FINGERPRINT VENDOR TECHNOLOGY EVALUATION 2003

APPENDIX E

TEST DESIGN AND ANALYSIS ISSUES

Table of Contents

Introduction

This Appendix discusses in detail several test design, methodology, and analysis issues introduced in the FpVTE Analysis Report.

1 Data Selection Issues

Section 3.2.4 of the FpVTE Analysis Report (Data Selection Issues) briefly highlighted key issues involved in data selection. Further explanation is included here.

1.1 Problems due to Controlled Collection of Data

Controlled collection of fingerprints results in better-than-operational quality data, which result in inflated measures of accuracy. Quality problems in operational data result from factors such as uncooperative subjects, time constraints, poor training, poor maintenance, poor facilities, stress, and overwork. In non-operational controlled data collection, these are not generally issues.

Even in the collection of test data in operational environments, using operational staff and equipment, the data quality will often improve due to a Hawthorn Effect: the knowledge that they are being tested will affect operators' behavior.

The inflated measures of accuracy that result from controlled collection of fingerprints may lead to unreasonably high expectations of real-world accuracy. In addition, such a test does not evaluate the ability of a matcher to process fingerprints of varying quality; such an evaluation is necessary for any system that might be implemented in an operational system. For that reason, exclusive use of non-operational fingerprints in system evaluation can bias results: systems that can ***only*** process good, non-operational fingerprints would not be differentiated from systems with broader capabilities.

In FpVTE, only one dataset (Ohio) was collected under controlled conditions. This was purposely included so that systems could have one set of measurements under optimal conditions.

1.2 Avoiding Matcher Bias in Data Selection

The problem of bias in data selection can be illustrated through a cautionary example from the IDENT/IAFIS Image Quality Study:

> *DS3 was collected using the IDENT matcher to determine the mates. Furthermore, of the mated pairs identified, generally those pairs with the highest matcher scores were selected. This resulted in a data set that was dramatically skewed toward matchability since the usual distribution of hard-to-match mates was not included in the data set. This bias rendered the data set virtually unusable. [NIST IQS, p 17]*

Matcher bias[1] is what happens when one matcher is used to select data that will then be used to evaluate other matchers. There are several problems with this:

- If the matcher used to select the data is included in the evaluation, then that matcher should be expected to perform unusually well on the evaluation.

[1] Also known as AFIS bias.

- If any matcher is used to select data, then all matchers (or at least all similar matchers) should be expected to have better-than normal results because using a matcher to select data eliminates all the mate comparisons for which that type of matcher is not effective.

Matcher bias may be direct or indirect:

- Direct matcher bias results when the matcher selects two specific fingerprint images as mates: those two fingerprint images are machine matchable by that type of matcher. Any similar or better matcher can be expected to match those fingerprints. Direct matcher bias is a problem even if the only purpose of the evaluation is to compare systems, since those matchers most similar to the matcher used for selection have a distinct advantage.
- Indirect matcher bias is more subtle: it results when a matcher is used to identify a person, and then a different set of that person's fingerprints are selected. At first it seems that there is no bias: the specific set of fingerprints were not selected by a matcher. What in fact happens is that the matcher selects a person whose fingerprints can be matched, effectively filtering out those people whose fingerprints are intrinsically difficult to match. Indirect matcher bias is very difficult to avoid when using large datasets. Indirect matcher bias should have little effect on the comparison of systems in an evaluation, but would be expected to inflate the true match rate slightly over an operational system.[2]

In short, direct matcher bias results in the selection of specific images that are easier to match than is true for the population as a whole. Indirect matcher bias results in the selection of individuals whose fingers are easier to match than is true for the population as a whole.

Whenever possible, mated fingerprints from operational databases that were used in FpVTE were selected using non-fingerprint means such as identification by name, FBI number, or other demographic information. Rolls and the corresponding slaps are known to be matches due to simultaneous collection. In FpVTE, mates that were known to be selected with a direct matcher bias were not included. Indirect matcher bias could not be avoided in some cases. One example of indirect matcher bias in FpVTE is in the selection of matching sets in the FBI 12k dataset: sets of rolled fingerprints were mated using the FBI's IAFIS, and the corresponding slap sets were included in FpVTE. The mate relationships between slap sets are indirectly biased in this case.

1.3 Avoiding Bias due to Data Filtering

If an evaluation is to get measurements that will correspond to operational performance, then the data used must correspond to real-world data. One frequent issue encountered when preparing datasets for evaluation is what to do with problem cases.

