Matching Performance for the US-VISIT IDENT System Using Flat Fingerprints

NISTIR 7110

C. L. Wilson, M. D. Garris, &
C. I. Watson

MAY 2004

Matching Performance for the US-VISIT IDENT System Using Flat Fingerprints

C. L. Wilson, M. D. Garris, and C. I. Watson

National Institute of Standards and Technology

ABSTRACT

This report discusses the flat-to-flat matching performance of the US-VISIT fingerprint matching system. Both one-to-many matching used to detect duplicate visa enrollments and one-to-one matching used to verify the identity of the visa holder are discussed. With the proper selection of an operating point, the one-to-many accuracy for a two-finger comparison against a database of 6,000,000 subjects is 95% with a false match rate of 0.08%. Using two fingers, the one-to-one matching accuracy is 99.5% with a false accept rate of 0.1%.

Keywords: fingerprint, flat, IDENT, identification, image quality, matching, performance, ROC, US-VISIT, verification

EXECUTIVE SUMMARY

One-to-Many Matching

The existing IDENT one-to-many matching system has been tested and NIST concludes:

1. Using Department of State (DOS) Mexican visa (BCC) data, the true accept rate (TAR) using index finger pairs is independent of background database size over the range from 100,000 entries to 6,000,000 entries. Using the IDENT operational thresholds of (1300, 1880), the TAR is 96%.

2. The false accept rate (FAR) using index finger pairs is linearly increasing with database size. At a gallery size of 6,000,000, current IDENT operational thresholds of (1300, 1880) achieve a FAR of 0.31%.

3. At the operating level used by the IDENT system, the trade-off between TAR and FAR is such that a large change in FAR results in only a small change in TAR. The trade-off curve is essentially flat with very small slope change. Therefore, NIST recommends IDENT thresholds of (1400, 2025), which achieve a FAR of 0.08% while maintaining a TAR of 95%. This represents a 4-fold reduction in FAR.

4. All the results given here require that the test data be consolidated and checked for correct ground truth by fingerprint examiners. From the original data, 1.7% was found to be incorrectly matched. Approximately 0.14% of the questioned data is of insufficient quality to be resolved by examiners. This 0.14% error rate is the minimum error limit detected in existing government fingerprint databases.

5. The Cogent[1] image quality is a good predictor of the IDENT one-to-many matching performance. The best quality images (quality 1), produce a TAR of 99% at a FAR of 1%. The worst quality images (quality 8), produce a TAR of 53% at a FAR of 1%.

6. Image quality distributions from BCC, the Atlanta pilot study, and Ohio were studied to determine how well the operational US-VISIT system could be expected to track BCC and Ohio results. The Atlanta data has slightly more quality 8 images and slightly less quality 1 images but should result in a TAR near BCC of 95%. The Ohio data has significantly less quality 8 images and significantly more quality 1 images, which is reflected in a TAR of 98% using the IDENT system.

7. The matcher used in this study achieves a match rate of 1,035,000 matches/second with shape filtering turned on, while it achieves a match rate of 734,000 matches/second with shape filtering turned off.

[1] Specific hardware and software products identified in this report do not imply recommendation or endorsement by the National Institute of Standards and Technology. It was necessary to study and report on these products as they were being used at the time in US-VISIT and other relevant programs.

Executive Summary (Continued)

One-to-One Matching

The proposed IDENT one-to-one matching system has been tested and NIST concludes:

1. Results on BCC data demonstrate that verification accuracy significantly improves when matcher scores from right and left index fingers are fused together. Using the simple method of adding together the two finger matcher scores, a threshold of 740 produces a TAR of 99.5% at a FAR of 0.1%. Similar accuracy should be achievable in US-VISIT.

2. These results were achieved using a Software Development Kit (SDK) supplied by Cogent that is the same algorithm used in US-VISIT.

3. Testing of 11 other SDK's proved that this algorithm is as accurate as any of the algorithms tested although further testing of additional algorithms is planned.

4. All algorithms tested have a relatively significant change in accuracy with image quality. The sensitivity to image quality decreases as the TAR of the specific algorithm increases. High accuracy algorithms are less sensitive to image quality than low accuracy algorithms.

5. Cogent image quality is a good rank statistic for all the algorithms tested for all the datasets used. The error rate of the best (quality 1) fingerprints is always lower than the error rate of any other image quality level, and the error rate of the worst (quality 8) fingerprints is always the highest. All other image quality levels result in the expected ordering of the error rates.

6. Consolidation results on various datasets available to NIST demonstrate that the errors obtained for one-to-one matching is less than the clerical error rate in most government databases. Clerical errors will be more common than biometric errors for one-to-one matching.

1. INTRODUCTION

This report discusses the flat-to-flat matching performance of the US-VISIT fingerprint matching system. Both one-to-many matching used to detect duplicate visa enrollments and one-to-one matching used to verify the identity of the visa holder are discussed. Both matching scenarios utilize flat[2] impressions of a person's left and right index fingers that were captured using a live scan device.

1.1 Previous NIST Recommendations

On February 4, 2003, a report titled, "Use of Technology Standards and Interoperable Databases with Machine-Readable, Tamper-Resistant Travel Documents [1]," was submitted to the Congress jointly by the Attorney General, Secretary of State, and NIST. (This report is informally referred to as the "303A Report" and was mandated by the U.S.A. Patriot Act [2] and the Enhanced Border Security Act [3]). It discusses measurements of the accuracy of both face and fingerprints as they relate to U.S. border entry and exit. The recommendations of the 303A Report are as follows[3].

1.1.1 Verification

Verification is defined as a one-to-one match used to decide if the individual (e.g. a traveler presenting a visa) is who they claim to be. The report states, "Our measurements indicate that a dual biometric system including two fingerprint images and a face image may be needed to meet projected system requirements for verification. Each fingerprint and the facial image should require 10 kilobytes or less of storage apiece. Therefore, a card capable of storing two fingerprints and one face image will require a 32K byte chip to fulfill these requirements."[4]

1.1.2 Identification

Identification is defined as a one-to-many match designed to determine if a specified user is in the database. The report states, "To perform background identifications, ten plain image impressions should be used for enrollment and retention. With the live scan fingerprint scanners currently available, the additional time required to capture the additional eight fingers will be insignificant."

1.2 Identification Test Issues

This report addresses the NIST evaluation of verification and identification using two index fingers. The primary test dataset used is based on fingerprints collected by the State Department

[2] The term "flat" is used to define a single finger plain impression.

[3] The conclusions of that report should be updated in light of NIST's recent findings that the VTB fingerprint matcher is substantially less accurate than commercial systems

[4] This result is modified in section 3.3 of this document.

(DOS) as part of the Mexican visa program. This data is referred to as BCC (border crossing card) data. These fingerprints were collected with single finger live scan readers. The total database size is approximately 6 million individuals. Of these, approximately 274,000 have one or more matching sets of fingerprints. The matching fingerprints were detected by matching names and geographic locations in the dataset. If fingerprint matching had been used, this would have biased the dataset to a particular fingerprint system.

1.2.1 Reliability of Test Data

Since the test data may contain examples of both unknown matches and incorrect matches, it was necessary that trained fingerprint examiners check all questionable candidate matches. This is discussed in more detail in Section 2.2.

1.2.2 Total Matching Scale for Identification

To accurately measure the false accept rate (FAR) of commercial automated fingerprint identification systems (AFIS), large numbers of probes need to be applied to a background gallery size in the millions. The results of experiments reported herein used a probe set of 60,000 finger pairs searched against a background gallery of 6 million. This resulted in a test that required 720 billion raw matches when using two index fingers. Filtering was used to reduce the fraction of the gallery that was actually matched by approximately 30%.

This demonstrates that special computer hardware matchers or special software implementations are essential for large scale AFIS testing. The commercial system used for these tests had a peak measured match rate just over 1M matches/second. This resulted in a total required CPU time for the primary test of approximately 504,000 seconds or around 6 days. The matcher test was run in 10 segments, each designed to take less than one day. Minutiae extraction was performed on a different array of computers and required an additional week of CPU time using 32 3GHZ processors.

1.3 Verification Test Issues

For verification testing, the required scale of testing is much smaller. Verification tests involved matching 6,000 probes against 6,000 gallery images. This resulted in a total of 36M matches and allowed the use of much slower matching software and slower computer systems. The matcher used for verification testing in this report was rated at 200 matches/second on a 3Ghz processor. This yielded a test time of 180,000 seconds or 50 hours. Testing on the 12 different datasets discussed herein took about 24 days.

2. DATA QUALITY ISSUES

The two primary data quality issues are image quality and the correctness of the identified matches used in the test. The test datasets used were derived from operational government databases. This means that there were variations in image quality due to different operating, sensor, and environmental conditions and that some of the matched fingerprint sets used in the test had duplicate matches or incorrect matches that had to be checked and corrected.

2.1 Effect of Image Quality on Commercial Systems

The IDENT matchers tested in this study were developed and marketed by Cogent Systems for use in the US-VISIT program. Cogent also provides software for the calculation of fingerprint image quality. Other commercial image quality algorithms are currently being tested at NIST. The Cogent image quality measure (IQM) is based on a scale from 1 to 8 where 1 is the best quality value and 8 is the worst image quality value. These quality values are interpreted operationally in the US-VISIT program as: 1-4 good but decreasing quality, 6-7 average quality, and 7-8 poor quality.

The IQM for three different datasets were computed as part of this study. These datasets were BCC, Ohio, and the Atlanta US-VISIT Pilot. The BCC database is discussed in more detail in the VTB report [4]. The Ohio database is discussed in detail in the ATB report [5].

A comparison of the Atlanta and BCC data IQM distributions is shown in Figure 1. As will be seen, the Cogent image quality is a good predictor of matcher performance. The strong similarity of the IQM functions between the Atlanta data (red '+'s) and the BCC probe and background curves (green 'x's and blue '*'s respectively) indicates that BCC data can be used as a proxy when predicting matcher performance for the data expected from US-VISIT.

