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

Analysis Report

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

FpVTE
Fingerprint Vendor Technology Evaluation

FINGERPRINT VENDOR TECHNOLOGY EVALUATION 2003

ANALYSIS REPORT

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

Sponsor

Justice Management Division, U.S. Department of Justice, IDENT/IAFIS Project Office

Partners

National Institute of Standards and Technology
U.S. VISIT Program, Department of Homeland Security
Federal Bureau of Investigation

Supporters

Office of the Chief Information Officer, U.S. Department of Justice
U.S. Department of Justice
Bureau of Immigration and Customs Enforcement
U.S. Department of Homeland Security
National Institute of Justice
Ohio Office of the Attorney General
U.S. Department of State
European Commission Services
Royal Canadian Mounted Police
U.K. Police Information Technology Organisation

Abstract

The Fingerprint Vendor Technology Evaluation (FpVTE) 2003 was conducted to evaluate the accuracy of fingerprint matching, identification, and verification systems. Eighteen different companies competed, 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. Accuracy varied greatly among the systems tested. The most accurate systems performed consistently well across a variety of tests. Many types and characteristics of fingerprints were analyzed; 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. As the fingerprint image quality improved, the systems' accuracy also improved.

The results of FpVTE are based on (1-to-Many) fingerprint matching technologies and capabilities from 2003 so they are very dated and should only be used accordingly. For results on more recent (1-to-1) fingerprint matching technology capabilities, readers should refer to results posted on the Proprietary Fingerprint Template (PFT) testing website: http://fingerprint.nist.gov/PFT. More recent (1-to-1) fingerprint matching technology capabilities using the exchange of standard minutiae templates are posted on the Ongoing Minutiae Exchange (MINEX) website: http://fingerprint.nist.gov/minex

Table of Contents

Executive Summary **6**

Section 1: Introduction **8**

1.1 Overview 8
1.2 Purpose 8
1.3 Personnel 9

Section 2: Key Concepts and Terminology **10**

Section 3: Test Description **15**

3.1 Overview of Tests 15
3.2 Evaluation Data 18
3.3 Similarity Scores and Matrices 23
3.4 Summary of Test Procedures 23

Section 4: Comparison of Systems **25**

4.1 Methods of Comparison 26
4.2 Multi-Finger Performance (LST) 28
4.3 Single-Finger Flat and Slap Performance (MST) 32
4.4 Single-Finger Flat Performance (SST) 36
4.5 System Anomalies 39

Section 5: Results **40**

5.1 Fingerprint Quality 40
5.2 Effect of Fingerprint Source 44
5.3 Effect of Number of Fingers 46
5.4 Comparison of Verification and Identification Results 51
5.5 Other Results 56

Section 6: Conclusions **70**

Section 7: Future Work **74**

Acknowledgements **76**

Glossary **77**

References **79**

Appendices

Announcement and Website Documents ..... Appendix A
System Description Documents ..... Appendix B
System-Specific Results ..... Appendix C
Details of Results ..... Appendix D
Test Design and Analysis Issues ..... Appendix E

Figures

Figure 1. Sample flat fingerprint. 12
Figure 2. Sample unsegmented 4-finger livescan slap image.. 13
Figure 3. Segmented slap images as used in FpVTE.. 13
Figure 4. Sample rolled fingerprint from a paper source. 14
Figure 5. Range of Accuracy over 27 Operational LST Partitions. 28
Figure 6. Range of accuracy over 17 controlled (Ohio)LST partitions.. 29
Figure 7. Range of accuracy across 7 MST partitions.. 33
Figure 8. ROC for MST Systems.. 35
Figure 9. ROC for MST Systems (Detail) 36
Figure 10. Range of Accuracy on Single-Finger Flats (SST). 37
Figure 11. Fingerprint Quality Distribution by Source in MST 42
Figure 12. Effect of Image Quality (MST) 44
Figure 13. Effect of data source (MST) 45
Figure 14. Effect of data source (LST) 46
Figure 15. Effect of Fingerprint Number and Other Variables in LST. 47
Figure 16. Effect of number of fingers on FBI 12k (slap livescan vs. rolled livescan) 48
Figure 17. Effect of number of fingers on IDENT-IAFIS (slap livescan vs. rolled livescan) 49
Figure 18. Effect of number of fingers on Cogent LST. 50
Figure 19. Comparison of 1:1 and rank-1 evaluation methods.. 52
Figure 20. Comparison of 1:1 and rank-1 1:N evaluation methods (Detail).. 53
Figure 21. Comparison of 1:1, rank-1 1:N, and rankless 1:N evaluation methods (Detail). 54
Figure 22. Rank-based identification performance for MST Systems 55
Figure 23. MST data by image type, controlling for data source.. 58
Figure 24. Flat vs. Slap performance on IDENT-IAFIS data (LST).. 59
Figure 25. Combinations of Slap/Rolled and Live/Paper (FBI 12k data).. 60
Figure 26. Combinations of Slap/Rolled and Live/Paper (Ohio data). 61
Figure 27. Segmented slap little fingers are more difficult to match than other fingers 62
Figure 28. Index fingers outperform thumbs in 2-finger IDENT-IAFIS data 63
Figure 29. No evidence of an effect on accuracy based on sex. 64
Figure 30. Accuracy is lower for older subjects (LST) 65
Figure 31. Accuracy is lower for older subjects (MST) 66
Figure 32. When poor-quality images are excluded, age has no clear effect 66
Figure 33. Effect of Multi-System Fusion of the Top Two MST Systems 68
Figure 34. Effect of Multiple Mates and True Imposters.. 69

Tables

Table 1. FpVTE Personnel....9
Table 2. Summary of FpVTE Tests....15
Table 3. Types and characteristics of the ten datasets that comprised LST (Datasets A-J)....17
Table 4. LST Test Structure. 17
Table 5. Types of fingerprints available from each source....20
Table 6. Sizes and descriptions of sources used in FpVTE. 20
Table 7. Distribution of sources in the FpVTE Datasets....21
Table 8. Types of devices used to collect the livescan fingerprints used in FpVTE. 21
Table 9. LST Systems 25
Table 10. MST Systems....26
Table 11. SST Systems 26
Table 12. Accuracy by threshold over operational LST partitions.. 30
Table 13. Accuracy by threshold over controlled LST Partitions....30
Table 14. Distribution of comparative ranks for 27 operational LST partitions....31
Table 15. Distribution of Comparative Ranks for 17 Non-Operational LST Partitions. 32
Table 16. Distribution of System Rank over 7 MST Partitions, with FAR = 10^{-4} 34
Table 17. Distribution of Comparative System Rank in SST Subtests where FAR = 10^{-3} 38
Table 18. System Anomalies 39
Table 19. Quality Distribution in MST and SST....43
Table 20. Comparison of rank-based identification performance and verification performance for MST systems. 56
Table 21. Correlation of Mate Scores and Ranks for Top MST Systems....67

Executive Summary

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 of the U.S. Department of Justice.

FpVTE 2003 evaluations were conducted from October through November 2003. Eighteen different companies competed, 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. 48,105 sets of fingerprints (393,370 distinct fingerprint images) from 25,309 individuals were used for analysis.

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 viability of fingerprint systems for near-term deployment in large-scale identification 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
- Evaluate scanners or other acquisition devices
- Directly measure performance against very large databases
- Take cost into consideration
- Address latent fingerprint identification

The FpVTE Analysis Report concludes:

1. Of the systems tested, those developed by NEC, SAGEM, and Cogent were shown to be the most accurate.
   - These systems were found to be the most accurate across all FpVTE tests.
   - These systems performed consistently well over a variety of image types and data sources.
2. There was a substantial difference in accuracy among the systems.
   - Many systems performed well on some types of data, particularly on ten-finger tests.
   - There was a clearly measurable difference in accuracy between the most accurate systems and the rest of the systems.
3. 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
4. Capture devices alone do not determine fingerprint quality
5. 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
6. Incorrect mating information is a pervasive problem for operational systems as well as evaluations, and limits the effective system accuracy
7. The most accurate fingerprint systems are far more accurate than the most accurate face recognition systems.

Section 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 (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 carried out in terms of numbers and variety of systems and fingerprints.

FpVTE 2003 used operational fingerprint data from a variety of U.S. and State Government sources. 48,105 sets of fingerprints (393,370 distinct fingerprint images) from 25,309 individuals were used for analysis.

FpVTE was composed of three separate tests, the Small-Scale Test (SST), Medium-Scale Test (MST), and the Large-Scale Test (LST). The SST and MST evaluated matching accuracy using individual fingerprint images. The LST evaluated matching accuracy using ***sets*** of fingerprints, in various combinations of flat, slap, and rolled fingerprint images, with up to ten images per fingerprint set.

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 viability of fingerprint systems for near-term deployment in large-scale identification 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
- Evaluate scanners or other acquisition devices

- Directly measure performance against very large databases
- Take cost into consideration
- Address latent fingerprint identification

1.3 Personnel

A number of people had roles in FpVTE. Table 1 lists the name, affiliations, and role of the staff members that designed and executed the test. In addition, please see Acknowledgements for a list of all people involved in the test, and the roles they played.

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

Section 2: Key Concepts and Terminology

Note: Please see the Glossary for definitions not discussed here.

Face Recognition Vendor Test (FRVT) 2002

FpVTE 2003 analyses and methodologies were built in part on the multi-agency Face Recognition Vendor Test (FRVT) 2002. [FRVT2002]

Query and Target, Probe and Gallery Sets

In any given test or subtest each system searched a Query Set (search set) against a Target Set (file set or fingerprint database). The Target Set represents the enrolled population. The Query Set represents the users of the system, both genuine and imposter.

Each time a system searches a Query Set against a Target Set, it produces an array of similarity scores known as a similarity matrix.

During analysis, a variety of subsets of the Query and Target sets are defined, based on some variable such as source, gender, or quality. The subsets used for a given analysis are generically referred to as dataset partitions. A subset of the Query Set is called the Probe Set; a subset of the Target Set is called the Gallery Set. Analysis was conducted on the resulting test partitions, which were part of the similarity matrix containing scores for the comparison of a Probe Set to a Gallery Set.

During analysis, a "standard" partition was defined for each test (or subtest). The Probe Set for each standard partition includes one fingerprint image (or set) for each user of the system (genuine or imposter); the Gallery Set includes one fingerprint image (or set) for each genuine user (distinct from that in the Probe Set). The standard partition is designed as a maximal subset of the test, i.e., it generally includes any other partitions defined.

Verification and Identification Performance

Fingerprint matching is central to a variety of operational tasks. FRVT 2002 identified three distinct tasks which were called Verification, Closed-Set Identification, and or Open-Set Identification. For each task, appropriate performance statistics were defined.

- In verification (1:1 matching), a subject presents his biometric image to the system and claims to be a person in the system's gallery. For evaluation, each probe image is compared to each gallery image independently. Two performance measures are computed: True Accept Rate (TAR), the fraction of true identity claims scoring above threshold; and False Accept rate (FAR), the fraction of false identity claims scoring above threshold. The resulting relationship between TAR and FAR, where each point is defined as a function of score threshold, may be graphed on a Receiver Operator Characteristic (ROC) curve.
- In closed-set identification (1:N matching), only subjects known to be in the gallery are searched. The system's ability to identify the subject is evaluated based on the fraction of

searches in which the probe image scored at rank k or higher. A probe has rank k if the correct match is the kth largest similarity score. No score threshold is used. The relationship between Identification rate and rank may be graphed on a Cumulative Match Characteristic (CMC) curve.

- In open-set identification[1] (1:N matching), each subject is searched against the gallery, and an alarm is raised if the subject occurs in the gallery. A subject is considered to be "in the gallery" if the probe image scored above the threshold at rank k or higher. In evaluation, the system's ability to detect and identify is measured as two rates: the true accept rate[2], and the false accept rate. An open-set identification ROC plots TAR vs. FAR. This may be generalized using rank, where the subject must be detected and identified at rank k or better.

Note that in a verification (1:1) task, the performance metrics are based on each comparison of a probe image to a gallery image, whereas in the identification (1:N) tasks, the performance metrics are based on each search of a probe image against the entire gallery. The primary evaluation method used in this report was that of verification. Throughout this report, unless otherwise stated, performance was evaluated using verification ROCs. These results were often summarized using "slice" charts, which are cross-sections of ROCs that report TAR at a specific FAR (10^{-4} unless otherwise stated). Slice charts allow cross-comparisons of multiple systems and one or more variables at a fixed FAR; ROCs show the full range of operational performance, but are impractical for showing the effects of variables across multiple systems.

Section 5.4 compares the effects of evaluating by the open set identification (1:N) method against the verification (1:1) method.

Operational identification systems with large databases require very low false match rates. The results of FpVTE are intended to address a number of issues, but projections to operational database sizes are explicitly not part of the scope of this report.

Failures to Enroll

Some systems are designed to reject some fingerprints due to poor image quality. The rate at which this occurs is generally known as the Failure to Enroll (FTE) rate. In FpVTE, Participant systems were not permitted to reject a probe as an FTE: every probe had to be processed by the Participant system. However, a Participant system was permitted to note which fingerprints would have been considered FTE. The results of this information, and its effect on performance, are reported in Section 5.1 Fingerprint Quality.

Flat Fingerprints

A flat fingerprint is a fingerprint image collected from a single-finger livescan device, resulting from the touching of one finger to a platen without any rolling motion. A flat fingerprint is also known as a single-finger plain impression or a touch print. Figure 1 is an

[1] In FRVT 2002, this was described as a "watch list" task.
[2] Also known as the detection and identification rate

example of a "flat" fingerprint image. In FpVTE, the term "flat" fingerprint always means an individual flat fingerprint and should not be confused with a "segmented slap," which is described below.

Figure 1. Sample flat fingerprint. The variation in the background was characteristic of some of the flats used in FpVTE.[1]

Slap Fingerprints

In FpVTE terminology, a segmented slap is an image of a single fingerprint that was segmented (cropped) from an image of a 4-finger slap (4-finger simultaneous impression), such as found at the bottom of a fingerprint card. Slaps may be from livescan devices or scanned from paper fingerprint cards. FpVTE segmented slaps were segmented using both automatic and manual processes. All segmentation was human verified; those images for which automatic segmentation failed were either manually segmented or excluded from the test.

Although flat and slap fingerprints are sometimes both known as plain impressions, the methods of collection and the collection devices used differ substantially, resulting in very different characteristics.

Figure 2 and Figure 3 show an example of a good-quality livescan slap image before and after segmentation.

[1] The sample images in this section were used as sample images in [NIST IQS], which was cleared for public release by DOJ.

Figure 2. Sample unsegmented 4-finger livescan slap image. Unsegmented images were not used in FpVTE: see Figure 3 for the corresponding segmented images. Note that part of the little finger was not included in the slap image: incomplete fingerprints such as this can sometimes occur with any finger in slap images, especially for images from paper sources or from livescans with smaller platens.

Figure 3. Segmented slap images as used in FpVTE. The white background was characteristic of many of the livescan slap and rolled images used in FpVTE.

Rolled Fingerprints

A rolled fingerprint, as illustrated in Figure 4, is a fingerprint image collected by rolling the finger across the livescan platen (or paper) from nail to nail. Rolls may be from livescan devices or scanned from paper fingerprint cards.

Figure 4. Sample rolled fingerprint from a paper source. The paper detail and pencil marks in the background were characteristic of most of the slap and rolled images from paper sources in FpVTE.

Section 3: Test Description

This is a brief description of FpVTE. The complete description and definition of FpVTE was provided on the FpVTE website (http://fpvte.nist.gov). Appendix A of this document includes all of the documents from that website.

3.1 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).

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 one to ten fingerprints collected from an individual subject at one time. The tests were designed so that the SST is a subset of the MST, allowing 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.

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 (21)	14 days

Table 2. Summary of FpVTE Tests

The size and structure of each test were determined to optimize among 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.

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

[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.

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: maximum throughput capacity; the relative proportion of time spent preprocessing images and matching images; the ability to increase throughput rates by decreasing accuracy; and the ability to increase throughput by adding additional hardware. Designing a well-balanced test to accommodate heterogeneous system architectures was a challenge.

To increase the 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, or the test would have been limited 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. Note that two systems that started LST did not complete it because they could not meet the throughput requirements. (See Section 4.5 System Anomalies for detail)

3.1.1 Large-Scale Test (LST)

LST was composed of a series of subtests to measure:

- Performance of rolled fingerprint sets against rolled fingerprint sets. Subtests of 10 fingers per set were conducted.
- Performance of segmented slap fingerprint sets against rolled fingerprint sets. Subtests of 1, 2, 4, 8, and 10 fingers per set were conducted.
- Performance of segmented slap fingerprint sets against segmented slap fingerprint sets. Subtests of 1,2,4,8, and 10 fingers per set were conducted.
- Performance of flat fingerprint sets against flat fingerprint sets. Subtests of 1 and 2 fingers per set were conducted. All fingers were index fingers (fingers 02 and 07).
- Performance of flat fingerprint sets against slap fingerprint sets. One subtest of 1 finger per set was conducted. All fingers were index fingers (fingers 02 and 07).

The rolled and segmented slap images came from livescan devices, or from paper fingerprint cards that were scanned on flatbed scanners. Images from paper cards in some cases included pencil marks, or printed lines and text from the card itself.

LST included ten distinct datasets, labeled A through J. Each dataset contained a different type of fingerprint data, as shown in Table 3. Datasets F, G, and H included a variety of finger positions: please see the Test Plan in Appendix A for more detail.

	Size	Type	# Fingers	Flat	Slap	Rolled	Livescan	Paper
A	8,000	2F	2	x			x	
B	3,000	1F	1	x			x	
C	9,000	10S-L	10		x		x	
D	4,000	10S-P	10		x			x
E	7,000	8S-L	8		x		x	
F	7,200	4S-L	4		x		x	
G	7,000	2S-L	2		x		x	
H	3,100	1S-L	1		x		x	
I	8,000	10R-L	10			x	x	
J	8,000	10R-P	10			x		x

Table 3. Types and characteristics of the ten datasets that comprised LST (Datasets A-J). Please note the image type abbreviations, which are used through this report. For example, 10S-P means 10 Slap fingerprints from Paper sources.

LST Participants were required to perform 31 subtests, using the ten datasets defined in Table 3. The structure of the LST subtests is shown in Table 4.

LST Subtests				Query Sets									
				A 2F 8,000	B 1F 3,000	C 10S-L 9,000	D 10S-P 4,000	E 8S-L 7,000	F 4S-L 7,200	G 2S-L 7,000	H 1S-L 3,100	I 10R-L 8,000	J 10R-P 8,000
Target Sets	A	2F	8,000	AxA	BxA	-	-	-	-	-	-	-	-
	C	10S-L	9,000	-	BxC	CxC	DxC	ExC	FxC	GxC	HxC	-	-
	D	10S-P	4,000	-	-	CxD	DxD	ExD	FxD	GxD	HxD	-	-
	I	10R-L	8,000	-	-	CxI	DxI	ExI	FxI	GxI	HxI	IxI	JxI
	J	10R-P	8,000	-	-	CxJ	DxJ	ExJ	FxJ	GxJ	HxJ	IxJ	JxJ

Table 4. LST Test Structure. LST systems generated a total of 31 similarity matrices, one for each subtest. Note that all ten datasets (A-J) were used as Query sets , but only five datasets were used as Target sets. For example, DxC means that D was the Query set and C was the Target set.

LST had a time limit of no more than 21 days (running continuously), not including setup and checkout. Two Participants that started LST did not complete in time (See Section 4.5, System Anomalies for more information).

3.1.2 Medium-Scale Test (MST)

MST was designed to evaluate matching accuracy using individual fingerprints, all of which were images from right index fingers. The test consisted of a single dataset containing 10,000 files, each of which contained one fingerprint image. 5,000 of the images were single-finger flats, and 5,000 of the images were single-finger segmented slaps. All of the images were from livescan devices.

The MST dataset was used as both the Query Set and the Target Set — in other words, all fingerprints in the dataset were compared against all other fingerprints in the dataset.

MST had a time limit of no more than 14 days (running continuously), not including setup and checkout. All MST Participants completed in time.

3.1.3 Small-Scale Test (SST)

The SST was designed for Participants whose throughput rates would not allow them to complete the MST.

The SST dataset was a subset of the flat fingerprints in the MST dataset. Other than the size and composition of the datasets, MST and SST were conducted in the same way. Slap fingerprints were not included because the size of SST was too small to allow measurement of an additional variable.

SST consisted of a single dataset containing 1,000 files. The SST dataset consisted exclusively of single-finger flat fingerprints from right index fingers. All of the images were from livescan devices.

SST had a time limit of no more than 14 days (running continuously), not including setup and checkout. All SST Participants completed in time.

3.2 Evaluation Data

More than 48,000 sets of fingerprints from more than 25,000 individuals were used in FpVTE. These fingerprints were selected from a pool of millions of sets of fingerprints in operational databases.

Each fingerprint (SST or MST) or fingerprint set (LST) was contained in an ANSI/NIST format file. ANSI/NIST is a standard format for fingerprint files [ANSINIST]. Each file contained

- an image of a single fingerprint (SST and MST), or
- a set of between one and ten fingerprint images (LST)

All images were WSQ compressed, 500 pixels per inch, 8-bit grayscale images. An MD5 message digest (hash) file was provided to Participants for each provided ANSI/NIST file. The MD5 files provided a means of proving that the ANSI/NIST files were correctly copied onto the Participants' systems.

For more information on file formats, please see the *FpVTE Data Format Specification*, in Appendix A.

All of the fingerprints in the FpVTE evaluation datasets were listed as Sensitive data, protected under the U.S. Privacy Act.

3.2.1 Types of Fingerprints

Flat Fingerprints

The flat fingerprints ranged from a minimum size of 368 pixels wide by 368 pixels high to a maximum size of 420 by 480.

All of the flat fingerprints were from the index fingers. In SST and MST, only the right index fingers were used. In LST, both index fingers were used, and the finger position was always noted in the ANSI/NIST file.

The images were usually upright, but were sometimes rotated up to about ±25 degrees, and rarely up to about ±45 degrees. The core was usually (but not always) centered in each image.

Slap Fingerprints

The size of segmented slap fingerprints ranged from a minimum size of 150 pixels wide by 150 pixels high to a maximum size of 500 by 600.

In MST, all slap fingerprints were from the right index finger. In LST, a variety of finger combinations was used, and the finger position was always noted in the ANSI/NIST file.

Slap images (except for thumbs) were usually rotated, as shown in Figure 3. Any rotation in the original image is retained in the segmented images. The average rotation is 20-25 degrees. Few images were rotated more than 45 degrees. Fingers from the left hand are usually rotated clockwise, and those from the right hand are usually rotated counterclockwise.

Rolled Fingerprints

The size of rolled fingerprints ranged from a minimum size of 500 pixels wide by 500 pixels high to a maximum size of 800 by 750.

Rolled fingerprints were only used in LST. The only datasets that included rolled fingerprints included all ten fingers, and the finger position was always noted in the ANSI/NIST file. The rolled images were usually upright, and the core was usually (but not always) centered in each image.

3.2.2 Sources of Fingerprints

The fingerprints in FpVTE were collected from a range of governmental sources:

- Federal Bureau of Investigation (FBI)
- U.S. Department of Homeland Security (DHS)
- U.S. Department of State (DOS)
- U.S. Department of Justice, IDENT/IAFIS Project (DOJ)
- Ohio Office of the Attorney General (Ohio)

Some of the fingerprints are representative of current operational data, others are representative of legacy data, and one set (Ohio) consists of fingerprints collected under controlled conditions. In practice, this means that the test data ranged from good to poor quality, and included a variety of different characteristics. The ability of a fingerprint matcher to accurately match a variety of types and qualities of fingerprints is paramount for a system that will be used in large-scale applications. It is important to note that even new fingerprint

identification systems may be required to search legacy data, and therefore may not have the luxury of being able to operate using a single type or quality of data.

