Fingerprint Vendor Technology Evaluation 2003: Summary of Results and Analysis Report

Summary of Results

NISTIR 7123

Charles Wilson [1]

R. Austin Hicklin [2]

Harold Korves [2]

Bradford Ulery [2]

Melissa Zoepfl [2]

Mike Bone [3]

Patrick Grother [1]

Ross Micheals [1]

Steve Otto [1]

Craig Watson [1]

[1] National Institute of Standards and Technology
[2] Mitretek Systems
[3] NAVSEA Crane Division

June 2004

FINGERPRINT VENDOR TECHNOLOGY EVALUATION 2003

SUMMARY OF RESULTS

Charles Wilson[1]

R. Austin Hicklin[2]
Harold Korves[2]
Bradford Ulery[2]
Melissa Zoepfl[2]

Mike Bone[3]
Patrick Grother[1]
Ross Micheals[1]
Steve Otto[1]
Craig Watson[1]

[1]National Institute of Standards and Technology
[2]Mitretek Systems
[3]NAVSEA Crane Division

Abstract

The Fingerprint Vendor Technology Evaluation (FpVTE) 2003 was conducted to evaluate the accuracy of fingerprint matching, identification, and verification systems. The FpVTE is one of the tests that NIST has conducted in order to fulfill part of its PATRIOT Act mandate. Additional evaluations include the testing of the FBI IAFIS system, the US-VISIT IDENT system and SDKs (Software Development Kits) from several vendors. Eighteen different companies competed in FpVTE, and 34 systems were evaluated. Different subtests measured accuracy for various numbers and types of fingerprints, using operational fingerprint data from a variety of U.S. Government sources. The most accurate systems were found to have consistently very low error rates across a variety of data sets. The variables that had the clearest effect on system accuracy were the number of fingers used and fingerprint quality. An increased number of fingers resulted in higher accuracy: the accuracy of searches using four or more fingers was better than the accuracy of two-finger searches, which was better than the accuracy of single-finger searches. The test also shows that the most accurate fingerprint systems are more accurate than the most accurate facial recognition systems, even when comparing the performance of operational quality single fingerprints to high-quality face images.

1 Introduction

1.1. Overview

The Fingerprint Vendor Technology Evaluation (FpVTE) 2003 was conducted to evaluate the accuracy of fingerprint matching, identification, and verification systems. FpVTE 2003 was conducted by the National Institute of Standards & Technology (NIST) on behalf of the Justice Management Division (JMD) of the U.S. Department of Justice. FpVTE 2003 serves as part of the NIST statutory mandate under section 403(c) of the USA PATRIOT Act to certify biometric technologies that may be used in the U.S. Visitor and Immigrant Status Indicator Technology (US-VISIT) Program.

FpVTE 2003 was conducted at the NIST Gaithersburg, MD facilities from October through November 2003. Planning for FpVTE started in May 2003, and analysis continued through April 2004. Eighteen different companies participated, with 34 systems tested, including the NIST Verification Test Bed fingerprint benchmark system. Each test had a time limit of two or three weeks, running continuously. It is believed that FpVTE 2003 was the most comprehensive evaluation of fingerprint matching systems ever executed, particularly in terms of the number and variety of systems and fingerprints.

Participants in the FpVTE 2003 test were required to assemble, configure, and run their own hardware and software at NIST's Gaithersburg, Maryland facility. The trials began in October 2003, with each participant running over a two- or three-week period according to a predetermined and staggered schedule. Testing of all eighteen different companies was completed in November 2003.

FpVTE 2003 included operational fingerprint data from a variety of U.S. and State Government sources. The test used 48,105 sets of flat slap or rolled fingerprint sets from 25,309 individuals, with a total of 393,370 distinct fingerprint images.

The FpVTE Analysis Report concludes:

1. Of the systems tested, NEC, SAGEM, and Cogent produced the most accurate results.
2. These systems performed consistently well over a variety of image types and data sources
3. These systems produced matching accuracy results that were substantially different than the rest of the systems
4. The variables that had the largest effect on system accuracy were the number of fingers used and fingerprint quality:
    - Additional fingers greatly improve accuracy
    - Poor quality fingerprints greatly reduce accuracy
5. Capture devices alone do not determine fingerprint quality
6. Accuracy can vary dramatically based on the type of data:
    - Accuracy on controlled data was significantly higher than accuracy on operational data
    - A biometric evaluation that only uses a single type of data is limited in how it can measure or compare systems
7. Incorrect mating information is a pervasive problem for operational systems as well as evaluations, and limits the effective system accuracy
8. With current technology, the most accurate fingerprint systems are far more accurate than the most accurate face recognition systems.

1.2 Purpose

The evaluations were conducted to:

- Measure the accuracy of fingerprint matching, identification, and verification systems using operational fingerprint data
- Identify the most accurate fingerprint matching systems
- Determine the effect of a wide variety of variables on matcher accuracy
- Develop well-vetted sets of operational data from a variety of sources for use in future research

The evaluations were ***not*** intended:

- To measure system throughput or speed
- To evaluate scanners or other acquisition devices
- To directly measure performance against very large databases
- To take cost into consideration
- To address latent fingerprint identification

1.3 Certification

For purpose of NIST PATRIOT Act certification this test certifies the accuracy of the participating systems on the datasets used in the test. This evaluation does not certify that any of the systems tested meet the requirements of any specific government application. This would require that factors not included in this test such as image quality, dataset size, cost, and required response time be included. Certifications of deployed government systems such as the FBI's IAFIS and US-VISIT's IDENT system are covered by references [ATB] and [IDENT].

1.4 Personnel

A number of people had roles in FpVTE. Table 1 gives the name, affiliations and role of the staff that designed and executed the test.

Manager	Charles Wilson	NIST
FpVTE Liaison	Steven Otto	NIST
Lead Test Agent	Mike Bone	NAVSEA Crane Division
Test Design and Analysis Team	Austin Hicklin Harold Korves Brad Ulery Melissa Zoepfl	Mitretek Systems
	Patrick Grother Ross Micheals Craig Watson	NIST

Table 1: FpVTE Personnel

2 Related Studies

NIST has or will release three related reports that provide additional information related to PATRIOT ACT certification of fingerprint systems.

2.1 Algorithmic Test Bed (ATB) Testing

NIST recently conducted a series of fingerprint matching studies using an experimental laboratory system called the Algorithmic Test Bed (ATB). The NIST ATB system is a lower capacity version of the FBI's Integrated Automated Fingerprint Identification System (IAFIS) and is being used to test the functional characteristics of IAFIS. The machine is configured with a gallery of nearly 1.2 million subjects and provides broad control over its operating modes and set points.

