Appendix B

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

Public Challenge Results

Contents

B-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.

B-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[1] 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
    - o L1 – image only
    - o L2 – image with EFTS-LFFS features (fields 9.014-9.023)
    - o L3 – image with EFS features (fields 9.300-9.373)
    - o L4 - EFS features alone
    - o L5 - EFTS-LFFS features alone
- Exemplar subtests
    - o E1 - 1000ppi rolled exemplars
    - o E2 - 500ppi[2] rolled exemplars
    - o E3 - 1000ppi plain exemplars (unsegmented slaps)
    - o E4 - 500ppi plain exemplars (unsegmented slaps)

---

[1] *1000 pixels per inch (ppi) is equivalent to 39.37 pixels per millimeter (ppmm)*
[2] *500 pixels per inch (ppi) is equivalent to 19.69 pixels per millimeter (ppmm)*

B-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.

B-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,[3] 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.

B-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.

B-4 Format of results

B-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.

---

[3] *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)

B-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.

B-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.
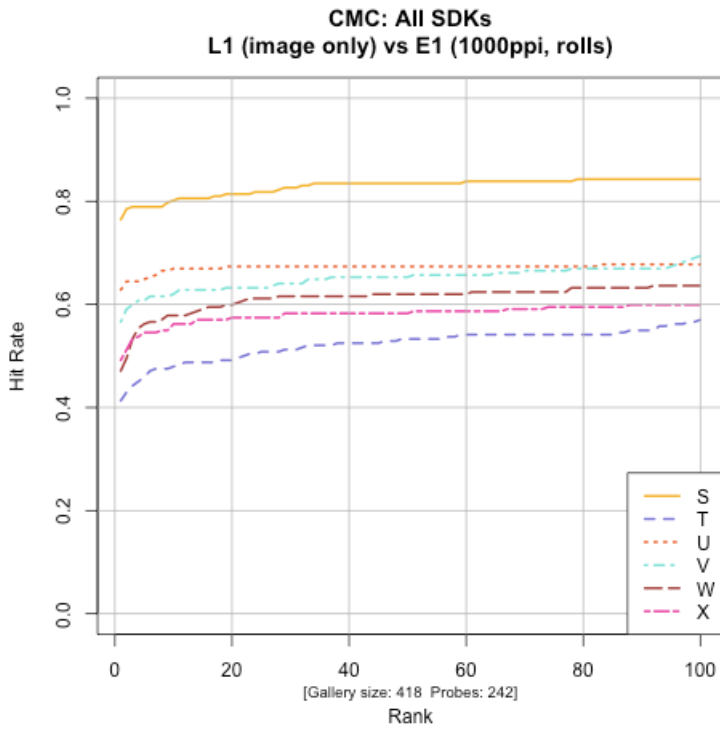
Note that not all participants returned results for all tests.
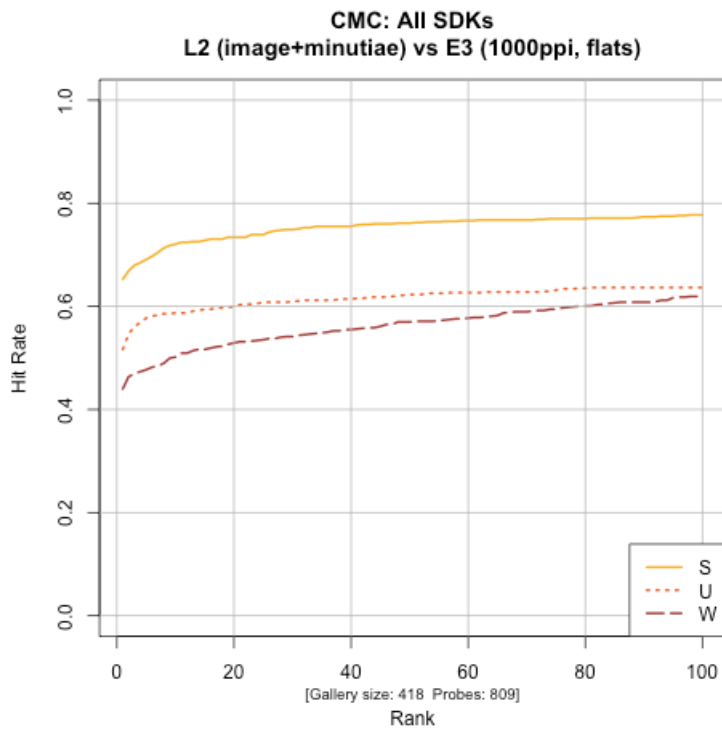
Table 1: Summary of rank-1 identification rates

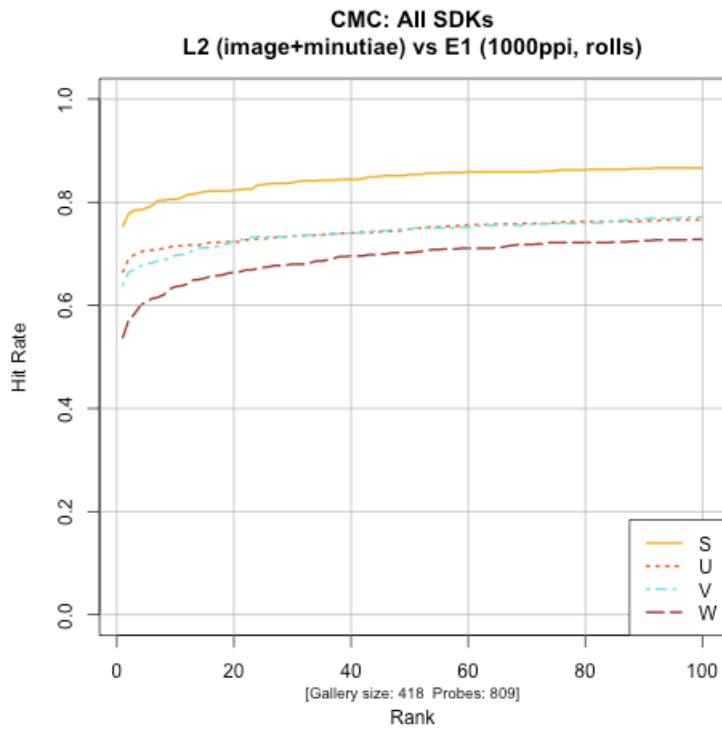|		Participant					
		S	T	U	V	W	X
E1: 1000ppi rolled	L1: Image only	0.764	0.413	0.628	0.566	0.471	0.492
	L2: Image + IAFIS	0.754	-	0.665	0.639	0.538	-
	L3: Image + EFS	0.868	-	0.779	0.648	0.663	-
	L4: EFS only	0.808	0.284	0.775	0.483	0.654	-
	L5: IAFIS only	0.576	-	0.460	0.481	0.396	-
E3: 1000ppi slap	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.
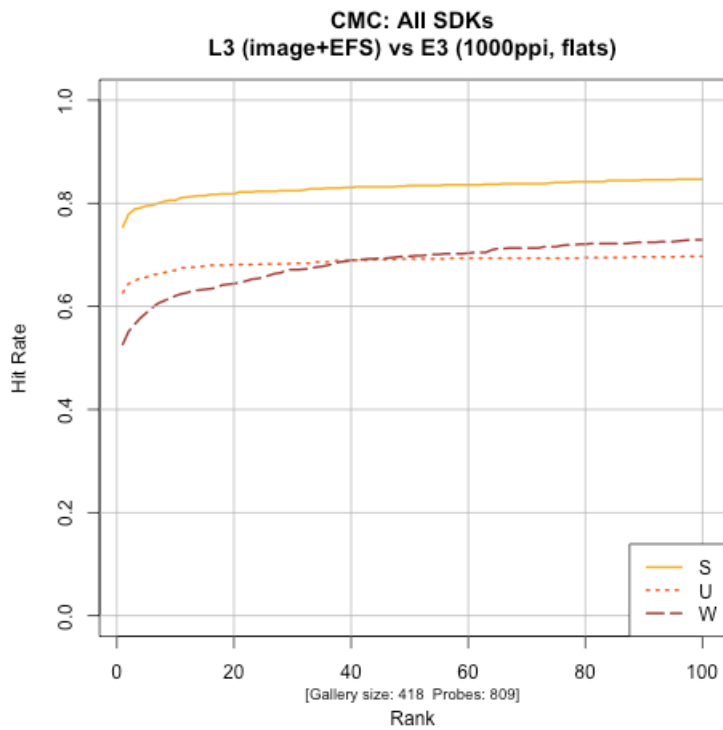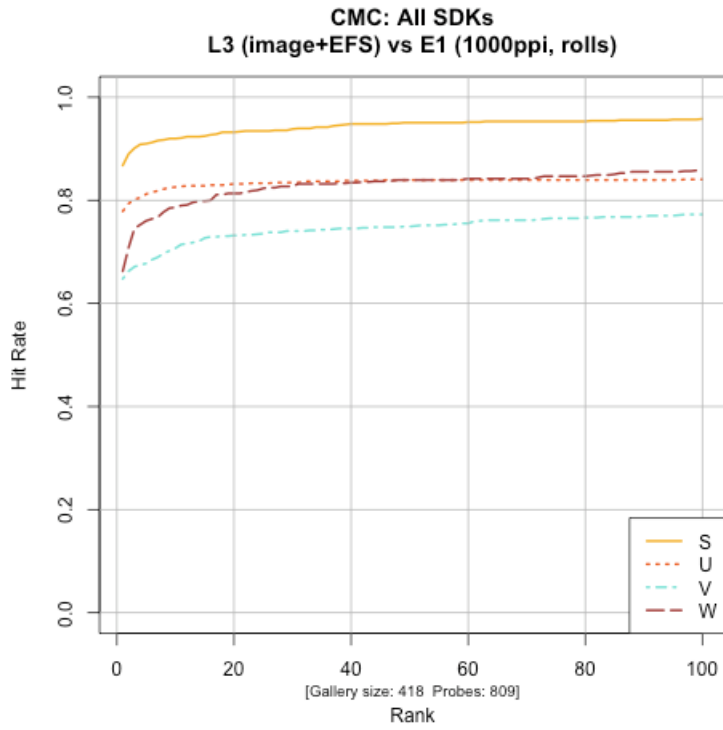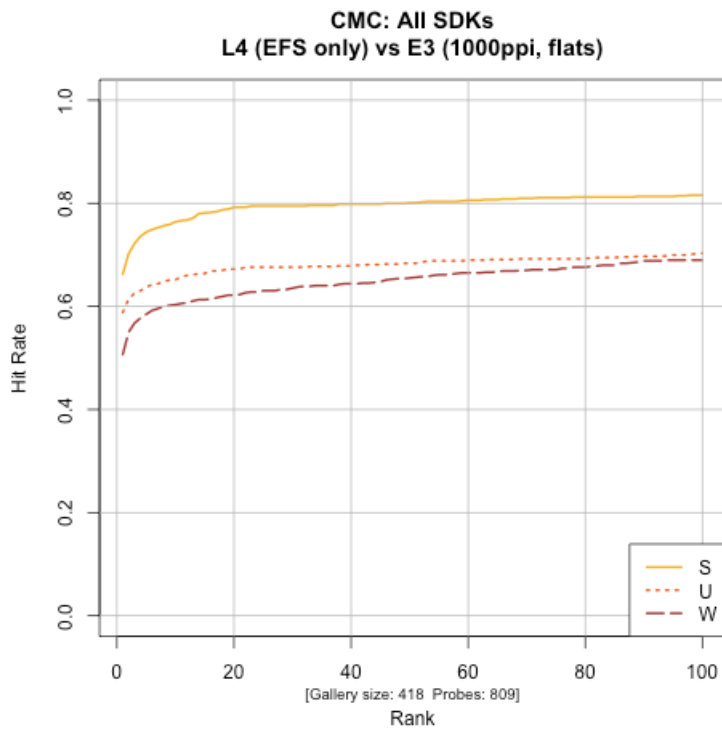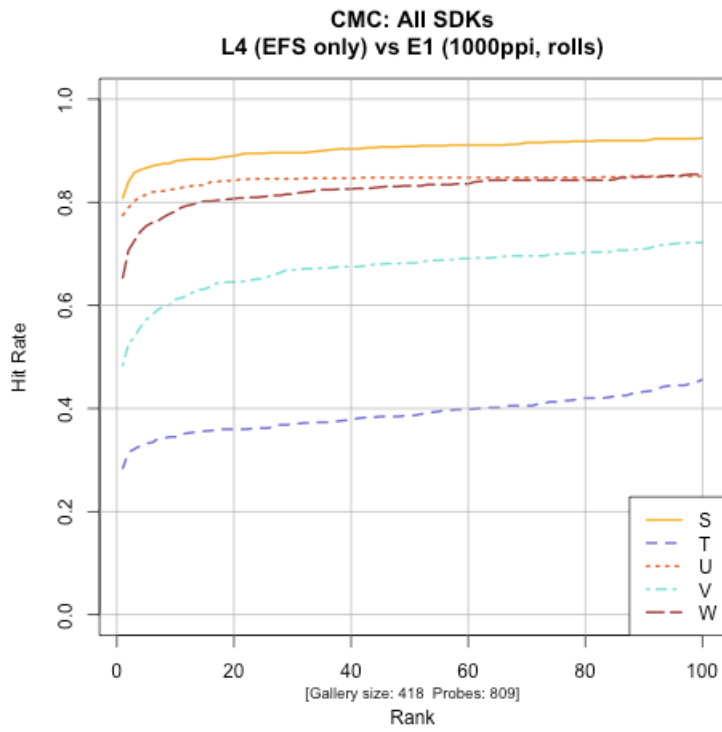
B-5.1   L1 – Image only