Table 5 and Table 6 summarize the sources of the fingerprints used in FpVTE.

Abbrev.	Name	# Fingers	Flat	Slap	Rolled	Livescan	Paper
DHS2	DHS Recidivist (Illegal Immigrant)	2	x			x	
DOS-BCC	State Department Border Crossing Card	2	x			x	
BEN	DHS Benefits	10		x	x	x	
Identlafis	IDENT/IAFIS Secondary processing	2, 10	x	x	x	x	
Ohio	Ohio	10		x	x	x	x
12k	FBI IAFIS Criminal and Civil	10		x	x	x	x
DHS10	DHS Criminal	10		x	x		x

Table 5. Types of fingerprints available from each source

	Fingerprint sets	Subjects	Description
12k	11,384	6,292	Rolled and slap sets collected from IAFIS workload in Jan-Feb 2000, including one livescan and one paper set per person. Includes civil and criminal subsets.
BEN	5,483	3,882	Rolled and slap livescan sets collected from BICE Benefits data. Operational fingerprints collected in an office environment. The fingerprints were from cooperative subjects applying for resident green cards, etc. from BICE.
DHS10	5,697	4,337	Rolled and slap sets collected from DHS Criminal data. These were livescan images that were printed onto 10-print cards and rescanned. This is a process known to degrade image quality.
DHS2	2,491	1,501	2-finger flat fingerprints, collected by the Border Patrol, often under difficult conditions. These are from recidivist illegal immigration cases, the majority of which are Mexican border crossing cases. Also known as IDENT Recidivist.
DOS-BCC	7,669	4,566	2-finger flat operational fingerprints collected by the U.S. Department of State (DOS) for Border Crossing Cards (BCCs) in US Consulates in Mexico, in an office environment. Also known as the Mexican Visa database. These are the fingerprints corresponding to the face images used in FRVT2002.
Identlafis	9,972	3,806	Rolled and slap livescan sets collected in secondary processing for IDENT/IAFIS. Also includes mated flat and slap livescan sets.
Ohio	5,409	925	Three sets of livescan slap or rolled/slap fingerprints collected from each of 925 prisoners under controlled conditions, along with two paper rolled/slap fingerprint sets.
Total	48,105	25,309	

Table 6. Sizes and descriptions of sources used in FpVTE. Note the variations in subject populations, in subject cooperation, and collection conditions.

Table 7 summarizes the distribution of source data in the various datasets.

	SST	MST	LST A	LST B	LST C	LST D	LST E	LST F	LST G	LST H	LST I	LST J
DHS2	30%	13%	15%	27%	-	-	-	-	-	-	-	-
DOS-BCC	70%	27%	75%	53%	-	-	-	-	-	-	-	-
BEN	-	8%	-	-	27%	-	21%	17%	17%	10%	19%	-
Identlafis	-	32%	10%	20%	32%	-	35%	32%	34%	18%	36%	-
Ohio	-	20%	-	-	16%	25%	15%	19%	16%	32%	10%	13%
12k	-	-	-	-	25%	37%	29%	32%	32%	40%	35%	35%
DHS10	-	-	-	-	-	38%	-	-	-	-	-	53%

Table 7. Distribution of sources in the FpVTE Datasets. The LST datasets A-J were summarized in Table 3.

Having a variety of sources of fingerprints is critical when attempting to evaluate operational accuracy. If a single source of fingerprints had been used in FpVTE, the results could have been very misleading, because the results would only be for a single type of fingerprint, yet could be easily misconstrued as applicable to all fingerprints.

Fingerprint quality is clearly related to the source of the fingerprints, as will be discussed in Section 5.1.

3.2.3 Fingerprint Collection Methods

The livescan fingerprints used in FpVTE were collected using a variety of livescan devices, as shown in Table 8.

Livescan Devices Used to Collect FpVTE Fingerprints				
Device	FBI Certification	Flats	Slaps	Rolls
CrossMatch 442	-		2.3%	
CrossMatch ID1000	F		33.1%	29.3%
DBI 1133S5	G		19.5%	17.5%
Identicator DFR-90	-	86.3%		
Identix TP2000	F		2.3%	10.0%
Identix TP600	G		0.2%	1.4%
Ricoh IS-510	-		1.9%	2.2%
Smiths Heimann LS2 Check	-		11.5%	
(Unknown)		13.7%	29.3%	39.6%
Total		*100%*	*100%*	*100%*

Table 8. Types of devices used to collect the livescan fingerprints used in FpVTE. Note that most of the slap and rolled livescan devices were certified by the FBI[1], but the single-finger flat devices were not. Also note that the scanner model was not known in many cases.

FpVTE had no role in the selection of these devices: the listing of makes and models does not imply a recommendation by NIST or FpVTE personnel, but simply recognizes the actual devices used by the agencies that contributed data to FpVTE.

[1] The FBI's Image Quality Specification is EFTS Appendix F; Appendix G was an interim certification. [EFTS] Only devices capable of capturing rolled prints are certified by the FBI. The list of certified devices is in [IAFISCert].

The type of scanner used to acquire each fingerprint was not revealed in the tests. This was because this information is often not available in operational databases.

The slap and rolled fingerprints collected from paper sources were usually from inked fingerprint cards. It is believed that most or all were scanned using FBI EFTS Appendix-F compliant flatbed scanners. However, one source of fingerprints (DHS10) consisted of livescan fingerprints that were printed, then rescanned on a flatbed scanner. This is a process known to degrade image quality, but that was used in the acquisition of some fingerprints in operational databases.

3.2.4 Data Selection Issues

A variety of issues were considered during the data selection process:

- Publicly available datasets (such as the NIST Special Databases [NIST SD29]) were inappropriate for evaluation.
- The fingerprints were randomly sampled from the original sources so that they were representative of the original datasets.
- A variety of sources were selected to allow measurement of the effect of sources, and to avoid dependence on any one source.
- The bulk of the fingerprints needed to be operationally collected, not collected under controlled conditions (see Section 1.1 of Appendix E for a discussion of problems due to controlled collection of data).
- The mated fingerprints must have been selected by means other than fingerprint matching (see Section 1.2 of Appendix E for a discussion of matcher bias).
- No fingerprints were excluded from the test because of poor quality (see Section 1.3 of Appendix E for a discussion of bias due to data filtering).
- All slap fingerprints were segmented by automatic or manual methods and visually verified by human review.

After the completion of the test, the results of all systems were analyzed, and anomalies were reviewed. This process, known as "exception analysis" or "groundtruthing," is critical to a technical evaluation because mating errors in the underlying datasets can potentially exceed error rates of the matchers themselves.

Many datasets, in operational systems or in evaluation databases, have mating errors. For example, after FRVT 2002 was published, the fingerprints associated with the facial images were used to double-check identities, and it was discovered that 1.7% of the images had incorrect mating information.

Section 1 of Appendix E discusses data selection and validation issues in detail. Of the 48,105 fingerprint sets in FpVTE, 5,174 were visually reviewed, in 3,177 pairs. Some datasets were more susceptible to some kinds of errors. All datasets had errors, including the controlled Ohio data.

124 consolidations (cases in which the same person has fingerprint sets under different names or IDs) were found and removed in FpVTE: this was 0.49% of the people in all the FpVTE datasets. However, most of these consolidations were due to the use of fingerprints

from different sources; people who were in multiple databases caused consolidations when the databases were merged into the FpVTE datasets. Only 24 of these consolidations (0.09%) were errors in the original sources.

119 misidentifications (cases in which fingerprints from different people are listed under the same name or ID) were found and removed in FpVTE: this was 0.47% of the people in all the FpVTE datasets. All of these misidentifications were errors in the original sources.

Note that if the 0.49% consolidation rate had not been found and corrected, then FAR could not have been accurately measured beyond 0.49% If the 0.47% misidentification rate had not been found and corrected, then TAR could not have been accurately measured beyond 99.53%.

3.3 Similarity Scores and Matrices

The output from each FpVTE test (or subtest) was a matrix of system-specific measures of similarity known as similarity scores. For most Participants, these measures correspond to matcher scores. Note that the systems did not return match vs. non-match determinations. Each similarity score measured the similarity of the fingerprints in an ANSI/NIST file to those in another ANSI/NIST file:

- In SST and MST, each ANSI/NIST file contained a single fingerprint, so a similarity score was a measure of the similarity of an individual fingerprint to another individual fingerprint.
- In LST, each ANSI/NIST file contained from one to ten fingerprints (collected at one time from an individual), so a similarity score was a single measure of the similarity of a set of fingerprints to another set of fingerprints. Put another way, in LST, when comparing a set of ten fingerprints to another set of ten fingerprints, one similarity score was put in the similarity matrix, not ten scores.

FpVTE analyses are based on distributions of similarity scores for match vs. non-match comparisons. A higher score necessarily means a higher degree of similarity.

The scale used for similarity scores was entirely up to each Participant: one system might have used a scale of 0.0 (no similarity) to 1.0 (identical), while another may have used a scale of -1,000,000 to 1,000,000. The exact format of a similarity matrix is defined in the *FpVTE Data Format Specification*, included in Appendix A.

In all three tests, some datasets were searched against themselves. Obviously such searches generated self-idents (comparisons of an image against itself), which were ignored in the analysis.

3.4 Summary of Test Procedures

Note: This is a brief summary of the FpVTE test procedures. The test procedures are carefully delineated in the FpVTE Test Procedures Document, which is included in Appendix A.

FpVTE 2003 tests were administered at the NIST facilities at Gaithersburg, MD, between September 29 and December 1, 2003. Not all Participants started the test on the same day.

FpVTE was designed with a variety of measures to safeguard the integrity of the test. In practice, these measures serve to guarantee that the test was fair, and that sensitive data could not be compromised.

Some of the test integrity measures taken in FpVTE include the following:

- The size and structure of the test were designed to balance the throughput limitations of the prospective participants as much as possible.
- The FpVTE Website and FAQs were created to facilitate open and impartial dialogue with the participants, and to ensure that all participants received information at the same time.
- Sample datasets were provided in advance to tentative Participants. These datasets were representative of the FpVTE evaluation datasets in format, but other characteristics, such as image quality and collection device differed because sensitive data from the evaluation data sources could not be released.
- After each Participant's equipment was set up, the Participant was required to run the FpVTE Trivial Datasets, which had the same structure as the actual Evaluation Datasets but had significantly fewer files. The trivial datasets provided a last-minute proof before the start of a test that the system was able to process the data.
- MD5 message digests (hash files) were provided for every fingerprint file as a means for the participants to prove that the fingerprints being processed could not have been subject to data transmission errors.
- A variety of administrative controls were put in place to minimize access to systems during the test. All subtests were required to run without human administration. Operators were allowed only very limited access, and only under the supervision of a test agent. All system accesses were logged by the test agents, and videotaped.
- Participants were required to produce an MD5 message digest (hash file) for each similarity file they produced as a means of verifying data integrity.
- Systems were purged of data at the completion of the test.
- Using the NIST SDK testbed, results of the evaluation can be corroborated at NIST; this has already been done for the most accurate FpVTE systems.

Section 4: Comparison of Systems

Eighteen companies provided systems for testing, in addition to NIST's fingerprint benchmark system, the VTB [VTB]. A total of 34 tests were successfully completed: 13 in LST, 18 in MST, and 3 in SST. The following tables 9 thru 11 provide an overview of the evaluated systems. BIO-key was scheduled to take MST and LST, but withdrew after the anonymous drop date. The reason given was: "Our hardware vendor failed to supply committed hardware in time for testing."

Participants were required to submit a *System Description Document* detailing each system. The contents of these documents are included in Appendix A.

The NIST VTB is being released as part of the NIST Fingerprint Image Software (NFIS), which is available free of charge, but is export controlled. [NFIS]

System ID	System Name	Software	Hardware
Antheus (LST)	Agora	FpVTE Application Software, Windows 2000	1 dual Pentium4
Biolink (LST)	Authenteon-based FpVTE	FpVTE Application Software v66, Windows Server 2003	2 Pentium4s, 24 Blades, 1 Pentium3
Cogent (LST)	Galaxy V8.60 for LST	Galaxy V8.60 for LST (minutae matching), Windows 2000	6 IBM xSeries 335 Server, dual CPU
Dermalog (LST)	DermalogFingerCode3	DermalogFingerCode3 kernel, Windows XP Professional	6 Pentium4s
Golden Finger (LST)	GAFIS	GAFIS 5.0 Engine Network Edition, Windows 2003 Enterprise Server	4 dual Xeon
Griaule (LST)	Griaule AFIS	Griaule AFIS	10 dual Xeon
Identix (LST)	BioEngine Software Developer's Kit (SDK)	BioEngine SDK v4.0 step 1, Windows 2000 Advanced Server	IBM X Series 8 Blade System
Motorola (LST)	Motorola Omnitrak	(Modified) Omnitrak Software, Windows 2000 Server	16 dual Xeon
NEC (LST)	NEC Cluster-PC Matching Server for LST	NEC Application Software, Windows 2000	10 dual Xeon
NIST VTB (LST)	VTB	Bozorth98 Matcher, Red Hat Linux 7.2	20 dual Xeon
Raytheon (LST)	RAYAFIS	RAYAFIS Software, Microsoft Server 2003	2 Pentium4; 2 dual Xeon; 1 Pentium
SAGEM L1 (LST)	SAGEM MetaMorpho	MetaMorpho Software version 3.1.2A.NST.3A, Windows 2000 Professional	7 Pentium4s
SAGEM L2 (LST)	SAGEM MetaMorpho	MetaMorpho Software version 3.1.2A.NST.3A, Windows 2000 Professional	3 Pentium4s

Table 9. LST Systems

System ID	System Name	Software	Hardware
123 ID M2 (MST)	Biometric System Search (BSS)	BSS	16 HP xw6000 search nodes + HP a330n
Antheus (MST)	Agora	FpVTE Application Software, Windows 2000	1 dual Pentium4
Avalon (MST)	Ultramatch AFIS	Ultramatch, Windows XP Professional	1 Pentium4
Biolink (MST)	Authenteon-based FpVTE	FpVTE Application Software v66, Windows Server 2003	1 Pentium4
Cogent (MST)	Galaxy V3.2 for MST	Galaxy V3.2 for MST (template matching), Windows 2000	1 IBM xSeries 335 Server, dual CPU
Golden Finger (MST)	GAFIS	GAFIS 5.0 Engine Network Edition, Windows 2003 Enterprise Server	4 dual Xeon
Identix (MST)	BioEngine Software Developer's Kit (SDK)	BioEngine SDK v4.0 step 1, Windows 2000 Advanced Server	IBM X Series 8 Blade System
Motorola (MST)	Motoral Omnitrak	(Modified) Omnitrak Software, Windows 2000 Server	16 dual Xeon
NEC (MST)	NEC Cluster-PC Matching Server for MST	NEC Application Software, Windows 2000	3 dual Xeon
Neurotechnologija M1 (MST)	VeriFinger	VeriFinger 4.2 software test version, Windows XP Professional	1 Pentium4
NIST VTB (MST)	VTB	Bozorth98 Matcher, Red Hat Linux 7.2	20 dual Xeon
Phoenix (MST)	AFIX Tracker	AFIX Tracker v4.4, Windows XP Professional	1 Pentium4
Raytheon (MST)	RAYAFIS	RAYAFIS Software, Microsoft Server 2003	2 Pentium4
SAGEM M1 (MST)	SAGEM MetaMorpho	MetaMorpho Software version 3.1.2A.NST.2A, Windows 2000 Professional	1 Pentium4
SAGEM M2 (MST)	SAGEM MetaMorpho	MetaMorpho Software version 3.1.2A.NST.2A, Windows 2000 Professional	1 Pentium4
Technoimagia (MST)	FP-Workstation	Fingerprint Aunthentication System Tool21 ("FAST21"), Windows XP	1 dual Xeon
UltraScan M1 (MST)	IDExpress Developer	IDExpress Developer version 1.5.3, Windows XP Professional	1 Pentium4
UltraScan M2 (MST)	IV&V	IVV version 2.00, Windows XP Professional	1 Pentium4

Table 10. MST Systems

System ID	System Name	Software	Hardware
Bioscrypt (SST)	Bioscrypt Core	Bioscrypt Core, Windows XP Professional	1 dual Xeon
Cogent (SST)	Galaxy V3.2T for SST	Galaxy V3.2T for SST (image matching), Windows 2000	1 IBM xSeries 335 Server, dual CPU
NIST VTB (SST)	VTB	Bozorth98 Matcher, Red Hat Linux 7.2	1 dual Xeon

Table 11. SST Systems

4.1 Methods of Comparison

Verification (1:1 matching) performance was used as a basis for comparing the accuracy of systems. Measures of identification performance were also used (see Section 5.4) but found to yield essentially the same conclusions. Several decisions were made in the analysis process that affected the final ordering of the systems, including the choice of the data partitions used for comparison, the choice of the operating point for comparison, and methods of

differentiation between systems. Section 2 of Appendix E discusses in detail how verification performance was applied to compare the systems.

In each test, systems were compared across a variety of distinct partitions of the data. These partitions were selected to reflect distributions of operational and controlled data, to expose system strengths and weaknesses by including various types of data, to limit redundant comparisons, and to favor statistically meaningful comparisons.

In LST, 44 partitions of the data were created to measure the effects of a variety of variables, including:

- Source
- Operational versus controlled data
- Number of fingers (from 1 to 10)
- Livescan versus paper
- Flat versus slap versus rolled fingerprints

The difference in performance between operational and controlled data was substantial enough that the LST results are reported separately for the 27 operational partitions and 17 controlled partitions.

In MST, seven distinct (non-overlapping) partitions based on source and fingerprint type (flat versus slap) were used to measure the range of performance. Six of the seven partitions contained operational data.

Due to the limited size of SST, only two distinct partitions, based on source, were available to measure the range of performance. Both of the partitions contained operational data.

These partitions are described further in Appendix D.

All comparisons were based on TAR as measured at FAR = .01%, except in SST, where comparisons were based on FAR = .1%. This does not imply that a FAR of .01%is necessarily appropriate for an operational system. A FAR of .01% is a reasonable point for comparison because results at that point can be directly measured without the need for projection or concerns about statistical significance, and because the rank order of the more accurate systems generally remained constant for lower values of FAR.

The charts in this section compare the range of accuracy for each system across all of the partitions used for comparison, and where possible differentiate between the results for operational and controlled data. For every partition, the systems were ranked in order by TAR. Tables in this section show the distribution of these ranks across all of the partitions.

When comparing systems, their relative accuracy depends on the choice of performance metrics and how they are applied. However, when a variety of comparisons are made and aggregated, the resulting distribution of rank order was found to be very stable, especially among the more accurate systems.

Details of all comparison results are included in Appendix D.

4.2 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 consistently accurate in their performance than others. Figure 5 shows the range of performance over 27 representative test partitions of operational data.

Each line depicts a summary statistic for the systems' performance over the 27 partitions, characterizing accuracy (TAR) as measured (or interpolated) at FAR = .01%. For example, the line labeled "Average" shows for each system the average of 27 separate TAR measurements, each at FAR = .01%. The maximum accuracy for each system was perfect or near-perfect. Since maximum and minimum values are often outliers, the 5th highest and 5th lowest accuracies over the 27 partitions are also shown.

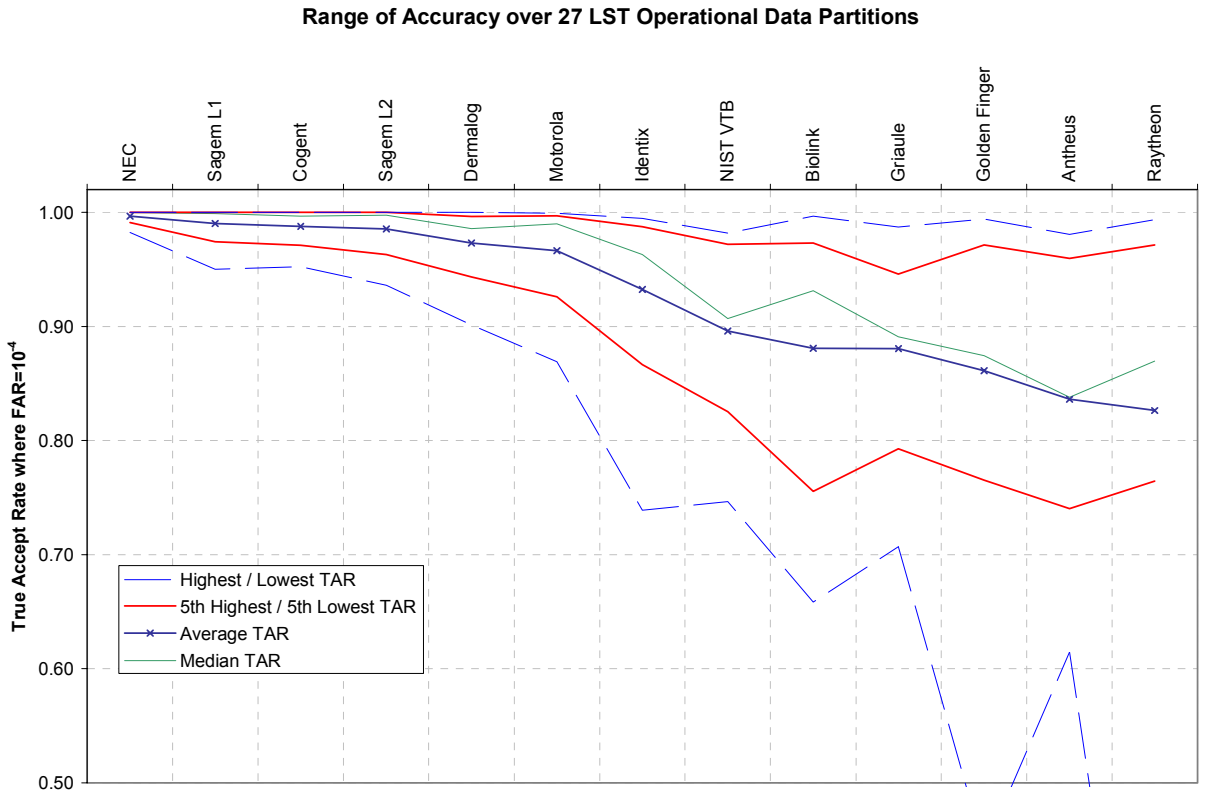

Figure 5. Range of Accuracy over 27 Operational LST Partitions. The systems are sorted by their average accuracy over the 27 partitions; note that sorting by median performance would change the order for some systems.

Figure 6 shows similar results for 17 partitions of controlled (Ohio) data. The red lines show the 3rd highest and 3rd lowest accuracies over the 17 partitions.

The Ohio fingerprints are of distinctly higher quality than the operational fingerprints. Note that all systems do better on controlled data, but the amount of this improvement varies from system to system.

In both charts, note the disparity between the highest and lowest accuracies, a span that is often more than an order of magnitude (in terms of false reject rate, which is 1 - TAR). Note also that the median TAR is higher than the average, indicating that a few problematic data sets drag the average down.

