ELFT-EFS Results

NIST Evaluation of Latent Fingerprint Technologies: Extended Feature Sets Evaluation #2

Michael Indovina, NIST
Vladimir Dvornychenko, NIST
Austin Hicklin, Noblis
George Kiebuzinski, Noblis

Overview

- ELFT-EFS is an open evaluation of the accuracy of automated latent fingerprint identification using features marked by experienced latent fingerprint examiners.
- ELFT-EFS is being conducted by NIST on behalf of the FBI and DHS
- Participants: SAGEM MorphoTrak, NEC, 3M Cogent, Warwick Warp, Sonda, and SPEX Forensics. (Results delayed for SPEX; not included here)
- 1066 latent images, with markup by certified latent examiners, were searched against a gallery of both rolled and plain exemplar fingerprints from approximately 100,000 subjects.
- The test evaluated the accuracy of searches when using: a) latent images alone; b) images combined with selected subsets of the Extended Feature Set (EFS) (defined in ANSI/NIST-ITL 2011); and c) selected features alone.

NIST Latent Testing – ELFT-EFS

- ELFT-EFS is part of NIST's Evaluation of Latent Fingerprint Technology (ELFT) testing program.
- ELFT-EFS evaluates the net accuracy of latent matching when using features marked by experienced human latent fingerprint examiners.
- ELFT-EFS is not per se a test of automatic EFS extraction (e.g. conformance).
- When image data are included in a search, automatic feature extraction may be used to any degree participants choose (e.g. in addition to, or in place of, manually specified features)
- A key result of ELFT-EFS is to measure the performance of differing "levels" of manually encoded feature-based searches, and because such markup is expensive, determine when image-only searching is sufficient

ELFT-EFS Testing -- Background

- ELFT-EFS Public Challenge (2009)
  - Anonymous open-book self-reported practice test on small dataset
  - Reported in Appendix to ELFT-EFS Evaluation #1 final report
- ELFT-EFS Evaluation #1 (2009-2011)
  - Technology evaluation using participant provided software, run on NIST hardware using sequestered datasets from multiple sources
  - Goal was to determine near-term benefits ("as currently implemented"), **not** long-term feasibility or accuracy
  - Final report (NISTIR 7775) published March 2011
- ELFT-EFS Evaluation #2 (2010-2012)
  - Same test plan and data as Evaluation #1 (with minor exceptions)
  - Evaluation #1 participants (and new participants) given chance to test revised algorithms
  - Final report (NISTIR 7859) published May 2012

Caveats

- The performance impact of any specific feature as measured in the this test may be limited for several reasons:
  - participants may not have yet developed approaches to utilize provided features
  - limited opportunity due to limited presence of such features in the data
  - human markup of such features may be ineffective for automated matching
- The results may not be applicable to other datasets and operational systems with different processing constraints.
- The relative performance of image-based and feature-based matching may be affected by differences in system resources (cost); image-only processing typically requires much greater resources.

Data

Sources of latent images

Source Name	# Latent Images	# Distinct Fingers	# Subjects	Description
Casework 1	368	368	272	Operational casework images
Casework 2	165	165	163	Operational casework images
WVU	440	440	383	Laboratory collected images
FLDS	93	93	15	Laboratory collected images
MLDS	*38*	*38*	*4*	*Laboratory collected images (small set of publicly releasable images for examples in reports)*
Total	1066	1066	826	

Latents

- All images were 8-bit grayscale, 1000ppi, uncompressed
- All latent images and their associated features were contained in ANSI/NIST records (no system specific transaction files used)
- Finger position was never specified
- Note: NIST is always receptive to more good latent data

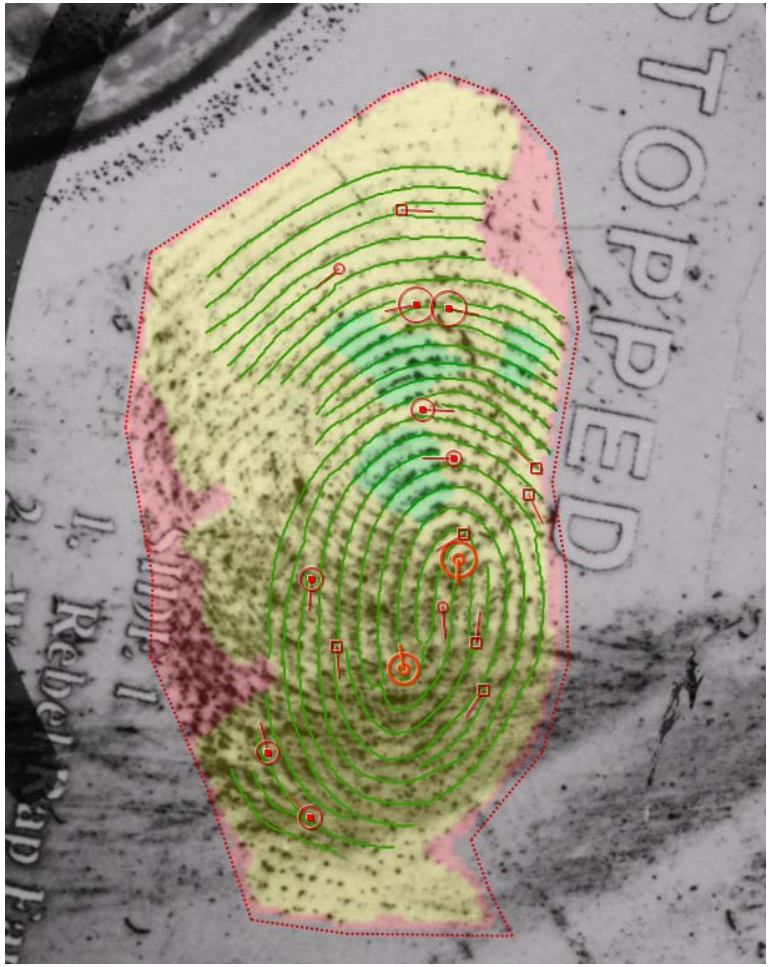
Examiner feature markup

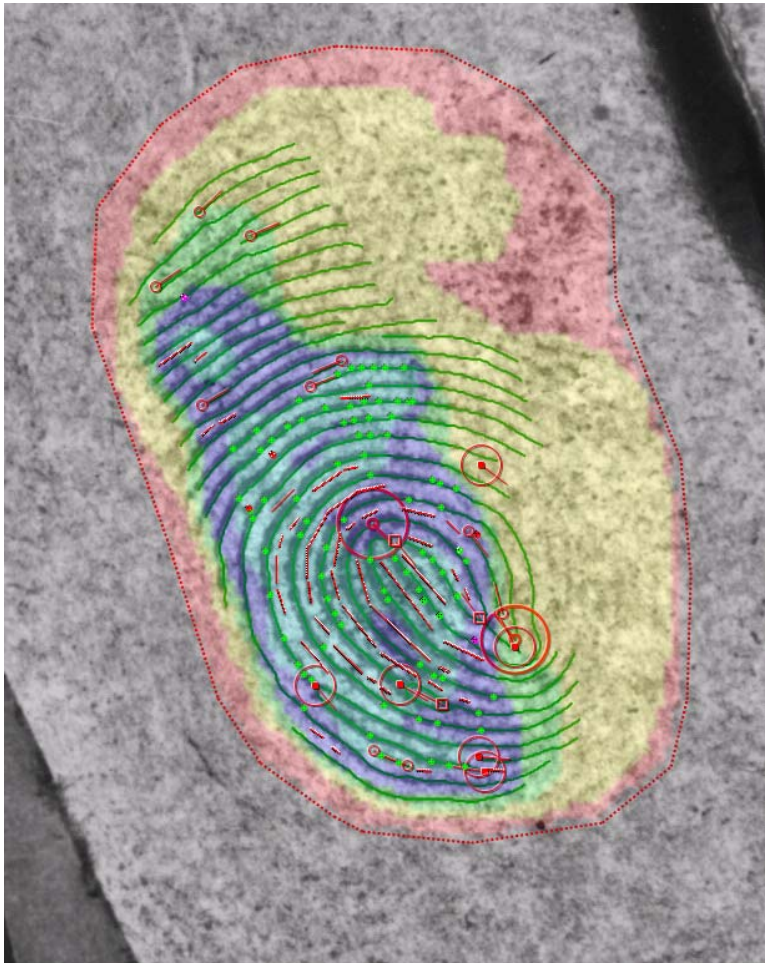
- The latent IAFIS/EFS features were marked by 21 IAI certified latent print examiners (CLPE)
- Features were defined in ANSI/NIST-ITL 1-2011, in the Extended Feature Set (EFS); and markup guidelines developed in “EFS Markup Instructions”
- EFS markup included: ridge quality maps, creases, dots, incipient ridges, ridge edge protrusions, and pores, minutiae with ridge counts, cores, deltas, and pattern class
- A subset of latents had skeletons marked (including associated ridge flow maps)
- No vendor-specific rules for feature encoding were used ; all encoding was made in compliance with the EFS specification
- Features were marked in latent images without reference to exemplars, except for the GroundTruth (GT) dataset
- Which features to use was left entirely to the participants (suppliers of SDKs); it is likely that all available features were not used by participants

Latent Feature Subsets

Subset	Description	Image
LA	Image only	With Image
LB	Region of Interest (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 (equivalent to IAFIS LFFS)	Without Image

Examples of image markup





ELFT-EFS Latent Datasets

- **Baseline**
  - 1,066 latents and associated feature markup
  - Used only for subsets LA (image only), LE (image + EFS) and LG (minutiae only)
- **Baseline-QA**
  - 418 latent subset of Baseline which underwent additional QA review
  - Used for all subsets
  - All latents have EFS skeleton (subset LF) available
- **GT (Ground Truth)**
  - Variation of the markup of the latents from Baseline-QA
  - Used only for subsets LA, LE, and LG
  - GT has “ideal” minutiae markup performed **with** reference to the exemplars

Exemplars

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

Exemplars

- All exemplars are 8-bit grayscale, 500ppi, compressed with WSQ
- All exemplars are in ANSI/NIST files
- Background (non-mated exemplars) images from operational databases
  - Sourced from ink and optical livescan; rolled and slap impressions
  - Plains segmented (by auto-segmentation) from slaps
- Foreground (mated exemplars)
  - Plains segmented using manual review and correction from slaps
  - Association with latent **not** determined by AFIS
- No features included with the exemplars at time of enrollment

Test Procedures

Latent Matching Software

- Each participant submitted a set of SDKs (Software Development Kits) that provided the interfaces defined by the ELFT-EFS-1 API
- 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 Linux or Windows, sequential or multithreaded, and utilize either 32 or 64-bit execution mode.

- All tests were run on NIST hardware (array of blade servers)

- All searches returned a candidate list with a fixed length of one hundred (100) candidates.

Timing

Exemplar feature extraction	**100 seconds per 10-finger exemplar set (rolled or pre-segmented slap)**
Latent enroll	**120 seconds per latent**
Search	For feature searches (subsets LC-LG): **25ms/exemplar set, per latent, per core** (400 single finger matches/sec; an exemplar set consists of either 10 rolled or 10 plain prints)
	For image searches (subsets LA-LB): **50ms/exemplar set, per latent, per core** (200 single finger matches/sec)

Rank-based Results

ELFT-EFS Evaluation #2 Participants

- A: SAGEM MorphoTrak
- B: NEC
- C: 3M Cogent
- D: Sonda Technologies
- E: Warwick Warp
- *F: SPEX Forensics*

NIST

noblis
For the best of reasons

Image only

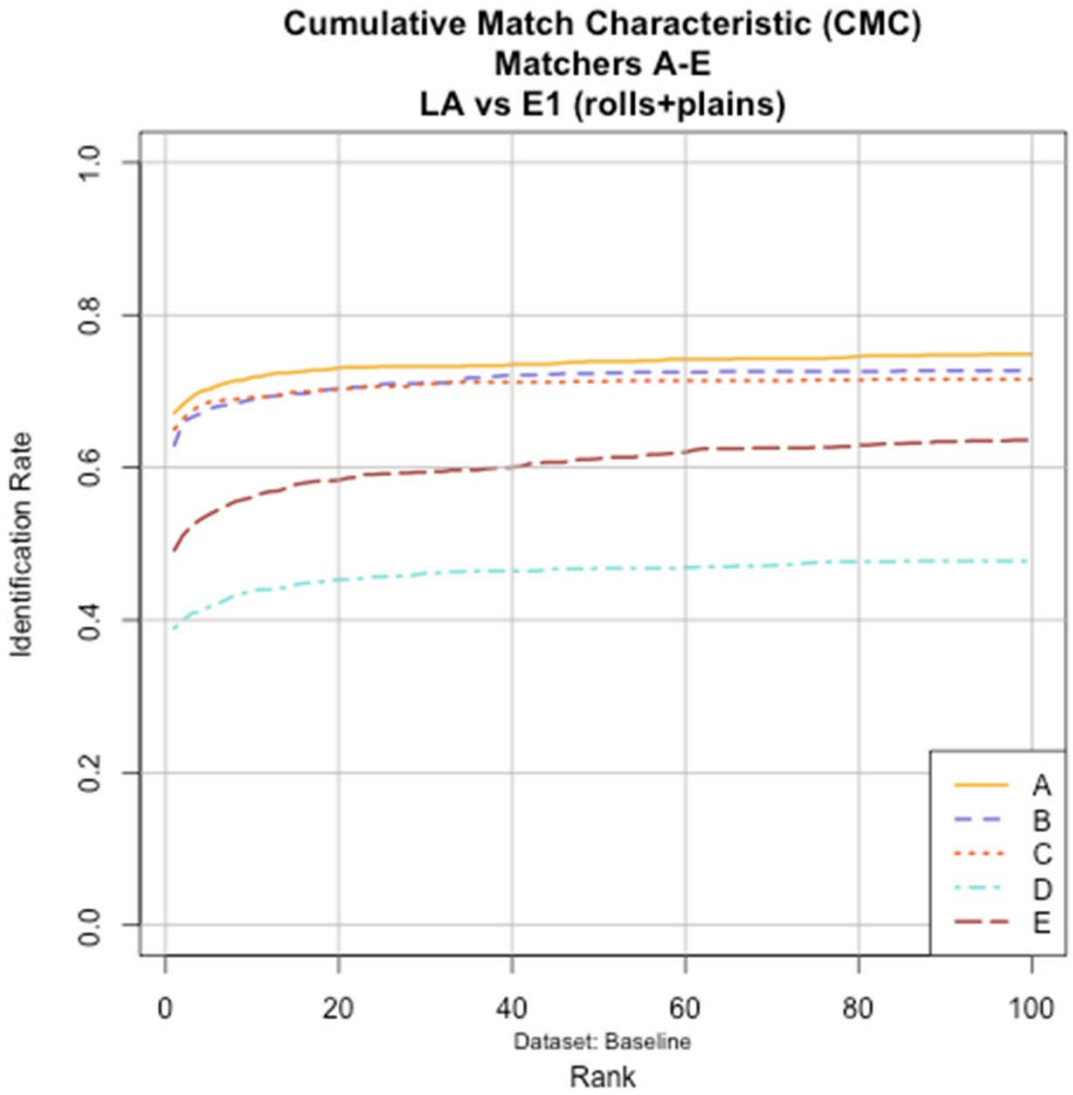
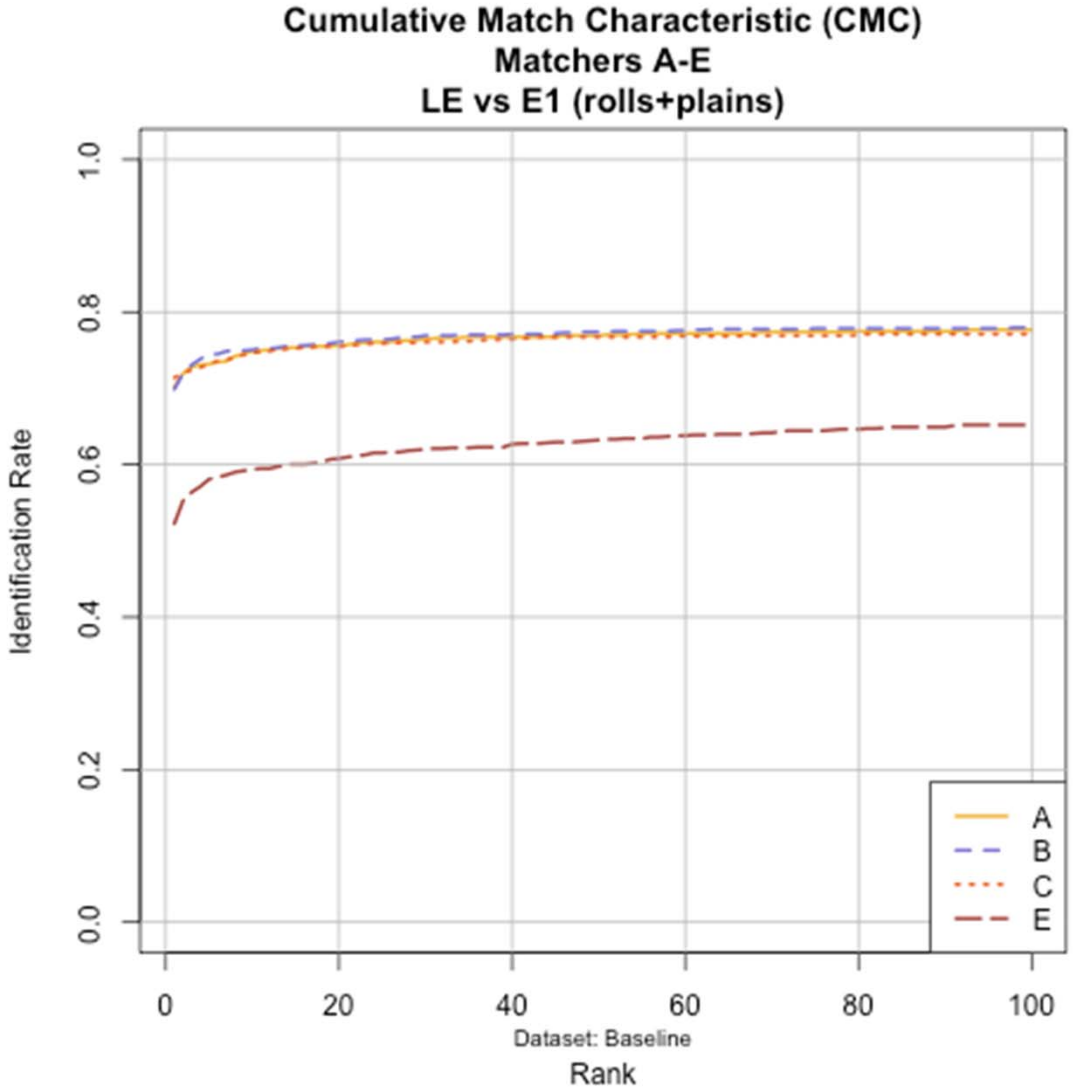
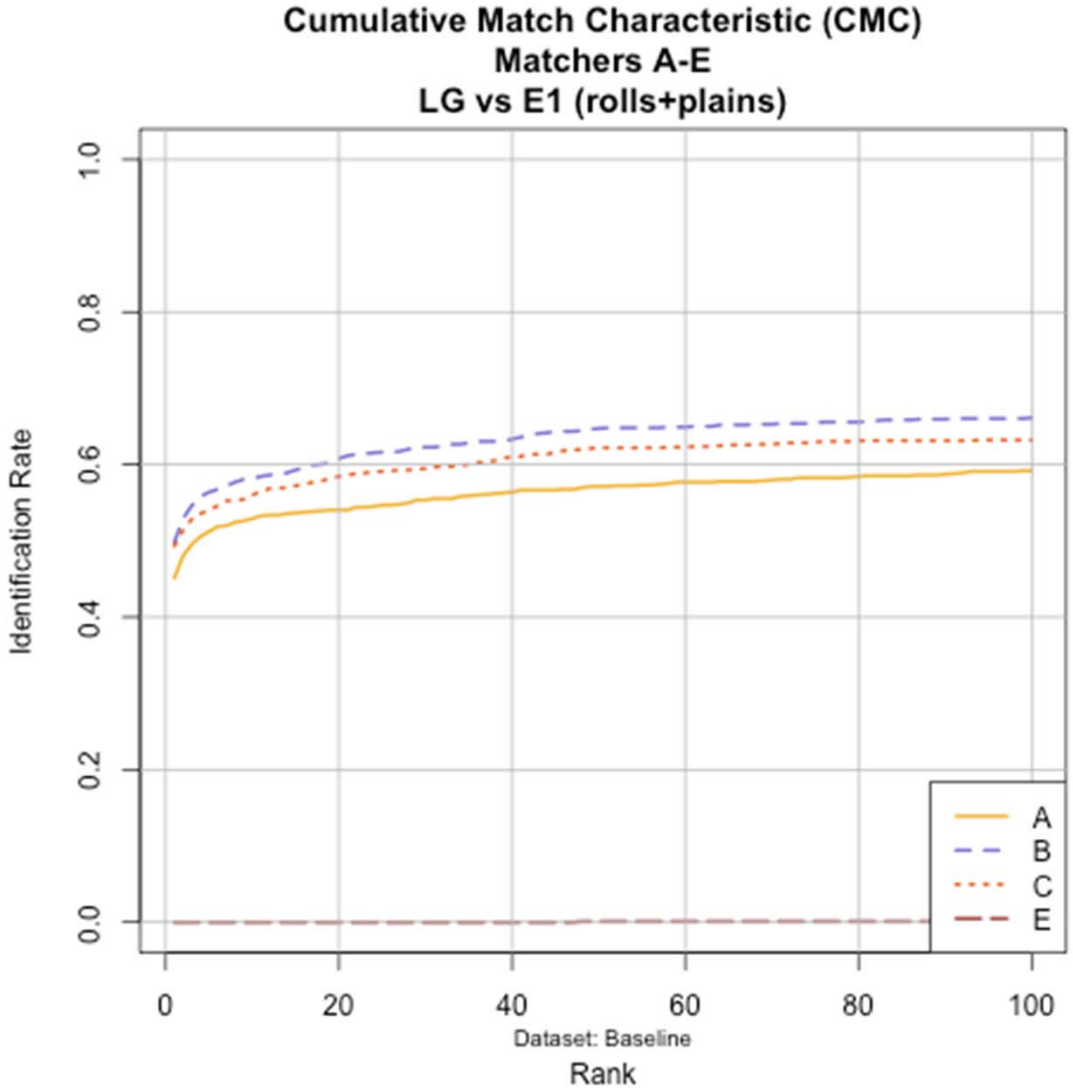

Cumulative Match Characteristic (CMC)
Matchers A-E
LA vs E1 (rolls+plains)
Identification Rate
0.0
0.2
0.4
0.6
0.8
1.0
0
20
40
60
80
100
Dataset: Baseline
Rank
A
B
C
D
E

Image + Extended feature set (no skeleton)



IAFIS minutiae + ridge counts (no image)



Summary of rank-1 results

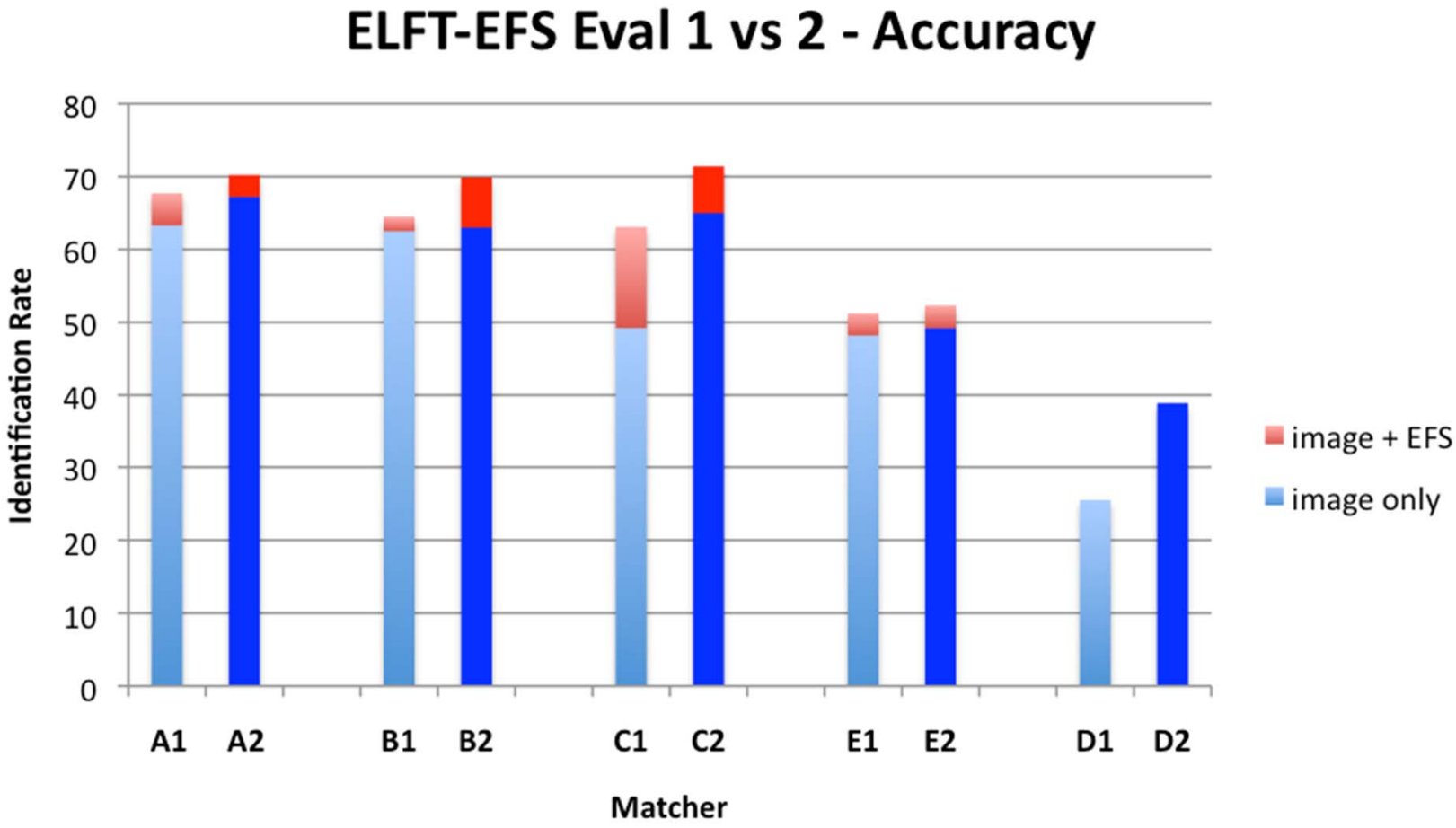
(searching E1: 100k roll+slap subjects; latent dataset Baseline)

1st place 2nd place 3rd place

	Latent Subset		
	LA	**LE**	**LG**
	Image only	Image + EFS	Minutiae only
A	67.2	70.2	45.1
B	63.0	69.9	49.8
C	65.0	71.4	49.3
D	38.9	*n/a*	*n/a*
E	49.2	52.3	0.0

Summary of rank-1 results

(searching E1: 100k roll+slap subjects; latent dataset Baseline)



Summary of rank-1 results

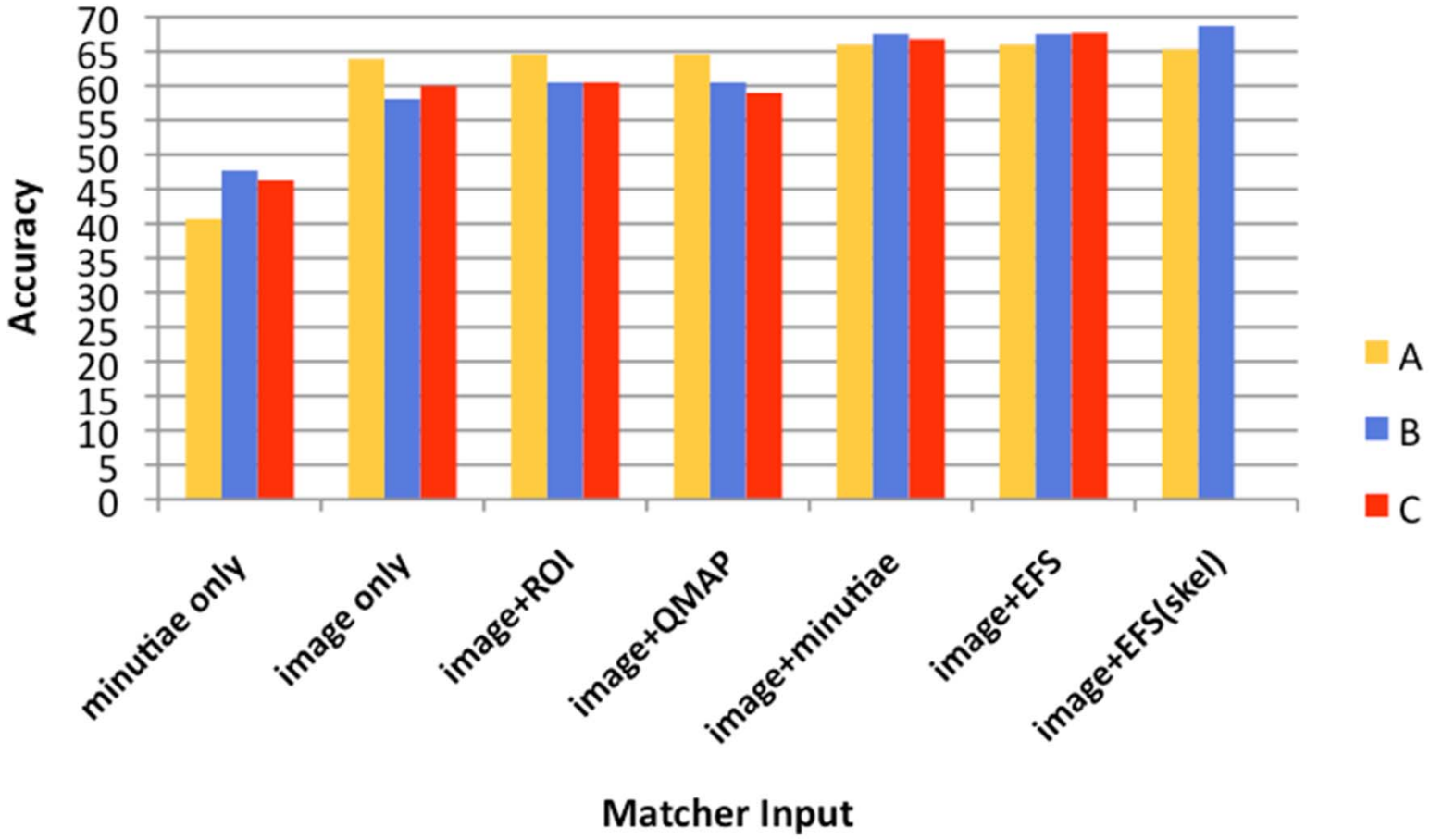
(searching E1: 100k roll+slap subjects; latent dataset Baseline-QA)

1st place 2nd place 3rd place

	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	63.4	64.1	64.1	65.6	65.6	64.8	40.4
B	57.7	60.1	60.1	67.0	67.0	68.2	47.4
C	59.6	60.1	58.6	66.3	67.2	*n/a*	45.9
D	31.8	23.9	*n/a*	*n/a*	*n/a*	*n/a*	*n/a*
E	44.0	46.9	47.1	46.9	47.1	48.3	0.0

Summary of rank-1 results

(searching E1: 100k roll+slap subjects; latent dataset Baseline-QA)

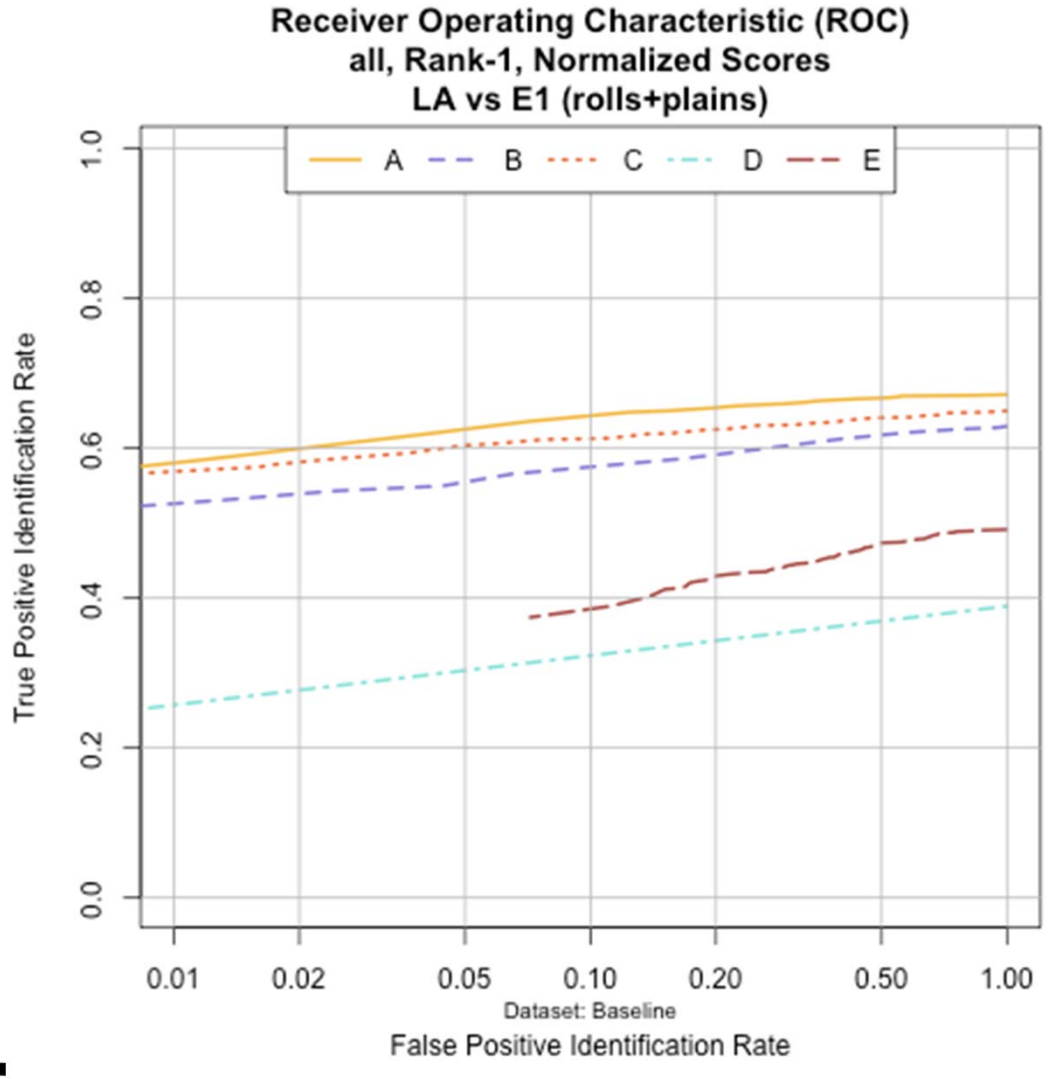


Summary of Rank 1 Results

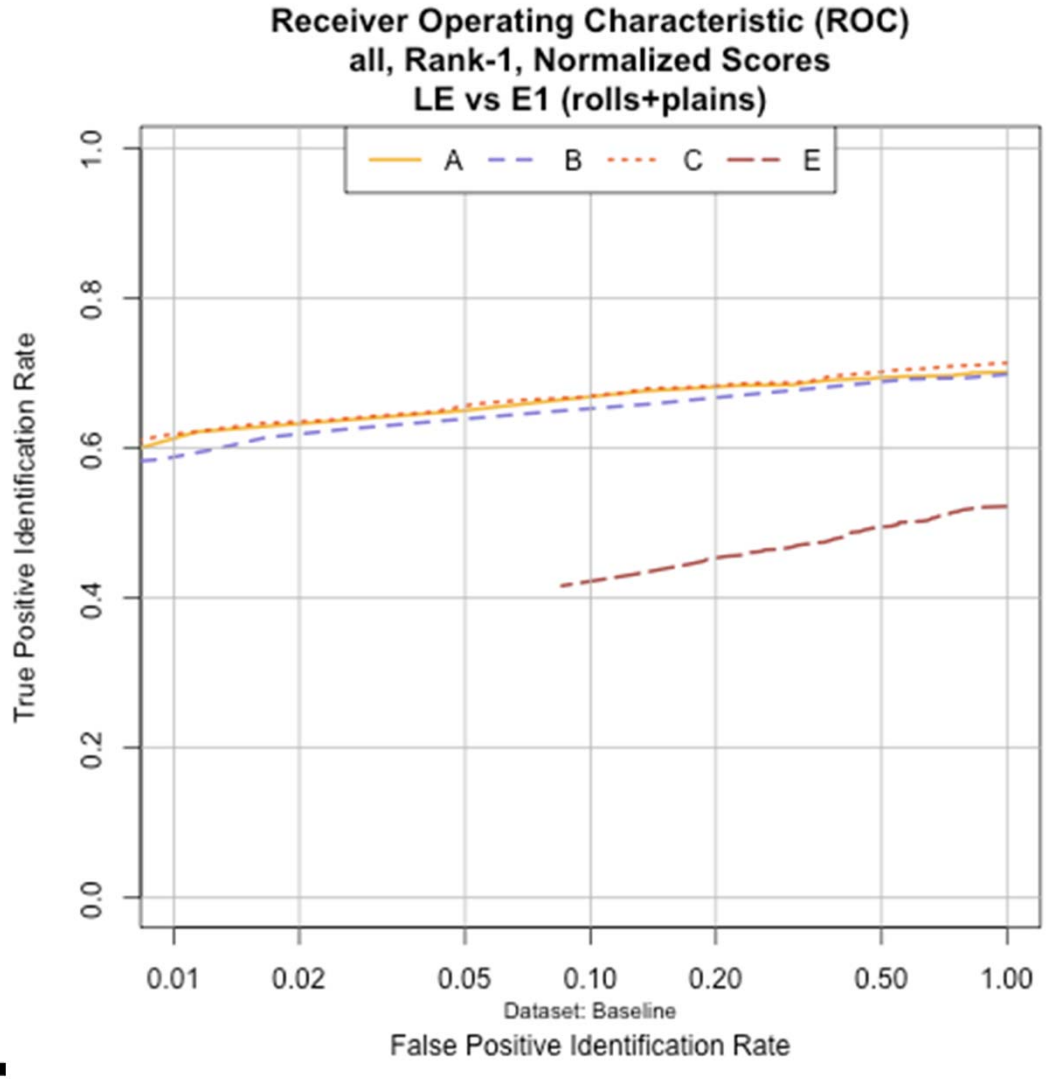
- **Minutiae-only searching performance:**
  - Minutiae-only < Image-only < Image+Features — for all matchers
- **Image-only searching:**
  - Cogent and Sonda improved about 13 – 16% over Evaluation #1
- **Image + Features searching:**
  - Best performance overall was 71.4% using image + EFS
  - The addition of features to the image (Minutiae or EFS) provided accuracy gains of 3-7% for the Baseline (1,066 latent) dataset
  - Apparently, Sagem & NEC did not utilize the additional EFS features (LD=LE performance)
  - Skeletons (LF) were only beneficial to NEC and Warwick (and made performance worse for Sagem)

Score-based results

LA: Image only



LE: Image + EFS


Receiver Operating Characteristic (ROC)
all, Rank-1, Normalized Scores
LE vs E1 (rolls+plains)
A
B
C
E
True Positive 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
Dataset: Baseline
False Positive Identification Rate

LG: IAFIS Minutiae + ridge counts (no image)

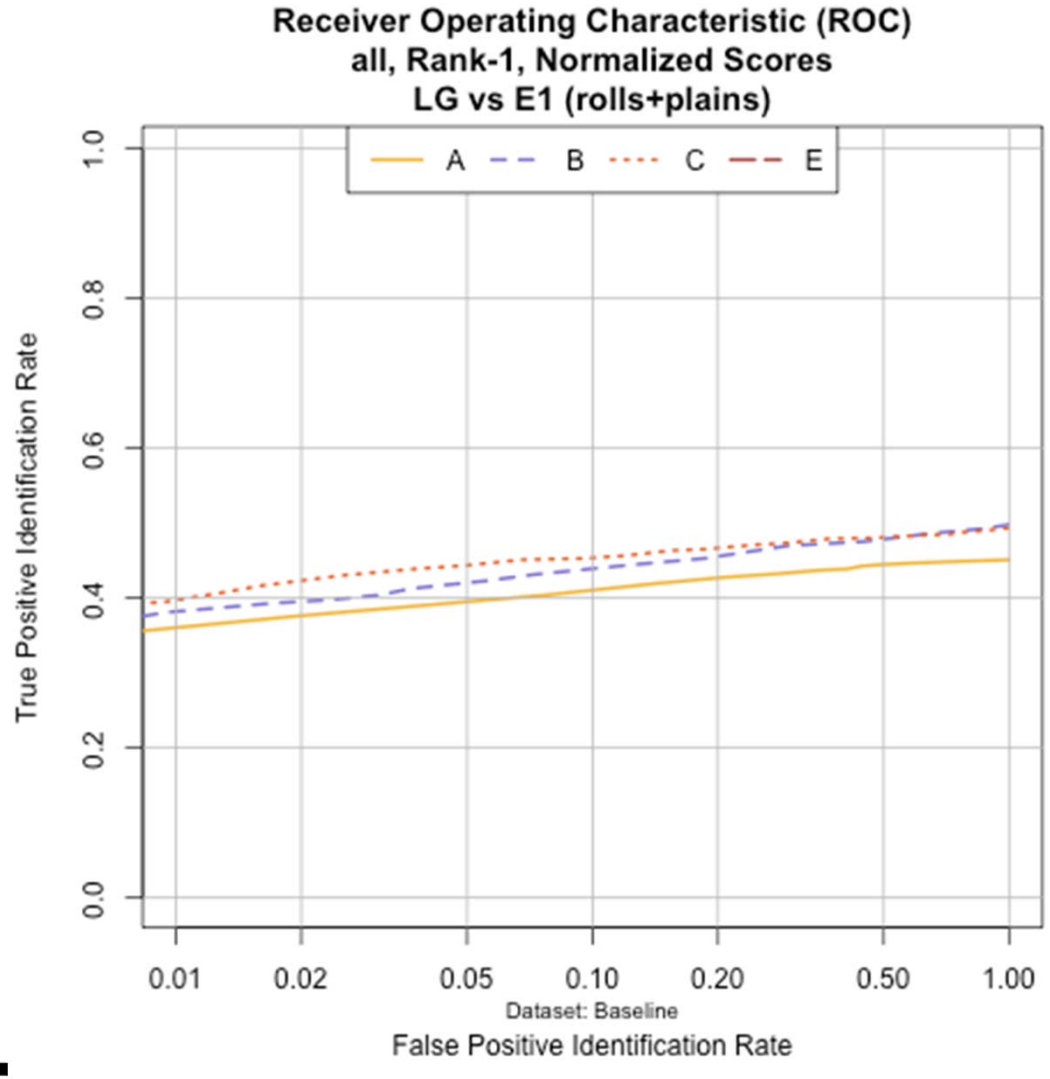
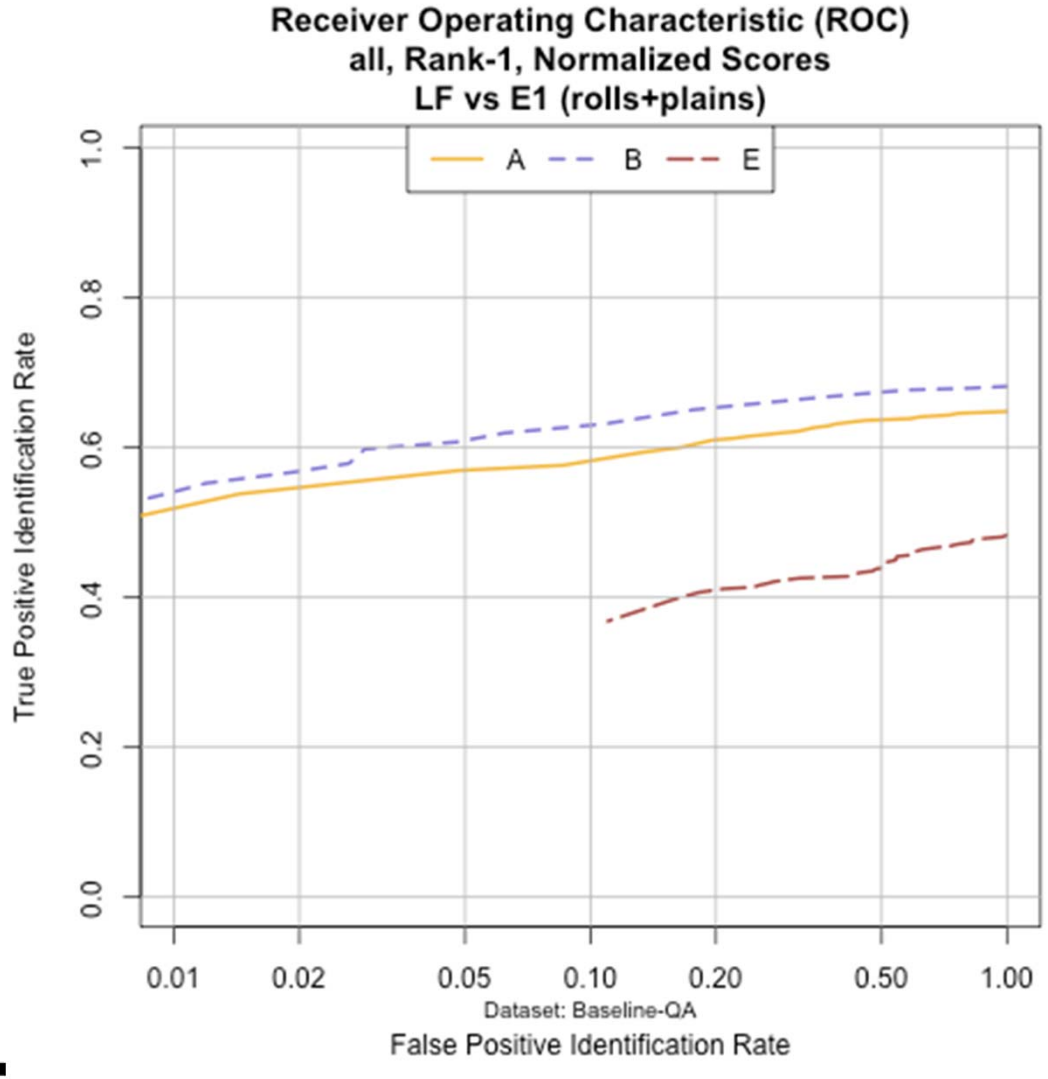


Image + Extended feature set + skeleton



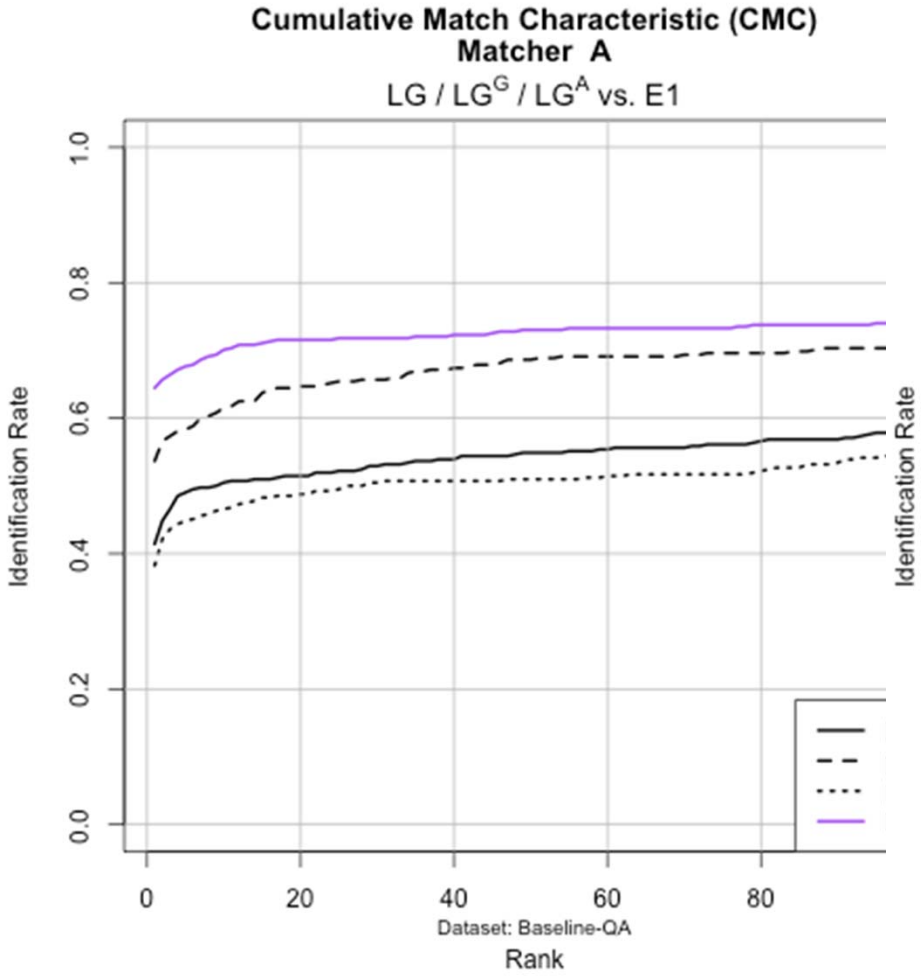
Summary of Score-based Results

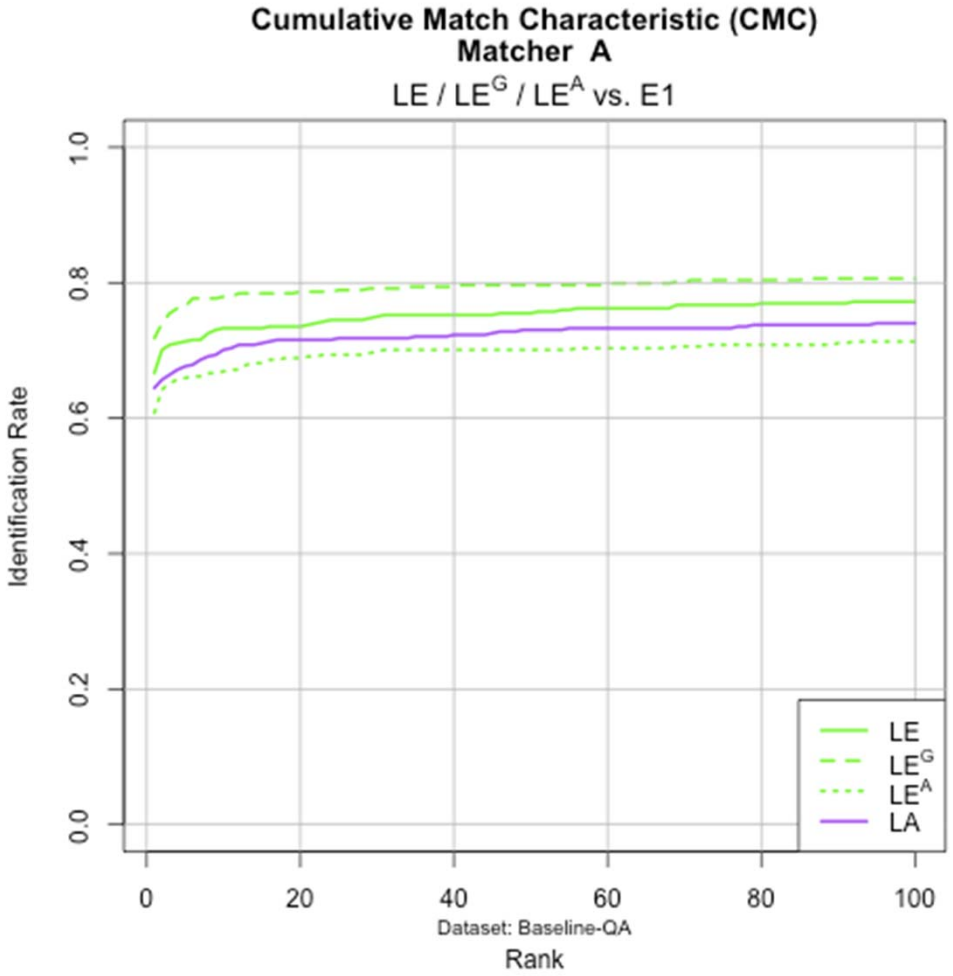
- Score-based results
  - are more scalable than rank-based results, and show how accuracy is affected by an increase in gallery size
  - show feasibility of automated candidate list reduction
  - show feasibility of improved candidate list management for reverse latent searches (i.e. unsolved latent file)

- At 1% FPIR only 1 in 100 candidate lists will contain a false candidate ; so that typically the candidate list is either empty, or contains only one candidate
- At 1% FPIR, identification rates for Baseline drop a average 9.0 pct-pts ; Least drop was for Image Only, greatest for Minutiae Only.

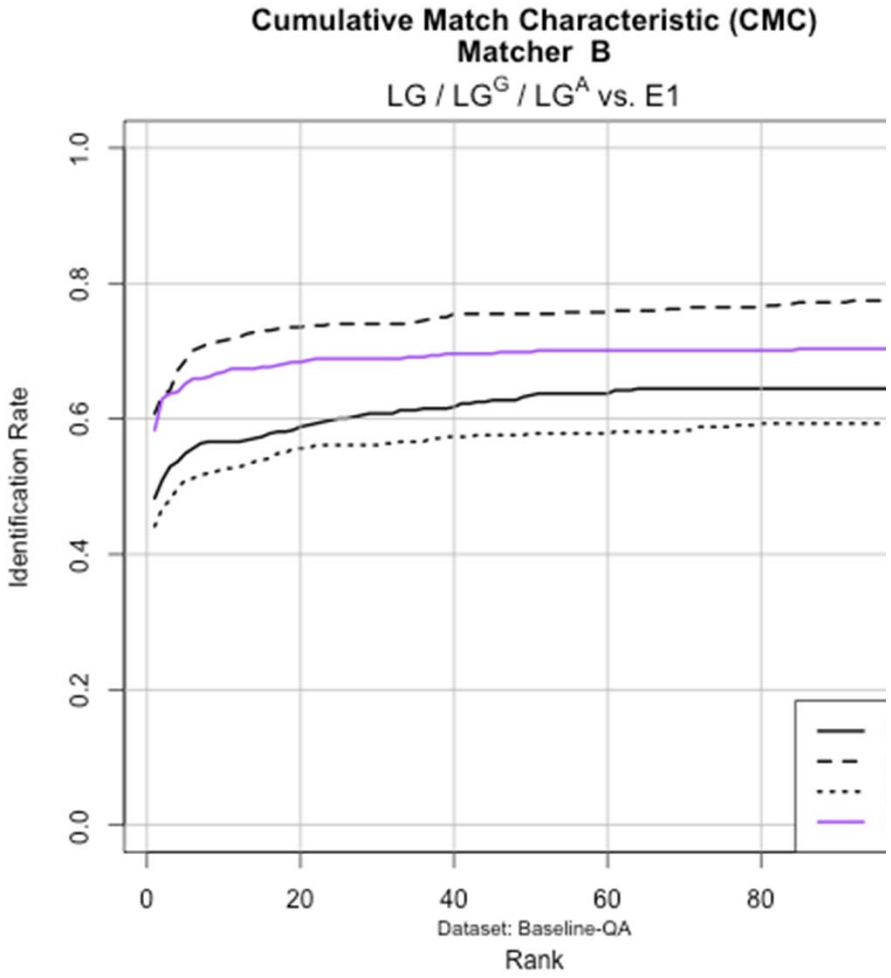
Effect of “Ground Truth” markup by SDK

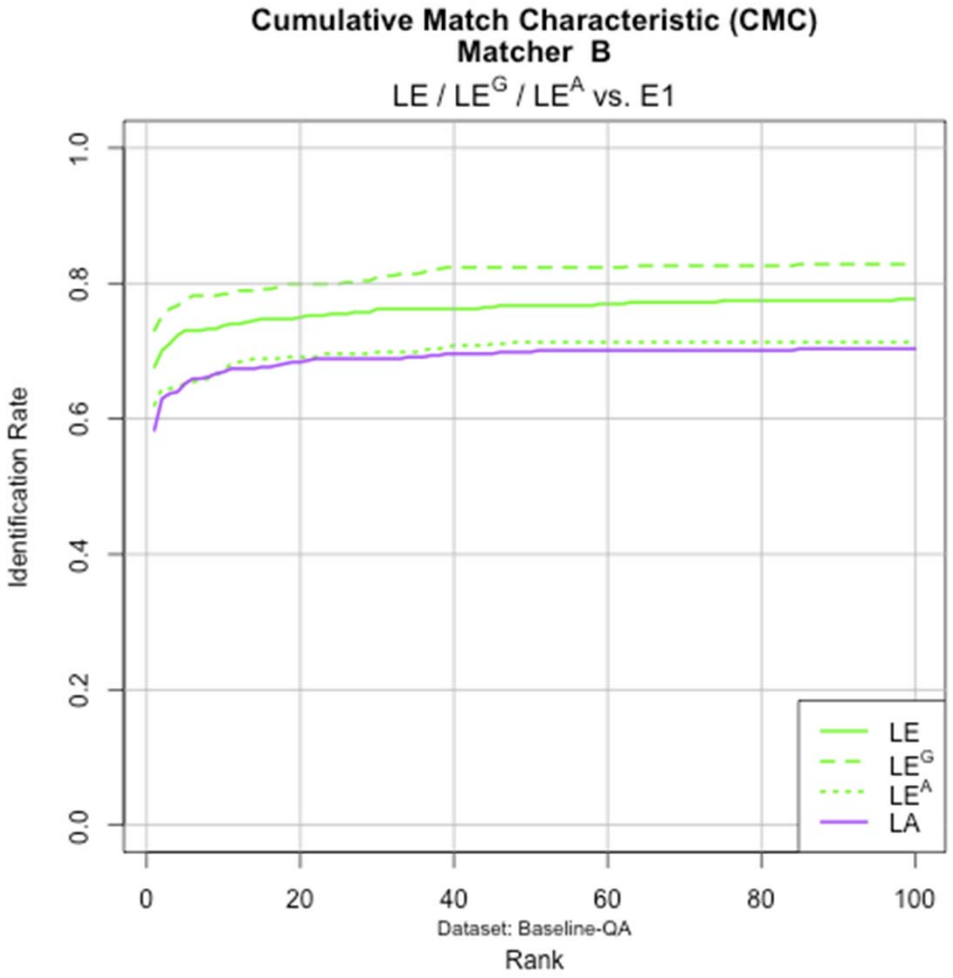
Matcher A: Effect of ground truth markup



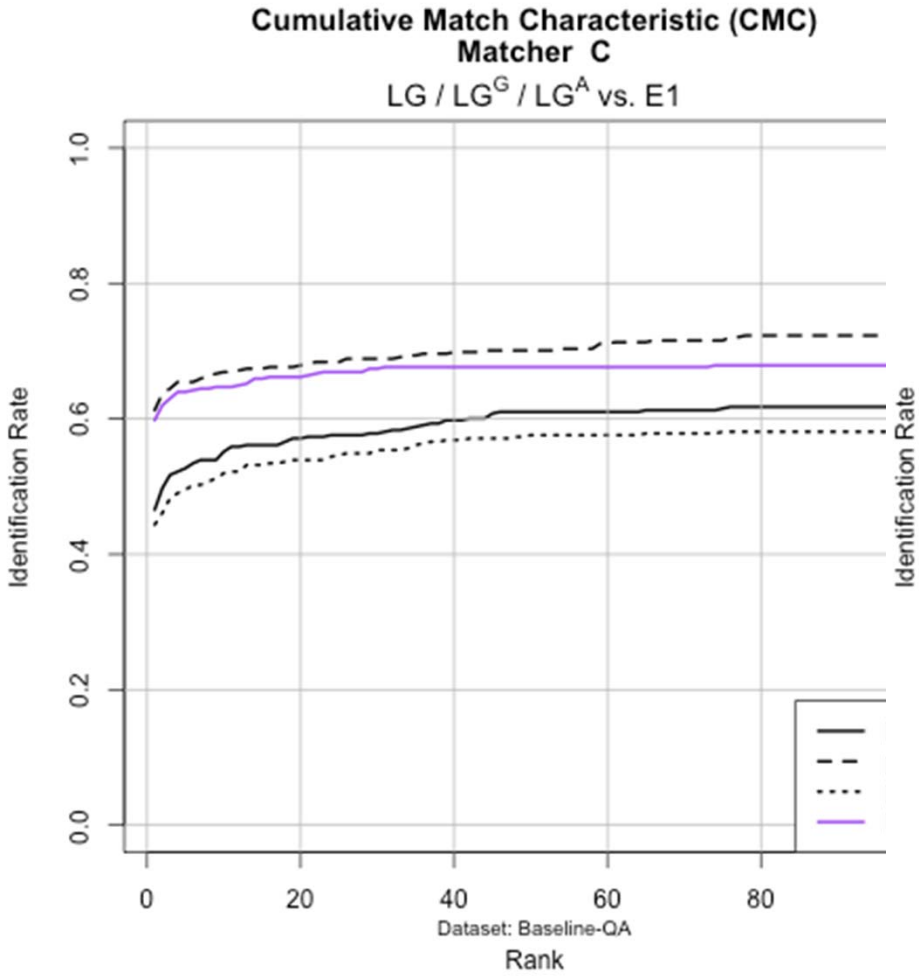


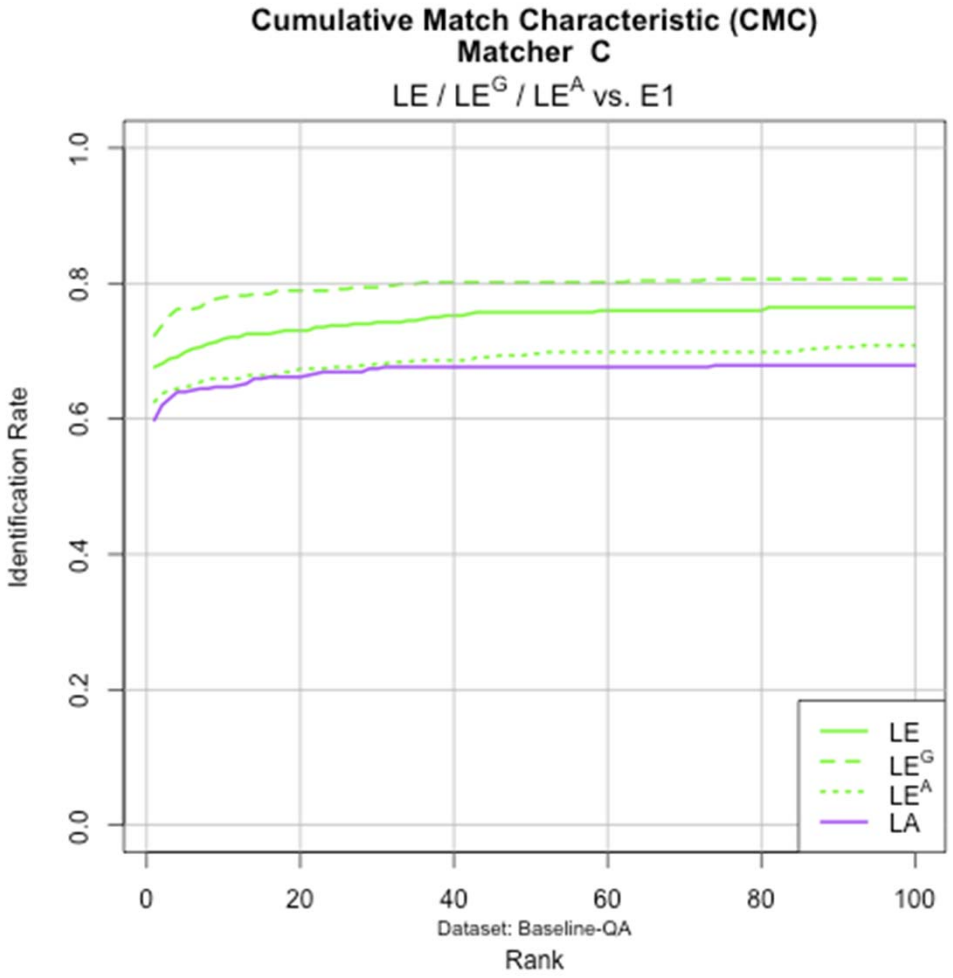
Matcher B: Effect of ground truth markup





Matcher C: Effect of ground truth markup





Summary of Effect of "Ground Truth" markup

- These results
  - show the difference between an operationally infeasible ideal markup, and actual markup by latent examiners
  - provide a measure of the impact of latent examiner variation in marking minutia on search performance

- The "GT" results were beneficial for SDKs 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 4-6%
- For latent subtest LG the difference in hit rate was about 12-15%: the differences between the markups had a direct impact on accuracy, since the matcher had no recourse to the image

Effect of Using *Rolled* vs. *Flat* Exemplar Types

Image only

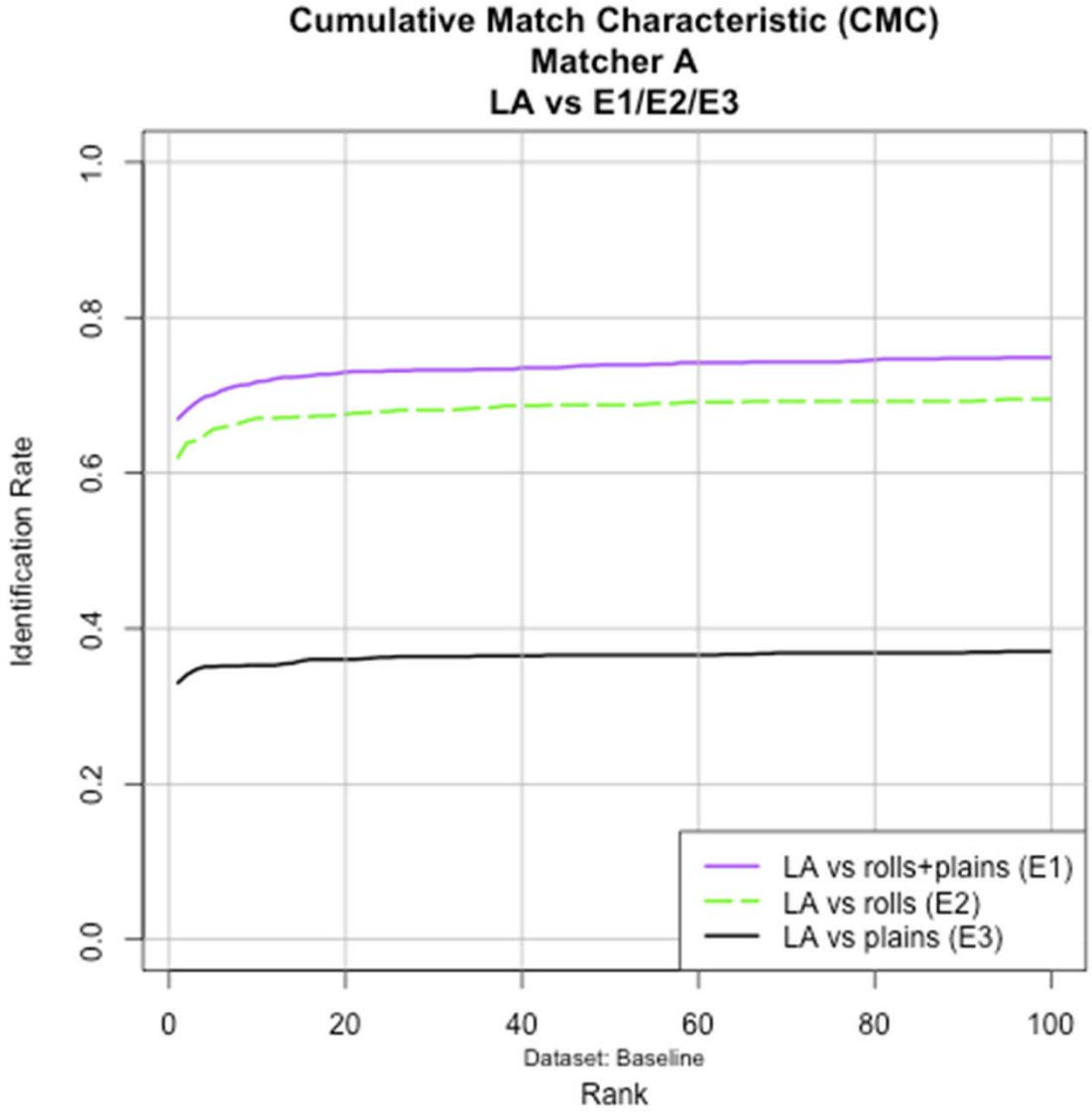

Cumulative Match Characteristic (CMC)
Matcher A
LA vs E1/E2/E3
Identification Rate
0.0
0.2
0.4
0.6
0.8
1.0
0
20
40
60
80
100
Dataset: Baseline
Rank
LA vs rolls+plains (E1)
LA vs rolls (E2)
LA vs plains (E3)

Image only

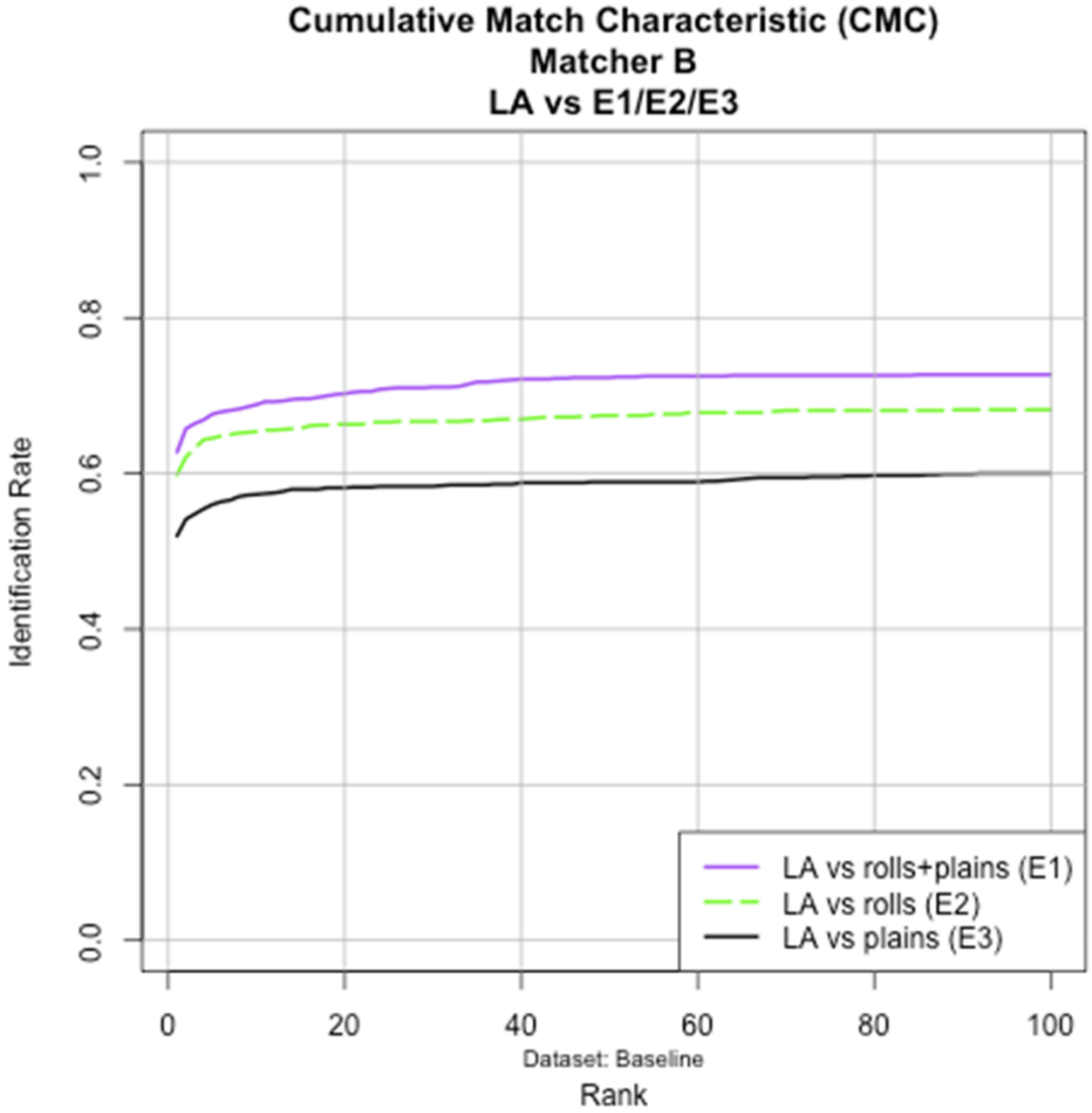

Cumulative Match Characteristic (CMC)
Matcher B
LA vs E1/E2/E3
Identification Rate
0.0
0.2
0.4
0.6
0.8
1.0
0
20
40
60
80
100
Dataset: Baseline
Rank
LA vs rolls+plains (E1)
LA vs rolls (E2)
LA vs plains (E3)

Image only

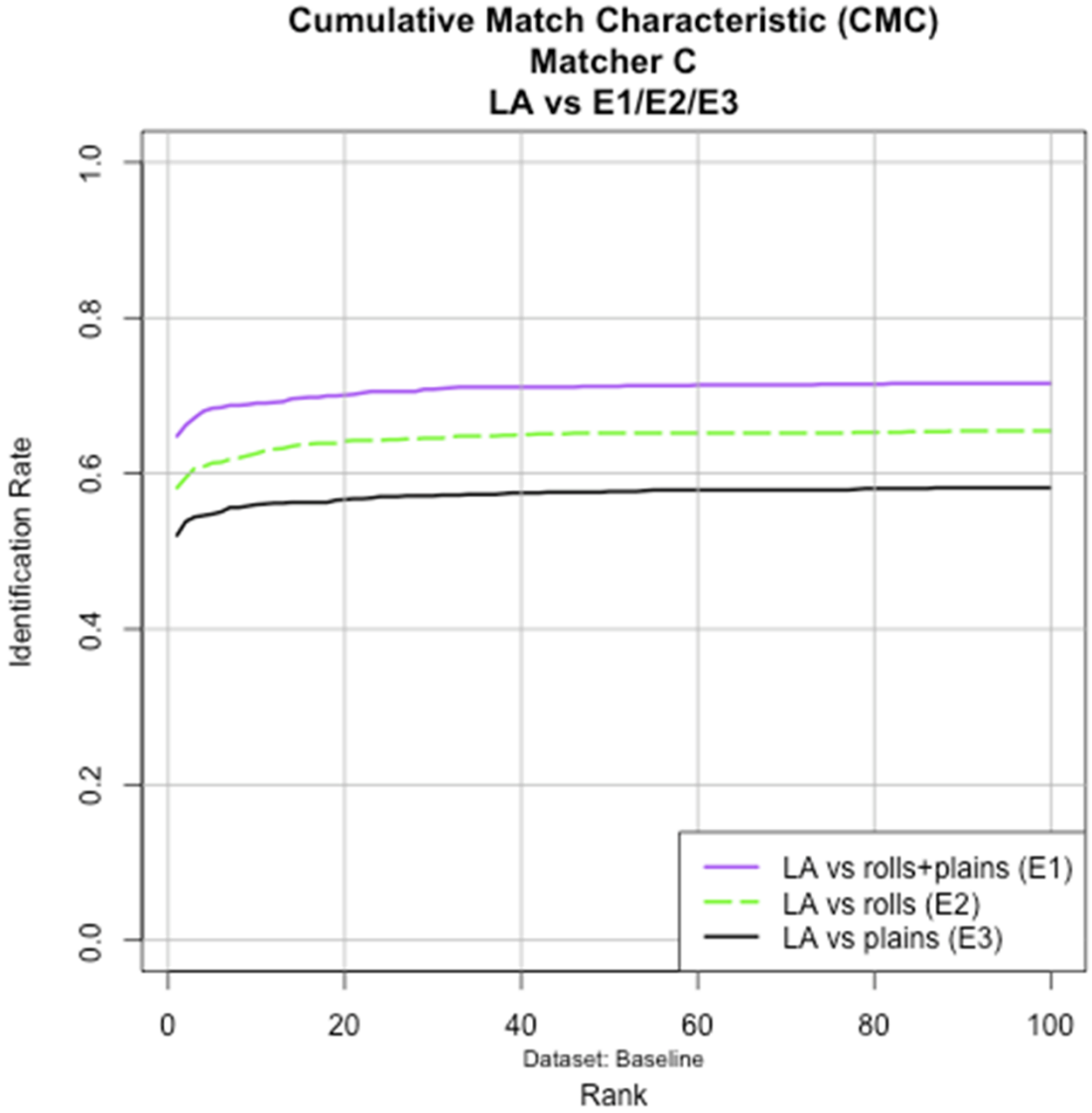

Cumulative Match Characteristic (CMC)
Matcher C
LA vs E1/E2/E3
Identification Rate
0.0
0.2
0.4
0.6
0.8
1.0
0
20
40
60
80
100
Dataset: Baseline
Rank
LA vs rolls+plains (E1)
LA vs rolls (E2)
LA vs plains (E3)

Image + EFS

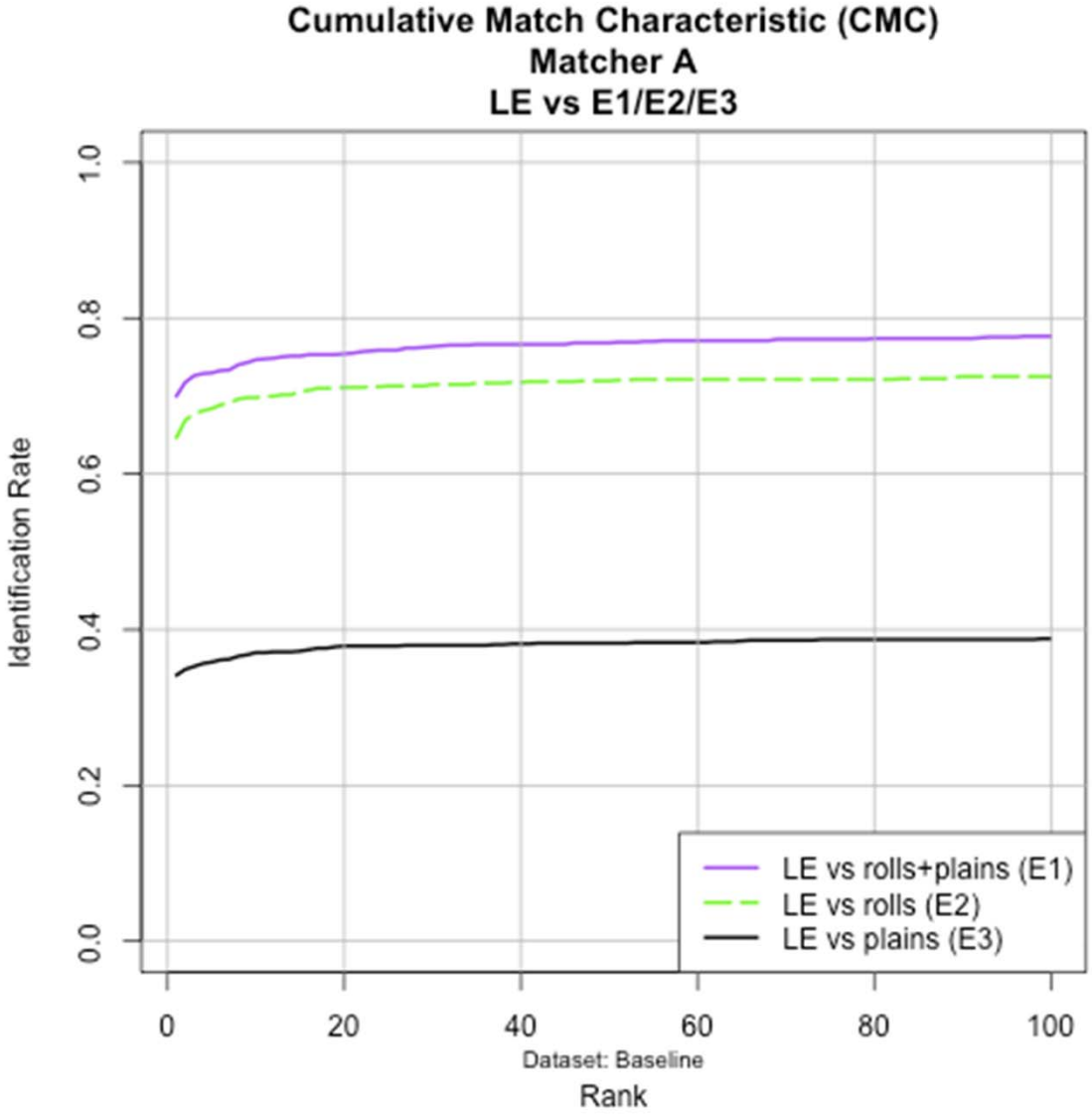

Cumulative Match Characteristic (CMC)
Matcher A
LE vs E1/E2/E3
Identification Rate
0.0
0.2
0.4
0.6
0.8
1.0
0
20
40
60
80
100
Dataset: Baseline
Rank
LE vs rolls+plains (E1)
LE vs rolls (E2)
LE vs plains (E3)

Image + EFS

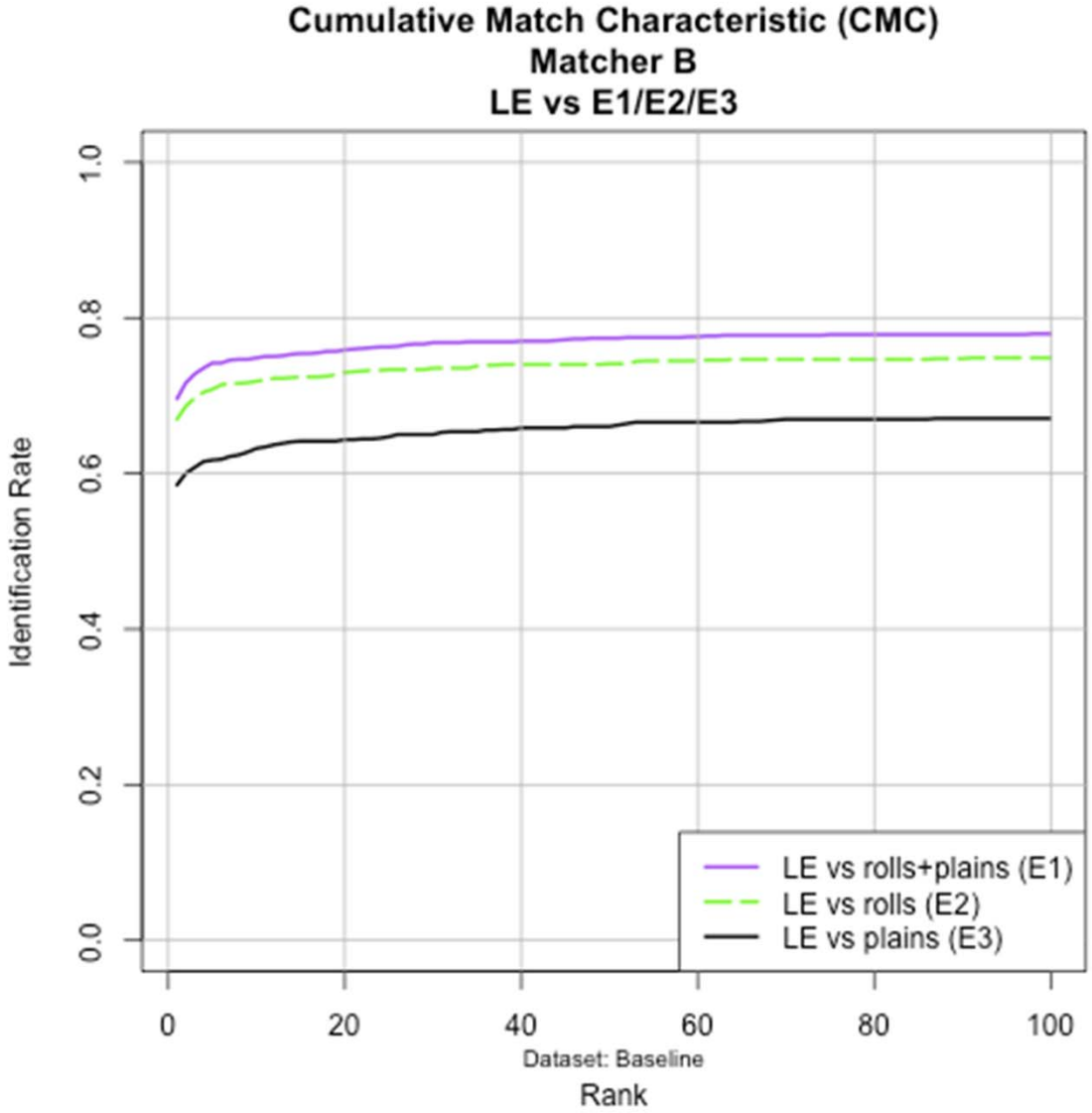
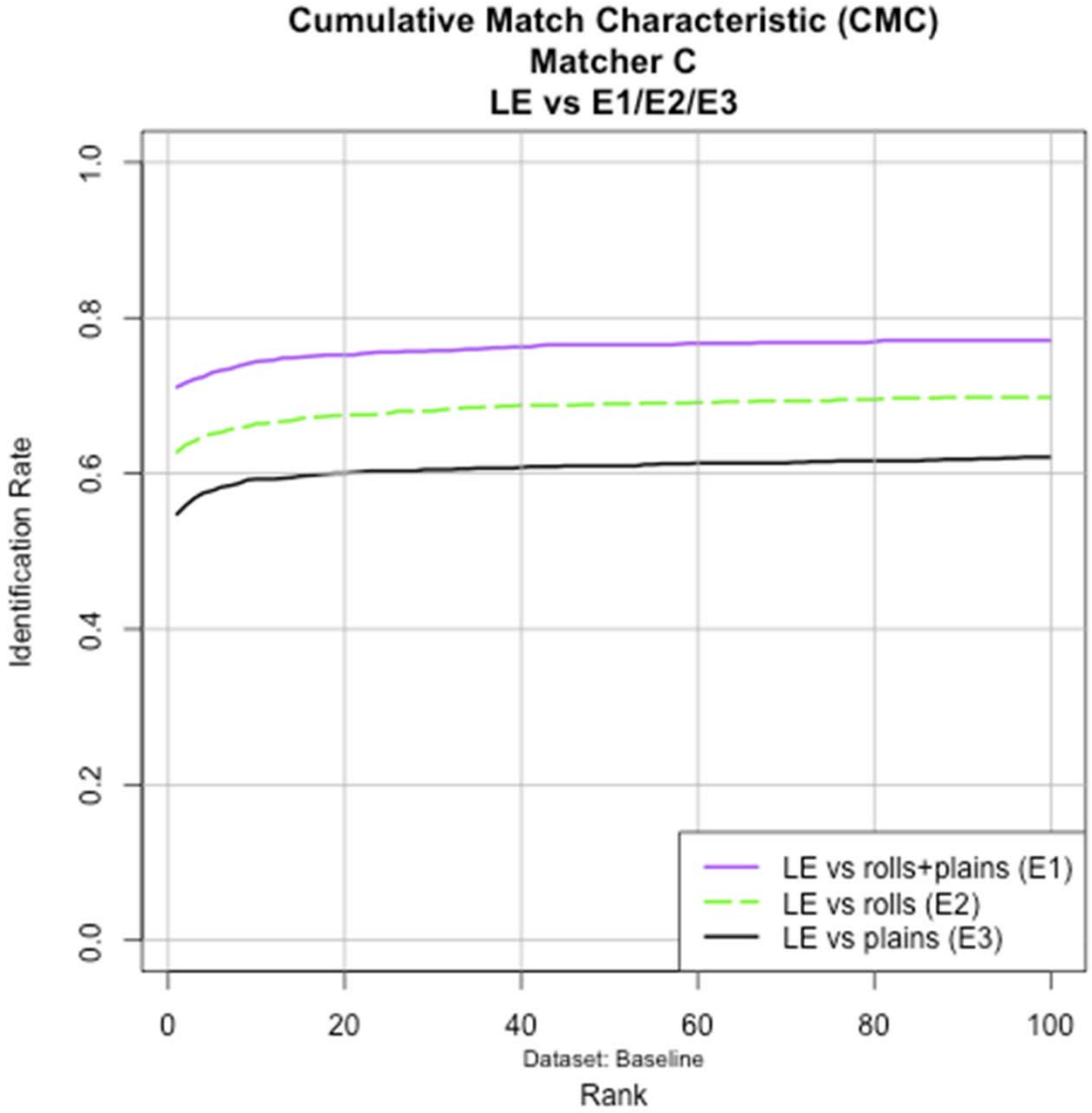
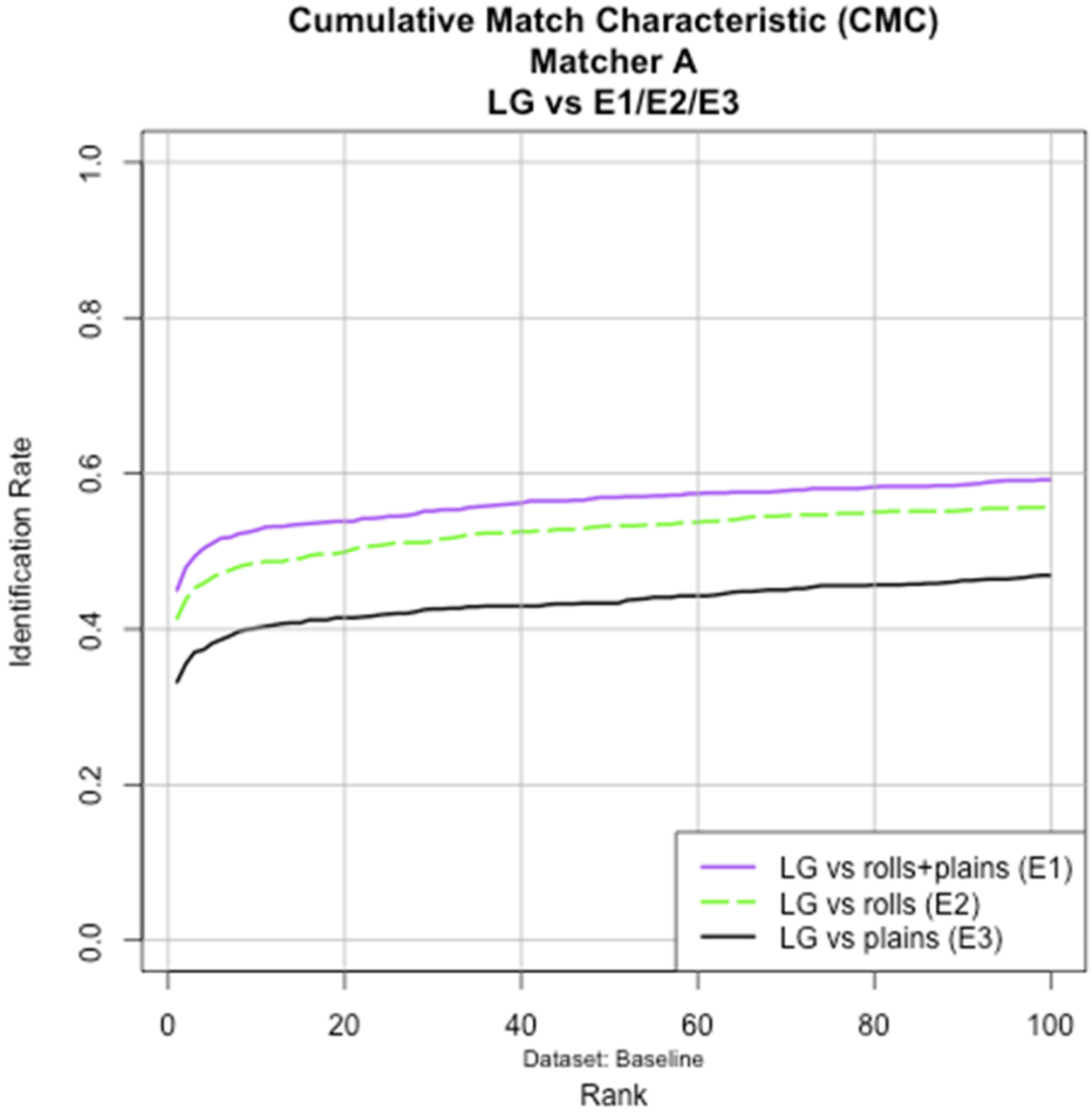

Cumulative Match Characteristic (CMC)
Matcher B
LE vs E1/E2/E3
Identification Rate
0.0
0.2
0.4
0.6
0.8
1.0
0
20
40
60
80
100
Dataset: Baseline
Rank
LE vs rolls+plains (E1)
LE vs rolls (E2)
LE vs plains (E3)

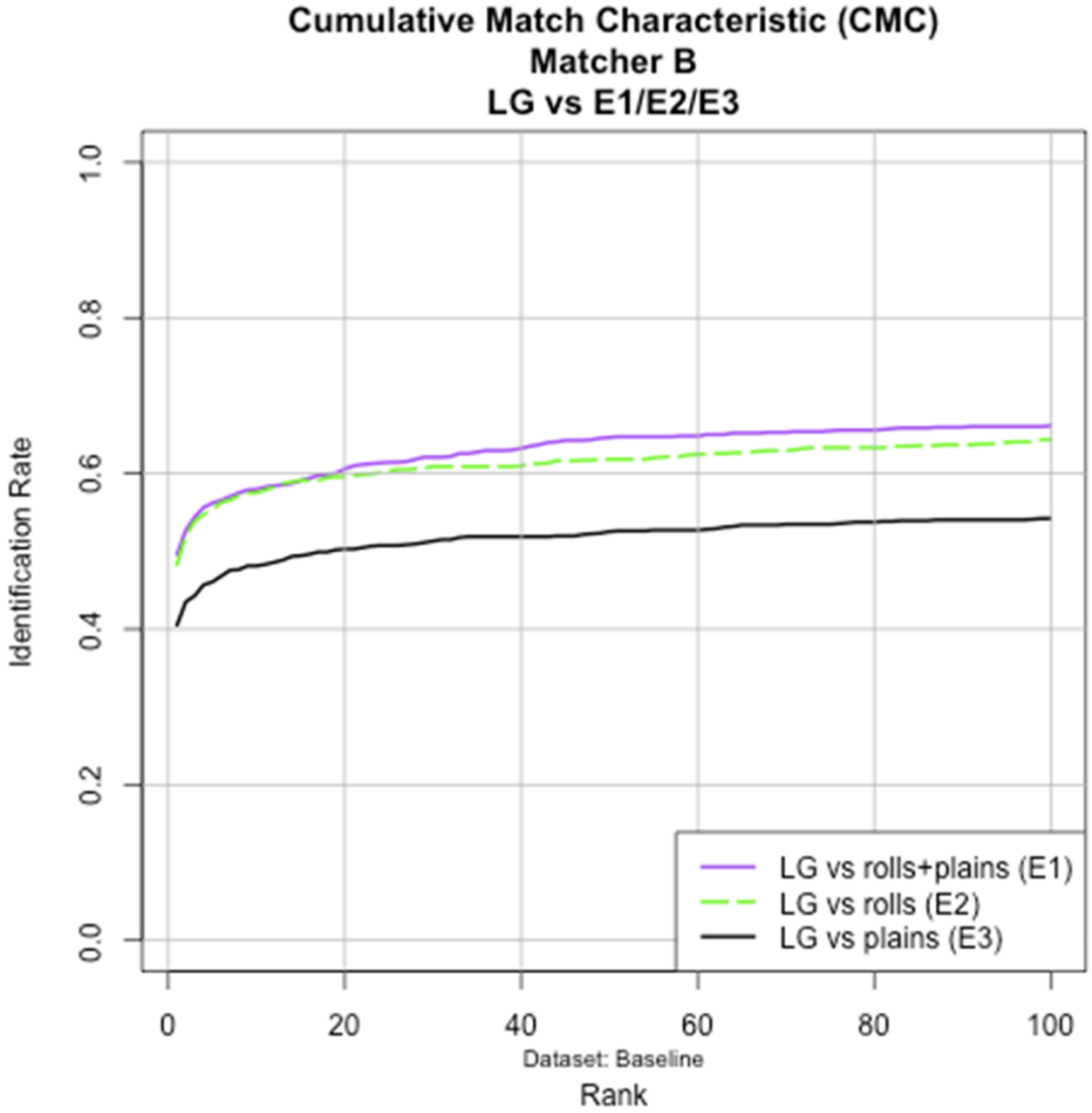
Image + EFS


Cumulative Match Characteristic (CMC)
Matcher C
LE vs E1/E2/E3
Identification Rate
0.0
0.2
0.4
0.6
0.8
1.0
0
20
40
60
80
100
Dataset: Baseline
Rank
LE vs rolls+plains (E1)
LE vs rolls (E2)
LE vs plains (E3)

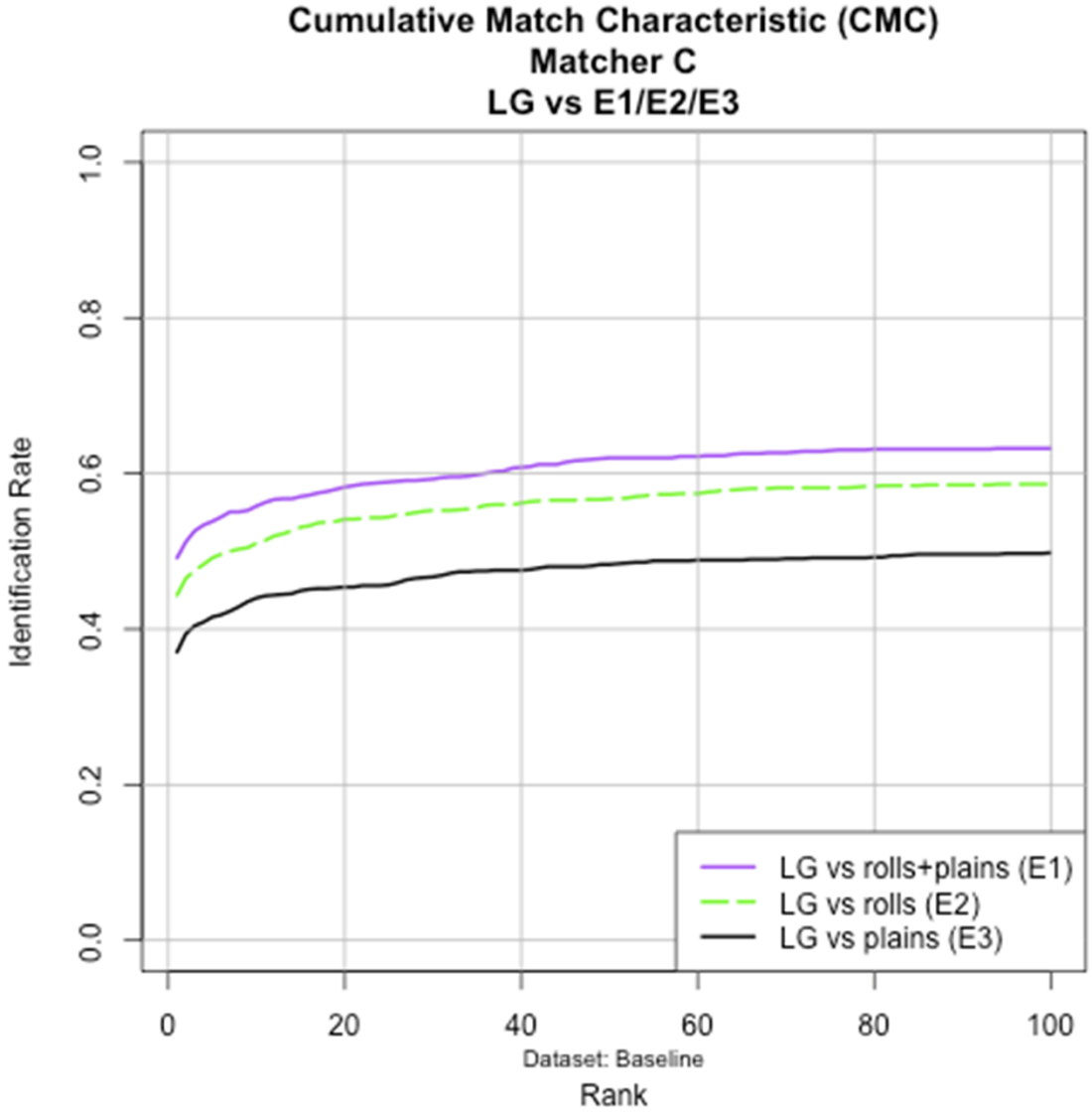
Minutiae Only



Minutiae Only



Minutiae Only



Summary of Rolled vs. Plain Performance

(E1, E2 & E3; Baseline latent dataset; rank 1; SDK C)

	Latent Subset		
	LA	**LE**	**LG**
	Image only	Image + EFS	Minutiae only
Rolls+Plains (E1)	65.0	71.4	49.3
Rolls Only (E2)	58.4	63.0	44.6
Plains Only (E3)	52.3	55.2	37.2

Effect of “Value” Determination

Effect of "Value" Determination

		All	No value	Limited value	Value
Count		1066	25	113	917
LA (Image only)	A	67.2%	**20.0%**	**34.5%**	72.6%
	B	63.0%	**8.0%**	**28.3%**	68.4%
	C	65.0%	**8.0%**	**30.1%**	70.8%
	D	38.9%	**4.0%**	**4.4%**	44.1%
	E	49.2%	0.0%	**10.6%**	55.0%
LE (Image + EFS)	A	70.2%	**20.0%**	**35.4%**	75.9%
	B	69.9%	**12.0%**	**31.0%**	76.2%
	C	71.4%	**20.0%**	**35.4%**	77.1%
	E	52.3%	0.0%	**17.7%**	57.9%
LG (Minutiae only)	A	45.1%	**4.0%**	**6.2%**	51.2%
	B	49.8%	0.0%	**4.4%**	56.8%
	C	49.3%	0.0%	**7.1%**	55.8%
	E	0.0%	0.0%	0.0%	0.0%

Additional Findings

- 17.8% of latents in the test were missed by all matchers (at rank 1) — nearly half of these could be individualized by a certified latent examiner

- 17.6% of latents in the test could not be individualized by a certified latent examiner— nearly half of these could be matched at rank 1 by one or more matchers

- Best matcher was 71% accurate vs. 82% “collective accuracy” – suggests room for improvement

- 90-94% of identifications recorded at rank 1 for the top 3 matchers

ELFT-EFS Conclusions

- EFS features effective as an interoperable feature set
  - Identical markup was used successfully by all vendors
  - EFS features usually result in a net gain
  - Marginal performance boost from non-minutiae features
- Accuracy of Image+features > Image-only > Features-only
- Accuracy of Image-only ~= image+features for higher quality images
- Unexpected ability to identify low feature content latents in certain cases
- Elimination of weak candidates via score thresholding could reduce workload significantly with minimal loss of accuracy
- Examiner assessed latent quality and minutiae count reasonably predictive of performance

Future Work

- Future ELFT evaluations (in planning)
  - Latent palmprint matching
  - