ELFT-EFS
NIST Evaluation of Latent Fingerprint Technologies: Extended Feature Sets
Evaluation #1

Preliminary Report
DRAFT FOR COMMENT

Michael Indovina, NIST

Austin Hicklin, Noblis

26 January 2010

NOTE: This is a preliminary report. Most, but not all of the data processing for ELFT-EFS Evaluation #1 has been completed and is reported in this document. Some of the data processing, most notably the multi-encounter galleries, will be completed and included in the final report. This document provides an opportunity to present the results available to date without awaiting additional processing and the final report.

The following changes are expected between now and the final report:

- Additional "GroundTruth" data will be run against the single-encounter gallery (see notes in Section 3.1.2 and Section 6.1).
- Data will be run against the multi-encounter galleries E2-E7 (see Section 3.2).
- For participant D, subset LE results had not been completed at the time this draft was written.
- A latent examiner review of the data is in process, particularly focusing on those latents that no participants successfully matched. This review will validate the association of identity between latents and exemplars, correcting that association if necessary. In addition, this review will provide insight into the possible reasons for failures to match.
- Additional analysis will be conducted on the results included here.
- The final report will include more extensive introductory/explanatory material.
- Because the results are not yet final, conclusions are not included in this draft; preliminary observations are included throughout the document.

Please send comments to latent-efs@nist.gov by 21 February 2010.

Contents

1 Introduction

The NIST Evaluation of Latent Fingerprint Technology — Extended Feature Sets (ELFT-EFS) is an independently administered technology evaluation of latent fingerprint feature-based matching systems. ELFT-EFS is being conducted by the National Institute of Standards & Technology (NIST).

ELFT-EFS is part of NIST's Evaluation of Latent Fingerprint Technology (ELFT) testing program. The ELFT evaluations to date, notably ELFT Phase II [3], have focused solely on automated feature extraction and matching (AFEM). ELFT-EFS is an evaluation of the accuracy of latent matching using features marked by experienced human latent fingerprint examiners. The purpose of this test is to evaluate the current state of the art in latent feature-based matching, by comparing the accuracy of searches using images alone with searches using different feature sets. The feature sets will include different subsets of the Extended Feature Set (EFS) features [1] proposed by CDEFFS[*]. A key result of the test is to determine when human feature markup is effective. Because human markup is expensive in terms of time, effort, and expertise, there is a need to know when image-only searching is adequate, and when the additional effort of marking minutiae and extended features is appropriate.

ELFT-EFS Public Challenge

> The ELFT-EFS Public Challenge was a practice evaluation: an open-book test on public data to validate formats and protocols. The results of the Public Challenge are included as Appendix B. Note that the ELFT-EFS Public Challenge was an evaluation of self-reported results from a small public dataset. The systems used and timing were not constrained. These results are appropriate for preliminary analysis, but are ***not*** appropriate for rigorous analysis or comparison: the ELFT-EFS Evaluation #1 is intended for those purposes. The participants in this evaluation are and will remain anonymous.

ELFT-EFS Evaluation #1

> The ELFT-EFS Evaluation #1 was conducted using participants' software on NIST hardware at NIST facilities. Datasets were from multiple sequestered sources, each broadly representative of casework. The ELFT-EFS evaluation #1 was run specifically to identify any near-term benefits, NOT to identify long-term feasibility/accuracy. Timing constraints, subtests, and analysis were based in part on the results and lessons learned from the ELFT-EFS Public Challenge.

Subsequent Evaluations

> Subsequent ELFT-EFS Evaluations will be conducted to identify long-term feasibility and respond to lessons learned.

A detailed description of the evaluation may be found in the Test Plan, included as Appendix A.

2 Participants

The participating organizations were

- A: SAGEM
- B: NEC
- C: Cogent
- D: Sonda
- E: Warwick

[*] *CDEFFS is the Committee to Define an Extended Fingerprint Feature Set. The current draft of the Extended Friction Ridge Features specification can be found at http://fingerprint.nist.gov/standard/cdeffs/.*

3 Evaluation Data

3.1 Latent Data

3.1.1 Sources of latent images

The latent images came from both operational and laboratory collected sources, as shown in Table 1. In none of the cases were the mates selected through the use of automated fingerprint matchers (known as AFIS bias), as was true in the ELFT Phase 2 evaluation.

Table 1: Sources of latent images in baseline dataset

Name	# Latents	Description
Casework 1	372	Operational casework images
Casework 2	165	Operational casework images
WVU	446	Laboratory collected images
FLDS	93	Laboratory collected images
MLDS	38	Laboratory collected images (small set of publicly releasable images for examples in reports)
Total	1114	

3.1.2 Features

Files containing latents had features defined, except in the case of image-only searches. Latent features were formatted in accordance with "Data Format for the Interchange of Extended Friction Ridge Features," [1] abbreviated here as the "EFS Spec" (Extended Feature Set Specification). The test evaluated different combinations of EFS fields, so not all EFS fields were present in any given search. The subsets of features used (defined as Subsets LA-LG) are defined in Table 2. The specific EFS fields included in each subset are listed in Appendix A (ELFT-EFS Test Plan), Section 7.

Table 2: Latent feature subsets

Subset	Description	Image
LA	Image only	With Image
LB	ROI	With Image
LC	ROI, Pattern Class, Quality Map	With Image
LD	Minutiae with ridge counts	With Image
LE	Extended features (no Skeleton)	With Image
LF	Extended features with Skeleton	With Image
LG	Minutiae with ridge counts	Without Image

Skeletons (subset LF) could not be marked for all images: a subset of the Baseline dataset labeled "Baseline-QA" includes skeleton markup.

All of the latent IAFIS/EFS features were provided with feature markup by human expert latent fingerprint examiners; all examiners are International Association for Identification Certified Latent Print Examiners (IAI CLPE). Note that human markup was conducted outside of ELFT-EFS and was not part of the evaluation.

The examiners based their markup on markup guidelines defined in [2]. EFS markup in subsets LE and LF included features such as dots, incipient ridges, ridge edge protrusions, and pores. Features were marked in latent images without reference to exemplars, with the sole exception of the GroundTruth (GT) dataset discussed in section 7.

Note that conformance testing of automatic extraction of CDEFFS features was not part of this test. In other words, the evaluation did not measure how close automatically extracted features were to examiner created features.

Automated algorithms can use the extended features defined for a latent search without explicitly computing them for the exemplar image, and thus it must be emphasized that automated extraction of the extended features on the exemplar is not necessarily the only, nor the best way, to use this information. For example, an examiner may mark an area as a scar; for the exemplar, the matcher would not necessarily have to mark the area as a scar, but may use that information to match against a corresponding area with many false minutiae and poor ridge flow.

It should be further noted that no vendor specific rules for feature encoding were used. All encoding was made in compliance with the CDEFFS standard.

3.1.3 Assessment of latents by value and minutiae count

The examiners who marked the data made value determinations at the time of markup, using the categories defined in EFS [1, Field 9.353, Examiner value assessment]. Table 3 shows the proportion of each value determination in the Baseline dataset. Note that 2% of the latents in the evaluation were marked as No value, and an additional 11% were marked as Limited value. The selected latents were not screened to exclude data based on the examiners assessment of value. Such a selection, which is often based on latent examiner experience with older AFIS technology, can bias the test in that it eliminates data that can and should be searched by an AFIS.

Table 3: Examiner value determinations

Latent value determination	Description	% of Baseline latents[*]
Value	The impression is of value and is appropriate for further analysis and potential comparison. Sufficient details exist to render an individualization and/or exclusion decision.	86.7%
Limited	The impression is of limited, marginal, value and may be appropriate for exclusion only.	11.1%
No value	The impression is of no value, is not appropriate for further analysis, and has no use for potential comparison.	2.3%

Table 4 shows the relation between minutiae count and value determination.

Table 4: Minutiae count statistics by value determination

	Value	Limited	No Value
Mean	25	8	4
StDev	16	4	3
Min	4	1	1
Q1	14	5	2
Median	20	7	3
Q3	31	9	5
Max	106	20	12

The relation between value determination and accuracy is reported in Section 10.

[*] *Draft note: 11 latents (out of 1114) did not have value determinations; all of those are expected to be found of value.*

Table 5 shows the minutiae distribution for the overall Baseline dataset, as well as for each of the data sources. These are the minutiae counts as marked by examiners.

Table 5: Minutiae count distribution by data source

	All (Baseline)	Casework 1	Casework 2	WVU	FLDS	MLDS
Count	1114	372	165	446	93	38
Mean	22	20	18	27	20	18
StDev	16	12	9	20	17	15
Min	1	3	5	1	1	4
Q1	12	12	11	13	8	9
Median	18	18	16	22	16	14
Q3	28	25	23	36	28	19
Max	106	79	48	106	85	69

Table 6: Minutiae count distribution by data source

	All	Casework 1	Casework 2	WVU	FLDS	MLDS
Count	1114	372	165	446	93	38
1-5	6%	5%	1%	5%	19%	8%
6-10	16%	15%	23%	13%	15%	24%
11-15	20%	21%	24%	18%	13%	26%
16-20	16%	19%	22%	10%	15%	24%
21-25	12%	17%	9%	11%	10%	3%
26-30	9%	10%	10%	9%	8%	0%
31-35	5%	4%	5%	7%	3%	3%
36-40	4%	3%	1%	6%	4%	5%
41-45	3%	2%	2%	5%	3%	3%
46+	33%	11%	2%	63%	29%	26%

The relation between minutiae count and accuracy is reported in Section 9.

3.1.4 Orientation

Latent fingerprint images varied in orientation from upright ±180°. Table 7 shows the distribution of the Baseline latents by orientation, as determined by latent examiners during markup.

Table 7: Orientation of Baseline latents. Angles are degrees from upright as labeled by latent examiners.

Orientation	% of Baseline latents
Unknown	7.0%
0-9 degrees	55.7%
10-19 degrees	12.3%
20-44 degrees	19.1%
45-89 degrees	4.9%
>90 degrees	1.0%

3.2 Exemplars

Exemplars came from optical livescan and inked paper sources. Exemplar sets always included all ten fingers. Table 8 shows the exemplar subsets used in the evaluation.

DRAFT NOTE: for this draft, the only exemplar subset used is E1; the other exemplar subsets will be included in the final report.

Table 8: Exemplar subsets

Exemplar subset	# subjects	Description
E1	100,000	10 rolled & 10 plain impressions each
E2	10,000	10 rolled impressions each
E3	10,000	10 plain impressions each
E4	10,000	2 sets of 10 rolled+plain impressions each
E5	10,000	3 sets of 10 rolled+plain impressions each
E6	10,000	4 sets of 10 rolled+plain impressions each
E7	10,000	5 sets of 10 rolled+plain impressions each

Note that an exemplar set may include rolls alone, plains alone, or both. Plain impressions were segmented from slap images. For the non-mated data, the slap segmentation was performed automatically; for the exemplars mated to the latent probes, human review was conducted to verify the accuracy of segmentation.

For cases in which multiple sets of exemplars associated with one person were included in the gallery, the association was made explicit in the exemplar enrollment stage: at the time of enrollment, exemplars that are known to belong to the same person will always share the same subject ID.

Files containing exemplars did not have any features defined.

3.3 Data Format

All images and data were contained in ANSI/NIST files.

All images were 8-bit grayscale. All latent images were 1000 pixels per inch, uncompressed. Exemplar images were 500 pixels per inch, compressed using WSQ.

Latent fingerprint images varied from 0.3”x 0.3” to 2.0” x 2.0” (width x height).

Exemplar images were approximately upright (retained in the same orientation as they were captured).

Exemplars were provided in complete 10-finger sets, with finger positions noted. The finger positions for latents were not noted – no searches were restricted to specific fingers.

4 Test Procedure

4.1 Latent Matching Software

Each participant submitted a set of SDKs (Software Development Kits) that provided the interfaces defined by the ELFT-EFS-1 API specified in Appendix A. The SDKs were provided as static or dynamic libraries to run on the NIST platform specified in Section 4.2. The ELFT-EFS API (Application Programmer Interface) was modeled after the API from ELFT Phase 2. The most notable differences from the ELFT Phase 2 API were that the exemplar and latent images and data provided to the SDK were contained in ANSI/NIST files; exemplar feature extraction processed a single exemplar per invocation (instead of the complete gallery); and the ELFT-EFS-1 API specified operational time limits on a per-processor core basis, rather than per-machine.

Each participant submitted

- one SDK for exemplar feature extraction and exemplar enrollment
- one SDK for latent feature extraction
- one SDK for latent 1-to-N search

SDKs were permitted to be sequential or multithreaded, and utilize either 32 or 64-bit execution mode.

4.2 Test Platform

The NIST ELFT-EFS Evaluation test platform consisted of an array of blade servers with the following hardware configuration:

Processor

- Dual 2.8 GHz/1MB Cache, Xeon (dual-core)
- 800 MHz Front Side Bus for PE 1855

Memory

- 16GB RAM (15GB available to 64-bit applications; 3GB available to 32-bit applications)

Secondary storage

- 300GB 15K RPM Ultra SCSI Hard drives

Operating systems

- RedHat Linux 3.1 64-bit
- Windows 2008 Server (64-bit or 32 bit)

Each SDK was allocated multiple blades/cores from the array, along with a subset of the test data in order to maximize (time) efficiency through parallel operation. Each SDK instance assigned to an individual blade or core operated on a subset of the data, using individual data copies (as needed) from a local storage device.

4.3 Format of results

All searches returned a candidate list. A candidate list has a fixed length of one hundred (100) candidates. Note that a given search may be associated with zero, one, or more subjects in the gallery, and the candidate list shall include all of them.

The candidate list consists of two parts, a required and an optional part.

The required part consists of:

- the index of the mating exemplar subject
- the matching finger number
- the absolute matching score
- an estimate of the probability of a match (0 to 100)

The optional part consists of:

- the number of good minutiae identified in the latent
- the number of latent minutiae which were successfully matched
- the quality estimate of the latent (0 to 100, 100 is best)
- the quality estimate of the candidate (0 to 100, 100 is best)

Draft note: analysis of the optional results has not been performed.

4.4 Timing

The ELFT-EFS Evaluation placed limits on the processing time of the major operations involving feature extraction and enrollment (exemplars and latents) and searching. The primary purpose is to measure performance at throughput rates comparable to large-scale operational scenarios; it is well understood that matching accuracy is typically inversely proportional to the time permitted.

The search time requirements specified below are for Subtests LC-LG (see Table 2). It is recognized that for some implementations, throughput for image-only searches (Subtest LA) may be slower. Therefore, it is allowable for throughput on Subtest LA (image only) and LB (image+ROI) to be slower by a factor of up to 2x than the stated nominally required search time.

Table 9: Timing requirements, per single CPU core

Exemplar feature extraction	100 sec/10-finger exemplar set (rolled or pre-segmented slap)
Latent enroll	120 sec/latent
Search	0.05 sec/exemplar set (20 exemplar sets/sec, per latent, assuming an exemplar set consists of 10 rolled and 10 segmented slap fingerprints)

5 Rank-based results

Overall accuracy results are presented in this section using rank-based metrics via Cumulative Match Characteristic (CMC) curves. A CMC curve shows how many latent images are correctly identified at rank 1, rank 2, etc. A CMC is a plot of identification rate (or hit rate) vs. recognition rank. Identification rate at rank *k* is the proportion of the latent images correctly identified at rank K or lower. A latent image has rank *k* if its mate is the kth largest comparison score on the candidate list. Recognition rank ranges from 1 to 100, as 100 was the (maximum) candidate list size specified in the API.

The following tables summarize the rank-1 hit rates for each of the participants for each of the latent subsets. Table 10 shows the results for the Baseline-QA dataset, the subset of the Baseline dataset for which skeletons were available. The Baseline-QA dataset is the portion of the Baseline dataset that had skeletons (LF) marked; for that reason, the Baseline-QA dataset is used to compare all of the latent subsets.

Draft note: For participant D, subset LE results had not been completed at the time this draft was written, but will be included in the final report.

Table 10: Summary of rank-1 hit rates for the Baseline-QA dataset (458 latents, subset of Baseline)

	Latent Subset						
	LA	**LB**	**LC**	**LD**	**LE**	**LF**	**LG**
	Image only	Image + ROI	Image + ROI + Pattern Class + Qual map	Image + Minutiae + Ridge Counts	Image + EFS	Image + EFS + Skeleton	Minutiae + Ridge Counts only
A	59.4	58.3	58.3	62.9	62.9	62.2	40.6
B	56.3	57.2	57.4	59.0	59.0	60.5	44.8
C	40.6	41.5	43.2	57.6	59.0	60.0	43.9
D	22.3	*n/a*[*]	*n/a**	14.0	*TBD*	14.4	10.9
E	42.6	44.3	46.5	44.5	47.2	31.2	24.2

Table 11 shows the rank-1 results for the complete Baseline dataset. Skeletons (LF) were not marked for the complete dataset. Due to limited available processing time only the subsets LA, LE, and LG were run for the complete Baseline dataset: LB/LC were omitted because they showed limited improvement over LA, and LD was omitted because the performance of LD and LE were so similar.

Table 11: Summary of rank-1 hit rates for the Baseline dataset (1114 latents)

	Latent Subset		
	LA	**LE**	**LG**
	Image only	Image + EFS	Minutiae + Ridge Counts only
A	62.2	66.7	44.0
B	61.2	63.3	48.3
C	48.3	62.0	47.8
D	25.1	*TBD*	4.5
E	47.2	50.3	29.4

Preliminary observations (Table 10 and Table 11):

- The performance for the complete Baseline dataset is better than for Baseline-QA by about 3-5% in most cases; this was an artifact of the process by which Baseline-QA was selected.
- Region of interest (latent subset LB) provided limited improvement in accuracy over image only (LA) for participants B/C/E, and were slightly counterproductive for participant A.
- Pattern class and quality map (latent subset LC) provided limited improvement in accuracy over image + ROI (LB) for participants B/C/E, but made no difference for participant A.
- Extended feature set (latent subset LE) provided slight improvement in accuracy over image + minutiae (LD) for participants C/E, but made no difference for participants A/B.
- Skeleton (latent subset LF) provided limited improvement in accuracy over image + EFS for participants B/C, was slightly counterproductive for participant A, and quite counterproductive for participant E.
- Image only (LA) was far more accurate for participants A/B/E than minutiae only (LG); this was not true for participant C.
- Adding minutiae or extended features (LD/LE) to the image (LA) provided limited but consistent improvement for participants A/B/E, but a substantial improvement to participant C.

[*] *Participant D informed NIST that their software did not utilize the features in subsets LB/LC, and therefore those subsets were not run.*

- Note that the overall accuracy is limited by the proportion of poor-quality latents, and that as discussed in Section 3.1.3, 2-14% of the Baseline dataset was marked as limited or no value. Overall accuracy would differ given a different distribution of poor-quality latents.

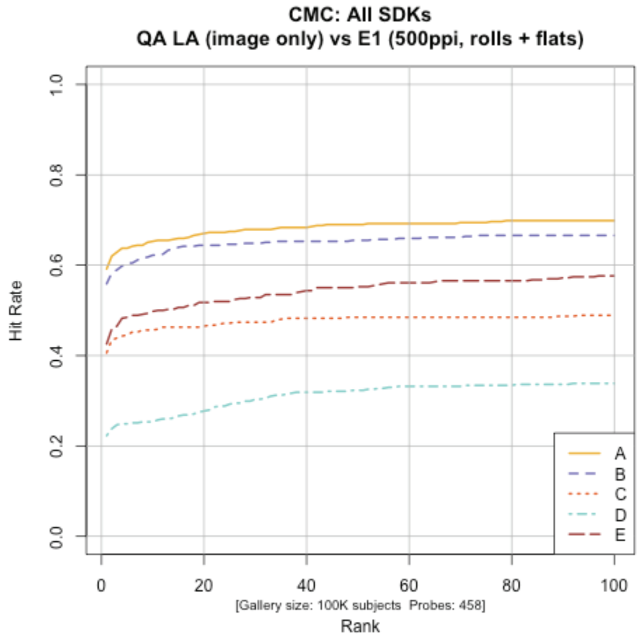
Draft note: the final report will include analysis of the datasets showing the effect on accuracy of including or excluding prints that examiners have labeled as limited or no value.

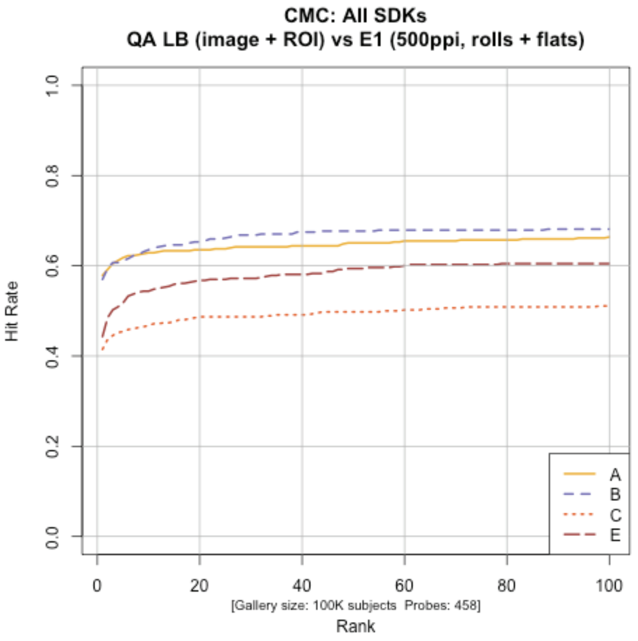
The following sections show the complete CMCs for the rank-based results, showing not just the rank-1 results included in Table 10 and Table 11, but on out to rank 100.

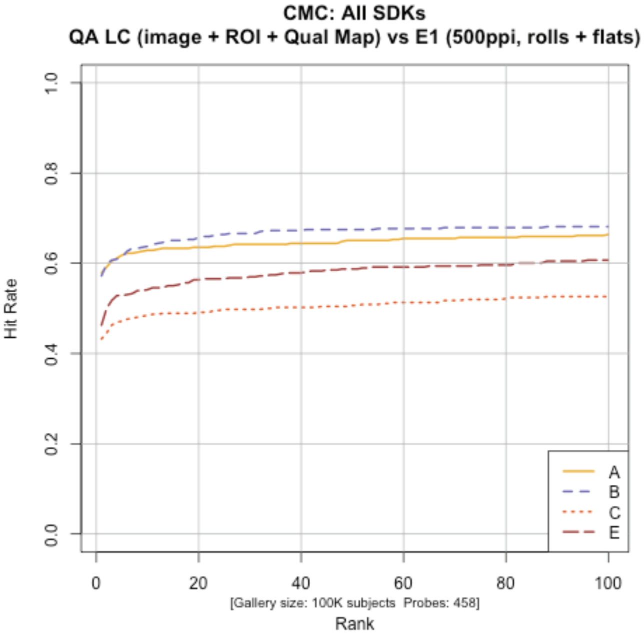
Preliminary observations (Sections 5.1-5.4):

- The CMC charts confirm the preliminary observations summarized in the rank-1 tables.
- The CMC curves for the different participants are generally quite parallel, so that the difference in performance between two matchers or subsets is nearly constant regardless of the rank.
- The CMC charts indicate substantial flattening which is consistent with maintaining a high performance rate at a greater gallery size. It should be noted the number of additional candidates in excess of 20 is very small for all of the leading vendors.
- The proportion of the total hits made by matchers that were rank 1 is an indication of scalability of performance:
  - For matchers A/B/C, 87-92% of hits are rank 1 for subset LA; 88-93% of hits are rank 1 for subset LE; 78-86% of hits are rank 1 for subset LG.
  - These results are reported in Appendix C (Additional Results).

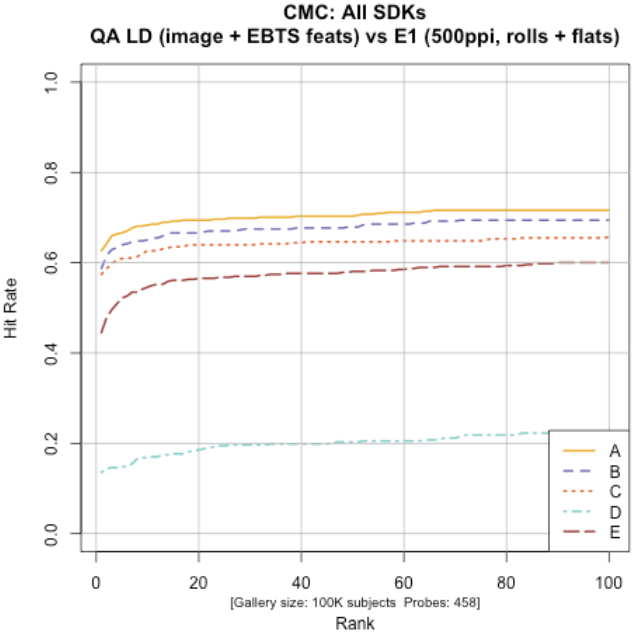
5.1 Results by latent subset, Baseline-QA dataset


CMC: All SDKs
QA LA (image only) vs E1 (500ppi, rolls + flats)

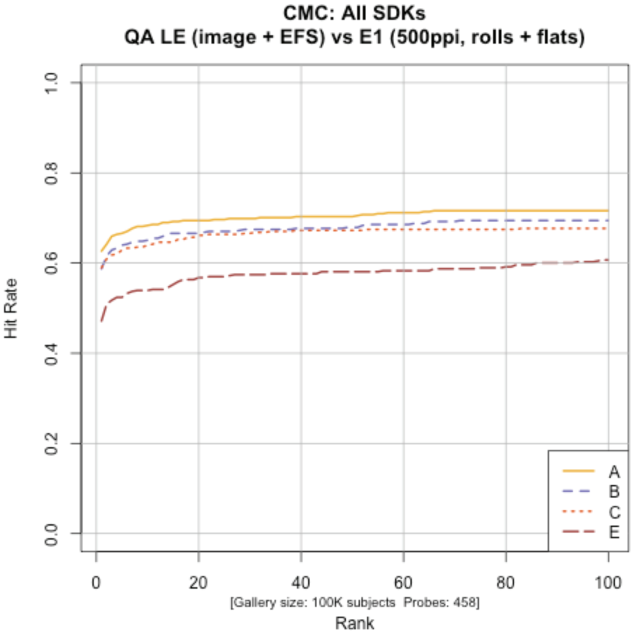

CMC: All SDKs
QA LB (image + ROI) vs E1 (500ppi, rolls + flats)



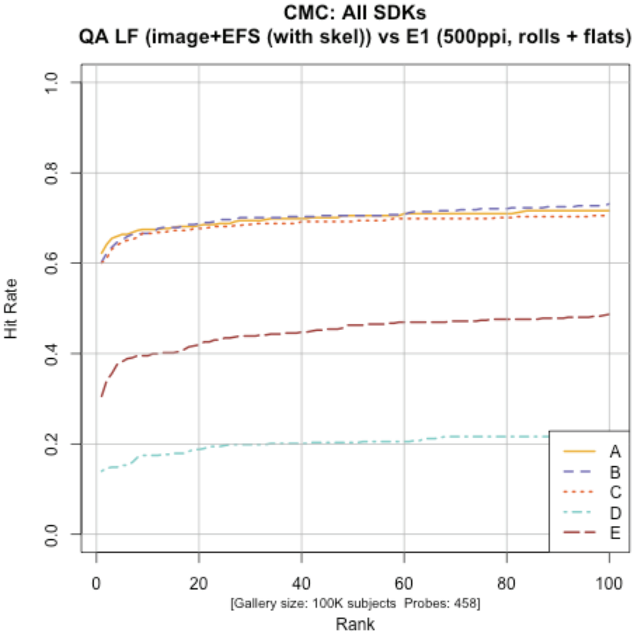
CMC: All SDKs
QA LC (image + ROI + Qual Map) vs E1 (500ppi, rolls + flats)



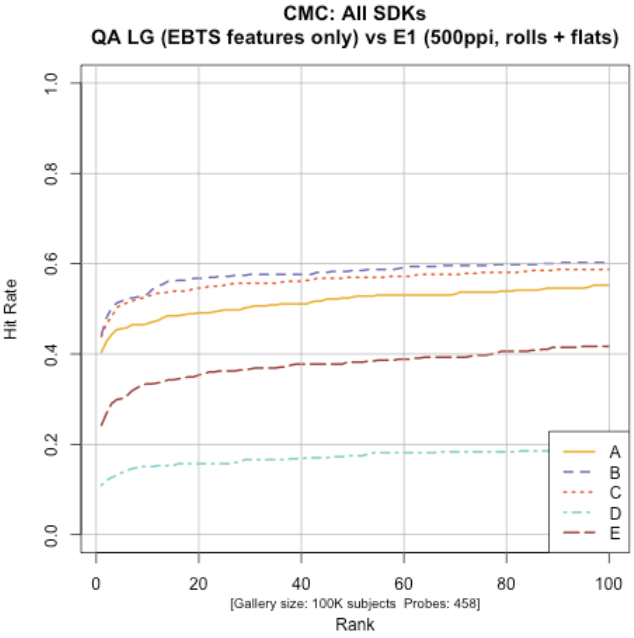
CMC: All SDKs
QA LD (image + EBTS feats) vs E1 (500ppi, rolls + flats)



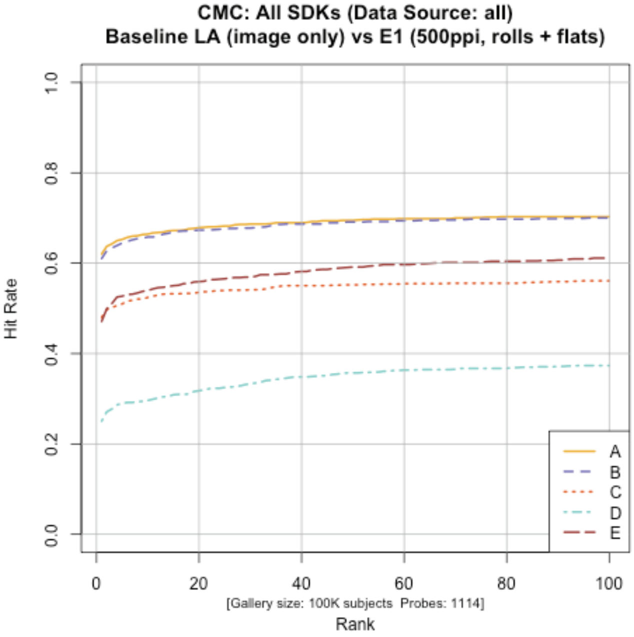
CMC: All SDKs
QA LE (image + EFS) vs E1 (500ppi, rolls + flats)

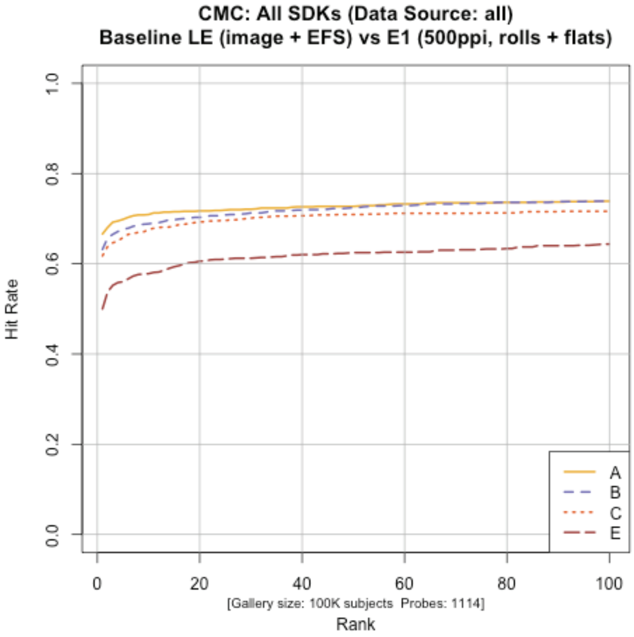


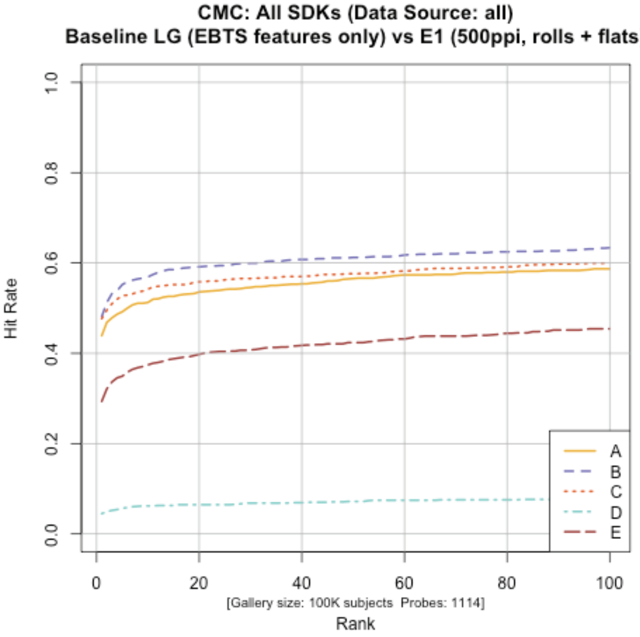
CMC: All SDKs
QA LF (image+EFS (with skel)) vs E1 (500ppi, rolls + flats)



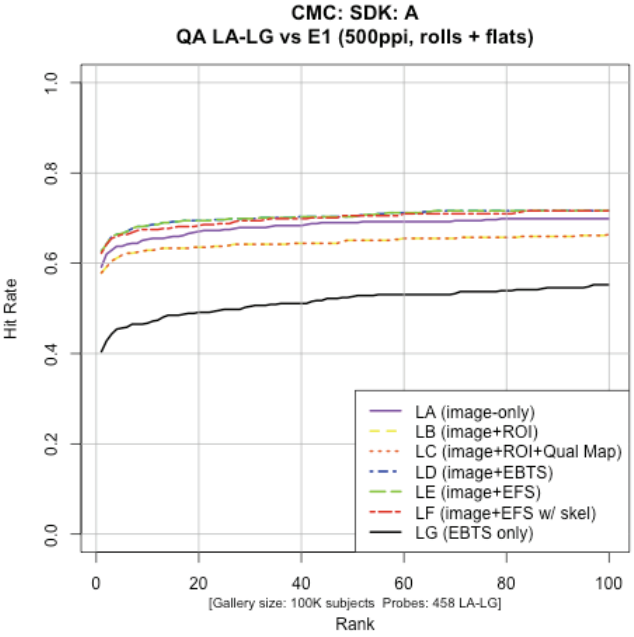
5.2 Results by latent subsets, Baseline dataset