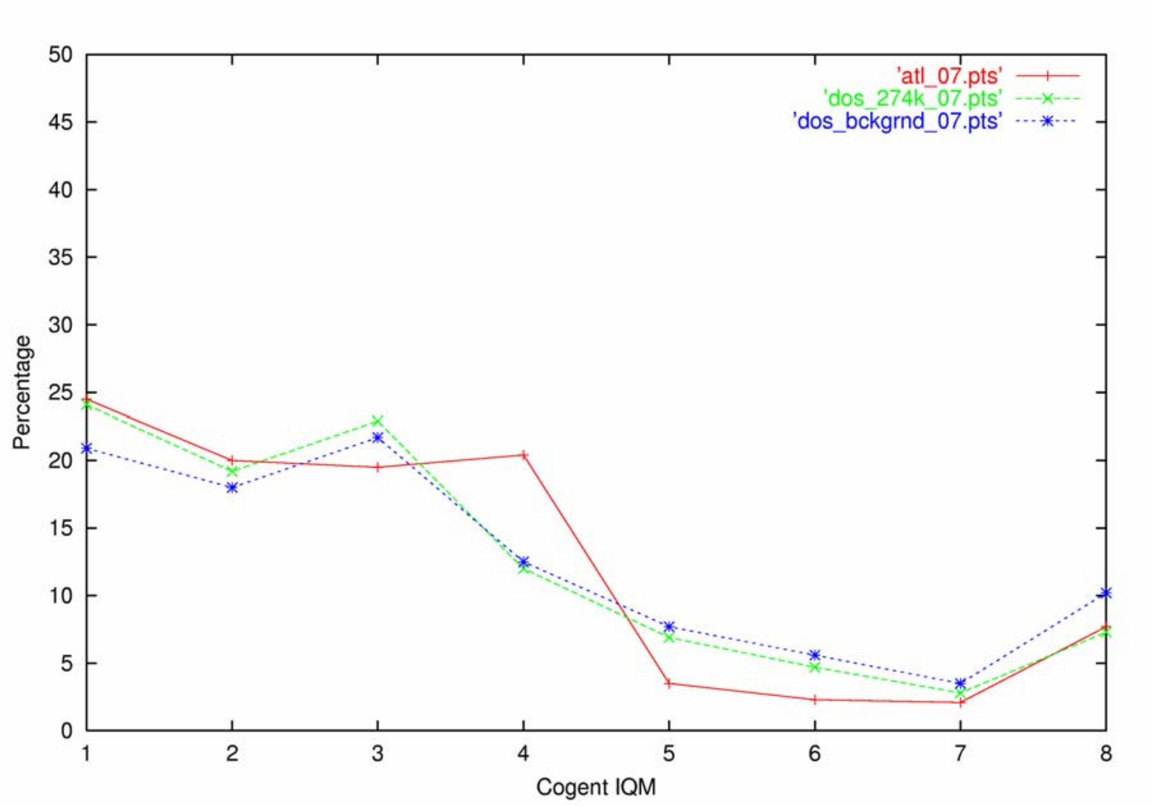

Figure 1. Cogent IQM distribution for left index fingers of BCC and Atlanta US-VISIT Pilot data

The Cogent IQM distributions for the Ohio data are shown in Figure 2. The IQM distribution for this data is significantly better than that shown in Figure 1. An increase in the fraction of quality

1 fingerprints from 25% in BCC data to 45% in Ohio data, and a decrease in the fraction of quality 8 data from 10% to 3%, can be seen by comparing the two figures. This IQM improvement is reflected in improved matching performance as tests discussed below will show.

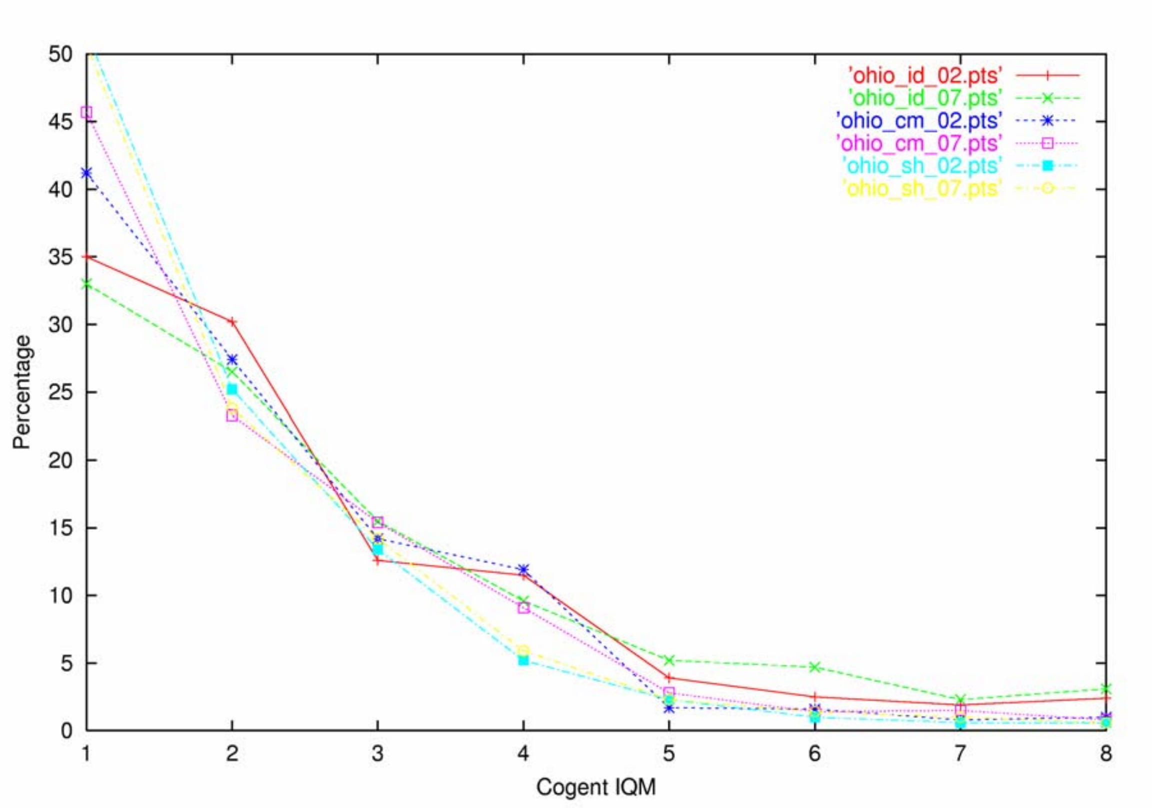

Figure 2. Cogent IQM distribution for left and right index fingers for three different fingerprint readers from Ohio data

2.2 Consolidation – Correctness of Matches

A critical component to the validity of a biometric test is the reliability of the data used. The tests conducted on the one-to-many IDENT matcher are designed around the assumption that a probe (a fingerprint used to search the background) has one, and only one, mate in the background. Given this assumption, successful matches and mismatches can be measured based on whether the mate was found in the background or not.

The data used to test the IDENT matcher came from operational data, and therefore contains image capture errors (such as swapped or substituted fingers), clerical errors (such as an individual being assigned more than one ID), and image quality issues that can render a fingerprint not verifiable due to a lack of coverage, smudging, occlusions, etc. To account for these sources of potential error and to support the assumption that each probe has one, and only one, mate in the background, a process called "consolidation" was developed.

2.2.1 The Consolidation Process

The goal of consolidation is to comb through large volumes of data, looking for a relatively small number of uncertain cases that require corrective action. A fingerprint matching system can be used to help pinpoint these cases. Actually, the system is used to determine which cases are clearly correct. The remaining "questionable" cases are then subject to review, in this case by professional fingerprint examiners.

The consolidation process is broken down into three major steps:

1. Generate review list of questionable cases
2. Submit these cases to human review
3. Consolidate cases as determined by human review

2.2.1.1 Generate Review List

To generate a list of questionable cases for review, a set of fingerprints wanting to be consolidated are searched against the background using a matching system. The result is a candidate report, listing for each probe all the likely candidates matched to the background. This report includes the probe's ID, the candidate's ID, and a match score. Questionable cases are represented by their pair of (probe, candidate) IDs and are identified according to the following two conditions:

1. If a probe's *alleged* mate is not reported as a candidate

   then review the probe and its *alleged* mate

2. If a candidate (not claimed to be the mate) is reported with too high a match score

   then review the probe and the *alleged* non-mate candidate

The second condition requires a parameterized threshold to be set in order to determine when a candidate's score is sufficiently high enough to raise suspicion as to whether the candidate may in fact be a mate. To determine this threshold, a training set of probes is searched by the matcher against the background, and a candidate report is produced. A distribution of all alleged mate scores is plotted and compared against the distribution of all alleged non-mate scores. The right tail of the non-mate distribution represents questionable cases that may very well be mates.

At this point consideration must be given to the capacity for which human reviews can be conducted. The lower the threshold is set, the more thorough the consolidation process will be, but at the expense of increased cases to be reviewed. The slope of the non-mate distribution is steep so that at a certain threshold point, even a small decrease in threshold will result in an overwhelming number of reviews to be generated. In the other direction, increasing the threshold will decrease the number of human reviews but increase the chances that an actual consolidation will be missed.

For the consolidations of the BCC data used in this report, a training set of 6K random probes was used to set the threshold. It was determined that, given the capacity for human reviews and the timeframe in which these reviews needed to be conducted, that a threshold should be chosen that generates a review list approximately one tenth the size of the number of probes wanting to be consolidated. So from the training set of 6K probes, a combined two finger matcher threshold of 1400 (Cogent Systems specific) was selected, generating a review list of approximately 600 cases.

Given this threshold, ten different random sets of 6K probes were then matched, candidate reports compiled, and ten sets of questionable review cases were generated for human review. Of the cases reviewed, approximately 30% checked alleged mates, while 70% checked alleged non-mates.

2.2.1.2 Human Review

NIST was provided contracted time from two full-time professional fingerprint examiners retired from the Federal Bureau of Investigation (FBI). These examiners inspected the images of the fingerprints associated with each case. The BCC and US-VISIT programs are based on capturing a traveler's two index fingers. So each case reviewed involved potentially inspecting two pairs of fingerprints, the probe and candidate right index fingers, and the probe and candidate left index fingers.

The examiners used the FBI's Universal Latent Workstation (ULW) [6] for viewing and comparing fingerprint pairs. This workstation includes fundamental image enhancement tools for controlling and changing brightness, contrast, scale, and rotation which greatly aided the examiners in making their comparisons.

The examiners were required to make their best professional judgment on each case reviewed. Comparing the prints between a probe and its candidate, they were asked to make the judgment of match, no-match, or reject. A match was to be decided if there was sufficient evidence to legally make an identification. A no-match was to be decided if there was sufficient evidence to say the probe and candidate did not match. A reject was to be decided if there were sufficient problems with the fingerprints to prevent either of the first two judgments. Categories of reject include not verifiable, finger positions switched, and one pair of fingerprints match while the other does not match.