A NIST report on these studies [ATB] was published in April 2004. The FpVTE study includes aspects of the ATB studies – that addressthe matching of plain to rolled, and plain to plain, fingerprint images

2.2 Software Development Kit (SDK) Testing

NIST has conducted a series of SDK (Software Development Kit) based verification tests intended to evaluate the accuracy of the one-to-one matcher used in the US-VISIT program. Fingerprint matching systems from six vendors not currently used in US-VISIT were also evaluated to allow benchmark comparisons of the current VISIT matcher with other commercially available products. Each SDK based verifier was tested using twelve different fingerprint data sets of varying difficulty. Each set consisted of 12,000 single-finger images from 6,000 persons.

The average measured true accept rate at a false accept rate of 0.01% exceeded 98% for the two highest scoring systems with the worst always greater than 94%.The findings of the SDK tests, including documentation of the data sets and testing procedures, are detailed in a separate report [SDK].

2.3 US-VISIT IDENT Testing

A third NIST study addressed the flat-to-flat matching performance of the operational US-VISIT fingerprint matching system. Different subsystems of IDENT perform both one-to-many matches (to detect duplicate visa enrollments) and one-to-one matches (to verify the identity of visa holders). With the proper selection of an operating point, the one-to-many true accept rate for a two-finger comparison against a database of 6,000,000 subjects is 95% with a false accept rate of 0.08%. Using two fingers, the one-to-one matching accuracy is 99.5% with a false accept rate of 0.1%.

A NIST report on this test [IDENT] was published in May 2004

3 Comparison of Face and Fingerprints

The report that was sent to Congress [303a] as part of NIST's PATRIOT Act mandate [PATRIOT, BorderSecurity] included a comparison of face recognition results from the FRVT 2002 study [FRVT2002] with single-finger results from the NIST VTB fingerprint system [VTB]. 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 the best commercial systems, and because the DHS2 data used for the 303a report were the poorest quality in any datasets used in FpVTE. In addition, although the images used in the FRVT 2002 test are of higher quality than many of those present in operational government data sets, they all fall short of the specifications of the draft Face Image standard [ISO/IEC].

Leading contemporary fingerprint systems are substantially more accurate than the face recognition systems tested in FRVT 2002. When all these factors are combined, the comparison of face and fingerprint accuracy needs to be revised. This conclusion holds even for face and fingerprint images categorized as high quality, however, it must also be considered that any advances in face recognition technology since the FRVT tests have yet to be evaluated. Further performance benefits associated with data collected to comply with ISO/IEC 19794-5 also remain unquantified.

The following entries summarize the verification performance documented in FpVTE 2003 and FRVT 2002. The most accurate face systems:

- 71.5% true accept rate @ 0.01% false accept rate
- 90.3% true accept rate @ 1.0% false accept rate.

The most accurate fingerprint system tested (NEC MST) using operational quality single fingerprints:

- 99.4% true accept rate @ 0.01% false accept rate
- 99.9% true accept rate @ 1.0% false accept rate

When multiple face images are available, the performance of face recognition can be improved [Grother3]. With four previous images in the gallery the error rates are substantially reduced

- 89.6% true accept rate @ 0.01% false accept rate
- 97.5% true accept rate @ 1.0% false accept rate

In FpVTE 2003, when four fingerprints were used for matching, the most accurate fingerprint system tested (NEC LST) always had true accept rates in excess of 99.9% at a FAR of 0.01%.

4 Overview of Tests

FpVTE was composed of three separate tests, the Large-Scale Test (LST) the Medium-Scale Test (MST), and the Small-Scale Test (SST). Table 2 compares parameters associated with each of the three tests.

SST and MST tested matching accuracy using individual fingerprints, all of which were images from right index fingers. This contrasts with LST, which evaluated matching accuracy using sets of fingerprint images, where each set includes anywhere from one to ten fingerprints collected from an individual subject at one session. The tests were designed so that the SST is a subset of the MST. As a consequence, this allows direct comparison of SST and MST Participants. LST Participants were encouraged to participate in the MST.[1]

Participants were permitted to enter more than one system in the evaluation.

[1] Eleven of the thirteen LST participants had valid MST results, but some of those had different system configurations in MST and LST.

Test	Compares	# Subtests	# Comparisons	# Systems Successfully Completed	Allowed Time
LST	Sets of 1-10 fingerprint images (Flat, Slap, and Rolled; various combinations of fingers)	31 (uses 10 datasets containing 64,000 fingerprint sets)	1.044 billion set-to-set comparisons	13	21 days
MST	Single images (Flat & Slap Right index)	1 (compares a single 10,000 image dataset against itself)	100 million single image comparisons	18	14 days
SST[2]	Single images (Flat Right index) (Subset of MST)	1 (compares a single 1,000 image dataset against itself)	1 million single image comparisons	3 (SST only) 21 (as a subset of MST)	14 days

Table 2. Summary of FpVTE Tests

The size and structure of each test were designed to optimize competing analysis objectives, available data, available resources, the Participants' responses to the *System Throughput Questionnaire* (see Appendix A), and the desire to include all qualified Participants.

In particular, the sizes of MST and LST were only determined after a great deal of analysis and consideration of a variety of issues. The systems in FpVTE differed in several significant ways, for example:

- maximum throughput capacity
- the relative proportion of time spent preprocessing images and matching images
- the ability to increase throughput rates by decreasing accuracy
- the ability to increase throughput by adding additional hardware.

Designing a well-balanced test to accommodate heterogeneous system architectures was a significant challenge.

The timing analysis performed by NIST suggests that to increase the total number of comparisons made by a factor of ten (which would have been the smallest meaningful increase in measurement precision), the LST test duration would have had to increase from three weeks to *thirty* weeks. The alternative of using larger datasets and three weeks test time would have limited the test to those systems that could trade accuracy for throughput. Extending the length of the test would have placed a greater burden on the Participants for personnel and hardware. Increasing the throughput requirements without extending the length of the test would have favored one type of system, may have favored Participants with specialized hardware, and would have limited the number of participants. Although software development kit tests (see section 2.2) offer the possibility to run tests over many weeks or months, they do so by requiring vendors' applications to run on standard hardware and operating system combinations.

[2] Three systems competed in SST, but since SST was a subset of MST, all of the MST participants can be compared directly in SST. Hence 21 systems successfully completed this subtest.

5 Summary of Results

FpVTE analysis had three interrelated goals:

- To compare the competing systems on a variety of fingerprint data, identifying the systems that were most accurate;
- To measure the accuracy of fingerprint matching, identification, and verification systems on actual operational fingerprint data; and
- To determine the effect of a variety of variables on matcher accuracy.