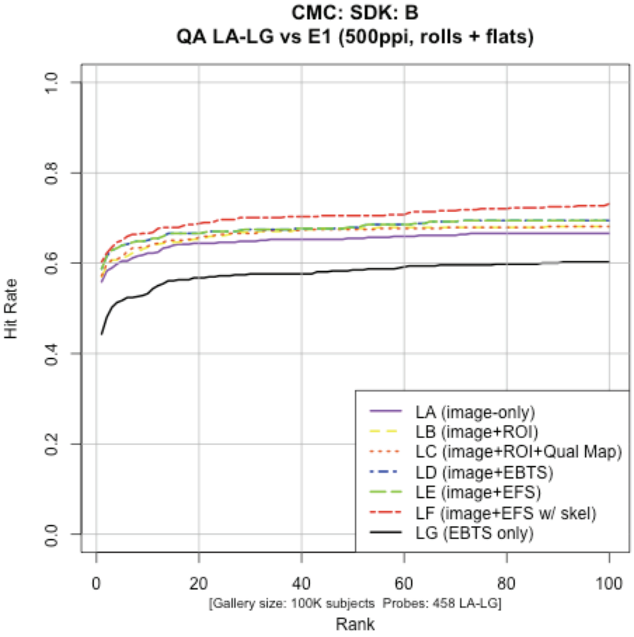


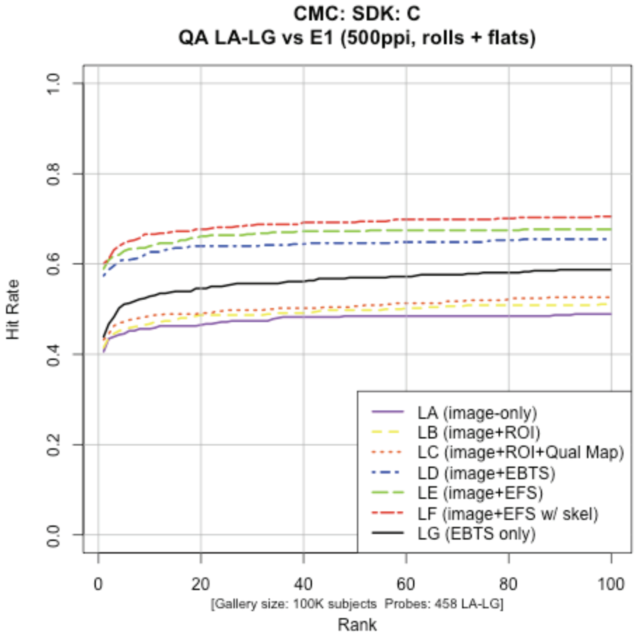


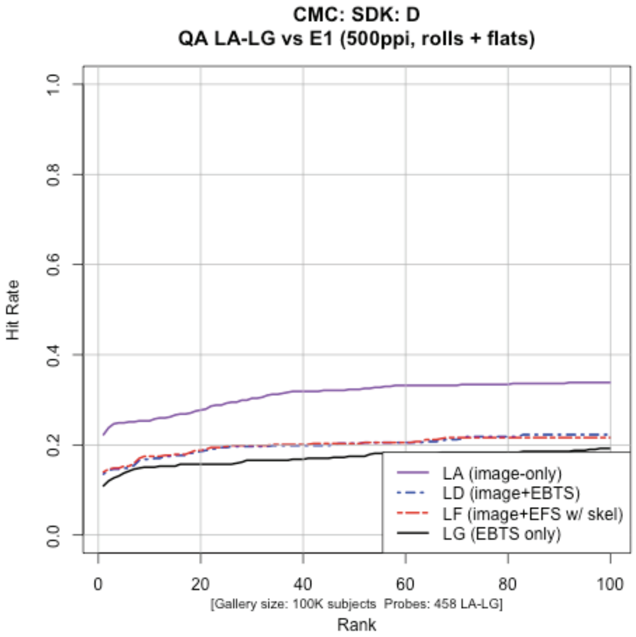


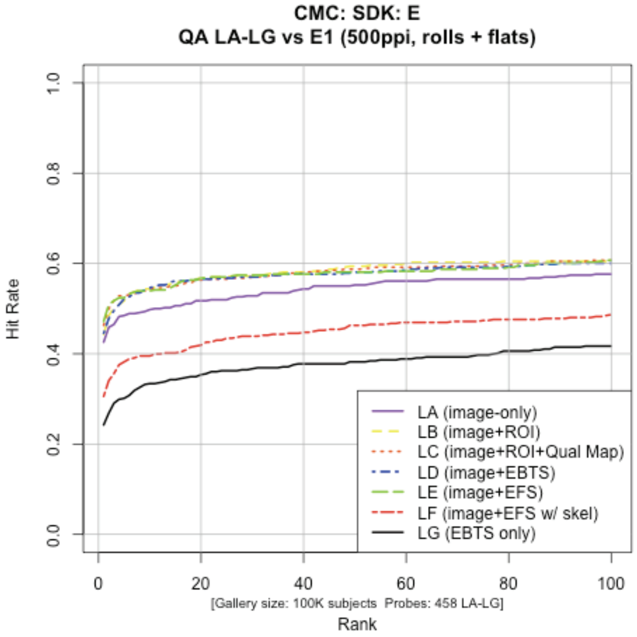
5.3 Results by participant, Baseline-QA Dataset









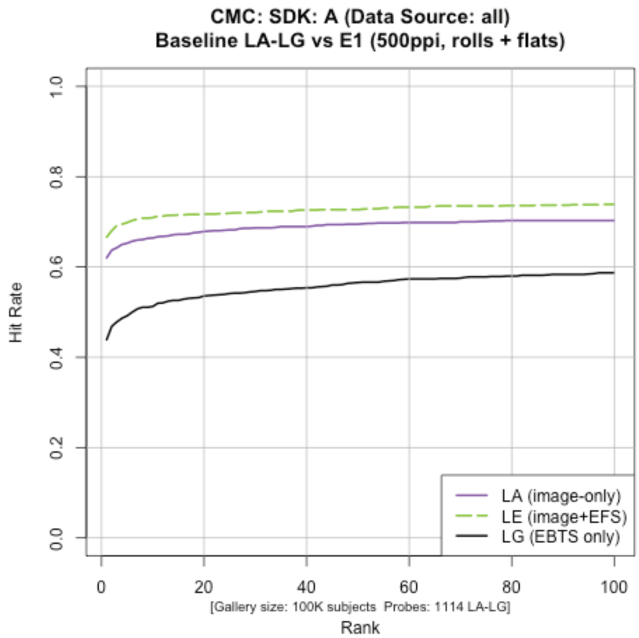


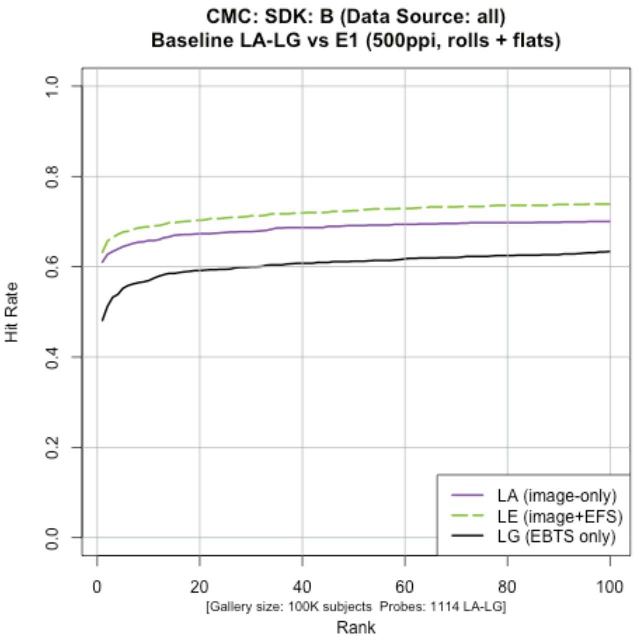
5.4 Results by participant, Baseline Dataset

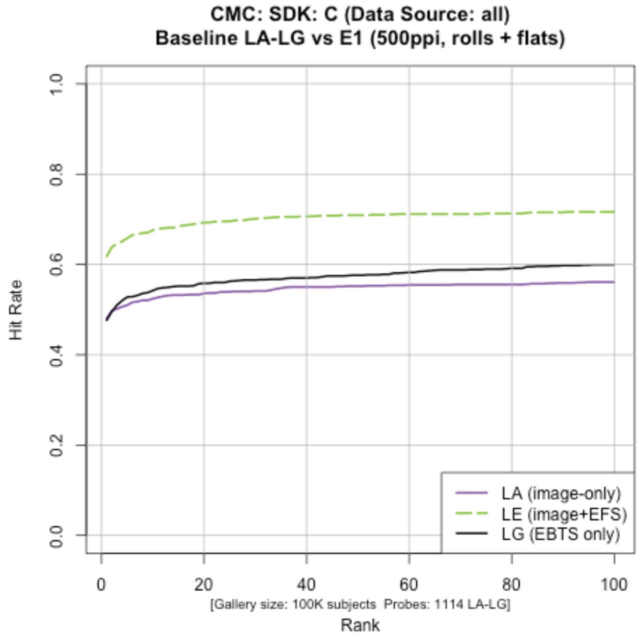
This section presents the Baseline dataset with images only (LA), images plus EFS (LE), and compares them with the legacy EBTS minutiae feature set (LG).

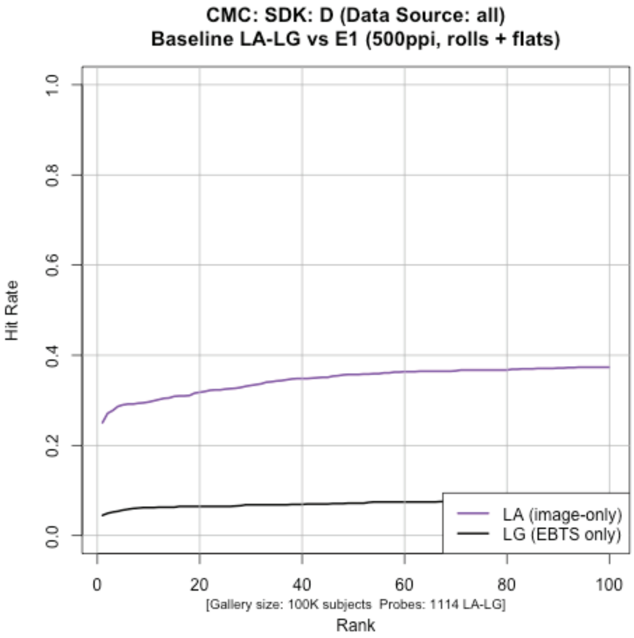
Preliminary observations:

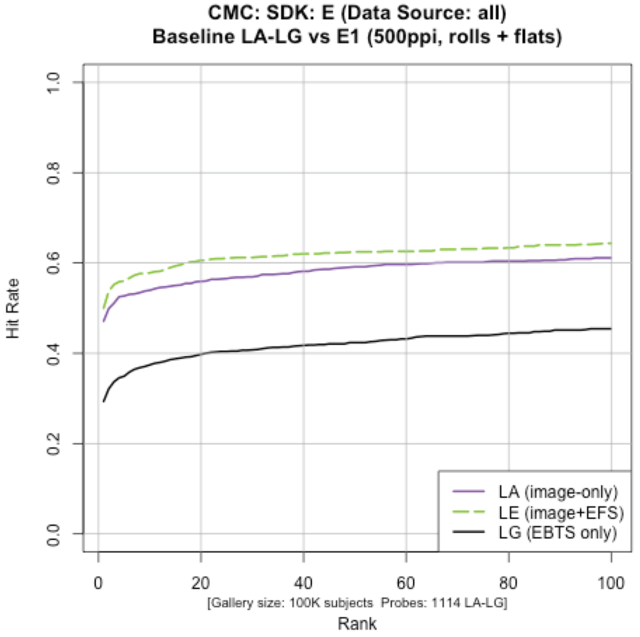
- For participants A/B/C/E, substantial improvement (10-30%) is shown for images plus EFS (LE), over the legacy EBTS minutiae feature set (LG).
- Participant C is the only one for whom the accuracy of the legacy EBTS minutiae feature set (LG) is higher than image only (LA).


CMC: SDK: A (Data Source: all)
Baseline LA-LG vs E1 (500ppi, rolls + flats)
Hit Rate
0.0
0.2
0.4
0.6
0.8
1.0
0
20
40
60
80
100
[Gallery size: 100K subjects Probes: 1114 LA-LG]
Rank
LA (image-only)
LE (image+EFS)
LG (EBTS only)


CMC: SDK: B (Data Source: all)
Baseline LA-LG vs E1 (500ppi, rolls + flats)
Hit Rate
0.0
0.2
0.4
0.6
0.8
1.0
0
20
40
60
80
100
[Gallery size: 100K subjects Probes: 1114 LA-LG]
Rank
LA (image-only)
LE (image+EFS)
LG (EBTS only)


CMC: SDK: C (Data Source: all)
Baseline LA-LG vs E1 (500ppi, rolls + flats)
Hit Rate
0.0
0.2
0.4
0.6
0.8
1.0
0
20
40
60
80
100
[Gallery size: 100K subjects Probes: 1114 LA-LG]
Rank
LA (image-only)
LE (image+EFS)
LG (EBTS only)


CMC: SDK: D (Data Source: all)
Baseline LA-LG vs E1 (500ppi, rolls + flats)
Hit Rate
0.0
0.2
0.4
0.6
0.8
1.0
0
20
40
60
80
100
[Gallery size: 100K subjects Probes: 1114 LA-LG]
Rank
LA (image-only)
LG (EBTS only)



6 Score-based results

The previous results reported rank-based identification performance. Here Receiver Operating Characteristic (ROC) curves were plotted using the methodology defined in ELFT Phase II ([3], Section 3.1.2 p 24.). All ROC curves in this analysis are limited to Rank 1 (limited to the highest scoring result in the candidate list).*

As defined for ELFT Phase II,

- The Identification Rate (IR) indicates the fraction of cases in which enrolled mates do not appear in the top position with a score greater than the threshold. (Note that the False Negative Identification Rate (FNIR = 1-IR) indicates the fraction of cases in which enrolled mates do not appear in the top position with a score greater than the threshold.)
- False Positive Identification Rate (FPIR) indicates the fraction of candidate lists (without enrolled mates) that contain a non-mate entry in the top position with a score greater than the threshold.

Note that a horizontal line is ideal, indicating no degradation in accuracy as non-mates are automatically excluded. Note also that when the FPIR=1.0, the raw score Identification Rate is the same as the rank-1 identification rate shown in the rank-based (CMC) analyses shown in Section 5.

In each case, participants returned a raw score and a normalized score estimating the probability of a match. Generally, the probability scores provided better results than the raw scores. A comparison of raw and probability scores is provided in Appendix C (Additional Results).

These results are of interest for multiple reasons:

- Score-based results are more scalable than rank-based results, providing a better indication of how accuracy would be affected by an increase in database size. The identification rate at 0.01 provides a rough projection of accuracy for an increase of 100x.

* *Note that ROC curves are used here instead of the DET curves used in ELFT Phase II. ROCs and DETs display the same information with the sole difference that ROCs display the true positive rate on the Y axis, while DETs display the inverse (the false negative/type-2 error rate is 1-true positive rate) on the Y axis, generally in log scale. DETs are effective at showing distinctions between small error rates, but are more difficult to interpret than ROCs for the accuracy levels reported here.*

- For reverse or unsolved latent matching, in which a gallery of latents is searched with newly acquired exemplars, potential candidates must be automatically screened to limit the impact on human examiners. Score-based results give an indication of the effectiveness of reverse matching.

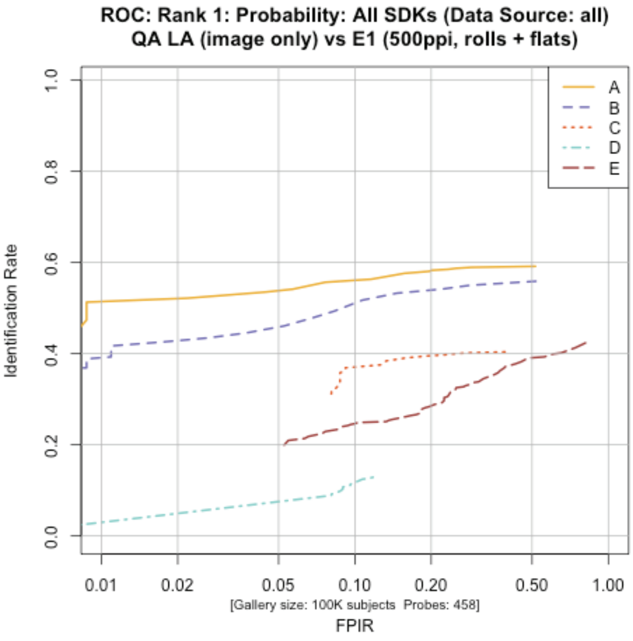
Draft note: the final report will include tables summarizing performance at different FPIR thresholds, such as 0.01, 0.1, etc.

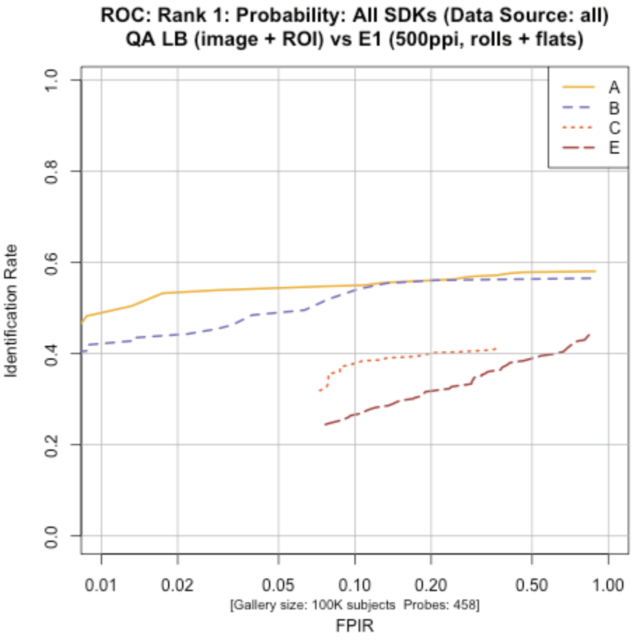
6.1 Results by latent subset, Baseline-QA dataset

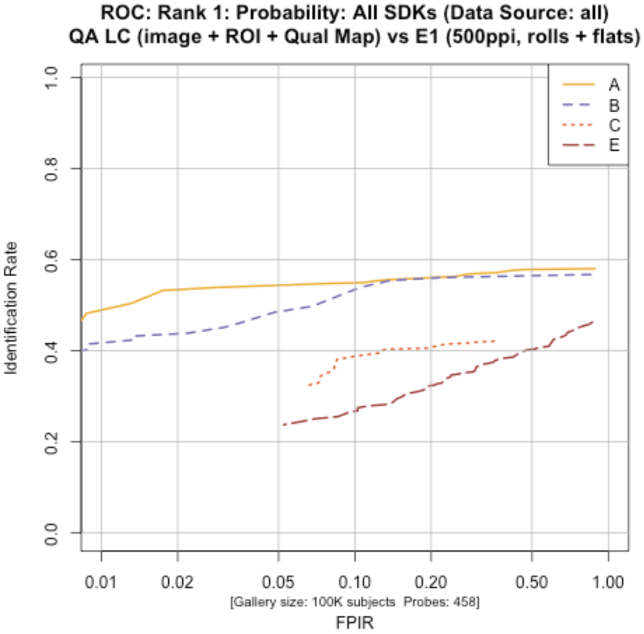
In practice, these charts show the effect of automatically eliminating candidates based on score. For example, participant A has a Rank-1 hit rate of 0.59 for latent subset LA (from Table 10). From the first chart below we see that the Identification Rate remains at 0.59 when FPIR=0.5: this means that if a score threshold were used to filter results, about half of the candidate lists that did not include a true mate could be eliminated without any impact on identification rate. If the score threshold is set to eliminate 99% of the candidate lists (FPIR=0.01), the identification rate would drop from 0.59 to 0.51, trading off a moderate drop in accuracy for a very substantial reduction in examiner effort. Note that for some matchers the curves do not extend fully across the charts; this simply means that the matcher scores did not fully populate the range of FPIR.

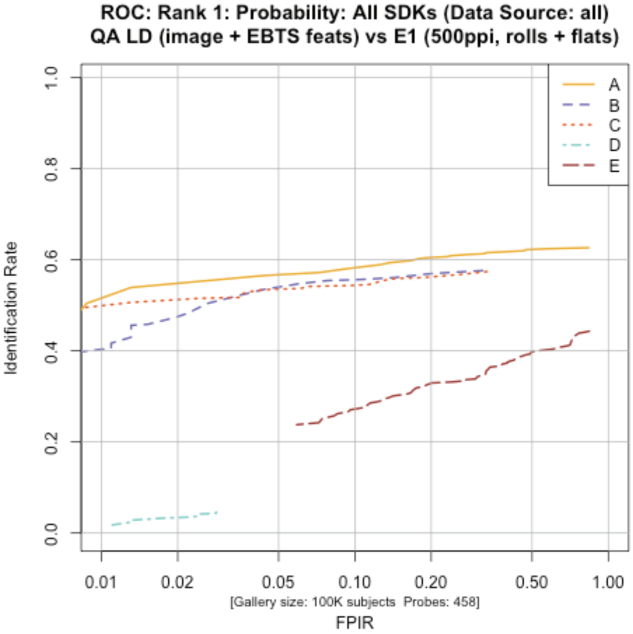
Preliminary observations:

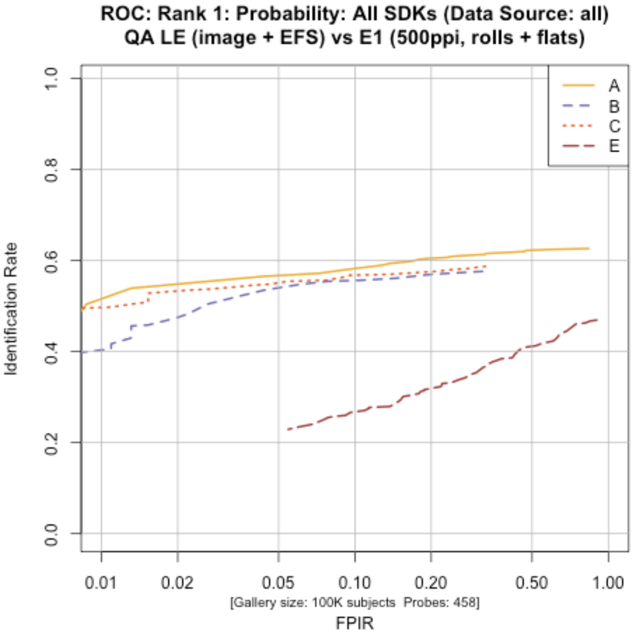
- The order of accuracy shown in the rank-based results does not necessarily correspond to these results. While the CMC curves were generally parallel, these curves often cross.
- Note that two participants can have similar performance at a high FPIR, but show substantial differences in accuracy as FPIR approaches 0.01. For example, compare participants A/B in subsets LA-LC, or participants A/B/C in subsets LD-LG.

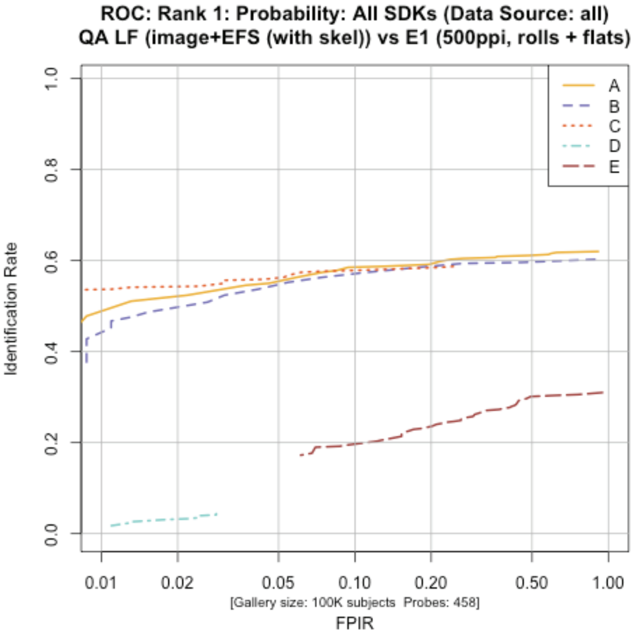


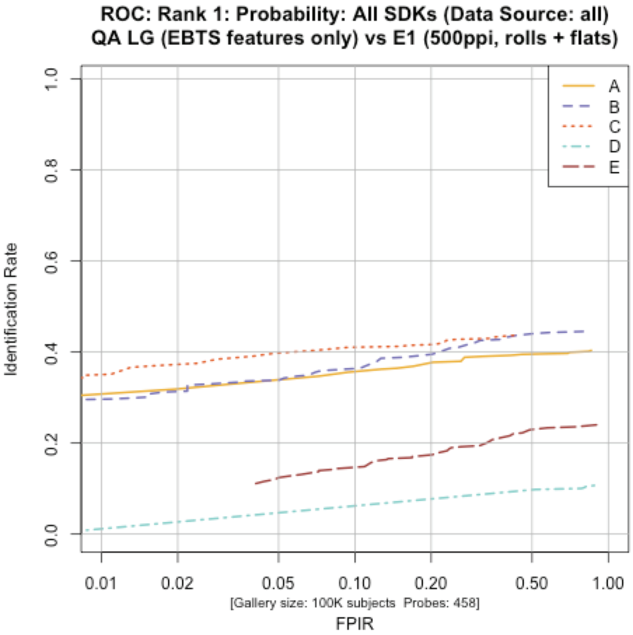



ROC: Rank 1: Probability: All SDKs (Data Source: all)
QA LC (image + ROI + Qual Map) vs E1 (500ppi, rolls + flats)

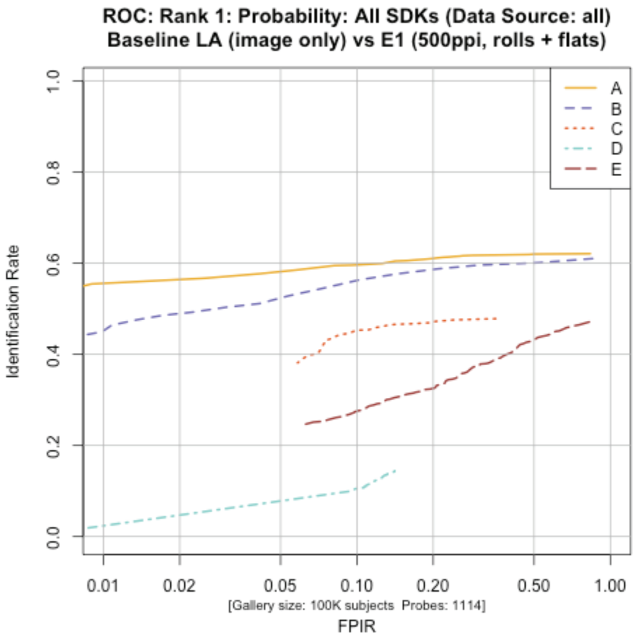

ROC: Rank 1: Probability: All SDKs (Data Source: all)
QA LD (image + EBTS feats) vs E1 (500ppi, rolls + flats)

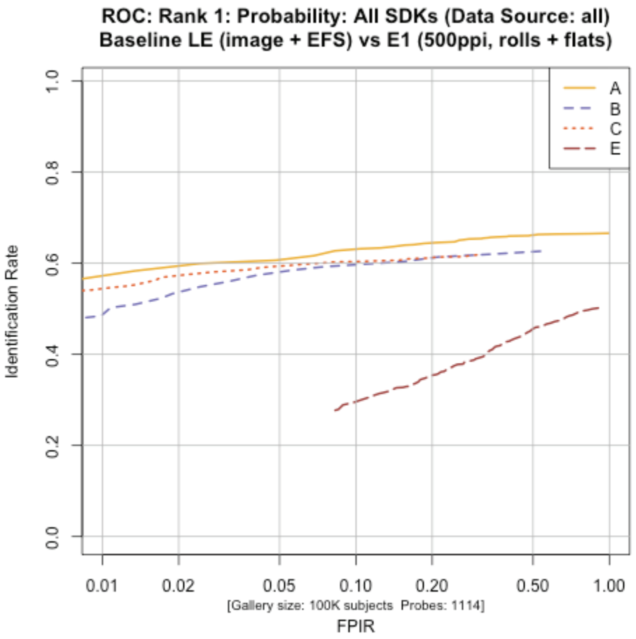

ROC: Rank 1: Probability: All SDKs (Data Source: all)
QA LE (image + EFS) vs E1 (500ppi, rolls + flats)

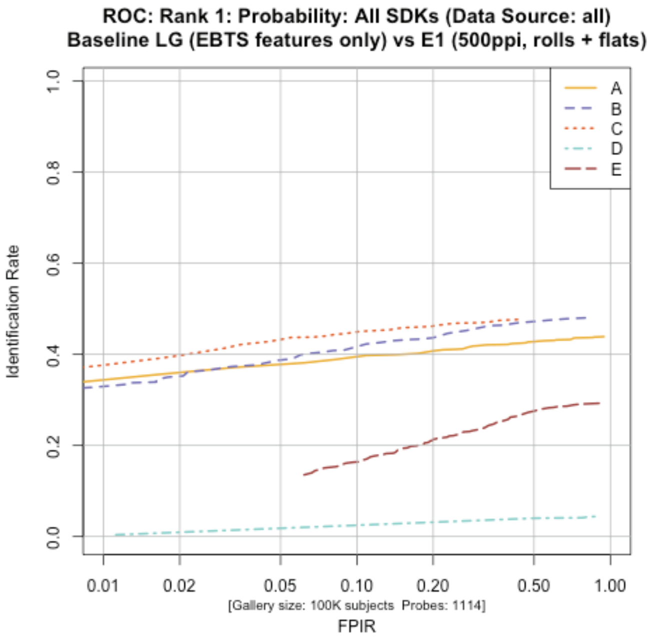

ROC: Rank 1: Probability: All SDKs (Data Source: all)
QA LF (image+EFS (with skel)) vs E1 (500ppi, rolls + flats)


ROC: Rank 1: Probability: All SDKs (Data Source: all)
QA LG (EBTS features only) vs E1 (500ppi, rolls + flats)
Identification Rate
0.0
0.2
0.4
0.6
0.8
1.0
0.01
0.02
0.05
0.10
0.20
0.50
1.00
[Gallery size: 100K subjects Probes: 458]
FPIR
A
B
C
D
E

6.2 Results by latent subset, Baseline dataset


ROC: Rank 1: Probability: All SDKs (Data Source: all)
Baseline LA (image only) vs E1 (500ppi, rolls + flats)


ROC: Rank 1: Probability: All SDKs (Data Source: all)
Baseline LE (image + EFS) vs E1 (500ppi, rolls + flats)


ROC: Rank 1: Probability: All SDKs (Data Source: all)
Baseline LG (EBTS features only) vs E1 (500ppi, rolls + flats)

7 Effect of "Ground Truth" markup

The results discussed so far in this report have been based on human examiner markup of each latent image, as might be expected in casework. This approach has the benefit of being realistic, but for the purposes of algorithmic performance evaluation there is a drawback in that the latent markup includes the variability one might expect from any human activity, including the results of differences of expertise and possible error. When evaluating matchers, one approach taken in the past was to mark only "GroundTruth" features, by referring to the exemplar(s) when marking each latent, so that the result would include no false features. This groundtruthing process obviously cannot be used operationally, but it has been used very effectively to provide test datasets that minimize human variability for the purposes of development and evaluation of fingerprint feature extraction and matching software. The dataset now know as NIST Special Database 27 (SD27) was collected in this way for the development of the FBI's IAFIS.

In every other section of this report, all latent images were marked using the latent alone. The results discussed in this section show the differences between operationally practical markup and groundtruth markup. The set of latent images in Baseline-QA were marked by the examiners using three different approaches, as defined in Table 12.

Table 12: Types of examiner markup

Markup type	Description
Baseline-QA	Features were marked in latent images without reference to exemplars.
AFIS	Derived from the Baseline-QA markup. The examiners removed debatable minutiae, to test the assumption that false minutiae are worse than missed minutiae for AFIS searching.*
GT	Ground Truth markup, derived from the Baseline-QA markup. Exemplar images (both rolled and plain) were consulted when marking latent features, so that all latent minutiae† marked were corroborated by one or more exemplars. Note that this is not possible operationally, but defines an upper bound for the accuracy of feature matching.

Draft note: Not all of the GT dataset has been run: this section includes partial results that are likely to change before the final report.

The charts below show the effect on accuracy of the different markup types, comparing them to image-only searches, for the latent subsets LE (image + EFS) and LG (minutiae only). In each case, the same set of latents is used, varying only by the markup.

These results can be seen as a way of depicting the difference in accuracy between theoretically ideal markup and operational human markup.