CMC: All SDKs
L1 (image only) vs E1 (1000ppi, rolls)

[Gallery size: 418 Probes: 242]

CMC: All SDKs
L1 (image only) vs E3 (1000ppi, flats)

[Gallery size: 418 Probes: 242]
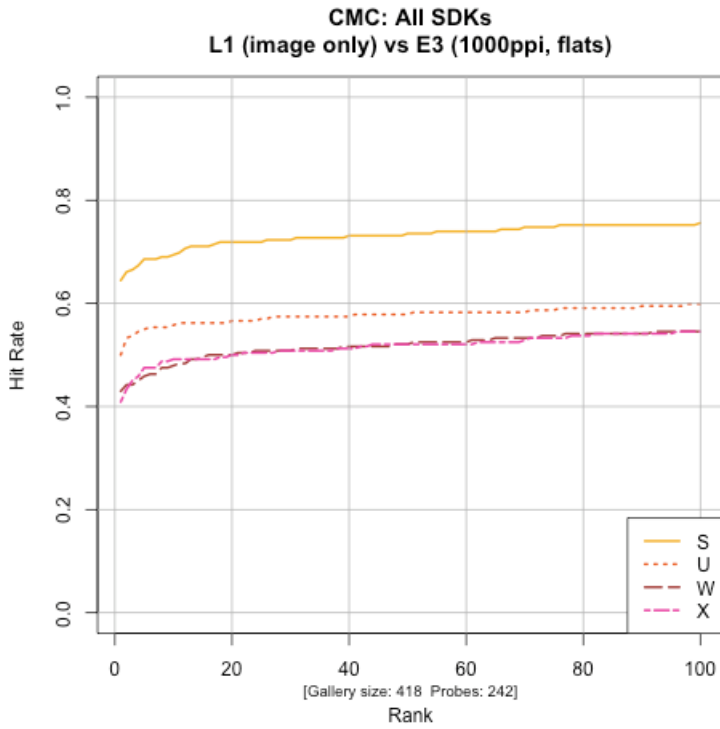
B-5.2 L2 – Image + IAFIS LFFS markup

CMC: All SDKs
L2 (image+minutiae) vs E1 (1000ppi, rolls)

[Gallery size: 418 Probes: 809]

CMC: All SDKs
L2 (image+minutiae) vs E3 (1000ppi, flats)

[Gallery size: 418 Probes: 809]

B-5.3 L3 – Image + Extended Feature Set markup

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

[Gallery size: 418 Probes: 809]

CMC: All SDKs
L3 (image+EFS) vs E3 (1000ppi, flats)

[Gallery size: 418 Probes: 809]

B-5.4 L4 – Extended Feature Set markup (no image)

B-5.5 L5 – IAFIS LFFS markup (no image)

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

[Gallery size: 418 Probes: 809]

CMC: All SDKs
L5 (minutiae only) vs E3 (1000ppi, flats)

[Gallery size: 418 Probes: 809]
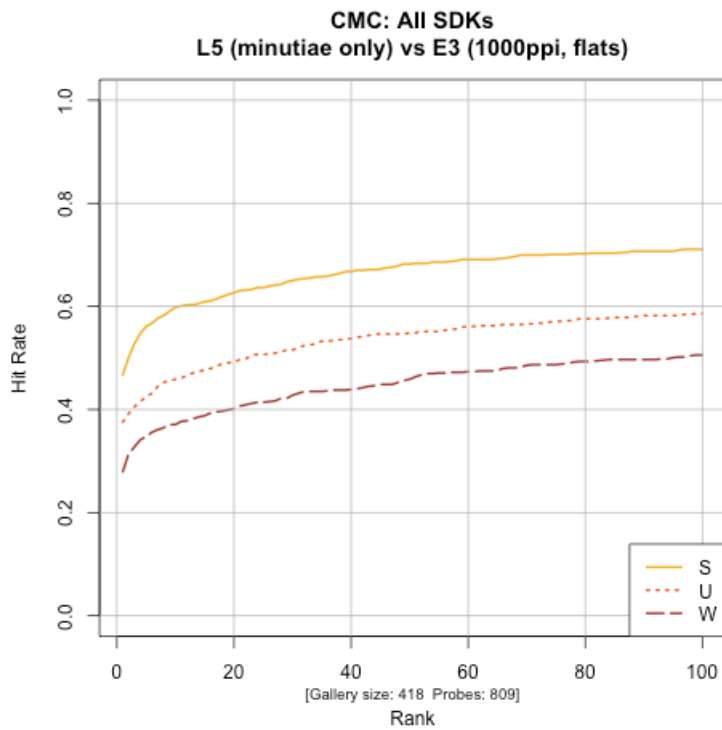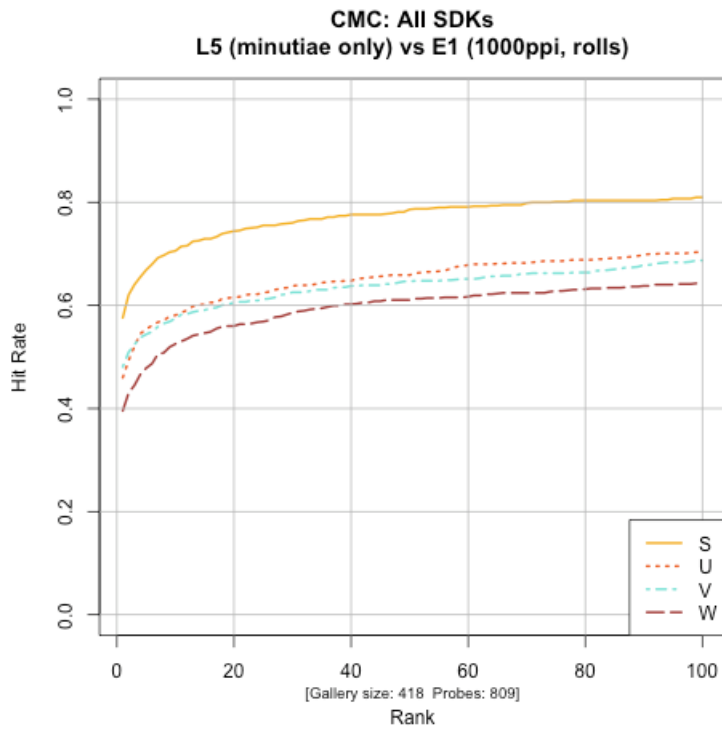
B-6 Results by participant

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

B-6.1 Participant S

B-6.2 Participant T

CMC: SDK: T
L1,L4 vs E1 (1000ppi, rolls)

[Gallery size: 418 Probes: 242 L1, 809 L2-L5]

B-6.3 Participant U

CMC: SDK: U
L1-L5 vs E1 (1000ppi, rolls)

[Gallery size: 418 Probes: 242 L1, 809 L2-L5]

- L1 (image only)
- L2 (image+minutiae)
- L3 (image+EFS)
- L4 (EFS only)
- L5 (minutiae only)

CMC: SDK: U
L1-L5 vs E3 (1000ppi, flats)

[Gallery size: 418 Probes: 242 L1, 809 L2-L5]

- L1 (image only)
- L2 (image+minutiae)
- L3 (image+EFS)
- L4 (EFS only)
- L5 (minutiae only)

B-6.4 Participant V

CMC: SDK: V
L1-L5 vs E1 (1000ppi, rolls)

B-6.5 Participant W

CMC: SDK: W
L1-L5 vs E1 (1000ppi, rolls)