As stated previously, the FpVTE analysis was not intended to take into consideration cost, throughput, equipment reliability or other factors that might be important in selecting a system for operational deployment.

5.1 Multi-Finger Performance (LST)

All of the LST systems achieved high accuracy on some of the data, especially in the ten-finger subtests. However, some of the LST systems were more consistent in their accuracy than others. Figure 1 shows the range of performance over 27 representative test partitions of operational fingerprint data. These partitions are discussed in the Analysis Report

Each line depicts a summary statistic for the systems' performance over the 27 partitions, characterizing the TAR accuracy as measured (or minimally interpolated) at FAR = 0.01%. For example, the line labeled "Average" shows for each system the average of 27 separate TAR measurements, each at FAR = 0.01%. The maximum accuracy for each system, also plotted on the graph, was quite high—100% accuracy or near-100% accuracy. Since maximum and minimum values are often outliers, the 5th highest and 5th lowest accuracies over the 27 partitions are also shown, to give a better indication of the spread of the data. Details of this subtest are discussed in Appendix D.

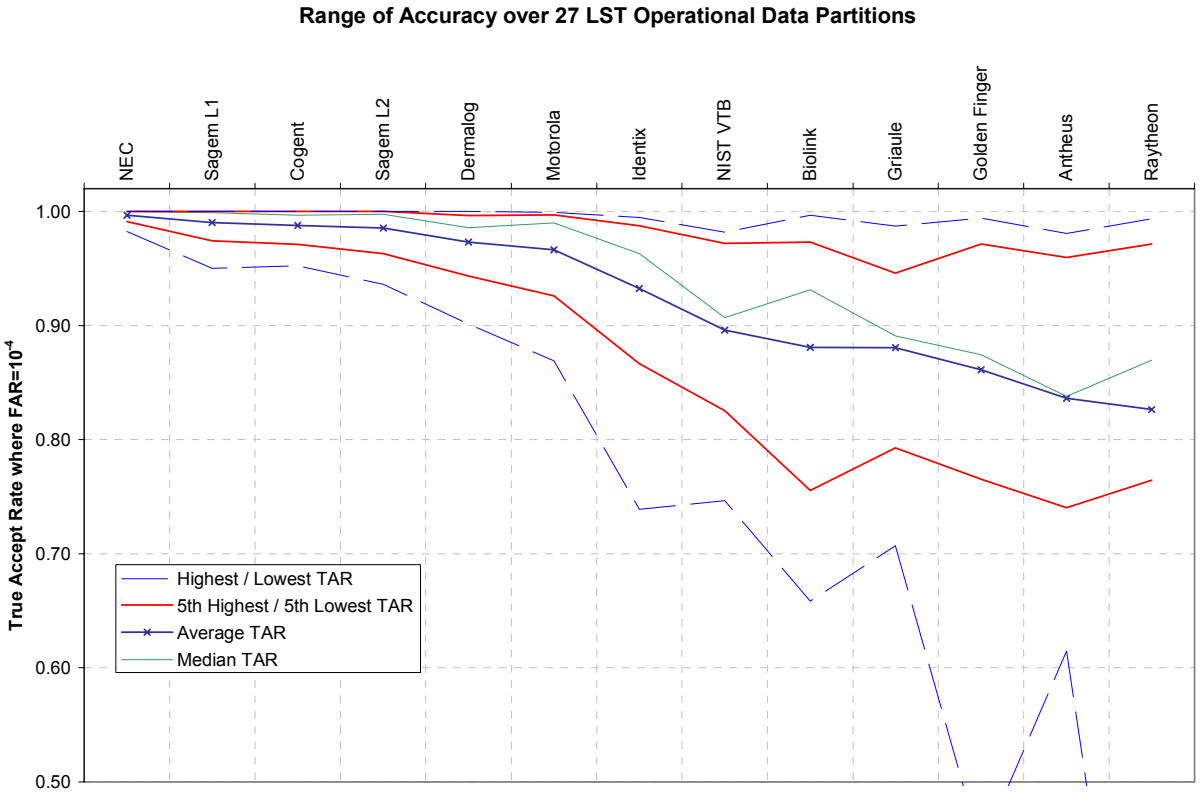

Figure 1. Range of Accuracy over 27 Operational LST Partitions. The systems are sorted by their average accuracy over the 27 partitions; note that sorting by median performance would change the order for some systems.

5.2 Single-Finger Flat and Slap Performance (MST)

In MST, the fingerprints were grouped by both source and type (as defined in the Analysis Report), yielding seven different combinations that were used to partition the data. The results for each participant for each partition were calculated and analyzed. The resulting range of accuracy on seven single-finger tests is shown in 3.

Since the highest and lowest true accept values are often outliers, the range between the second highest and second lowest is also shown. In the LST comparison, the highest TAR was often 100% with a minimal difference between the top several applicants. In MST, there was also a substantial difference between the highest and second highest values for many systems. For most systems, the highest value was achieved on the one partition that was collected under highly controlled conditions (Ohio dataset).The remaining six partitions contained only operational data, so the difference in the highest and second-highest scores are indicative of the difference between data collected via operational systems versus data collected under highly controlled conditions.

Details of this subtest are discussed in Appendix D.

