag{13}$$

where M is the cdf of the match scores. We also include the effect of z-normalizing the fused scores:

$$z_k^\dagger = Z(s_k^\dagger) \tag{14}$$

where Z represents the operation defined in eq. (9).

We report multi-sample fusion results on two subpopulations of the FRVT corpus. First is the 1678 persons with five or more images; second is the 4678 persons with four or more images. The former is a subset of the latter. This allowed fusion of, respectively, $1 \ldots 4$ and $1 \ldots 3$ scores to be conducted, i.e. those similarity values resulting from the comparison of a probe to each of four and three gallery images. The results of using the various combinations of scaling and normalization are shown in Table 3, notes on which follow:

1. **Population** The two subpopulations are comprised of a significantly different population than that of the whole FRVT 2002 corpus. Note that for $K = 1$, i.e. no fusion, the verification rates are significantly higher than those observed in FRVT 2002 which are included in this report in Figures 1 ($N = 1$) and 5. This arises because the frequent visa applicants, those with five or more images, are more often men (65.8% here compared with 48.8% in the 37437 person FRVT corpus) and more ethnically diverse (10.3% are Chinese here versus 1.7% in FRVT

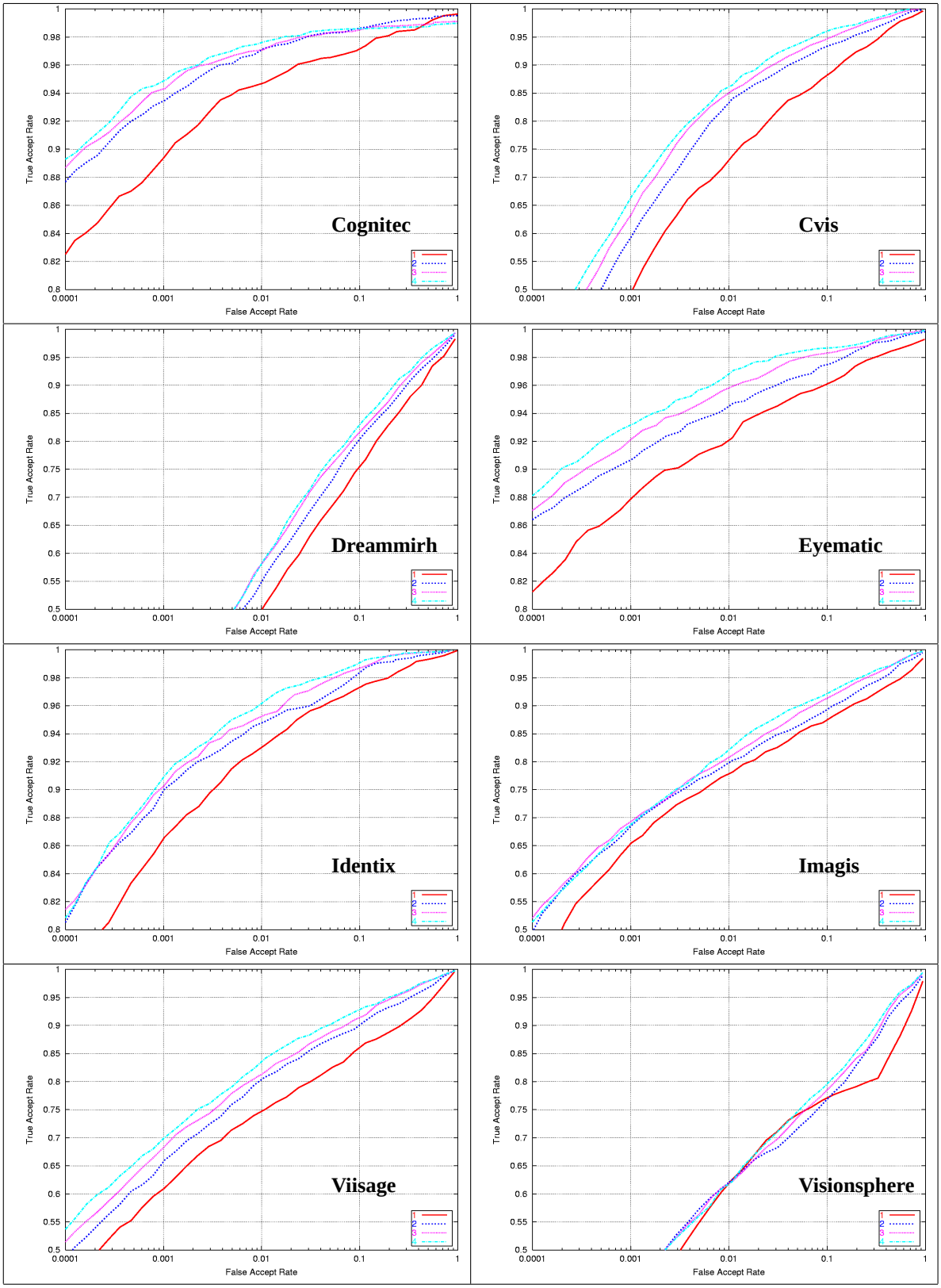


Figure 17: Verification performance after fusion of multiple images' **raw** scores, with post-fusion z-normalization. The four curves in each plot correspond to the fusion of 1, 2, 3 and 4 scores. Note the different y-axis scales.

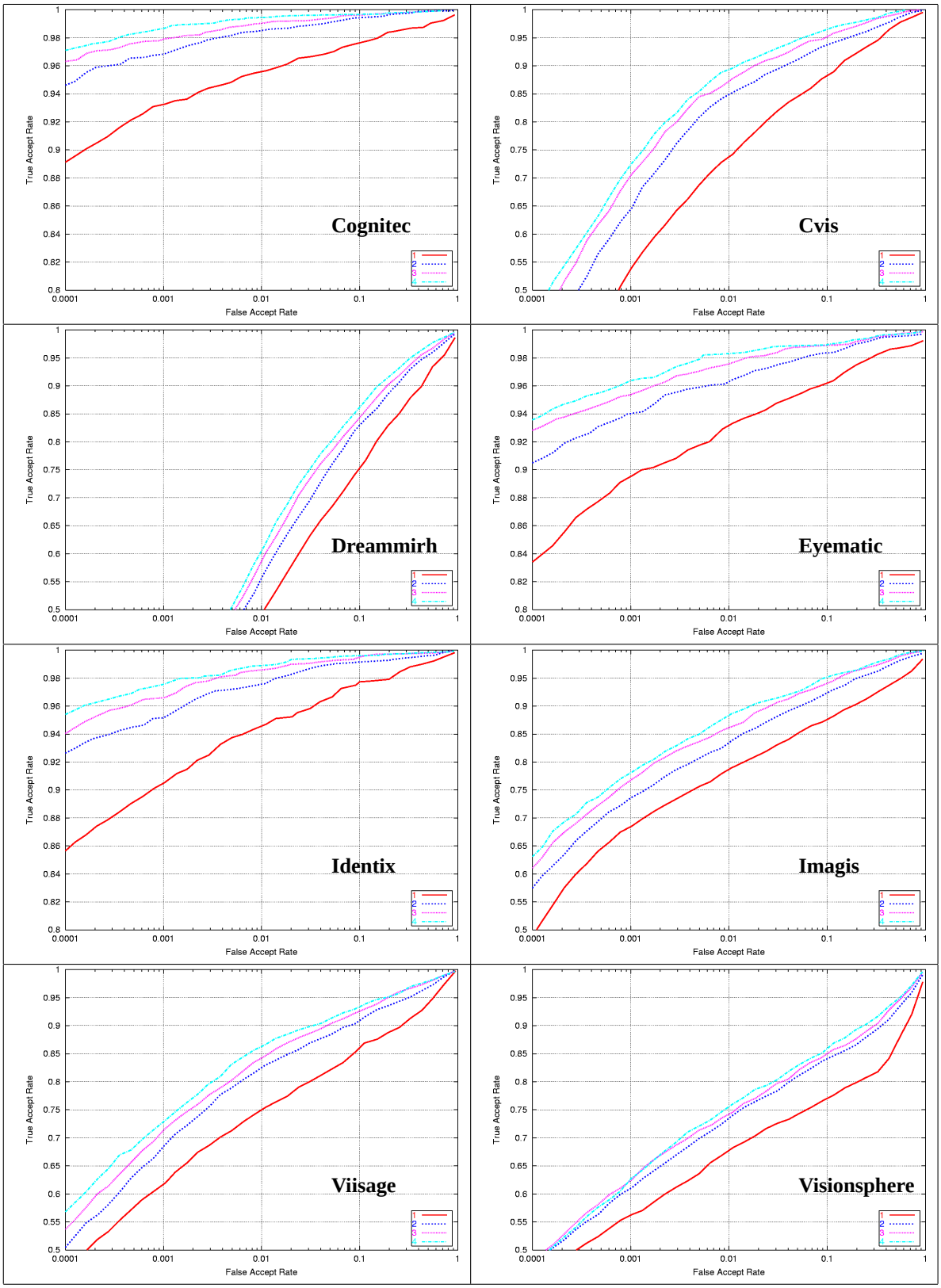
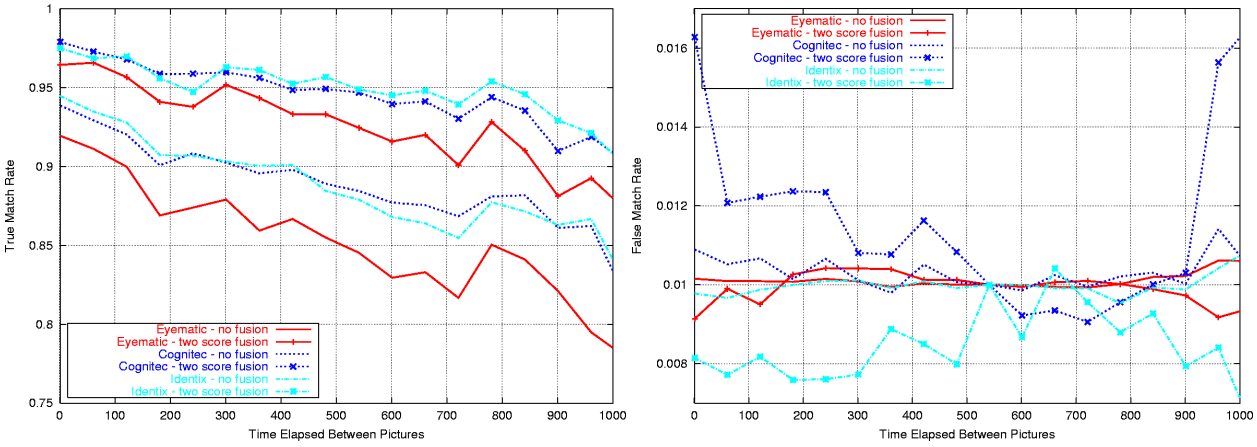