For one set of 6K probes, 6 (0.1%) alleged mates were determined to not be mates; 99 (1.65%) alleged non-mates were determined to be mates; and 23 (0.38%) cases were rejected, 21 of which were determined to be not verifiable. In the process of developing the consolidation process documented in this section, an alternative method was studied. The alternative identified a similar set of consolidations with the exception of two cases. This lends evidence that the process deployed detects a majority of the consolidations residing in the BCC data. A rate of 2 in 6000 indicates data integrity between 10^{-3} and 10^{-4}.

2.2.1.3 Consolidate Data

The examiners pass judgments case by case without any knowledge of whether they are validating an alleged mate or an alleged non-mate. As can be seen from the statistics above, the majority (98%) of the cases reviewed actually confirmed BCC data records and required no action. In general, the logical conditions requiring corrective action are:

1. If probe and alleged mate reviewed

   and determined to be no-match or reject,

   then remove probe from search set

2. If probe and alleged non-mate candidate reviewed

   and determined to be match or reject

   then remove candidate from the background

The purpose of consolidations is to improve the reliability of data used in testing and thereby increase the integrity of any test results. This impact is observed in that the complete 60K BCC test set originally achieved a TAR of 94.9% at a FAR of 1.5%. Applying consolidations, this same test set achieved a TAR of 95.2% at a FAR of 0.09%, when the same system thresholds are applied. This represents a nearly 17-fold reduction in FAR. This demonstrates how important reliable data is to making sound engineering and policy decisions.

3. VERIFICATION PERFORMANCE

The verification performance of the IDENT system was tested by evaluating the single finger performance of the Cogent "feedback" matcher on a randomly selected sample of 6000 BCC subjects. A full similarity matrix of 36M matches was computed for the matching of right index fingers, and a second similarity matrix was computed for left index fingers. The IDENT system uses both index fingers for one-to-one matching. Simple methods for combining the match scores of right and left index fingers were evaluated to predict this performance.

3.1 Single Finger Performance

3.1.1 BCC

Figure 3 and Figure 4 show the receiver operator characteristic (ROC) curves for BCC data using the Cogent feedback one-to-one matcher. The right and left index finger performances are significantly different. At a FAR of 0.01%, the TAR for the right finger is 97.5 %, and for the left finger, the TAR is 95.3 %. This performance difference has not been explained. Consolidation makes a significant difference in the right finger results but does not change the left finger results for FARs less than 0.3 %.

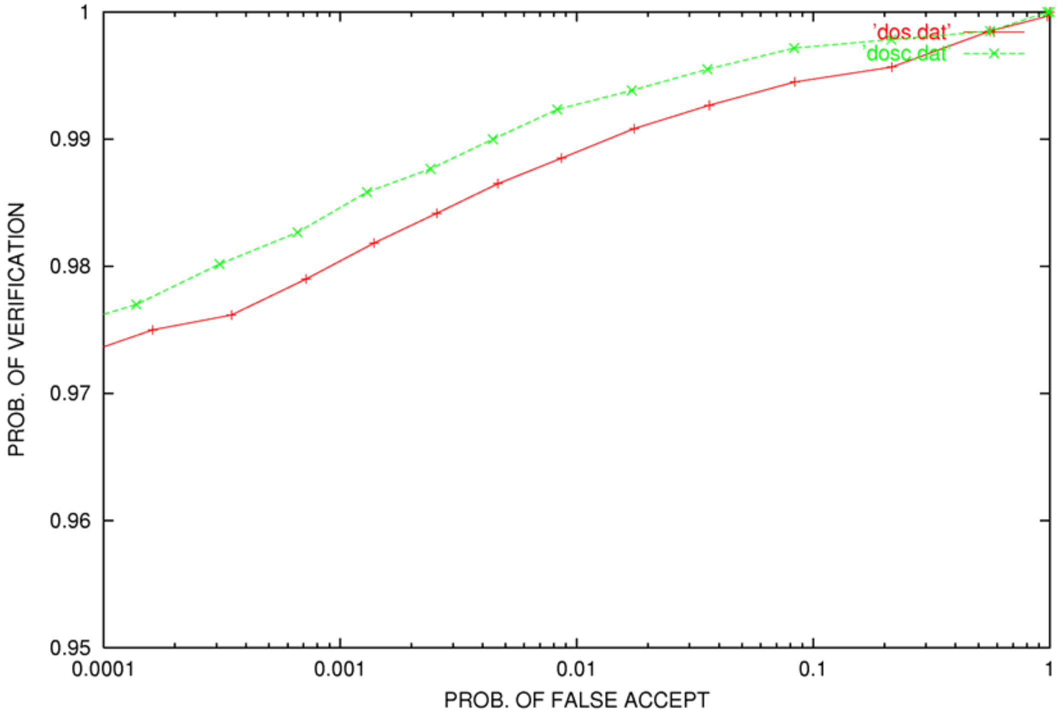

Figure 3. ROC plot for BCC right index fingers using consolidated and unconsolidated data

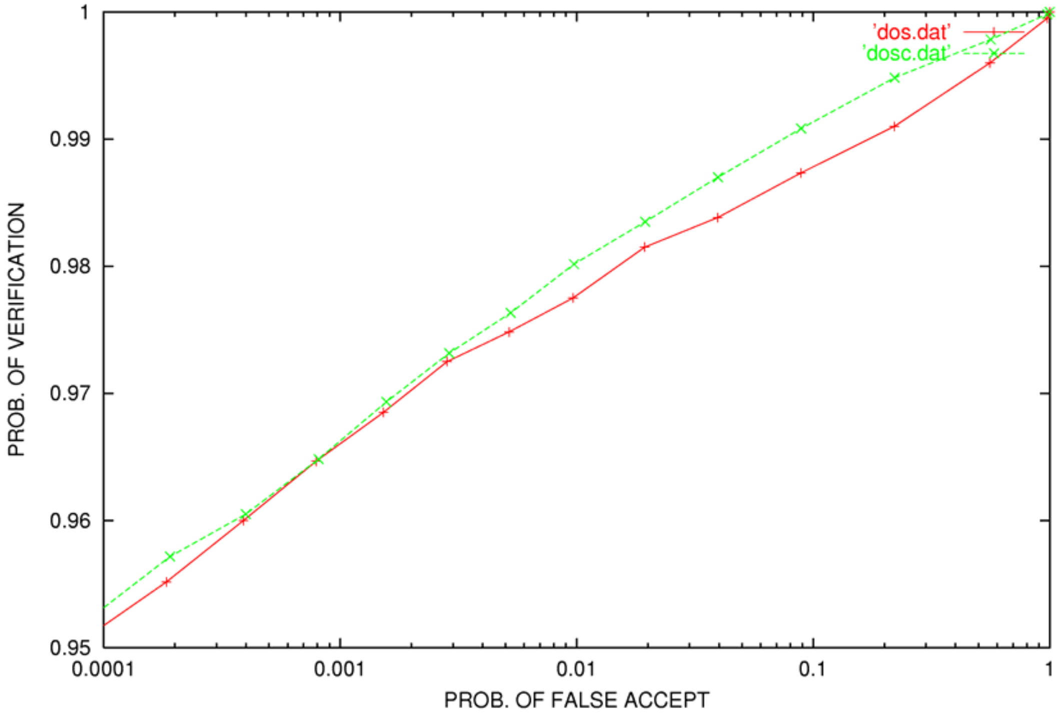

Figure 4. ROC plot for BCC left index fingers using consolidated and unconsolidated data

3.1.2 Other SDK Test Issues

The complete results of the SDK (Software Development Kit) tests are the subject of another report, but certain observations from these tests can be used to put the results in Figure 3 and Figure 4 in perspective. Figure 5 shows ROC curves from eight different algorithms on twelve different single finger datasets. The vertical scale of the plot has been expanded from 1.0 to 0.95 in Figure 4 to 1.0 to 0.7 in Figure 5 to cover a much wider range of results. This figure shows that for a range of FARs between 0.01% and 100%, different combinations of data and algorithm give most of the possible results that fit on the graph. If you try enough algorithms and enough datasets you can get as wide a range of results as are possible.

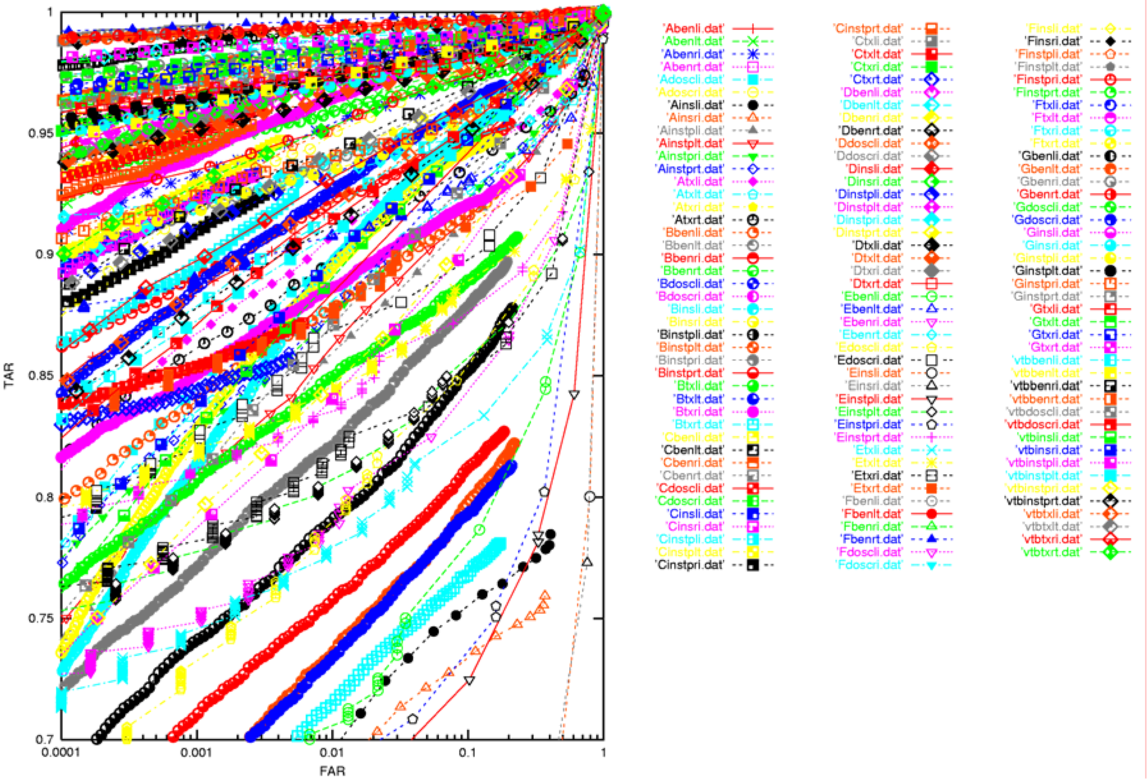

Figure 5. ROC plots from the SDK test of eight algorithms and 12 different datasets

Figure 6 shows the data subset from Figure 5 that characterizes the Cogent feedback matcher. Except for the two most difficult datasets, the TARs at FARs above 0.01 % are all greater than 95 %. This matcher is one of the three best tested in the SDK test yet it still retains a large sensitivity to image data quality.