There are different types of problem cases that can occur when analyzing a fingerprint database:

- The mate relationships are found to be in error (covered in detail in Section 1.4)

[2] Indirect matcher bias can have some effect on the comparison of systems in some circumstances. For example, a pure topological matcher can match fingerprints from people who have few or no minutiae (a very small part of the population). Using a minutiae-based matcher to select mates would not select mates from this portion of the population, which could be matched using a pure topological matcher.

- The intrinsic fingerprints or the fingerprint images are of poor quality. If the fingerprints have matching fingerprints in the evaluation, there are several possibilities:
  - The images are poor quality, but can be verified by a human expert
  - The images are poor quality, cannot be verified by a human expert, but other means allow verification (such as the other fingers in a 10-finger set)
  - The images are poor quality, and cannot be verified by any means
- The fingerprint sets are not valid
  - The images are blank
  - The fingerprints are incorrectly segmented from fingerprint cards or slap images
  - There are image processing errors
  - The fingers are out of order (such as left/right hand swaps)

The instinctive reaction of most people is to remove all problem cases from the datasets. To some extent this is more fair for the Participants, since it is clearly unreasonable to expect a matcher to process a completely blank image.

The problem with removing all problem cases is that operational databases have such problems. If the evaluation datasets are representative samples of the operational databases, such problem cases must be included in the evaluation to depict operational performance accurately. In addition, subjective removal of fingerprints due to poor quality can remove fingerprints that some matchers can match, and therefore bias the comparison of systems: those matchers that cannot process poor quality fingerprints will have inflated performance.

In FpVTE, the operational data was left as undisturbed as possible:

- Erroneous mate relationships were corrected.
- Unmated poor quality fingerprints were left in the datasets.
- Mated poor quality fingerprints were left in the datasets if the mating could be verified by any means, but particularly poor quality fingerprints were noted.
- Invalid fingerprints were ignored during the analysis of similarity scores.

1.4 Groundtruthing and Manual Validation of Data

In any biometric evaluation, the performance of a system is compared against the known mate and non-mate relationships in the data. The evaluation requires comparing measured results to known results. The problem is that it is difficult to achieve a high degree of certainty about mating.

Simply accepting the stated mating information is problematic, because many datasets, both operational and controlled, have mating errors. After FRVT 2002 was published, the fingerprints associated with the face images were used to double-check identities, and it was discovered that 1.7% of the images had incorrect mating information.

If such errors are not corrected, the mating error rates of the underlying data define bounds for the effective error rates for the system. For example, if 1% of the mates in an evaluation are actually non-mates, then no system could possibly achieve more than 0.99 TAR. An evaluation that attempts to measure high precision results cannot succeed if the mating error rates exceed those operating points.

PROCEDURE

The original source for the *a priori* mating information was the Government source of the fingerprints. The analysis team, with a great deal of help from fingerprint examiners for the U.S. Department of Homeland Security, revised the mating information based on visual analysis of selected pairs. The selection of pairs for examiner review was performed in an iterative manner, as unbiased as possible. To minimize bias, each system's high-scoring non-mates and low-scoring mates were selected for examination; this process was repeated iteratively after mating corrections were made.[3] The analysis software reprocessed each participant system's similarity matrices with the revised groundtruth.

TYPES OF PROBLEMS

A consolidation is a case in which two fingerprint sets are labeled as belonging to different people when in fact they belong to the same person. Consolidations are very common when multiple datasets are merged.

A misidentification (or misident) is a case in which two fingerprint sets are incorrectly labeled as belonging to the same person. Misidents can often be attributed to typographical or other administrative errors during processing.

Sequence errors occur if fingers are switched, or the fingers are out of order. Sequence errors are not a mating error, but they have similar effects.

Self-identifications (self-idents) are cases in which two fingerprint sets contain exactly the same images. Self-idents were excluded during analysis. Self-idents are common when multiple datasets are merged.

Fingerprints noted by the examiners as particularly poor quality were labeled "Examiner poor" — they were not removed from the datasets, but were considered Quality F for image quality analysis.

RESULTS

The results of the groundtruthing process are summarized in Table E-1. 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. Note that consolidations and misidentifications were corrected by correcting the mating information for the test, but that fingerprint sets with other errors were ignored during analysis.