Figure 18: Verification performance after fusion of multiple images' M **scaled** scores, with post-fusion z-normalization. The four curves in each plot correspond to the fusion of 1, 2, 3 and 4 scores. Note the different y-axis scales.

K	$\Sigma_k^K x_k$	$\Sigma_k^K M(x_k)$	$z\left(\Sigma_k^K x_k\right)$	$z\left(\Sigma_k^K M(x_k)\right)$
Cognitec				
1	0.941	0.941	0.946	0.956
2	0.963	0.971	0.971	0.985
3	0.962	0.977	0.970	0.990
4	0.964	0.979	0.976	0.994
Cvis				
1	0.707	0.707	0.732	0.736
2	0.773	0.788	0.833	0.848
3	0.793	0.811	0.851	0.875
4	0.789	0.823	0.860	0.893
Dreammirh				
1	0.447	0.447	0.500	0.491
2	0.458	0.449	0.546	0.555
3	0.455	0.442	0.583	0.587
4	0.432	0.430	0.581	0.605
Eyematic				
1	0.888	0.888	0.920	0.932
2	0.908	0.930	0.947	0.963
3	0.914	0.939	0.958	0.976
4	0.916	0.945	0.970	0.983
Identix				
1	0.920	0.920	0.931	0.946
2	0.940	0.957	0.948	0.975
3	0.938	0.969	0.952	0.986
4	0.947	0.975	0.962	0.989
Imagis				
1	0.706	0.706	0.779	0.788
2	0.714	0.743	0.798	0.834
3	0.707	0.752	0.808	0.863
4	0.697	0.762	0.819	0.880
Viisage				
1	0.733	0.733	0.747	0.750
2	0.781	0.802	0.802	0.823
3	0.787	0.819	0.815	0.842
4	0.812	0.837	0.836	0.861
Visionsphere				
1	0.638	0.638	0.618	0.677
2	0.610	0.693	0.620	0.735
3	0.590	0.689	0.617	0.743
4	0.578	0.696	0.616	0.756

K	$\Sigma_k^K x_k$	$\Sigma_k^K M(x_k)$	$z\left(\Sigma_k^K x_k\right)$	$z\left(\Sigma_k^K M(x_k)\right)$
Cognitec				
1	0.930	0.930	0.942	0.951
2	0.955	0.961	0.967	0.980
3	0.960	0.970	0.972	0.986
Cvis				
1	0.682	0.682	0.724	0.730
2	0.739	0.760	0.809	0.830
3	0.762	0.785	0.836	0.862
Dreammirh				
1	0.426	0.426	0.489	0.484
2	0.425	0.417	0.528	0.538
3	0.429	0.427	0.538	0.558
Eyematic				
1	0.864	0.864	0.908	0.913
2	0.888	0.919	0.936	0.958
3	0.893	0.925	0.948	0.969
Identix				
1	0.919	0.919	0.934	0.938
2	0.937	0.958	0.951	0.973
3	0.938	0.972	0.952	0.984
Imagis				
1	0.672	0.672	0.762	0.768
2	0.670	0.718	0.793	0.837
3	0.668	0.721	0.796	0.855
Viisage				
1	0.702	0.702	0.719	0.720
2	0.749	0.769	0.764	0.787
3	0.769	0.792	0.792	0.817
Visionsphere				
1	0.611	0.611	0.589	0.641
2	0.553	0.651	0.572	0.693
3	0.529	0.657	0.575	0.712