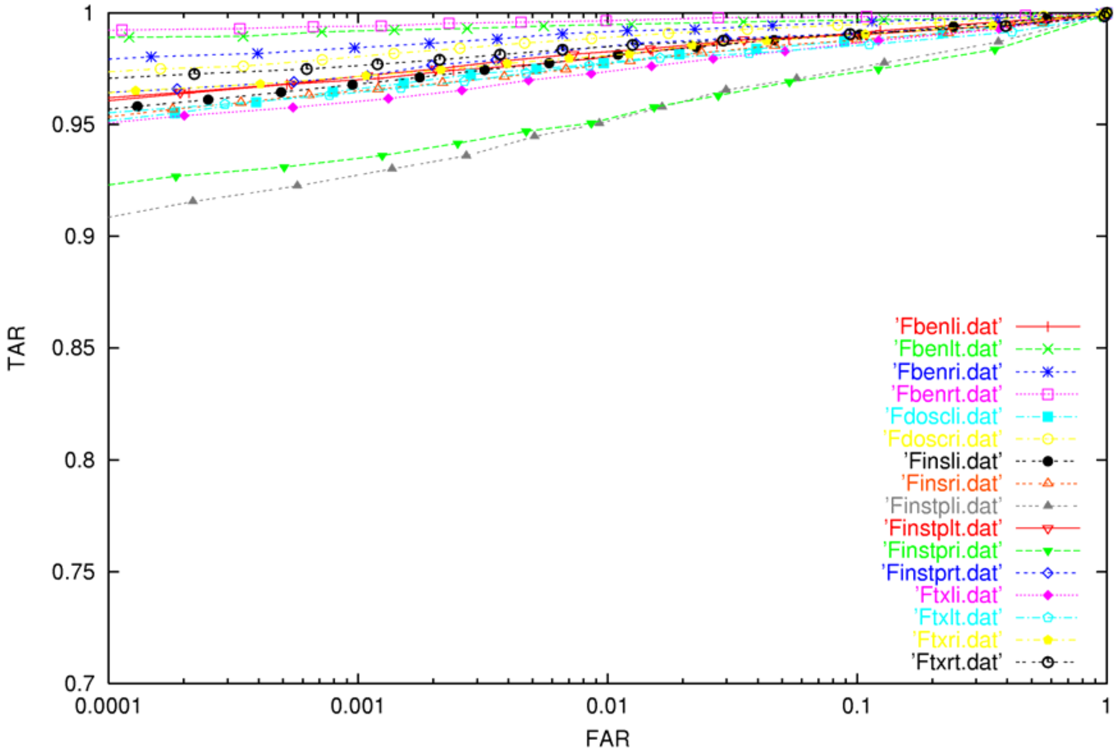

Figure 6. ROC plots from the SDK test for the Cogent feedback matcher on 12 different datasets

3.2 Methods for Combining Fingers

The results shown in Figure 3 and Figure 4 need to be combined to produce a single match / non-match decision when two pairs of index fingers are compared. These decisions are achieved in IDENT by combining two single finger scores. The relationship between these two scores is shown as a scatter plot in Figure 7. Figure 7 also indicates the range of scores that can result in ambiguous decisions by a pair of bounding lines.

3.2.1 Adding Single Finger Scores

The simplest way to combine scores is to add match scores from both fingers and apply a threshold in the range indicated in Figure 7. Using this method, a threshold of 740 produced a TAR of 99.5% at a FAR of 0.1%. This is near the center of the decision region in Figure 7. The resulting ROC cure for the combined two fingers is shown in Figure 8.

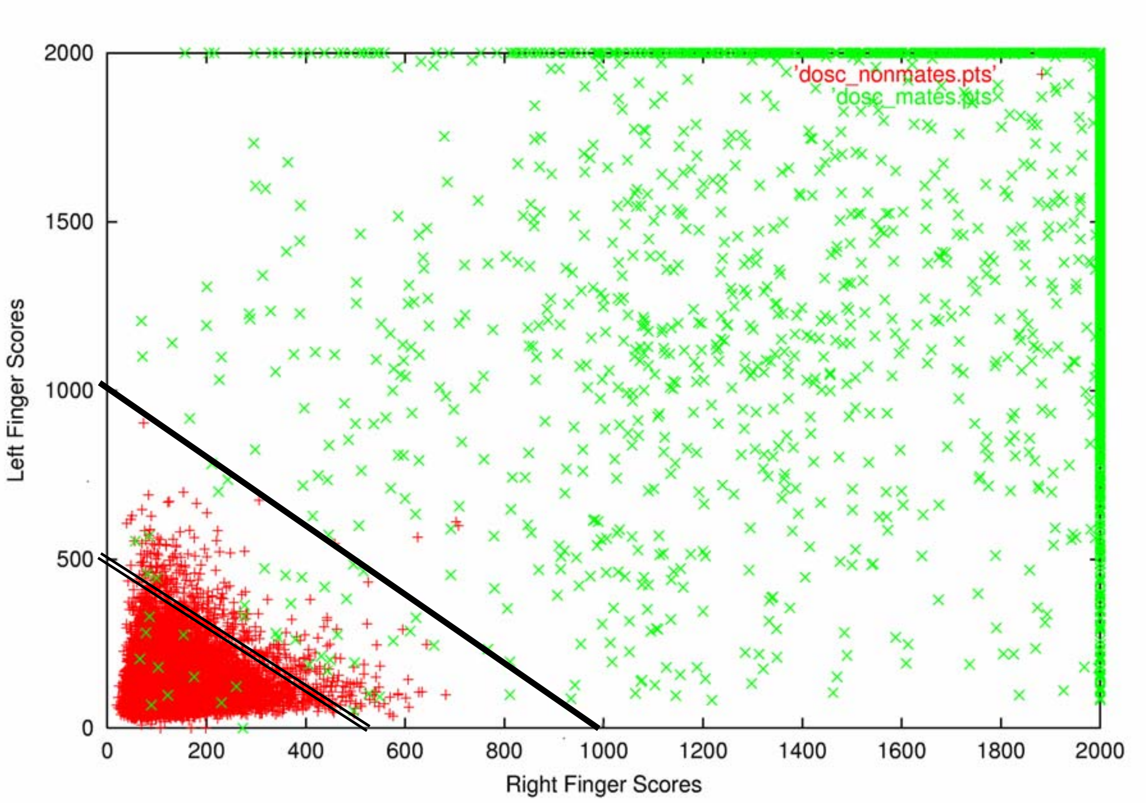

Figure 7. Scatter plot for BCC right and left index fingers using consolidated data

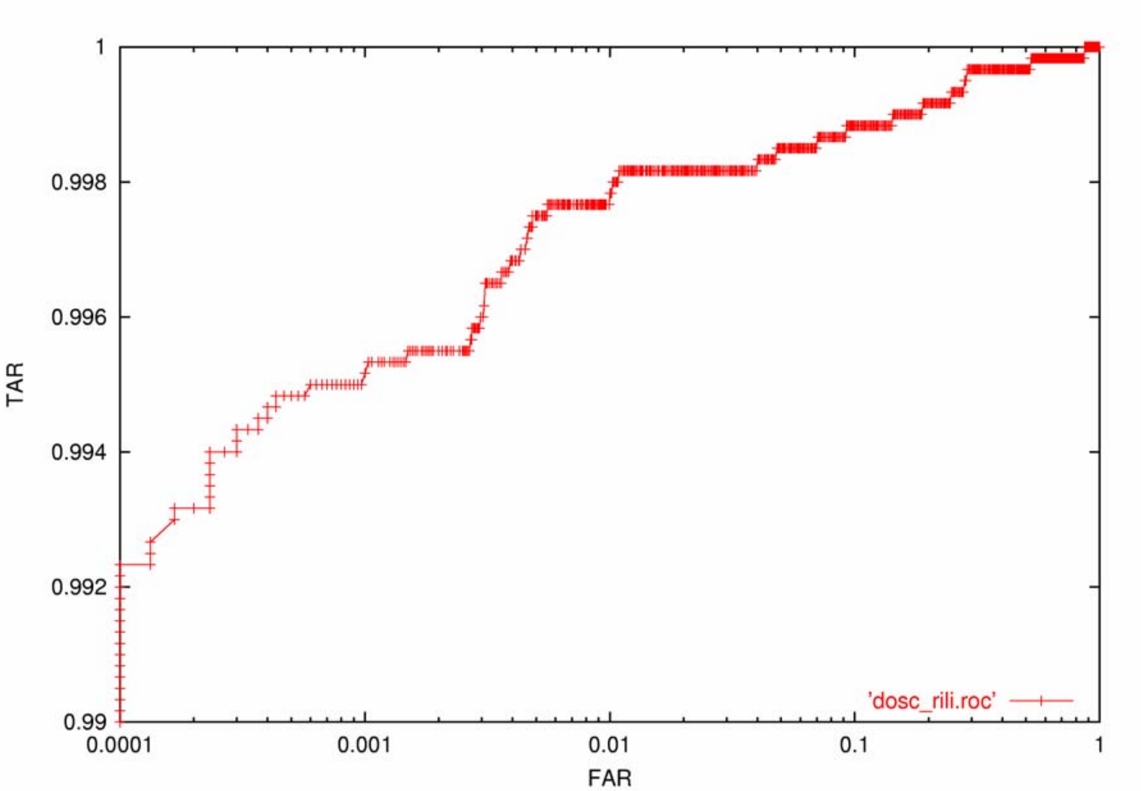

Figure 8. ROC plot for BCC right and left index fingers combined by adding scores

3.3 Comparison with Face Recognition

The report that was sent to Congress [1] as part of NIST PATRIOT Act mandate [2, 3] compared face recognition results from the FRVT 2002 study [8] with single-finger studies conducted using the NIST VTB fingerprint system [4]. The conclusions of that report should be updated in light of NIST's recent findings that the VTB fingerprint matcher is substantially less accurate than commercial systems, and because the INS2 data used for [1] has lower quality than the BCC data used here. In addition, the images used in the FRVT 2002 test are of higher quality than those in operational government data sets.