Based on our review of the data, the FpVTE datasets used for analysis contained very few if any residual misidents, consolidations or self-idents. Any missed consolidations would only have affected results for FAR values far less than the standard reporting level of 10^{-4}.

[3] Systems with very large numbers of comparisons tied for highest or lowest scores did not have all of them considered.

Groundtruthing Results				
Total people in FpVTE	25,309			
Total fingerprint sets in FpVTE	48,105			
Fingerprint sets viewed	5,174	fingerprint sets	10.76%	
Visual comparisons (Unique pairs)	3,177	pairs of fingerprint sets		
Issue			**% of whole**	**Removed from test**
Consolidations between different datasets	100	people	0.40%	No
Consolidations within a dataset	24	people	0.09%	No
MisIdents	119	people	0.47%	No
Examiner Poor	49	fingerprint sets	0.10%	No
Sequence Errors	28	fingerprint sets	0.06%	Yes
Missegmentation	17	fingerprint sets	0.04%	Yes
Other Unusable	2	fingerprint sets	0.00%	Yes
Self ident, same person	126	people	0.50%	Yes
Self ident, different people, same dataset	4	pairs of fingerprint sets		Yes
Self ident, different datasets	94	pairs of fingerprint sets		Yes

Table E-1. Groundtruthing Results

2 Methods of Comparison

In the FpVTE Analysis Report, the verification (1:1 matching) performance statistic was used as a basis for measuring and comparing the accuracy of systems. Measures of identification accuracy were found to yield essentially the same conclusions.[4]

This section discusses how the verification performance statistic was applied to rank the systems. Several decisions were made in this process that affected the final ordering of the systems:

- Choice of the data partitions used for comparison
- Choice of the operating point for comparison
- Methods of differentiation between systems

2.1 Performance Statistics

Biometric systems can be used in a variety of applications, including identification and verification. Depending on the application and test design, different evaluation metrics are appropriate. The verification ROC — the primary analysis tool in this evaluation — evaluates systems based on their ability to do 1:1 comparisons. Large fingerprint matchers such as automated fingerprint identification systems (AFISs) typically perform identifications using 1:N strategies, often filtering out many of the candidates quickly to minimize resource expenditures. Following this approach on a verification test can be disadvantageous. Some applications assign a high cost to false matches (e.g., causing innocent persons to be detained, or extra work for human operators) and some assign a high cost to false non-matches (e.g., criminals going undetected). These differences in applications make some systems better suited to one type of application over the other. Numerous methods for evaluating biometric systems have been documented.

2.2 Data used for Comparison

The choice of data partitions used for comparison affects ranking in several ways:

- Some systems are relatively more accurate on some types of data, less accurate on others. The rank order of some systems will change depending on the mix of data included in the ranking metric. For instance, changing the proportion of high vs. low quality data in the test, or the proportion of single-finger vs. ten-finger comparisons will favor certain systems over others.
- Some systems have very specific weaknesses. Including certain types of data can severely reduce the rank of these systems.
- Some sample datasets result in relatively easy tests, where the top systems achieve perfect or near-perfect scores. Observed differences among the top systems on these tests may be more spurious than meaningful.

The choice of sample data affects not only the competitive ranking order of participants, but also the overall impression of what accuracy fingerprint systems can achieve, and the extent to which systems differ in capability.

[4] See Section 5.4 of the FpVTE Analysis Report (Comparison of Verification and Identification Results)

For the purpose of ranking systems, all comparisons were based on the same data. Sample data was selected to reflect typical operational distributions; to expose strengths and weaknesses by including a variety of types of data; and to favor statistically meaningful comparisons.

LST

The LST systems were ranked based on their performance on a variety of test partitions. These partitions were not just the 31 subtests included in LST, because each subtest included data from different sources. Since systems perform differently on different sources, it was important to isolate this effect.

Each of the 31 subtests was partitioned by data source, yielding 95 partitions.

It is important to understand that the principle of unbiased, random sampling was applied to the design of each dataset, and hence to each test partition, but not to the entire collection of subtests and multiple partitions. Each test partition is qualitatively distinct from the others. So, for instance, there is no rule to say how many single-finger tests and how many ten-finger tests should have been performed.