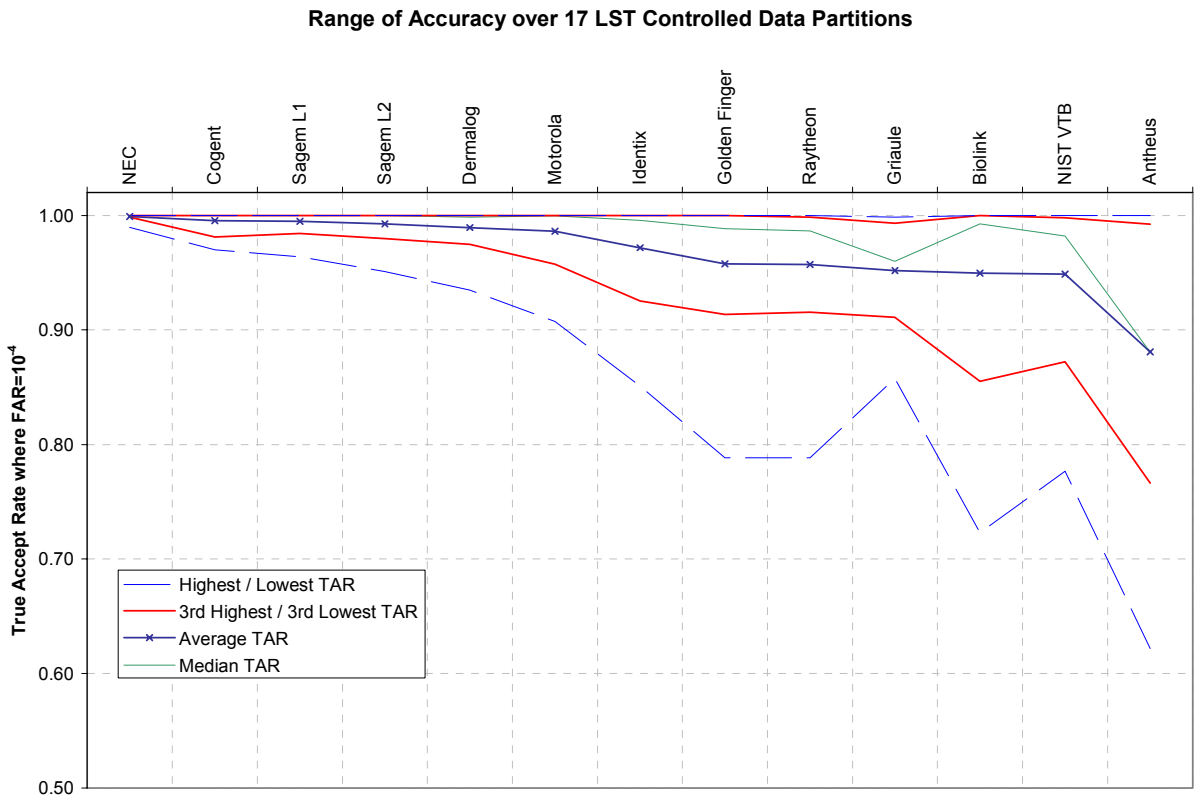

Figure 6. Range of accuracy over 17 controlled (Ohio)LST partitions. The systems are sorted by their average accuracy over the 17 controlled partitions; note that when compared to Figure 5, Cogent and SAGEM L1 changed positions, as well as every system to the right of Identix.

Table 12 and Table 13 provide a different perspective of the same results, by showing the accuracy of the LST systems over the operational and non-operational partitions by TAR threshold. For example, for the operational partitions, NEC had a TAR of 100% for 13 of the 27 partitions. Note that for the operational partitions, three systems always had a TAR above 95%. For the non-operational partitions, four systems always had a TAR above 95%.

Accuracy Over 27 Operational LST Partitions (FAR=10^{-4})	Number of partitions where TAR is			
	= 1.0	>= 0.99	>= 0.95	>= 0.90
NEC	13	23	27	27
Sagem L1	8	19	27	27
Cogent	8	17	27	27
Sagem L2	7	17	24	27
Motorola	0	13	20	25
Dermalog	0	12	21	27
Identix	0	4	16	19
Biolink	0	1	12	14
Golden Finger	0	1	9	13
Raytheon	0	1	8	12
Antheus	0	0	6	10
NIST VTB	0	0	5	15
Griaule	0	0	4	10

Table 12. Accuracy by threshold over operational LST partitions. This shows the number of the 27 operational partitions for which the TAR was at or above 1.0, 0.99, 0.95, and 0.90.

Accuracy Over 17 Controlled LST Partitions (FAR=10^{-4})	Number of partitions where TAR is			
	= 1.0	>= 0.99	>= 0.95	>= 0.90
NEC	13	16	17	17
Sagem L1	12	14	17	17
Sagem L2	12	13	17	17
Cogent	11	14	17	17
Motorola	8	12	15	17
Dermalog	5	12	16	17
Identix	5	11	13	15
Biolink	4	10	12	13
Golden Finger	3	8	12	15
Raytheon	2	8	12	15
NIST VTB	2	5	12	14
Antheus	1	3	7	8
Griaule	0	4	12	15

Table 13. Accuracy by threshold over controlled LST Partitions. This shows the number of the 17 controlled partitions for which the TAR was at or above 1.0, 0.99, 0.95, and 0.90.

To facilitate comparison between systems, all of the LST systems were ranked based on TAR as measured at FAR = .01%

, for each of the 27 operational and 17 controlled partitions. For example, the system with the highest TAR (or tied for the highest TAR)[1] for a partition was given a rank 1 for that partition; the system with the lowest TAR was given a rank of 13. The distributions of these comparative ranks are shown in Table 14 and Table 15.

[1] In these results, a tie for rank 1 means that the tied systems had a TAR of 1.0 for that partition.

The systems are sorted by the average rank in each table. Note that this ordering differs somewhat from that in Figure 5 and Figure 6, which are sorted by the average TAR.

These rankings are useful for comparison, but do not measure absolute accuracy: the lowest TAR (therefore rank 13) for a partition ranged from 11.6% to 99.7% ..00. One controlled partition had 11 systems tied for rank 1 with a TAR of 1.0. For these reasons, the best and worst values in these tables should be considered outliers.

In Table 14 and Table 15, note that

- NEC was the most accurate system for almost every operational partition, and for every controlled partition. (NEC was in rank 1 for 42 of the 44 partitions)
- For operational partitions, Dermalog and Motorola were almost always in rank 5 or 6.
- For operational partitions, NIST VTB, Griaule, Golden Finger, Biolink, and Raytheon alternated positions, but were usually in ranks 8 through 12.

Summary of Rank over LST Partitions where FAR = 10^{-4} 27 LST Partitions of Operational Data	Best	5th Best	Average	Median	5th Worst	Worst
NEC	1	1	1.2	1	1	4
Sagem L1	1	1	2.1	2	3	4
Cogent	1	1	2.8	3	4	6
Sagem L2	1	1	2.9	3	4	5
Dermalog	3	5	5.2	5	6	6
Motorola	4	5	5.6	6	6	6
Identix	7	7	7.5	7	8	11
NIST VTB	7	8	9.8	10	11	13
Griaule	7	8	9.9	10	12	13
Golden Finger	8	8	10.0	10	12	12
Biolink	7	8	10.1	10	12	13
Raytheon	8	8	10.3	10	12	13
Antheus	8	9	11.9	13	13	13

Table 14. Distribution of comparative ranks for 27 operational LST partitions. For example, NEC had the highest TAR (or was tied for highest TAR) for almost all of the partitions, but had the 4th highest TAR for at least one partition.

Summary of Rank over LST Partitions where FAR = 10^{-4}						
17 LST Partitions of Controlled (Ohio) Data						
	Best	3rd Best	Average	Median	3rd Worst	Worst
NEC	1	1	1.0	1	1	1
Sagem L1	1	1	1.5	1	3	3
Cogent	1	1	1.6	1	3	4
Sagem L2	1	1	1.9	1	4	4
Motorola	1	1	3.6	5	6	7
Dermalog	1	1	4.8	5	8	9
Identix	1	1	5.3	7	8	9
Biolink	1	1	7.6	8	12	12
Golden Finger	1	1	8.3	9	11	11
Raytheon	1	9	8.9	10	11	11
NIST VTB	1	8	9.4	10	12	13
Griaule	6	7	10.2	12	13	13
Antheus	1	11	11.6	13	13	13

Table 15. Distribution of Comparative Ranks for 17 Non-Operational LST Partitions. For example, NEC had the highest TAR (or was tied for highest TAR) for every partition.

4.3 Single-Finger Flat and Slap Performance (MST)

In MST, the fingerprints were grouped by source and type to create seven partitions that were used to measure the range of performance. The results for each participant for each source were calculated and analyzed. The resulting range of accuracy on seven single-finger tests is shown in Figure 7.

Since the best and worst values are often outliers, the space between the second best and second worst is instructive. In the LST comparison, the highest TAR was often 1.0, and there was a minimal difference between the top several values. Here, however, there is a substantial difference between the best and second best values. The reason for this is that one of the seven partitions here is composed of the controlled Ohio data, which is the best source for almost all systems: the other six partitions are all operational data, so the second best source is a better measure of the limit of performance on operational data.

Descriptions of the partitions and details of all comparison results are included in Appendix D.

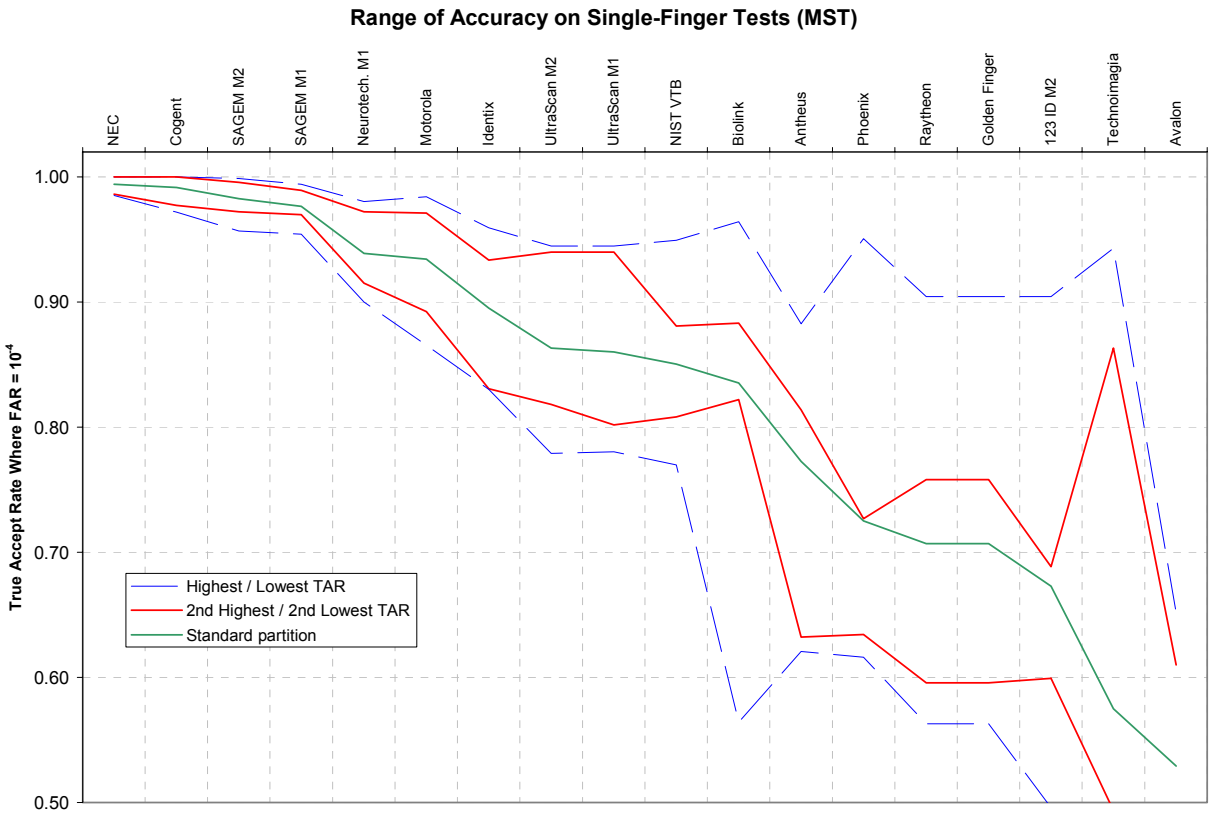

Figure 7. Range of accuracy across 7 MST partitions. These systems are sorted by the systems' performance on the standard MST, which is simply the combination of the other partitions.[1]

To facilitate comparison, all of the MST systems were ranked in order of TAR where FAR was .01%[···], for each of the seven partitions. The distribution of these ranks is shown in Table 16. These systems are sorted by the average rank over all seven partitions.

[1] 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.

Summary of Rank over 7 MST Partitions where FAR = 10^{-4}						
System	Best	2nd Best	Average	Median	2nd Worst	Worst
NEC	1	1	1.1	1	1	2
Cogent	1	1	1.7	2	2	3
SAGEM M2	2	3	2.9	3	3	3
SAGEM M1	4	4	4.0	4	4	4
Neurotech. M1	5	5	5.4	5	6	6
Motorola	5	5	5.6	6	6	6
Identix	7	7	8.1	8	9	10
UltraScan M2	7	7	8.7	9	10	12
UltraScan M1	7	7	8.7	9	10	12
Biolink	7	8	9.7	9	11	15
NIST VTB	8	10	10.4	11	11	12
Antheus	12	12	13.4	13	15	17
Technoimagia	11	11	13.7	13	18	18
Phoenix	9	14	13.9	14	16	16
Raytheon	11	13	14.3	15	16	16
Golden Finger	11	13	14.3	15	16	16
123 ID M2	13	14	15.3	15	17	17
Avalon	17	17	17.7	18	18	18

Table 16. Distribution of System Rank over 7 MST Partitions, with FAR = 10^{-4}

Several details in Table 16 should be noted:

- The order of the top four systems is stable and the accuracy of these systems was clearly separated. Note that SAGEM M2 is 3rd in 6 of 7 tests, and SAGEM M1 is 4th for every partition.
- Neurotechnologija and Motorola are difficult to differentiate: they are 5th or 6th for every partition.
- Although the two Ultrascan systems are identical in rank, their ROCs are different.
- The Golden Finger and Raytheon MST systems produce identical ROCs for every partition.
- Technoimagia has an abrupt drop in performance. Its relative position would worsen dramatically if a smaller FAR were chosen.

Figure 8 shows the ROC for all MST systems for the complete MST (with detail of the same graph in Figure 9).

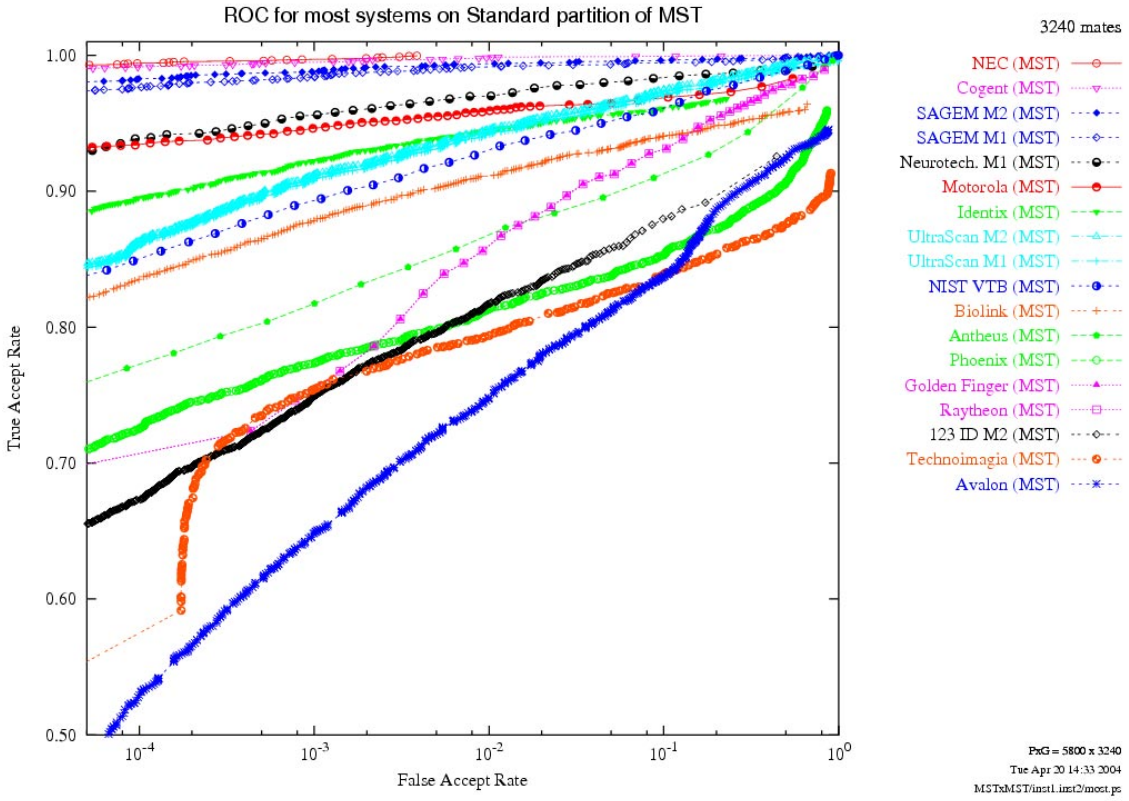

Figure 8. ROC for MST Systems. The "Standard partition" line from Figure 7 is a cross-section of this chart at FAR=10^{-4}.

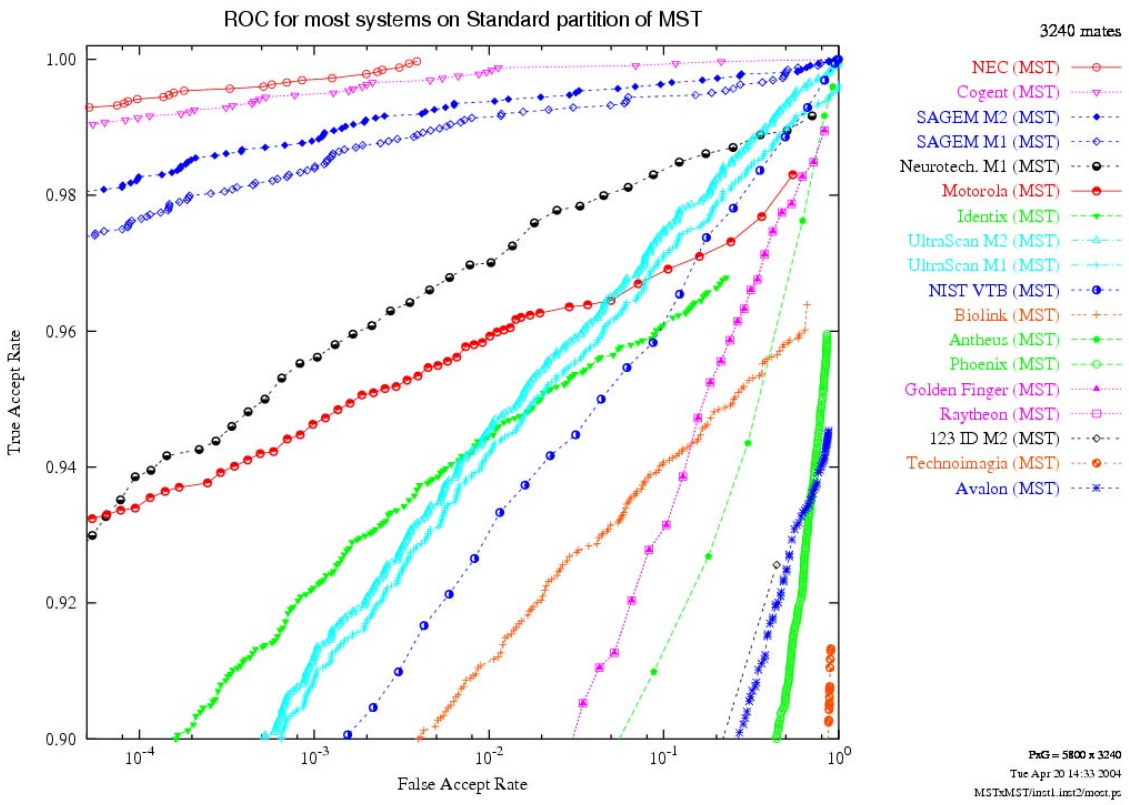

Figure 9. ROC for MST Systems (Detail, with TAR of 0.90 and above)

4.4 Single-Finger Flat Performance (SST)

SST was a small test that only included 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 Figure 10. These systems are sorted by the systems' performance on the SST standard partition, which is simply the combination of the other two partitions.

Note that these results are at FAR=.1%, ***not*** .01% as is true for most of the charts in this report. This is due to the size of SST.

Descriptions of the partitions and details of all comparison results are included in Appendix D.

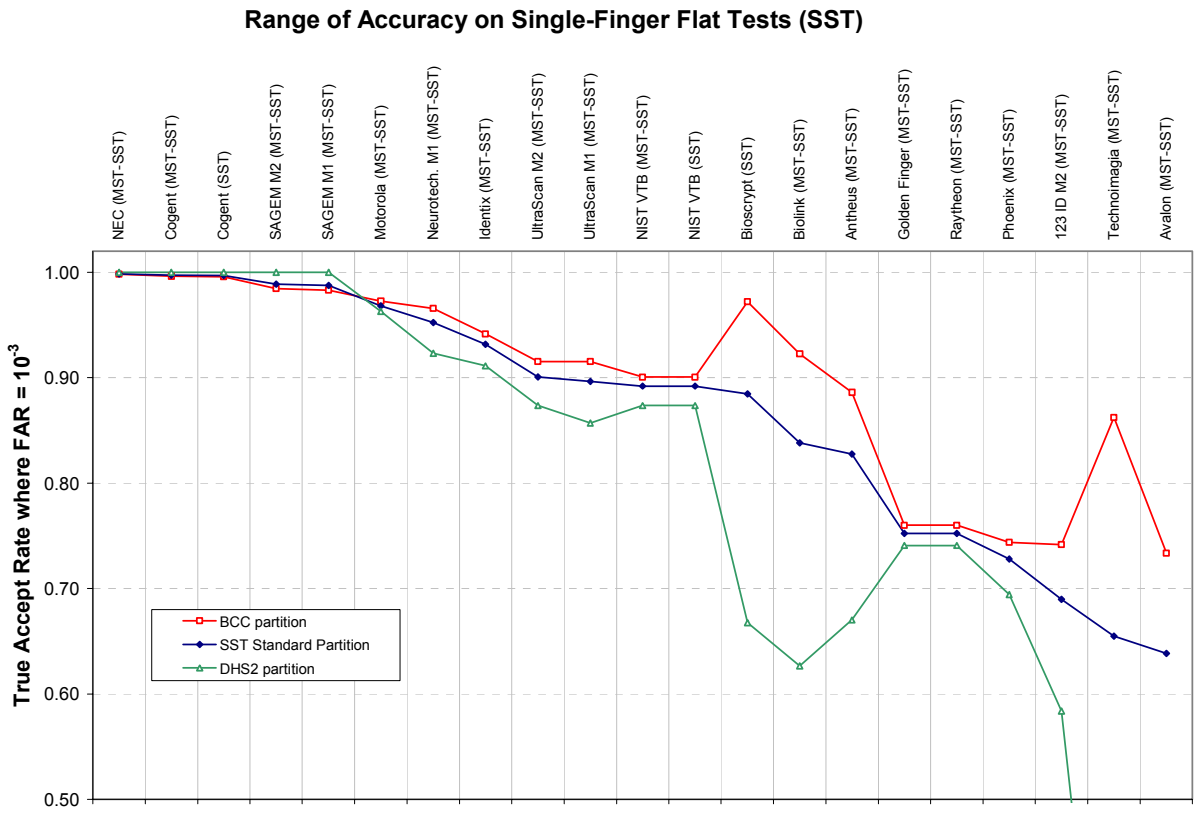

Figure 10. Range of Accuracy on Single-Finger Flats (SST). These systems are sorted by performance on the SST standard partition. Note that these results are reported at FAR = 10^{-3}

Several details should be noted:

- Only three systems took the SST by itself. However, the NIST SST and MST systems were identical.
- The Cogent MST and SST systems had nearly identical results
- A few systems have a dramatic difference between their BCC and DHS2 results. Since DHS2 has a greater proportion of poor-quality fingerprints, this difference may be viewed as a particular sensitivity to poor-quality data.
- Despite the poor quality of DHS2, the NEC, Cogent, and SAGEM systems all achieved a TAR of 100% at this FAR (a 5-way tie for rank 1).
- Bioscrypt performs well on the BCC data; in a test composed solely of BCC-type fingerprints, its rank would have been between Motorola and Neurotechnologija.