When all these factors are combined, the comparison of face and fingerprint accuracy needs to be substantially revised. Contemporary fingerprint systems are substantially more accurate than the face recognition systems tested in FRVT 2002. This conclusion holds even for face and fingerprint images categorized as high quality. This conclusion should be qualified by the observation that any advances in face recognition technology, since the FRVT test, have yet to be evaluated.

The two-fingerprint accuracy at 1% FAR discussed here is 99.8% while the best FRVT 2002 face result at 1% FAR was 90% using controlled illumination. When outdoor illumination was used in FRVT 2002, the best TAR at 1% FAR was 54%. Even under controlled illumination, which is not used in US-VISIT, the error rate of face is 50 times higher than the two-fingerprint results discussed here. If the case of uncontrolled illumination is considered, this factor would be 250. This means that face recognition is useful only for those cases where fingerprints of adequate quality cannot be obtained.

4. IDENTIFICATION PERFORMANCE

The primary purpose of the work behind this report is to certify the biometric technologies used in US-VISIT. US-VISIT, as a process, can be viewed as a multi-staged workflow, involving a procedurally dynamic interaction of the traveler with US Custom and Border Protection inspectors, and biometric acquisition (face and index fingerprints) with behind the scene one-to-one fingerprint verification, one-to-many fingerprint identification, and human fingerprint verification by professional fingerprint examiners. In this section, the identification performance (or the accuracy of the one-to-many search) is reported.

The following method for measuring identification performance was developed to better reflect specific decision points within US-VISIT. In this section, the true accept rate (TAR) is defined as:

Given a set of probes known TO BE IN the background,

TAR = (# probes hit / # probes searched)

where a hit is defined when a probe's mate is reported on the

candidate list.

whereas the false accept rate (FAR) is defined as:

Given a set of probes known NOT TO BE IN the background,

FAR = (# probes hit / # probes searched)

where a hit is defined as anything reported in the

candidate list.

All TAR and FAR results reported in this section have been computed according to the above definitions, and all results are based on the consolidated BCC dataset except where otherwise explicitly noted.

4.1 Threshold Performance

Cogent developed a one-to-many fingerprint matching system called the Elite, which uses special purpose hardware to achieve an effective throughput of 1M matches per second. One of these systems was sent to NIST for testing, and it was configured to be consistent to Elite systems which were at that time being used operationally in US-VISIT. Lockheed Martin (LMCO) is the integrator for US-VISIT and has engineered a workflow around the Cogent matchers in order to support the processing of visa carrying travelers. LMCO, in coordination with Cogent, has selected an operating pair of matcher score thresholds; the first is a single finger threshold; the second is a two finger threshold. These are referred to as the IDENT thresholds. NIST conducted a study to measure the impact that changing these thresholds have on performance and to validate the current operational settings.

Recall that a potential match for US-VISIT involves two pairs of fingerprints, one pair from the probe and the other pair from the candidate reported from the background. Matching the left index fingers to each other and matching the right index fingers to each other result in separate matcher scores. If the maximum matcher score derived from either the right or left index fingers exceeds the single finger threshold, then the candidate is considered a match to the probe. If the sum of the two matcher scores exceeds the two finger threshold, then the candidate is considered a match to the probe. These two threshold conditions are combined using a logical OR, so that only one of the two conditions need be met to be considered a match.

The FAR generated by the system is reduced as the thresholds are increased, but this comes as a cost in decreased TAR. To date, a threshold pair of (1300, 1880) has been used operationally in US-VISIT, where 1300 is the single finger threshold, and 1880 is the two finger threshold. Using these thresholds on Elite output from the 60K set of BCC probes, a TAR of 95.9% is achieved with a FAR of 0.31%. The question arises, might a different threshold setting achieve even more desirable results? To answer this question, a range of thresholds was studied to determine the expected TAR at a FAR of 0.1% and what operational thresholds would give this level of performance.

The results from a range of surveyed thresholds are shown in Figure 9. The results reported were derived from searching the same 60K consolidated probe set against the 6M BCC background, and applying a series of different thresholds. Two points are labeled in the plot. The first corresponds to the current operational point produced by the thresholds (1300, 1880). As can be seen, the second point produced by the thresholds (1400, 2025) yields a FAR of 0.08% with a TAR of 95.3%. Based on these results, NIST recommends incrementally adjusting the US-VISIT operational IDENT thresholds to (1400, 2025).

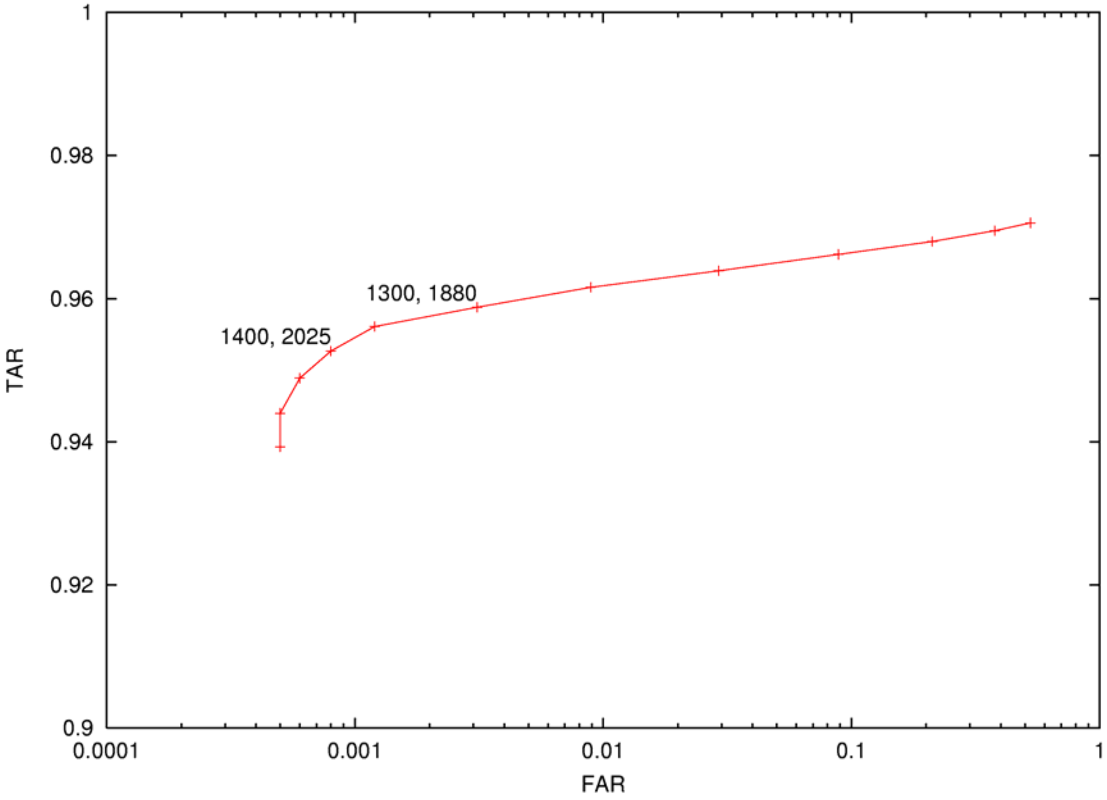

Figure 9. FAR vs. TAR response curve across a range of thresholds for 60K BCC probes

4.2 Shape Filter Performance

Another significant parameter, in addition to the IDENT thresholds, that stands to impact performance is shape filtering. When turned on, this feature conducts a quick pattern-based comparison (looking at coarse ridge flow and structure) and drops unlikely candidates from the search list and subsequent (slower) matching. The goal is to quickly and accurately reduce the number of matches made across the background, thus speeding up the throughput of the overall system. A trade-off is introduced whenever filtering, such as this, is used. The quick decision to drop candidates from the search list introduces the possibility that a mate in the background may be mistakenly dropped, in which case a probe that should hit the background is incorrectly missed.

Experiments were conducted to analyze the impact of shape filtering on the Elite. In this case, a representative sample of 6K probes from BCC was used. The probes were searched against the consolidated background of 6M, the first time with shape filtering turned off, the second time with shape filtering turned on. Two sets of IDENT thresholds were then applied to the resulting candidate lists; the current operational thresholds of (1300, 1880) were applied as well as the NIST recommended (1400, 2025). Resulting TAR, FAR, and effective throughputs are reported in Table 1.

	Thresholds 1300, 1880		Thresholds 1400, 2025		
Shape Filter	FAR	TAR	FAR	TAR	Matches per Second
Off	0.30%	96.3%	0.07%	95.6%	734K
On	0.32%	96.1%	0.07%	95.5%	1035K

Table 1. Effect of shape filter on performance

Looking at the table, first notice the difference in performance between the two sets of thresholds. Changing thresholds from (1300, 1880) to (1400, 2025) reduced FAR by more than a factor of 4 with a reduction in TAR of about 0.5%. The TAR and FAR reported here for the 6K probe set, using (1400, 2025), are comparable to the 60K results plotted in Figure 9, where TAR is 95.2% and FAR is 0.09% for the same threshold settings.

Now look at the differences in performance when shape filter is turned off to when shape filter is turned on. For thresholds (1300, 1880), there is a small decrease in TAR and a negligible increase in FAR. For thresholds (1400, 2025), there is a negligible decrease in TAR and no change in FAR. While little change is observed between shape filtering being on or off, there is a significant different in effective throughput. The Elite achieves a throughput of 734K matches per second when shape filter is off, while it achieves a throughput of 1035K matches per second when shape filter is turned on. It is concluded from these results that the throughput gained by using shape filter far exceeds any trade-off in the performance of TAR and FAR.

4.3 Trading FRR for FAR

Figure 9 and Table 1 demonstrate how TAR and FAR respond as the IDENT thresholds are changed. As thresholds are increased, TAR decreases (which is bad) while FAR decreases (which is good). Given the definitions of TAR and FAR used in this section, the false reject Rate (FRR) is defined as (1 – TAR), so as IDENT thresholds are increased, FRR increases while FAR decreases. This means there is an inverse relationship, or trade-off, between the achievable levels of these two types of system errors.