CMC: SDK: W
L1-L5 vs E3 (1000ppi, flats)

B-6.6   Participant X

CMC: SDK: X
L1 vs E1 (1000ppi, rolls)

[Gallery size: 418 Probes: 242 L1, 809 L2-L5]

— L1 (image only)

CMC: SDK: X
L1 vs E3 (1000ppi, flats)

[Gallery size: 418 Probes: 242 L1, 809 L2-L5]

— L1 (image only)

B-7 Multi-encounter [4]

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.[5] 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.
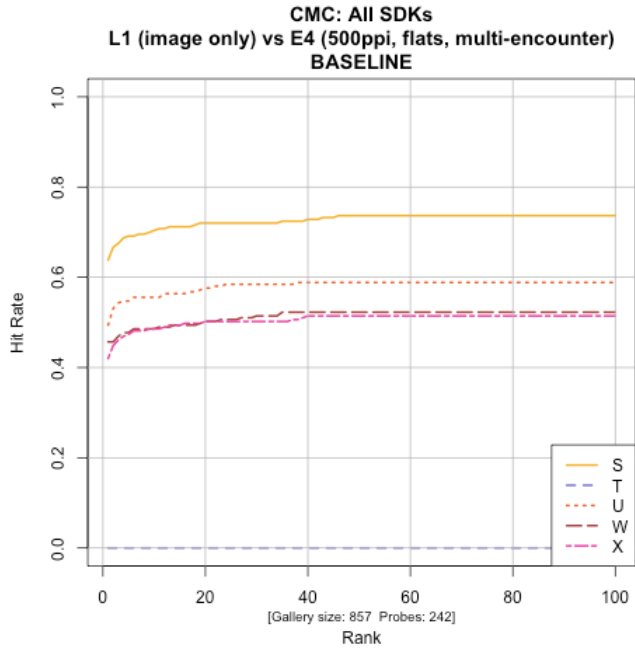
---