Table 17 shows the system rank information corresponding to Figure 10.

Summary of Rank for SST Partitions where FAR = 10^{-3}			
	System Rank		
System	Standard	BCC	DHS2
NEC (MST-SST)	1	1	1
Cogent (MST-SST)	2	2	1
Cogent (SST)	3	3	1
SAGEM M2 (MST-SST)	4	4	1
SAGEM M1 (MST-SST)	5	5	1
Motorola (MST-SST)	6	6	6
Neurotech. M1 (MST-SST)	7	8	7
Identix (MST-SST)	8	9	8
UltraScan M2 (MST-SST)	9	11	11
UltraScan M1 (MST-SST)	10	11	12
NIST VTB (MST-SST)	11	13	9
NIST VTB (SST)	11	13	9
Bioscrypt (SST)	13	7	17
Biolink (MST-SST)	14	10	18
Antheus (MST-SST)	15	15	16
Golden Finger (MST-SST)	16	17	13
Raytheon (MST-SST)	16	17	13
Phoenix (MST-SST)	18	19	15
123 ID M2 (MST-SST)	19	20	19
Technoimagia (MST-SST)	20	16	21
Avalon (MST-SST)	21	21	20

Table 17. Distribution of Comparative System Rank in SST Subtests where FAR = 10^{-3}

4.5 System Anomalies

Some systems did not successfully complete FpVTE. Sometimes these systems were able to successfully retake the test.[1] The results of the successful run are included in the body of the report. The results of the unsuccessful runs are not included in the body of the report. Details of these anomalies are included in Appendix C, and a summary is listed in Table 18.

Appendix C also lists the various operational issues encountered by systems during FpVTE (in Section 21: Test Processing Issues).

System ID	System Name	Software	Hardware	Exception	Explanation
123 ID (MST) (Error)	Biometric System Search (BSS)	BSS	4 Search nodes + administrator	Error	Participant stated that results were in error (overflows); test was successfully rerun using modified software and hardware
Dermalog (LST) (Error)	DermalogFinger Code3	DermalogFingerCode3 kernel Windows XP Professional	6 Pentium4s	Error	Some similarity matrices were not successfully saved; those subtests were successfully rerun
Dermalog (MST) (Error)	DermalogFinger Code3	DermalogFingerCode3 kernel Windows XP Professional	6 Pentium4s	Error	Resulting similarity matrices were blank (determined during analysis)
Neurotechnologi a M2 (MST) (Error)	VeriFinger	VeriFinger 4.2 software test version`	1 Pentium4	Error	Much of similarity matrix appeared to be random (determined during analysis). This run was made after th initial run could not decompress some images
Avalon (LST) (Halted)	Ultramatch AFIS	Ultramatch, Windows XP Professional	4 Celerons + 2 Pentium4s	Halted	Test halted by participant when it became clear it would not be completed in time
SAGEM M2 (MST) (Halted)	SAGEM MetaMorpho	MetaMorpho Software version 3.1.2A.NST.2A, Windows 2000 Professional	3 Pentium4	Halted	Test halted by Participant when it became clear that the distribution of processing was not working correctly; test rerun as SAGEM M2 (MST) with single Pentium4 configuration
BIO-Key (LST) (Late Dropout)	N/A	N/A	N/A	Late Dropout	"Our hardware vendor failed to supply committed hardware in time for testing"
BIO-Key (MST) (Late Dropout)	N/A	N/A	N/A	Late Dropout	"Our hardware vendor failed to supply committed hardware in time for testing"
123 ID (LST) (Overtime)	Biometric System Search (BSS)	BSS + Virtual Print Signature Technology	36 Search nodes + administrator	Overtime	Test ran over the allotted time
Technoimagia (MST) (Redundant)	FP-Workstation	Fingerprint Aunthentication System Tool21 ("FAST21"), Windows XP	1 dual Xeon	Redundant	Participant ran two identical systems to ensure that at least one finished successfully; systems both finished successfully and generated identical results

Table 18. System Anomalies

[1] All of these cases were able to retake the test within the original allotted time.

Section 5: Results

This section investigates the effects of several independent variables – characteristics of the images provided and demographics of the subjects from which the images were taken – on matcher accuracy. The following independent variables were investigated:

- Source – which government agency provided the fingerprints
- Quality – a weighted combination of image quality metrics from multiple Participants
- Number of fingers – 1, 2, 4, 8 or 10 fingers provided for the search
- Type of fingerprints -- Flat, Slap and Rolled; Livescan or Paper
- Finger combinations – which individual finger or combination of fingers was compared
- Age – subject age when the search image was collected
- Sex – male or female, when provided

Effects of these variables on accuracy were studied across systems, for general trends. The degree to which individual systems conform to these trends varies, as can be seen in the charts (some exceptions are noted in the text). The general trends are important in many ways, such as interpreting claims of accuracy based on different datasets, setting expectations for future scenarios before all operational and design factors are known, designing meaningful benchmark tests, and identifying areas for research.

Of the variables listed above, Source, Quality, Number of Fingers, Finger Combinations, and Age were found to have a significant effect on accuracy. The others were often difficult to measure, rather than found to be inconsequential. Some of these variables have a significant effect on the rank order of the systems, but this occurs primarily among the lower-ranked systems.

To say that a variable has a significant effect does not necessarily imply that a large change in operational accuracy will be observed. A very small effect is statistically significant when the test is sufficiently large to confidently observe the effect. The more accurate the system, the more difficult it is to observe an effect. On the other hand, in large operational systems, a small effect may be very beneficial or costly.

Statistically, it is important to note that these variables often are not independent and may in fact be surrogates for other phenomena. For example, in the next section it can be seen that Fingerprint Source is closely tied to Fingerprint Quality. It should also be noted that identifying a significant variable does ***not*** imply an understanding of causality.

5.1 Fingerprint Quality

FpVTE Participants were given the option to provide fingerprint quality information for the images in the FpVTE datasets. As part of FpVTE analysis, the Participants' quality metrics were aggregated and used to measure the effect of fingerprint quality on system accuracy.

5.1.1 Background

FpVTE datasets included a range of fingerprint qualities, including some poor quality fingerprints. This is part of the nature of operational data: fingerprints collected in real-world operational scenarios will not be as uniform in quality as fingerprints taken in controlled settings. This can be seen in the performance of the Ohio data, which is the only non-operational data included in the tests.

Some systems are designed to reject some fingerprints due to poor image quality. The rate at which this occurs is generally known as the Failure to Enroll (FTE) rate. The systems evaluated in FpVTE were required to generate similarity scores for all fingerprints: fingerprints could not be ignored due to poor quality. FpVTE provided an optional method for Participants to indicate which fingerprints would have been rejected as FTE in an operational system.

Seven FpVTE Participants chose to provide image quality information for the MST fingerprints. Each also provided an FTE threshold below which images would be considered unsatisfactory for enrollment. Since SST is a subset of MST, these quality measures could be used for both MST and SST.

An insufficient number of Participants provided image quality information for LST for analysis to be productive. This was not a particular surprise, since effectively aggregating multiple fingerprint quality measurements to define a single quality measure for a set of fingerprints is problematic.

Previous work in fingerprint image quality [IQS, NIST_IQS] has shown that the relationship between image quality and matcher accuracy is highly correlated for poor-quality fingerprints, but is less definitive for higher-quality fingerprints. In addition, it is clear that a wide variety of fingerprint quality measures exists, and a single fingerprint quality metric may correlate closely with the performance of some matchers, but not all. A well-designed proprietary image quality metric will be tuned to that vendor's matching algorithm. It is important to select a quality metric that is as broad-based as possible to avoid biasing the results.

For these reasons, it was decided that an effective method of analysis of fingerprint quality would be to treat the seven systems' FTE information for each MST image as votes: this measure of image quality for a fingerprint is based on the percent of the MST systems that labeled the fingerprint as an FTE. Note also that this method is broad-based, and is not dependent on one single image quality metric. There is a subtle distinction between this and most quality metrics: this does not even attempt to differentiate between good and mediocre-quality fingerprints, but focuses on levels of poor quality.

The FTE-based quality method used here is based on the number of FTE votes the image received from the seven systems:

- A: None of the systems labeled the image as FTE
- B: 1-25% of the systems (1/7) labeled the image as FTE
- C: 26-50% of the systems (2/7 or 3/7) labeled the image as FTE
- D: 51-75% of the systems (4/7 or 5/7) labeled the image as FTE

- F: All of the systems labeled the image as FTE ***or*** a human fingerprint examiner during groundtruthing labeled the image as particularly poor

5.1.2 Distribution of Fingerprint Quality

Fingerprint quality is clearly related to the source of the fingerprints, as shown in Figure 11 and Table 19. Note that the distribution of good-quality (A) fingerprints varies dramatically.

DHS Recidivist fingerprints (DHS2) are clearly poorer quality than any of the other sources; they were collected under very difficult operational circumstances, which clearly had an impact on quality. It is interesting to note that quality as measured here is not tied to scanner type: DHS2 and DOS-BCC were collected using the same scanner model (DFR-90) and software, but the quality of DOS-BCC is far superior.

The Ohio fingerprints were the only fingerprints that were collected in controlled (non-operational) settings. It should be no surprise that the quality is the best of all of the sources.

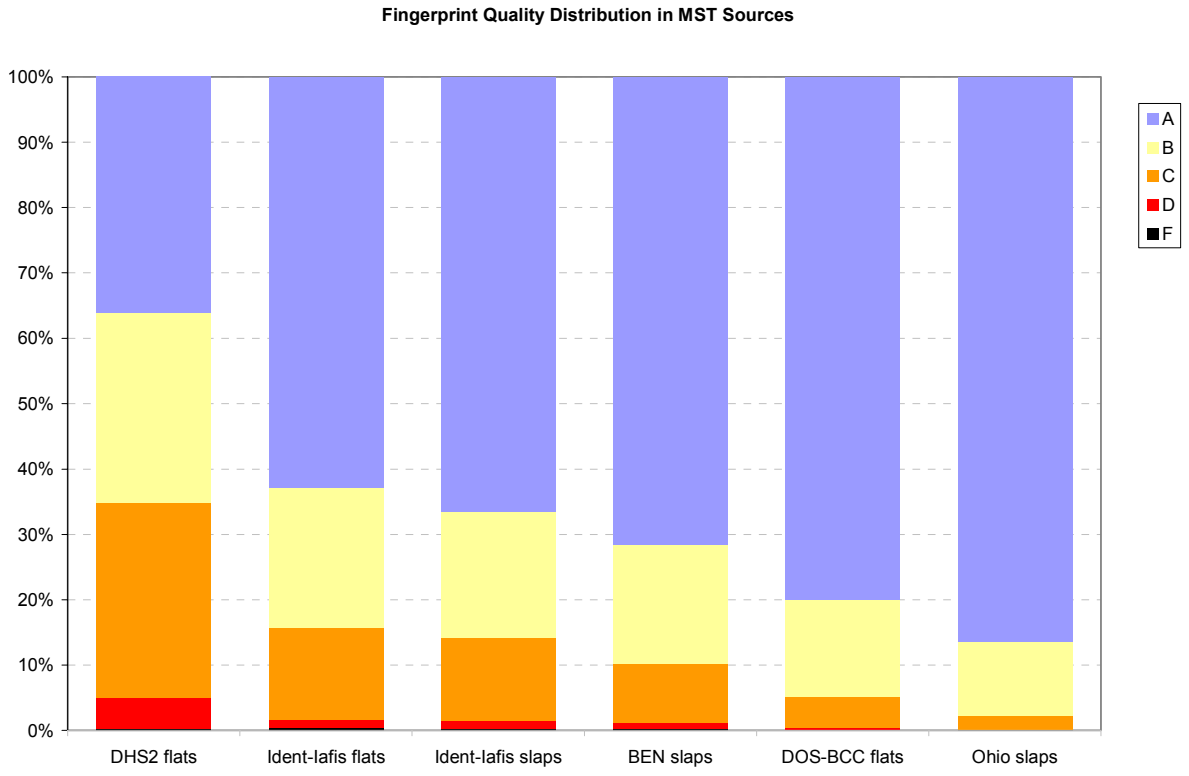

Figure 11. Fingerprint Quality Distribution by Source in MST

Table 19 shows that the quality distribution of the SST sources is not substantially different from the corresponding MST flat sources. Fingerprint quality was not measured for LST, so results were not available for the 12k and DHS10 sources.

Quality Distribution in MST							
	Flats			Slaps			MST
	DHS2	DOS-BCC	Identlafis	Identlafis	BEN	Ohio	Total
A	36.1%	80.0%	62.9%	66.4%	71.6%	86.4%	70.3%
B	29.0%	14.8%	21.4%	19.4%	18.2%	11.4%	17.9%
C	30.0%	4.8%	14.1%	12.8%	8.9%	2.1%	10.5%
D	4.7%	0.2%	1.2%	1.3%	1.0%	0.1%	1.2%
F	0.2%	0.2%	0.4%	0.2%	0.3%	0.1%	0.2%

Quality Distribution in SST			
	Flats		SST
	DHS2	DOS-BCC	Total
A	36%	83%	69%
B	28%	13%	17%
C	31%	4%	12%
D	6%	<1%	2%
F	<1%	<1%	<1%

Table 19. Quality Distribution in MST and SST

5.1.3 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 12: 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 12 also shows that some systems are extremely sensitive to image quality.

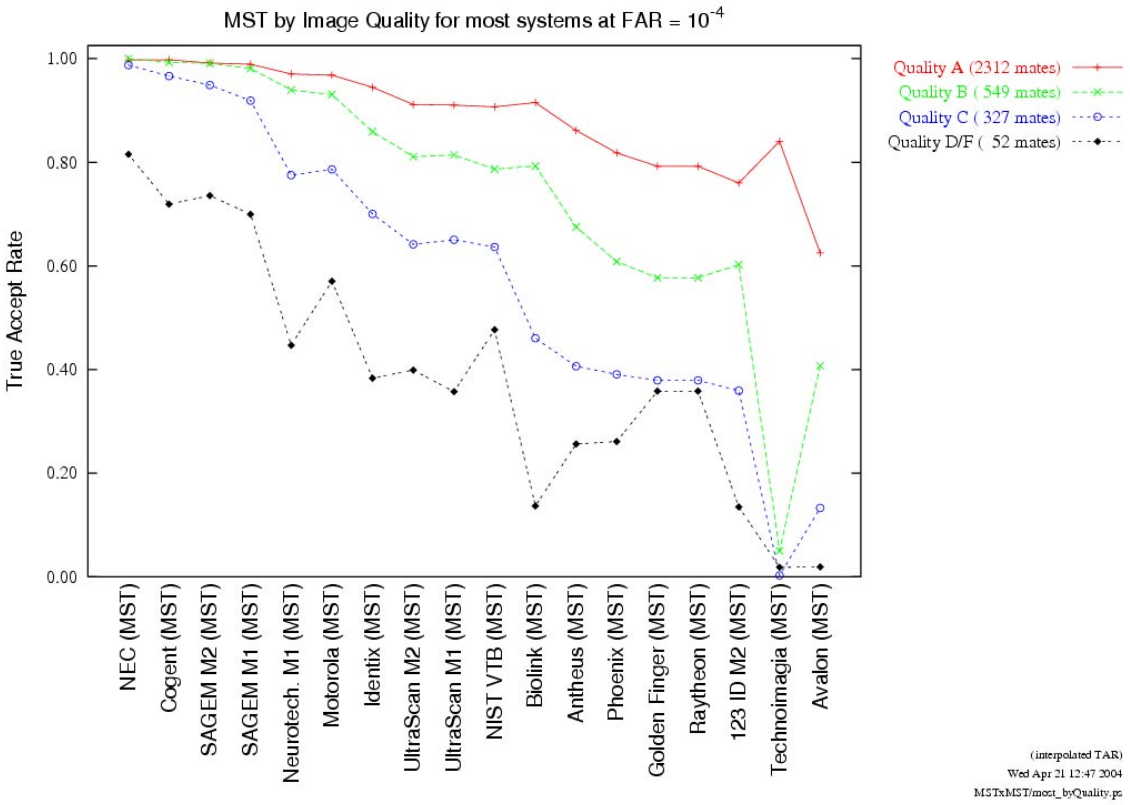

Figure 12. Effect of Image Quality (MST)

5.2 Effect of Fingerprint Source

As described in Section 3.2.2, the fingerprints for FpVTE were sampled from eight distinct sources. Each source is characterized by subject population demographics, operational procedures and technology, and other factors. The use of real operational data in this evaluation provides meaningful performance measures, but often makes it difficult to attribute differences in accuracy to specific factors such as image type or capture technology.

Source is an important factor in predicting operational accuracy: some sources are more difficult to match than others. Accuracy on the non-operational Ohio dataset is significantly higher than accuracy on the operational datasets. There is some agreement among the systems as to the relative difficulty of the operational sources, but the results are not definitive.

Figure 13 shows that single-finger results for Ohio are significantly more accurate than those for BCC, Benefits, DHS2, or IDENT-IAFIS Slaps for the majority of MST systems. It is notable that the relative difficulty of the operational partitions varies by system.

For the most accurate systems, the Benefits and IDENT-IAFIS Flats partitions are too small to allow meaningful differentiation.

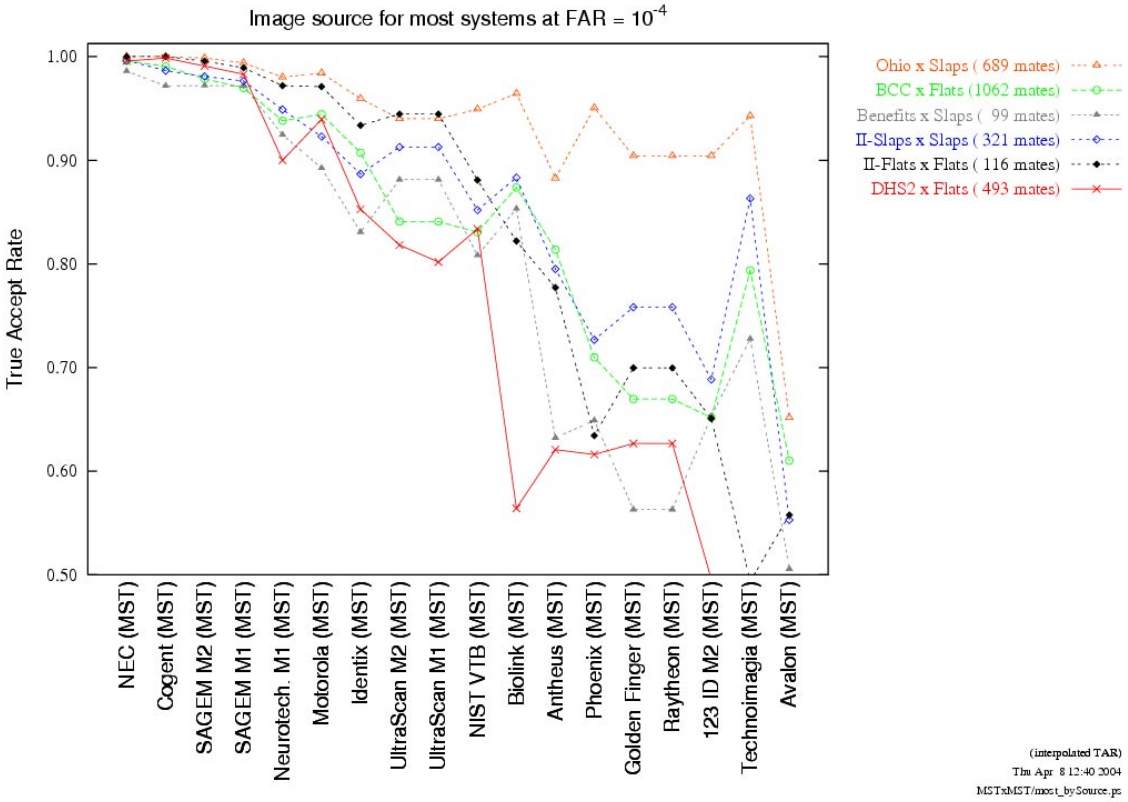

Figure 13. Effect of data source (MST)

The results from LST are consistent with the MST results, as shown in Figure 14. Since this test (LST BxA) only includes single-finger flat:flat comparisons, the differences between sources are clearer.

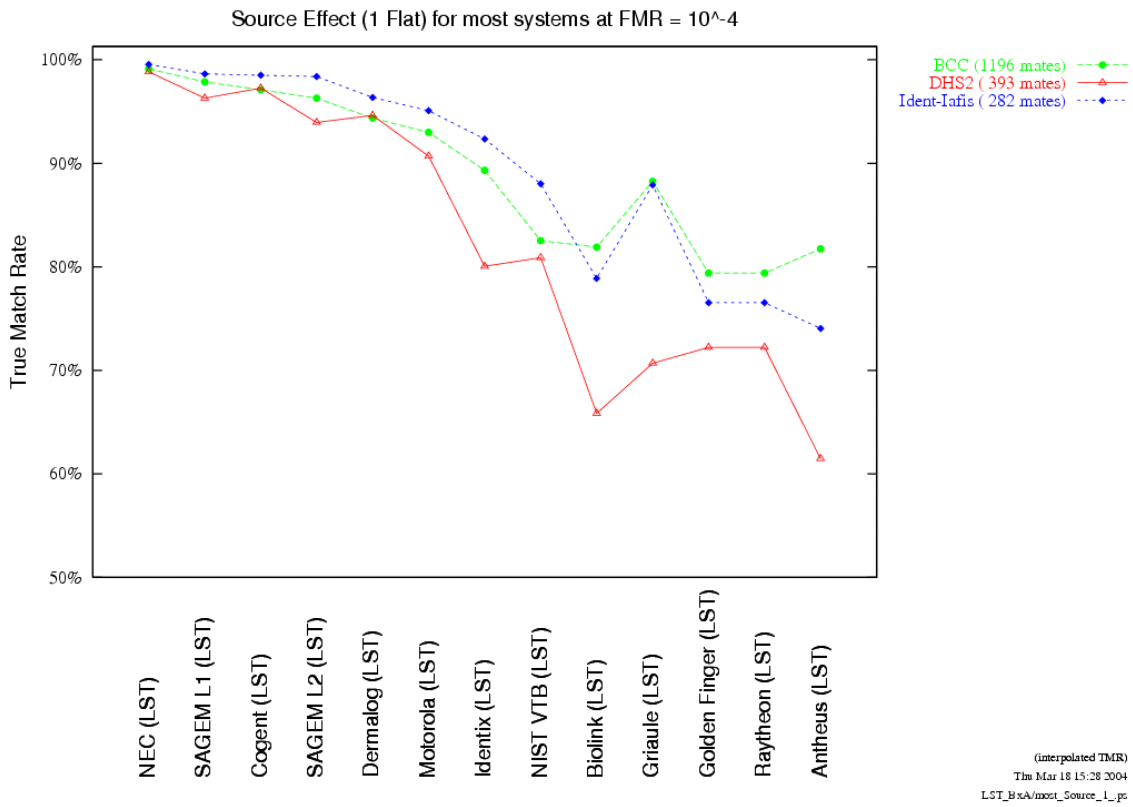

Figure 14. Effect of data source (LST)

For multi-finger tests, the effect of source becomes less clear. Many of the more accurate systems achieve perfect or near-perfect TARs in many multi-finger subtests, so the effect of data source is not measurable.