To help understand the difference between FRR and FAR, we can use the example of a watch list application. A false reject occurs when a person *known to be* on the watch list presents his fingerprints to the biometric system but is not correctly identified. There are two conditions under which this can happen. Either the system remains silent and does not return any candidates, or the candidates returned do not include the person's mate. In the first case, when the system returns no candidates, the person will pass on through primary inspection. In the second case, the person is redirected to secondary inspection while the candidates reported are reviewed. Because the person's mate is not on the candidate list, no identification can be made, and the person will be permitted to pass on through.

In the same watch list application, a false accept occurs when a person *known not to be* on the watch list presents his fingerprints to the biometric system, and the system reports back a candidate list. In this case, an innocent traveler is unnecessarily detained and inconvenienced rather than being permitted to pass on through.

For an enrollment control application, a one-to-many search is conducted to determine if a person is already enrolled in a system. If they are, then the current encounter is associated with the account already on file. If the person is not already enrolled (in other words it is the person's first encounter with the system), then a new account is generated and associated with the person on future encounters. A false reject occurs when a person (known to be enrolled), presents his fingerprints to the biometric system, and the system either does not return any candidates, or the candidates returned do not include the person's mate. In this case, a new account is mistakenly generated for the person (who is already enrolled in the system) and now the person has two accounts in the system. A false accept occurs when a person (known to not be enrolled), presents his fingerprints to the biometric system, and the system reports back a candidate list. In this case, the current encounter of a person (new to the system) may be mistakenly associated with another person's account. The probability of this happening is reduced if the candidates are reviewed by a human fingerprint examiner[5].

It is important to understand that these two types of errors (FRR & FAR) are traded off against each other, and policy makers should consider this when determining acceptable levels of performance, and engineers should consider this as they are setting thresholds to achieve these specified levels. It is also important to note that in testing an operational system it is only possible to measure FAR (the mistakes made in alleged identifications). It is not possible to measure FRR, as it is impossible to determine who the system has missed (those mistakenly granted access). This is a significant reason why testing biometric systems on seeded/consolidated backgrounds is necessary to certify biometric performance.

4.4 Performance vs. Background Size

A very important question regarding biometric system is how well do they scale up. In other words, how is performance impacted as the size of the background is increased? The Cogent Elite results from searching the 6M consolidate BCC background with 60K probes were analyzed to determine the impact of background size.

One way to conduct this study is to set up incrementally increasing backgrounds on the identification system, and for each size, resubmit the probe searches and record the results.

[5] This is the current design of the US-VISIT system

While this is a very straight forward approach, it is very costly in terms of system set-up time and search time, taking weeks if not months to complete.

An alternative approach was developed whereby a single candidate report was generated by searching the entire 6M BCC background once with the 60K probe set. Search results were then computed for a series of simulated decreasing backgrounds. Non-mate records in the background were randomly selected and their corresponding match results reported as candidates in the candidate report were removed. Performance on the augmented candidate report was then conducted and recorded based on the definitions of TAR and FAR above.

Figure 10 shows how FAR relates to gallery size. In this study, the gallery size was incrementally reduced by half, starting at 6M and continuing down to where the only records remaining in the gallery were the 60K mates to the probe set. There are two curves in the figure. The top, red curve corresponds to the performance achieved with the current operational IDENT thresholds (discussed above) set to (1300, 1880). The bottom, green curve corresponds to the performance achieved with the NIST recommended thresholds of (1400, 2025). Note that in both cases, there is a linear relationship between gallery size and FAR. Also note that at a gallery size of 6M, there is 4-fold difference in FAR between the two threshold settings. It should also be noted that there was no measured changed in TAR across the range of gallery sizes sampled.

From these results, it is concluded that for fingerprint identification systems, such as the Cogent Elite, TAR remains constant as the gallery size increases up to 6M, while FAR increases linearly as the gallery size increases. Further tests will be completed to shows the effects of greater gallery sizes.

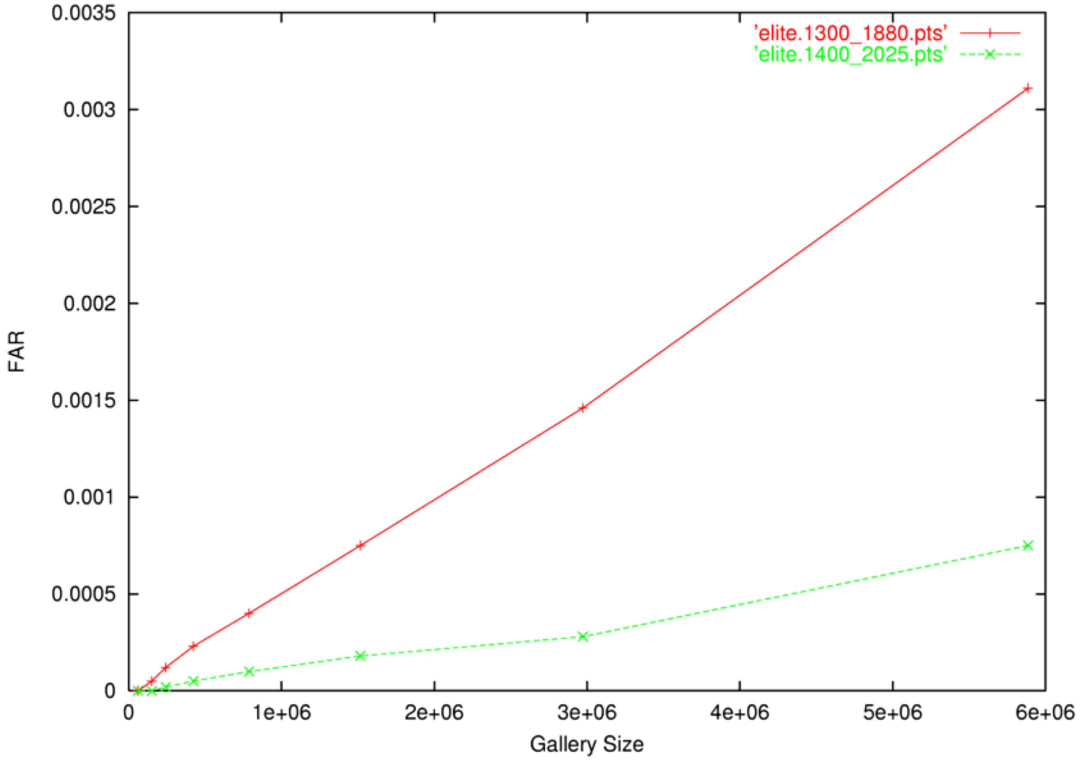

Figure 10. Impact of gallery size on Elite performance (FAR)

4.5 Effect of Image Quality

4.5.1 BCC Study

One of the fundamental conclusions of this report is that the most critical single factor impacting the performance of fingerprint recognition systems is image quality. If a reasonably engineered system is performing poorly, look at the quality of the image being processed by the system. If you desire to improve performance of the system, analyze and determine what steps can be taken to improve image quality. Improvement in image quality will translate into improvement in system performance.

As mentioned earlier, Cogent has developed a proprietary method for calculating image quality called an image quality measure (IQM). The method computes a quality value on the scale from 1 to 8, where 1 is the best quality and 8 is the worst. Using this method, an IQM was computed for each image used in the BCC studies presented in this report. Quality distribution profiles were generated and compared. One such quality distribution analysis is plotted in Figure 11.

Three different distribution curves are plotted in this figure. The first curve (with red '+'s) plots the quality distribution for all 274K right index probes and their mates in the BCC dataset. The second curve (with green 'x's) plots the quality distribution for all the left index probes and their mates. Comparing these two curves at IQM levels 1 and 2, it is observed that right index fingers tend to be of higher quality than left index fingers. To date, there has been no firm explanation for this observed difference between the fingers.

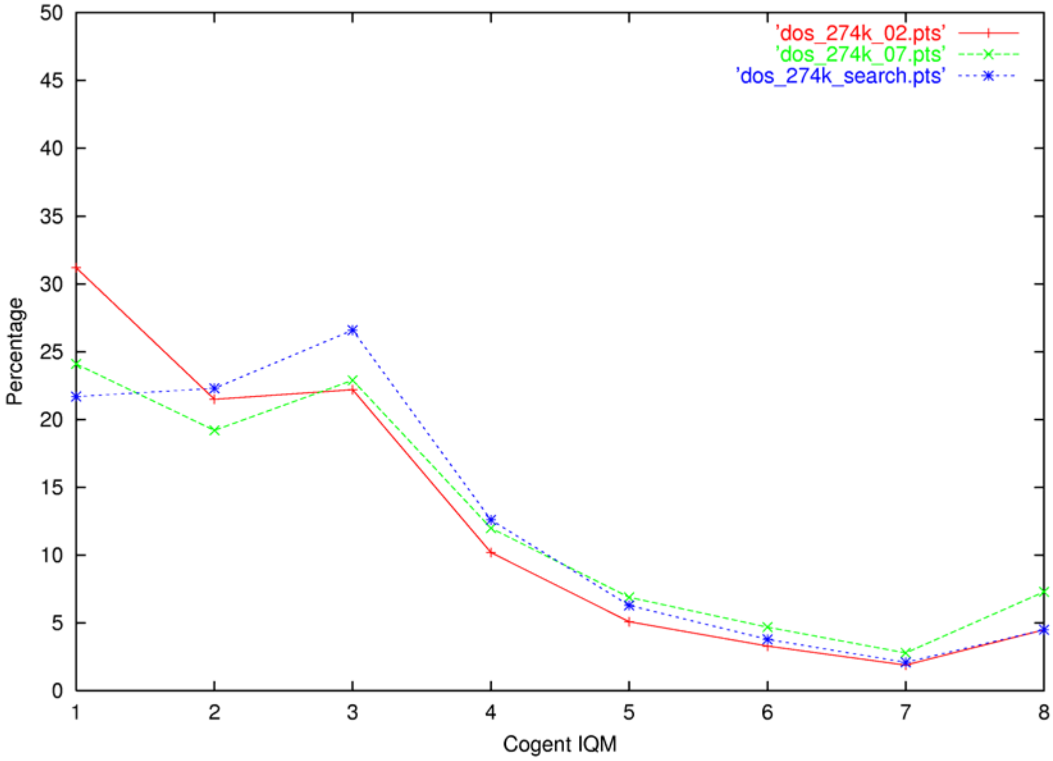

Figure 11. Cogent IQM distributions for 274K BCC right and left index fingers used as probes and their mates and their combined search quality distribution

Now turn your attention to the third curve (with blue ‘*’s) in the figure. The Cogent systems used in US-VISIT are set up as two finger systems. This means the background contains, and is searched with, one right index fingerprint and one left index fingerprint per person. As a result, there are four fingerprints directly involved when making a true match in the system. There is the person’s pair of right and left index fingerprints used as the probe, and there is the person’s mated pair of right and left index fingerprints stored in the background.