[4] *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.*
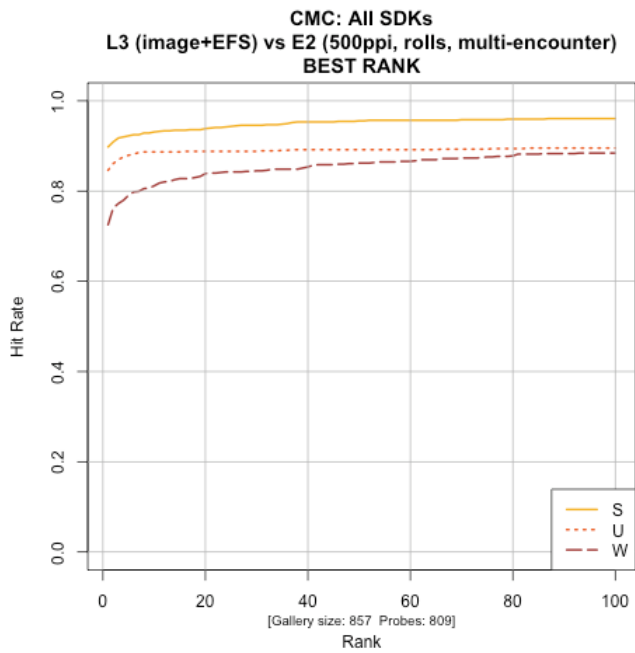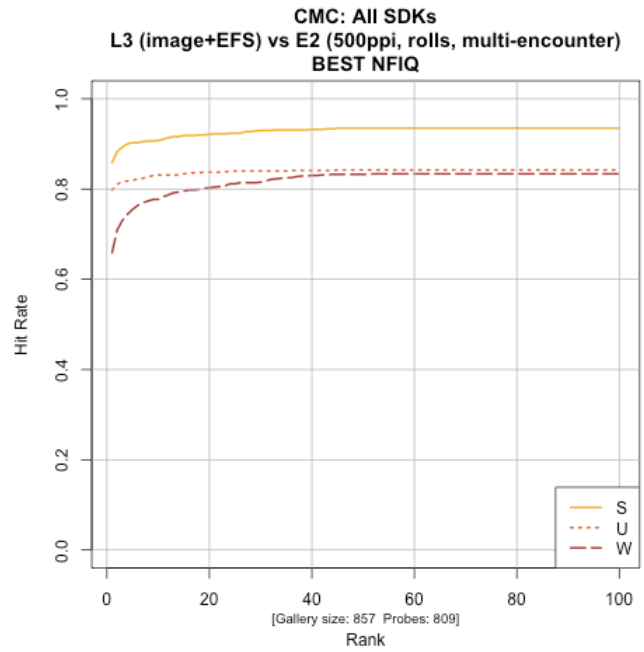
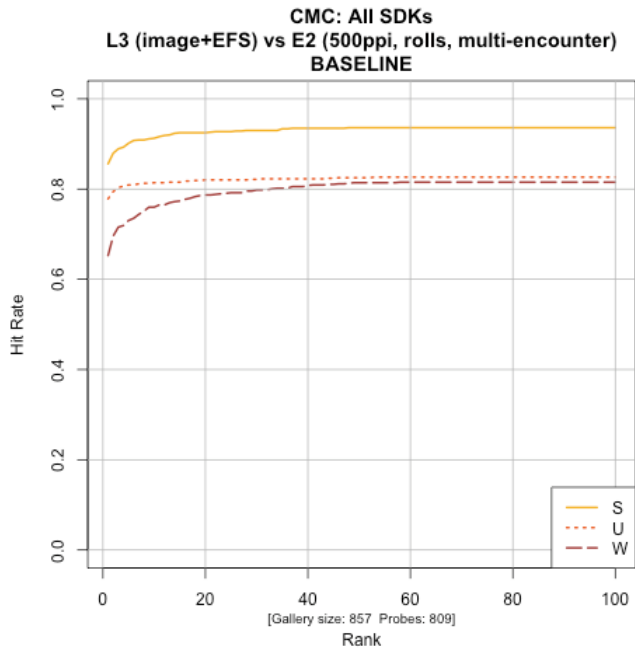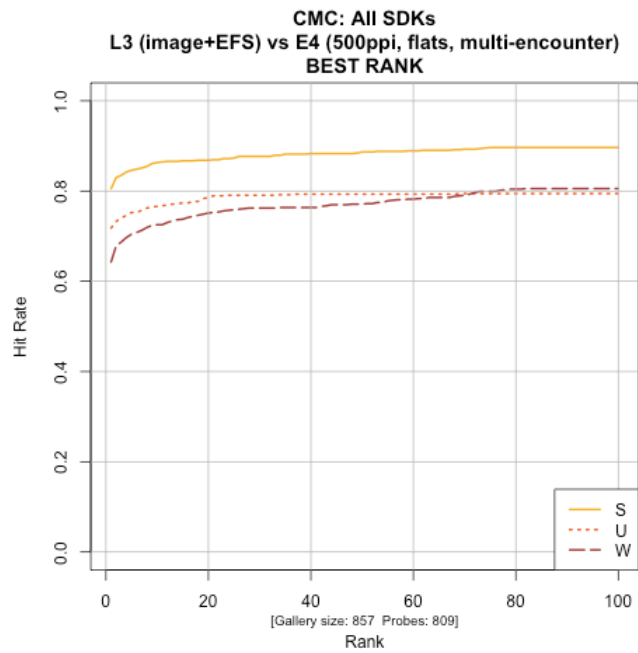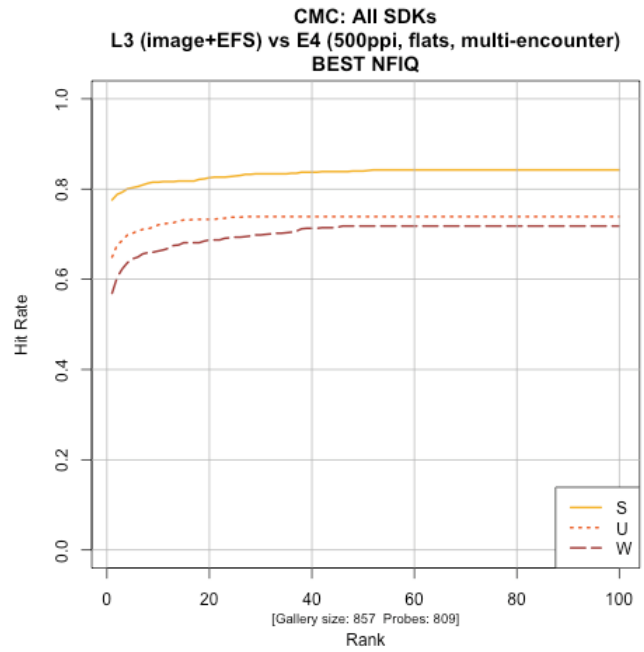[5] *NIST Fingerprint Image Quality*
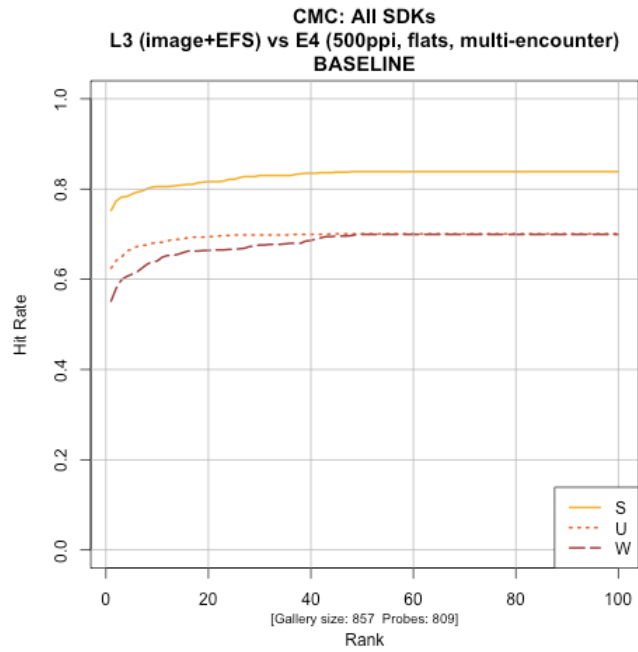
B-7.1 L1 (Image only) x E2 (500ppi rolls)
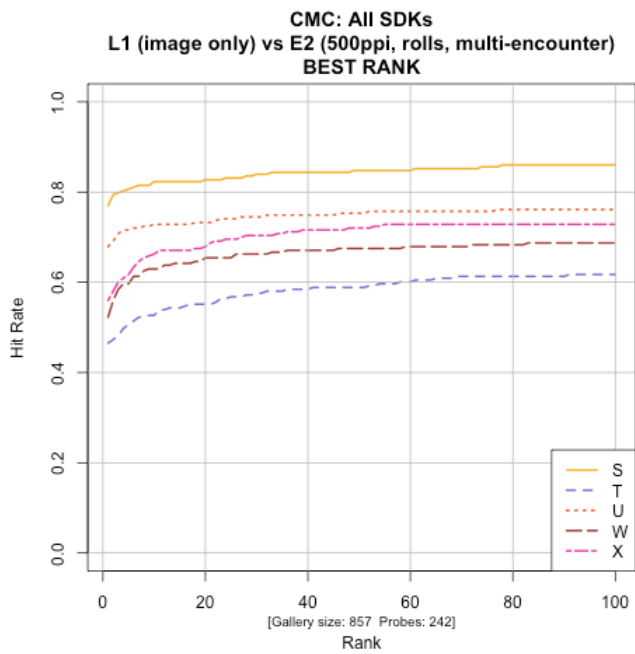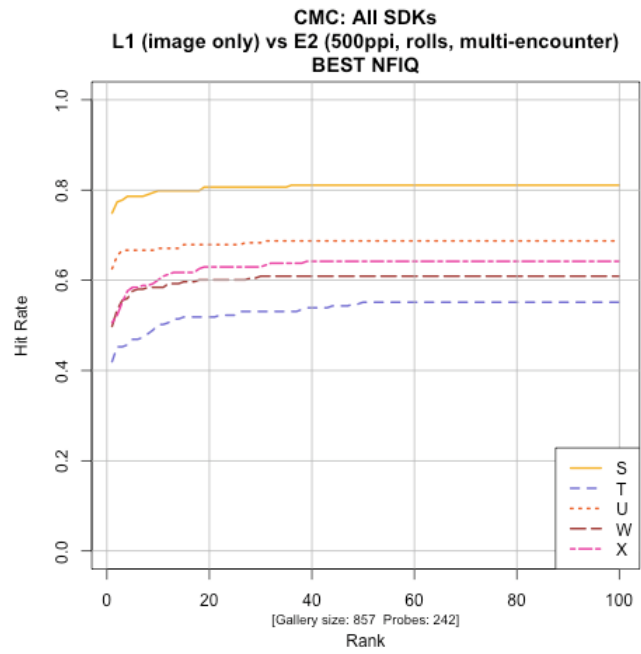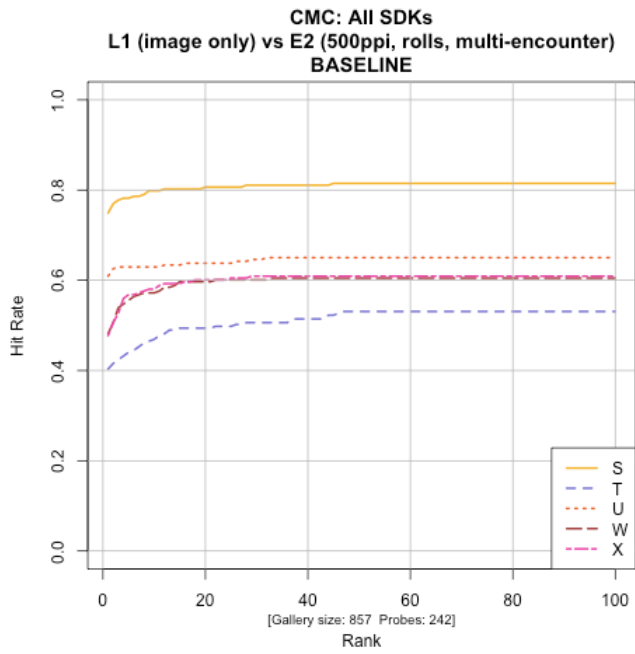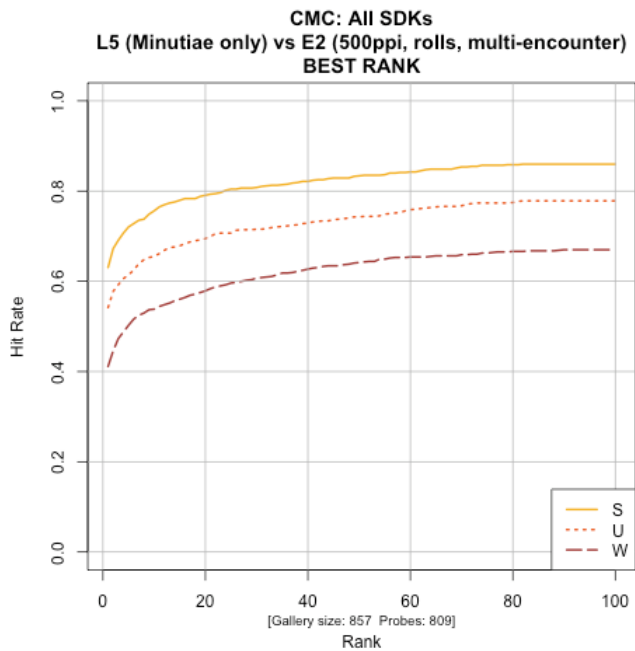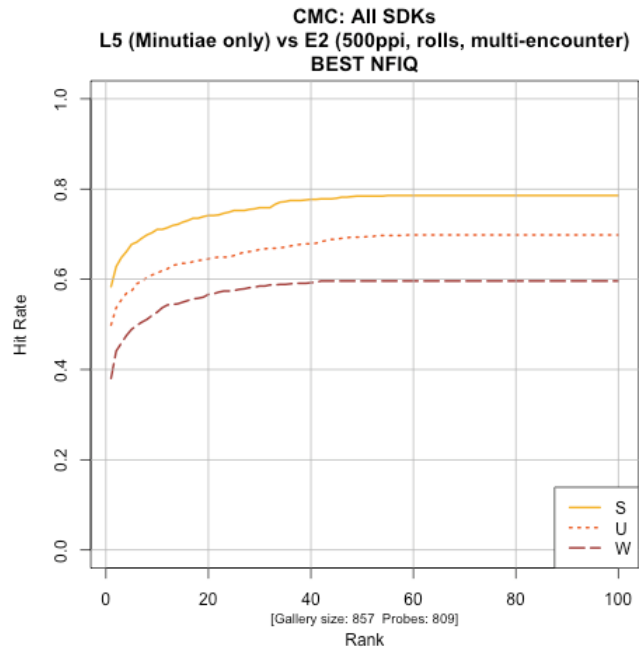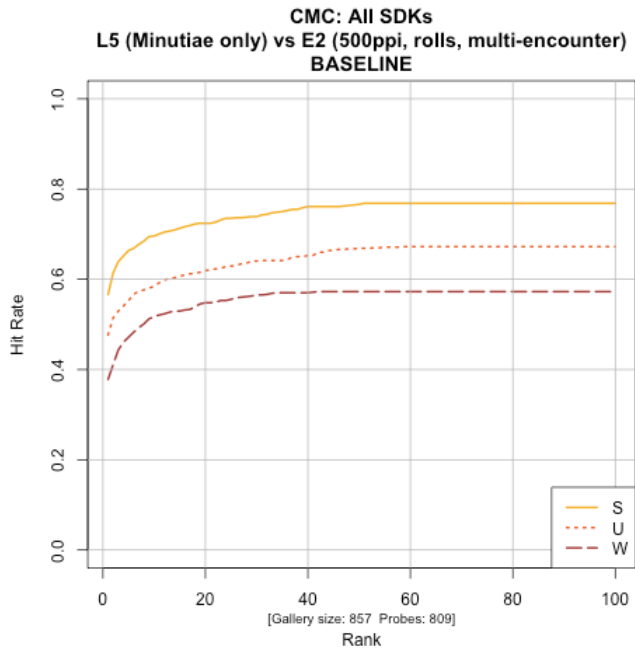
B-7.2 L1 (Image only) x E4 (500ppi flats)
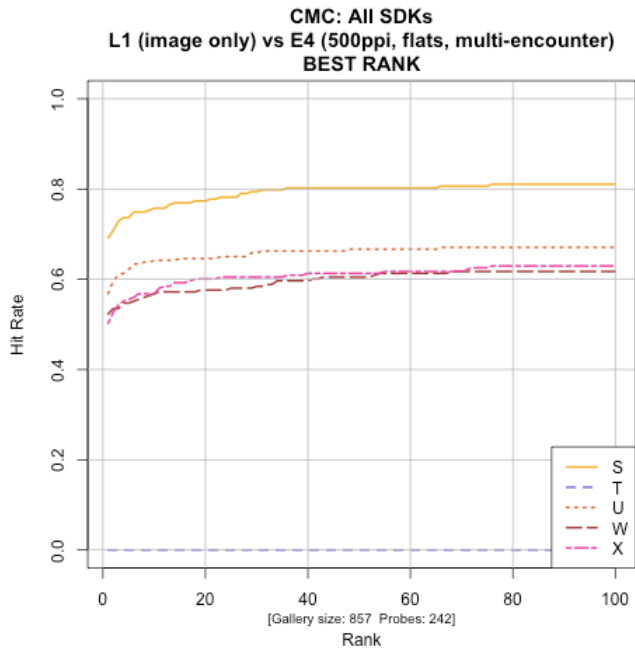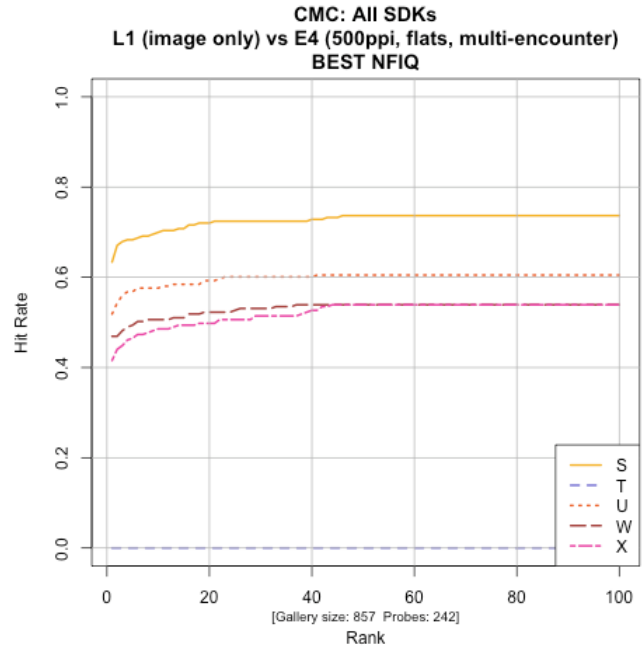
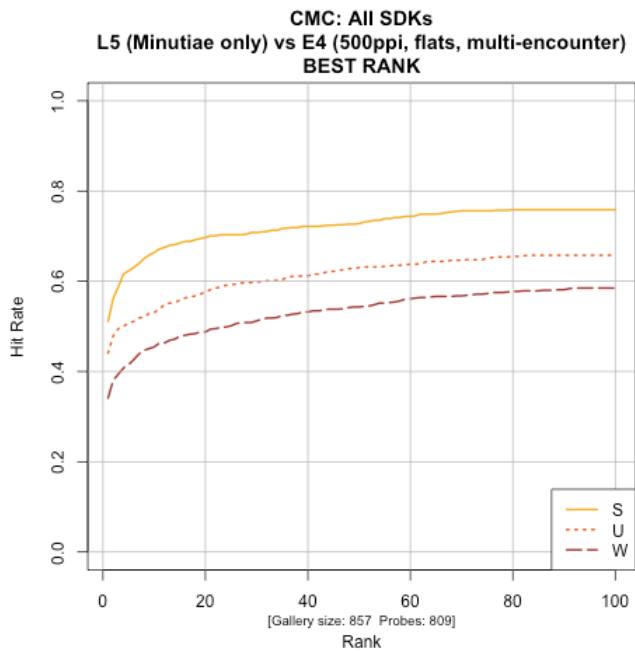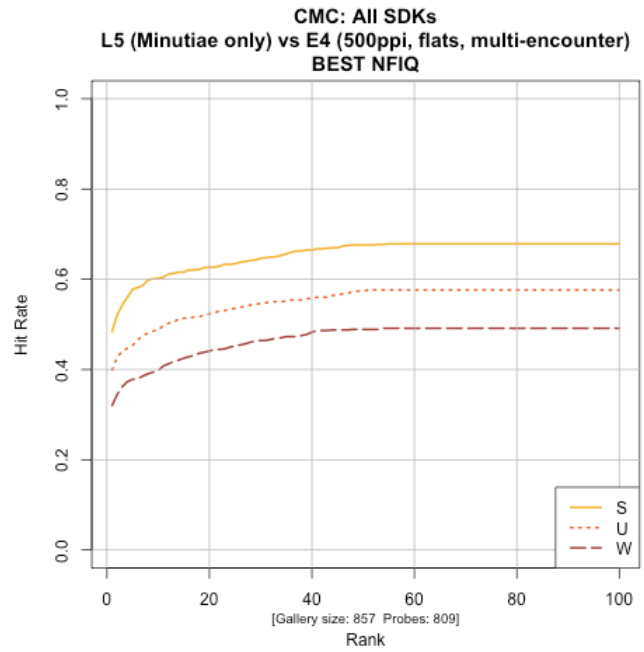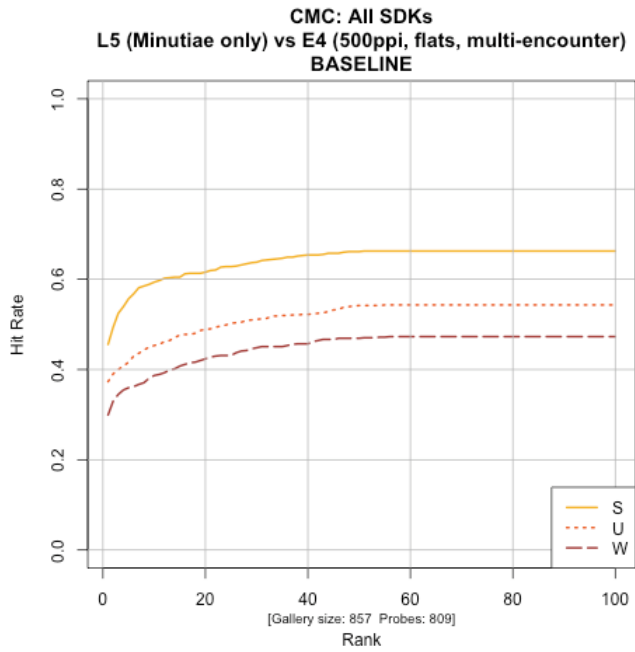**B-7.3   L3 (Image + Extended Feature Set markup) x E2 (500ppi rolls)**