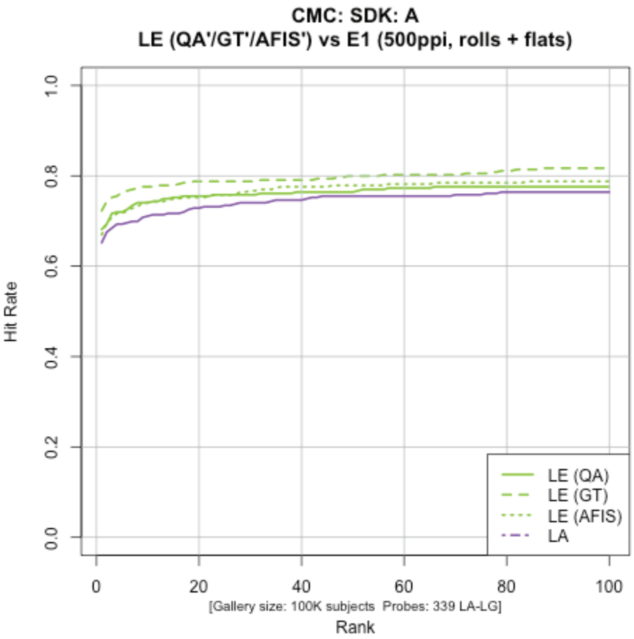
Preliminary observations:

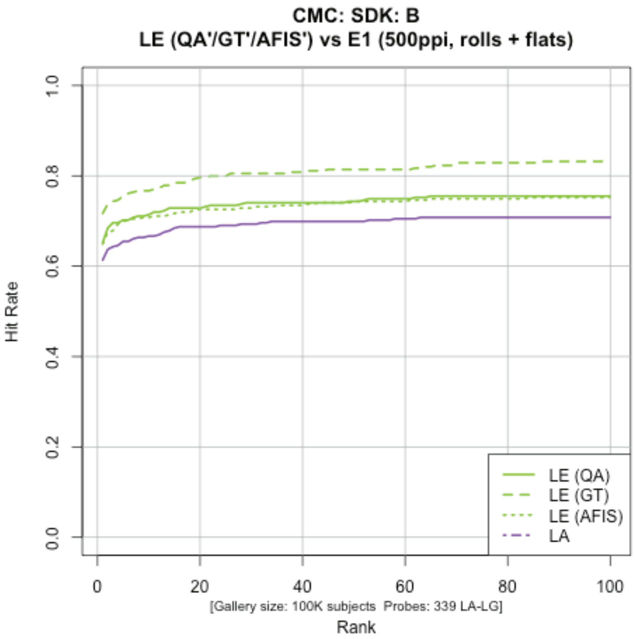
- The "GT" results were beneficial for participants A/B/C using latent subset LE, but were dramatically beneficial for participants A/B/C/E using latent subset LG.
  - For latent subset LE the difference in hit rate between Baseline-QA and GT was limited to about 3-7%.
  - For latent subset LG the difference in hit rate was about 9-15%: the differences between the markups had a direct impact on accuracy, since the matcher had no recourse to the image.
- The "AFIS" markup approach provided no substantive benefit, and was counterproductive in some cases.

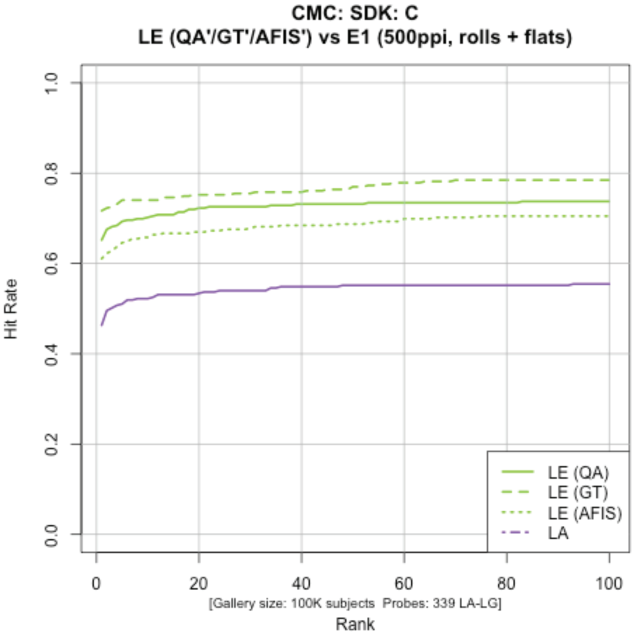
* *A review was conducted to derive a generic set of rules for AFIS feature markup, considering guidance such as excluding minutiae on short ridges or short enclosures, excluding minutiae near the core or in high curvature areas, excluding isolated minutiae, or excluding separated clusters of minutiae. For each of these, it was determined that the guidance was not vendor-neutral, and therefore had the potential to benefit some participants to the detriment of others. The resulting guidance was solely to remove the most debatable minutiae.*

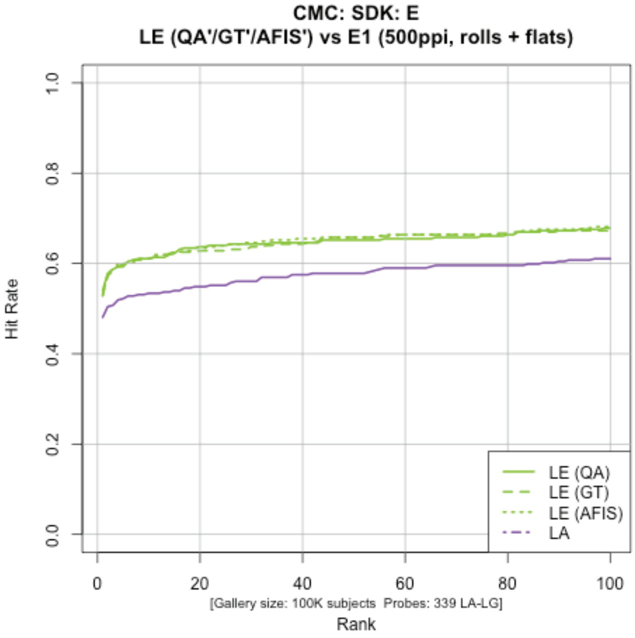
† *Note that the minutiae were groundtruthed against the exemplars, but not the other extended features, such as incipients or skeletons.*

7.1 Effect of GroundTruth markup — subset LE (Image + EFS)





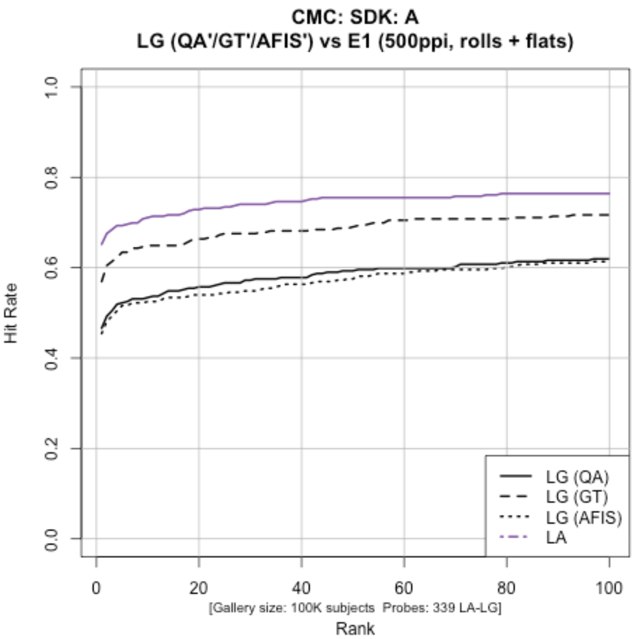


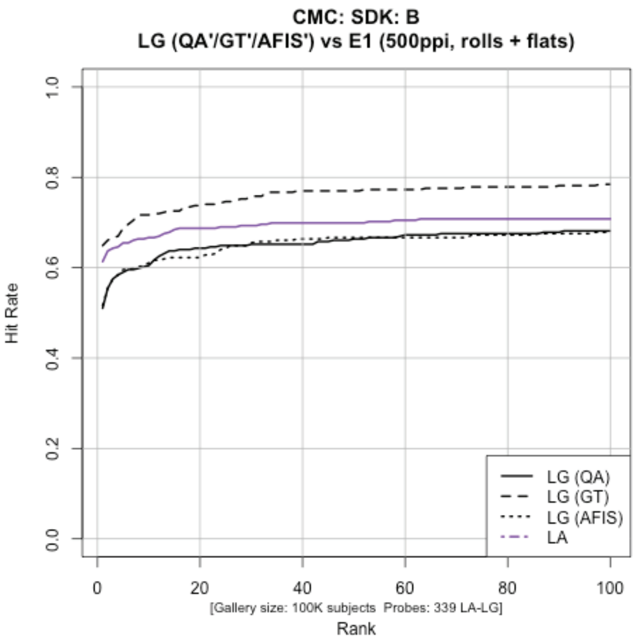


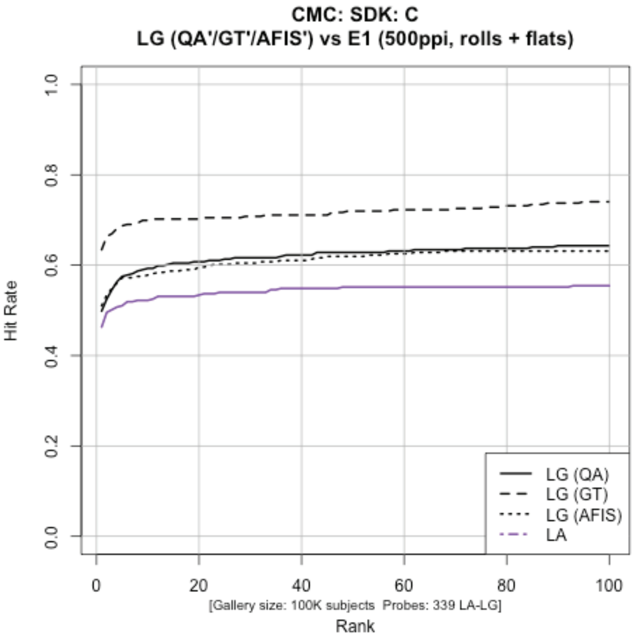
7.2 Effect of GroundTruth markup — subset LG (Minutiae only)

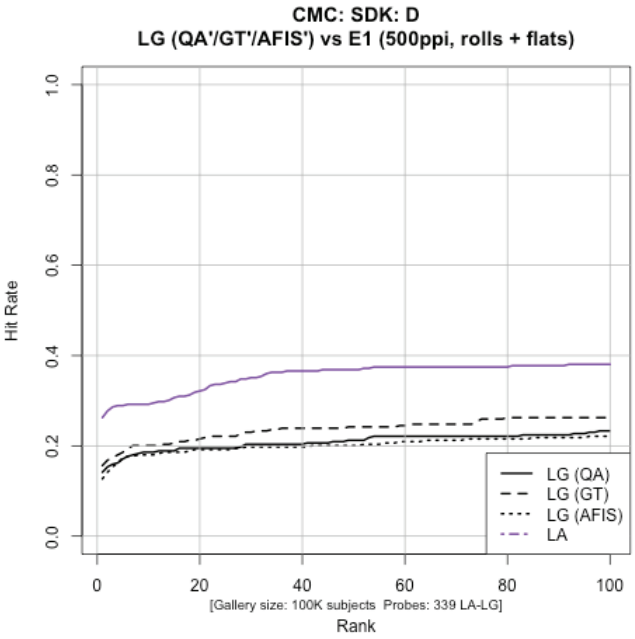
Preliminary observations:

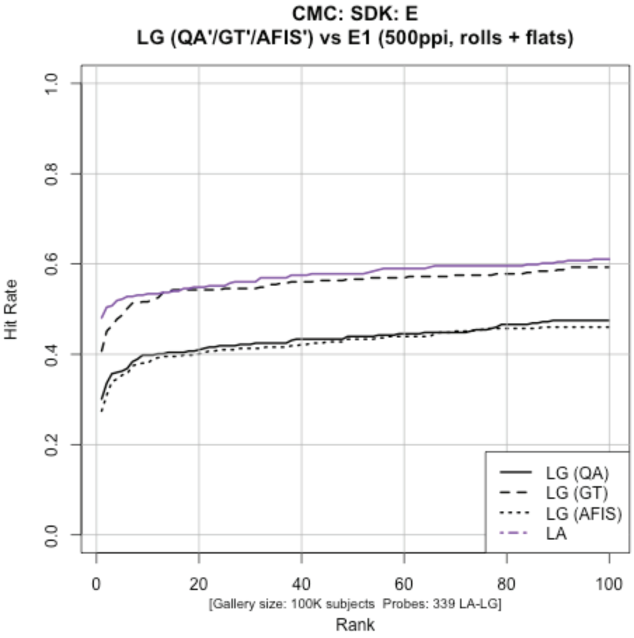
- Note that participants A/D/E have higher accuracy using image only (LA) than even GroundTruth minutiae only (LG).





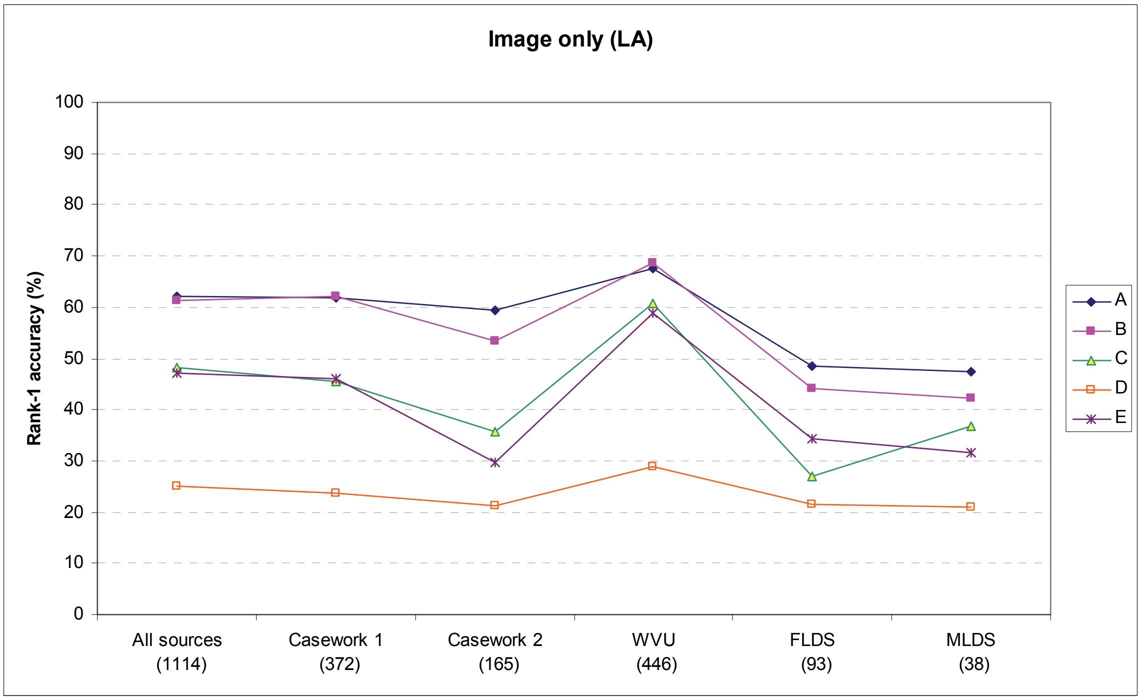


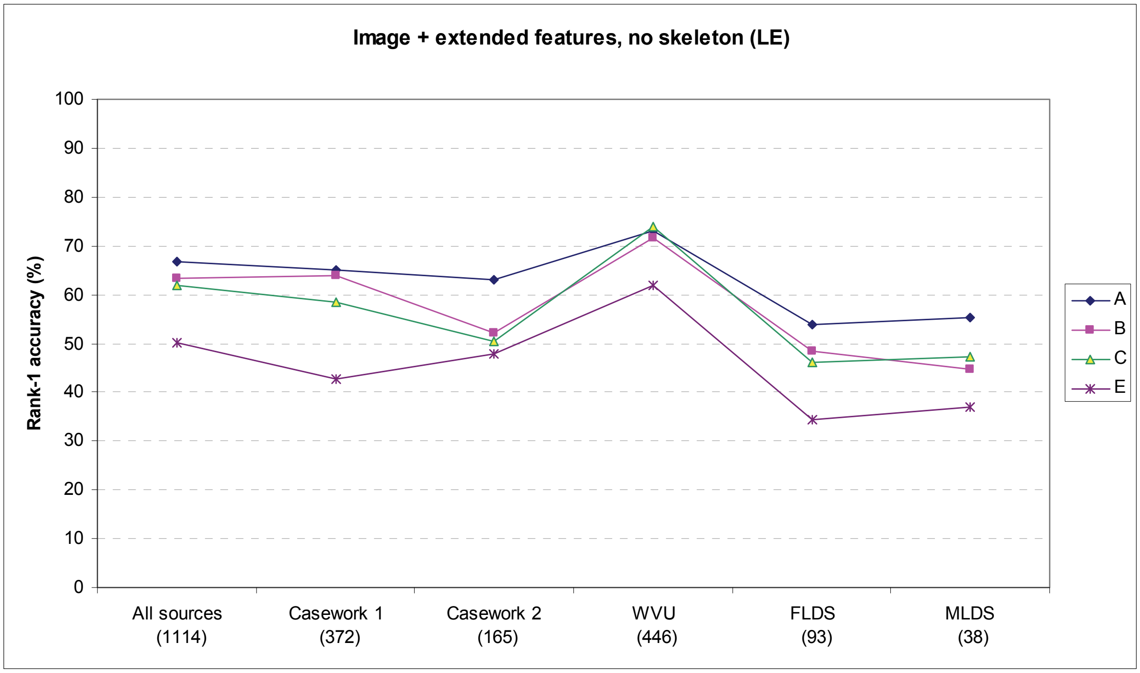


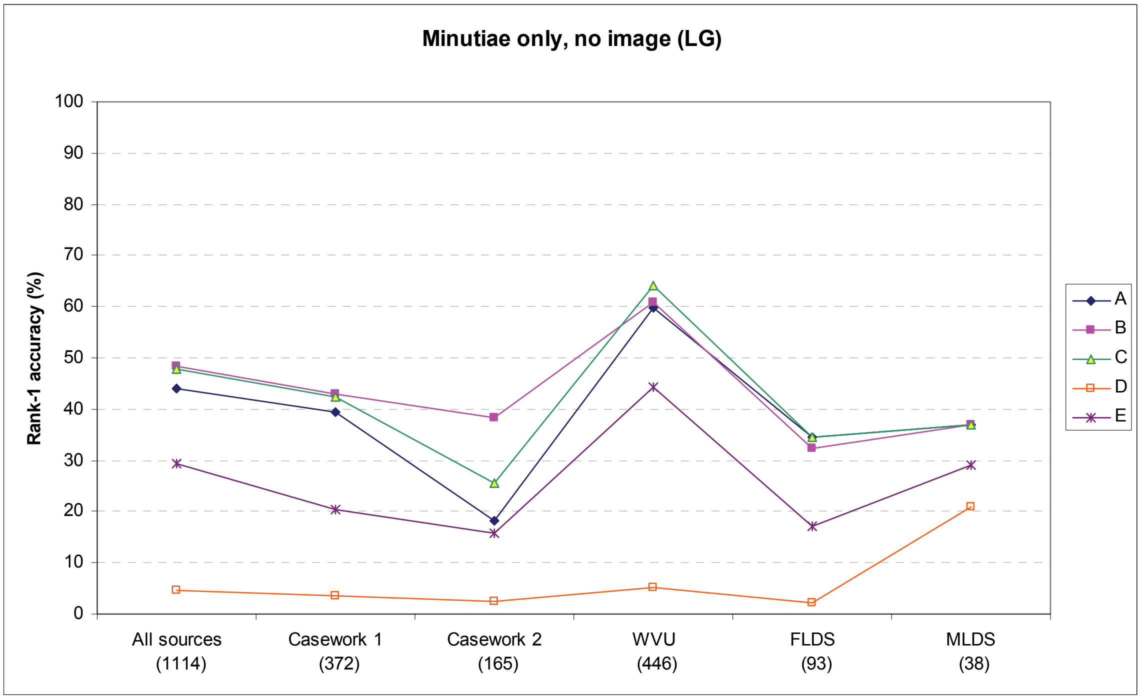


8 Effect of Latent Source

The latents were collected from disparate sources. The following charts show the difference in rank-1 accuracy between the sources of latents for the Baseline dataset. As shown above in Table 1, Casework 1 and 2 are from operational casework, while the others were collected in laboratory conditions.





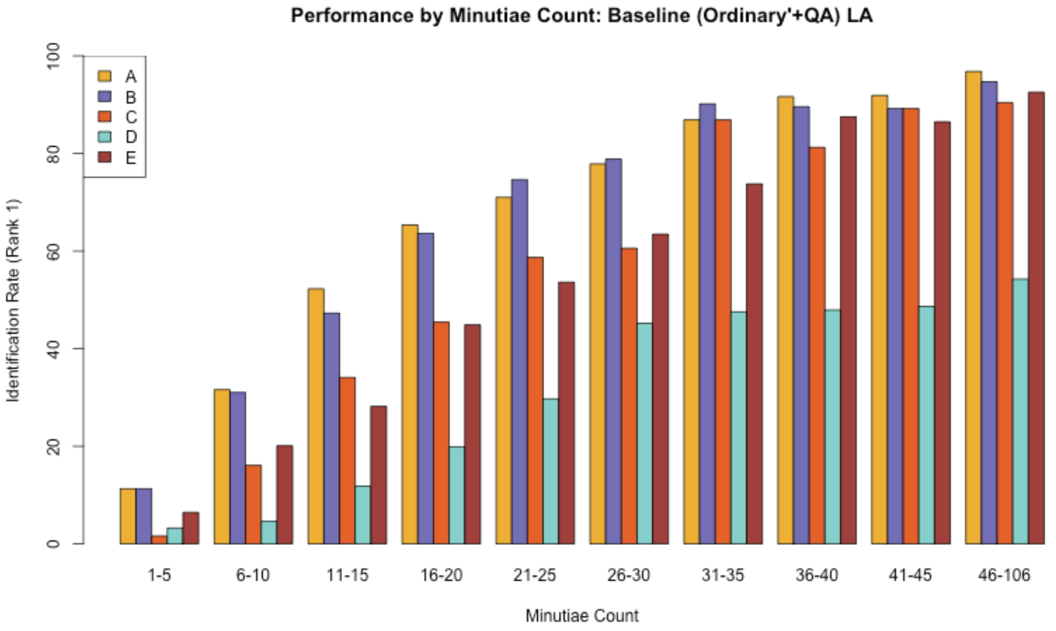


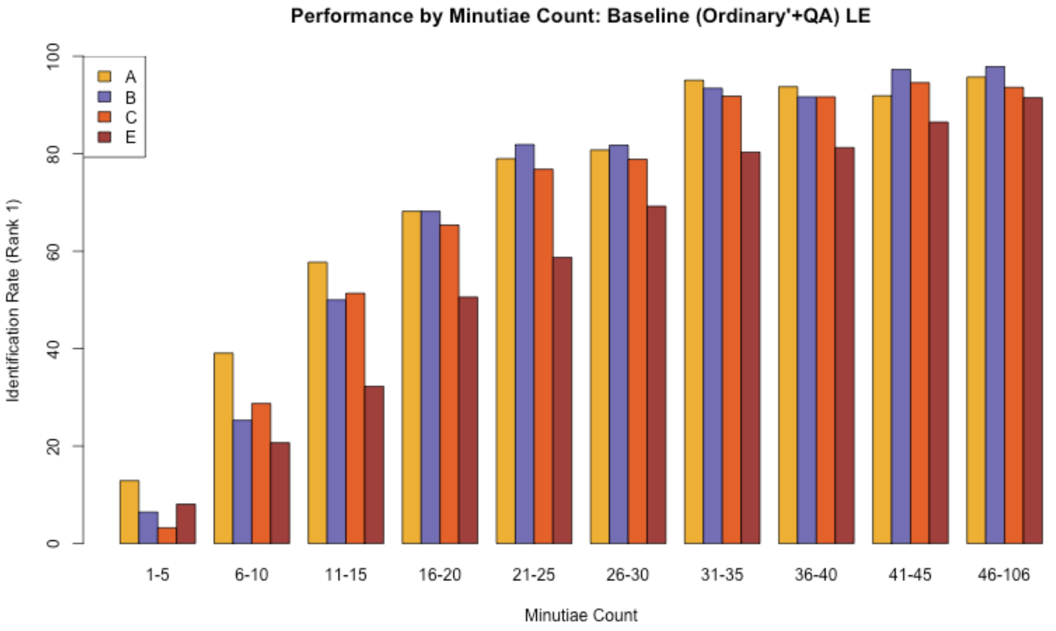
9 Effect of Minutiae Count

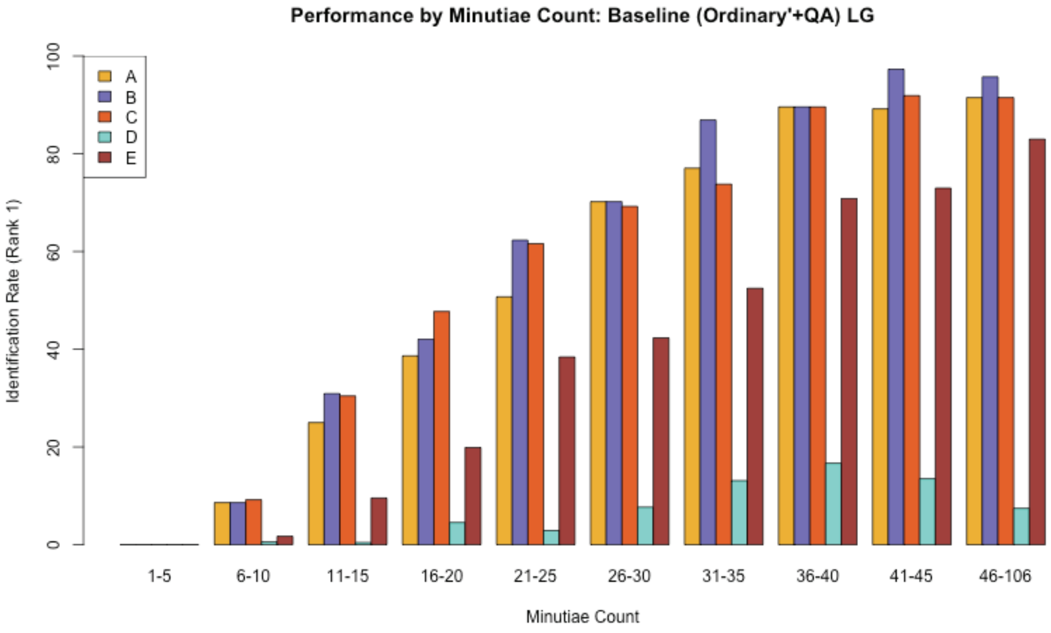
The following charts show rank-1 identification rate broken into bins by minutiae count, for latent subsets LA (image only), LE (image + EFS), and LG (minutiae only).

Preliminary observations:

- Some of the matchers achieve a 10%+ hit rate even on latents with 1-5 minutiae, for subsets LA/LE. On review by an examiner, some but not all the cases with 1-5 minutiae could arguably have had 1-3 additional minutiae marked; note that minutiae count is based on the examiner-marked minutiae, not groundtruthed minutiae.


Performance by Minutiae Count: Baseline (Ordinary'+QA) LA
Identification Rate (Rank 1)
0
20
40
60
80
100
A
B
C
D
E
1-5
6-10
11-15
16-20
21-25
26-30
31-35
36-40
41-45
46-106
Minutiae Count


Performance by Minutiae Count: Baseline (Ordinary'+QA) LE
Identification Rate (Rank 1)
0
20
40
60
80
100
A
B
C
E
1-5
6-10
11-15
16-20
21-25
26-30
31-35
36-40
41-45
46-106
Minutiae Count



10 Effect of Value Determination

As discussed in Section 3.1.3, the examiners who marked the latent images made determinations of Value, Limited Value, or No Value at the time of markup. Table 13 shows the relationship between value determination and rank-1 accuracy.

Preliminary observations:

- As expected, accuracy is very clearly related to value determination, with much greater accuracy for the latents determined a priori to be of value.
- The notable and surprising result is that participant A's hit rate for no value latents is 20% on subsets LA/LE, and even higher on limited value latents.
- The results for participants A/B/C/E show that matching is practical even for limited or no value latents, given lower expectations of accuracy.
- The decision to include no value and limited value latents in the test has been justified.

Table 13: Rank-1 identification rate by value determination

		All	No value	Limited value	Value
Count		1114[*]	25	122	956
LA (Image only)	A	62.2%	20.0%	26.2%	67.9%
	B	61.2%	4.0%	19.7%	67.9%
	C	48.3%	-	14.8%	53.9%
	D	25.1%	-	3.3%	28.7%
	E	47.2%	-	13.9%	52.5%
LE (Image + EFS)	A	66.7%	20.0%	31.2%	72.5%
	B	63.3%	8.0%	22.1%	70.0%
	C	62.0%	8.0%	20.5%	68.6%
	E	50.3%	8.0%	12.3%	56.3%
LG (Minutiae only)	A	44.0%	-	4.9%	50.0%
	B	48.3%	4.0%	3.3%	55.2%
	C	47.8%	4.0%	7.4%	54.0%
	D	4.5%	-	0.8%	5.0%
	E	29.4%	-	1.6%	33.8%

11 Comparison with ELFT Phase II

NIST ELFT Phase II was an evaluation of automatic feature extraction and matching (AFEM) in latent identification, a process directly comparable to the image-only latent subset LA in ELFT-EFS. ELFT Phase II results were published as NISTIR 7577 [3] in April 2009. As shown in Table 14, Figure 1, and Figure 2, the results are substantially different: ELFT-EFS results for image-only matching (AFEM) are far less accurate than in ELFT Phase II, even though three of the participants were the same in both tests.

Table 14: Comparison of ELFT-EFS and ELFT Phase II rank-1 results

ELFT-EFS	
LA - Image only	
Rank 1 against 1M fingerprints	
Sagem	62.2
NEC	61.2
Cogent	48.3
Warwick	47.2
Sonda	25.1

ELFT Phase II	
Rank 1 against 100K fingerprints	
NEC	97.2
Cogent	87.8
SPEX	80.0
Motorola	79.3
L1 Identity Solutions	78.8
Peoplespot	67.9
Sonda	28.5
BioMG	27.5

[*] *Draft note: 11 latents (out of 1114) did not have value determinations; all of those are expected to be found of value.*

Figure 1: Comparison of ELFT-EFS and ELFT Phase II CMC results

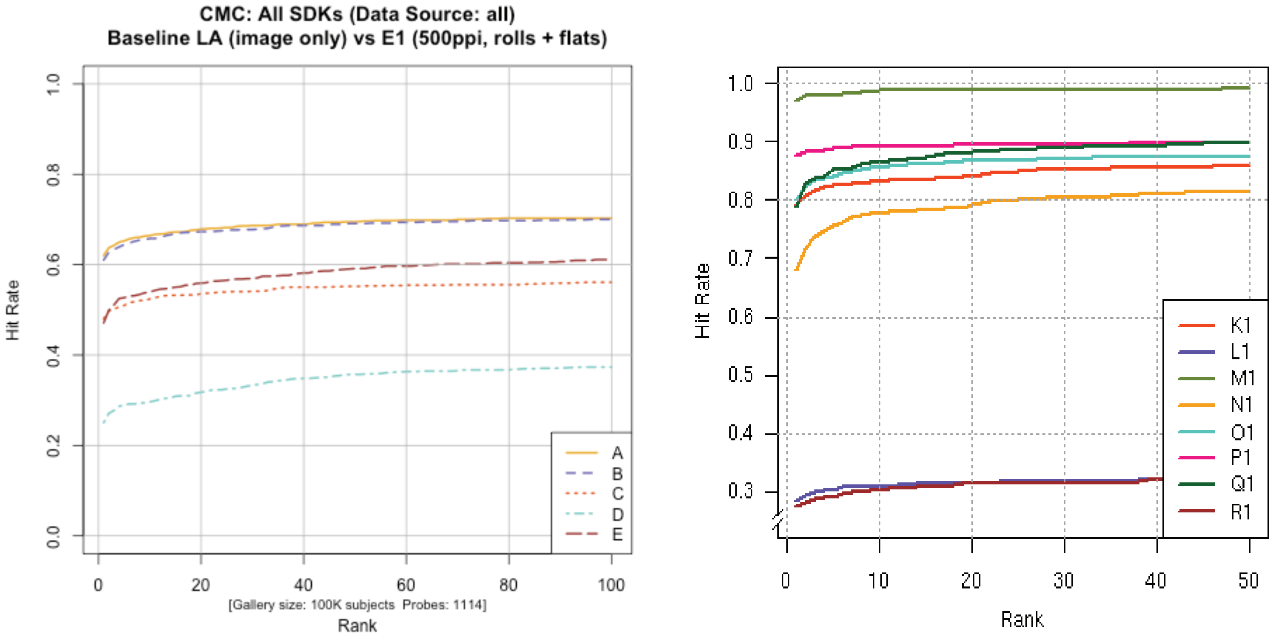
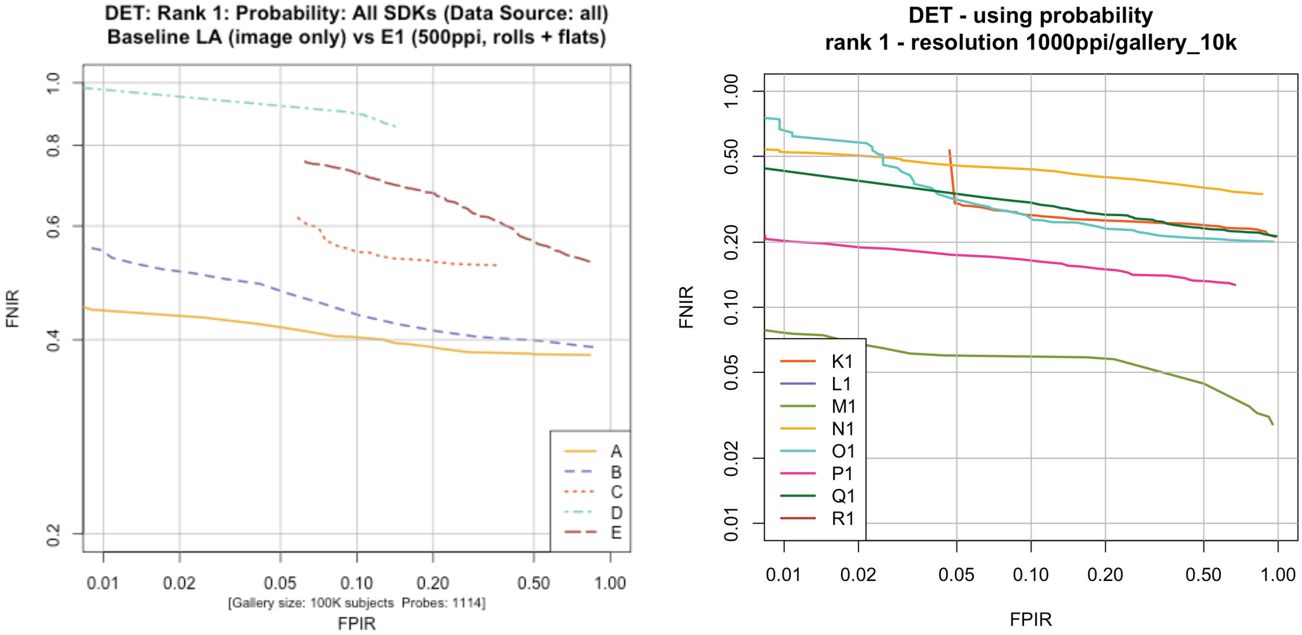


Figure 2: Comparison of ELFT-EFS and ELFT Phase II DET results



The differences between ELFT-EFS and ELFT Phase II results can be attributed to a variety of causes, but the most obvious causes would be differences in data and timing:

- Data
  - The datasets used in ELFT-EFS and ELFT Phase II differed markedly in their sources, selection, difficulty, and proportion of poor-quality images.
  - The latent images used in ELFT Phase II were selected based on successful feature-based seaches of an AFIS. Therefore, the results for ELFT Phase II were not characteristic of latents in general, but instead

    served to quantify what portion of latent images that had successfully been searched as feature searches could have been searched as image searches.
- Timing
    - Matcher accuracy is directly related to processing speed. ELFT-EFS had much more restricted throughput requirements than did ELFT Phase II.