The quality of these four fingerprints will greatly dictate whether the system will be successful in matching these prints together. The right index probe is matched against right index fingers in the background. The success of matching the right index probe to its mate in the background will be greatly limited by the fingerprint of worse quality. For example, even if the probe has an image quality of 1, but if the mate in the background has an image quality of 8, there is very little chance the pair will strongly match each other. The same is true when matching the left index probe to its mate in the background. Based on this relationship, the image qualities of the four fingerprints (the probe pair and the corresponding mate pair) are combined to represent the image quality of the probe when matched. This combined quality is referred to as the “search quality.”

Search quality S is calculated as:

```
    Given right and left probe index fingers ($r_p$, $l_p$)
        corresponding IQM values are (qual($r_p$), qual($l_p$))
    Given right and left mate index fingers ($r_m$, $l_m$)
        corresponding IQM values are (qual($r_m$), qual($l_m$))
    $S$ = min(max(qual($r_p$), qual($r_m$)), max(qual($l_p$), qual($l_m$)))
```

This formula begins with two pair-wise comparisons between the right probe and right mate, then left probe and left mate, recording for each pair the worse quality value. This represents the limiting quality within right index finger matching and the limiting quality within left index finger matching. These limiting qualities are then compared, and the better quality of the two is selected to represent the search quality. In US-VISIT, if a single finger match score is sufficiently high, then the other finger’s match score is ignored, no matter how bad; therefore, the best of the limiting qualities is selected.

The third curve (with blue ‘*’s) in Figure 11 plots the search quality distribution combining the qualities represented in the other two curves.

Given the definition and assignment of search quality above, the 60K BCC probes used in this report were sorted into their respective eight quality bins. Performance was then computed for each quality bin, and the results are reported in Figure 12. Performance was recorded at each of the IDENT thresholds surveyed in Figure 9.

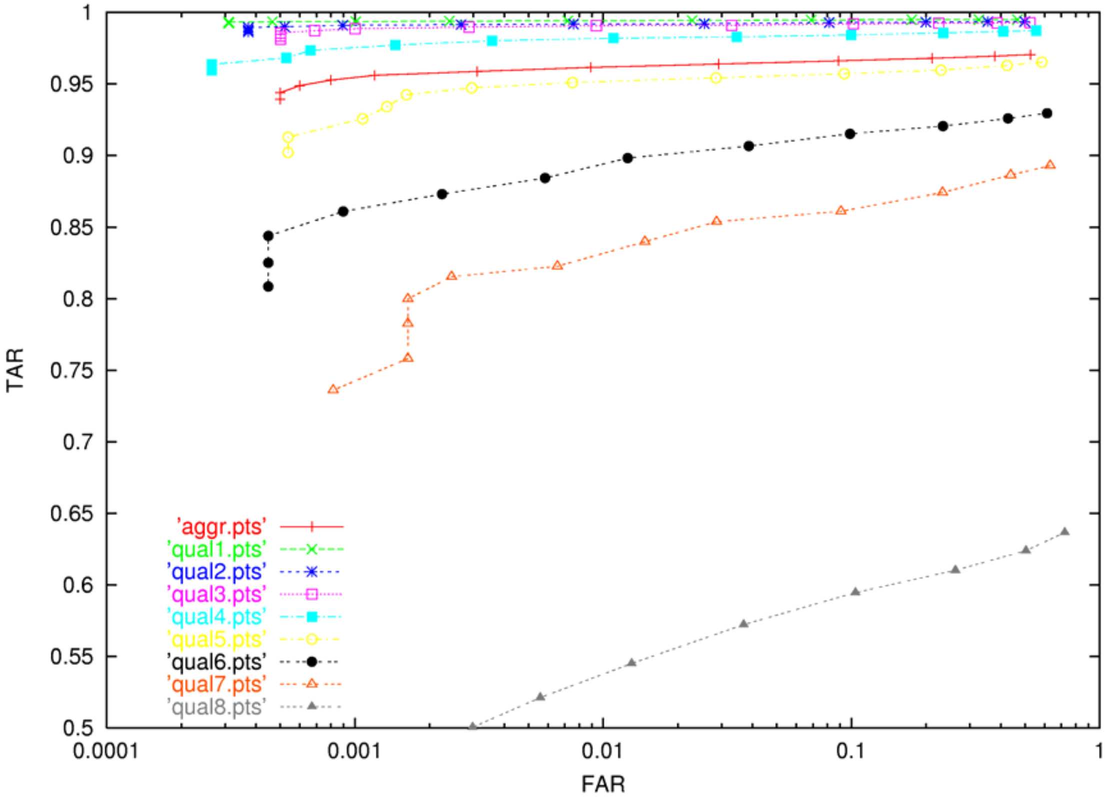

Figure 12. Effect of image quality on Elite performance across a sample range of IDENT thresholds

The first curve to note in the figure is the aggregate curve (with red '+'s). This is the same curve plotted in Figure 9 and is included here for comparison purposes. Looking at the other eight curves, the vertical order of the curves from top to bottom corresponds to the eight quality bins in increasing order. The aggregate curve lies between the curves for image quality 4 and 5. The curves for quality 1, 2, and 3 are tightly grouped at the top of the graph. Qualities 4 through 7 follow below and are increasingly spread apart vertically. Image quality 8 is known to be a "catch-all-remaining" category, and as a result, its curve is significantly lower than quality 7's curve. As can be seen, fingerprints of quality 8 will match very poorly, if at all.

4.5.2 Ohio Study

The results shown in Figure 12, illustrate the impact of image quality on performance. The image qualities were analyzed within the natural distribution of the BCC dataset shown in Figure 11. One concludes that overall identification performance improves as image quality improves. To further verify this quality vs. performance relationship, a study was designed whereby the BCC background was seeded with mates and searched with probes of higher quality than those naturally occurring in the BCC application. The results of this study are presented in this section.

The State of Ohio ran a pilot study which examined the effects of capturing live scan plain images and using them to search the FBI's Criminal Master File which is a legacy collection of rolled ink fingerprints. [7] As part of this study, several different live scan devices were used and compared, and fingerprints from the same subject were captured on these different devices. These devices included an Identix TP600/2000, CrossMatch CMT 442, and Smiths Heimann LS2-Check RJ0444.

The fingerprint images captured off these three devices were released to NIST for performance evaluation purposes. In total, there were 864 people who had fingerprints captured on all three devices. The quality distribution profiles for the fingerprints captured from each of these devices are shown in Figure 13. Looking at the figure's legend, the qualities are broken out by right index finger (labeled with finger position "02") and by left index finger (labeled with finger position "07"). Identix qualities are labeled "id", CrossMatch qualities are labeled "cm", and Smiths Heimann qualities are labeled "sh". In terms of highest quality, Smiths Heimann has the greatest number of quality 1's, next is CrossMatch, and then Identix. The results are intermixed at quality 2.

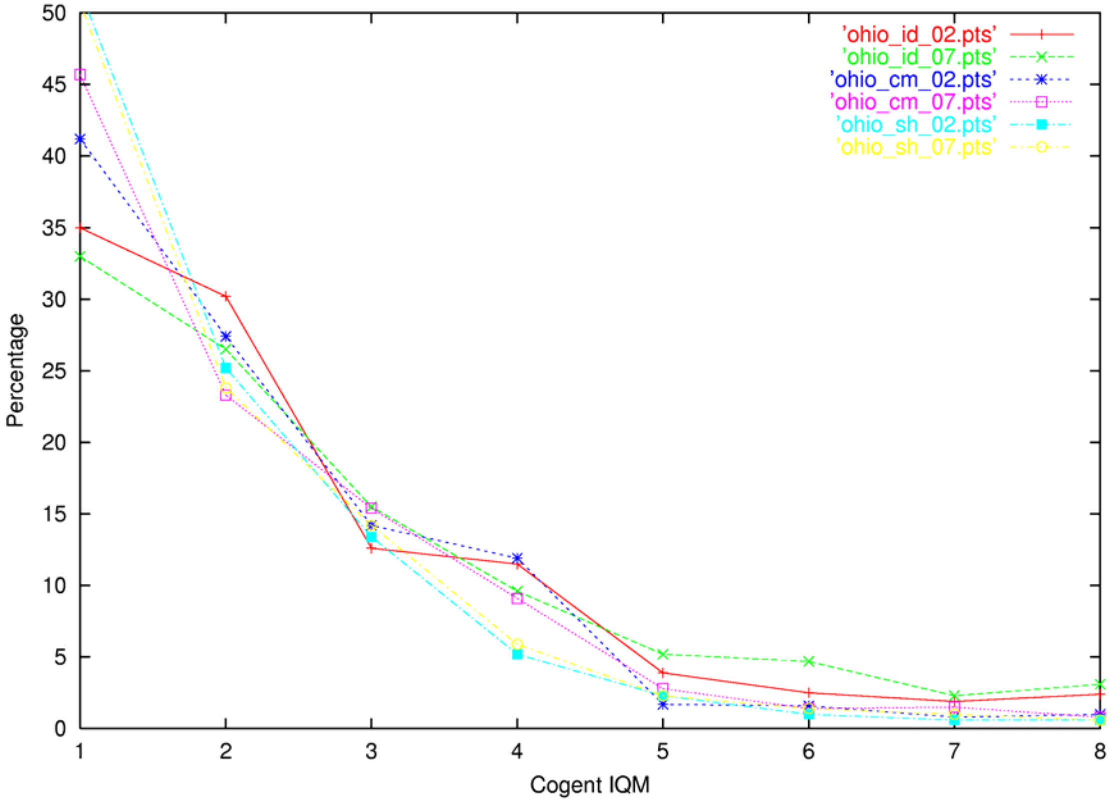

Figure 13. Cogent IQM distributions for Ohio fingerprints

There is a significant difference observed when comparing the Ohio quality distributions in Figure 13 to the BCC quality distributions in Figure 11. The majority of Ohio fingerprints have a quality equal to or better than 5, with the greatest amount at quality 1 and monotonically decreasing to 5. The BCC quality distribution has a lower frequency at qualities 1 & 2, an increased frequency in qualities 5, 6, & 7, with a relative increase in quality 8. These differences indicate that the Ohio data is of significantly better quality than the BCC data. This is expected as the Ohio subjects were prisoners and significant care (and as much time as was necessary) was taken to capture good quality fingerprints. It is generally accepted that the Ohio data is as good of quality as can be expected given the state of the art of live scan devices used at the time of that study.