Table 3: Verification performance at a fixed 1% false match rate, for various combinations of scaled and z-normed summations of similarity scores. The left and right tables pertain, respectively, to the 1678 and 4678 person populations with five and four or more images per person. The number of fused scores, K, corresponds to the comparison of a single probe with K gallery entries. The four math expressions at the head of the columns refer, respectively, to summation of raw scores, summation of M-scaled scores, z-normalization of the summation of raw scores, and z-normalization of the summation of M-scaled scores.



Time	1	61	121	181	241	301	361	421	481	
Number	1560	1533	1673	2193	2389	2836	3499	2869	2820	
Time	541	601	661	721	781	841	901	961	1021	1081
Number	2594	2429	2515	2626	1559	1209	965	605	533	497

Figure 19: The effect of time delay on the efficacy of multi-sample fusion. The curve plots true match rate at a fixed threshold as a function of the time elapsed between the gallery image and the probes. For the fused result the time refers to the latter of the two that were fused. The false match rate, shown at right, varies little from its nominal 1% value. The results were computed from normalized similarity values, except for Cognitec.

generally), both of which help most of the systems. The same is true, but to a a lesser extent for the four-image persons. Nevertheless the benefits of fusion, *stated relative to the unfused, single-image case*, are significant.

2. **Effect of K** Performance improves substantially with K. Only Visionsphere shows an anomalous decline in performance. For the leading systems much of the improvement is realized for $K = 2$. Thereafter returns typically diminish. For example, looking at the Identix system for summed raw scores, for the five image population, the false non-match rate (i.e. $1 - P_{TA}$) decreases from 8% ($K = 1$) to 6% ($K = 2$) with only a further reduction of 0.7% for two more scores.

3. **Effect of Scaling** Performance without fusion, $K = 1$, is unchanged by the scaling operation because $M(x)$ is a monotonic function of x. However the distribution of the scores does change and therefore z-normalization does change the result for $K = 1$.

   With multiple scores, i.e. $K \geq 2$, the use of scaling improves performance except in the case of Dreammirh. This is to be expected on the basis that M is adding information.

4. **Effect of z-Normalization** The best results are obtained with z-normalization. The highest verification rate, $P_{TA} = 99.4\%$ at $P_{FA} = 1\%$, is obtained from the Cognitec system after z-normalization of the sum of four scaled scores. The technique is effective for scaled and unscaled scores.

5. **Mechanism of Action** Fusion essentially integrates information from multiple images. The process can potentially average away any abnormally low match values arising from single-image quality defects. However the likely reason for the success of the technique stems from the fact that face recognition systems depend on accurate localization of facial features, in particular the eyes. The incorporation of multiple images effectively reduces localization errors via averaging. Systems based on eigenface techniques should reap significantly more benefit from such information than other published algorithms such as LFA.

   The extent to which a single high resolution image would solve the localization difficulties in small images is not known. Certain performance improvements may be available by giving greater weight to recent images, or by factoring in per-image assessments of image quality.

6. **Full ROCs** Figures 17 and 18 show the full ROCs for the two kinds of scaling. Both include z-normalization.

Vendor	Eyematic	Cognitec	Identix	Cvis	Viisage	Dreammirh	Imagis	Visionsphere
Eyematic	0.870 0.809	**0.949**	**0.937**	0.835	0.852	0.731	0.849	0.793
Cognitec	**0.905**	0.899 0.901	**0.954**	0.860	0.867	0.757	0.879	0.822
Identix	**0.901**	**0.932**	0.903 0.878	0.851	0.862	0.754	0.873	0.814
Cvis	0.773	0.812	0.810	0.645 0.595	**0.770**	0.633	**0.765**	**0.704**
Viisage	**0.825**	0.850	0.847	**0.727**	0.656 0.637	0.634	**0.767**	**0.712**
Dreammirh	0.638	0.669	0.676	0.545	0.559	0.389 0.335	0.630	**0.554**
Imagis	0.756	0.799	0.805	**0.678**	**0.711**	0.516	0.689 0.591	0.688
Visionsphere	0.744	0.788	0.781	**0.644**	**0.681**	0.462	**0.604**	0.495 0.530

Table 4: Verification performance at $P_{FA} = 0.01$ for systems fused in pairs **without** weighting. The inputs to all fusions are non-normalized scores. The above diagonal elements are computed from similarities that have been z-normed after fusion. The on-diagonal elements give the plain verification scores with no pre- or post-normalization. Bold entries indicate an improvement over *both* systems run alone.