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

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

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

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

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

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

[Gallery size: 857 Probes: 809]
Rank

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

[Gallery size: 857 Probes: 809]
Rank

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

[Gallery size: 857 Probes: 809]
Rank

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

CMC: All SDKs
L5 (Minutiae only) vs E4 (500ppi, flats, multi-encounter)
BASELINE

CMC: All SDKs
L5 (Minutiae only) vs E4 (500ppi, flats, multi-encounter)
BEST NFIQ

CMC: All SDKs
**L5 (Minutiae only) vs E4 (500ppi, flats, multi-encounter)**
**BEST RANK**

B-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]
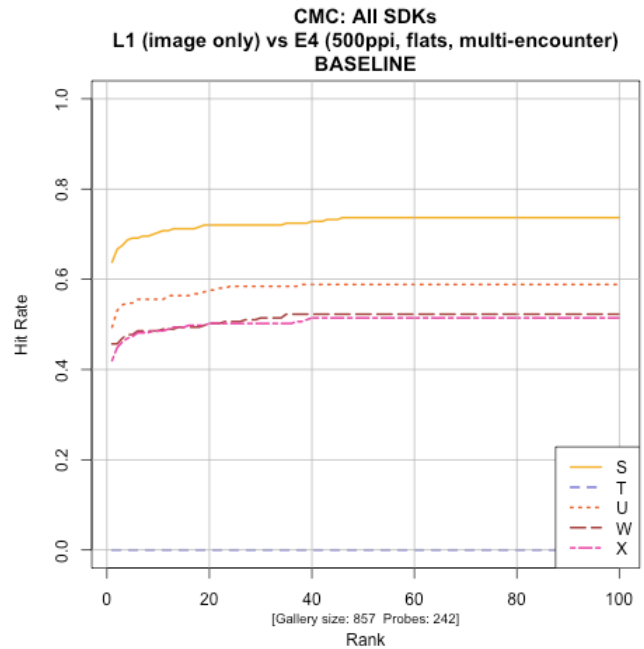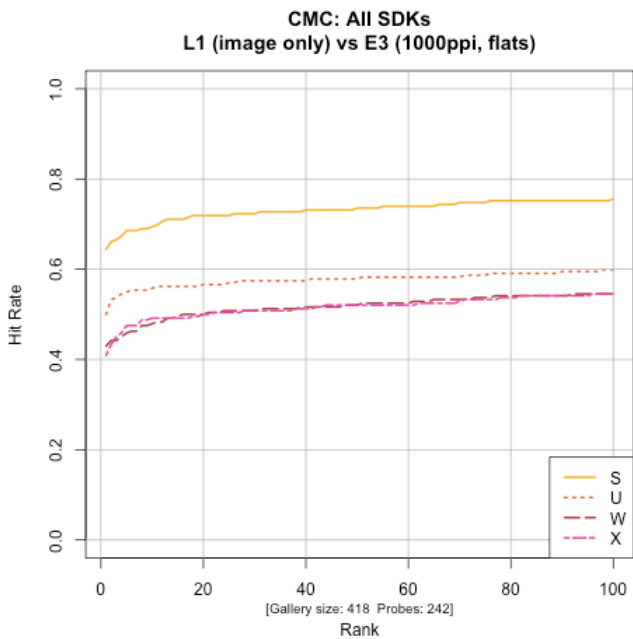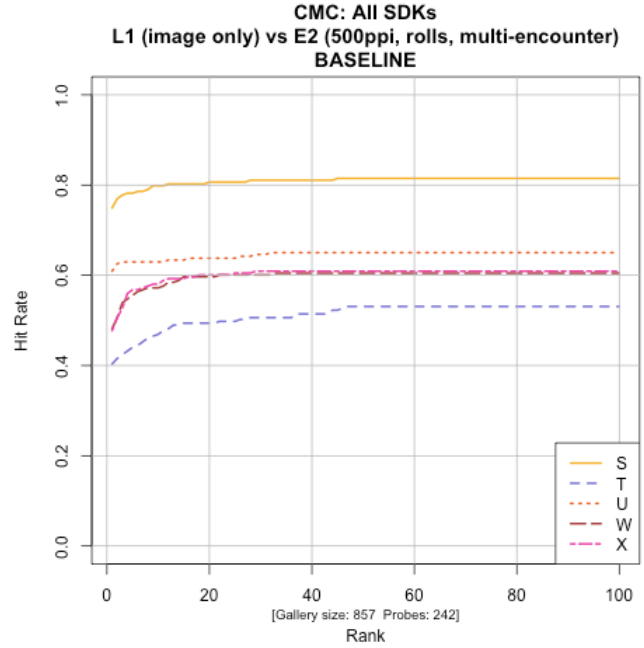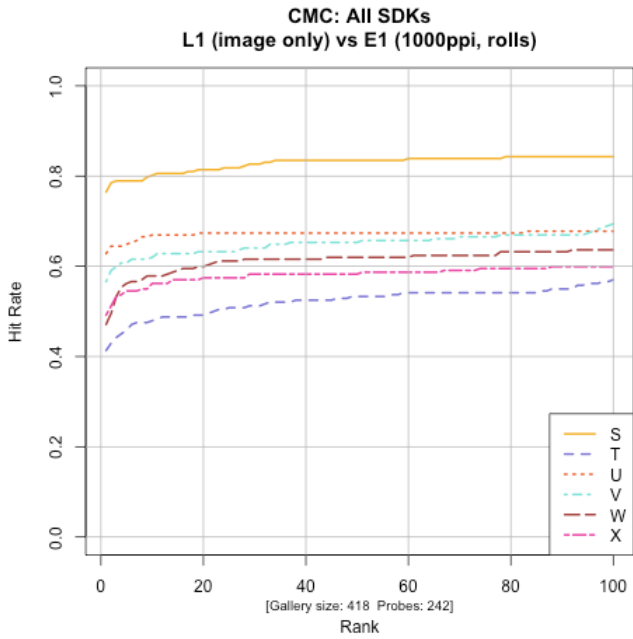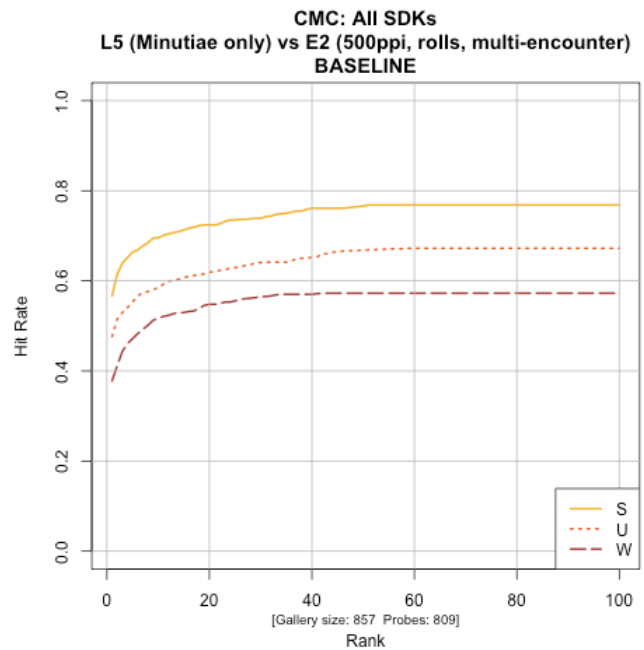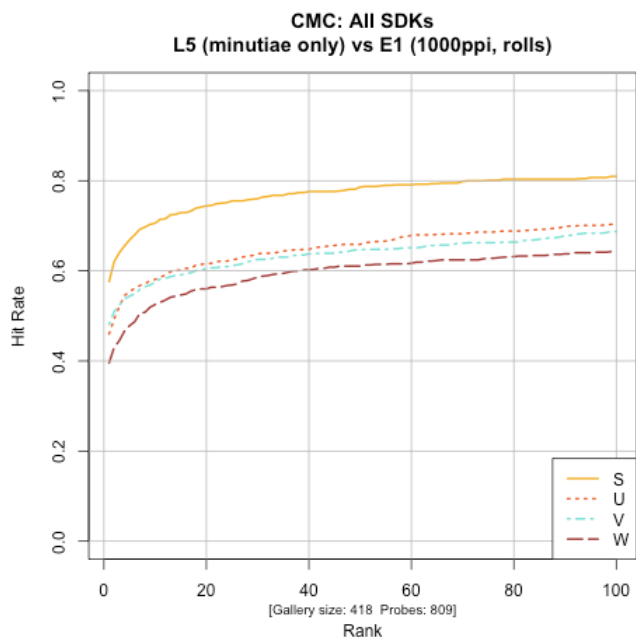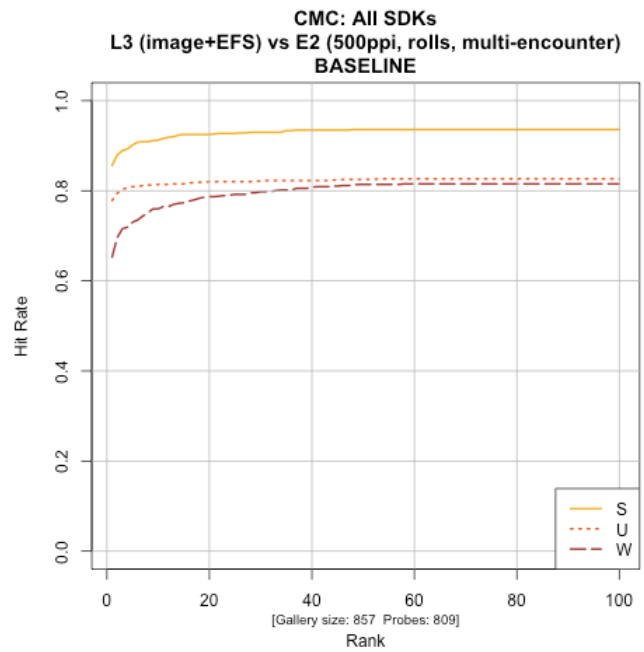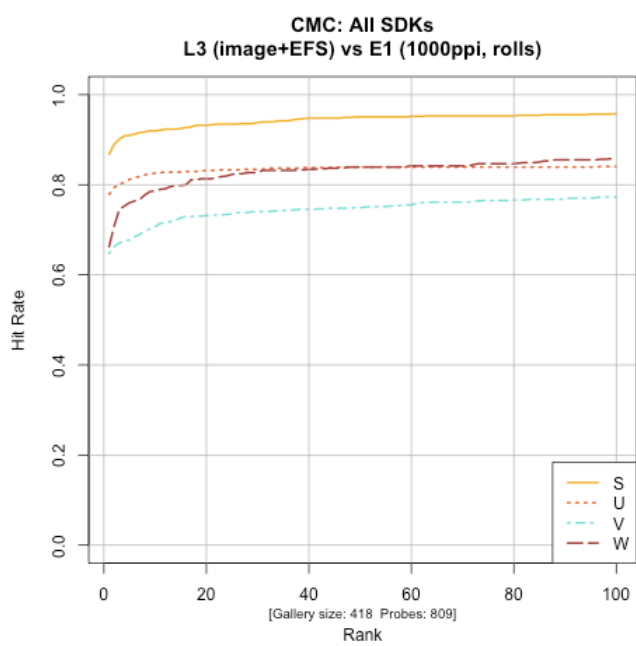
B-9 Score-based results