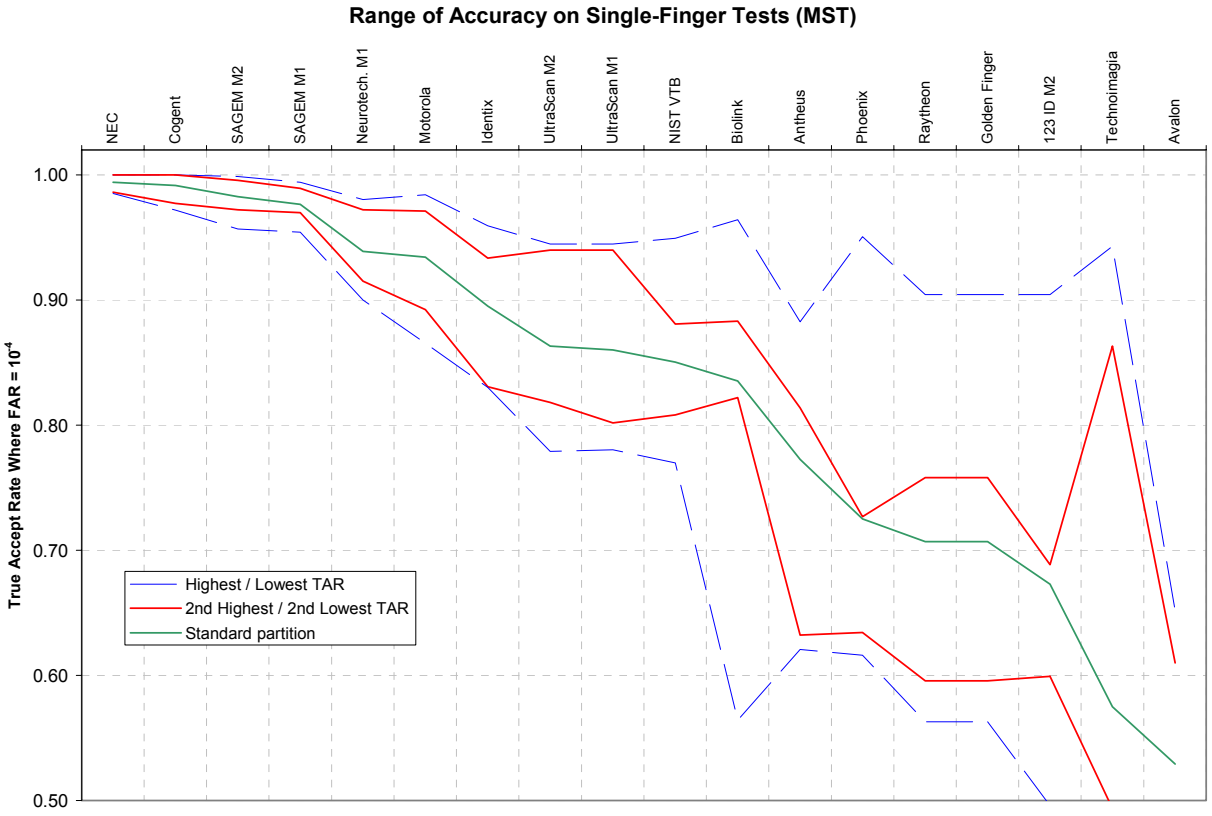

Figure 2. Range of accuracy across 7 MST partitions. These systems are sorted by the systems' performance on the standard MST, which is simply the combination of the seven partitions.[3] The large difference between the highest and second highest TARs is attributable to the presence of data collected in a controlled (highest) versus operational (second highest) setting.

To facilitate comparison, all of the MST systems were ranked in order of TAR at a 0.01% FAR, for each of the seven partitions these systems are sorted by the average rank over all seven partitions.

5.3 Single-Finger Flat Performance (SST)

SST was a small test that included only a single type of data (single-finger flats), from two sources. SST was a subset of MST, so any SST partitions are (by definition) partitions of MST. The results for each SST and MST participant for each source were calculated and analyzed. The resulting range of accuracy is shown in 5. The SST systems are sorted by the systems' performance on the SST standard partition, which is simply a combination of the other two partitions.

[3] Since the seven partitions differ in size, the results for the MST standard partition are not quite the same as the average of the seven partitions.

Due to the smaller size of the SST, these results are presented at a false accept rate of 0.1% and *not* 0.01% as is true for most of the other figures in this report.

Details of the SST are included in Appendix D.

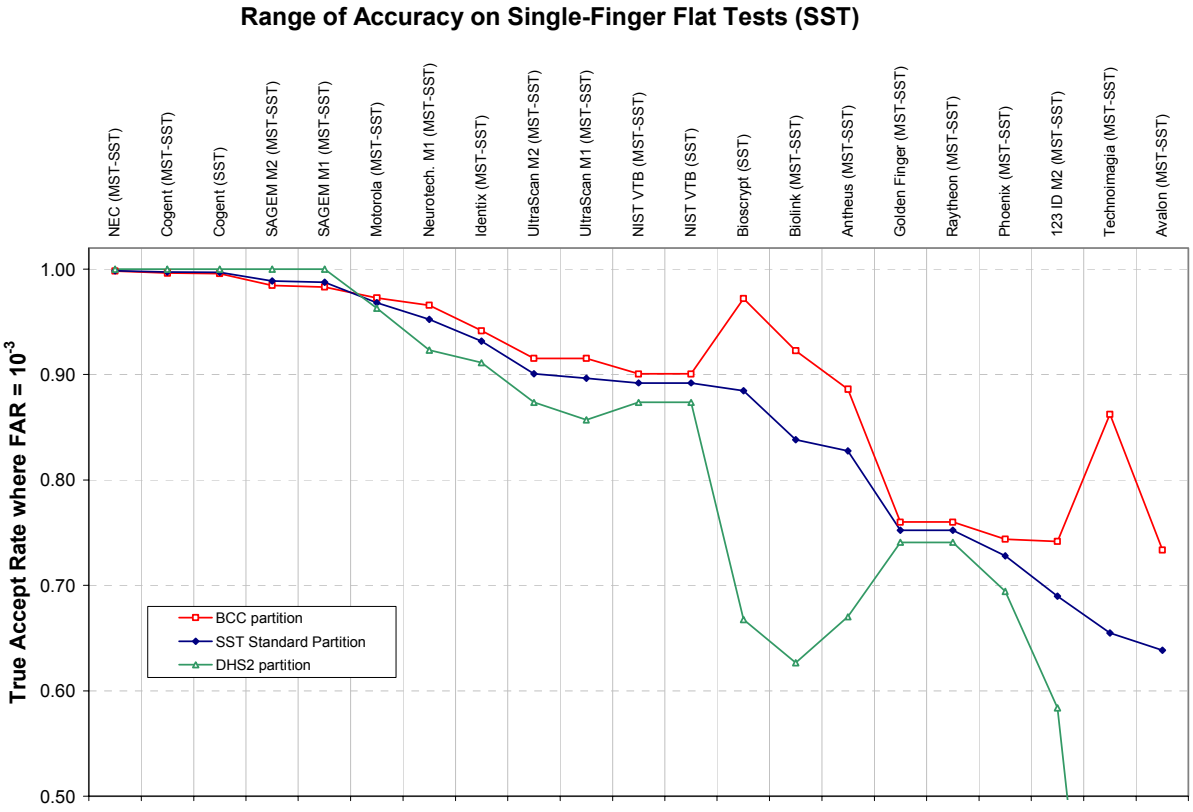

Figure 3. Range of Accuracy on Single-Finger Flats (SST). These systems are sorted by performance on the SST standard partition. Note that these results are reported at FAR=0.1%, in contrast to most of the figures in this report, which are based on FAR=0.01%.