6.3.1 Effect of Elapsed Time

The multi-sample fusion studies above make use of images obtained over a period of several years. Many applications, however, would benefit if the same performance improvements were all obtained on the same day. Although a full study of this is beyond the FRVT 2002 design, Figure 19 shows the relative benefit of fusion for increasingly time lapsed gallery and probe-pair images. It plots 1:1 verification performance against the time elapsed between the gallery image and the *latter* of the two probe images whose two scores are fused. It compares this with the result reported in the FRVT primary report showing single probe (i.e. no fusion) verification as a function of time. The conclusion is that the benefit, i.e. the fractional reduction in false non-match rate $1 - P_{TA}$, is marginally higher for short elapsed times.

6.4 Multi-system Fusion

The issue of whether recognition systems fail on the same images is worthy of some investigation. For now, we assume that there is limited overlap in the failures and proceed to fuse the scores from multiple systems, first pairwise and then in a triple. Scaling is necessary so the fusion of K systems extends the multi-sample equations (13) and (14) to include system-specific scaling functions:

$$z^{\dagger} = Z\left(\sum_{k=1}^{K} M_k(s_k)\right) \tag{15}$$

Figure 4 shows the effect of pairwise fusion of the FRVT 2002 systems participating in the High Computational Intensity test. Because the fusion is symmetric, the above diagonal elements would be redundant and are therefore replaced by those that result if the fused scores are z-normalized. The data sets used for this study are eleven disjoint sets each of size $|\mathcal{G}| = 3000$, $|\mathcal{P}_{\mathcal{G}}| = |\mathcal{P}_{\mathcal{N}}| = 6000$. The best result, from Identix and Cognitec, with post-fusion z-normalization, has $P_{TA} = 95.4\%$ at $P_{FA} = 1\%$ representing a reduction in verification errors of more than a factor of two over the best FRVT 2002 result, $P_{TA} = 90.1\%$.

6.5 Weighted Fusion

When certain systems are more capable than others a weighted fusion is one method of accounting for this:

Vendor	Eyematic	Cognitec	Identix	Cvis	Viisage	Dreammirh	Imagis	Visionsphere
Eyematic	0.870 0.809	**0.953**	**0.936**	**0.891**	**0.899**	**0.878**	**0.884**	**0.879**
Cognitec	**0.919**	0.899 0.901	**0.957**	**0.936**	**0.938**	**0.931**	**0.933**	**0.932**
Identix	**0.906**	**0.934**	0.903 0.878	**0.910**	**0.916**	**0.908**	**0.910**	**0.908**
Cvis	**0.815**	**0.902**	**0.881**	0.645 0.595	**0.770**	**0.670**	**0.770**	**0.710**
Viisage	**0.848**	**0.911**	**0.894**	**0.729**	0.656 0.637	**0.669**	**0.771**	**0.716**
Dreammirh	0.809	**0.901**	**0.878**	**0.604**	**0.642**	0.389 0.335	**0.708**	**0.571**
Imagis	**0.809**	0.900	**0.879**	**0.678**	**0.711**	**0.591**	0.689 0.591	**0.716**
Visionsphere	0.809	0.900	**0.878**	**0.647**	**0.686**	**0.532**	**0.610**	0.495 0.530

Table 5: Verification performance at $P_{FA} = 0.01$ for systems fused in pairs **with** weighting. The inputs to all fusions are non-normalized scores. The above-diagonal elements are computed from similarities that have been z-normed after fusion. The lower triangle uses the un-normalized fused results. The on-diagonal elements give the plain verification scores with no pre- or post-normalization. Bold entries indicate an improvement over *both* systems run alone.

Vendor	Eyematic	Cognitec	Identix	Cvis	Viisage	Dreammirh	Imagis	Visionsphere
Eyematic	1	3.01	1.61	0.12	0.19	0.07	0.11	0.09
Cognitec	3.14	1	0.47	0.08	0.08	0.01	0.11	0.02
Identix	1.85	0.73	1	0.15	0.10	-0.03	0.10	-0.02
Cvis	0.12	0.07	0.11	1	0.92	0.38	1.48	0.60
Viisage	0.46	0.21	0.20	1.12	1	0.31	1.50	0.74
Dreammirh	0.00	-0.07	-0.00	0.23	0.13	1	6.91	4.00
Imagis	0.05	-0.05	0.13	1.07	1.03	9.69	1	0.28
Visionsphere	0.03	-0.05	0.01	0.82	0.77	8.53	0.44	1

Table 6: Weights applied in equation 13 to the pairwise system fusion. For entries below the diagonal the weight applies to the system given in that row, the other system has weight 1. For entries above the diagonal the weight is applied to the system indicated by the column name, the other system has weight 1. The above-diagonal elements applies to the scores that have been z-normed after fusion. Those below the diagonal do not use post normalization.