Note also that the gallery for ELFT-EFS was ten times the size of the ELFT Phase II gallery, and contained linked rolled and plain fingerprints for each subject.

12 Comparison with ELFT-EFS Public Challenge

The ELFT-EFS Public Challenge results are included in Appendix B.

The ELFT-EFS Public Challenge was a practice evaluation in preparation for ELFT-EFS Evaluation #1, essentially an open-book test on public data to validate formats and protocols. The ELFT-EFS Public Challenge was an evaluation of self-reported results from a small public dataset. The systems used and timing were not constrained. The results are appropriate for preliminary analysis, but are not appropriate for rigorous analysis or comparison. The participants in the ELFT-EFS Public Challenge are and will remain anonymous.

The results for the Public Challenge can be expected to differ from the results included here for the following reasons:

- The dataset used in the Public Challenge was a public dataset that has been used heavily for research and development of fingerprint systems for fifteen years.
- The gallery used in the Public Challenge was very small (428 subjects).
- Processing time was not constrained for the public challenge.

13 Caveats

- ELFT-EFS was an evaluation of an emerging draft specification (Extended feature set specification [1]). The file format and syntax and semantics of the features were not familiar to the participants, and therefore software for parsing and using the features had to be developed with limited opportunity for testing.
- The schedule was extremely demanding for the participants. The timetable for the test did not permit time for extensive research and development, and instead may have been limited to use or modification of existing software. Therefore, the evaluation is a measurement of the status quo, and in no way measures what the effect of additional research and development might be.

This is a preliminary draft: results and conclusions may change before the final report.

References

1 Committee to Define an Extended Fingerprint Feature Set (CDEFFS); "Data Format for the Interchange of Extended Friction Ridge Features"; Proposed Addendum/Revision to *ANSI/NIST-ITL 1-2007 Data Format for the Interchange of Fingerprint, Facial, & Other Biometric Information*; Draft Version 0.4; 12 June 2009. (http://fingerprint.nist.gov/standard/cdeffs)

2 Hicklin; "Guidelines for Extended Feature Set Markup of Friction Ridge Images" ; Working Draft Version 0.3, 12 June 2009. (http://fingerprint.nist.gov/standard/cdeffs)

3 Indovina, et al; *ELFT Phase II - An Evaluation of Automated Latent Fingerprint Identification Technologies; NISTIR 7577*; April 2009. (http://fingerprint.nist.gov/latent/NISTIR_7577_ELFT_PhaseII.pdf)

Appendix A

NIST ELFT-EFS
Evaluation of Latent Fingerprint Technology — Extended Feature Sets

Test Plan

Contents

1 Overview

The NIST Evaluation of Latent Fingerprint Technology — Extended Feature Sets (ELFT-EFS) is an independently administered technology evaluation of latent fingerprint feature-based matching systems. ELFT-EFS is being conducted by the National Institute of Standards & Technology (NIST).

ELFT-EFS is a complement to NIST's Evaluation of Latent Fingerprint Technology (ELFT) testing program. The ELFT evaluations to date have focused solely on automated feature extraction and matching (AFEM) in the context of latent fingerprint identification.

ELFT-EFS will evaluate the accuracy of latent matching using features marked by experienced human latent fingerprint examiners. The purpose of this test is to evaluate the current state of the art in latent feature-based matching, by comparing the accuracy of searches using images alone with searches using different feature sets. The features sets will include the current IAFIS latent feature set, and different subsets of the Extended Feature Set (EFS) features proposed by CDEFFS[1]. A key result of the test is to determine when human feature markup is effective. Because human markup is expensive in terms of time, effort, and expertise, there is a need to know when image-only searching is adequate, and when the additional effort of marking minutiae and extended features is appropriate.

The following summarizes the planned test:

- The evaluation will involve 1:N searches using latent 1000ppi images provided with human markup of EFS features.
- Exemplars for the gallery will be images only. Exemplars will be 500ppi.
- The test will be an SDK-type test, in that participants will provide software, and all processing will take place on NIST hardware.
- Different tests will be run for the following search types:
    - Image only
    - Image with region of interest markup
    - Image with minutiae (IAFIS EFTS LFFS equivalent)
    - Image with EFS features
    - Minutiae only (IAFIS EFTS LFFS equivalent)

Test results will be made publicly available in a NIST report after the conclusion of the test.

2 Participation

Participation in Evaluation #1 is limited to all organizations which participated in the ELFT-EFS Public Challenge that submitted results by the 28 June 2009 deadline.

All systems must comply with the ELFT-EFS-1 API, outlined in Section 5. Anonymous participation will not be permitted. The Application form includes details regarding application and qualification.

[1] *CDEFFS is the Committee to Define an Extended Friction Ridge Feature Set. The current draft of the Extended Friction Ridge Features specification can be found at http://fingerprint.nist.gov/standard/cdeffs/.*

3 Data

3.1 Datasets

Validation Dataset

A Validation Dataset will be provided to participants before the evaluation to verify the correct operation of participants' software before and after delivery to NIST.

Evaluation Dataset

The Evaluation Dataset will contain sequestered data, formatted in the same manner as the Sample Dataset. The Evaluation Dataset will contain Privacy Act or FOIA Protected Information and will not be released to the participants or the public. The Evaluation Dataset will to the extent permitted by law be protected under the Freedom of Information Act (5 U.S.C 552) and the Privacy Act (5 U.S.C. 552a) as applicable.

3.2 Format

All images and data will be contained in ANSI/NIST files. All images will be 8-bit grayscale.

Each latent ANSI/NIST file in the evaluation will contain one Type-1 record, one Type-2 record, zero or one Type-9 records, and one Type-13 record. All latent images will be in Type-13 records, in uncompressed format.

Each exemplar ANSI/NIST file in the evaluation will contain one Type-1 record, and ten Type-14 records (one for each finger, with finger positions identified). All exemplar images will be in Type-14 records. 500ppi exemplar images will be compressed using WSQ.

3.3 Features

Files containing exemplars will not have any features defined: no Type-9 record will be present.

Files containing latents may or may not have any features defined: zero or one Type-9 records will be present. There will be tests comparing the accuracy of three types of searches:

- Image-only searches, in which the latent image will not be accompanied by a type-9 record.
- Feature-based searches, in which the latent image will be accompanied by a type-9 record with features defined in fields 9.300-9.372, formatted in accordance with "Data Format for the Interchange of Extended Fingerprint and Palmprint Features," abbreviated here as the "EFS Spec" (Extended Feature Set Specification). The test will evaluate different combinations of EFS fields, so not all EFS fields may be present in any given search. The subsets of features used (defined as Subsets LA-LG) are defined in Section 7.

Note: The current EFS Spec version is 0.4 (June 2009).

All of the latent IAFIS/EFS features will be provided with feature markup by human experts. Note that all human markup will be conducted outside of ELFT-EFS and is not part of the evaluation.

Note also that conformance testing of automatic extraction of CDEFFS features is not part of this test. In other words, the evaluation will not be measuring how close automatically extracted features are to examiner created features. Automated algorithms can use the extended features defined for a latent search without explicitly computing them for the exemplar image, and thus it must be emphasized that automated extraction of the extended features on the exemplar is not necessarily the only nor the best way to use this information. For example, an examiner may mark an area as a scar; for the exemplar, the matcher would not necessarily have to mark the area as a scar, but may use that information to match against a corresponding area with many false minutiae and poor ridge flow.

3.4 Resolution

All latent images will be 1000 pixels per inch.

Exemplar images will be at 500 pixels per inch. This resolution will be contained in field 14.009 (Horizontal pixel scale), which will be identical to field 14.010 (Vertical pixel scale).

3.5 Dimensions and orientation

Latent fingerprint images may vary from 0.3"x 0.3" to 2.0" x 2.0" (width x height), all at 1000ppi. 1st & 3rd quartiles are about 700-1200 pixels (width) or 900-1400 pixels (height).

Exemplar images will be approximately upright (in the same orientation as they were captured).

Neither latent nor exemplar images will be larger than 2.0" in either width or height.

Latent fingerprint images may vary in orientation from upright ±180°. Images accompanied by EFS fields may or may not include the orientation direction and uncertainty fields (9.301).

3.6 Exemplar types

All exemplars will include rolled and/or plain (segmented slap) fingerprints. The impression types will include optical livescan and inked paper sources. The impression type will be noted in field 14.003.

Exemplars will always include all ten fingers.

Note that an exemplar set may include rolls alone, plains alone, or both.

In some cases, multiple sets of exemplars associated with one person will be included in the gallery. This association will be made explicit in the exemplar enrollment stage: at the time of enrollment, exemplars that are known to belong to the same person will always share the same subject ID.

3.7 Finger positions

Exemplars will be provided in complete 10-finger sets, all contained within a single ANSI/NIST file, with finger positions noted.

The finger positions for latents will not be noted – no searches will be restricted to specific fingers.

3.8 Dataset size

The largest size gallery used for Evaluation #1 will contain 100,000 subjects having two 10-finger exemplar sets (rolled and plain impressions) per subject.

The total number of unique latent images is approximately 1,500, with the number of latent searches based on section 4.2.

4 Evaluation Criteria

4.1 Performance Metrics

Performance metrics will be based on rank and matcher score:

- Rank will be reported by the number of true matches reported in each position in the candidate list. For example, the Rank-1 metric is the proportion of searches in which the correct mate appears in the top position on the candidate list. CMC[2] curves will also be reported to show how many latent images are correctly identified at rank 1, rank 2, etc. A CMC is a plot of identification rate vs. recognition rank. Identification rate at rank k is the proportion of the latent images correctly identified at rank K or lower. A latent image has rank k if its mate is the kth largest comparison score on the candidate list. Recognition rank ranges from 1 to 50, as 50 is the (maximum) candidate list size specified in the API.

- Matcher score metrics are evaluated in terms of DET/ROC[3] performance, by plotting False Positive Identification Rate (FPIR) and False Negative Identification Rate (FNIR) for all score values. Note that this approach requires that a given matcher score be comparable between different latent searches. Both the absolute matcher score and the probability of true match values (see Section 5.6) will be used for DET analysis.

4.2 Evaluation Subtests

The Evaluation is composed of the following subtests. For precise definitions of which features will be present for each subtest: see Section 7. All latents in each subtest may or may not be searched against all exemplars (galleries).

- Latent Subtests
  - LA – image only
  - LB – image + ROI
  - LC – image + ROI + Pattern Class + Quality Map
  - LD – image + IAFIS/EFTS equivalent features
  - LE – image + baseline EFS[4]
  - LF – image + baseline EFS + Skeleton
  - LG – IAFIS/EFTS equivalent features only
- Exemplar Subtests
  - E1 – 100,000 subjects; 10 rolled & 10 plain impressions each; 500ppi
  - E2 – 10,000 subjects; 10 rolled impressions each; 500ppi
  - E3 – 10,000 subjects; 10 plain impressions each; 500ppi
  - E4 – 10,000 subjects; 2 sets of 10 rolled+slap impressions each; 500ppi
  - E5 – 10,000 subjects; 3 sets of 10 rolled+slap impressions each; 500ppi
  - E6 – 10,000 subjects; 4 sets of 10 rolled+slap impressions each; 500ppi
  - E7 – 10,000 subjects; 5 sets of 10 rolled+slap impressions each; 500ppi

[2] *Cumulative Match Characteristic*

[3] *Detection Error Trade-off/Receiver Operating Characteristic*

4.3 Reporting of Results

The ELFT-EFS Final Report will contain descriptive information concerning the evaluation, descriptions of each experiment, aggregate test results across all participants, and individual test results for each participant. All results will be reported for each participating system, with the exception of results for different combinations of EFS features. Because not all participating systems may implement all of the EFS features, results from those evaluations will be stated in generic terms so that participants cannot deduce which features are used by other systems.

Note that the application form stipulates that each participant consents to the disclosure of its performance.

Enrollment, feature extraction and search timing information will also be reported, with the explicit caveat that speed of execution, for both enrollment and latent search, is of secondary importance. The report will specify the hardware specifications used in the evaluation, and will also note that operational latent searching algorithms are likely to be implemented in more sophisticated hardware.

5 Latent Matching Software

5.1 Overview

Participants shall submit a set of SDKs (Software Development Kits) that provide the interfaces defined by the ELFT-EFS-1 API specified below. The SDKs shall be provided as static or dynamic libraries to run on the NIST platform specified below. The ELFT-EFS API (Application Programmer Interface) is modeled after the API from ELFT Phase 2. The most notable differences from the ELFT Phase 2 API are that the exemplar and latent images and data provided to the SDK will be contained in ANSI/NIST files, and exemplar feature extraction will process a single exemplar per invocation (instead of the complete gallery). Also, the ELFT-EFS-1 API specifies operational time limits on a per-processor core basis, rather than per-machine.

Each participant shall submit

- one SDK for exemplar feature extraction and exemplar enrollment
- one SDK for latent feature extraction
- one SDK for latent 1-to-N search

NIST recognizes the proprietary nature of the participant's software and will take all reasonable steps to protect this. The software submitted will be in an executable library format, and no algorithmic details need be supplied. NIST agrees not to use the Participant's software for purposes other than indicated above, without express permission by the Participant.

5.2 Test Platform

The NIST ELFT-EFS Evaluation test platform consists of an array of blade servers having a hardware configuration similar to:

Processor

- Dual 2.8 GHz/1MB Cache, Xeon (dual-core)
- 800 MHz Front Side Bus for PE 1855

Memory

- 16GB RAM (15GB available to applications)

Secondary storage

- 300GB 15K RPM Ultra SCSI Hard drives

The operating systems available (in order of preference) are:

- RedHat Linux 3.1 64-bit
- Windows 2008 Server 64-bit
- (Windows Server 32-bit may be available on request)

The available RAM for 64-bit SDKs will be no more than 15GB total. The available RAM for 32-bit SDKs will be no more than 3GB per process.

5.3 Execution protocol

Each SDK tested will be allocated multiple blades/cores from the array, along with a subset of the test data in order to maximize (time) efficiency through parallel operation.

Each SDK instance assigned to an individual blade or core will operate on a subset of the data, using individual data copies (as needed) from a local storage device.

For purposes of execution, there are two classes of SDKs, (1) sequential and (2) multithreaded. And each class the SDK may utilize either 32 or 64-bit execution mode. *Note that each SDK submitted (i.e. either of the two SDKs per participant) may be of a different class and execution mode. For example, the Exemplar feature extraction / enrollment SDK may be sequential 32-bit and the Latent feature extraction / search SDK may be multithreaded 64-bit.*

It is highly recommended that SDKs implement multithreading using 64-bit execution mode. However, if some participants are unable or unwilling to submit multithreaded or 64-bit SDKs, we support other modes of operation as outlined below.

5.3.1 Sequential

An advantage of sequential (i.e. non-multithreaded) SDKs is the ability to "manually" parallelize SDK execution for a given test by executing multiple instances per blade server (e.g. one per core). A potential drawback is that individual 64-bit SDK instances have the potential to over-allocate available RAM, which may result in "swapping," decreasing overall execution speed. Another potential drawback is contention for resources given that each instance is executing independently (i.e. without coordinated resource usage). For this reason NIST does not recommend the submission of sequential SDKs.

As a simple example, the execution of a sequential SDK for a subtest requiring M latent searches against N exemplars (i.e. Gallery size N), may allocate M searches amongst K available cores such that each core is executing M/K searches total. The primary choice here is whether or not to allocate all cores available on a given blade server, or a subset thereof. How much memory is allocated by the SDK (limited by whether it is 32 or 64-bit mode) is a primary consideration.

Sequential SDKs which run in 32-bit execution mode shall have access to no more than 3GB per process. NIST will execute four (4) SDK instances (one instance per core) on each available blade server, in order to maximize processor and memory utilization.

Sequential SDKs which run in 64-bit execution mode shall have access of up to 15GB per process, and the participant should inform NIST at submission time as to the SDK's memory usage requirements. It is strongly recommended that the SDK perform most efficiently when executed as four (4) instances (one per core) on each blade server, where each instance allocates no more than a quarter of available RAM (i.e. 3.75GB), as opposed to when executed as a single (1) instance on each blade server which allocates all available RAM (i.e. 15GB). If more than 3.75GB is allocated per instance, the number of cores which can be utilized per blade server (without swapping) is essentially 15GB divided by the amount of RAM allocated per SDK instance (rounded to the nearest whole number).

5.3.2 Multithreaded

An advantage of multithreaded SDKs is the automatic utilization of available processor and memory resources through parallelization (without need for "manual" scheduling). Another advantage is coordinated access (of each thread) to resources such as disk I/O. For this reason NIST strongly recommends that submitted SDKs utilize multithreading aimed at maximizing usage of 4 cores and run in 64-bit mode in order to have access of up to 15GB of RAM.

As a simple example, the execution of a multithreaded SDK for a subtest requiring M latent searches of N exemplars (i.e. Gallery size N), will allocate M searches amongst K available blades such that each blade is executing M/K searches total.

Multithreaded SDKs which run in 64-bit mode have full access to all cores and memory (15GB) on each allocated blade. This approach clearly makes use of processing resources, and has the potential to mitigate contention issues through a coordinated use of parallelism.

Multithreaded SDKs which run in 32-bit mode will be limited to 3GB of RAM per process, which may limit their performance. Another option which exists here is for a multithreaded SDK to use no more than 2 threads, where each SDK instance uses the maximum 3GB of RAM. If informed, NIST could allocate two such SDKs per blade server in order to more fully utilize RAM.

5.4 API

The software undergoing testing will be hosted on NIST-supplied computers. The executable software under test will be built up from two sources: participant-supplied (SDK) and NIST test driver. The SDKs being tested shall contain the following function entry points:

- Exemplar SDK
  - o extract_exemplar() — Extracts features from a single exemplar ANSI/NIST file (containing 10 images), creating a file containing the vendor-specific features for that exemplar set.
  - o create_gallery() — Associates a set of extracted exemplar features into a Gallery.
- Latent extraction SDK
  - o extract_latent() — Extracts features from a latent image, creating a file containing the vendor-specific features for that latent.
- Latent matching SDK
  - o set_gallery() — Selects the Gallery that latents will be searched against.
  - o latent_search() — Searches the latent against the select Gallery for potential mates.

extract_exemplar()

Inputs: exemplarFilename (string), outputDirectory (string)

Outputs: Writes extracted features to outputDirectory/exemplarFilename.feat

Notes: Takes the ANSI/NIST file *exemplarFilename*, containing ten rolled OR 10 segmented slap fingerprint images, and saves the result as *exemplarFilename*.feat. No other files other than the an2 file may be read; no other files other than the feat file may be written. The contents of *outputDirectory* (structure and other contents) are not relevant. All paths include Unix-style forward slashes.

Example: if *exemplarFilename*=”/input/dir/E9999_1.an2” and *outputDirectory*=”/output/directory”, processes the file “/input/dir/E9999_1.an2” and creates the file “/output/directory/E9999_1.feat”

create_gallery()

Inputs: *exemplarFeatFilenames* (list of strings), *galleryDirectory* (string)

Outputs: Writes exemplar features to *galleryDirectory*, and associates exemplar feature sets that share the same ID

Notes: The format of the enrolled gallery is at the discretion of the SDK provider. Subdirectories and multiple files may be created within *galleryDirectory*. The exemplar feature filename will be formatted “E”*subjectID* “_” *instance* “.feat”, where subjectID is the numeric ID for the subject, and the instance is a 1-based arbitrary numeric index to differentiate between multiple exemplar sets belonging to the same subject.

Example: if *exemplarFeatFilenames* = [“/input/path/E9999_1.feat” “/input/path/E12345_1.feat” “/input/path/E12345_2.feat”] and *galleryDirectory*= “/this/gallery/”, the software inserts the exemplar feature sets into the stated gallery, associating the two files with the subject number 12345.

extract_latent()

Inputs: *latent_fileID* (string), *output_features_dir* (string)

Outputs: Writes extracted features to *output_features_dir* as “*latent_fileID*.feat”

Notes: *latent_fileID* is the filename of an ANSI/NIST file

Example: if *latent_fileID* = “/in/path/L12ABC.an2”, and *output_features_dir*=”/out/path”, creates ”/out/path/L12ABC.an2”.

set_gallery()

Inputs: *galleryDirectory* (string)

Outputs: Sets the latent matching SDK to use specified *galleryDirectory* for all subsequent latent_search() calls.

Notes: *galleryDirectory* was defined using previous *create_gallery* calls.

latent_search()

Inputs: *latent_feature_fileID* (string), *candidate_dir* (string)

Outputs: Writes candidate list file to *candidate_dir ;* name and format specified in Section 5.6.

Notes: *latent_feature_fileID* is a filename of a vendor-specific feature file

Example: if *latent_feature_fileID*="/directory/L123.feat" and *candidate_dir*="/cand/list/dir", searches the gallery selected in the previous *set_gallery* call, and creates a candidate list in the file "/cand/list/dir/L123.CL".

5.5 Software execution process

The execution process will take place in three passes:

- Exemplar feature extractions and Gallery creation
- Latent image feature extractions
- Latent searches against each Gallery

5.6 Format of Candidate List

The result of the latent_search() function is a candidate list, saved as a tab-delimited text file. The candidate list has a fixed length of one hundred (100) candidates. The candidate list consists of two parts, a required and an optional part.

The required part consists of:

- the ID of the mating exemplar subject
- the matching finger number
- the absolute matching score
- an estimate of the probability of a match (0 to 100)

The optional part consists of:

- the number of minutiae identified in the latent
- the number of latent minutiae which were successfully matched

Sample Candidate List						
Required Part					Optional Part	
Rank	Mate ID	Finger No.	Abs. Score	Prob. Of True Match	No. Latent Minutiae	Minutiae Matched
1	73141	2	3513	93	18	12
2	10316	2	605	5	18	5
3	14334	3	513	4	18	5
…						
100	20792	9	422	1	18	4

Table 1: Sample candidate list

The candidate list is ordered based upon the absolute score, with the highest score in the first position.

The parameter *Probability of True Match* is an estimate of the probability that the candidate is a true match. Its values range from 0 to 100.

Each candidate list will be stored in an individual tab-delimited ASCII text file having the extension ".CL" and the base filename of the *latent_feature_fileID* specified to the latent_search() function (e.g. the candidate list for a search of "L9999.feat" will be named "L9999.CL"). Within the candidate list file, all required and optional parts for an individual candidate entry (i.e. row) should be written one per-line in the order shown above, with each part (i.e. column) separated by a single tab character. Note that "Mate ID" shall be written as the numeric portion of the exemplar filename specified to the create_gallery() function (e.g. if "E99999_12.feat" was enrolled to the gallery being searched, the Mate ID shall be "99999" without quotes). Note also that the candidate list refers to a subject and finger position, not a specific exemplar impression.

5.7 Validation

As discussed in Section 3.1, a Validation Dataset will be provided to verify the correct operation of participants' software before and after delivery to NIST. Using this data and the submitted SDK, identical outputs must be generated by NIST to those submitted by participants in order for the submitted SDK to be accepted. Acceptance of the submitted SDK must occur prior to the deadlines specified in section 6.

The Validation Dataset will be a small subset of the ELFT-EFS Public challenge dataset.

5.8 Timing Requirements

The ELFT-EFS Evaluation test must place limits on the processing time of the major operations involving feature extraction and enrollment (exemplars and latents) and searching. There are two purposes for such limits. The first is to enable practical execution of the test within an acceptable period of time. The second is to measure performance at throughput rates comparable to large-scale operational scenarios. Our sponsors have interest in relevance of results to near-term operational requirements. The size of the test will be dictated to a large extent by these throughput numbers.

SDK time limits are specified on a "per-core" basis, meaning that the specified operational rates are for a single core – in other words, rates will be specified from the perspective of a sequential process executing on a single CPU core. For example, if the specified rate for latent search is R exemplars per second, then a multithreaded SDK instance operating on 4 cores must achieve an aggregate rate of 4 x R. All time limits below are averages with respect to the hardware used on the NIST test platform specified above.

The search time requirements specified below are for Subtests LC-LG: see Section 7 for details. It is recognized that for some implementations, throughput for image-only searches (Subtest LA) may be slower due to less effective screening. It is allowable for throughput on Subtest LA (image only) and LB (image+ROI) to be slower by a factor of up to 2x than the stated search time.

Time limits for the ELFT-EFS Evaluation are (per single CPU core):

Exemplar feature extraction	100 sec/10-finger exemplar set (rolled or pre-segmented slap)
Latent enroll	120 sec/latent
Search	0.05 sec/exemplar set (20 exemplar sets/sec, per latent, assuming an exemplar set consists of 10 rolled and 10 segmented slap fingerprints)

Table 2: Timing requirements

6 Schedule and Software Submission Requirements

To enable enrolling the gallery before the evaluation itself takes place, we are requesting the exemplar feature extraction/enrollment SDKs prior to the latent feature extraction/search SDKs. For each SDK, we have both early and final deadlines: we will accept SDKs as early as the early deadline, and will use the period from receipt of the SDKs until the final deadline to validate correct operation of the SDKs, but must have fully operational software by the final deadline. Between the early and final deadlines, we will report any software issues encountered, and will accept software replacements.

If major software problems arise during the execution of the evaluation (i.e. after the submission deadline), reasonable attempts will be made to resolve the issue(s) through reporting and receipt of replacement software. However replacement software must not include algorithm enhancements beyond those addressing the specific problem(s) reported.

Registration/Withdraw

- Registration form online: 13 July 2009
- Registration deadline: 27 July 2009
- Deadline for anonymous withdraw: 16 August 2009

Exemplar feature extraction / enrollment SDKs:

- Early deadline: 2 August 2009
- Final deadline: 16 August 2009
- Preparation of galleries will start when SDKs are validated, but no later than Monday 17 August

Latent feature extraction / search SDKs:

- Early deadline: 16 August 2009
- Final deadline: 30 August 2009
- Latent evaluations to start post SDK validation, but no later than Monday 31 August

7 EFS Fields Used

			Subtest combinations for ELFT-EFS Evaluation #1						
Abb.	#	Field Name	Subtest LA: Image only	Subtest LB: ROI	Subtest LC: ROI, Pattern Class, Quality Map	Subtest LD: IAFIS/ EFTS equivalent	Subtest LE: Baseline EFS	Subtest LF: Baseline EFS with Skeleton	Subtest LG: IAFIS/ EFTS equivalent
			With Image						Without Image
LEN	9.001	Logical Record Length	Yes	Yes	Yes	Yes	Yes	Yes	Yes
IDC	9.002	Image Designation Character	Yes	Yes	Yes	Yes	Yes	Yes	Yes
IMP	9.003	Impression Type	Yes	Yes	Yes	Yes	Yes	Yes	Yes
FMT	9.004	Minutiae Format	Yes	Yes	Yes	Yes	Yes	Yes	Yes
ROI	9.300	Region of Interest		Yes	Yes	Yes	Yes	Yes	Yes
ORT	9.301	Orientation		Yes	Yes	Yes	Yes	Yes	Yes
FPP	9.302	Finger/Palm Position(s)							
PAT	9.307	Pattern Classification			Yes	Yes (**)	Yes	Yes	Yes (**)
RQM	9.308	Ridge Quality Map			Yes		Yes	Yes	
RQF	9.309	Ridge Quality Map Format			Yes		Yes	Yes	

RFM	9.310	Ridge Flow Map						Yes	
RFF	9.311	Ridge Flow Map Format						Yes	
RWM	9.312	Ridge Wavelength Map							
RWF	9.313	Ridge Wavelength Map Format							
TRV	9.314	Tonal Reversal		Yes	Yes	Yes	Yes	Yes	Yes
PLR	9.315	Possible Lateral Reversal							
FQM	9.316	Friction Ridge Quality Metric							
PGS	9.317	Possible Growth or Shrinkage							
COR	9.320	Cores				Yes	Yes	Yes	Yes
DEL	9.321	Deltas				Yes	Yes	Yes	Yes
CDR	9.322	Core-Delta Ridge Counts				Yes	Yes	Yes	Yes
CPR	9.323	Center Point of Reference					Yes	Yes	
DIS	9.324	Distinctive Characteristics					Yes	Yes	
NCR	9.325	No Cores Present					Yes	Yes	
NDL	9.326	No Deltas Present					Yes	Yes	
NDC	9.327	No Distinctive Areas Present					Yes	Yes	
MIN	9.331	Minutiae				Yes (*)	Yes	Yes	Yes (*)
MRA	9.332	Minutiae Ridge Count Algorithm							
MRC	9.333	Minutiae Ridge Counts				Yes	Yes	Yes	Yes
NMP	9.334	No Minutiae Present					Yes	Yes	
RCC	9.335	Ridge Count Confidence					Yes	Yes	
DOT	9.340	Dots					Yes	Yes	
INR	9.341	Incipient Ridges					Yes	Yes	
CLD	9.342	Creases and Linear Discontinuities					Yes	Yes	
REF	9.343	Ridge Edge Features					Yes	Yes	
NPP	9.344	No Pores Present					Yes	Yes	
POR	9.345	Pores					Yes	Yes	
NDT	9.346	No Dots Present					Yes	Yes	
NIR	9.347	No Incipient Ridges Present					Yes	Yes	
NCR	9.348	No Creases Present					Yes	Yes	
NRE	9.349	No Ridge Edges Present					Yes	Yes	
MFD	9.350	Method of Feature Detection							
COM	9.351	Comments							
LPM	9.352	Latent Processing Method							
EAA	9.353	Examiner Analysis Assessment					Yes	Yes	
EOF	9.354	Evidence of Fraud							
LSB	9.355	Latent Substrate							
LMT	9.356	Latent Matrix							
LQI	9.357	Local quality issues					Yes	Yes	
AOC	9.360	Area of Correspondence							
CPF	9.361	Corresponding Points or Features							
ECD	9.362	Examiner Comparison Determination							
SIM	9.372	Skeletonized Image						Yes (***)	
RPS	9.373	Ridge Path Segments							

NOTES	
*	IAFIS/EFTS equivalent minutiae will include X,Y,Theta, Type; but NOT radius of uncertainty or direction uncertainty.

**	IAFIS/EFTS pattern class is limited to the General classification information item, NOT the subclassification or delta relationship.
***	Skeleton and ridge flow are only available for a subset of images.

Appendix B

ELFT-EFS
NIST Evaluation of Latent Fingerprint Technologies: Extended Feature Sets

Public Challenge Results

Contents

1 Introduction

ELFT-EFS is an evaluation of automated latent fingerprint matching software. The purpose of this evaluation is to determine the effectiveness of human latent examiner-marked fingerprint features on latent fingerprint search accuracy, specifically with respect to the comparative accuracy of image-only searches, image+minutiae searches, and image+extended feature searches.

ELFT-EFS Public Challenge

> The ELFT-EFS Public Challenge is a practice evaluation: an open-book test on public data to validate formats and protocols. Note that the ELFT-EFS Public Challenge was an evaluation of self-reported results from a small public dataset. The systems used and timing were not constrained. These results are appropriate for preliminary analysis, but are ***not*** appropriate for rigorous analysis or comparison. The ELFT-EFS Evaluation #1 is intended for those purposes. The participants in this evaluation are and will remain anonymous.

ELFT-EFS Evaluation #1

> NIST will conduct the ELFT-EFS Evaluation #1 using participants' software on NIST hardware at NIST facilities. Datasets will be from multiple sequestered sources, each broadly representative of casework. The ELFT-EFS evaluation #1 will be run specifically to identify any near-term benefits, NOT to identify long-term feasibility/accuracy. The ELFT-EFS 1st Evaluation timing constraints, subtests, and analysis are being based in part on the results and lessons learned from the ELFT-EFS Public Challenge.

Subsequent Evaluations

> Subsequent ELFT-EFS Evaluations will be conducted to identify long-term feasibility and respond to lessons learned.

2 Overview of challenge problem

The challenge problem will be conducted at the participants' facilities, using the public challenge data, with self-reported results.

The challenge problem will involve 1:N searches using latent 1000ppi images provided with human markup of CDEFFS features. Each latent search will result in a list of candidates, with scores, across all exemplars in the subtest, including all fingerprint sets for each individual and all finger positions. Normalized/probability scores shall be provided in addition.

The challenge is composed of the following subtests. Participants are requested to do all 20 combinations (e.g. L1E1 .. L5E4), but may choose to do only some combinations.