5.3 Effect of Number of Fingers

System accuracy was very sensitive to the number of fingers compared. 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. This can be seen clearly in Figure 15.

This figure also shows that there is great variability within the data for a given number of fingers. Much of this variability can be explained by variations in data source, quality, and type. The effect of these is difficult to isolate.

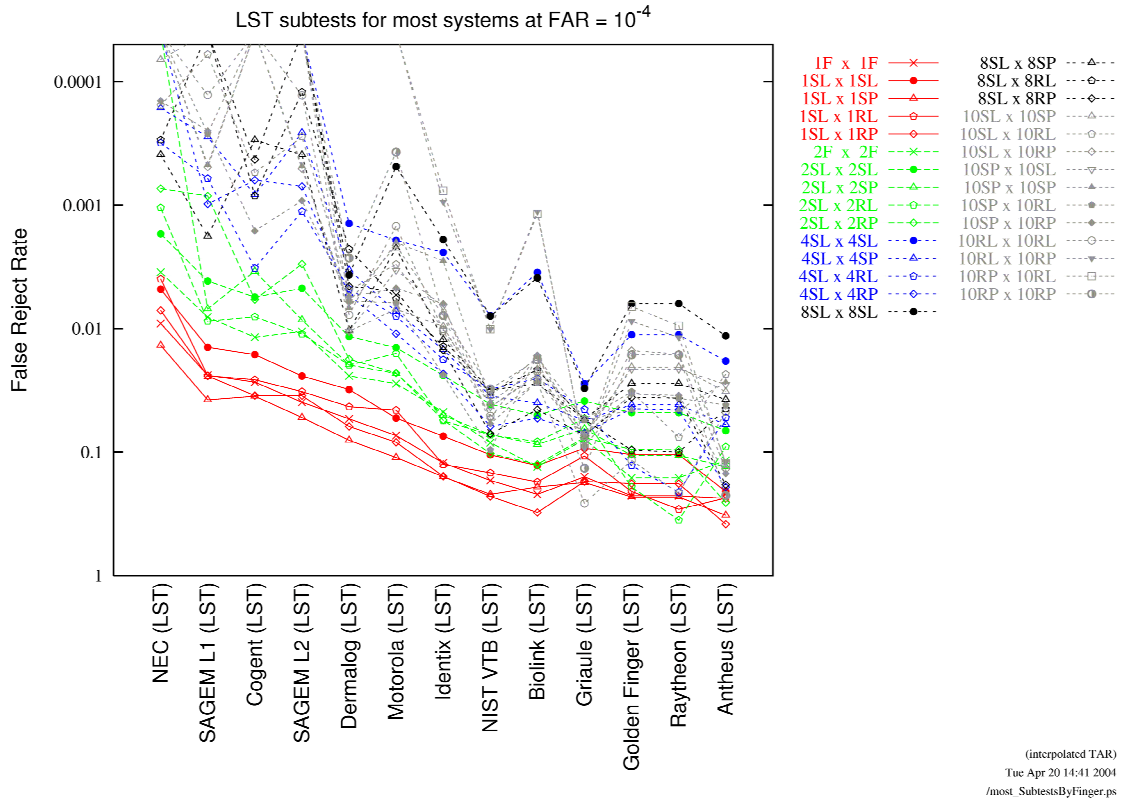

Figure 15. Effect of Fingerprint Number and Other Variables in LST. The Y scale is the log of False Reject Rate, which is 1 – TAR. Note that the single-finger searches (red) are clearly separated from the two-finger searches (green), but the four, eight, and ten-finger searches are intermingled. At the test sizes used, accuracy of four, eight, and ten-finger searches is difficult to differentiate. The lines off of the top of the chart are for FRR=0 (perfect results), which cannot be represented in log scale.

In order to minimize the effects of confounding variables, data source and image type were controlled for this analysis. Given the factored test design and the availability of operational data from the various sources, there was sufficient data for 10 separate 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. However, the degree of improvement is not consistent, and is apparently strongly affected by other variables.

The following charts show examples of the effect of number of fingers, where data source and type of fingerprint are held constant. Figure 16 shows results for FBI 12k, slap livescan vs. rolled livescan, 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 1.0 on one finger.

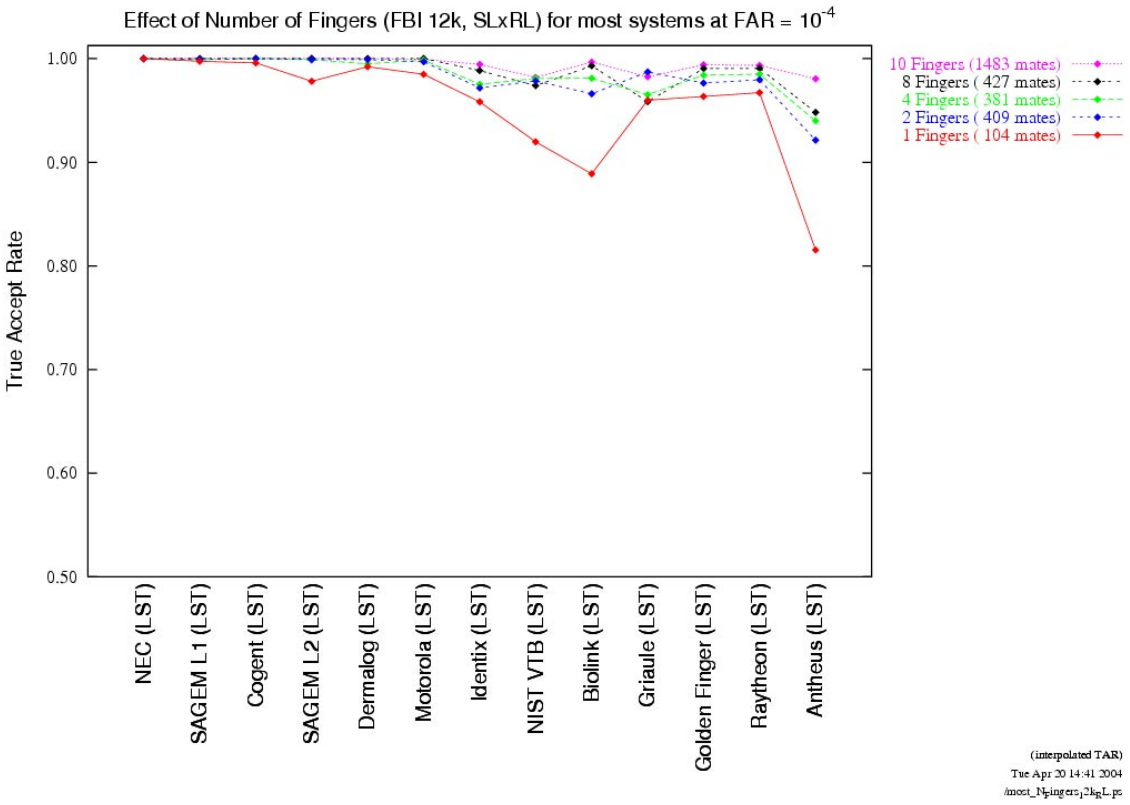

Figure 16. Effect of number of fingers on FBI 12k (slap livescan vs. rolled livescan): effect is not measurable when 1-finger TAR approaches 1.0

Figure 17 shows typical results for a dataset that is more difficult to match: IDENT-IAFIS (slap livescan vs. rolled livescan). The 1-finger and 2-finger results are clearly separated, but the distinctions between 4, 8, and 10 fingers can be observed clearly only in the less accurate systems.

Note that Griaule's accuracy does not improve greatly with more than 2 fingers. Golden Finger and Raytheon had unusual difficulties on the IDENT-IAFIS SL vs. RL tests.

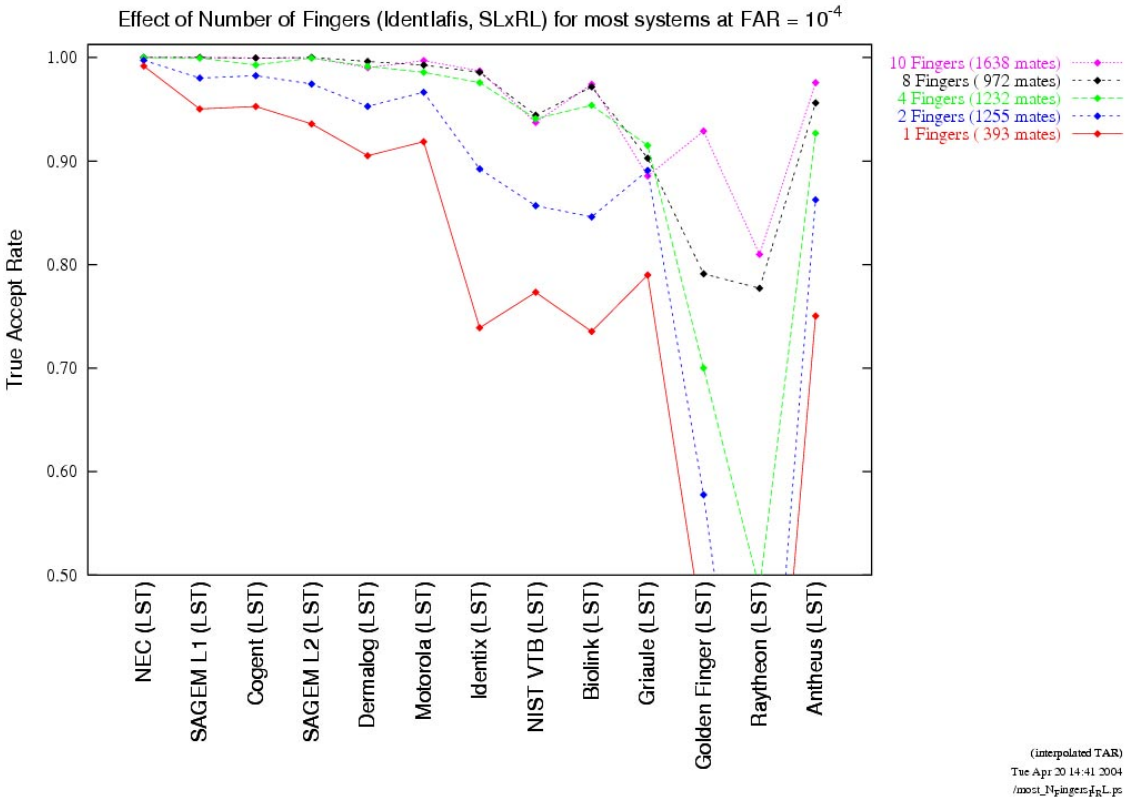

Figure 17. Effect of number of fingers on IDENT-IAFIS (slap livescan vs. rolled livescan): this data clearly shows the benefit to comparing more than 2 fingers

Even for some of the more accurate systems, a difference in performance can be seen between 2 and 4-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. Cogent, however, shows a distinct separation between 4 and 8-finger performance, as shown in detail in the ROCs in Figure 18.

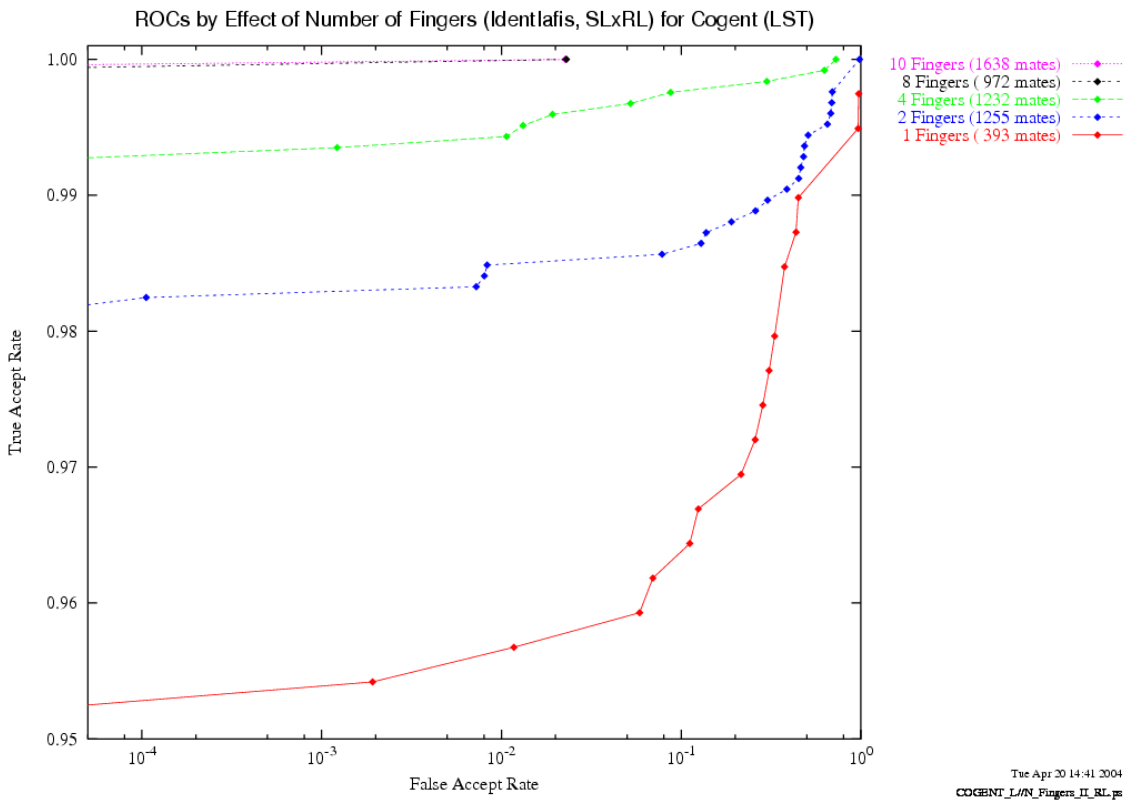

Figure 18. Effect of number of fingers on Cogent LST: the 4-finger line is clearly separate from the 8 and 10-finger lines. This is a detail of the same data shown in Figure 17.

Several observations can be made from these analyses:

- The ability of the most accurate systems to discriminate mates from non-mates was perfect or nearly perfect on all of the four, eight and ten-finger tests.
- All systems achieve greater accuracy when multiple fingers are provided for comparison than when only one finger is provided. The difference is large and consistent.
- As the number of fingers increases, the expected improvements are generally observed, but not always. There are several likely explanations for this:
  - The sample size is too small to measure small variations in TAR, especially for the top systems where TAR is at or near 1.0.
  - The results include a variety of uncontrolled variables, which are problematic when using operational data.
  - Some systems may have had fusion algorithms or performance tuning that limited the number of fingers actually compared.

5.4 Comparison of Verification and Identification Results

As discussed in Section 2, FRVT 2002 defined performance statistics for verification and identification tasks [FRVT2002]. This section examines whether the findings presented in this report are sensitive to the choice of performance statistic.

Some biometric models assume that the false accept rate grows linearly with gallery size when true accept rate is kept constant. This assumption was tested by comparing the results of verification and open-set identification ROCs.

The FAR used in an open-set identification ROC does not correspond directly to a 1:1 FAR, but is measured against a specific gallery size. The assumption is that for a gallery of size N, $FAR_{1:N} \approx 1 - (1-FAR_{1:1})^N$. Therefore, FARs can be rescaled in this way:

$$\frac{FAR_{1:N}}{N} \approx FAR_{1:1}$$

Figure 19 compares standard verification (1:1) ROCs to open-set rank-1 identification at rank-1 (1:N) ROCs rescaled in this way[1]. The 1:N FAR scale is shown on the top axis, and the 1:1 FAR scale is shown on the bottom axis. Both sets of ROCs were computed on the standard partition of MST, which has a Probe Set size of 5800 and a Gallery Set size of 3240. The lines are shown superimposed such that 1:N FAR = 1 (top axis) is aligned with 1:1 FAR = 1/3240 = $3.1*10^{-4}$ (bottom axis).

[1] On the verification (1:1) ROC, FAR is computed from all non-mate comparisons; TAR is computed from all mate comparisons. On the open-set identification (1:N) ROC, FAR is computed from probes that have no mate in the gallery ("imposters"); TAR is computed from probes that have a mate in the gallery

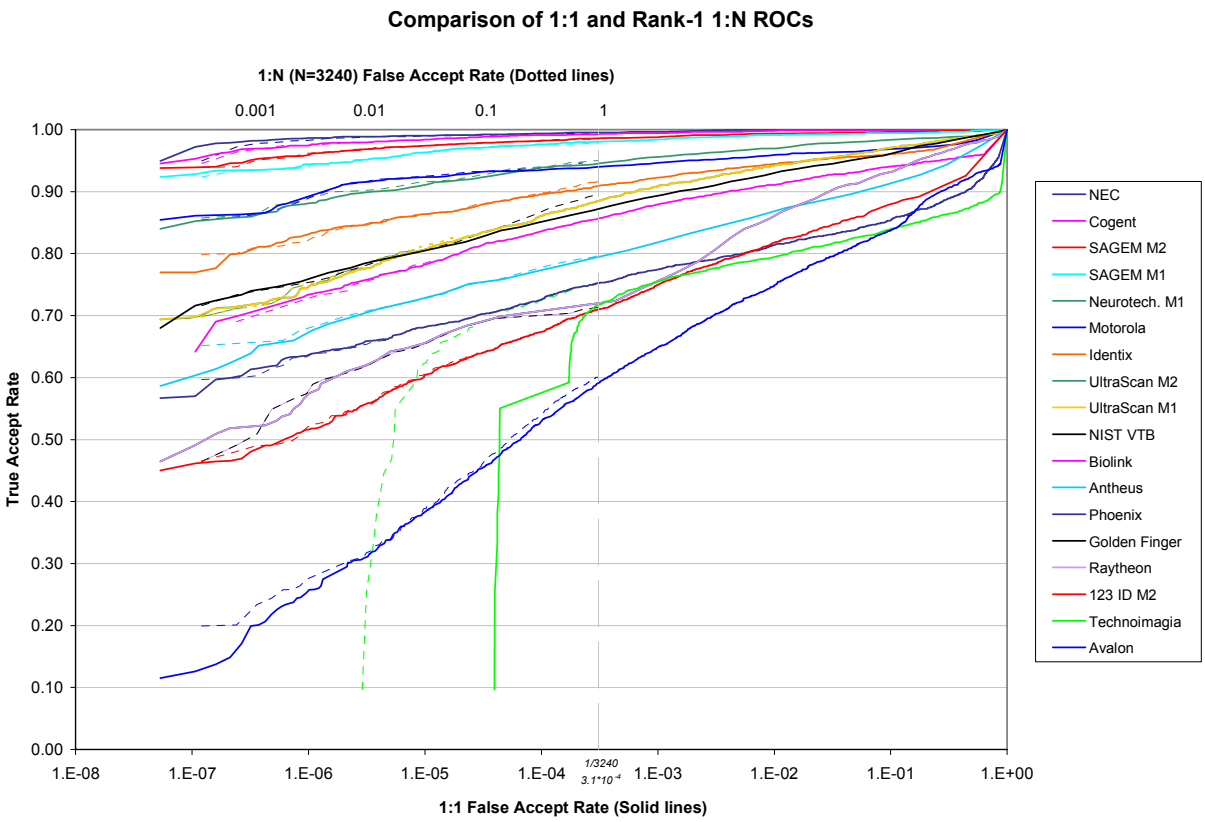

Figure 19. Comparison of 1:1 and rank-1 evaluation methods. The 1:N ROCs were rescaled by the gallery size (3240). The 1:1 ROCs are solid lines; the 1:N ROCs are dotted lines.

In most cases, the solid (1:1) and dotted (1:N) lines are closely aligned, with the obvious exception of Technoimagia. This alignment can be better seen in Figure 20, which shows detail from the same graph. At this scale, it can be seen that the difference in TAR as measured by the two methods rarely exceeds .5%, but that the rightmost points on the NIST VTB ROCs do differ by about 2.0%. Note also that neither type of ROC is consistently above the other.

I THINK THAT THE CAPTION ABOVE FIGURE 20 SHOULD READ "N=3240" NOT 3520

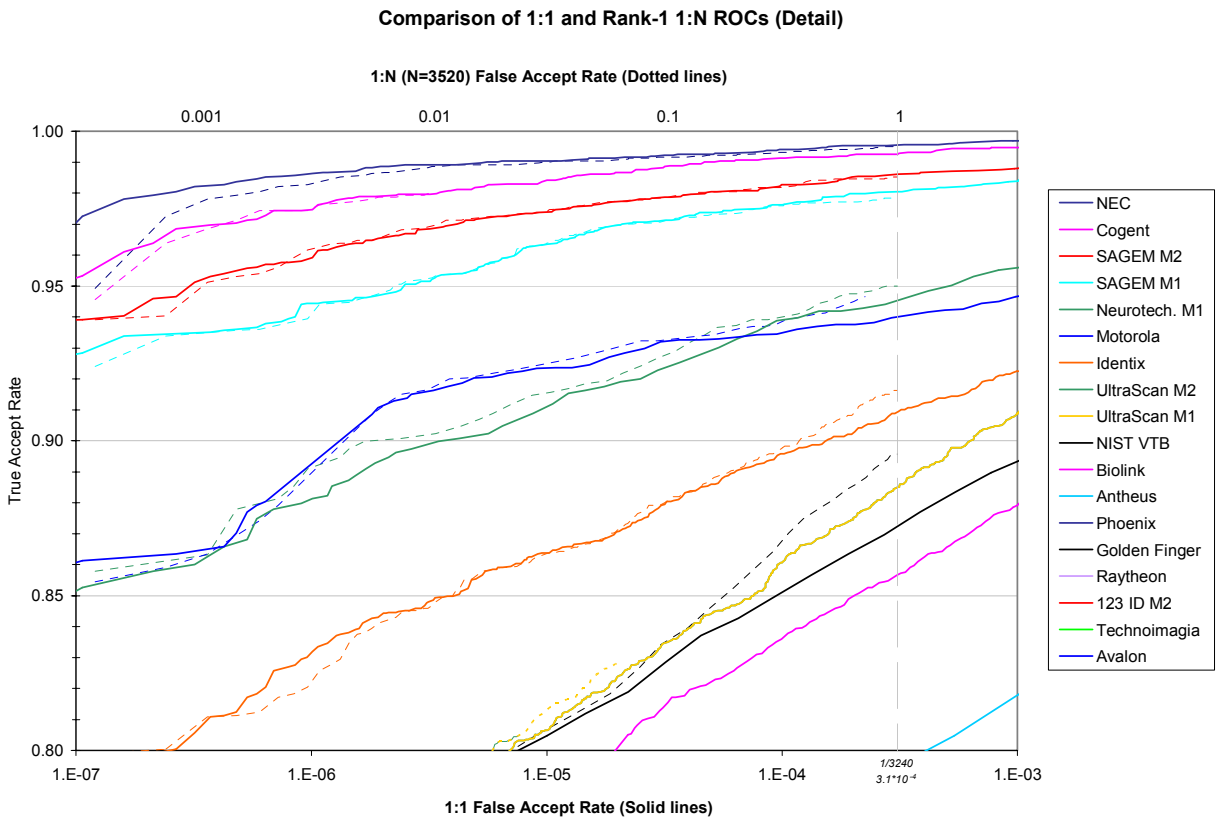

Figure 20. Comparison of 1:1 and rank-1 1:N evaluation methods (Detail). 1:1 and rank-1 1:N ROCs are usually closely aligned.