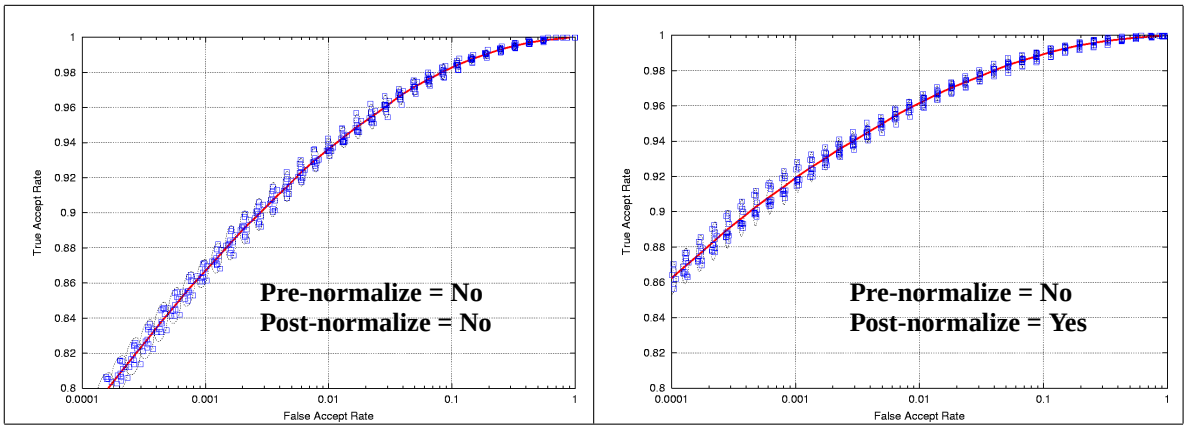


Figure 20: Multi-System Fusion: Verification performance of the system that results from the weighted sum-rule fusion of the Cognitec, Eyematic and Identix scores. The ROCs are computed from 11 disjoint sets of $|\mathcal{G}| = 3000, |\mathcal{P}| = 6000$ images.

$$z^{\dagger} = Z\left(\sum_{k=1}^{K} w_k M_k(s_k)\right) \tag{16}$$

where, without loss of generality, $w_K = 1$. The weights w_k are obtained as that set that minimizes the area under the ROC curve

$$A = \int_{-\infty}^{\infty} N(x)m(x)dx \tag{17}$$

where $N(x)$ is the cumulative distribution function of the fused non-match scores and $m(x)$ is the density of the matches. In practice these are not known so the area is approximated by numerical integration of the above using empirical estimates for the distributions. The optimizer itself is the 1D Brent optimization algorithm [12]. For three or more systems a multidimensional principal axis method known as *praxis* [5] is used. In neither case is derivative information required.

A distinct set of 3000 x 6000 images is used for this training. The results, shown in Table 5, are only moderately better than for the unweighted fusion, except for those many cases where one of the leading systems is fused with one of the less capable systems. Table 6 shows the weights applied.

6.6 Three System Fusion

Taking this approach further, Figure 20 shows ROCs for the weighted fusion of the scores from the three leading FRVT participants, Eyematic, Identix and Cognitec. The verification result, $P_{TA} = 0.962$ at $P_{FA} = 0.01$, is superior to that obtained by pairwise fusion of Identix and Cognitec, namely $P_{TA} = 0.957$.

6.7 Multi-system and Multi-image Fusion

The method of fusion generalizes to the fusion of multiple images *and* multiple systems. Several strategies may be entertained for doing this: straightforward fusion of all image scores from all systems; of images *then* systems; and finally of systems *then* images. With unweighted additive models these schemes are equivalent, but in the general case the results differ. Although the results described below were obtained with the first "all in" strategy, a more realistic integration would be to form a black box hybrid system, which internally would direct an image through the separate systems in parallel, fuse the output scores, and emit only the one result. This architectural difference does not significantly alter the results.