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. [6] 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.
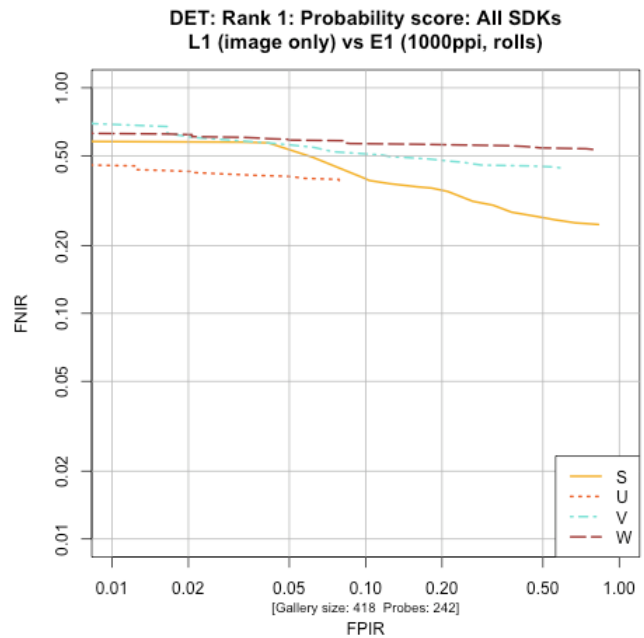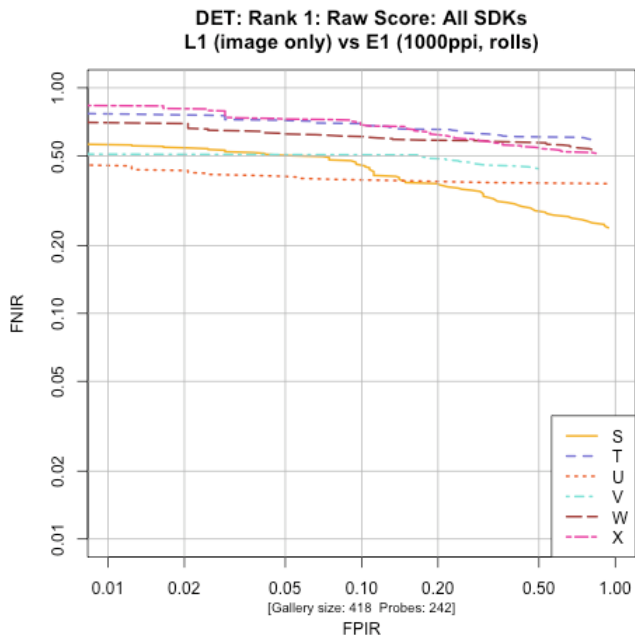
---

[6] *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.
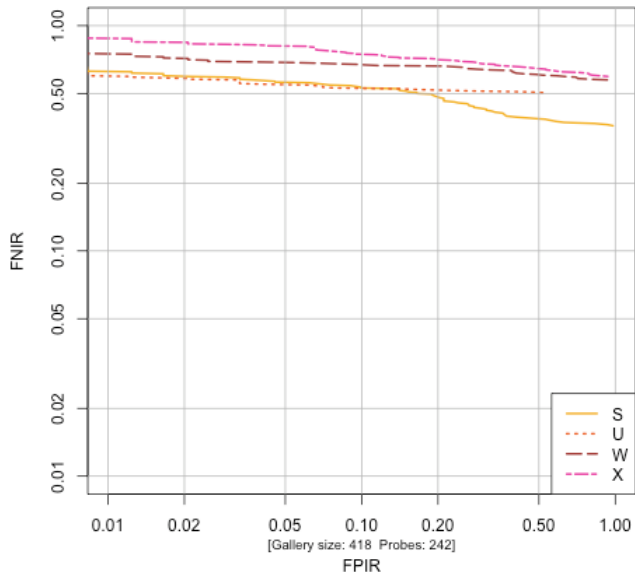
In practice, these charts show the effect of automatically eliminating candidates based on score. For example, in the first chart (Raw score DET for L1 vs E1), for participant S, the FNIR=0.236 @ FPIR=1.0, and reduces to FNIR=0.5 @ FPIR=0.05. What this means is that if a score threshold is set so that 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.
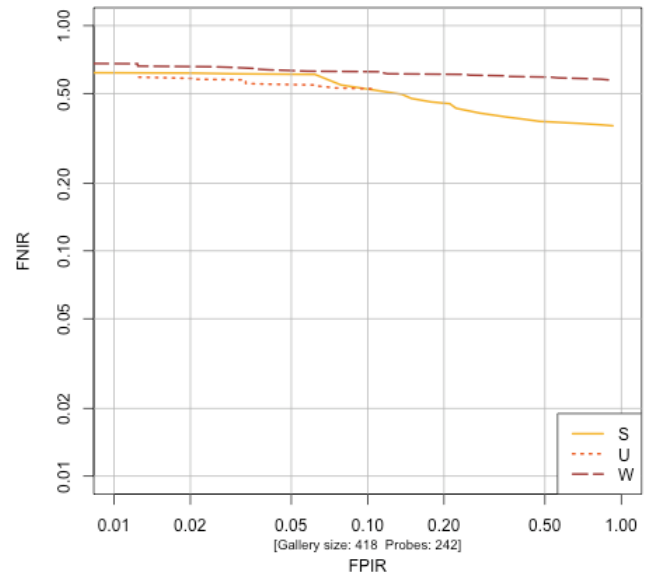
In each case, participants returned a raw score and a normalized score estimating the probability of a match. Not all participants returned probability scores.
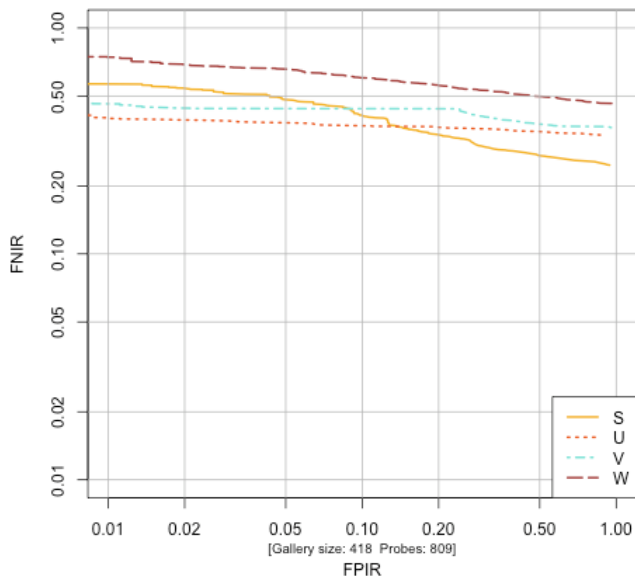
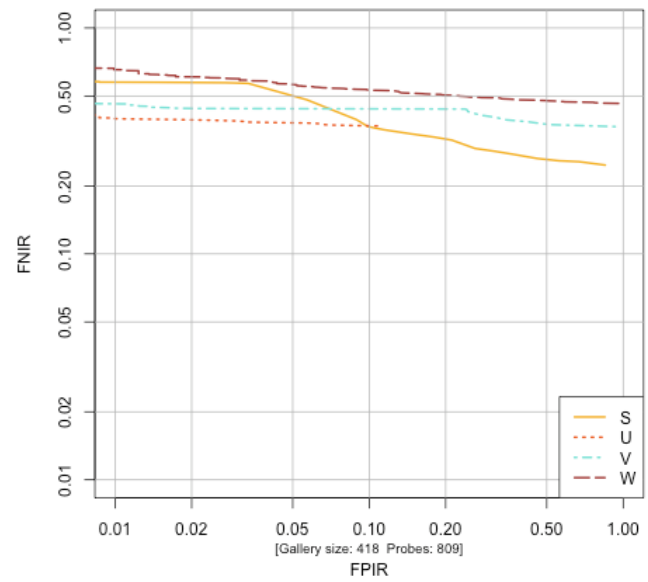DET: Rank 1: Raw Score: All SDKs
L1 (image only) vs E3 (1000ppi, flats)