The 95 were based on specific criteria: data source, number of fingers, and type of image. We had the ability to control other variables, so additional partitions might have been used. The 95 partitions also were not completely independent, and some of the 95 partitions were statistically small. The exclusion of many of the 95 was done to show performance over a range of conditions, specifically trying to avoid misrepresentations that can result from redundant test partitions and small datasets

A subset of those 95 partitions was selected based on the following criteria:

- All test partitions having fewer than 200 mated pairs were excluded, to avoid statistical anomalies due to small sample sizes (28 partitions).
- Partitions from redundant sets were excluded: DxC (redundant with CxD), and JxI (redundant with IxJ) (3 partitions).
- Partitions involving 4 and 8 fingers were excluded. By excluding this data, which was in many ways redundant with many of the remaining partitions, greater weight was given to types of data that were only represented in the 1 and 2-finger datasets. (20 partitions)
- Operational and non-operational (controlled) data were separated so that this effect would not dominate the outlier statistics (minimum and maximum accuracy).

This process yielded the 27 operational and 17 controlled partitions used for comparison of systems in Section 4 of the FpVTE Analysis Report. See Appendix D (Section 2.1) for more detail, and for a list of the partitions.

MST

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. See Appendix D (Section 2.1) for more detail, and for a list of the partitions.

SST

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. See Appendix D (Section 2.1) for more detail, and for a list of the partitions.

2.3 Operating Point

The choice of operating point affects ranking:

- An ROC shows a system's ability to trade off False Accepts for accuracy (True Accepts). When two systems' ROCs intersect, the rank order of the systems depends on the operating point selected. On FpVTE 2003, cross-overs seldom occurred at FAR < 10^{-4}.
- The most appropriate operating point for comparison depends on the target application. Large operational systems may be set to high values of TAR or low values of FAR, depending on the relative costs of the two types of errors.
- It is possible to compare systems by reporting FAR at a fixed TAR. However, when comparing systems across multiple datasets, a fixed TAR would often correspond to values of FAR that are either higher than typical operating values or lower than what could be meaningfully measured on these tests.

For the purpose of ranking systems, all comparisons were based on TAR at FAR = 10^{-4}. Although FpVTE allowed measurements of TAR at much lower values of FAR, rank order of the most accurate systems generally remained constant for lower values of FAR.

2.4 Presentation and Summary Statistics

Figure E-1 shows example ROCs for an MST subtest for seven of the most accurate (MST) systems. The choice of BCC data determines overall characteristics of the chart, most notably the overall difficulty of the test. Choosing to summarize this data at a specific FAR can affect competitive rank order of the systems (as seen here with Neurotechnologia and Motorola). The choice of a linear y-axis reveals a small absolute difference in TAR among the top systems and a clear separation in absolute performance between the most and least accurate systems; a log-scale y-axis would emphasize seemingly small differences among the most accurate systems that can be very significant under large-scale operating conditions. Because the x-axis is on a logarithmic scale, comparing the systems at a fixed TAR reveals differences in FAR that are measured in orders of magnitude. On this sample data, at TAR = 0.98, NEC's FAR is unmeasurably small; Cogent's FAR is 10^{-5}; SAGEM M2's FAR is 10^{-4}; and Neurotechnologija's FAR is 10^{-2}.