The result, shown in Figure 21, gives the verification result obtained by fusing scores from the three leading systems, i.e. Cognitec, Eyematic and Identix. The four traces on each graph correspond to the use of one, two, three and four

[5] Code and documentation is via http://gams.nist.gov

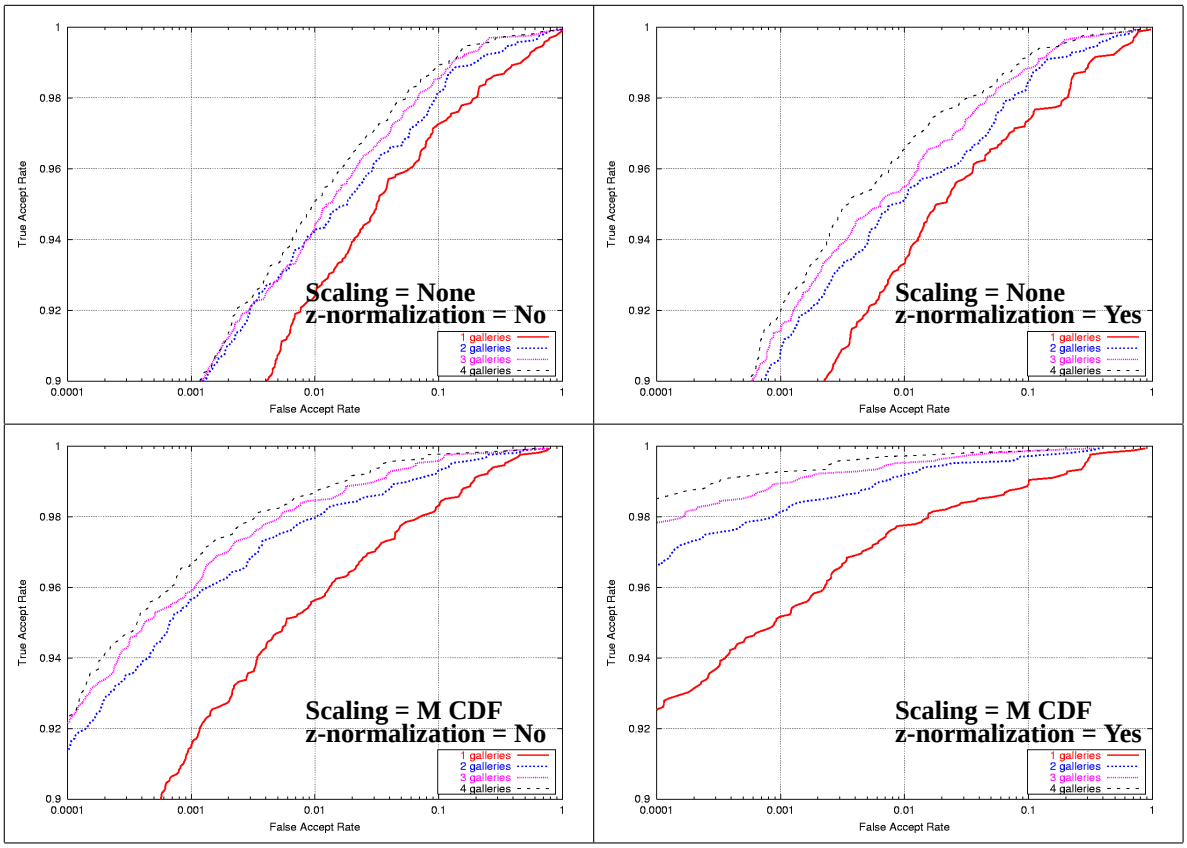


Figure 21: Multi-System and Multi-Image Fusion. Verification performance obtained after scaled, unweighted, sum-rule fusion of the Cognitec, Eyematic and Identix scores from the comparison of a single gallery image with 1,2,3 and 4 probe images per user.

probe images compared with the gallery. The fusion is unweighted. It cannot be compared with the pairwise fusion in Table 4 because here only the 1678 person population has been used. Instead the results are comparable with the multi-image, single-system results shown in Table 3. The conclusion is that the three systems together usually outperform any one system alone. The best result, a true match rate of 99.7% at 1% false match rate, obtained with four-score fusion, scaling and z-normalization, is an improvement from 99.4% for Cognitec alone. The extent to which weighting the systems would improve this further is not known at this time.

7 Correction

The primary FRVT report [11] presents a watch list ROC (cf. Figure 1 here) in Figure 12. The caption, and the relevant passage in the main text, should reflect a gallery size of 800, not 3000.

8 Acknowledgements

The FRVT 2002 team members, responsible for design implementation, execution, and the primary analysis were, Jonathon Phillips, Patrick Grother, Duane Blackburn, Ross Micheals, Elham Tabassi, and Mike Bone. The author is grateful to the other teams members for on-going advice and ideas.

The author also wishes to thank David Casasent for helpful insight on the normalization topic.

References

[1] S. Bengio, J. Mariéthoz, and S. Marcel. Evaluation of biometric technology on xm2vts. Technical report, Dalle Molle Institute for Perceptual Artificial Intelligence, 2001.