- Latent Subtests
  - L1 – image only
  - L2 – image with EFTS-LFFS features (fields 9.014-9.023)
  - L3 – image with EFS features (fields 9.300-9.373)
  - L4 - EFS features alone
  - L5 - EFTS-LFFS features alone
- Exemplar subtests
  - E1 - 1000ppi rolled exemplars
  - E2 - 500ppi rolled exemplars
  - E3 - 1000ppi plain exemplars (unsegmented slaps)
  - E4 - 500ppi plain exemplars (unsegmented slaps)

3 Data

The ELFT-EFS Public Challenge dataset is a dataset of latent images and corresponding exemplars. This dataset was collected from the same initial source as the Universal Latent Workstation GroundTruth or NIST SD27 datasets, but is neither a subset nor superset of those.

3.1 Public Challenge Latent Dataset

This dataset contains 255 latent images from 214 subjects (distinct individuals). 173 subjects have one latent per subject; 41 subjects have two latents per subject.

The latent fingerprints were collected from case work in the mid-1990s and captured as photographic images. The physical photographs were rescanned in 2008,[1] resulting in these 1000ppi images.

Each latent image is provided with multiple markups to show inter-examiner variation. The majority of the images were marked up three times by IAI-certified latent examiners:

- by two examiners, each working alone;
- subsequently by a "jury" team of two other examiners based on a review of the individual markups.

Note that the feature markups were based solely on analysis of the latent image, as compared with the ULW GT/SD27 “Ideal” markup, which used both the latent and exemplar images to create a best-case feature markup. These feature markups therefore may be seen as more representative than the Ideal markup, but are also likely to be less accurate.

Feature markup in each file is saved as Extended Feature Set (EFS) fields, (fields 9.300-9.373) and as EFTS-LFFS features (fields 9.014-9.023, compliant with FBI EFTS 7.1). The EFTS-LFFS features were automatically converted from the EFS features, which is appropriate since EFS is a superset of EFTS-LFFS.

The Good/Bad/Ugly quality designation from ULW GT/SD27 is retained in these files and has not been changed.

3.2 Public Challenge Exemplar Dataset

Corresponding (mated) exemplars

202 of the 214 subjects have rolled and plain (slap) exemplars available as 1000ppi images of inked paper cards. The slap images are not segmented into separate fingers. Each of these 1000ppi exemplar images is also included as a 500ppi image.

111 of the subjects have more than one exemplar set per subject (up to 18 sets per subject). The multiple exemplar sets are only available as 500ppi images, include both rolled and slap images, and include a mix of inked paper and livescan originals.

Background (unmated) exemplars

This dataset includes an additional 214 subjects for use as background. The same images were rescanned for the 500ppi and 1000ppi datasets.

4 Format of results

4.1 Candidate Lists

All searches shall return a candidate list. A candidate list has a fixed length of one hundred (100) candidates. Note that a given search may be associated with zero, one, or more subjects in the gallery, and the candidate list shall include all of them.

[1] *The latents were scanned at 2000ppi, 16-bpp grayscale and downsampled to 1000ppi, 8-bpp grayscale.*

The candidate list consists of two parts, a required and an optional part.

The required part consists of:

- the index of the mating exemplar subject
- the matching finger number
- the absolute matching score
- an estimate of the probability of a match (0 to 100)

The optional part consists of:

- the number of good minutiae identified in the latent
- the number of latent minutiae which were successfully matched
- the quality estimate of the latent (0 to 100, 100 is best)
- the quality estimate of the candidate (0 to 100, 100 is best)

4.2 Timing

In addition, timing information for exemplar enrollment and latent search was reported as "wall clock" elapsed time (not CPU time) measurements, including the time to retrieve, process, and output all test data and results.

5 Rank-based results by subtest

Overall accuracy results are presented in this section using rank-based metrics via Cumulative Match Characteristic (CMC) curves. A CMC curve shows how many latent images are correctly identified at rank 1, rank 2, etc. A CMC is a plot of identification rate (or hit rate) vs. recognition rank. Identification rate at rank *k* is the proportion of the latent images correctly identified at rank K or lower. A latent image has rank *k* if its mate is the kth largest comparison score on the candidate list. Recognition rank ranges from 1 to 100, as 100 was the (maximum) candidate list size specified in the API.

The results in this section are based on the 1000-ppi exemplars; the 500-ppi exemplars show very similar results, as shown in Section 8.

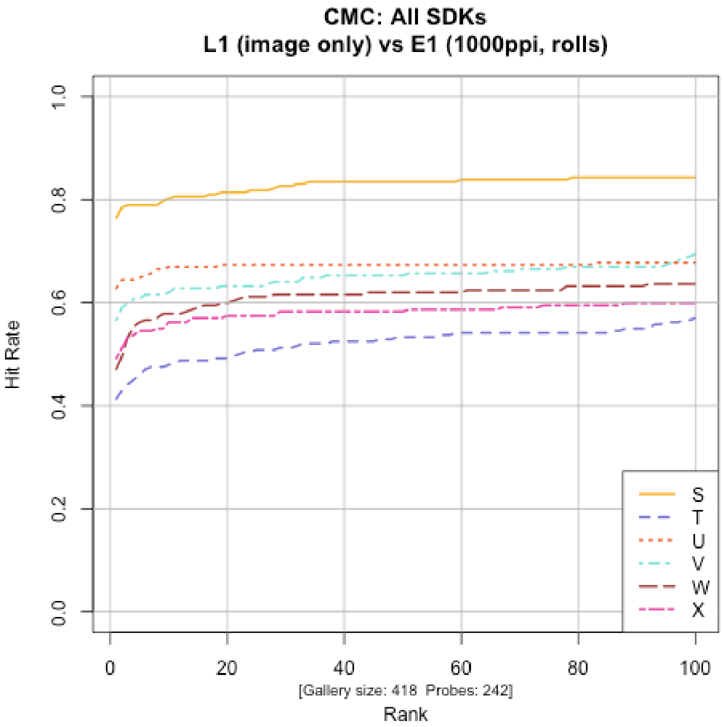
Note that not all participants returned results for all tests.

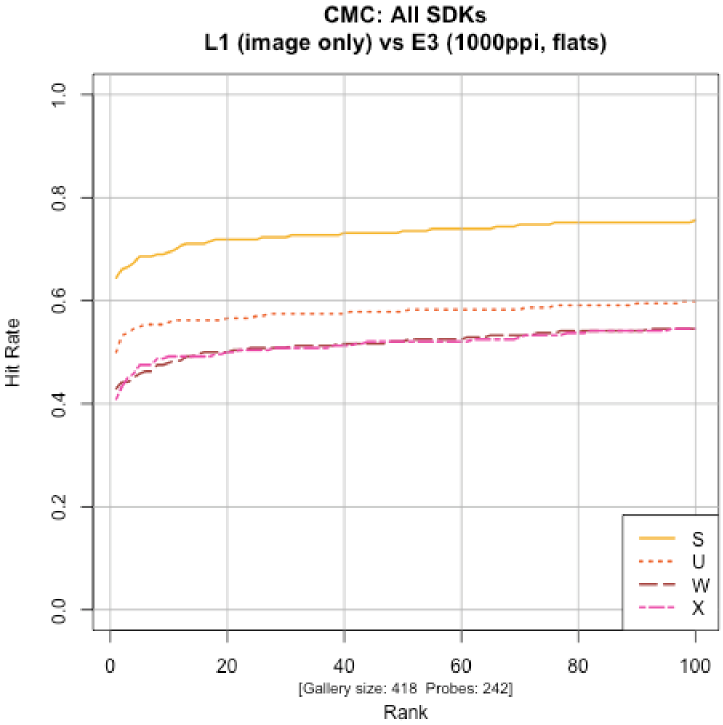
Table 1: Summary of rank-1 identification rates

		Participant					
		S	T	U	V	W	X
E1: 1000ppi rolled	L1: Image only	0.764	0.413	0.628	0.566	0.471	0.492
	L2: Image + IAFIS	0.754	-	0.665	0.639	0.538	-
	L3: Image + EFS	0.868	-	0.779	0.648	0.663	-
	L4: EFS only	0.808	0.284	0.775	0.483	0.654	-
	L5: IAFIS only	0.576	-	0.460	0.481	0.396	-
E3: 1000ppi slap	L1: Image only	0.645	-	0.500	-	0.430	0.409
	L2: Image + IAFIS	0.653	-	0.517	-	0.440	-
	L3: Image + EFS	0.754	-	0.627	-	0.527	-
	L4: EFS only	0.663	-	0.588	-	0.507	-
	L5: IAFIS only	0.467	-	0.376	-	0.279	-

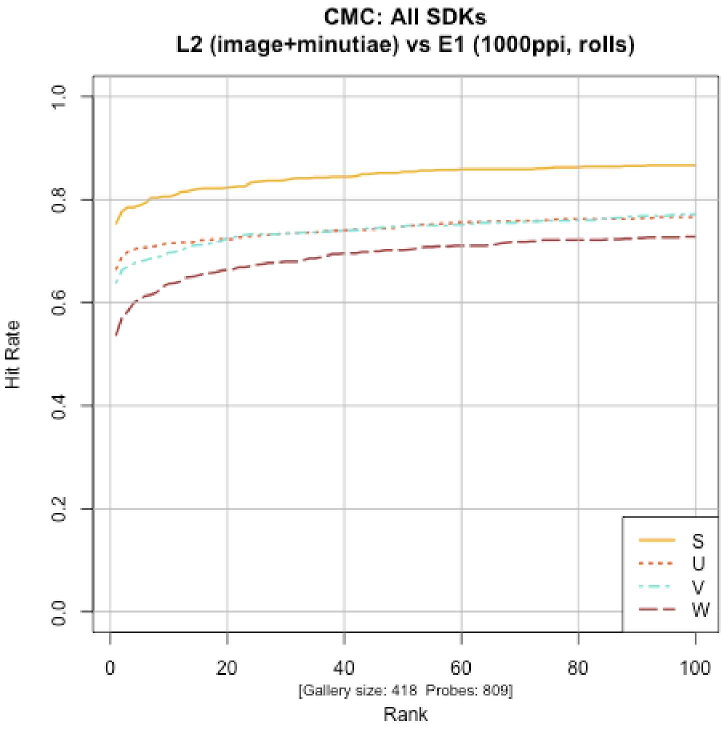
The gallery size was 418 for subtests E1 and E3, and 857 for subtests E2 and E4 (including multiple exemplar sets per subject). There were 242 distinct latent images with multiple markups per image, so there were 242 probes for the image-only subtest (L1), and 809 probes for the other subtests.

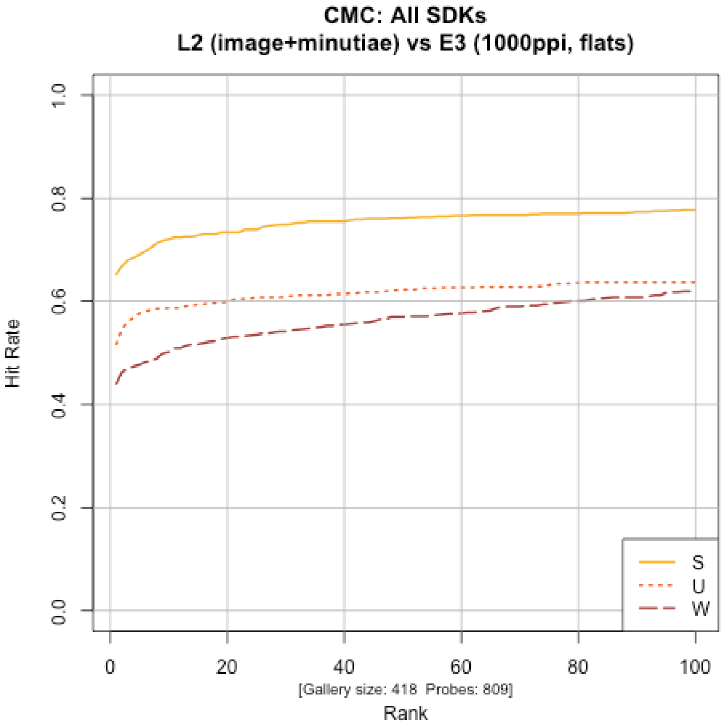
5.1 L1 – Image only



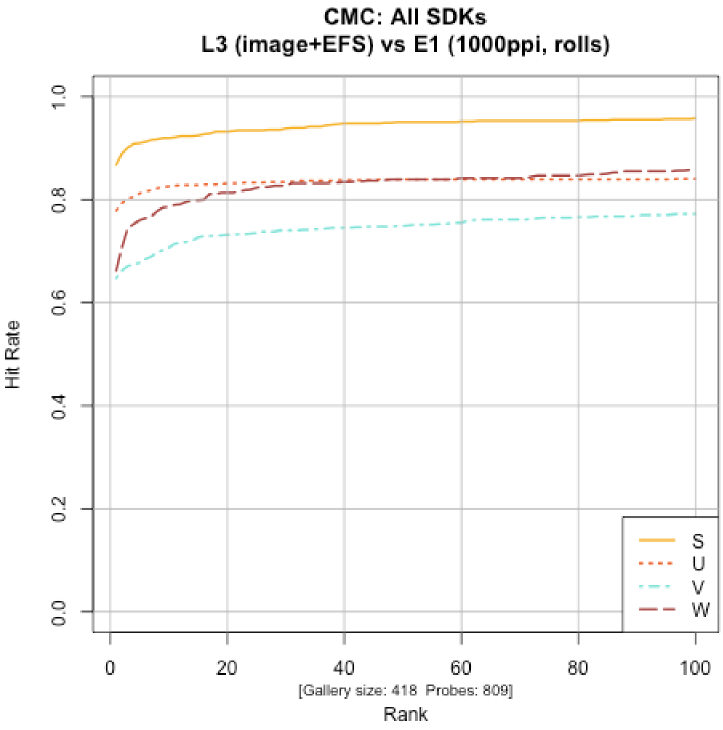


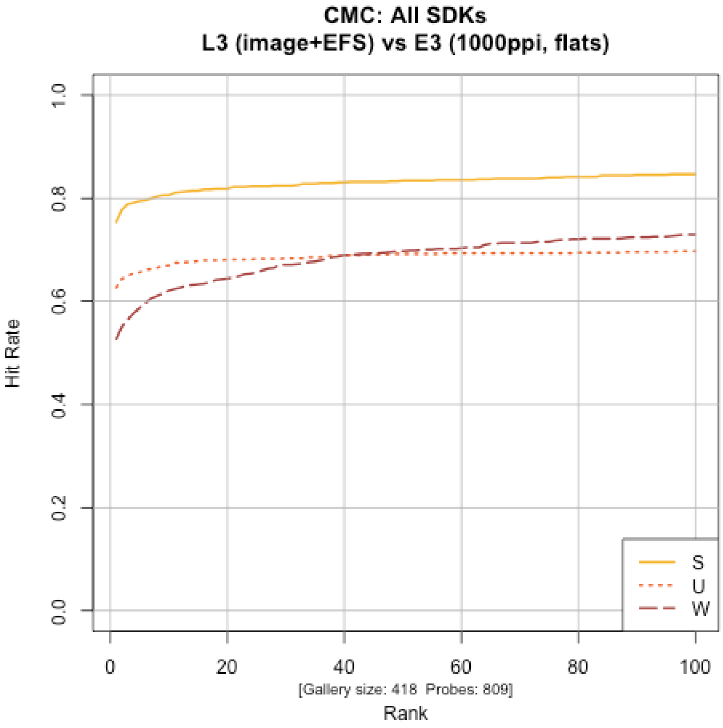
5.2 L2 – Image + IAFIS LFFS markup



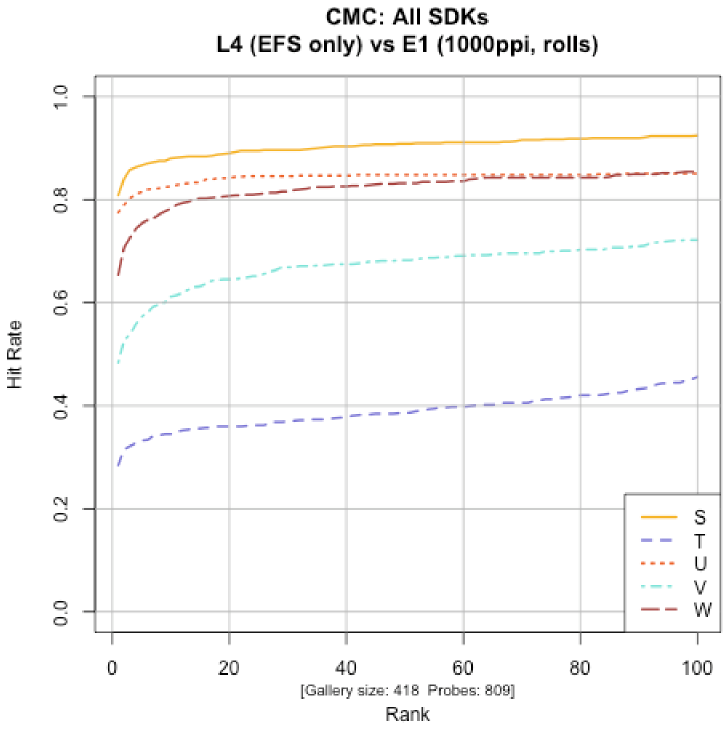


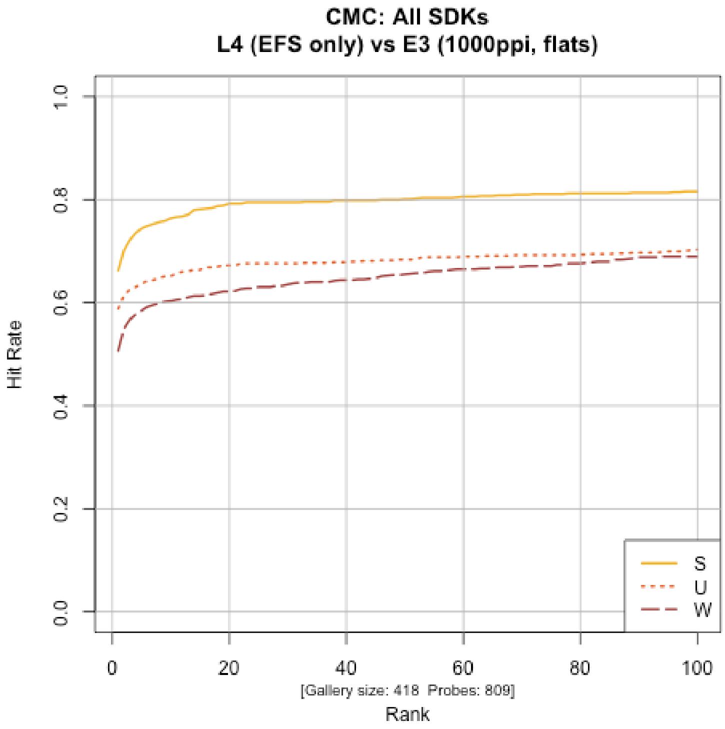
5.3 L3 – Image + Extended Feature Set markup



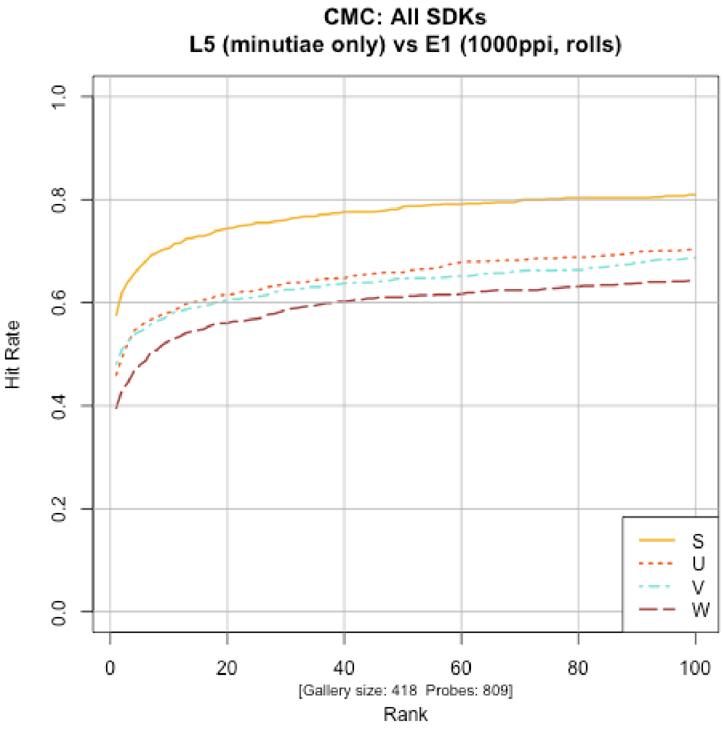


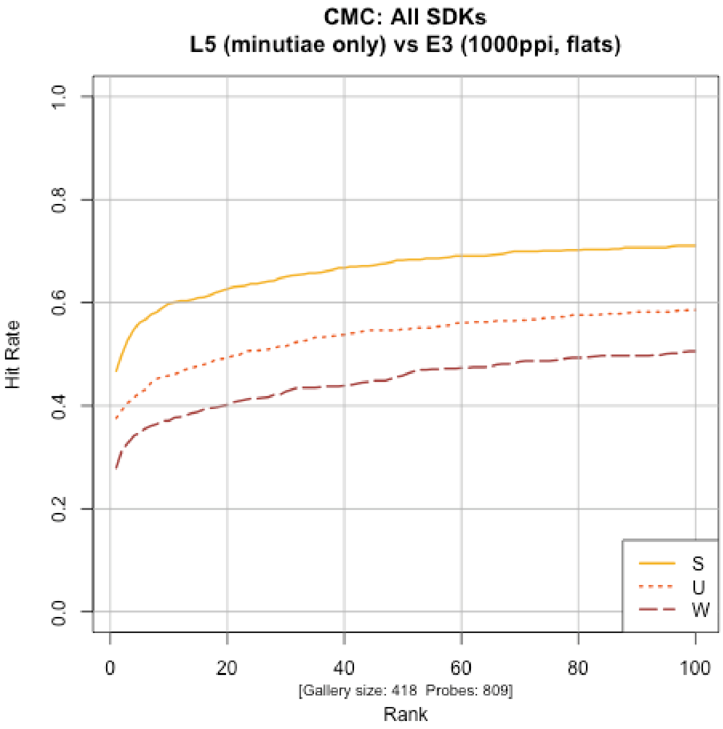
5.4 L4 – Extended Feature Set markup (no image)





5.5 L5 – IAFIS LFFS markup (no image)

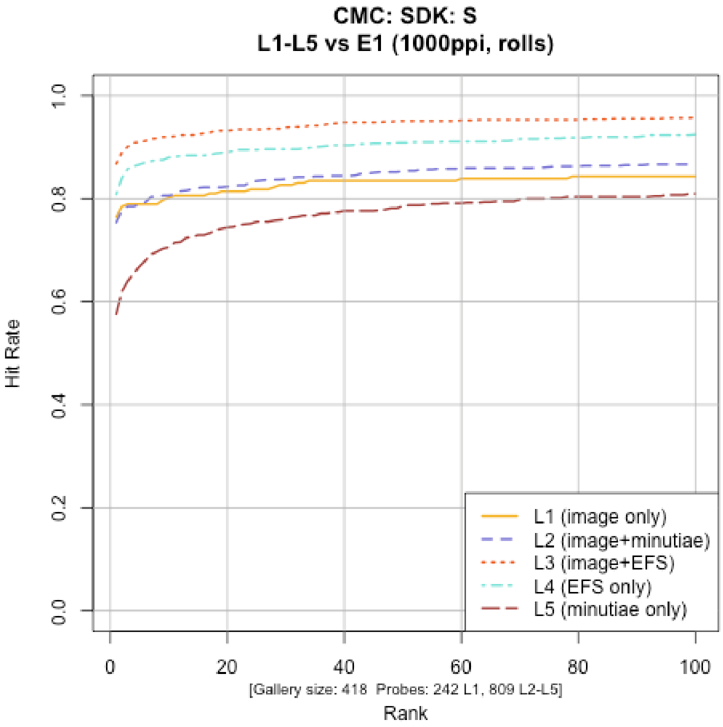


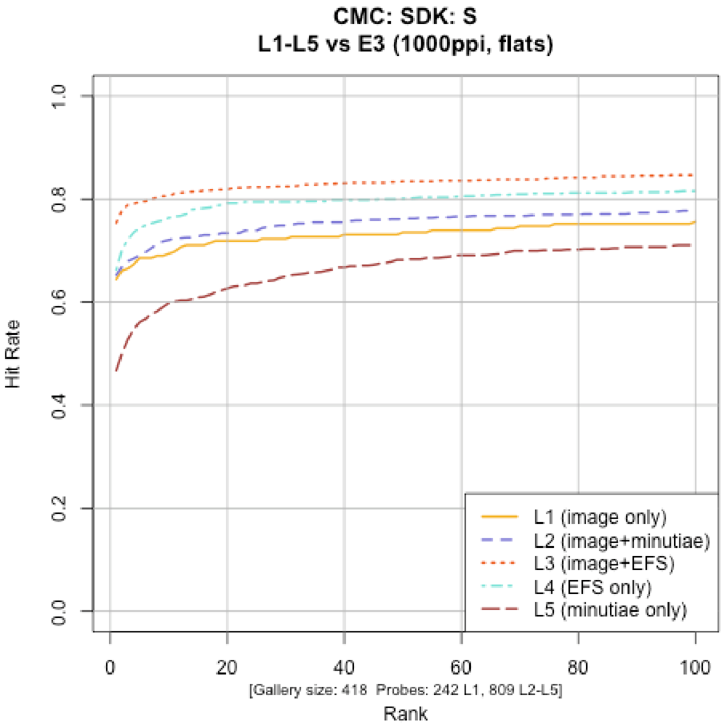


6 Results by participant

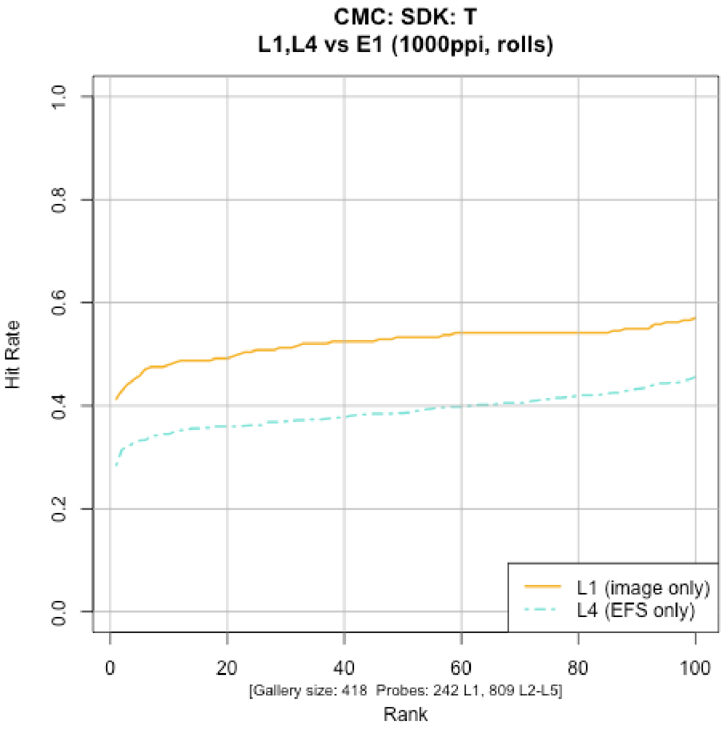
This section reports the same results as the previous section, but with charts grouped by participant.