The overlap of the shifted rank-1 open-set identification (1:1) ROCs and the verification 1:1 ROCs in Figure 19 and Figure 20 is consistent with the observation that the false accept rate grows linearly with gallery size, and the true accept rate remains constant. This behavior is expected if the match score depends only on the two matching fingerprints. Note, however, that because our definition of *true accept rate* requires that the match be at rank one implies a dependence on the other gallery elements too. That the curves nevertheless overlap occurs because the match is often at rank one (as will be shown below). The rank criterion is therefore secondary to the usual threshold criterion.

The similarity between 1:1 and 1:N ROCs does *not* necessarily mean that this model would be valid for other biometrics. It should be noted that the similarity between 1:1 and rank-1 1:N ROCs should be expected to decrease for very low FAR values, where the rank-1 requirement is likely to decrease the TAR for the 1:N matches.

The above comparisons are based on matches made at rank 1. Figure 21 repeats data from Figure 20 for three systems, and shown in Figure 20, but adds 1:N ROCs disregarding rank. Since an ROC by definition must go to (1, 1), the rankless 1:N lines necessarily rise above the rank 1:N lines as FAR increases.

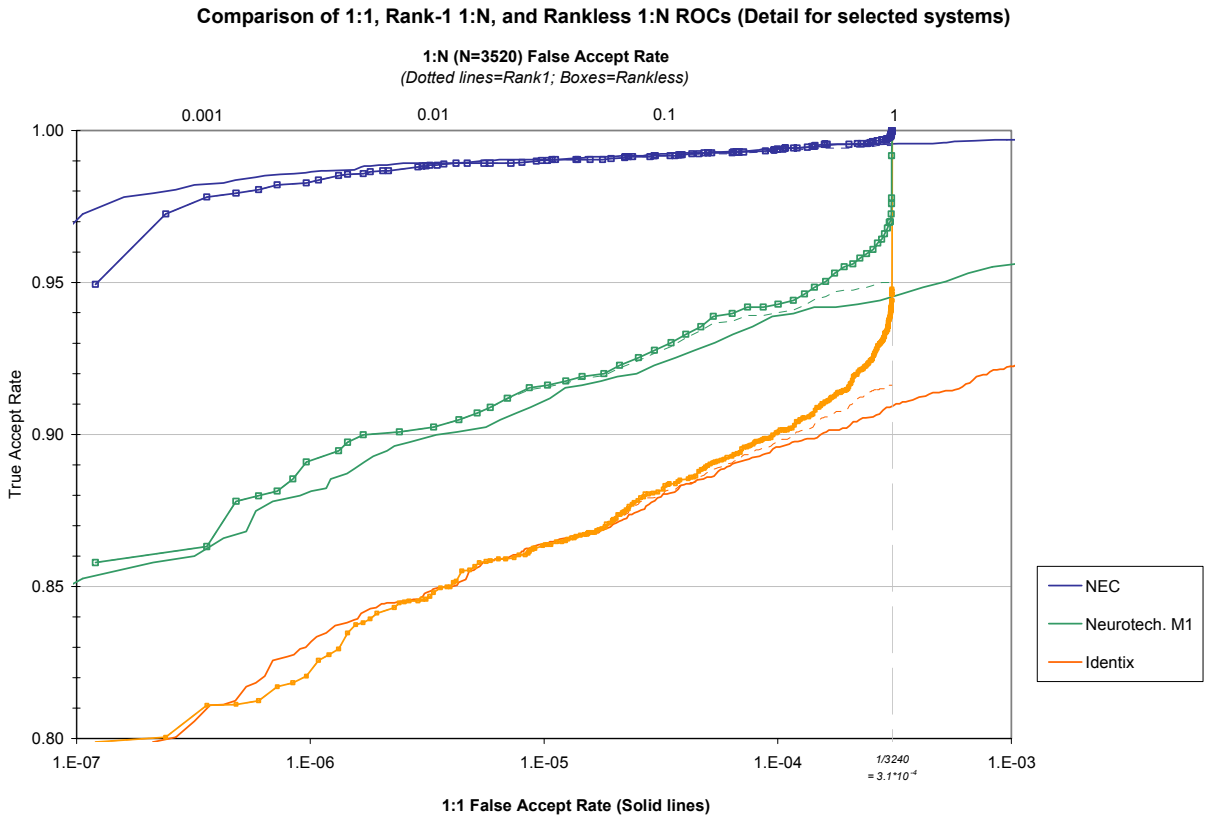

Figure 21. Comparison of 1:1, rank-1 1:N, and rankless 1:N evaluation methods (Detail). Solid lines are 1:1, dotted lines are 1:N (Rank-1), and lines with boxes are 1:N (any rank).

In accounting for the small differences resulting from these methods of evaluation, it should be noted that in FpVTE 2003, systems were not constrained to perform true 1:1 matches[1]: systems were provided with Probe and Gallery sets, and were allowed to search each probe against the entire gallery in order to produce the similarity scores for that probe. Test guidelines were deliberately flexible on this point in order to accommodate various existing system designs and scoring strategies:

- Some of the participating systems were designed as multi-stage fingerprint matchers (in which multiple matchers and indexing algorithms are fused in series or parallel). A possible consequence of using filtering or prescreening stages for processing efficiency is that the algorithm used to compute each similarity score may depend on other images in the gallery.
- Systems in FpVTE (and FRVT) were permitted to normalize their results, i.e., apply a mathematical transformation to the set of scores resulting from one search. Normalized scores are not based strictly on 1:1 comparisons.

[1] About half of the systems in MST were shown to have performed true 1:1 matches. See Appendix D for details.

Identification performance can be measured in terms of rank alone.[1] The results are graphed on a CMC. Figure 22 shows CMCs for all systems for the standard MST partition.

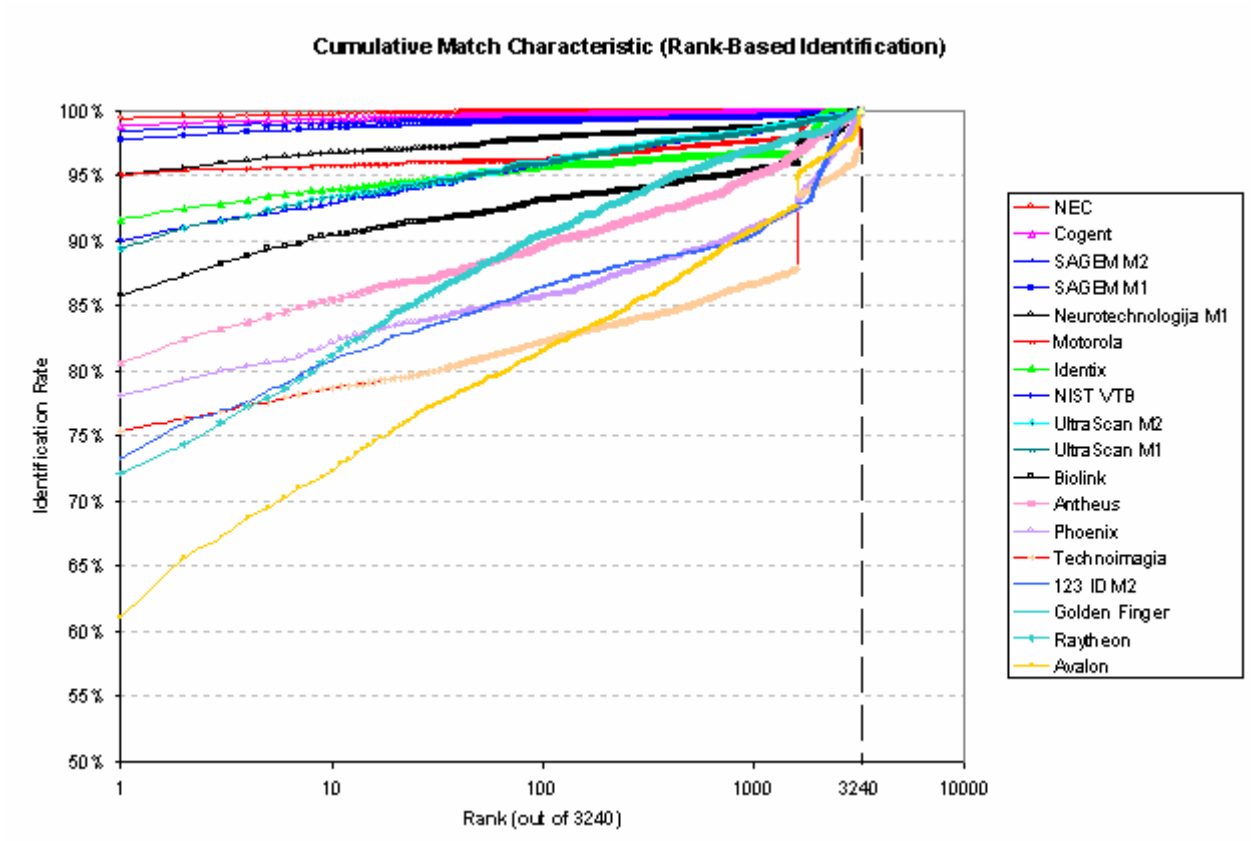

Figure 22. Rank-based identification performance for MST Systems

Table 20 shows the corresponding rank 1 identification rates, which is shown in Figure 22 as the point at which the system's CMC meets the left y-axis. It also compares them to the corresponding points on the ROC. Note how the rank 1 CMC identification rate is similar (but not identical) to the ROC at 1/(gallery size) (1/3240, or $3.1 * 10^{-4}$).

[1] This is sometimes described as closed-set identification.

MST System	Identification Rate at Rank 1	TAR where FAR=3.1*10^{-4} (1/3240)
NEC	99.4%	99.6%
Cogent	98.9%	99.3%
SAGEM M2	98.4%	98.6%
SAGEM M1	97.8%	98.1%
Neurotech. M1	95.1%	94.5%
Motorola	95.1%	93.9%
Identix	91.6%	90.9%
NIST VTB	90.0%	87.2%
UltraScan M1	89.4%	88.5%
UltraScan M2	89.4%	88.8%
Biolink	85.8%	85.6%
Antheus	80.6%	79.4%
Phoenix	78.1%	75.2%
Technoimagia	75.4%	71.7%
123 ID M2	73.3%	71.0%
Golden Finger	72.1%	72.0%
Raytheon	72.1%	72.0%
Avalon	61.0%	59.1%

Table 20. Comparison of rank-based identification performance and verification performance for MST systems. Rank-1 identification rate is close to verification TAR at FAR = 1/(gallery size).

5.5 Other Results

This section presents findings on the effects on accuracy of fingerprint type, finger position and combinations, sex, and subject age. Also presented are limited findings on the correlation of system results and potential for fusion, and the effects of multiple mates.

Although not detailed in this report, the effects of scanner type and of Civil vs. Criminal records were also analyzed. Scanner type data was available only for the Ohio datasets, for three high-end slap and rolled livescan systems. The FBI 12k dataset was partitioned into Civil and Criminal records. Neither analysis found a significant correlation between the variable studied and matcher accuracy.

5.5.1 Effect of Fingerprint Type

FpVTE included flat, slap and rolled fingerprint images collected on livescan devices and on paper. These image types differ in terms of resolution, distortion, background, etc. This analysis investigated the effects of image type on accuracy, both within group (i.e., same type used for probe and gallery) and across groups (i.e., different types for probe and gallery).

Effect of Flat, Slap, and Rolled Fingerprints

Conventional wisdom would lead us to expect that

- Rolled fingerprints would be easier to match than slap or flat fingerprints due to increased size

- Slap fingerprints would match with higher accuracy than flat fingerprints due to image quality
- Accuracy drops when probe and gallery images are of different types

The FpVTE results show that such statements cannot be generally stated. Flat, slap, and rolled fingerprints have these effects on accuracy:

- The effect of fingerprint type varied dramatically from system to system. Few general statements about fingerprint type apply to all systems.
- Rolled fingerprints did not show increased accuracy at these operating points.
- Slap fingerprints do match with higher accuracy than flat fingerprints, but that effect can be attributed to the presence of the controlled (Ohio) slap data and the poor-quality DHS2 flat data. When those sources are disregarded, there is no clear difference between slap and flat fingerprints (given the same number of fingers).
- Accuracy often drops when probe and gallery images are of different types.

In operational data, image type data is generally confounded with other uncontrolled variables (capture technology correlates with operational procedures, demographics, etc). Although accuracy varies significantly by dataset, attributing this variance to specific image types may not be possible.

Figure 23 shows the relevant partitions of MST data. There is not a clear distinction between Flat vs. Flat and Slap vs. Slap performance, especially if the non-operational Ohio results are disregarded. The mixed image types (Flat vs. Slap) are generally less accurate, but even this trend is not consistent for all systems.

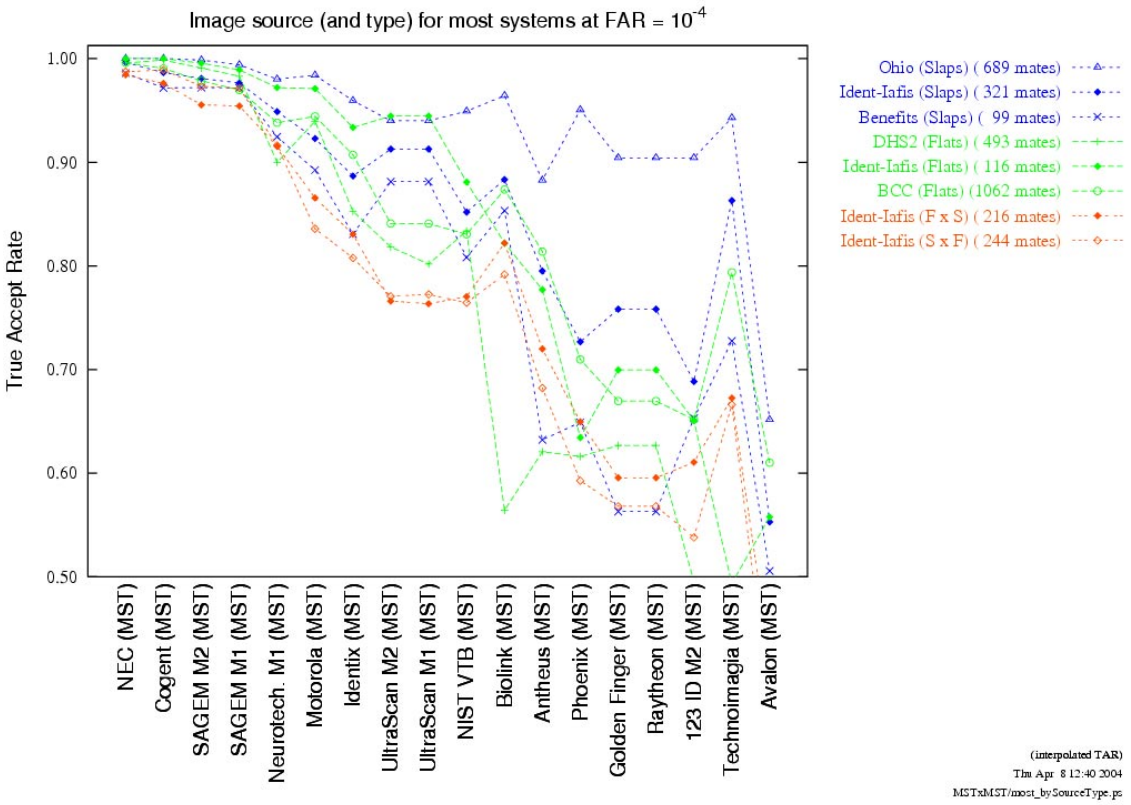

Figure 23. MST data by image type, controlling for data source. Note that flats vs. flats (flats searched against flats) are in green, slaps vs. slaps are in blue, and cross-type comparisons (flat vs. slap & slap vs. flat) are in orange.

Figure 24 provides another comparison of Flat, Slap, and Rolled data, this time from LST. Note that the flat vs. slap results do not differ in accuracy from the flat vs. flat or slap vs. slap results, so a general statement cannot be made that cross-type matches degrade accuracy. The lower accuracy for slap vs. rolled matches holds for all systems.

These results suggest that the effects of fingerprint type are small in comparison to other variables, but that some types of cross-type matches result in lower accuracy.

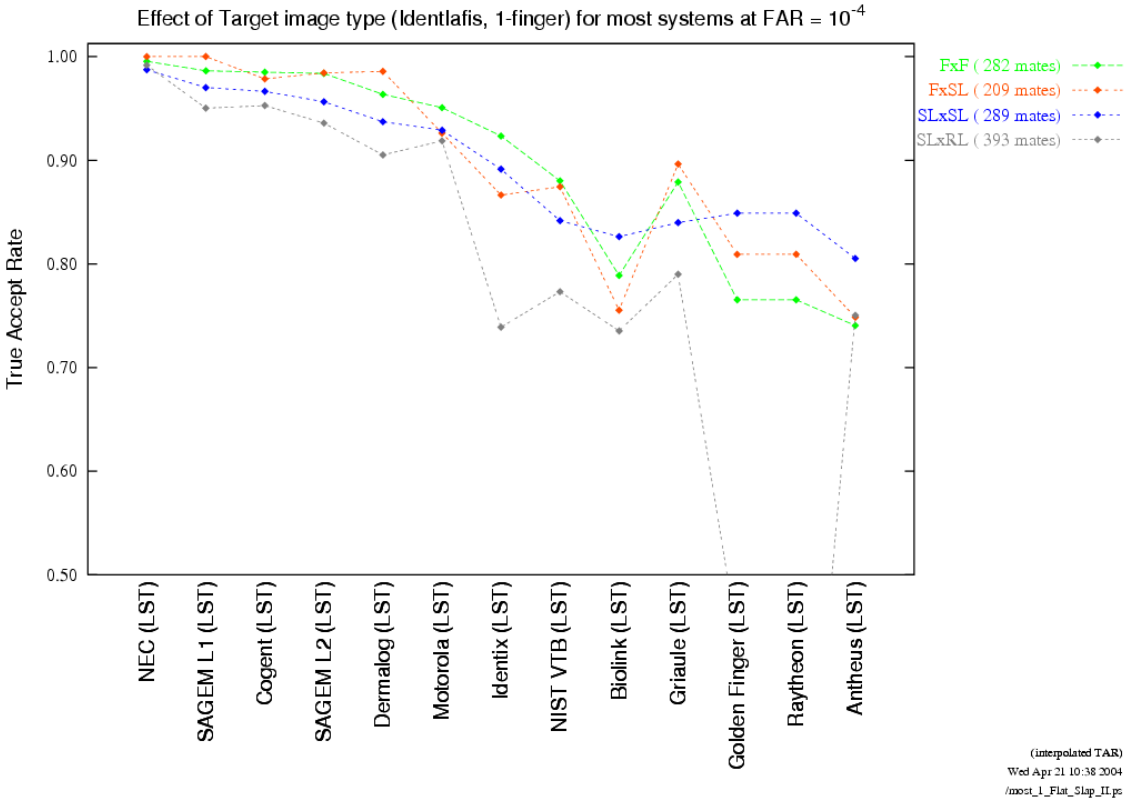

Figure 24. Flat vs. Slap performance on IDENT-IAFIS data (LST). Note that the results are mixed, except that the Slap Livescan vs. Rolled Livescan results are least accurate for every system.

Effect of Livescan vs. Paper Fingerprints

There was a clear effect on accuracy when comparing fingerprints that came from livescan and paper sources. Searches in which the probe and gallery were both livescan were more accurate than searches in which livescan was searched against paper. This was true for all systems.

Figure 25 suggests a very pronounced effect on accuracy favoring livescan vs. livescan comparisons over livescan vs. paper in the 12k data. Comparable SL vs. SL data was not available from this source.

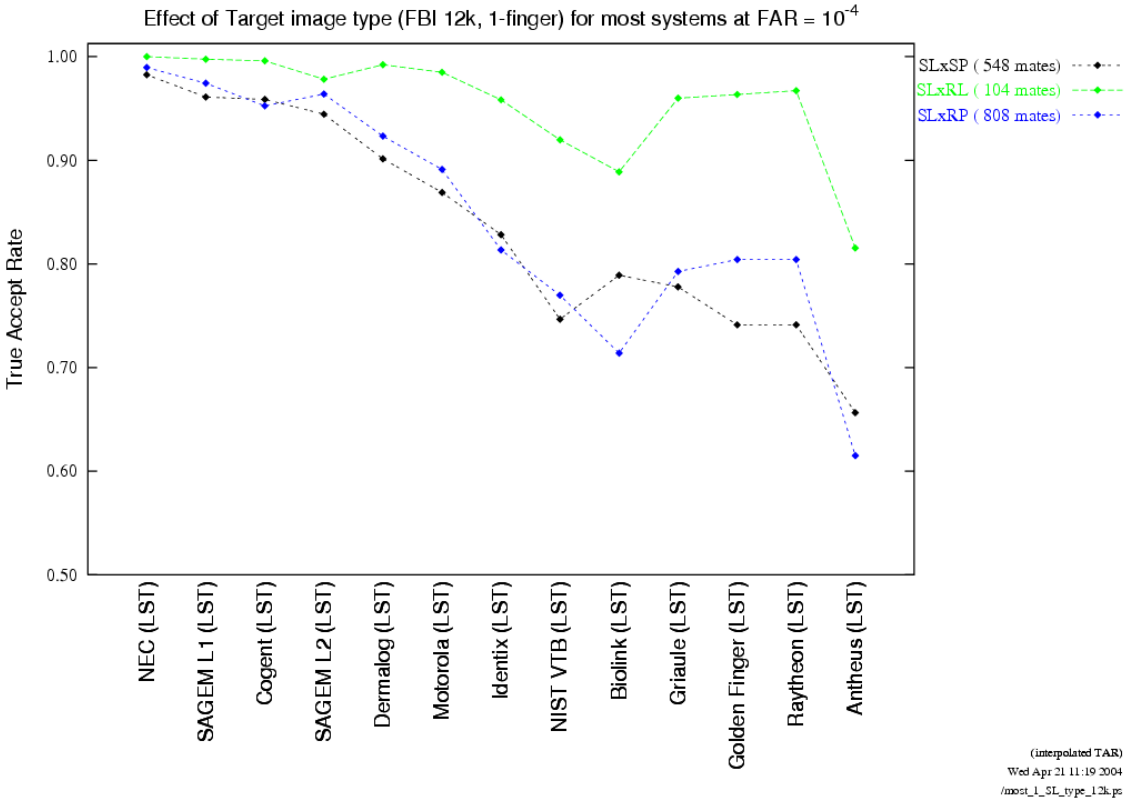

Figure 25. Combinations of Slap/Rolled and Live/Paper (FBI 12k data). Note the results in which livescan is searched against paper are substantially less accurate than the livescan-only results.

Figure 26 makes a similar comparison using the Ohio data. It is significant to note that the Ohio findings are both consistent with the FBI 12k findings and involve large sample sizes. These results show examples of what appears to be a general trend of lower accuracy on comparisons of livescan vs. paper as compared to livescan vs. livescan.