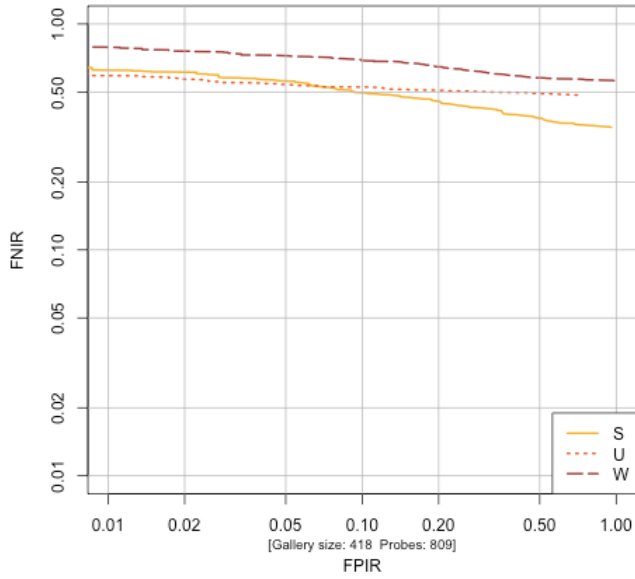DET: Rank 1: Probability score: All SDKs
L1 (image only) vs E3 (1000ppi, flats)

DET: Rank 1: Raw Score: All SDKs
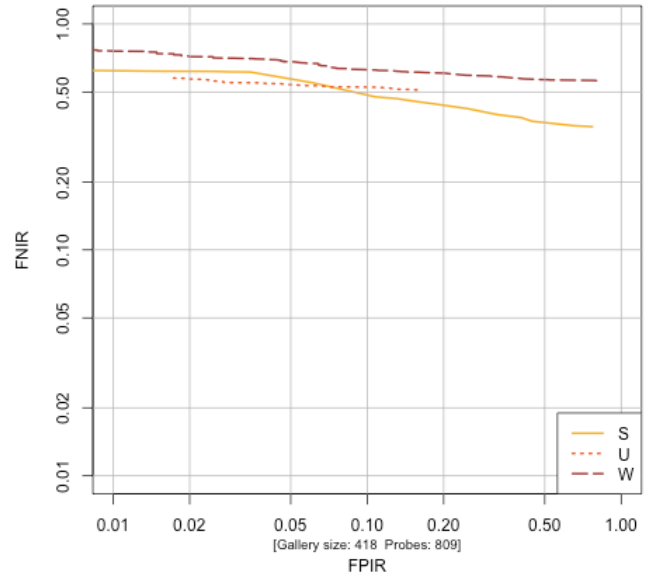L2 (image+minutiae) vs E1 (1000ppi, rolls)

DET: Rank 1: Probability score: All SDKs
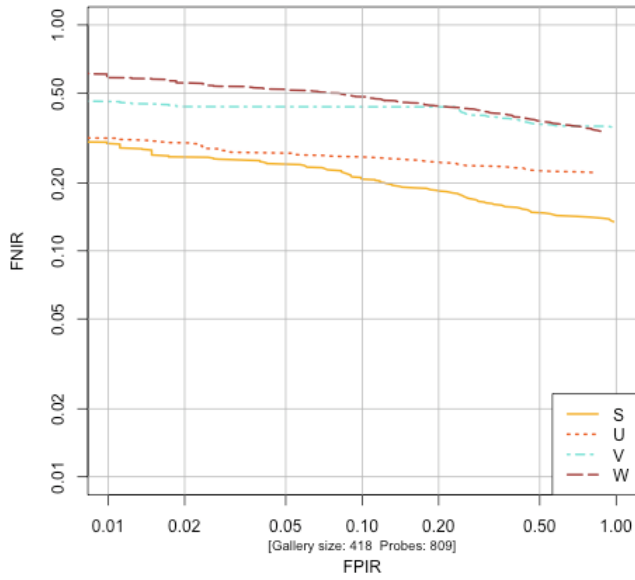L2 (image+minutiae) vs E1 (1000ppi, rolls)

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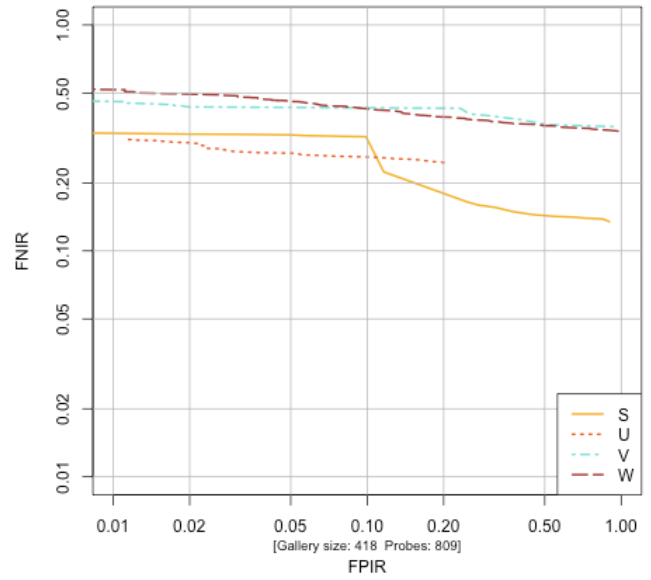**L2 (image+minutiae) vs E3 (1000ppi, flats)**

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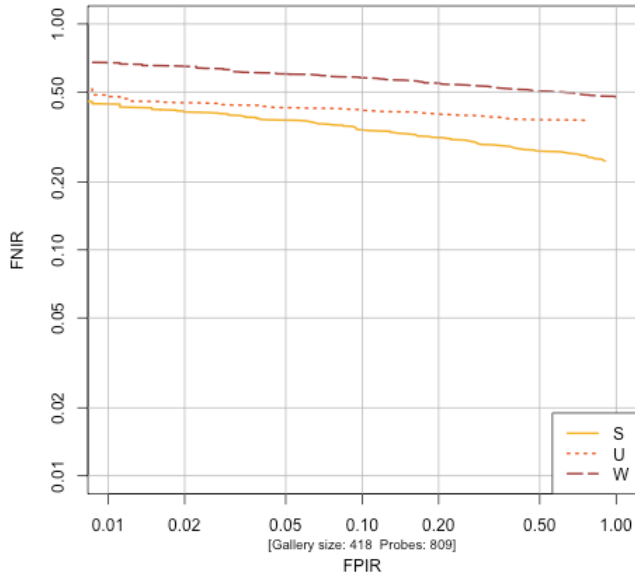**L2 (image+minutiae) vs E3 (1000ppi, flats)**

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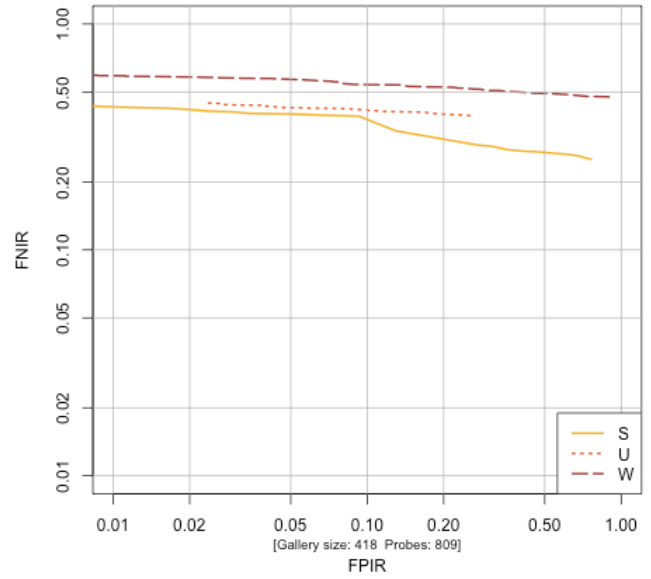**L3 (image+EFS) vs E1 (1000ppi, rolls)**