6.1 Participant S

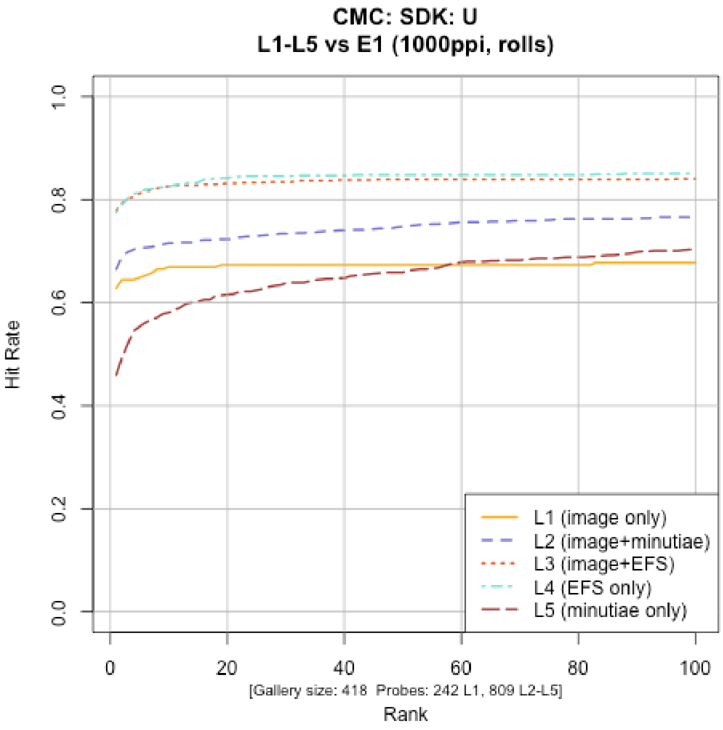


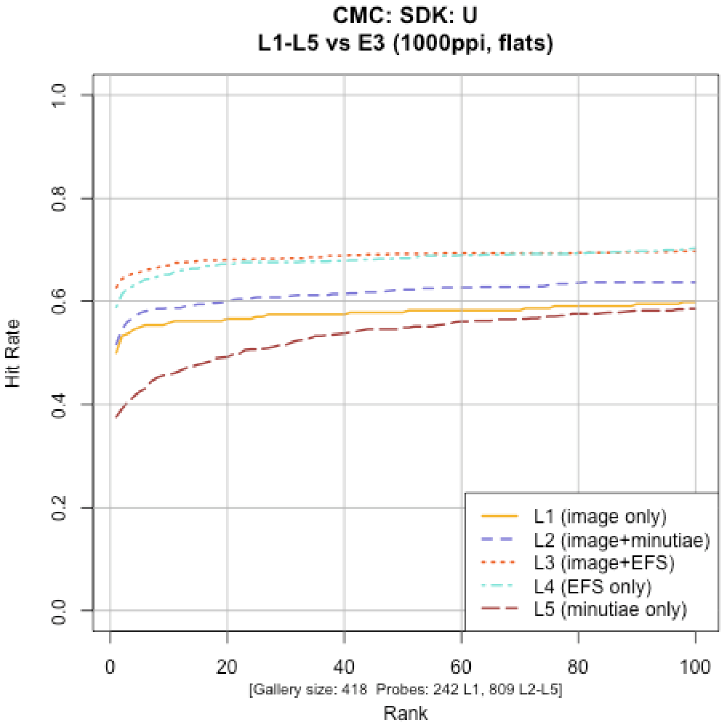


6.2 Participant T

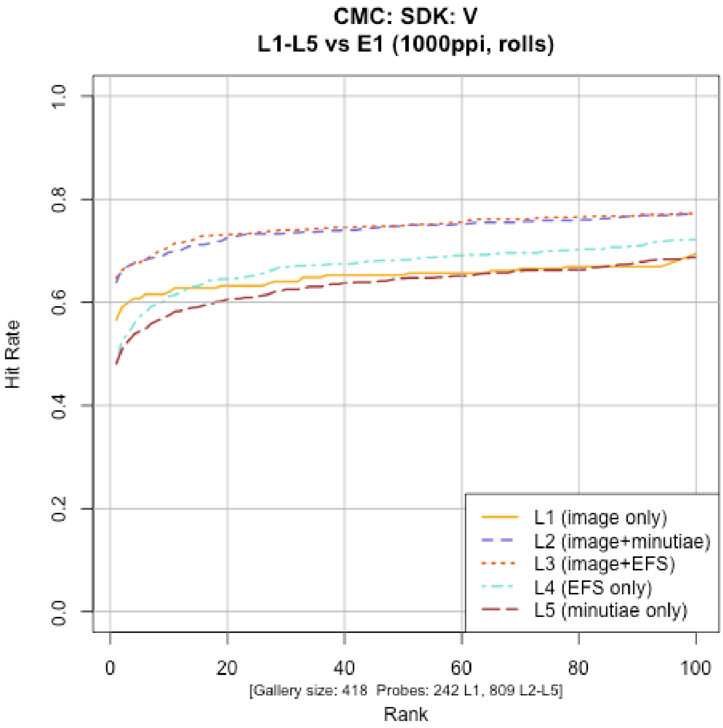


6.3 Participant U

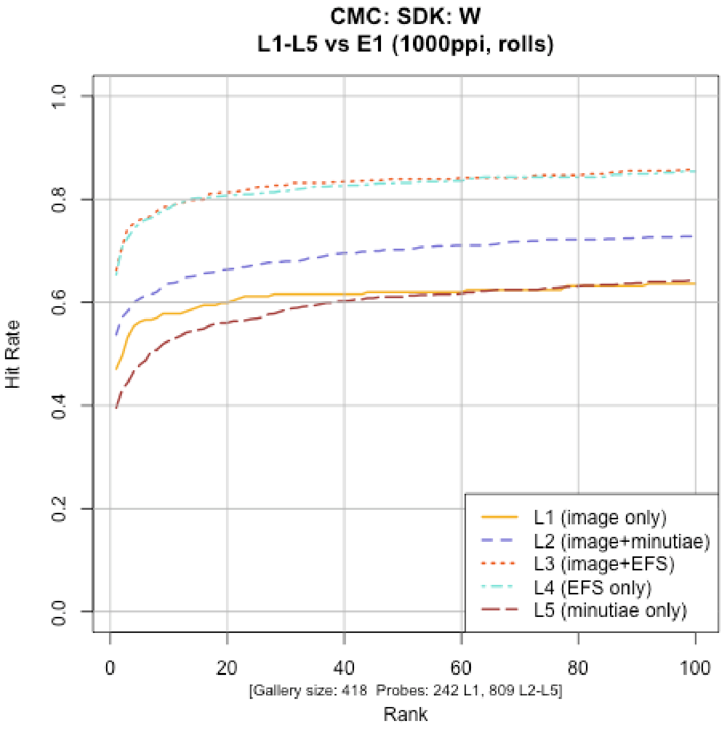


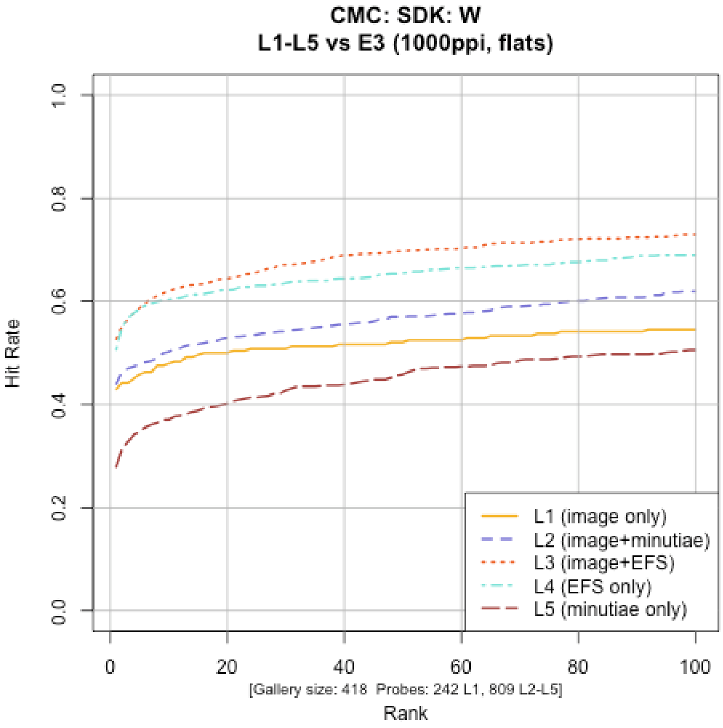


6.4 Participant V

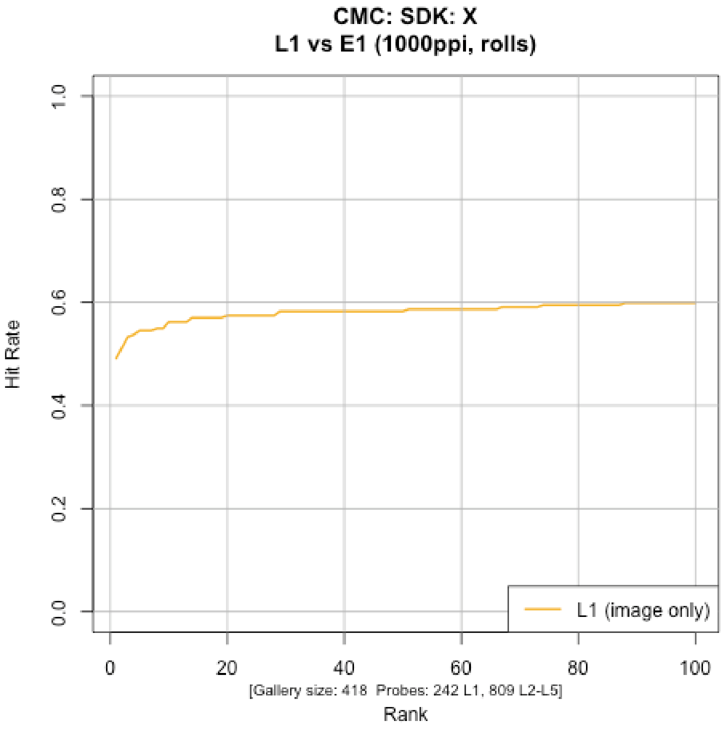


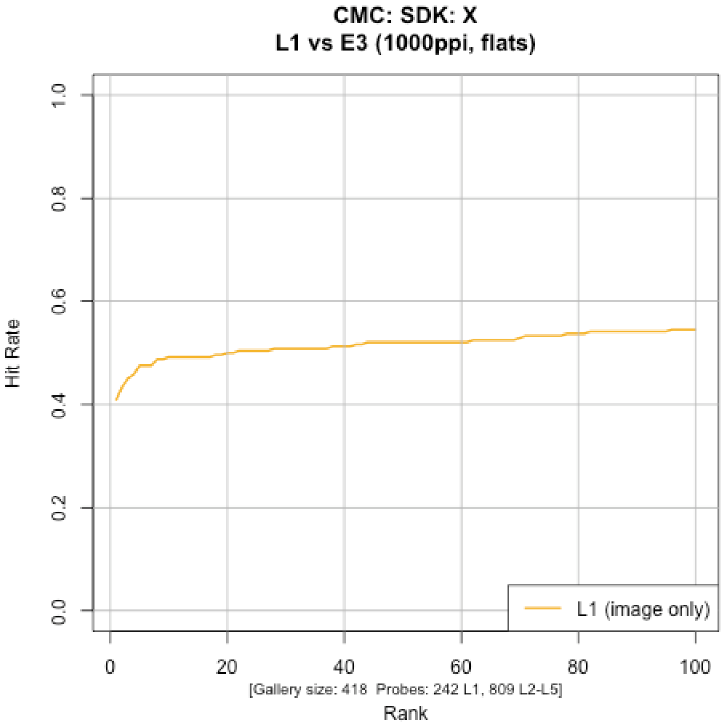
6.5 Participant W





6.6 Participant X





7 Multi-encounter [2]

For the 500ppi exemplars, 112 of the 213 subjects had more than one exemplar set per subject, as outlined in the following table. The multiple exemplar sets included a mix of inked paper and livescan originals. In the Public Challenge, the participants were instructed to treat the exemplar sets as if they were all from different subjects.

Exemplar sets per subject at 500ppi	Count
0	10
1	91
2	34
3	25
4	11
5	11
6	4
7	4
8	4
9	4
10	4
11	3
12	2
13	2
14	1
15	1
16	0
17	1
18	1
Total	213

In comparing the multi-exemplar results, three methods were used to assess performance. In each case, only one exemplar per subject was selected from the candidate list, and the others were ignored. (The same selection method was used for mated or background gallery subjects)

Baseline

The selected exemplar set is a 500ppi subsample of the 1000ppi exemplar set. This shows the effect if the gallery only contained a single (arbitrary) exemplar set per subject.

Best NFIQ

The selected exemplar set is a composite record containing the highest-quality image available for each finger position, as measured by NFIQ.[3] This shows the effect of an often-used operational approach.

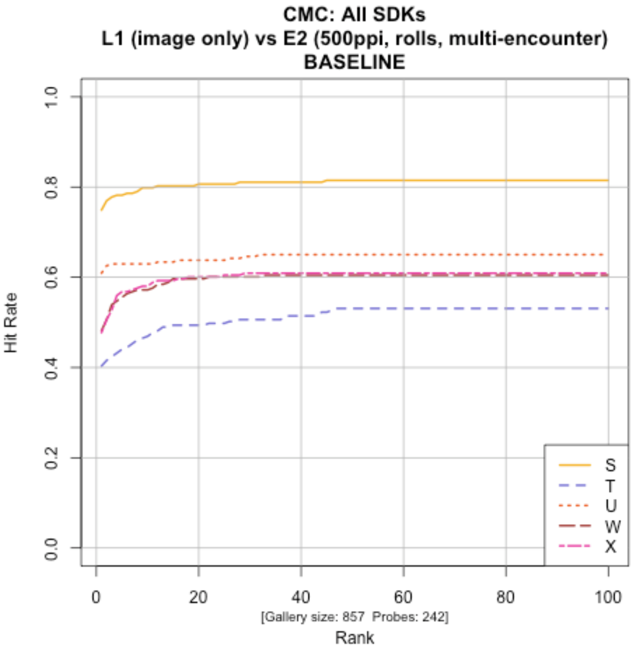
Best Rank

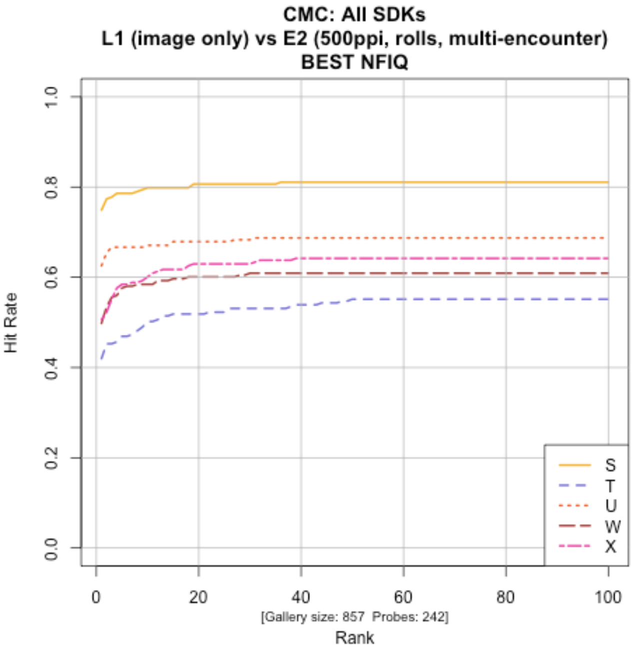
The selected exemplar is simply the highest-ranking result returned for each subject:finger combination. This shows the effect of retaining all exemplars in the gallery and using the highest-scoring results.

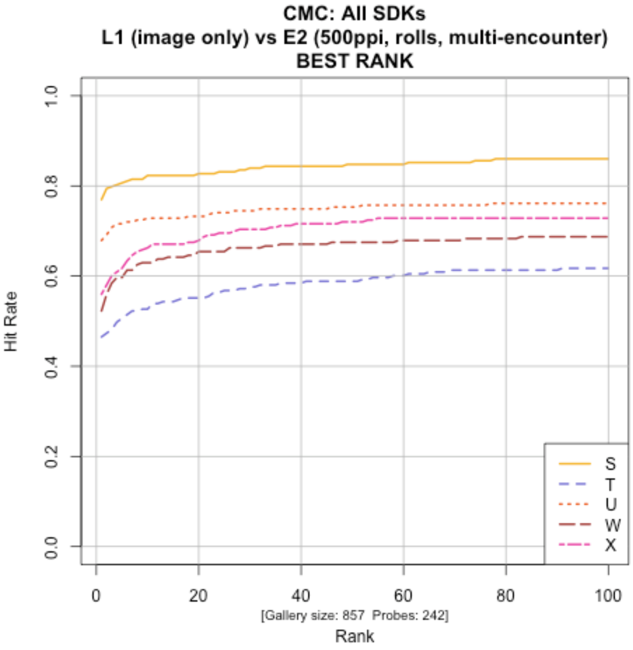
[2] *Ed. Note: the multi-exemplar 500ppi charts in this draft show curves that extend all the way to rank 100. Because of the pruning approach used to handle multi-exemplar data, the candidate lists were almost always reduced to fewer than 100 candidates (generally about 50). In subsequent reporting, this will be corrected.*

[3] *NIST Fingerprint Image Quality*

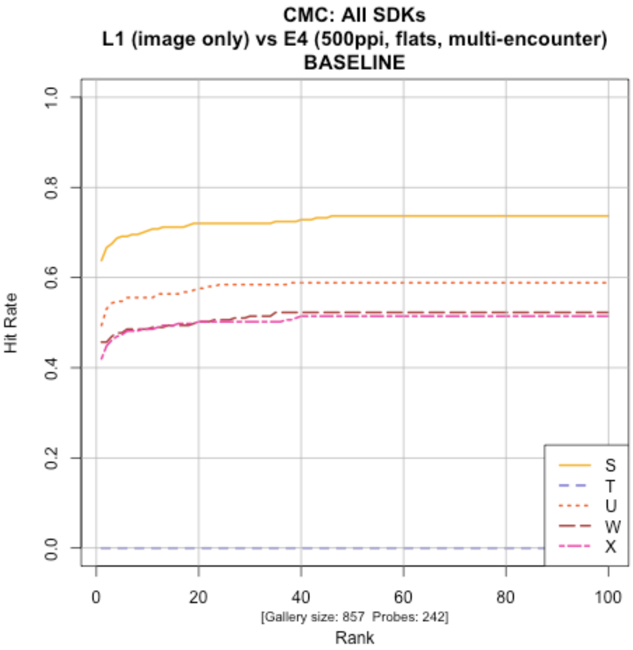
7.1 L1 (Image only) x E2 (500ppi rolls)

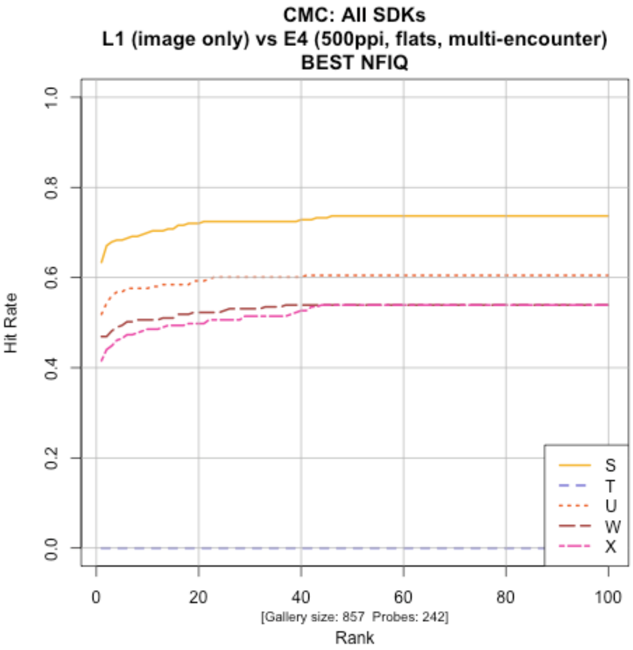


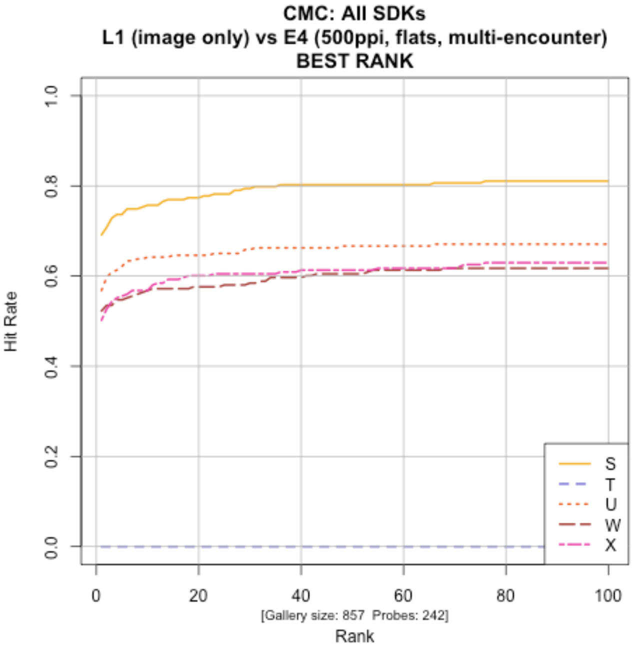




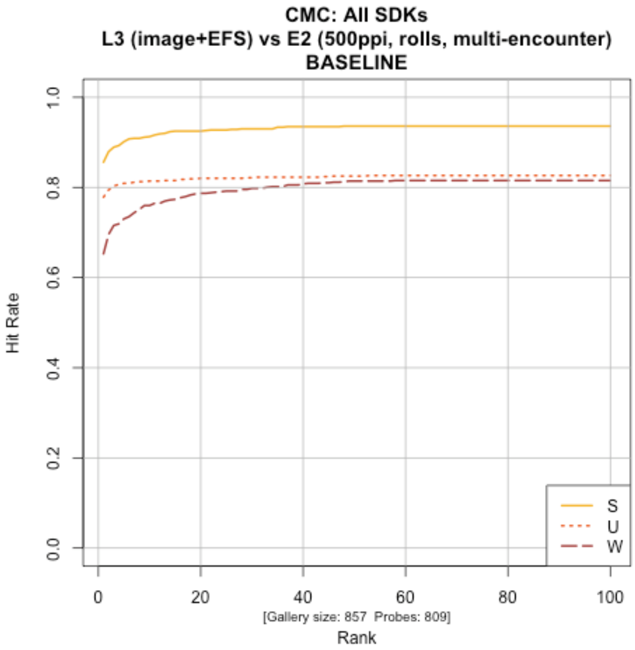
7.2 L1 (Image only) x E4 (500ppi flats)

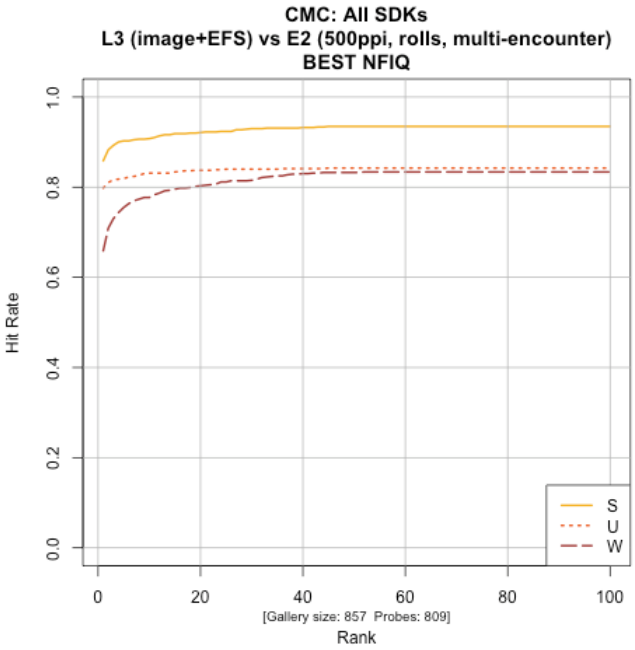


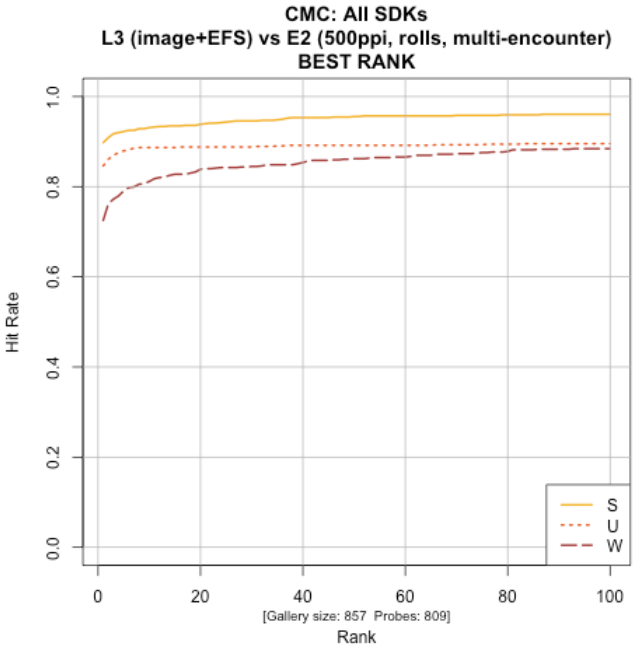




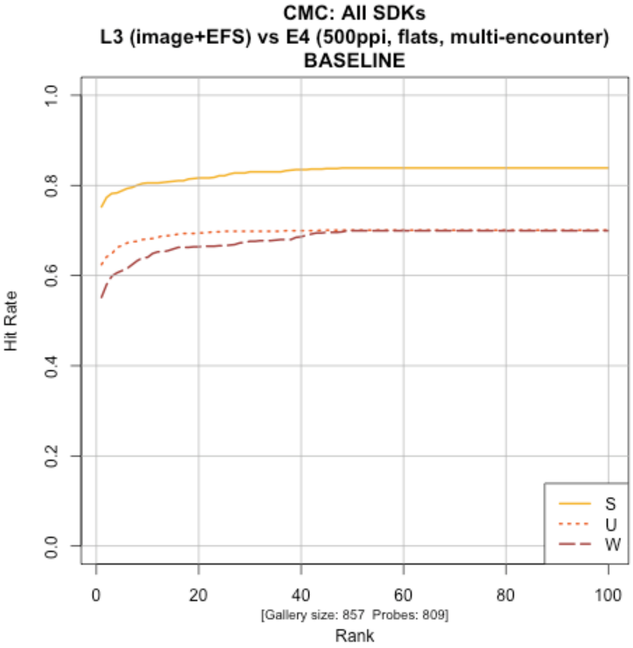
7.3 L3 (Image + Extended Feature Set markup) x E2 (500ppi rolls)

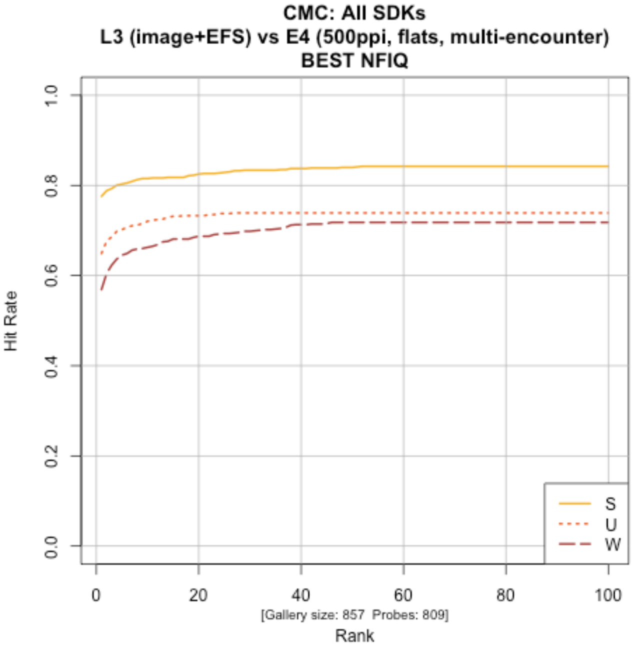


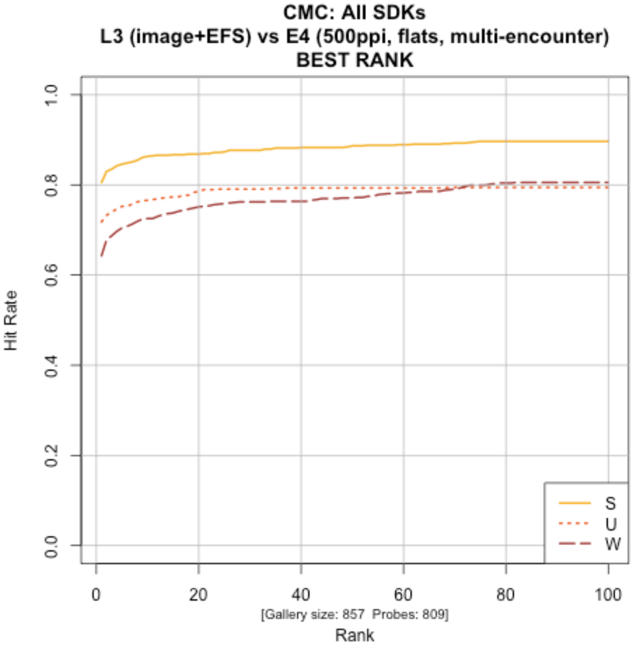




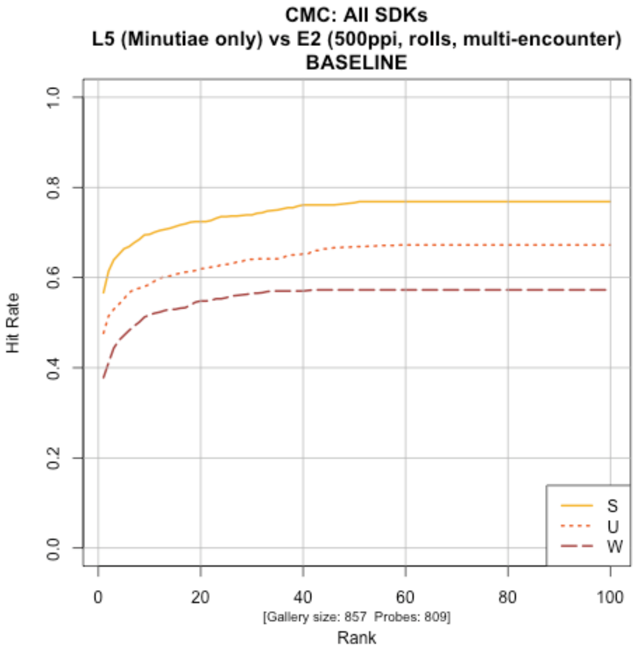
7.4 L3 (Image + Extended Feature Set markup) x E4 (500ppi flats)

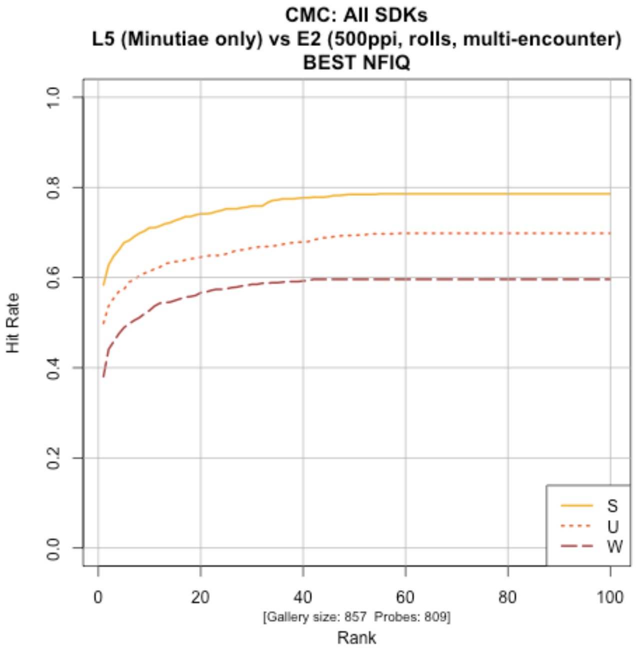


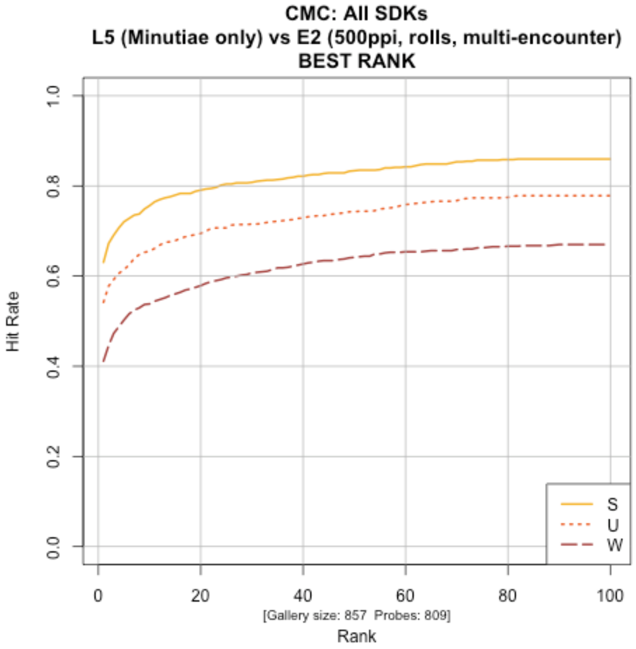




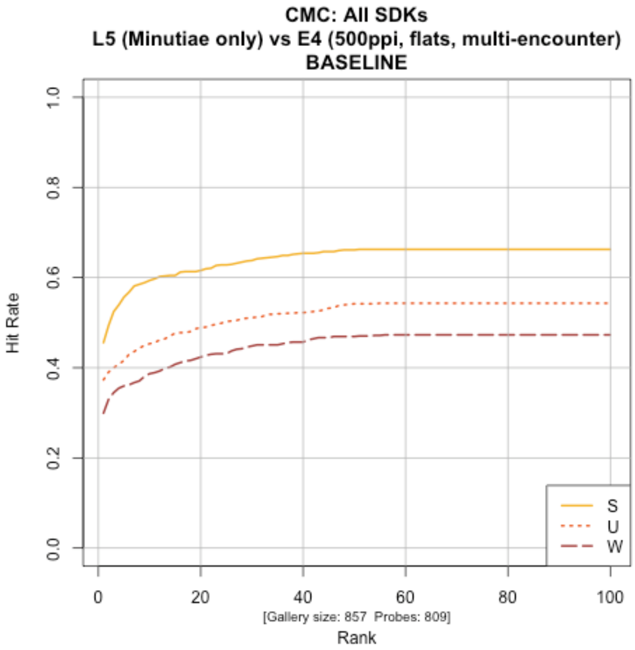
7.5 L5 (IAFIS LFFS markup, no image) x E2 (500ppi rolls)

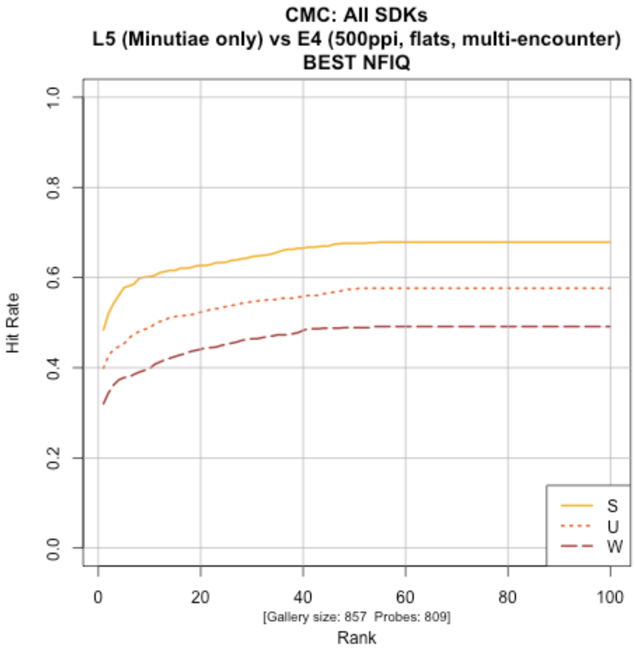


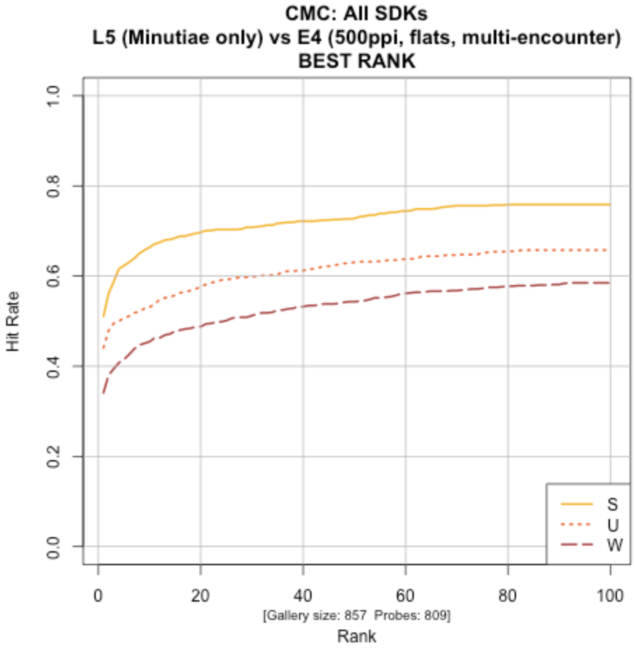




7.6 L5 (IAFIS LFFS markup, no image) x E4 (500ppi flats)



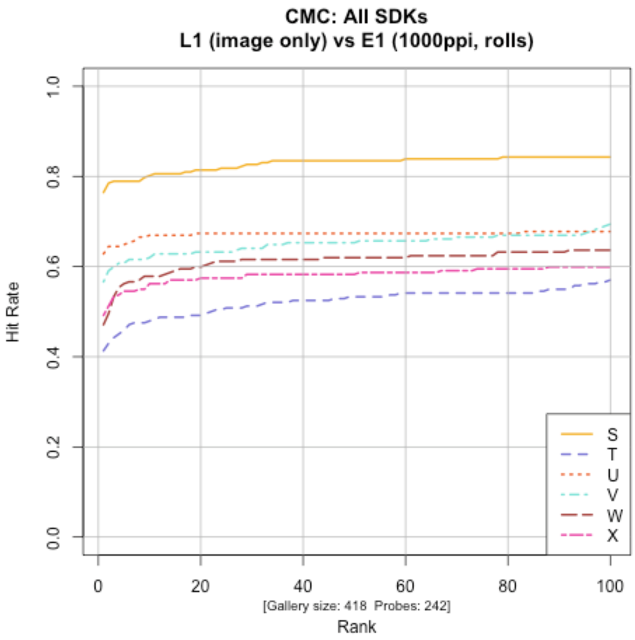


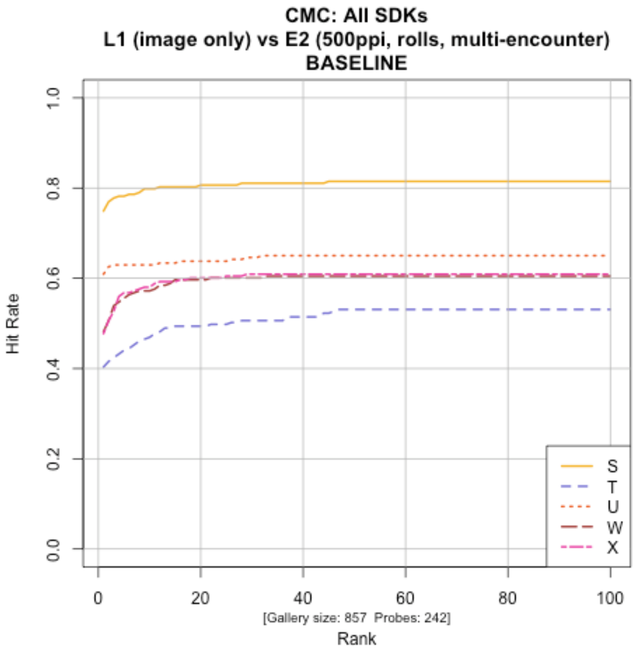


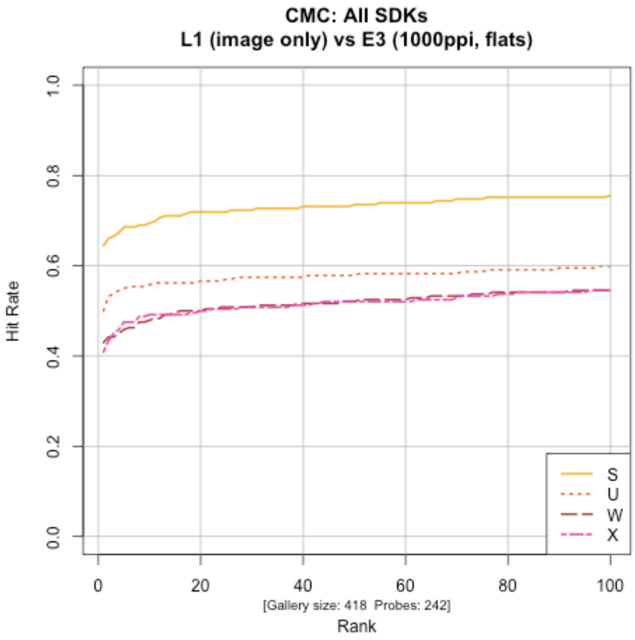
8 Resolution

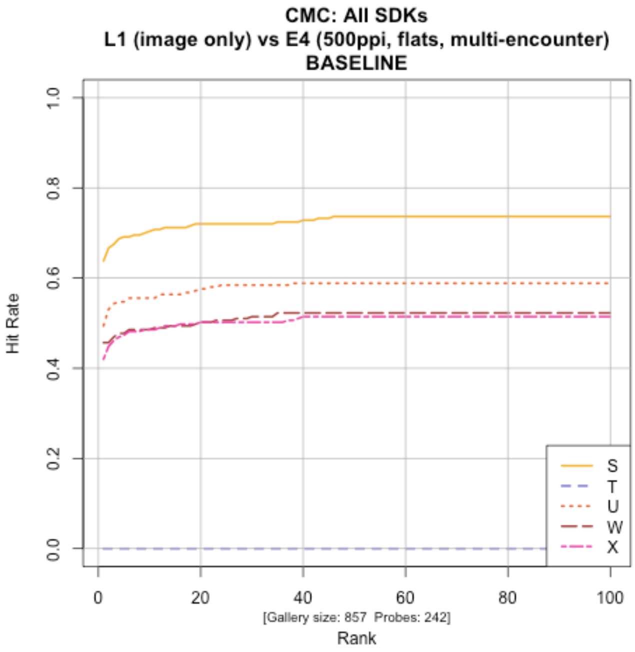
The following charts (repeated from elsewhere in this report) compare the effects for 1000ppi and 500ppi images.