[2] G. H. Givens, J. R. Beveridge, B. A. Draper, and D. Bolme. Using a generalized linear mixed model to study the configuration space of a pca+lda human face recognition algorithm. Technical report, Colorado State University, Fort Collins, 2002.

[3] G. H. Givens, J. R. Beveridge, B. A. Draper, and D. Bolme. A statistical assessment of subject factors in the pca recognition of human faces. In *CVPR 2003 Workshop on Statistical Analysis in Computer Vision Workshop*, June 2003.

[4] P. Griffin. Face recognition format for data interchange. Committee Draft ISO/IEC 19794-5 SC37 M1/03-0494 SC37 Document 342, Identix Corporate Research Center, August 2003.

[5] P. J. Grother, R. J. Micheals, and P. J. Phillips. Performance metrics for the frvt 2002 evaluation. In *Proceedings of Audio and Video Based Person Authentication Conference*, June 2003.

[6] P. J. Grother and P. J. Phillips. Models of large population recognition performance. In *Publication Pending*, 2004. Contact: patrick.grother@nist.gov.

[7] T. Hartmann. Machine Readable Travel Documents, Annex B. Technical report, 2003.

[8] M. Indovina, U. Uludag, R. Snelick, A. Mink, and A. Jain. Multimodal biometric authentication methods: A cots approach. In *Workshop on Multimodal User Authentication*, Santa Barbera, CA., December 2003.

[9] R. Micheals and T. Boult. Is the urn well mixed? discovering false cofactor homogeneity assumptions in evaluation. In *Publication Pending*, 2004. Contact: rossm@nist.gov.

[10] R. J. Micheals, P.J. Grother, and P. J. Phillips. The human id evaluation framework. In *Proceedings of Audio and Video Based Person Authentication Conference*, June 2003.

[11] 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 or www.frvt.org, March 2003.

[12] W. H. Press, B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling. *Numerical Recipes*, chapter 10. Cambridge University Press, 1989.

[13] D. Reynolds. Comparison of background normalization methods for text independent speaker verification. In *European Conference on Speech Communication and Technology (EUROSPEECH 97)*, pages 963–966, 1997.

[14] H. Schneiderman and T. Kanade. A statistical method for 3d object detection applied to faces and cars. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 746–751, 2000.

Appendix P

Participant Comments on FRVT 2002: Supplemental Report

FRVT 2002 Participants were provided a pre-release copy of the *FRVT 2002: Supplemental Report*, and were invited to submit a 2-page (maximum) document that serves as their "position paper" on the supplemental report.

Participant documents are the opinions of the Participants and are provided in this text for reference. Inclusion of these documents does NOT imply that FRVT 2002 Authors/Sponsors/Supporters agree with any statements or derivative results within those documents. The identification of any commercial product or trade name does not imply endorsement or recommendation by the National Institute of Standards and Technology or any other FRVT 2002 Author/Sponsor/Supporter.

Suite 1630
1075 W. Georgia Street
Vancouver, BC, Canada V6E 3C9
Tel: +1.604.684.2449
Fax: +1.604.684.9314
URL: www.imagistechnologies.com

February 27, 2004

Attn: Mr. Patrick Grother
Face Recognition Vendor Test Committee

RE: Response to Face Recognition Vendor Test 2002 Supplemental Report NISTIR 7083

Imagis Technologies Inc. is pleased to comment on the recent supplemental FRVT 2002 report. In particular, we were encouraged to read the findings of the report indicating that the fusion of multiple FR systems leads to (sometimes markedly) improved scores. These findings validate our strategy of enabling biometric fusion within our Briyante Integration Server, which will allow our customers to implement supplemental biometrics and logic to create streamlined, hybrid systems in a web-service environment.

We were intrigued by the findings of the report indicating that accuracy can depend on ethnicity. The results of tests on the Chinese sub-population were quite interesting. It would be, in our view, highly valuable to try and isolate some of the main physical attributes that contribute to this accuracy improvement. Perhaps the architecture of the Chinese face reduces the effects of surface rendering under changes in light source position. Furthermore, the consistent dark iris, surrounded by a consistent skin tone may also be a factor, since these conditions are quite friendly to any system employing circular gradient detectors over the iris.

Imagis would like to thank the FRVT organizers for highlighting the effectiveness of match score normalization. We appreciate that normalization, implemented at runtime during matching, can be a lengthy operation due to sample mean and std. dev. calculation. However, we expect that normalizing at runtime will not be necessary since there are opportunities to normalize the data earlier in the process. This effectively avoids any runtime speed degradation during matching.

Imagis continues to improve our facial recognition technology and solutions, both in their ability to operate independently, and in their ability to fuse multiple biometrics and disparate data in a Web Service environment.

Regards,

"Andy Amanovich"
Senior Technology Strategist
Imagis Technologies Inc.