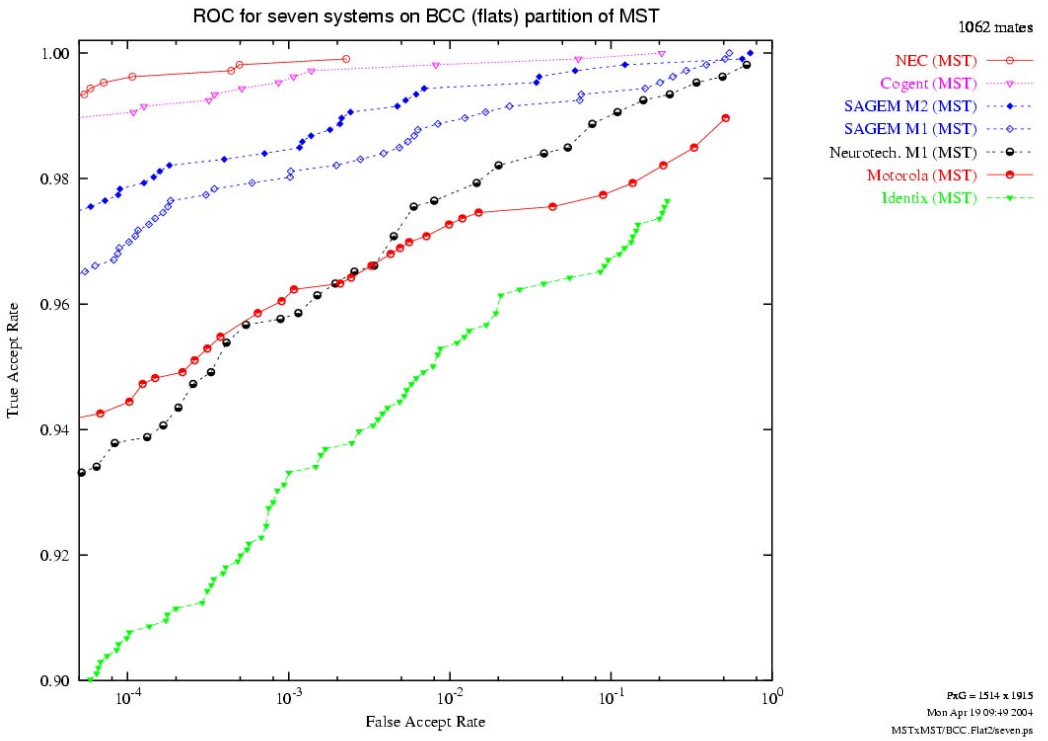

Figure E-1. Example ROCs on BCC data

Figure E-2 shows a "slice" chart that includes data from Figure E-1. This data is derived from the ROCs by interpolating TAR at FAR = 10^{-4}. For instance, in Figure E-2, the data point summarizing Neurotechnologija's accuracy on the BCC test partition (green, open circle) is just less than 0.94; this value is computed as the log-linear interpolation (dashed line) between the Neurotechnologija's third and fourth data points from the left in Figure E-1.

The summary statistics in Section 4 are computed over several of these "slice" points. For instance, data for four of the lines shown in Figure 7 of the main report can be read off Figure E-2. One of the lines in Figure 7, the "Standard Partition" (i.e., complete test), amounts to a weighted average of the data shown in Figure E-2, i.e., whereas an average would weight each data source equally, the Standard Partition reflects the actual mix of data in the original test (primarily Ident-Iafis and Ohio).