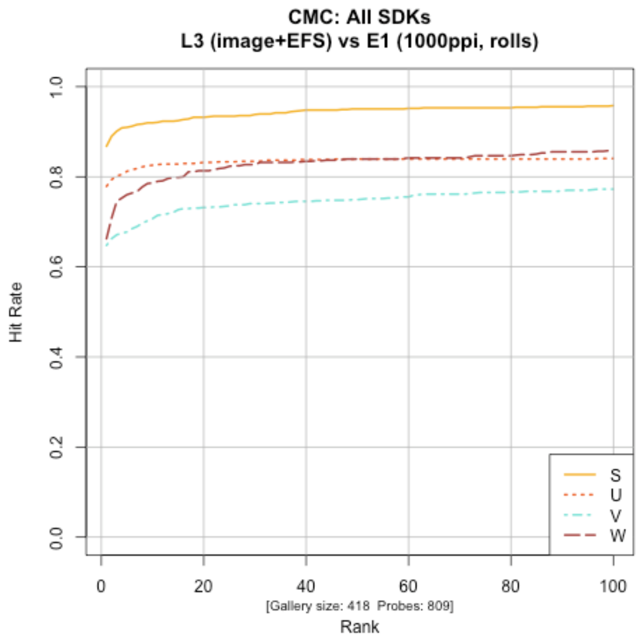
Ed. Note: the multi-exemplar 500ppi charts in this draft show curves that extend all the way to rank 100. Because of the pruning approach used to handle multi-exemplar data, the candidate lists were almost always reduced to fewer than 100 candidates (generally about 50). In subsequent reporting, this will be corrected. In these charts, differences to the right of about rank 50 should be ignored. [TBD]



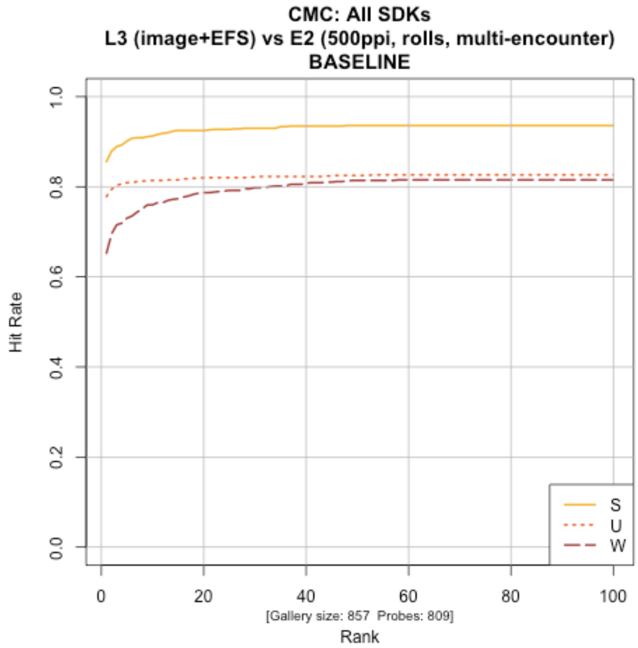




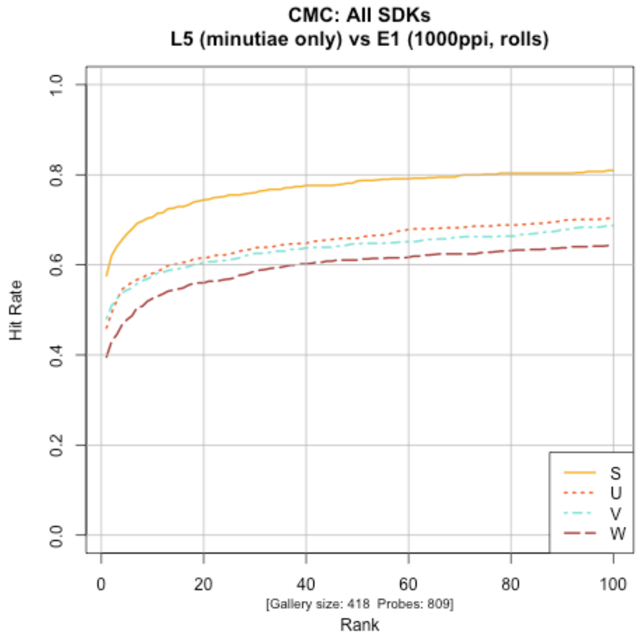




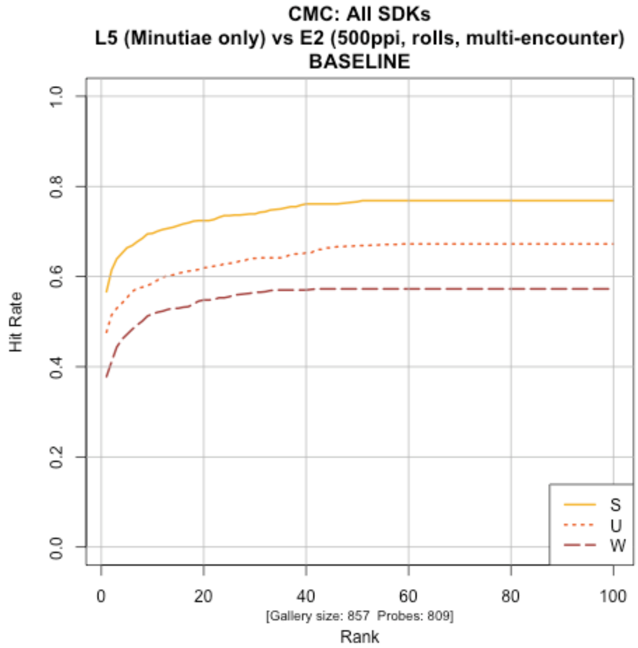
CMC: All SDKs
L3 (image+EFS) vs E1 (1000ppi, rolls)



CMC: All SDKs
L3 (image+EFS) vs E2 (500ppi, rolls, multi-encounter)
BASELINE



CMC: All SDKs
L5 (minutiae only) vs E1 (1000ppi, rolls)



CMC: All SDKs
L5 (Minutiae only) vs E2 (500ppi, rolls, multi-encounter)
BASELINE

9 Score-based results

The previous results reported rank-based identification performance. Here Detection Error Trade-off (DET) curves were plotted using the methodology defined in ELFT Phase II. [4] All DET curves in this analysis are limited to Rank 1 (limited to the highest scoring result in the candidate list).

As defined for ELFT Phase II,

- False Negative Identification Rate (FNIR) indicates the fraction of cases in which enrolled mates do not appear in the top position with a score greater than the threshold.

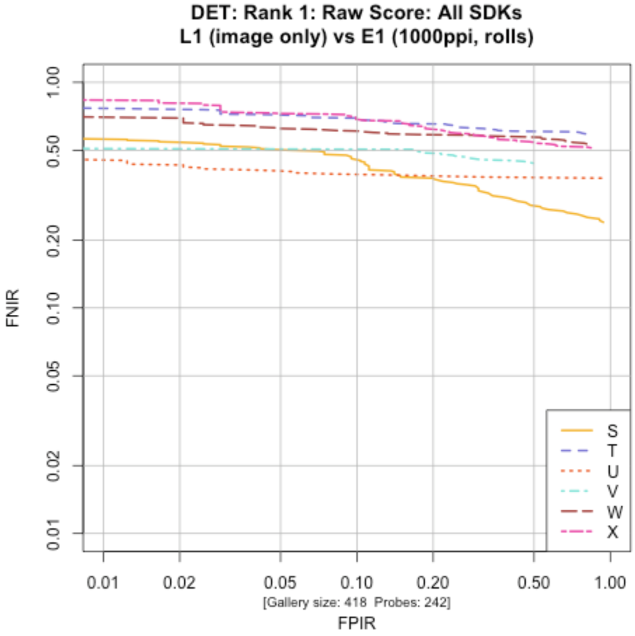
[4] *Indovina, et al; ELFT Phase II - An Evaluation of Automated Latent Fingerprint Identification Technologies; NISTIR 7577; Section 3.1.2 p 24.*

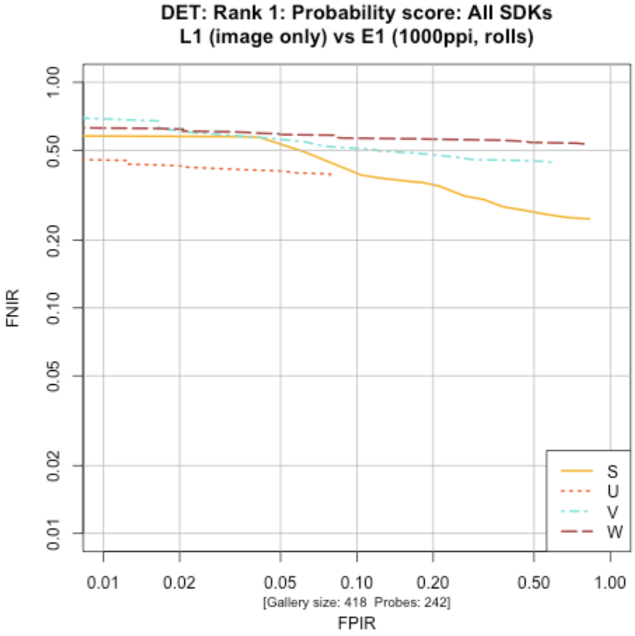
- False Positive Identification Rate (FPIR) indicates the fraction of candidate lists (without enrolled mates) that contain a non-mate entry in the top position with a score greater than the threshold.

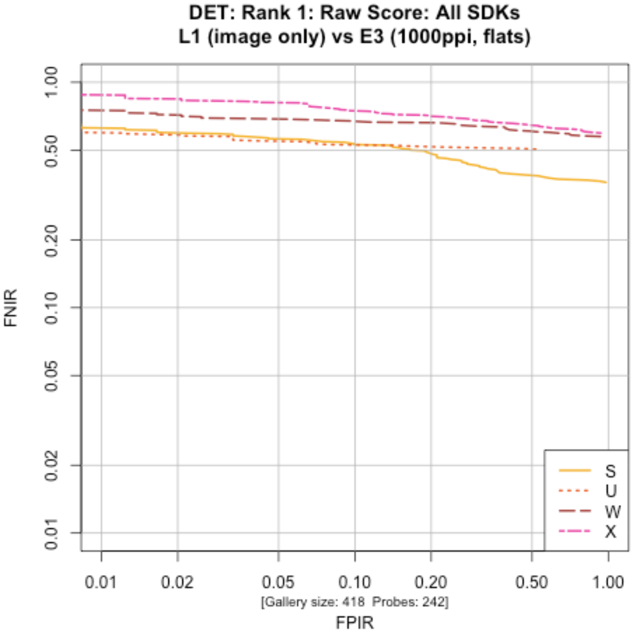
In practice, these charts show the effect of automatically eliminating candidates based on score. For example, in the first chart (Raw score DET for L1 vs E1), for participant S, the FNIR=0.236 @ FPIR=1.0, and reduces to FNIR=0.5 @ FPIR=0.05. What this means is that if a score threshold is set so that in 95% of cases no candidates are returned, the accuracy (1-FNIR) reduces from 76.4% to 50%. While (obviously) this is not acceptable for high-priority cases, this is of great interest for some uses such as reverse searches (unsolved latent processing), or automatic processing of low-priority cases.

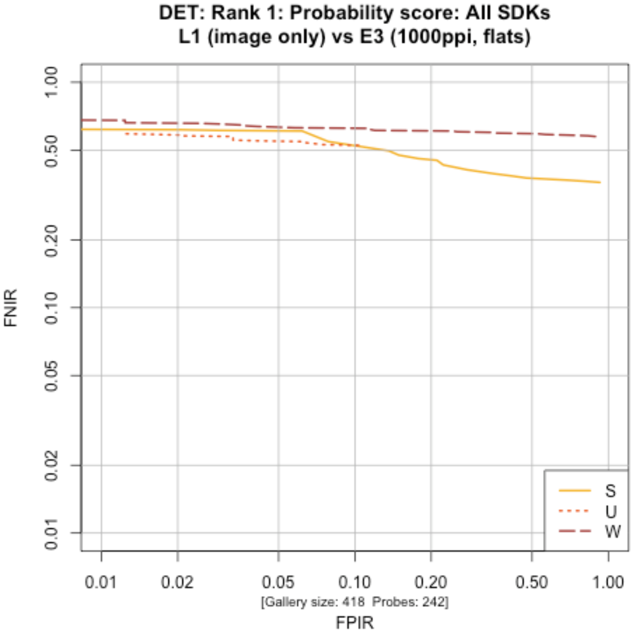
Note that a horizontal line is ideal, indicating no degradation in accuracy as non-mates are automatically excluded. Note also that when the FPIR=1.0, the raw score FNIR I the same as the rank-1 identification rate shown in Table 1 and the CMC analyses above.

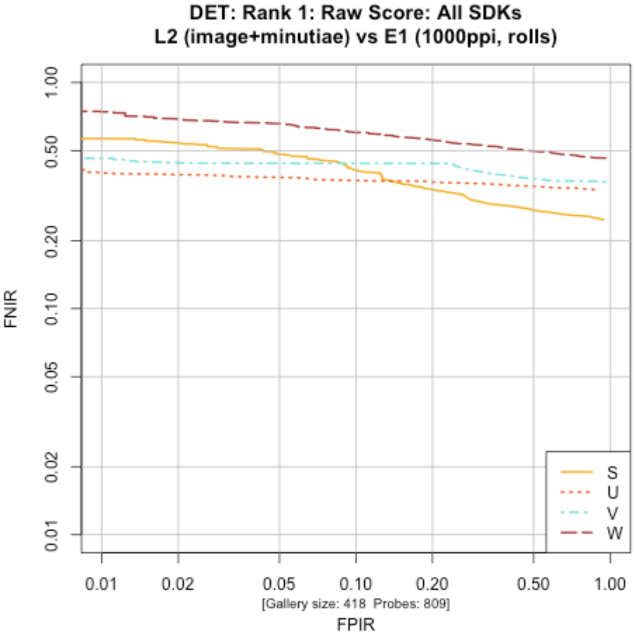
In each case, participants returned a raw score and a normalized score estimating the probability of a match. Not all participants returned probability scores.

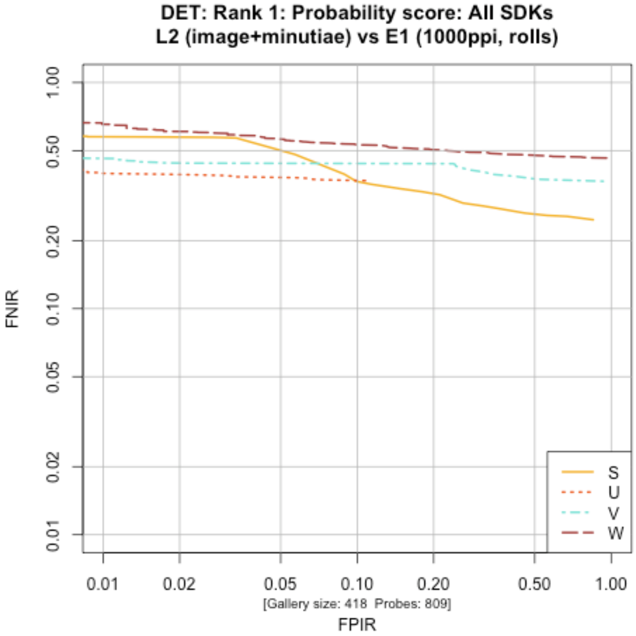

DET: Rank 1: Raw Score: All SDKs
L1 (image only) vs E1 (1000ppi, rolls)

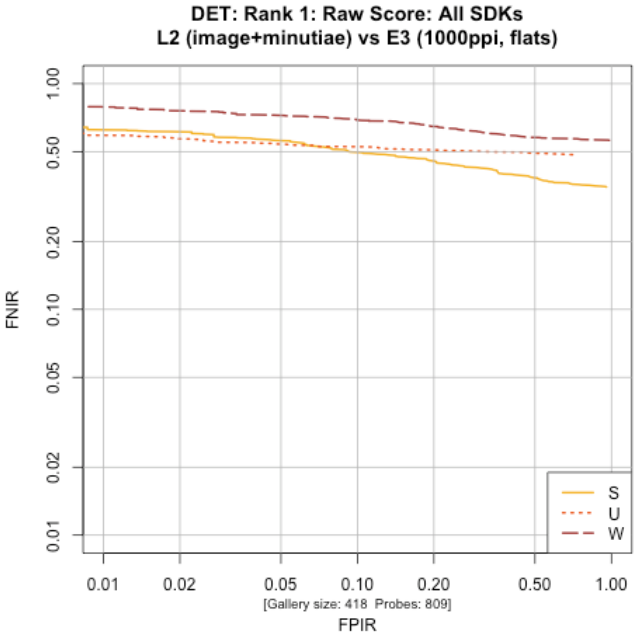

DET: Rank 1: Probability score: All SDKs
L1 (image only) vs E1 (1000ppi, rolls)


DET: Rank 1: Raw Score: All SDKs
L1 (image only) vs E3 (1000ppi, flats)
FNIR
1.00
0.50
0.20
0.10
0.05
0.02
0.01
0.01
0.02
0.05
0.10
0.20
0.50
1.00
[Gallery size: 418 Probes: 242]
FPIR
S
U
W
X

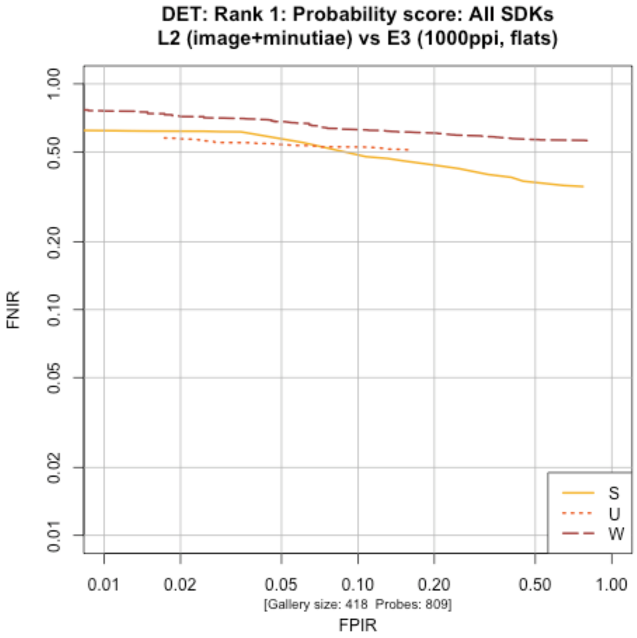

DET: Rank 1: Probability score: All SDKs
L1 (image only) vs E3 (1000ppi, flats)
FNIR
1.00
0.50
0.20
0.10
0.05
0.02
0.01
0.01
0.02
0.05
0.10
0.20
0.50
1.00
[Gallery size: 418 Probes: 242]
FPIR
S
U
W


DET: Rank 1: Raw Score: All SDKs
L2 (image+minutiae) vs E1 (1000ppi, rolls)
FNIR
1.00
0.50
0.20
0.10
0.05
0.02
0.01
0.01
0.02
0.05
0.10
0.20
0.50
1.00
[Gallery size: 418 Probes: 809]
FPIR
S
U
V
W

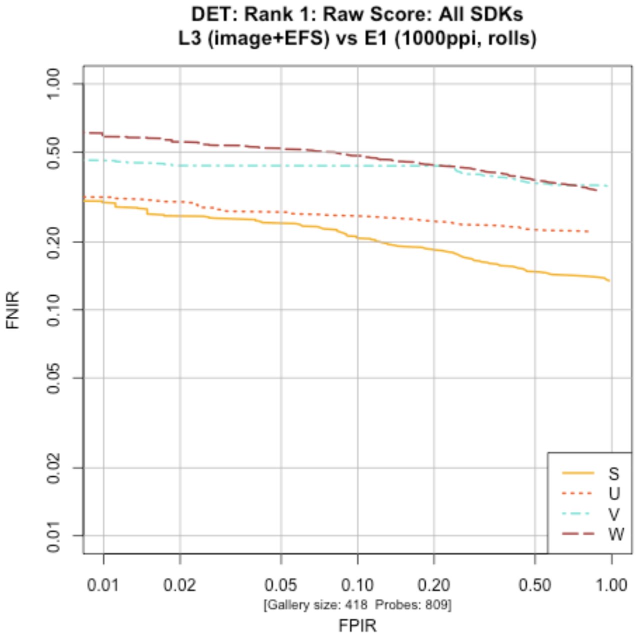

DET: Rank 1: Probability score: All SDKs
L2 (image+minutiae) vs E1 (1000ppi, rolls)
FNIR
1.00
0.50
0.20
0.10
0.05
0.02
0.01
0.01
0.02
0.05
0.10
0.20
0.50
1.00
[Gallery size: 418 Probes: 809]
FPIR
S
U
V
W



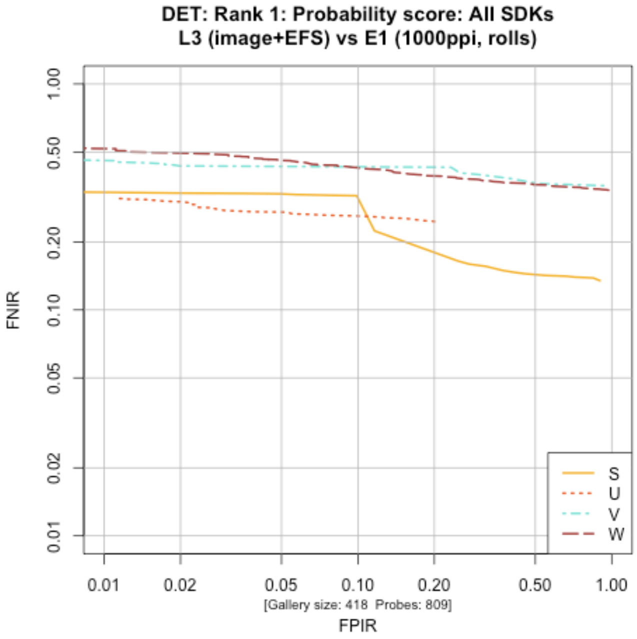
DET: Rank 1: Raw Score: All SDKs
L2 (image+minutiae) vs E3 (1000ppi, flats)



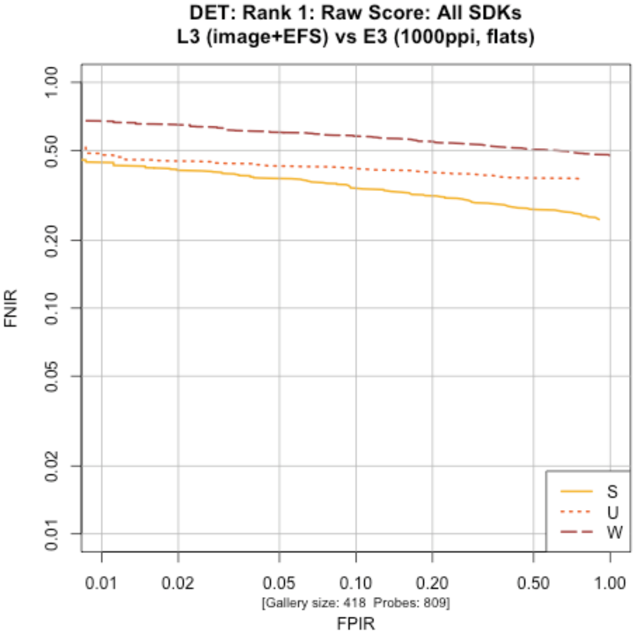
DET: Rank 1: Probability score: All SDKs
L2 (image+minutiae) vs E3 (1000ppi, flats)

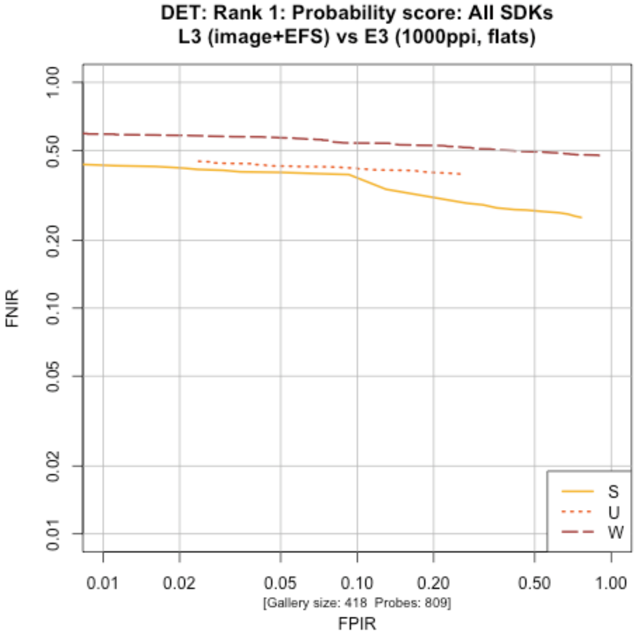


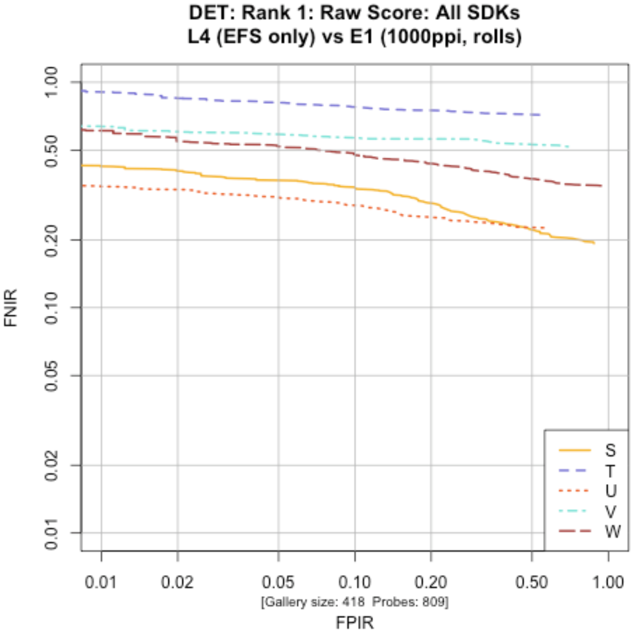
DET: Rank 1: Raw Score: All SDKs
L3 (image+EFS) vs E1 (1000ppi, rolls)

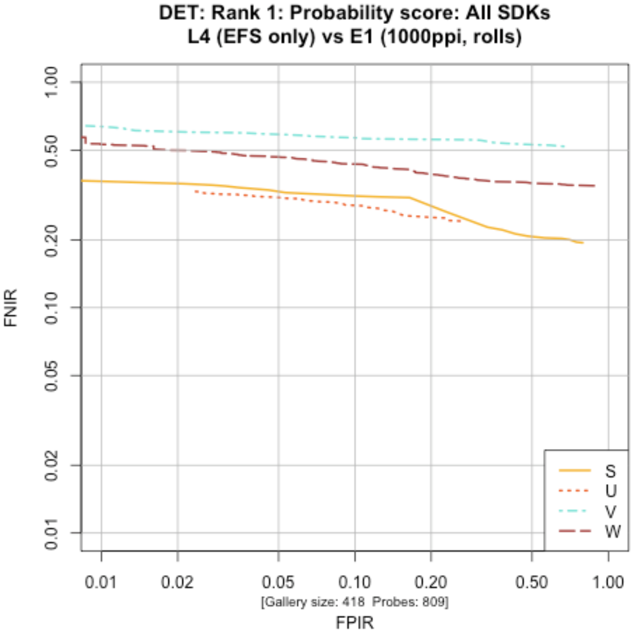


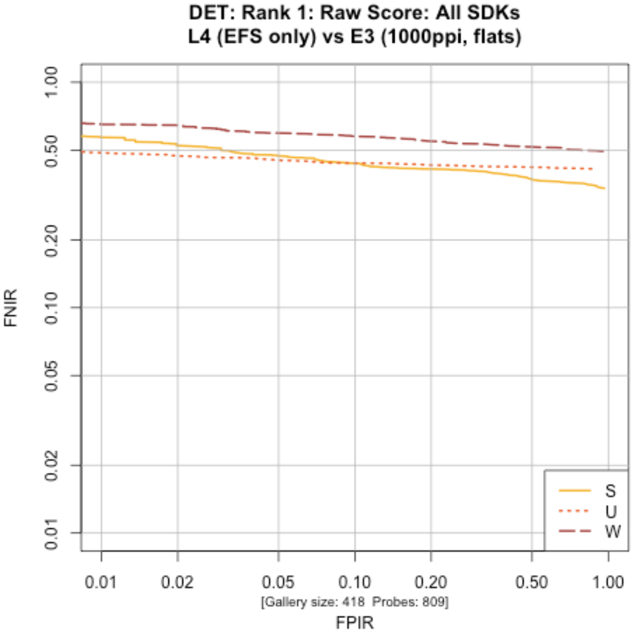
DET: Rank 1: Probability score: All SDKs
L3 (image+EFS) vs E1 (1000ppi, rolls)

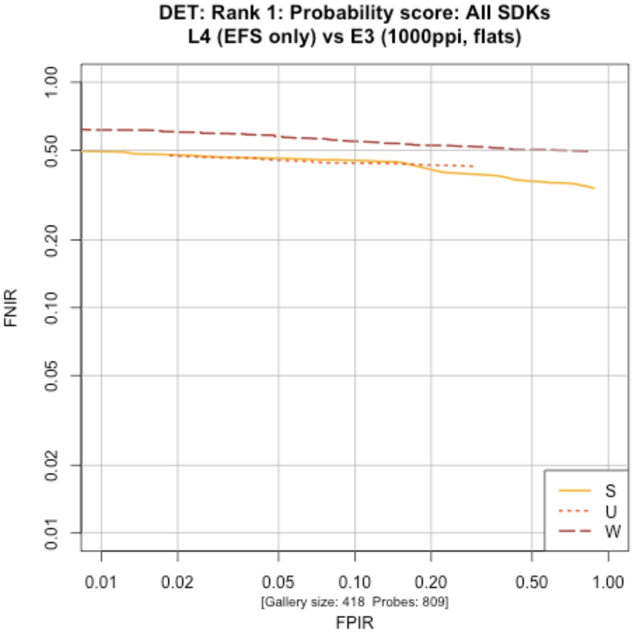

DET: Rank 1: Raw Score: All SDKs
L3 (image+EFS) vs E3 (1000ppi, flats)
FNIR
1.00
0.50
0.20
0.10
0.05
0.02
0.01
0.01
0.02
0.05
0.10
0.20
0.50
1.00
[Gallery size: 418 Probes: 809]
FPIR
S
U
W

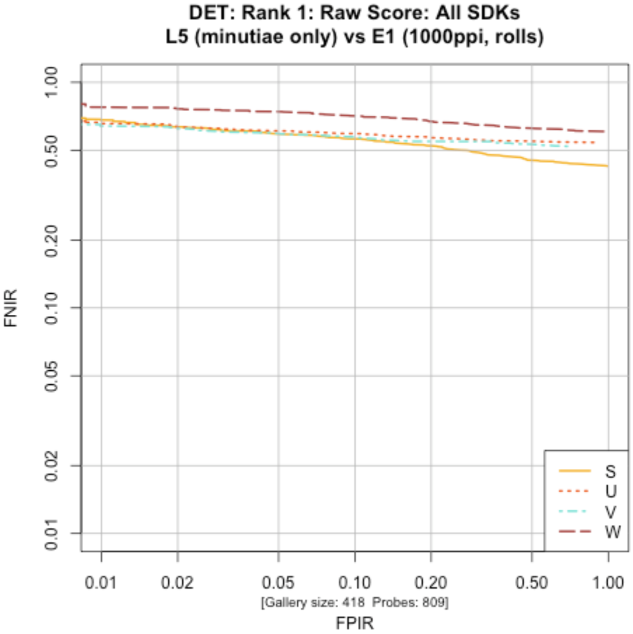

DET: Rank 1: Probability score: All SDKs
L3 (image+EFS) vs E3 (1000ppi, flats)
FNIR
1.00
0.50
0.20
0.10
0.05
0.02
0.01
0.01
0.02
0.05
0.10
0.20
0.50
1.00
[Gallery size: 418 Probes: 809]
FPIR
S
U
W