5.4 Effect of Fingerprint Quality on Matcher Accuracy

It is well known that poor quality fingerprints are universally difficult to match. The effects of fingerprint quality are clear and dramatic, as shown in Figure : without exception, accuracy on good quality images was much higher than accuracy on poor quality images. This finding is important for several reasons:

- Operational procedures can be used to control fingerprint quality to a large extent;
- System designers can model the effect of different distributions of fingerprint quality on matcher accuracy to predict system cost and performance;
- Systems can use fingerprint quality to predict search reliability (low quality leads to false non-matches);
- The relevance of tests is limited if the distribution of fingerprint quality is not known in the test sets;
- The outcome of tests can vary significantly if fingerprint quality is not controlled.

Note that the sample sizes for the poorer quality images are very small, but the results are as expected and consistent across systems. Figure also shows that some systems are extremely sensitive to image quality.

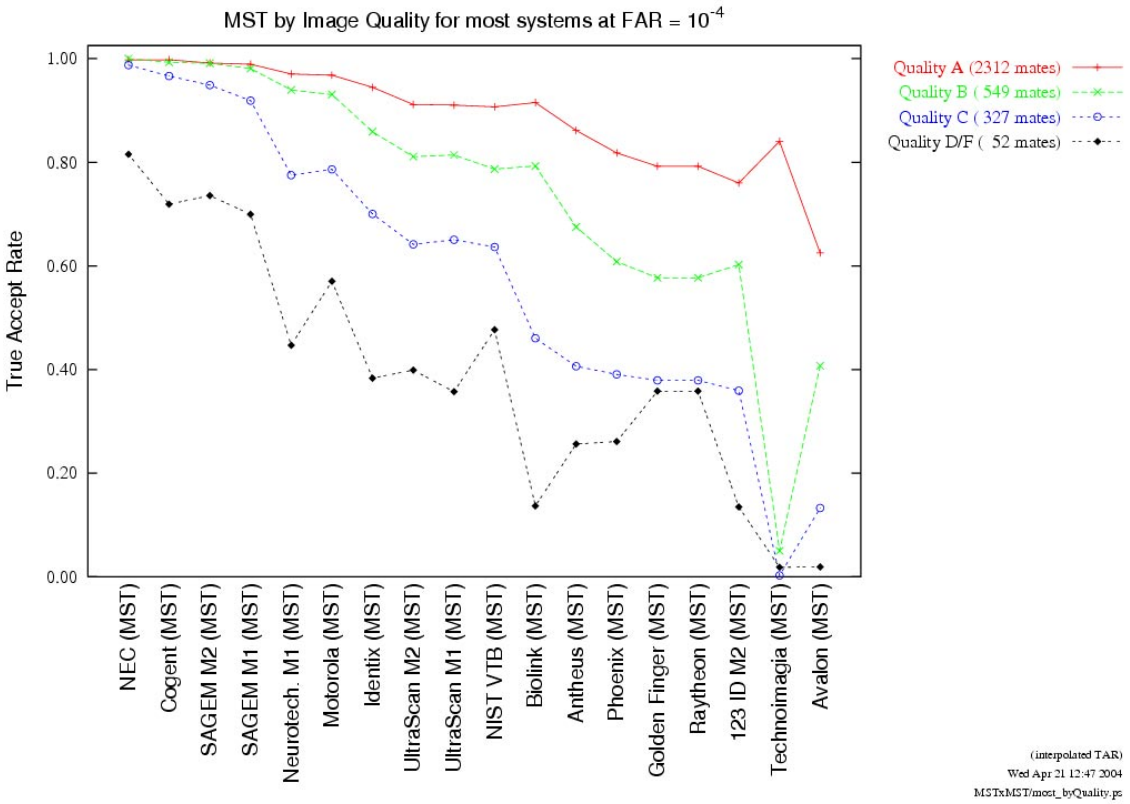

Figure 4. Effect of Image Quality (MST)

The image quality metrics used are discussed in Section 5.1 of the Analysis Report.

5.5 Effect of Number of Fingers

System accuracy was *highly* sensitive to the number of fingers compared. This can be seen clearly in Figure 4, which shows false reject rates at a fixed false accept rate of 0.01%. This figure compares different numbers of both plain and rolled fingerprints from both livescan and paper. Different colors are used to represent the number of fingers, while the letters denote rolled (R) and slap (S) fingerprints from paper (P) and livescan (L). Thus, the last aqua/grey curve as listed in the legend applies to the comparison of ten rolled paper prints with ten rolled paper prints (10RP vs. 10RP).

The error rates for each vendor typically vary by a factor of 100. The first vendor, NEC, falsely rejects 1 in 100 of the most difficult single fingers but fewer than one in ten thousand of the easiest ten finger sets. The last entry, Antheus, falsely rejects 40%of the hardest single fingerprints and 1% of the easiest multi-finger sets.

Figure 5 clearly shows that single finger matching is less accurate than two-finger matching, that two-finger matching is less accurate four-finger matching, and that four-finger matching is less accurate than eight-finger matching. The test sample size is not large enough to separate the eight and ten finger results. Thus the major conclusion from the figure is that each doubling of the number of fingers produces a fixed factor reduction in false rejection errors. For NEC, the error rates are 1% 1%, 0.2%, 0.05% and 0.01% for one, two, four and eight fingers respectively. Therefore, the errors reduce by approximately a factor of five as the number of fingers is doubled. Similar ratios of accuracy apply to the other vendors on the left side of the graph.

The figure also shows that there is great variability within the data for a given number of fingers. Much of this variability can be attributed to variations in data source, quality, and type.