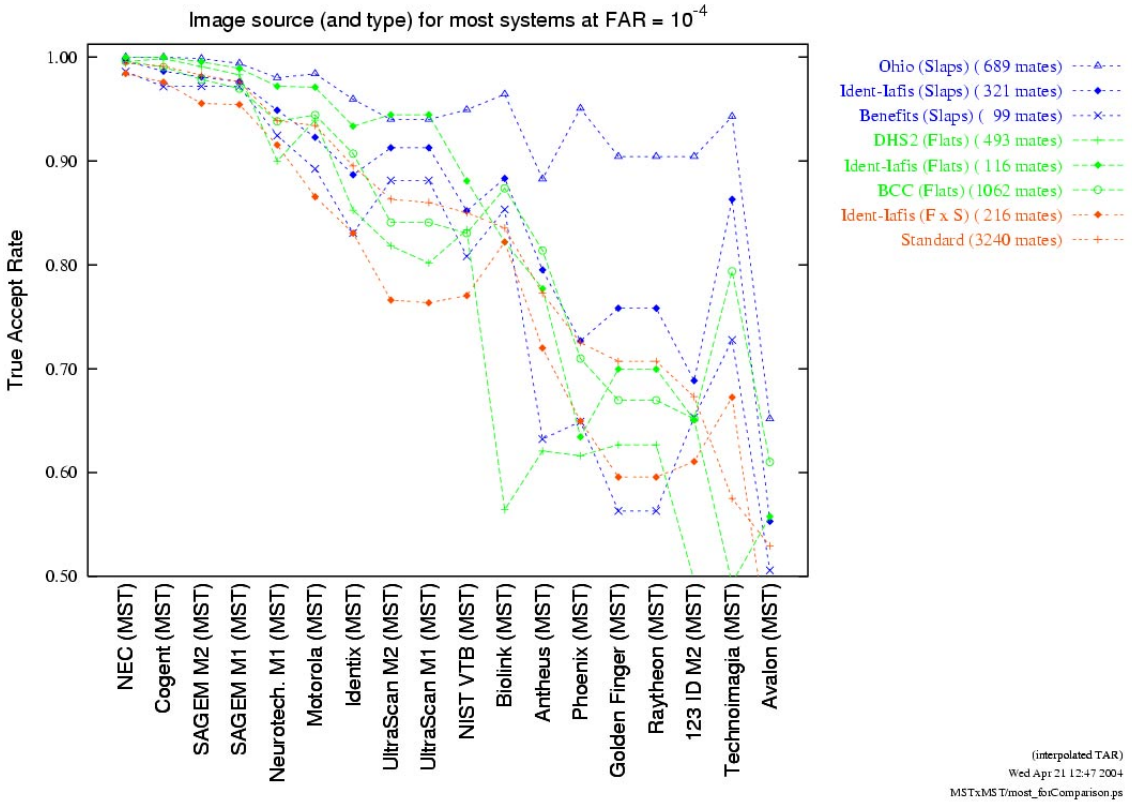

Figure E-2. Slice chart including data corresponding to Figure E-1

Which system is "best" depends on the requirements for a specific application, and must consider factors other than accuracy. The comparisons presented in Section 4 of the Analysis Report were not highly sensitive to the subjective decisions discussed in this section. Rank order among the more accurate systems was more stable relative to these decisions than rank order among the less accurate systems.

3 Summary of Analysis Procedures

The 34 systems that were evaluated in FpVTE compared more than 390 thousand fingerprint image pairs, and returned a total of 424 similarity matrices, containing a total of more than half a trillion similarity scores. The fingerprints were not of a single type or from a single source, and had a variety of other characteristics that needed to be analyzed. The FpVTE analysis team designed and developed the analysis procedures and software to process all of this data. The tasks of the analysis effort included the following:

- Determine the measured accuracy of each system
- Define methods to compare the performance of systems
- Determine and isolate the effect of four major interrelated variables and a number of minor variables on matcher performance
- Identify and correct fingerprint pairs whose *a priori* mating information might be in error, or that contained sequence errors

Satisfying each goal required automation of a variety of analyses and intensive computational effort to process the similarity matrices. This first goal was essential to ensure that the accuracy measurements were as correct as possible and to provide a well-vetted set of operational data that can be used for future analyses. Processing of similarity matrices yielded results for thousands of subtest partitions and permitted analysts to achieve the second and third goals. Systematic solutions to all these challenges were required in order to minimize the chance of error. A database was developed to keep track of information about the fingerprints used in the analysis, such as source, type, finger number, sex and age if known, mating information, and subtests which used the fingerprint.

FpVTE 2003 tools include analysis software and an SQL experiment management database. The analysis software was developed to provide repeatable, consistent results, to fully automate analysis of the similarity matrices associated with the identified partitions, and to analyze the impact of selected parameters. The SQL experiment management database tool helped in selecting partitions and managing the groundtruthing process.

Although existing FRVT 2002 software was used or modified when practical, the scale of the iterative FpVTE 2003 analysis effort required new analysis software solutions. For example, for every complete analysis run for LST, thousands of ROCs are automatically generated due to the number of participants, subtests, partitions, and variables considered during analysis. The software to score similarity matrices was not new with FpVTE 2003; however, software to fully automate the entire scoring process was developed for FpVTE 2003. For each partition test, the analysis software automatically generates numerous data files for every participant system including ROCs (and DETs), information on scores for mated pairs and false accepts, statistics, etc. ROC charts are available for different ranges of TAR beginning at 0, 0.50, 0.90, or 0.95, and the DET charts are also available. Additionally, the analysis software output includes "slices" at different FAR values such as this report's depictions of partition test results for each participant system at the specific FAR value of 10^{-4}. FRVT 2002 software was used to generate CMCs and open-set identification (watch list) ROCs.

Iterative groundtruthing, a resource-intensive process (both machine and human), required careful planning to minimize project impact and was on-going during the FpVTE 2003 analysis effort. A relational database was developed to track groundtruthing information

such as which prints had been examined, the indications of which prints were mated or not mated, and the subtests in which those prints occurred.

Analysis was necessarily iterative: often, preliminary findings led to the definition of more controlled analyses that could better measure perceived effects with a newly defined partition (e.g., to reduce the effect of a confounding variable). Some preliminary findings were misleading or failed to discern an effect through statistical noise. Subsequent analyses were needed to identify and control influential variables to more accurately observe effects. For example, MST analysis required more than 70 dataset partitions to support analyses. A pair of dataset partitions was selected to define the probe and gallery for each experiment. A relational database associating fingerprint images with metadata and test structure was developed to generate the partitions.

Considerable effort was required to summarize and present key findings without overly distorting an inherently complex set of relations among the variables.