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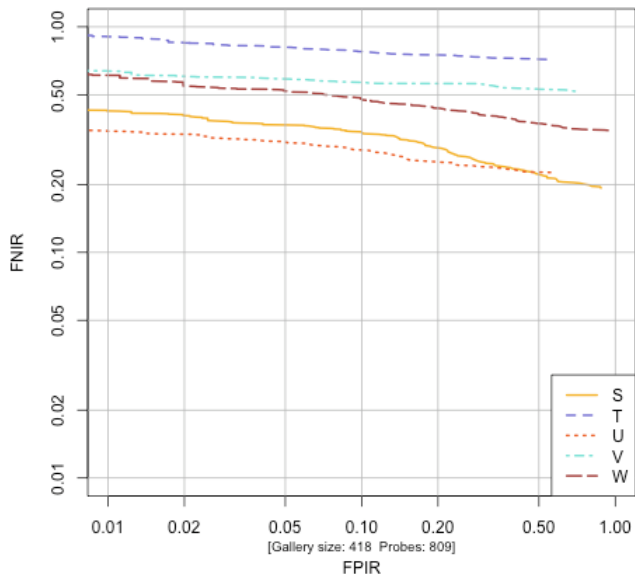**L3 (image+EFS) vs E1 (1000ppi, rolls)**

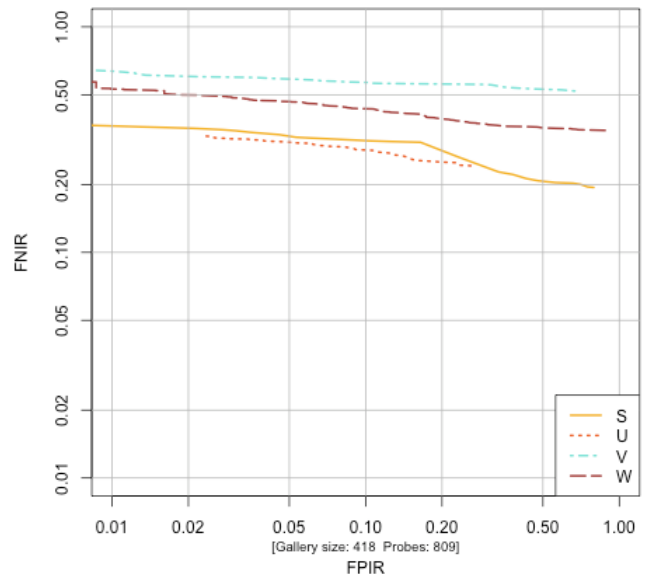DET: Rank 1: Raw Score: All SDKs
L3 (image+EFS) vs E3 (1000ppi, flats)

DET: Rank 1: Probability score: All SDKs
L3 (image+EFS) vs E3 (1000ppi, flats)

DET: Rank 1: Raw Score: All SDKs
L4 (EFS only) vs E1 (1000ppi, rolls)

DET: Rank 1: Probability score: All SDKs
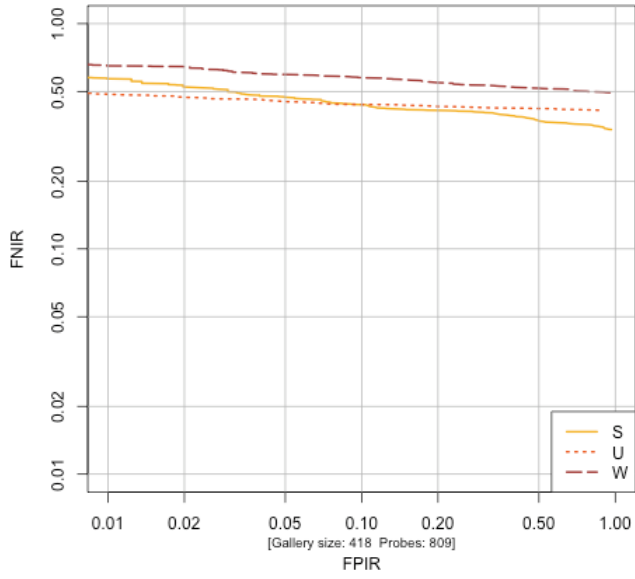L4 (EFS only) vs E1 (1000ppi, rolls)

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
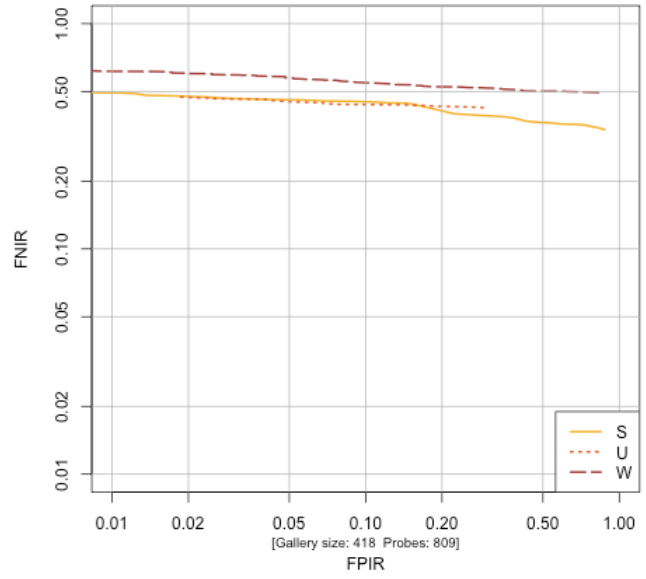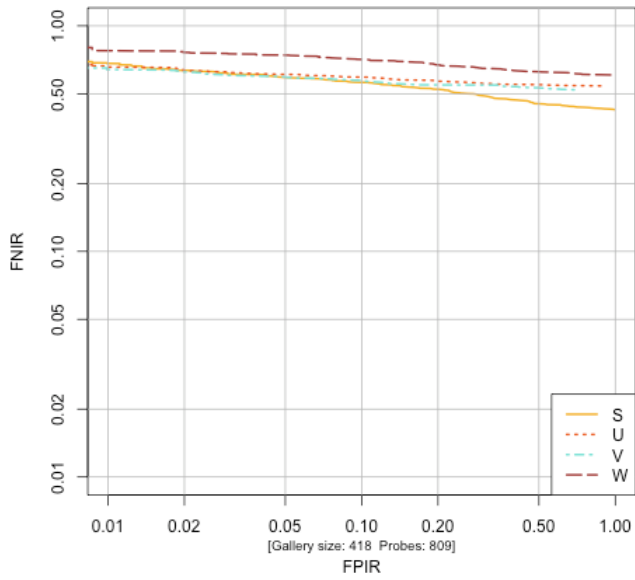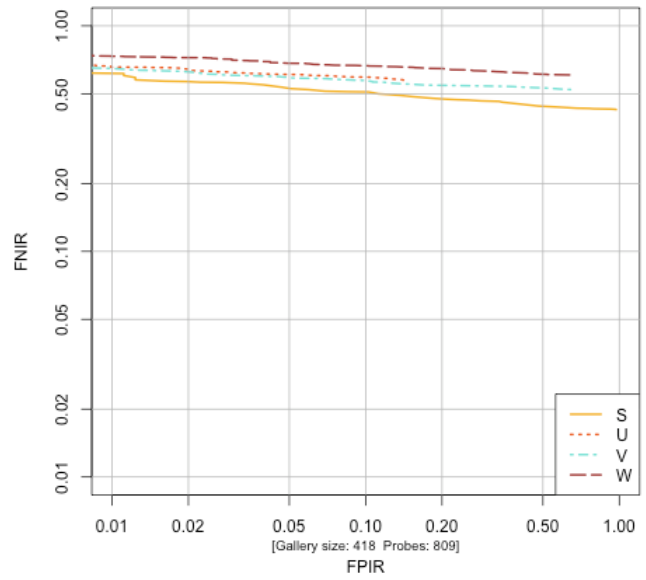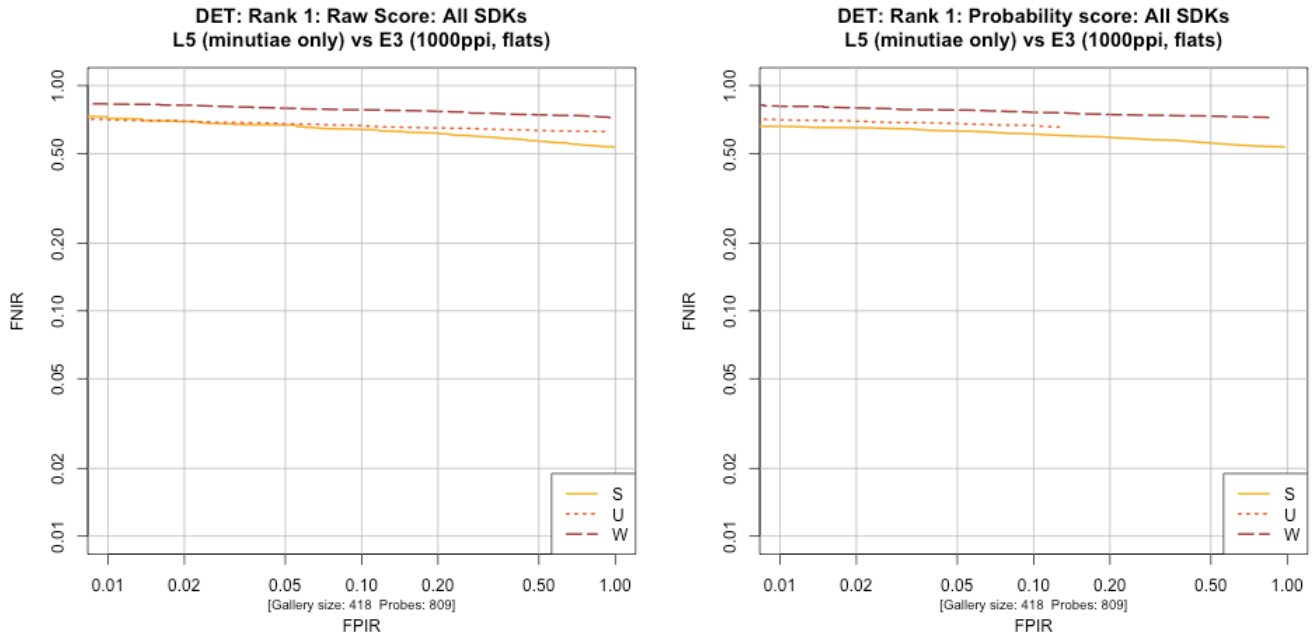**L5 (minutiae only) vs E1 (1000ppi, rolls)**

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

DET: Rank 1: Raw Score: All SDKs
L5 (minutiae only) vs E3 (1000ppi, flats)

DET: Rank 1: Probability score: All SDKs
L5 (minutiae only) vs E3 (1000ppi, flats)

B-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	-