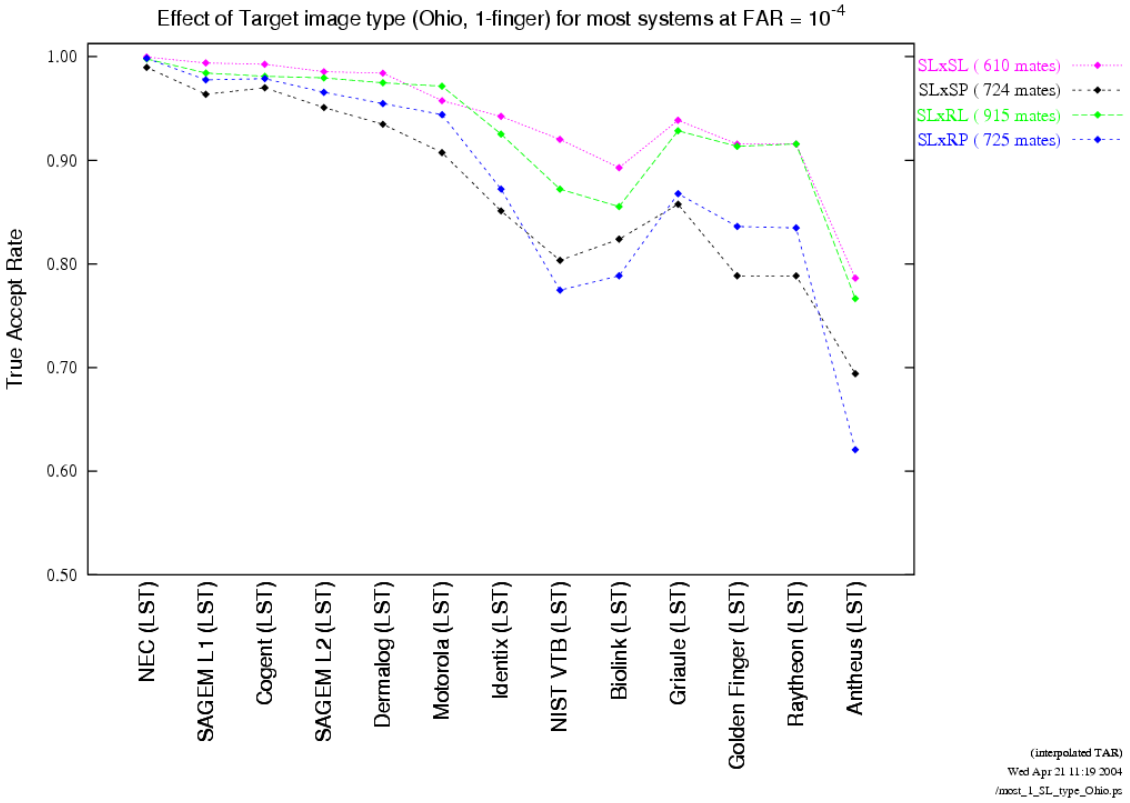

Figure 26. Combinations of Slap/Rolled and Live/Paper (Ohio data) Note the results in which livescan is searched against paper are less accurate than the livescan-only results.

5.5.2 Effect of Finger Position and Combinations

If identifications are to be based on fewer than 10 fingers, the question arises as to the relative effectiveness of different fingers or combinations of fingers. For single-finger comparisons, accuracy on segmented slap little fingers is significantly lower than on other fingers. No consistent pattern of preference among the other four fingers was detected. In general, the data source accounts for more variance than which finger is used, except for the little finger.

Eight single-finger LST partitions were analyzed. In these analyses, data from left and right hands were combined to increase sample sizes. The Ohio dataset included all fingers, whereas the IDENT-IAFIS and FBI 12k data included only index and thumb. Each partition has unique characteristics that might be attributable to variations in data source and image type as well as to limited sample sizes. For example, the two IDENT-IAFIS partitions showed better results on the index finger, whereas the two FBI 12k partitions showed better results on thumbs. Figure 27 shows a typical result, in which the little finger is clearly separated from the other fingers, but no other clear trend is discernable.

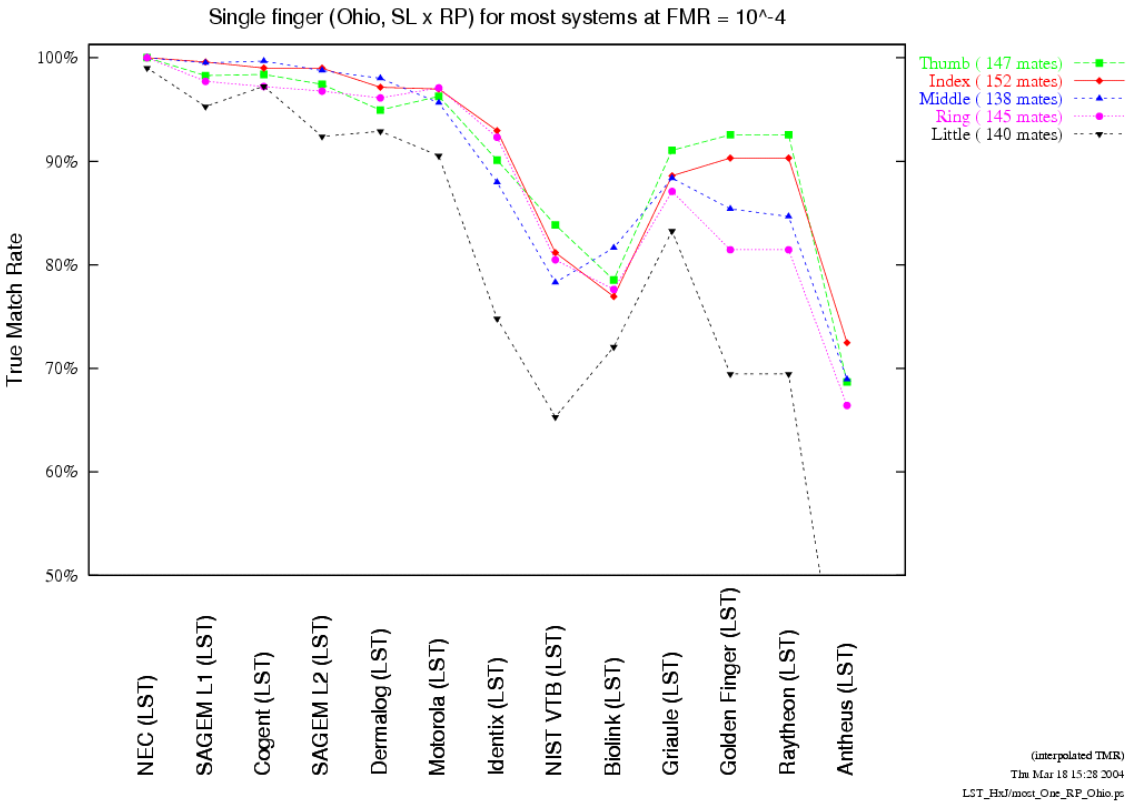

Figure 27. Segmented slap little fingers are more difficult to match than other fingers

Four 2-finger LST partitions (index and thumb only) were analyzed. Again the sample sizes are too small to be conclusive in most cases, but the results are similar: IDENT-IAFIS gives more accurate results for pairs of index fingers; FBI 12k favors pairs of thumbs; Benefits favors pairs of index fingers; and Ohio shows no effect. As shown in Figure 28, the unexpected finding for IDENT-IAFIS is supported by a relatively large sample size (note the image type for each line).

Several 4-finger combinations were also analyzed (index-thumb pairs, index-middle pairs, and right slap for IDENT-IAFIS and Ohio), but no significant results could be detected.

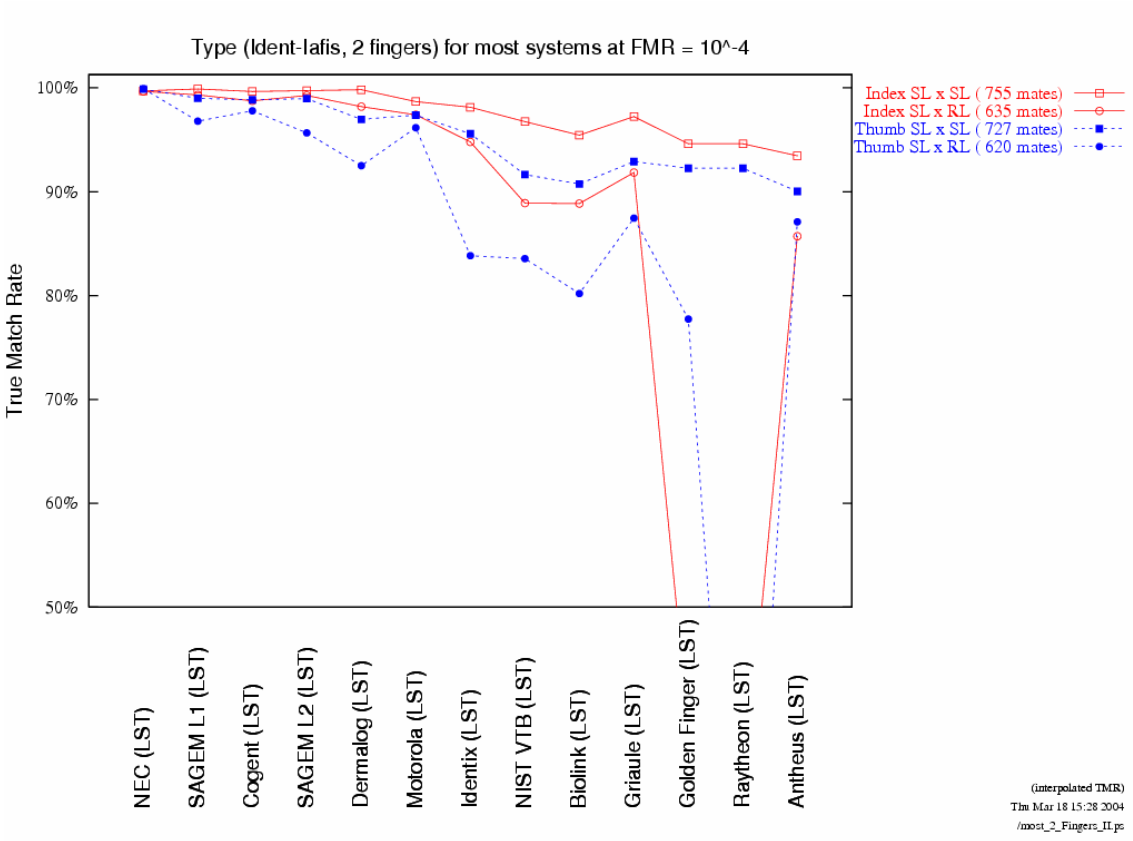

Figure 28. Index fingers outperform thumbs in 2-finger IDENT-IAFIS data

5.5.3 Effect of Sex

Females are known to have characteristically different fingerprints than males (e.g., fewer minutiae and more closely spaced ridges). Preliminary analysis results appeared to confirm that females are harder to match than males. Closer analysis, however, revealed that there were proportionally more males in the easier partitions, and that no clear effect of sex could be measured.

To pursue the question further, five partitions were constructed from the LST data, again controlling data source and image type. These partitions were selected based on the number of mated female pairs; partitions where many systems achieved accuracy near 1.0 were excluded. Figure 29 shows that in each partition, neither sex is consistently more difficult across all systems.

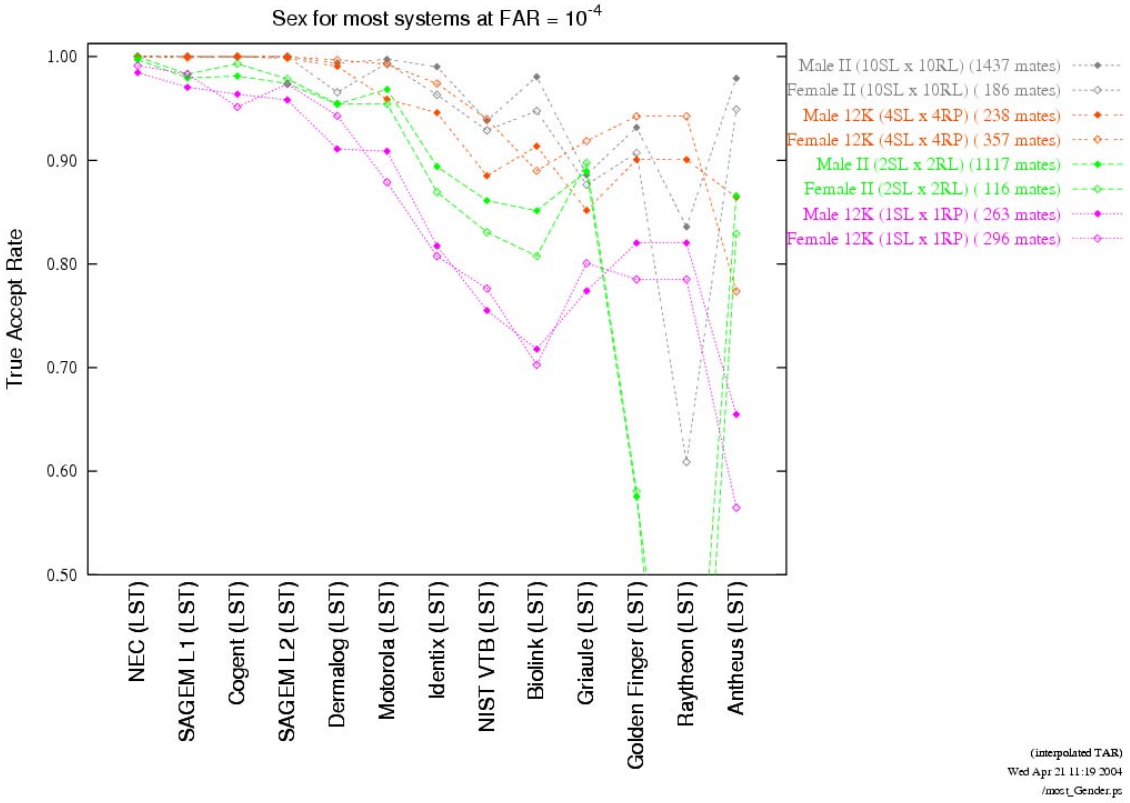

Figure 29. No evidence of an effect on accuracy based on sex.

5.5.4 Effect of Subject Age

The capture date and/or date of birth were not known for many of fingerprints in FpVTE. In the cases for which the capture date was known, the images were captured over the period from 1995 through 2003.

Figure 30 confirms a general trend where accuracy drops as subject age at time of capture increases, especially for subjects over 50 years of age. The sample sizes for this data are small, but the results are largely consistent with expectations and across systems.

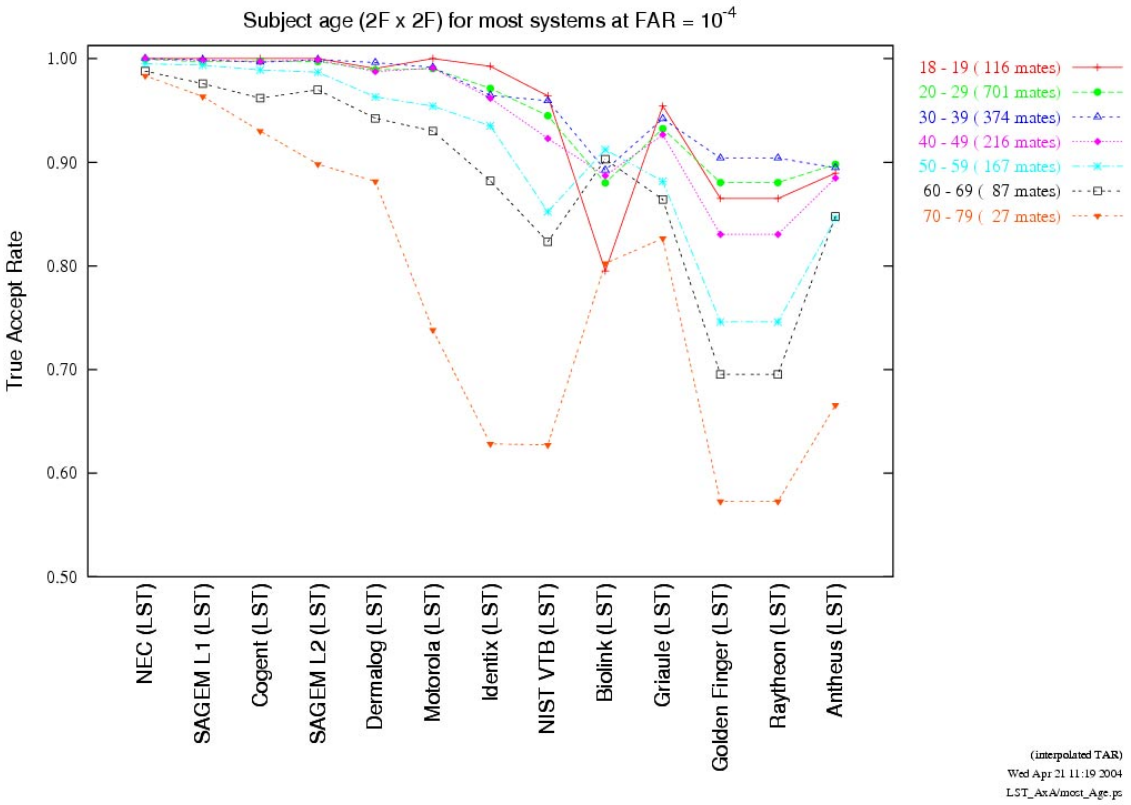

Figure 30. Accuracy is lower for older subjects (LST)

The observed effect of decreased accuracy on older subjects might be explained in terms of image quality. Since image quality information was not available in LST, this effect is shown using MST results. Figure 31 shows MST results by age group; Figure 32 shows the same results, but limits the data to Quality A images. Note that the distinctions between age groups lessen dramatically when poorer-quality images are ignored.

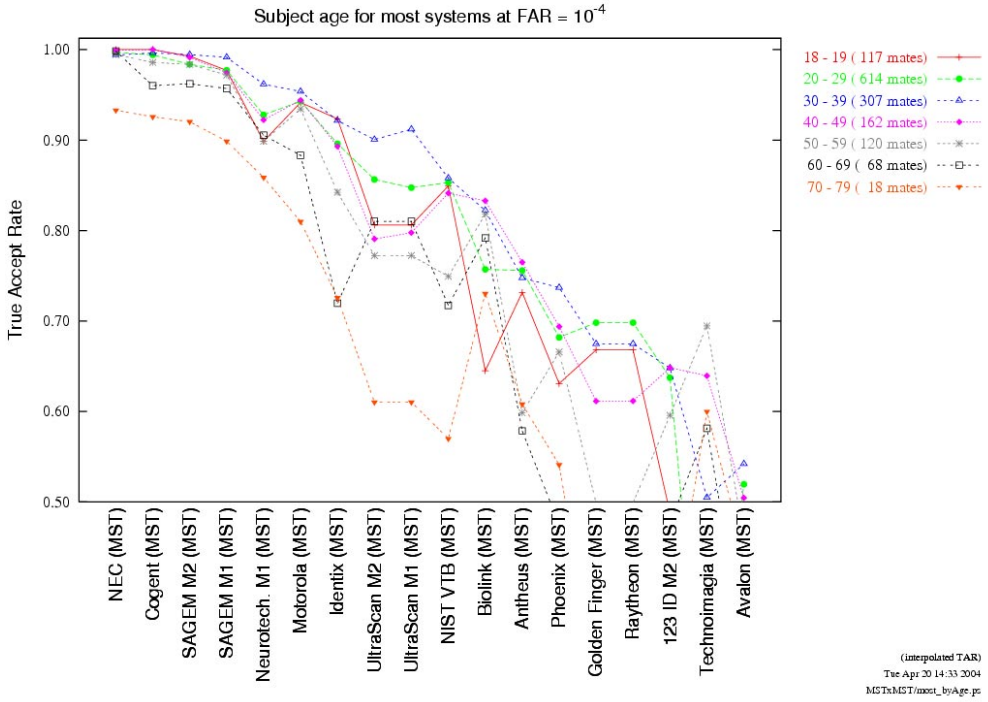

Figure 31. Accuracy is lower for older subjects (MST)

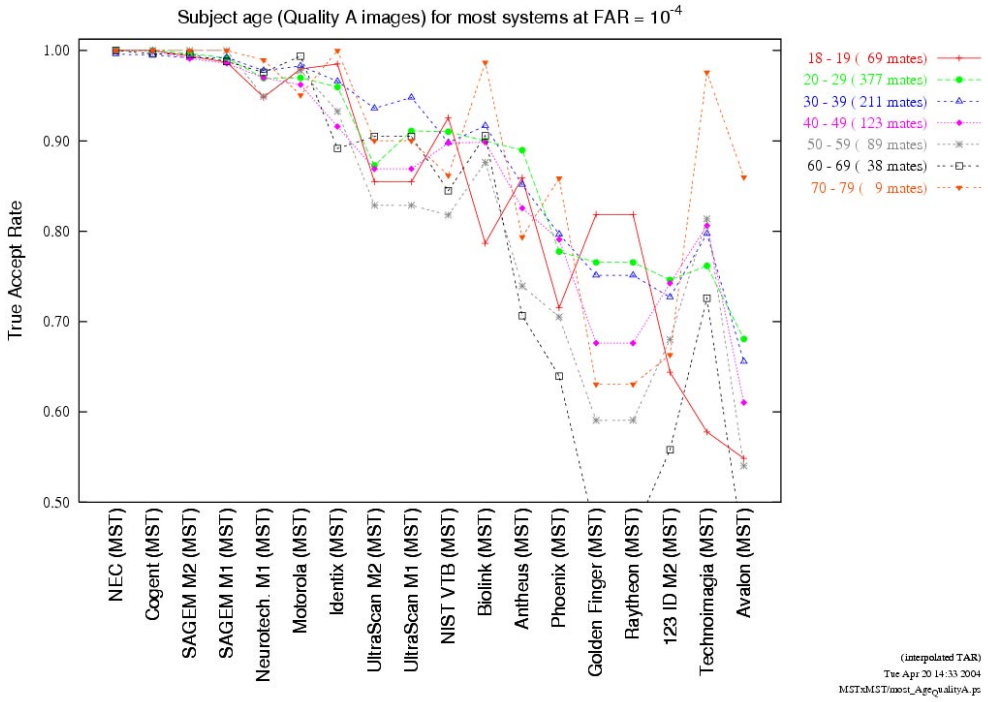

Figure 32. When poor-quality images are excluded, age has no clear effect

5.5.5 Correlation of System Results and the Potential for Fusion

Determining the level of correlation between matchers is important because it is an indication of the potential for fusion of the matchers. If the scores or ranks of two matchers are not strongly correlated, there might be a potential for a fused system that could have results better than either of the two separately.

This section examines the correlation of three top MST systems, and shows that a fused system can perform better than even the most accurate systems in FpVTE. This is in no way intended to be an exhaustive examination of the complex methods of implementing multi-system fusion: it merely shows that there is a good potential for improving results through fusion.

Table 21 shows the correlation between the NEC, Cogent, and SAGEM M2 MST systems. Correlations are shown both for the mate scores and mate ranks. The Cogent and SAGEM systems show a fairly high correlation, but the NEC system shows a surprisingly low correlation to the other two, given that all three systems perform as well as they do.