Having established a comparative baseline of very good quality, the Ohio data was seeded into the BCC background and probed, and identification performance was computed. To accomplish this, the 864 pairs of fingerprints captured by Identix were added to the BCC background as mates. Two sets of probes were then searched against this seeded background, the first corresponding to fingerprint pairs captured by CrossMatch, and the second captured by Smiths Heimann. The performance results are listed in Table 2.

Results are reported in Table 2 for the operational IDENT thresholds of (1300, 1880) and the NIST recommended thresholds of (1400, 2025). Performance is slightly better for the Smiths Heimann probes than for the CrossMatch probes for both threshold settings. FAR significantly improves when using the thresholds (1400, 2025). Compare the first two rows of results to that of the last row. The set of 60K BCC probes achieves results comparable in FAR to that of CrossMatch. However, the improved image quality of the Ohio data is seen in the difference of TAR values. BCC probes achieved a TAR greater than 95%, while both probe sets of Ohio data achieved a TAR greater than 98%. Although the size of the Ohio dataset is quite small, making it difficult to adequately determine the difference in performance between the different live scan devices, this study conclusively demonstrates that improvements in image quality will translate into improvements in identification performance.

	Thresholds **1300, 1880**		**Thresholds** **1400, 2025**	
Probe Set	**FAR**	**TAR**	**FAR**	**TAR**
CrossMatch	0.46%	98.4%	0.12%	98.4%
Smiths Heimann	0.23%	98.8%	0.00%	98.8%
BCC	0.46%	95.9%	0.09%	95.2%

Table 2. Performance of Ohio probes seeded into BCC background compared to BCC probes

5. COMPARISON TO ATB

The ATB (Algorithm Test Bed) is a fingerprint matching system, built by Lockheed-Martin, which allows algorithms identical to those used in the FBI IAFIS system to be tested.[5] The performance of the ATB duplicated that of IAFIS on accuracy. Running tests on the ATB with data identical to that used to test the IDENT system allows the accuracy of the two systems to be directly compared and conclusions drawn to IAFIS.

The ATB algorithms were designed and optimized for matching 10 rolled fingerprints. This is the condition that allows the system to perform with greatest efficiency and accuracy. This is not the only possible way to operate the system. The primary matcher of the ATB is a two index finger matching algorithm. The comparisons presented here are for two index fingers which degrades the system throughput performance by a factor of 40.

5.1 Filtering

In the ATB and IDENT systems, the fingerprint matchers use a process called filtering that decreases the fraction of the database required to be searched and increases the effective speed of the system. In the ATB with 10 rolled fingerprints, the nominal filter rate is 2% and filter rates of 1.5% or less are common. For a filter rate of 2%, this increases the effective match speed by a factor of 50. The 98% of the database is eliminated by comparing the patterns of all ten fingers with those in the database. Only those fingerprints that have the same sequence of finger patterns are matched.

The IDENT system was designed to use much less aggressive filtering based on the shape of the ridge flow of the fingerprint. Using BCC data, 30% of the database is dropped by the shape filter, and only the remaining 70% is searched.

The different filtering strategies used by IAFIS and IDENT were optimized for different expected applications. The ATB/IAFIS algorithm is optimized for ten print matching and makes use of the ten available fingers to filter aggressively. The ATB algorithm is not as effective on two fingers as the IDENT algorithm that was designed specifically for two finger use.

5.2 Two Finger Matching

Both the ATB and IDENT systems use two index fingers for the high speed matching operation that searches a large section of the database. In earlier versions of IAFIS and in the current IDENT system, the primary database search uses special purpose hardware. The high-speed section of each system produces a candidate list of fingerprints that are then matched by slower but more accurate matching methods. In the ATB, the matcher uses additional fingers. In the IDENT system the matcher uses additional features.

5.3 Sensitivity Image Quality

The sensitivity to image quality of both the ATB and the IDENT matcher was measured for BCC quality data. The impact of applying Cogent IQM is shown in Table 3.

The results for both systems are strikingly similar. At quality 1, both systems had a TAR over 98%. At quality 8, both systems had a TAR approaching 50%. The algorithms used in the two systems are different as discussed above, so one concludes that the impact of image quality is greater than the impact of the difference in algorithms.

Cogent IQM	IDENT FAR	IDENT TAR	ATB FAR	ATB TAR
1	1.0%	99.4%	0.6 %	98.2%
2	1.0%	99.2%	0.7%	97.9%
3	1.0%	99.1%	0.8%	97.1%
4	1.0%	98.2%	0.8%	94.9%
5	1.0%	95.2%	1.0%	90.9%
6	1.0%	89.3%	0.9%	84.0%
7	1.0%	83.0%	1.2%	77.7%
8	1.0%	53.6%	1.4%	47.0%

Table 3. Sensitivity of ATB and IDENT to Cogent image quality levels

NIST has also developed an open (non-proprietary) algorithm for determining the quality of a finger print image. This algorithm, named NIST Fingerprint Image Quality (NFIQ), is based on quality statistics produced by NIST's minutia detector, which are turned into feature vectors and classified by a neural network. The details of this algorithm are reported in Reference [9]. An export-controlled free source code distribution of NFIQ (and a whole host of other fingerprint technologies) is available on CD-ROM [10] by contacting one of the authors of this report.

NFIQ categorizes image quality into five quality levels, with level one representing best quality down to level five representing worst quality. The accuracies from both the ATB and the IDENT matcher on BCC data were resorted according to the search qualities produced by NFIQ, and the sensitivity to NFIQ qualities is shown in Table 4.

NIST NFIQ	FAR	IDENT TAR	ATB TAR
1	1.0%	99.4%	98.0%
2	1.0%	98.4%	95.4%
3	1.0%	88.1%	80.4%
4	1.0%	59.4%	53.0%
5	1.0%	27.8%	24.4%

Table 4. Sensitivity of ATB and IDENT to NIST fingerprint image quality levels

Once again, one observes a clear and consistent degradation in the performance of both the ATB and IDENT matchers. However, comparing Table 4 to Table 3, it is clear that NFIQ discriminates much more significantly between its five quality levels than does Cogent IQM across its eight levels. This also provides supporting evidence that NFIQ is matcher independent.

6. RECOMMENDATIONS

NIST and DHS have developed a set of testing procedures and test data that is suitable for testing both the verification and identification functions used for fingerprints in the US-VISIT program. This process has involved substantial effort and cost. We recommend that this procedure and its associated test data be used to develop and conduct ongoing tests for future modifications of the biometric algorithms and their settings in US-VISIT.

6.1 Existing Algorithms

The existing algorithms deployed in US-VISIT have been tested on BCC data. We recommend that these tests be conducted on operational US-VISIT data when a sufficiently large sample becomes available. The Department of State (DOS) is using different readers and different client software in their consular offices. We recommend that this data be used for testing against the Phase-I US-VISIT data. When the collection of data at land port of entries (POEs) begins, we also recommend further testing of the US-VISIT algorithms on this data as well.

6.2 Algorithm Modifications

During the course of developing the tests used in this report, NIST worked with three sets of one-to-one and three sets of one-to-many algorithms. Each set of algorithms was more accurate than the previous set. This demonstrates that some system developers can make improvements in biometric algorithms very rapidly and that significant improvements in algorithms are possible in the near future. We recommend that the test procedures developed here be used to monitor and evaluate these improvements.

6.3 Adjustments for Image Quality

This report has shown that fingerprint image quality is critical for the accurate matching of fingerprints. The US-VISIT program has implemented an extensive image quality monitoring program. We recommend that programs of this type, where the image quality is monitored for all the sources of data use in the program, be extended to all planned sources of US-VISIT fingerprints.

6.4 Adding Single Finger Scores for Verification

Results in this report have demonstrated that verification accuracy is significantly improved when matcher scores from right and left index fingers are fused together. Using the simple method of adding together the two finger matcher scores, a threshold of 740 yielded a TAR of 99.5% at a FAR of 0.1%. We recommend incorporating this method into US-VISIT.

REFERENCES

[1] "Use of Technology Standards and Interoperable Databases with Machine-Readable, Tamper-Resistant Travel Documents – Appendix A;" PDF document at http://www.itl.nist.gov/iaui/894.03/fing/fing.html; November 2002.

[2] Public Law 107-56 (USA PATRIOT ACT); 107th United States Congress, Washington, D.C.; 26 October 2001

[3] Public Law 107-173 (Enhanced Border Security and Visa Entry Reform Act of 2002); 107th United States Congress, Washington, D.C.; 14 May 2002

[4] C. Wilson, C. Watson, M. Garris, A. Hicklin; "Studies of Fingerprint Matching Using the NIST Verification Test Bed (VTB)" NISTIR 7020; National Institute of Standards and Technology; Gaithersburg Maryland; 07 July 2003

[5] S. Wood and C. Wilson, "Studies of Plain-to-Rolled Fingerprint Matching Using the NIST Algorithmic Test Bed (ATB)" NIST IR 7112; National Institute of Standards & Technology, Gaithersburg Maryland, April 2004..

[6] "The Universal Latent Workstation (ULW);" For FBI information on the Web: http://www.fbi.gov/hq/cjisd/ulw.htm.

[7] "National WebCheck Program Pilot Project Preliminary Report;" Ohio Attorney General Bureau of Criminal Identification and Investigation; London, Ohio; December 2003.

[8] P. J. Phillips, P. Grother, R. Micheals, D. Blackburn, E. Tabassi, and M. Bone, 'Face recognition vendor test 2002, NIST IR 6965, National Institute of Standards & Technology, Gaithersburg Maryland, March 2003.

[9] E. Tabassi, C. Wilson, C. Watson, "Fingerprint Image Quality," Technical Report 7151, August 2004. Appendices for NISTIR 7151 can be found at http://fingerprint.nist.gov/NFIS.

[10] C. Watson, M. Garris, E. Tabassi, C. Wilson, R. M. McCabe, S. Janet, "*NIST Fingerprint Image Software 2 (NFIS2)*: NFSEG, PCASYS, MINDTCT, NFIQ, BOZORTH3, AN2K, & IMGTOOLS," free export-controlled CD-ROM, October 2004.