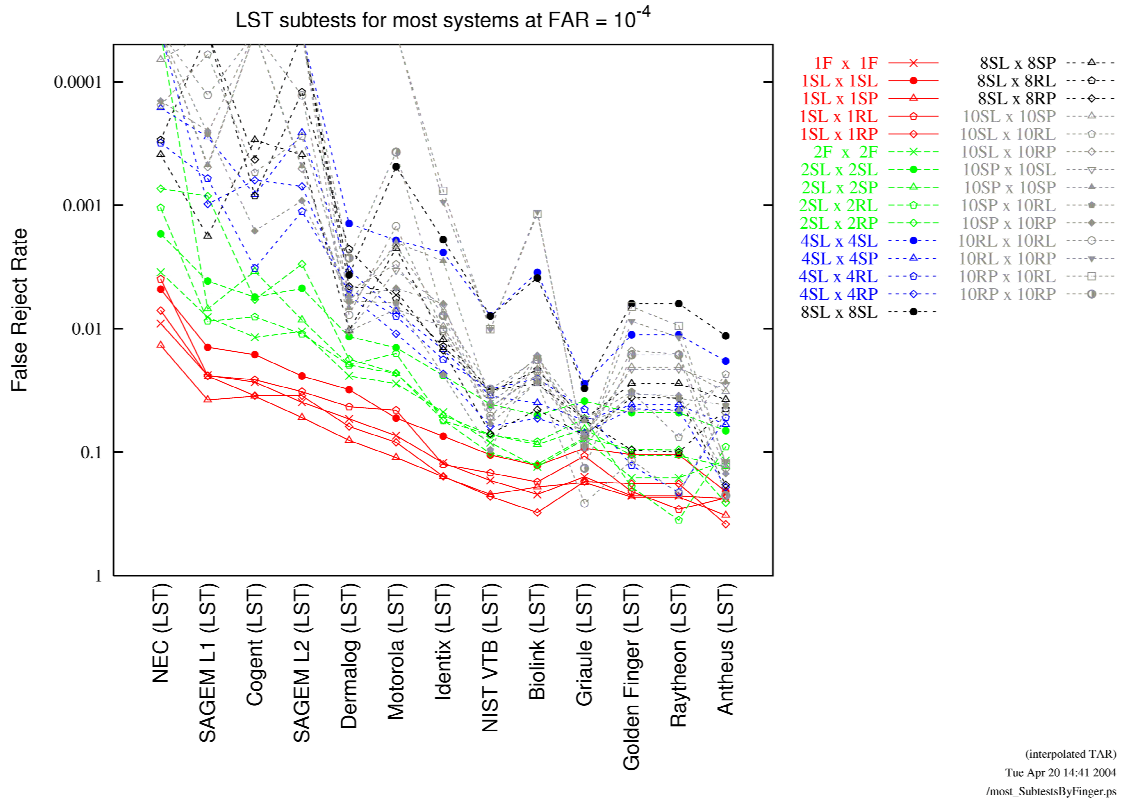

Figure 4. Effect of Fingerprint Number and Other Variables in LST. The Y scale is the log of False Reject Rate, which is 1 – TAR. Note that the single-finger searches (red) are clearly separated from the two-finger searches (green), but the four, eight, and ten-finger searches are intermingled. At the test sizes used, accuracy of four, eight, and ten-finger searches is difficult to differentiate and depends, to some extent, on the type of fingerprints used. The lines off of the top of the chart are for TAR=100% (no false rejects), which cannot be represented in log scale.

In order to minimize the effects of confounding variables, data source and image type were controlled in additional analyses. These analyses involved slap livescan probes compared to four different gallery types (slap livescan, slap paper, rolled livescan, and rolled paper), with data from four distinct sources. In general, the results showed that for most systems accuracy clearly improves as the number of fingers increases.

The following two charts show examples of the effect of number of fingers, where data source and type of fingerprint are held constant. Figure 5 shows results for the FBI's 12k[4], slap livescan vs. rolled livescan data set, which most systems match with high accuracy. Note that for the more accurate systems the results provide no evidence that more fingers improve accuracy on a dataset such as this, because TAR is already at or near 100% for a single finger.

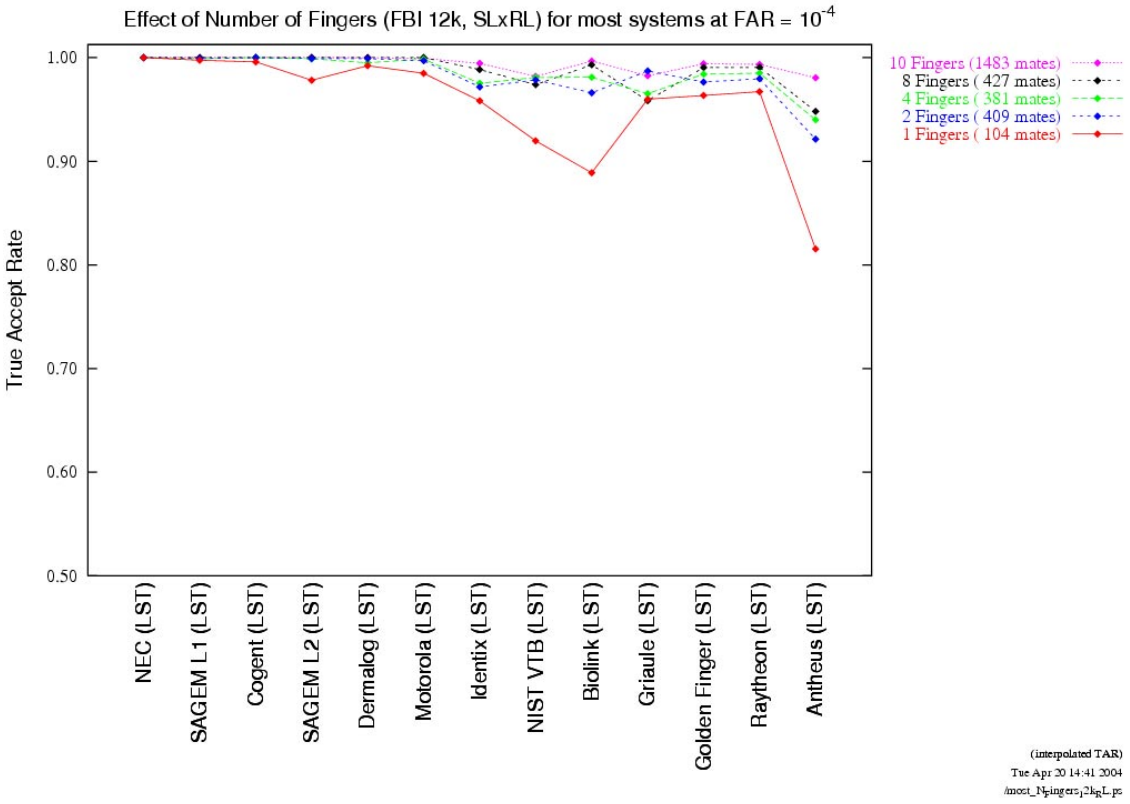

Figure 5. Effect of number of fingers on FBI 12k (slap livescan vs. rolled livescan). The effect is not measurable when the one-finger TAR approaches 100%.

[4] The "FBI 12k" data is described in more detail in the analysis report, with others.