Correlation of Mate Scores			
	NEC	Cogent	SAGEM M2
NEC	1	0.51	0.55
Cogent	0.51	1	0.77
SAGEM M2	0.55	0.77	1

Correlation of Mate Ranks			
	NEC	Cogent	SAGEM M2
NEC	1	0.56	0.51
Cogent	0.56	1	0.78
SAGEM M2	0.51	0.78	1

Table 21. Correlation of Mate Scores and Ranks for Top MST Systems

To test whether a fused system could improve accuracy, a particularly simple method of fusion was used to fuse the NEC and Cogent results: each score in the NEC similarity matrix was multiplied with the corresponding Cogent score, and a new similarity matrix was generated. The results of even this simple test are indicative of an improvement, as shown in Figure 33. Other, more sophisticated methods of fusion may improve upon this result.

Any observed improvement in performance does not necessarily imply that such an improvement may be realizable, and would require further evaluation.

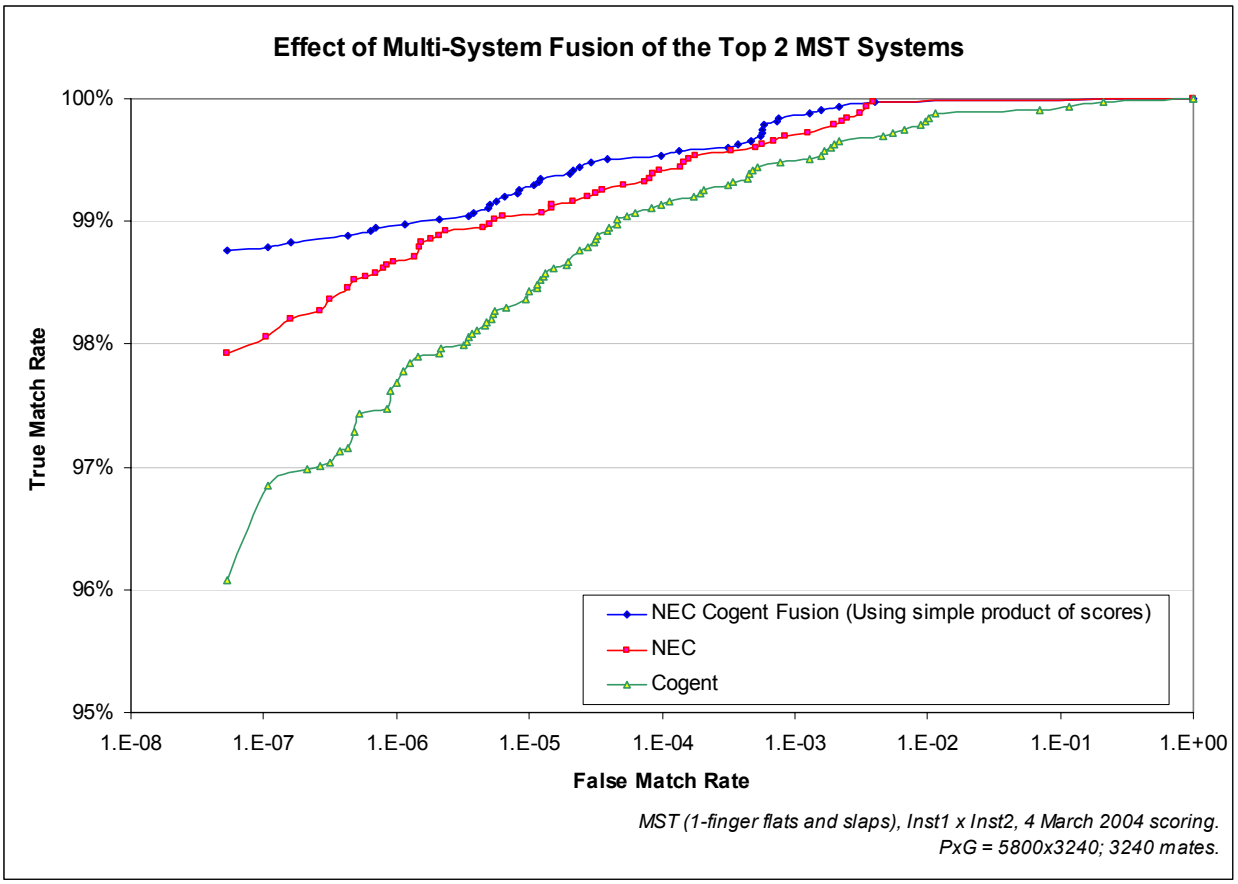

Figure 33. Effect of Multi-System Fusion of the Top Two MST Systems

5.5.6 Effect of Multiple Mates and True Imposters

In every FpVTE test, when searching a Query set against a Target set, the fingerprints in the Query set had zero, one, or more matching fingerprints in the Target set. In FpVTE analysis, results were only reported for single-instance partitions, so that each probe had at most one mate in the gallery. The effect of multiple mates is of interest in the biometric community. FpVTE results showed that allowing multiple mates in a test had a small but distinct effect on accuracy. This is shown in Figure 34. The Standard (multi-instance) line includes a large number of multiple mates, while the red and blue lines are single-instance. For every system, there is a small increase in accuracy if multiple mates are included.

The effect of "true imposters" is of great interest in the biometric community. A true imposter is a person who is in the probe set and does not have a mate in the gallery. Figure 34 shows the effect of including true imposters. The "with imposters" line was constructed from 5800 probes, 3240 of which have a mate, and 2560 which have no mate; these were searched against a gallery of those 3240 mates. The "no imposters" line was constructed from the same 3240 probes, each having exactly one mate in the gallery; these were searched against the same gallery of 3240 mates. The addition of more than 2500 true imposters had no discernable effect on accuracy: the red and blue lines are effectively identical.

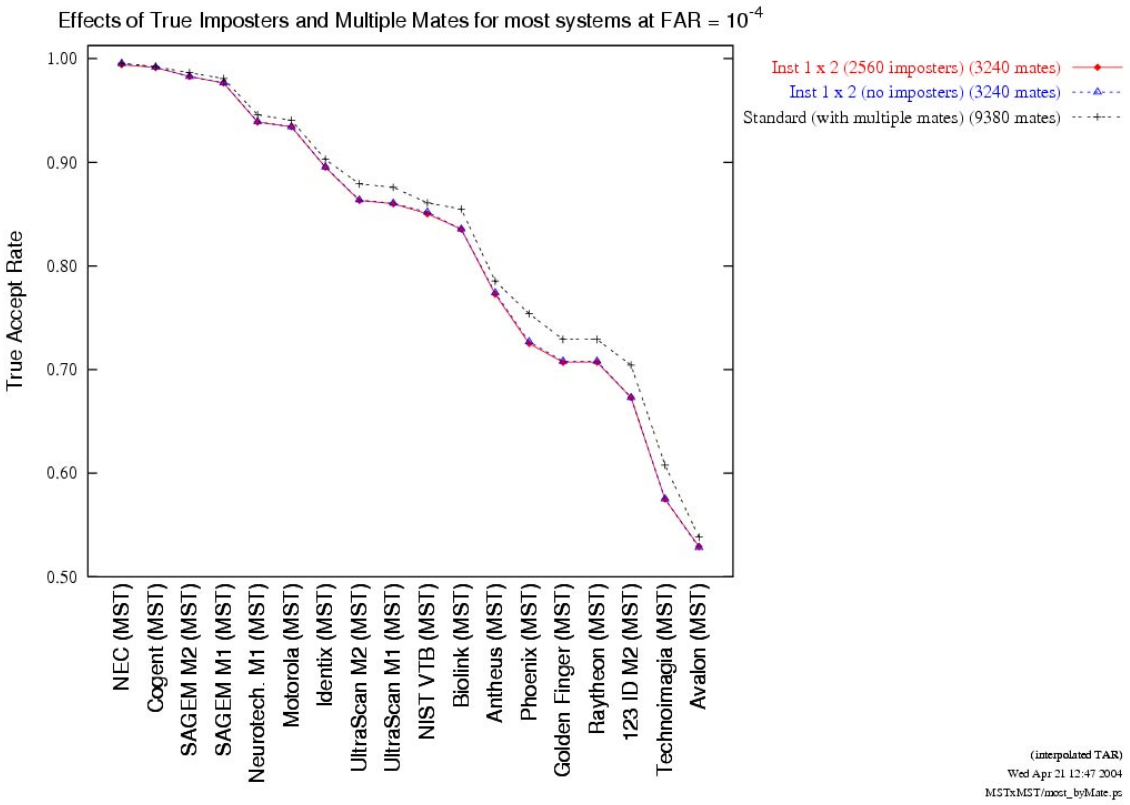

Figure 34. Effect of Multiple Mates and True Imposters. Multiple mates have a small effect on accuracy: this can be seen by comparing the Standard line with either of the other lines. True imposters have no discernable effect on accuracy: this can be seen in the fact that the red and blue lines do not differ.

Section 6: Conclusions

1. The systems developed by NEC, SAGEM, and Cogent were highly accurate

On 44 test partitions defined by fingerprint type, number, and source, the NEC LST system was capable of identifying more than 98% of the mates in *every* subtest, with a false accept rate of .01%.

Given a false accept rate of .01% and the NEC LST system:

- Every single-finger subtest had a true accept rate higher than 98.6%
- Every 2-finger subtest had a true accept rate higher than 99.6%
- Every 4, 8, or 10-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 .01%.

1a. These systems were found to be the most accurate across all FpVTE tests

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, 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.

In LST, the most accurate of the other systems were developed by Dermalog and Motorola, which had comparable performance.

In 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.

1b. These systems performed consistently well over a variety of image types and data sources

The most accurate systems maintained high accuracies even on data on which other systems performed poorly.

2. There was a substantial difference in accuracy among the systems

2a. Many systems performed well on some types of data, particularly on ten-finger tests

2b. There was a clearly measurable 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 of participating systems. This corresponds to a False Reject Rate much more than 10 times that of the best 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 difference is large and consistent. 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.

As a rough rule of thumb, the false reject rate was measured to be nearly 10 times greater for single-finger comparisons than for two-finger comparisons; the false reject rate for two-finger comparisons was 10 times greater than for 10-finger comparisons. Actual differences varied by dataset and by system, but the general trend was quite consistent.

At the test sizes used, the accuracy of four, eight, and ten-finger searches was often difficult to differentiate. This does not mean that four, eight, and ten-finger searches would be equivalent in a larger test or an operational system. It is highly likely that further improvements can be achieved; however, the test data size would need to be increased substantially.

3b. Poor quality fingerprints greatly reduce accuracy

Accuracy on good quality images was much higher than accuracy on poor quality images for all systems. Some systems were particularly sensitive to poor image quality. For example, the Technoimagia MST accuracy of 84% for the highest-quality fingerprints dropped to 2% for the lowest quality fingerprints[1]. NEC MST achieved an accuracy of 99.8% for the highest-quality fingerprints, which dropped to 82% for the lowest quality fingerprints.

[1] Quality D and F combined

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 Dept. Border Crossing Card (BCC) data and the DHS Recidivist (DHS2) data used the same scanners and software, but are substantially different in quality. Using the FpVTE image-quality metric, 80% of BCC is good quality, but only 45% of DHS2. For most systems, there is a clear difference in accuracy between the two datasets.

The subject population, 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 type of data

Performance on one type of data is not necessarily similar to performance on another type of data. The False Reject Rate for a system often varied by a factor of 2 or more between different datasets.

Some systems showed an unusually high sensitivity to the sources or types of fingerprints; while other systems did not. For example, in SST Cogent had a true accept rate of 99.6% for BCC data and 100% or DHS2, at a false accept rate of .1%. The NIST VTB had a true accept rate of 90.1% for BCC data and 87.4% for DHS2. Bioscrypt had a true accept rate of 97.2% for BCC data and 66.8% for DHS2.

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. A biometric evaluation that only uses a single type of data is limited in how it can measure or compare systems

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 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 data. Mating errors were found 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 mating errors were corrected in FpVTE as part of analysis.

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%.

7. The most accurate fingerprint systems are far more accurate than the most accurate face recognition systems.

The most accurate fingerprint systems are far more accurate than the most accurate facial recognition systems, even when comparing the performance of operational quality single fingerprints to good quality facial images

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 (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 in the gallery, the performance of face recognition improves.[1] With four images in the gallery:

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

When four fingerprints are used for matching, the most accurate fingerprint system (NEC LST) always has true accept rates in excess of 99.9%.

[1] Results of fusion from [FRVTSupp].

Section 7: Future Work

FpVTE 2003 answered a number of questions about the state of the art of fingerprint matching. However, an immense amount of data was collected that has yet to be analyzed: a number of areas still require further analysis. This section identifies key areas for future work.

Projection of results to operational sizes

FpVTE 2003 was conservative in its choice of operating points, and did not attempt to project results to the very large database sizes required by operational systems. IAFIS currently has criminal fingerprint sets from 46 million individuals; it is expected that U.S. VISIT will build an even larger database in a rather short time. Using appropriate techniques, the results from FpVTE might be projected to larger database sizes, at least to the extent of determining bounds on performance. A variety of methods are known for such projections, such as extreme value statistics or various methods of extrapolation, as well as methods for determining confidence intervals associated with the projections. This area for future work would include investigation and application of various projection methods in order to take full advantage of the results from FpVTE.

Fingerprint quality analysis

Only a limited study of the effect of fingerprint quality was conducted in FpVTE, but the results have shown that fingerprint quality has a clear effect on matcher accuracy. Operational systems need effective image quality metrics for real-time operator feedback, and as a method of identifying sources of poor quality fingerprints.

The quality of FpVTE 2003 fingerprints, collected from a range of governmental sources, varied from good to poor quality and included a variety of different characteristics. An investigation of performance is needed to characterize the performance of the most accurate systems and to plan for operations in large database environments. This area for future work would involve an analysis of the efficacy of proprietary image quality metrics, as well as open-source fingerprint quality metrics now in development.

Further testing of most accurate systems

NIST is currently conducting further testing of a variety of systems, including ones identified as most accurate in this evaluation. This ongoing effort uses participant-provided Software Development Kits (SDKs) for additional analysis beyond the scope of FpVTE, and preliminary results have corroborated the results of FpVTE. This testing involves larger datasets for better discrimination among the systems, to better measure multi-finger accuracy, and to measure accuracy at lower false match rate settings.

Evaluation of slap segmentation algorithms

In FpVTE, all slap fingerprints were automatically segmented, then manually verified. Slap segmentation algorithms are imperfect. If slap fingerprints are used operationally without a good understanding of the efficacy and error rates associated with slap segmentation algorithms, slap segmentation may be a source for fingerprint quality problems and degraded performance. Future work should be performed to evaluate slap segmentation processes and their effect on accuracy.

Fusion of results from multiple systems

As shown in this report, there is the potential to improve accuracy by the fusion of results from multiple systems. The fusion method used was effective on the sample data, but was a particularly simple algorithm for match score level fusion. Future work on more sophisticated approaches and repeatability would lead to improved results. Further work in the fusion of the results from different matchers could have a dramatic payoff in increased accuracy for operational systems.

Evaluation of matching using minutiae templates instead of images

Several of the FpVTE Participants have expressed interest in a follow-on "minutia exchange" test, in which each system would create minutiae templates (using M1 and/or FBI formats) and search using the templates created by all systems. Further work is necessary to determine the efficacy of such an approach.

Evaluation of the efficacy of latent searches of flats and slaps

Conventional wisdom is that a database of rolled fingerprints is necessary for effective searching of latent fingerprints. Since some databases may move to the collection of slap fingerprints, it is especially important to determine the effect this move may have on latent fingerprint searches. Few operational databases (such as Los Angeles) conduct latent searches against slaps.

Measurement of system throughput (speed)

FpVTE purposely limited its scope to the analysis of accuracy, not system throughput. However, for any operational system, throughput is a key parameter. Future work should analyze tradeoffs between accuracy and system throughput.

Acknowledgements

The authors gratefully acknowledge the sponsors and supporters of FpVTE:

- Justice Management Division, US Department of Justice, IDENT/IAFIS Project
- National Institute of Standards and Technology
- Bureau of Immigration and Customs Enforcement (U.S. Department of Homeland Security)
- Federal Bureau of Investigation
- U.S. Department of State
- U.S. VISIT Program (U.S. Department of Homeland Security)
- Ohio Office of the Attorney General
- European Commission Services
- Office of the Chief Information Officer, U.S. Department of Justice
- Royal Canadian Mounted Police
- U.K. Police Information Technology Organisation (PITO)
- U.S. Department of Homeland Security
- U.S. Department of Justice

The authors extend their thanks to:

- Chris Surmacz, Abhishek Sindhwani (Northrop Grumman), Keith Casey, John Lewington (UTA) for all of their work as Test Agents
- Jackie Bell, Bob Gallup, Tom Martin, Jim Parker, Ambrose Sampson, and Ed Sears (DHS) for their incredible diligence and patience as fingerprint examiners in the excruciating groundtruthing process.
- Steve Wood, Mike Garris, and Mike McCabe at NIST for providing assistance and insights
- Frank Boyle, Robert Scott and Mike Archer for their guidance
- Brian Finegold (UTA) and Frank Torpey (Northrop Grumman) for their assistance
- Rama Krishnan (Lockheed Martin) for his assistance in obtaining some of the IDENT/IAFIS data
- Nirav Desai, Ted Unnikumaran, Mike Weinreb, and Eeshat Ansari at Mitretek for their assistance in slap segmentation verification.
- The FRVT 2002 team for providing a good and useful basis for evaluation, administration, and analysis methodologies.
- Duane Blackburn at the FBI for his suggestions on test design and administrative issues.
- Jim Wayman (San Jose State University) and Joseph Campbell (MIT) for their advice and expertise.

Glossary

ANSI/NIST	A file format for fingerprint files compliant with NIST Special Publication 500-245, Data Format for the Interchange of Fingerprint, Facial, & Scar Mark & Tattoo (SMT) Information (ftp://sequoyah.nist.gov/pub/nist_internal_reports/sp500-245-a16.pdf) The FBI's Electronic Fingerprint Transmission Specification (EFTS) is based on ANSI/NIST. Fingerprint files that are EFTS compliant are necessarily ANSI/NIST compliant. In FpVTE, all images embedded in ANSI/NIST files use WSQ compression.
Consolidation	A mating error in which the same person has fingerprint sets under different names or IDs
Controlled data	Data collected under controlled conditions; generally higher quality than operational data.
Dataset	A collection of multiple fingerprint sets.
EFTS	The FBI's *Electronic Fingerprint Transmission Specification* [EFTS] file formats and transactions for fingerprints. It is based on ANSI/NIST. Fingerprint files that are EFTS compliant are necessarily ANSI/NIST compliant.
Fingerprint set	A single ANSI/NIST file containing multiple fingerprint images from a single individual, collected at one time. The fingerprint positions (finger numbers) are noted in the file. The finger positions are not repeated in the file: no more than one fingerprint per position is included.
Flat fingerprint	A fingerprint image collected from a single-finger livescan device, resulting from the touching of a finger to a platen without any rolling motion. Also known as a single-finger plain impression.
Match	Two fingerprint images match if they came from the same finger of a person. Equivalent to Mate.
Mate	Two fingerprint images are mates if they came from the same finger of a person. Equivalent to Match.
Misidentification	A mating error in which fingerprints from different people are listed under the same name or ID
Normalization	A mapping function that operates on the entire list of scores generated by each Query and maps the raw similarity scores to "normalized" scores. Normalization blurs the distinction between 1:1 and 1:N matching. In one common implementation, the mean and standard deviation of the set of raw similarity scores associated with a query are determined and the normalization function adjusts each score such that the resulting, normalized distribution has a mean of 0.0 and a standard deviation of 1.0. Another

	common implementation is to use the median of raw similarity scores rather than the mean.
Operational data	Data from an operational database, collected under real-world conditions. The quality of operational data may vary greatly.
Partition	A portion of a Query Set or Target Set defined by some variable, such as sex, quality, etc. A Probe Set is a partition of a Query Set; a Gallery Set is a partition of a Target Set.
Preprocessing	Also known as Characterization or Feature Extraction. The process of creating a machine representation of a fingerprint image. A few matchers do not perform preprocessing.
Query Set	The dataset containing the searches for a given test or subtest: an experiment searches a Query Set against a Target Set. Also known as a Search set.
ROC	A Receiver Operator Characteristic graph shows the tradeoffs between True Match Rate and False Match Rate. The False Match Rate is traditionally in the X axis, in log scale.
Rolled fingerprint	A fingerprint image collected by rolling the finger edge to edge across the livescan platen (or paper) from nail to nail. Rolls may be from livescan devices or scanned from paper fingerprint cards.
Segmented slap	An image of a single fingerprint that was segmented (cropped) from an image of a 4-finger slap (4-finger simultaneous impression), such as found at the bottom of a fingerprint card. Slaps may be from livescan devices or scanned from paper fingerprint cards. FpVTE segmented slaps have been segmented using automatic and/or manual processes; all segmentation has been human verified.
Self-ident	The special case of a fingerprint set (or an individual fingerprint) being compared against itself. Self-idents are ignored during analysis. When a dataset is compared against itself and a square matrix of scores is generated, the scores on the diagonal are self-idents.
Similarity matrix	A matrix of Participant-specific matcher scores, which compare each member of a Query Set against each member of a Target Set. The file format for a similarity matrix is defined in the Data Format Specification.
Subject	An individual person
Target Set	The dataset being searched against in a given test or subtest: an experiment searches a Query Set against a Target Set. Also known as a File set or just fingerprint database. A Gallery is a subset of a Target Set.
WSQ	Wavelet Scalar Quantization. The standard image compression method used for fingerprint images stored in ANSI/NIST format files.

References

[ANSI/NIST] *Data Format for the Interchange of Fingerprint, Facial, & Scar Mark & Tattoo (SMT) Information*; NIST Special Publication 500-245. (ftp://sequoyah.nist.gov/pub/nist_internal_reports/sp500-245-a16.pdf)

[BestPractices] A. J. Mansfield and J. L. Wayman. *Best Practices in Testing and Reporting Performance of Biometric Devices*, Version 2.01, August 2002

[EFTS] FBI CJIS. *Electronic Fingerprint Transmission Specification*. Version 7. 29 January 1999. http://www.fbi.gov/hq/cjisd/iafis/efts70/cover.htm

[FRVT2002] Phillips, Grother, Micheals, Blackburn, Tabassi, Bone. *Face Recognition Vendor Test 2002*. March 2003. http://www.frvt.org

[FRVTSupp] Grother. *FRVT 2002: Supplemental Report*. February 2004. NISTIR 7083. http://frvt.org/FRVT2002/documents.htm

[Grother1] Patrick Grother et al, Face *Recognition Vendor Test 2002 Performance Metrics*. March 2003. NIST IR 6982.

[Grother2] Patrick Grother, *Open-Set 1:N Measurement for Policy Formulation* (Briefing). July 2003. (http://www.mitretek.org/biometricslides/16P.Grother_102103.ppt)

[IAFISCert] *Products Certified for Compliance with the FBI's Integrated Automated Fingerprint Identification System Image Quality Specifications*. 30 January 2004. http://www.fbi.gov/hq/cjisd/iafis/cert.htm

[IQS] D'Amato, Hicklin, Khanna, Kiebuzinski, Nadel, Splain. *IDENT/IAFIS Image Quality Study*. December 2000.

[NIST IQS] Hicklin, Reedy. *Implications of the IDENT/IAFIS Image Quality Study for Visa Fingerprint Processing*. October 2002.

[NFIS] NIST Fingerprint Image Software (NFIS). http://www.itl.nist.gov/iad/894.03/databases/defs/nist_nfis.html

[NIST SD29] National Institute of Standards and Technology, NIST Special Database 29, *Plain and Rolled Images from Paired Fingerprint Cards*. http://www.nist.gov/srd/nistsd29.htm

[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)