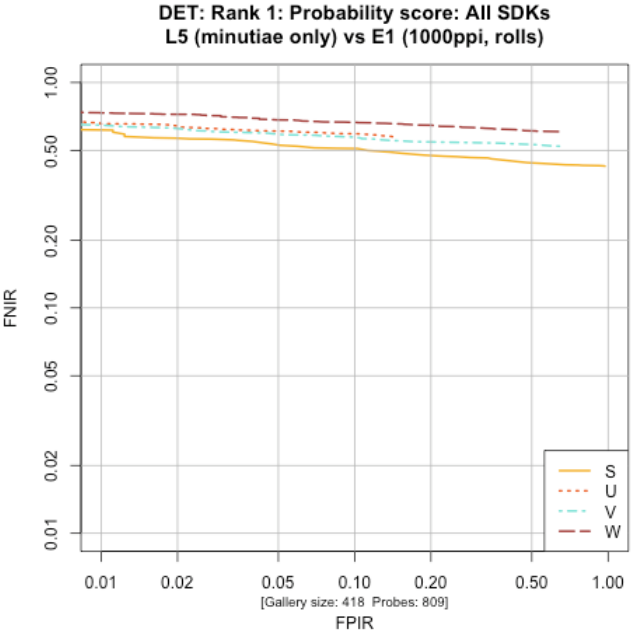

DET: Rank 1: Raw Score: All SDKs
L4 (EFS only) vs E1 (1000ppi, rolls)
FNIR
1.00
0.50
0.20
0.10
0.05
0.02
0.01
0.01
0.02
0.05
0.10
0.20
0.50
1.00
[Gallery size: 418 Probes: 809]
FPIR
S
T
U
V
W

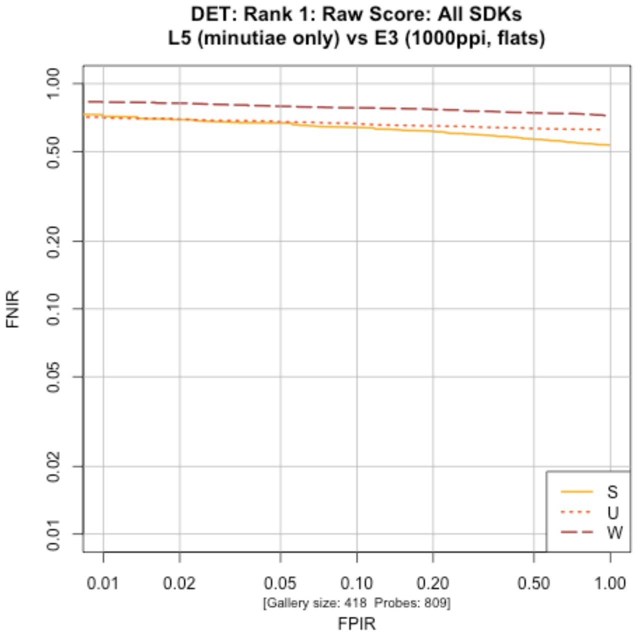

DET: Rank 1: Probability score: All SDKs
L4 (EFS only) vs E1 (1000ppi, rolls)
FNIR
1.00
0.50
0.20
0.10
0.05
0.02
0.01
0.01
0.02
0.05
0.10
0.20
0.50
1.00
[Gallery size: 418 Probes: 809]
FPIR
S
U
V
W

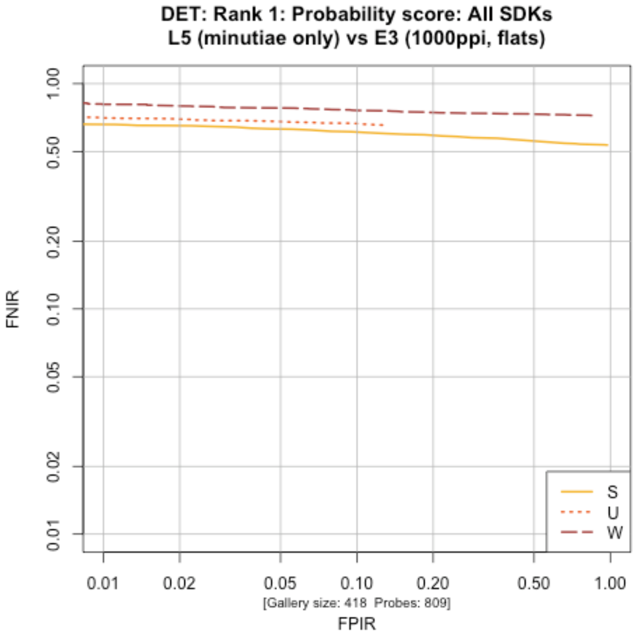

DET: Rank 1: Raw Score: All SDKs
L4 (EFS only) vs E3 (1000ppi, flats)


DET: Rank 1: Probability score: All SDKs
L4 (EFS only) vs E3 (1000ppi, flats)


DET: Rank 1: Raw Score: All SDKs
L5 (minutiae only) vs E1 (1000ppi, rolls)


DET: Rank 1: Probability score: All SDKs
L5 (minutiae only) vs E1 (1000ppi, rolls)





10 Timing

Processing time was not constrained for the public challenge, but participants were requested to return system and timing information, as discussed in Section 4.2.

Table 2: Systems used by participants

Participant	System
S	Intel(R) Core(TM)2 Quad CPU, *2.66GHz, *3.24GB RAM, single thread, per core
T	n/a
U	All timings are reported on one core of a Xeon 5450 @ 3GHz.
V	n/a
W	Intel(R) Xeon(R) E5410 @ 2.33 Ghz - Dual Processor Quad Core, Memory: FB-DDR2 332.5 MHz - 32GB, 1 thread, 1 process, per core
X	Xeon 2.33 Ghz machine with 8 cores, 4GB RAM running 32-bit Linux, per core

Table 3: Processing time for exemplar enrollment (sec per 10-print set)

	S	U	T	V	W	X
E1	104.12	29	-	-	57.94	16.6
E2	108.25	21	-	-	62.70	5.46
E3	89.62	26	-	-	26.80	18.14
E4	91.57	15	-	-	26.13	6.62

Table 4: Processing time for latent matching, per latent per 10-print exemplar set

	S	U	T	V	W	X
L1vsE1	0.46	0.06	0.24	0.40	0.92	0.52
L1vsE2	0.31	0.04	-	-	0.73	0.24
L1vsE3	0.44	0.04	-	-	0.75	0.32
L1vsE4	0.28	0.03	-	-	0.48	0.20
L2vsE1	0.48	0.07	-	0.28	0.51	-
L2vsE2	0.31	0.05	-	-	0.44	-
L2vsE3	0.45	0.04	-	-	0.29	-
L2vsE4	0.28	0.03	-	-	0.27	-
L3vsE1	0.45	0.09	-	0.28	0.23	-
L3vsE2	0.29	0.07	-	-	0.20	-
L3vsE3	0.42	0.05	-	-	0.33	-
L3vsE4	0.26	0.04	-	-	0.31	-
L4vsE1	0.08	0.04	0.45	0.08	0.48	-
L4vsE2	0.07	0.03	0.20	-	0.33	-
L4vsE3	0.06	0.03	-	-	0.39	-
L4vsE4	0.06	0.02	-	-	0.28	-
L5vsE1	0.08	0.01	-	0.07	0.08	-
L5vsE2	0.07	0.01	-	-	0.07	-
L5vsE3	0.06	0.01	-	-	0.04	-
L5vsE4	0.06	0.01	-	-	0.04	-

Appendix C

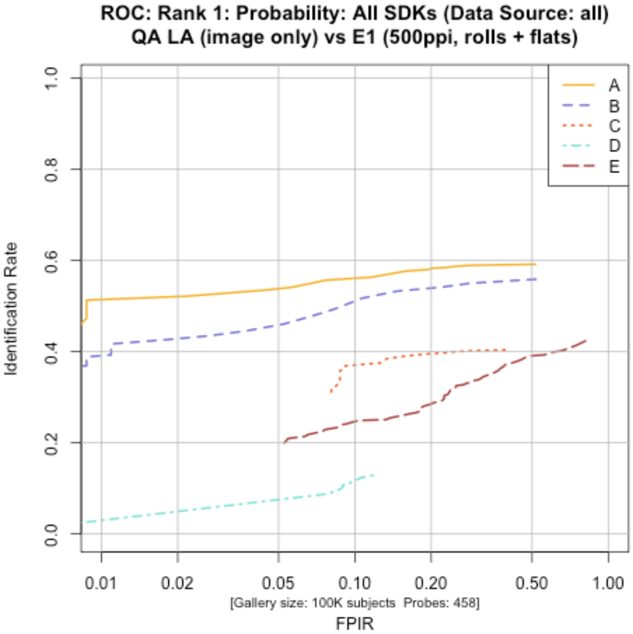
ELFT-EFS
NIST Evaluation of Latent Fingerprint Technologies: Extended Feature Sets

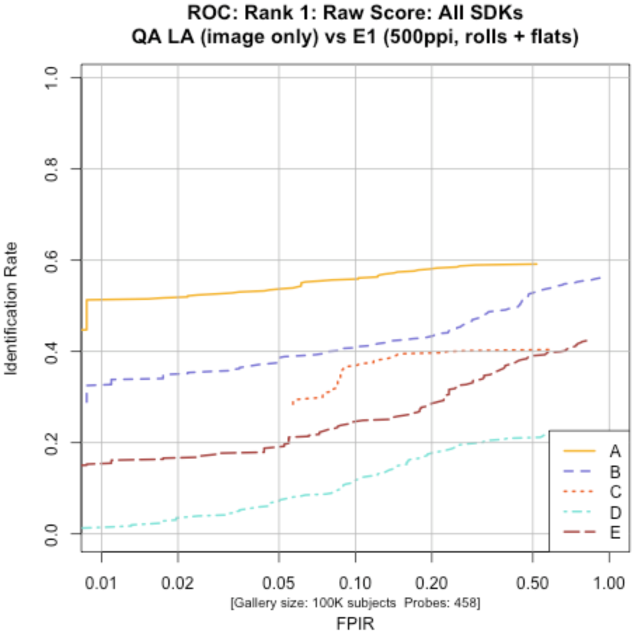
Additional Results

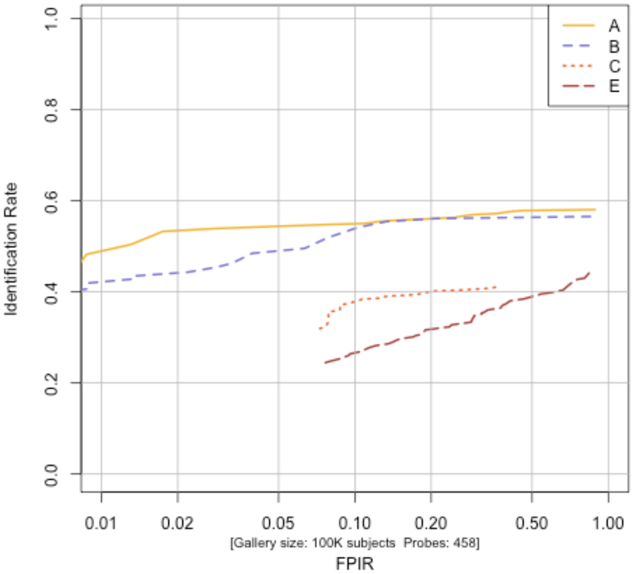
Contents

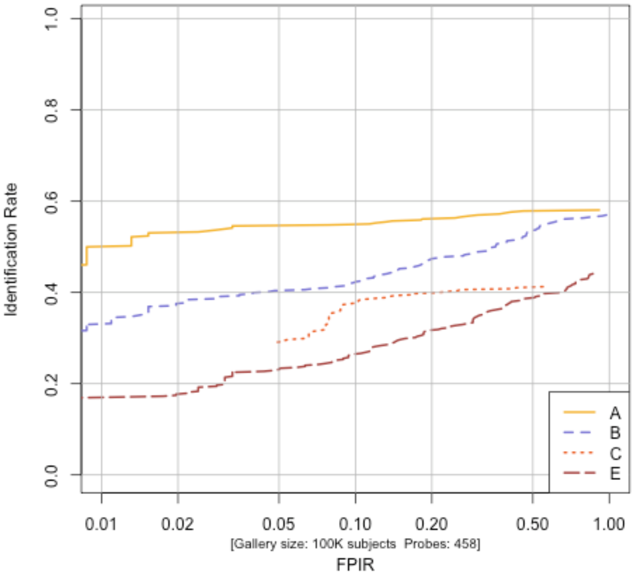
1 Raw score and probability score results

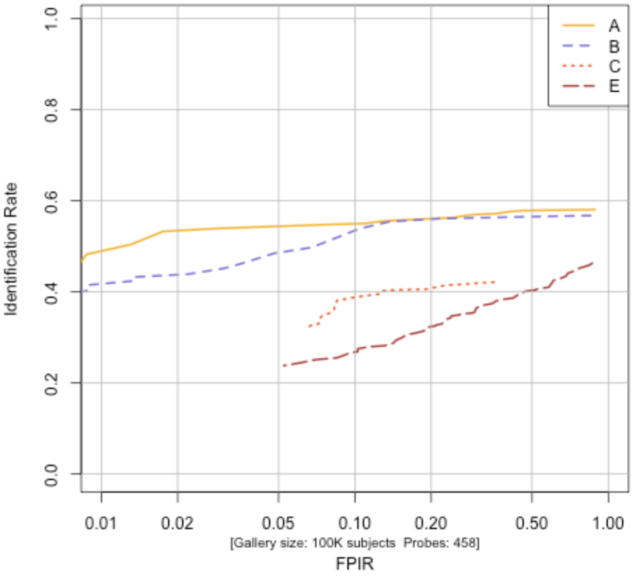
The following ROCs show the differences between the normalized "Probability" scores and the raw scores. In each case, participants returned a raw score and a normalized score estimating the probability of a match. In almost all cases, the probability scores provided better results than the raw scores. In each case, the probability score is on the left and the corresponding raw score on the right.

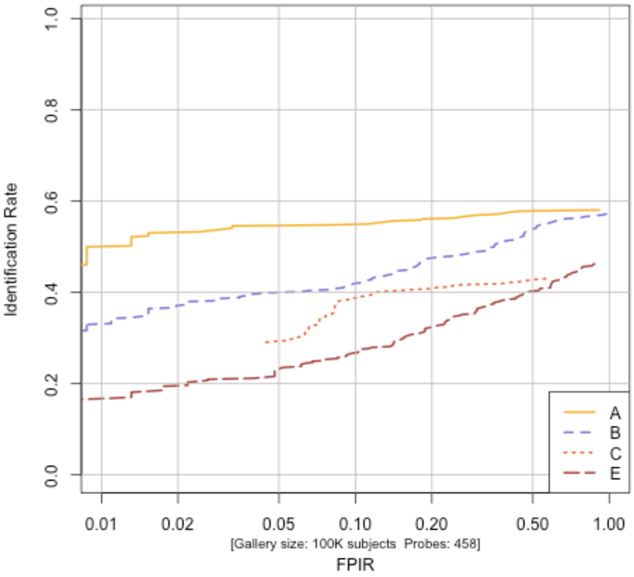




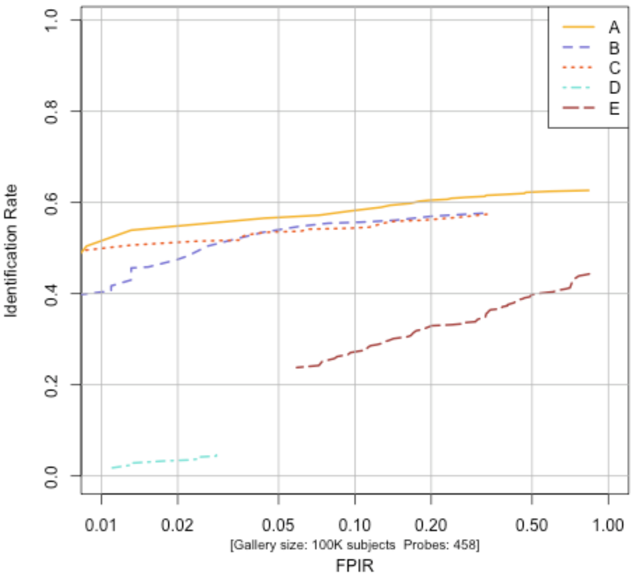

ROC: Rank 1: Probability: All SDKs (Data Source: all)
QA LB (image + ROI) vs E1 (500ppi, rolls + flats)


ROC: Rank 1: Raw Score: All SDKs
QA LB (image + ROI) vs E1 (500ppi, rolls + flats)

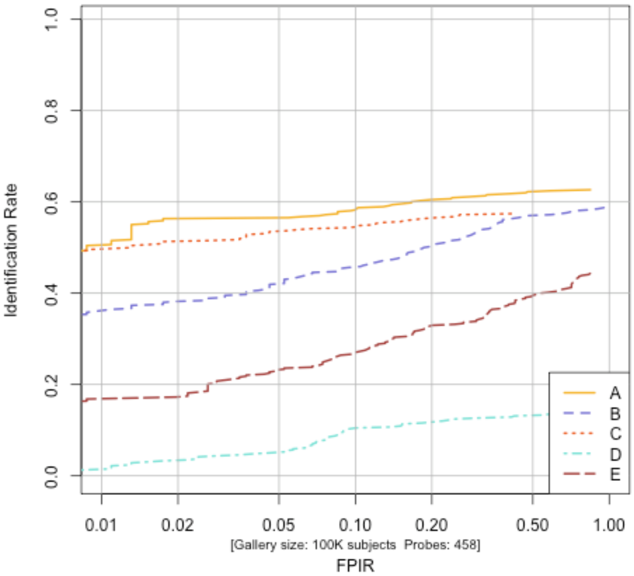

ROC: Rank 1: Probability: All SDKs (Data Source: all)
QA LC (image + ROI + Qual Map) vs E1 (500ppi, rolls + flats)


ROC: Rank 1: Raw Score: All SDKs
QA LC (image + ROI + Qual Map) vs E1 (500ppi, rolls + flats)

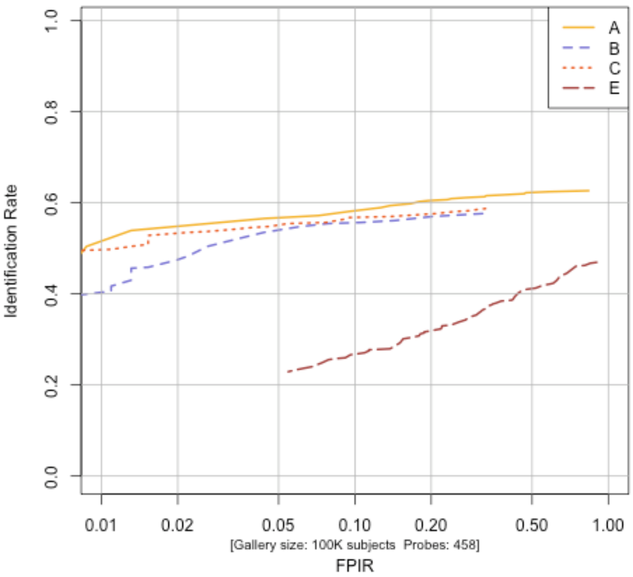
ROC: Rank 1: Probability: All SDKs (Data Source: all)
QA LD (image + EBTS feats) vs E1 (500ppi, rolls + flats)



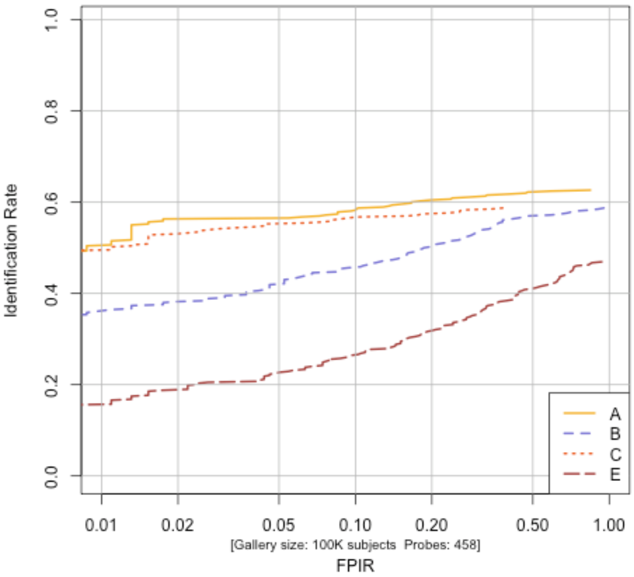
ROC: Rank 1: Raw Score: All SDKs
QA LD (image + EBTS feats) vs E1 (500ppi, rolls + flats)



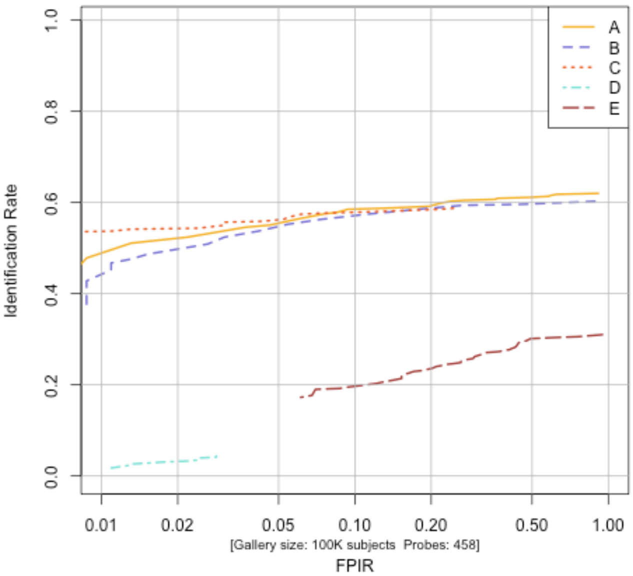
ROC: Rank 1: Probability: All SDKs (Data Source: all)
QA LE (image + EFS) vs E1 (500ppi, rolls + flats)



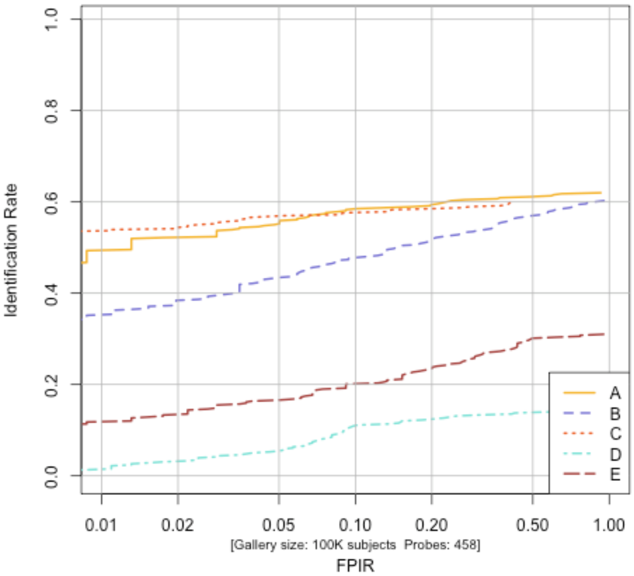
ROC: Rank 1: Raw Score: All SDKs
QA LE (image + EFS) vs E1 (500ppi, rolls + flats)



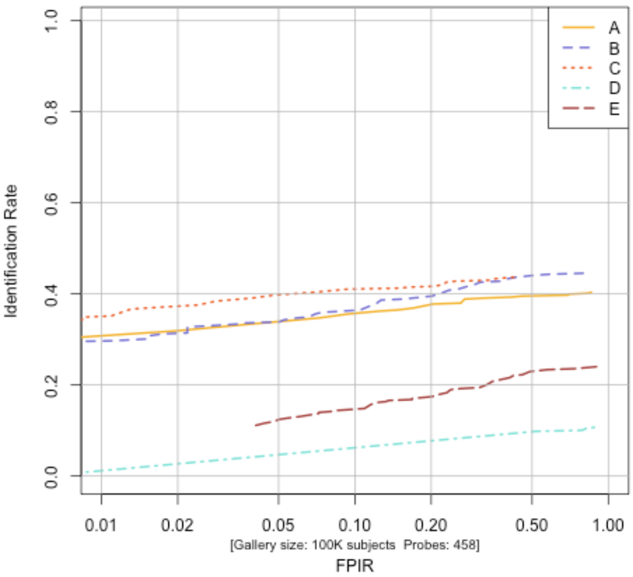
ROC: Rank 1: Probability: All SDKs (Data Source: all)
QA LF (image+EFS (with skel)) vs E1 (500ppi, rolls + flats)



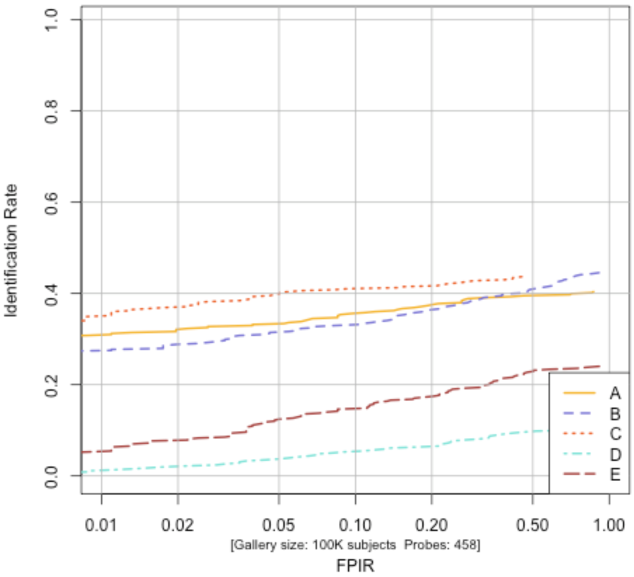
ROC: Rank 1: Raw Score: All SDKs
QA LF (image+EFS (with skel)) vs E1 (500ppi, rolls + flats)



ROC: Rank 1: Probability: All SDKs (Data Source: all)
QA LG (EBTS features only) vs E1 (500ppi, rolls + flats)



ROC: Rank 1: Raw Score: All SDKs
QA LG (EBTS features only) vs E1 (500ppi, rolls + flats)



2 Proportion of hits at rank 1

The following tables show the proportion of the total hits made by a matcher at any rank (rank ≤ 100) that were rank 1. This is sometimes known as the "Ray Moore statistic". AFIS pioneer Ray Moore observed that this tended to be about 83% for minutiae searches at the time.

Draft note: For participant D, subset LE results had not been completed at the time this draft was written, but will be included in the final report.

Table 1 Proportion of hits at rank 1 for the Baseline-QA dataset (458 latents, subset of Baseline)

	Latent Subset						
	LA	**LB**	**LC**	**LD**	**LE**	**LF**	**LG**
	Image only	Image + ROI	Image + ROI + Pattern Class + Qual map	Image + Minutiae	Image + EFS	Image + EFS + Skeleton	Minutiae only
A	88%	91%	91%	90%	90%	91%	82%
B	87%	87%	87%	88%	88%	88%	78%
C	87%	85%	88%	90%	89%	89%	80%
D	80%	*n/a*[*]	*n/a**	73%	*TBD*	75%	69%
E	82%	78%	82%	79%	83%	73%	68%

Table 2: Proportion of hits at rank 1 for the Baseline dataset (1114 latents)

	Latent Subset		
	LA	**LE**	**LG**
	Image only	Image + EFS	Minutiae only
A	92%	93%	82%
B	90%	90%	81%
C	89%	89%	86%
D	78%	*TBD*	69%
E	84%	83%	74%

3 Potential for fusion

This section addresses the potential accuracy if pairs of matchers were combined.

3.1 Rank 1

LA

All matchers combined hit 798 unique latents (71.63% identification rate)
SDK A had the best hit rate with 693 hits out of 1114 (62.21% identification rate)
SDK B had the 2nd best hit rate with 682 hits out of 1114 (61.22% identification rate)
Combining SDK A and SDK B yields 777 hits out of 1114 (69.75% identification rate)

- A (693 hits) and B (682 hits) = 777 (69.75% identification rate)
- A (693 hits) and E (526 hits) = 728 (65.35% identification rate)
- A (693 hits) and C (538 hits) = 724 (64.99% identification rate)
- B (682 hits) and E (526 hits) = 712 (63.91% identification rate)
- A (693 hits) and D (280 hits) = 712 (63.91% identification rate)
- B (682 hits) and C (538 hits) = 711 (63.82% identification rate)
- B (682 hits) and D (280 hits) = 693 (62.21% identification rate)
- C (538 hits) and E (526 hits) = 632 (56.73% identification rate)

[*] *Participant D informed NIST that their software did not utilize the features in subsets LB/LC, and therefore those subsets were not run.*

- C (538 hits) and D (280 hits) = 572 (51.35% identification rate)
- D (280 hits) and E (526 hits) = 571 (51.26% identification rate)

LE

All matchers combined hit 839 unique latents (75.31% identification rate)
SDK A had the best hit rate with 743 hits out of 1114 (66.70% identification rate)
SDK B had the 2nd best hit rate with 705 hits out of 1114 (63.29% identification rate)
Combining SDK A and SDK B yields 801 hits out of 1114 (71.90% identification rate)

- A (743 hits) and B (705 hits) = 801 (71.90% identification rate)
- A (743 hits) and C (691 hits) = 800 (71.81% identification rate)
- A (743 hits) and E (560 hits) = 777 (69.75% identification rate)
- B (705 hits) and C (691 hits) = 772 (69.30% identification rate)
- B (705 hits) and E (560 hits) = 755 (67.77% identification rate)
- C (691 hits) and E (560 hits) = 750 (67.32% identification rate)

Draft note: For participant D, subset LE results had not been completed at the time this draft was written, but will be included in the final report.

LG

All matchers combined hit 614 unique latents (55.12% identification rate)
SDK B had the best hit rate with 538 hits out of 1114 (48.29% identification rate)
SDK C had the 2nd best hit rate with 532 hits out of 1114 (47.76% identification rate)
Combining SDK B and SDK C yields 599 hits out of 1114 (53.77% identification rate)

- B (538 hits) and C (532 hits) = 599 (53.77% identification rate)
- A (490 hits) and B (538 hits) = 578 (51.89% identification rate)
- A (490 hits) and C (532 hits) = 571 (51.26% identification rate)
- B (538 hits) and E (327 hits) = 550 (49.37% identification rate)
- C (532 hits) and E (327 hits) = 544 (48.83% identification rate)
- B (538 hits) and D (50 hits) = 541 (48.56% identification rate)
- C (532 hits) and D (50 hits) = 534 (47.94% identification rate)
- A (490 hits) and E (327 hits) = 506 (45.42% identification rate)
- A (490 hits) and D (50 hits) = 493 (44.25% identification rate)
- D (50 hits) and E (327 hits) = 340 (30.52% identification rate)

3.2 Rank 100

LA

All matchers combined hit 909 unique latents (81.60% identification rate)
SDK A had the best hit rate with 783 hits out of 1114 (70.29% identification rate)
SDK B had the 2nd best hit rate with 780 hits out of 1114 (70.02% identification rate)
Combining SDK A and SDK B yields 861 hits out of 1114 (77.29% identification rate)

- A (783 hits) + B (780 hits) = 861 (77.29% identification rate)
- A (783 hits) + E (681 hits) = 839 (75.31% identification rate)
- B (780 hits) + E (681 hits) = 835 (74.96% identification rate)
- B (780 hits) + C (625 hits) = 814 (73.07% identification rate)
- A (783 hits) + D (416 hits) = 811 (72.80% identification rate)

- A (783 hits) + C (625 hits) = 811 (72.80% identification rate)
- B (780 hits) + D (416 hits) = 806 (72.35% identification rate)
- C (625 hits) + E (681 hits) = 758 (68.04% identification rate)
- D (416 hits) + E (681 hits) = 731 (65.62% identification rate)
- C (625 hits) + D (416 hits) = 672 (60.32% identification rate)

LE

All matchers combined hit 934 unique latents (83.84% identification rate)
SDK B had the best hit rate with 823 hits out of 1114 (73.88% identification rate)
SDK A had the 2nd best hit rate with 823 hits out of 1114 (73.88% identification rate)
Combining SDK B and SDK A yields 885 hits out of 1114 (79.44% identification rate)

- A (823 hits) + B (823 hits) = 885 (79.44% identification rate)
- A (823 hits) + C (798 hits) = 884 (79.35% identification rate)
- B (823 hits) + E (717 hits) = 882 (79.17% identification rate)
- B (823 hits) + C (798 hits) = 878 (78.82% identification rate)
- A (823 hits) + E (717 hits) = 872 (78.28% identification rate)
- C (798 hits) + E (717 hits) = 859 (77.11% identification rate)

Draft note: For participant D, subset LE results had not been completed at the time this draft was written, but will be included in the final report.

LG

All matchers combined hit 795 unique latents (71.36% identification rate)
SDK B had the best hit rate with 706 hits out of 1114 (63.38% identification rate)
SDK C had the 2nd best hit rate with 668 hits out of 1114 (59.96% identification rate)
Combining SDK B and SDK C yields 759 hits out of 1114 (68.13% identification rate)

- B (706 hits) + C (668 hits) = 759 (68.13% identification rate)
- A (654 hits) + B (706 hits) = 750 (67.32% identification rate)
- A (654 hits) + C (668 hits) = 732 (65.71% identification rate)
- B (706 hits) + E (506 hits) = 723 (64.90% identification rate)
- B (706 hits) + D (88 hits) = 708 (63.55% identification rate)
- C (668 hits) + E (506 hits) = 700 (62.84% identification rate)
- A (654 hits) + E (506 hits) = 683 (61.31% identification rate)
- C (668 hits) + D (88 hits) = 669 (60.05% identification rate)
- A (654 hits) + D (88 hits) = 654 (58.71% identification rate)
- D (88 hits) + E (506 hits) = 515 (46.23% identification rate)