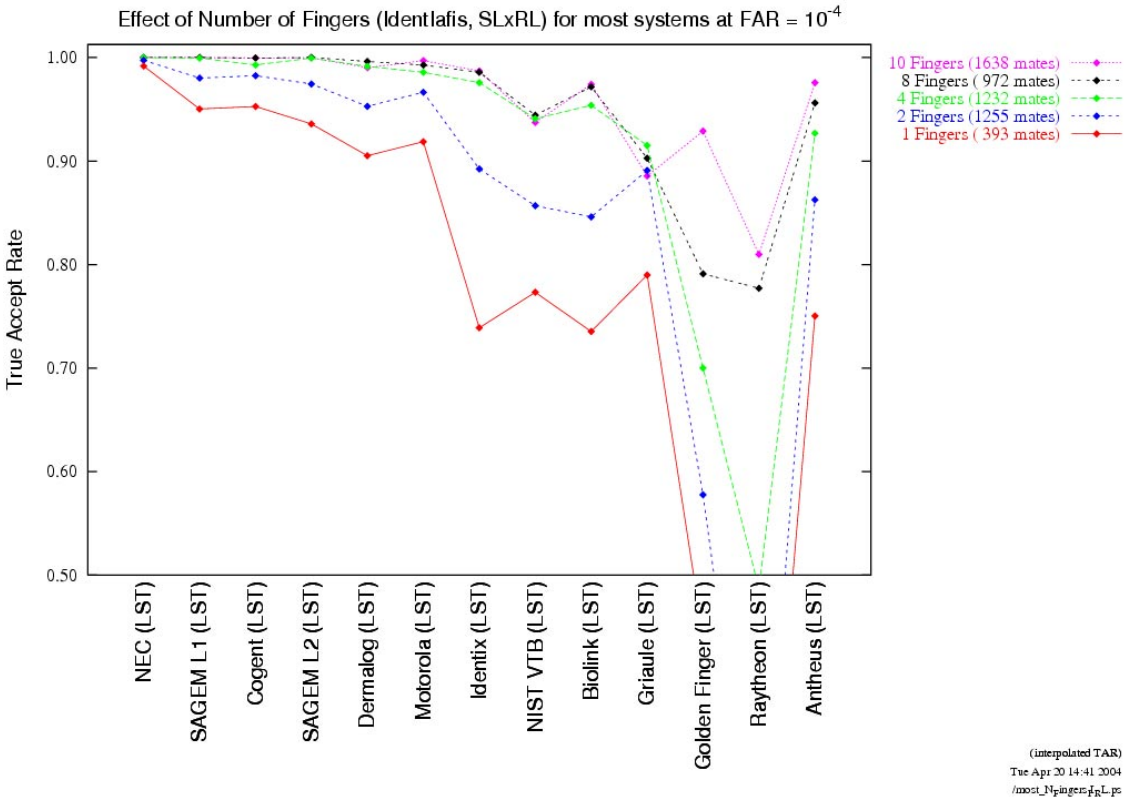

Figure 6. Effect of number of fingers on IDENT-IAFIS (slap livescan vs. rolled livescan). This data clearly shows the significant benefit when comparisons are done with more than two fingers.

Even for some of the more accurate systems, a difference in performance can be seen between two- and four-finger comparisons on the IDENT-IAFIS data (slap livescan vs. rolled livescan). Since NEC and SAGEM L1 achieved TARs of 100% with 4 fingers, they cannot be expected to differentiate at this level.

5.6 Other Results

Other important results discussed in the Analysis Report include:

- Accuracy on controlled data was significantly higher than accuracy on operational data.
- Some systems were highly sensitive to the sources or types of fingerprints, but this was not true of all systems.
- Accuracy dropped as subject age at time of capture increased, especially for subjects over 50 years of age. This effect may be due largely to image quality, which is known to vary by age.
- The choice of finger was not found to have a substantial effect on accuracy, except that segmented slap little fingers performed poorly.
- The following variables were ***not*** found to have a substantial effect on accuracy - gender, or criminal vs. civil records.

- In any operational government database, the performance attributable to source, fingerprint type (rolled, flat or slap images) and livescan vs. paper cannot be fully separated.

5.7 Implications for Operational Systems

When discussing the implications of the FpVTE results for operational systems, several issues need to be emphasized:

- Real world operational results for a system may be better or worse than the results reported here. Differences may arise from factors such as the operational environment, sources and types of fingerprint data, capture devices, operators and their training, hardware and software architecture and implementations, throughput requirements, and gallery size. One important conclusion of FpVTE is that such factors have a clear but complex effect on the performance of fingerprint systems.
- Operational systems are likely to use different operating points than are cited here, with correspondingly different error rates.
- Operational systems can be tuned to maximize performance given a particular concept of operations.
- Many systems have the ability to trade off accuracy for throughput: different throughput requirements will result in different levels of accuracy. Very high throughput requirements may be attained through a drop in accuracy.
- System cost, which was not addressed in FpVTE, must always be considered for operational systems.
- The error rates associated with slap segmentation were not addressed in FpVTE.

6 Conclusions

Overall, FpVTE makes six major conclusions regarding state-of-the-art, COTS and GOTS fingerprint systems.

1. **The systems were that performed most accurately developed by NEC, SAGEM, and Cogent**

In single and multi-finger tests (LST), NEC was the most accurate system (or tied for most accurate) in 42 out of 44 distinct combinations of data, including tests of mixed image type, and those from a variety of operational and controlled sources. The SAGEM and Cogent systems were the next most accurate LST systems.

In single-finger tests (MST), NEC was the most accurate system (or tied for most accurate) in 6 out of 7 distinct combinations of data, from both operational and controlled sources. The Cogent and SAGEM systems were the next most accurate MST systems.

Following the tier of the most accurate systems tested, the most accurate of the other systems tested were developed by Dermalog and Motorola, which had comparable performance.

Similarly, in the MST, the most accurate of the other systems were developed by Neurotechnologija and Motorola, which had comparable performance.

The SST results corresponded to the MST results.

2. The most accurate systems were highly accurate

On 44 test partitions defined by fingerprint type, number, and source, the most accurate LST system (NEC) was capable of identifying more than 98% of the mates in *every* subtest, with a false accept rate of 0.01%.

Given a false accept rate of 0.01% the results for NEC LST system showed that:

- Every single-finger subtest had a true accept rate higher than 98.6%
- Every two-finger subtest had a true accept rate higher than 99.6%
- Every four, eight, or ten-finger subtest had a true accept rate higher than 99.9%

SAGEM L1 and Cogent had true accept rates in excess of 95% on all single and multi-finger LST tests, at a false accept rate of 0.01%.

2a. The most accurate systems performed consistently well over a variety of image types and data sources

The most accurate systems maintained high accuracy even on data on which other systems performed with significantly less accuracy.

2b. There was a substantial difference in accuracy between the most accurate systems and the rest of the systems

The most accurate systems were more accurate than the rest of the systems for almost every metric examined.

On single-finger tests (MST and LST), accuracies below 80% were typical among the lower third (by rank) of participating systems. This corresponds to a False Reject Rate much more than ten times that of the high-accuracy systems. This ratio was even greater for multi-finger tests.

3. The variables that had the largest effect on system accuracy were the number of fingers used and fingerprint quality

3a. Additional fingers greatly improve accuracy

All systems achieve greater accuracy when multiple fingers are provided for comparison than when only one finger is provided. The improvement is both large and consistent. Although the actual benefits were found to vary by dataset and by system, the general trend was quite consistent. The accuracy of searches using four or more fingers was higher than the accuracy of two finger searches, which was higher than the accuracy of single-finger searches.

As a rough rule of thumb, at a fixed false accept rate the false reject rate was found to decrease by up to an order of magnitude when using two fingers rather than one, and again when using ten fingers rather than two. Actual differences varied by dataset and by system, but the general trend was quite consistent.

It should be acknowledged, however, that given accurate systems and a relatively limited number of images, a *precise* quantification of the benefit of using of four, eight, and ten-finger sets was not possible in FpVTE 2003. The utility of using an increased number of prints (four or more) is in suppressing false accepts when either a large one-to-many search is needed or when aggregate image quality is reduced.

3b. Poor quality fingerprints greatly reduce accuracy

For all systems, accuracy on high-quality images was much higher than accuracy on low-quality images. Some systems were particularly sensitive to low image quality. For example, at the standard false accept rate of 0.01% the Technoimagia MST accuracy of 82% for the highest-quality fingerprints dropped to 2% for the lowest quality fingerprints[5]. NEC MST achieved an accuracy of 99.8% for the highest-quality fingerprints, which dropped to 84% for the lowest quality fingerprints.

4. Capture devices alone do not determine fingerprint quality

Different operational fingerprint sources can use the same type of collection hardware *and* software and yet result in substantially different performance. The State Department Border Crossing Card (BCC) data and the DHS Recidivist (DHS2) data used the same scanners and software, but are substantially different in overall quality. Using the FpVTE image-quality metric (see Analysis Report), 80% of BCC is high quality, but only 45% of DHS2. Consequently, for most systems, there is a clear difference in accuracy between the two datasets.

Therefore, the subject populations, collection environment, staff training, and equipment maintenance are some of the other factors that are believed to have a substantial impact on fingerprint quality.

5. Accuracy can vary dramatically based on the characteristics, or type, of the data

Performance on one type of data is not necessarily similar to performance on another type of data. The False Reject Rate (one minus the TAR) for a system often varied by a factor of two or more between different datasets.

Some systems showed an unusually high sensitivity to the sources or types of fingerprints; the most accurate systems did not. For example, in SST Cogent had a true accept rate of 99.6% for BCC data and 100% for DHS2, at a false accept rate of 0.1%. At the same false accept rate Bioscrypt had a true accept rate of 97.2% for BCC data and 66.8% for DHS2.

[5] Quality D and F combined. Performing with near-zero errors on low quality prints may indicate a system has an effective mechanism for electing not process such images. If revealed such events are included in a failure to acquire rate. FpVTE ignores FTA by demanding systems return a result no matter what. This yields system level performance.

Any predictions of operational accuracy must account for this important source of variability. Projections from measurements on one type of data to operational performance on another type of data are questionable.

5a. Accuracy on controlled data was significantly higher than accuracy on operational data

All systems were more accurate on the controlled Ohio fingerprints, which were of distinctly higher quality than the operational fingerprints.

5b. Biometric evaluations that only use a single type of data are limited in how systems can be measured or compared

An evaluation that uses a single type of data can measure the accuracy only on that type of data, and may give a misleading impression of overall performance. Likewise, it is not safe to assume that operational performance will closely resemble performance on test data.

In addition, the relative performance of different systems varies by the type of data, so a comparison of systems using one type of data may be very different from a comparison using different data. Rank order among systems was sometimes sensitive to which dataset was selected for comparisons; for this reason, comparisons were based on an aggregate of results.

6. Incorrect mating information is a pervasive problem for operational systems as well as evaluations, and limits the effective system accuracy

The *effective* accuracy of a system is bounded by the mating error rate of the underlying ground truth data. Mating errors were found by trained examiners in every source used in FpVTE. The initial mating errors in most of the datasets used in this evaluation exceeded the matching error rates for the most accurate systems. These ground truth errors were corrected before formal scoring.

Minimizing mating errors in evaluation data is essential to correctly evaluating the accuracy of systems, especially at very low false accept rates or very high true accept rates.

For example, the number of consolidations (cases in which the same person has fingerprint sets under different names or IDs) found and removed in FpVTE was 0.49%. If these had not been found and corrected, then FAR could not have been measured below 0.5%.

References

[303a] "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.

[ATB] Stephen S. Wood and Charles L. Wilson, "Studies of Plain-to-Rolled Fingerprint Matching Using the NIST Algorithmic Test Bed (ATB)" NIST IR7112, April 2004; National Institute of Standards & Technology, Gaithersburg Maryland.

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

[FpVTE2003] C. Wilson, R. A. Hicklin, H. Korves, B. Ulery, M. Zoepfl, M. Bone, P. Grother, R. Micheals, S. Otto, C. Watson, Fingerprint Vendor Technology Evaluation 2003 Analysis Report.

[FRVT2002] P. J. Phillips, P. Grother, R. J. Micheals, D. M. Blackburn, E. Tabassi, and M. Bone, 'Face recognition vendor test 2002, NIST IR 6965, National Institute of Standards & Technology, Gaithersburg Maryland, March 2003

[Grother] Patrick Grother, Face *Recognition Vendor Test 2002 Supplemental Report*. February 2004. NIST IR 7083.

[IDENT] C. L. Wilson, M. D. Garris, and C. A. Watson ,"Matching Performance for the US-VISIT IDENT SystemUsing Flat Fingerprints,"DRAFT NISTIR; National Institute of Standards & Technology, Gaithersburg Maryland.

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

[SDK] Craig Watson, Charles Wilson, Karen Marshall, Mike Indovina, and Rob Snelick, "Studies of One-to-One Fingerprint Matching with Vendor SDK Matchers," DRAFT NISTIR; National Institute of Standards & Technology, Gaithersburg Maryland .

[VTB] Wilson, Watson, Reedy, Hicklin. *Studies of Fingerprint Matching Using the NIST Verification Test Bed (VTB)*; NISTIR 7020; 7 July 2003. (ftp://sequoyah.nist.gov/pub/nist_internal_reports/ir_7020.pdf.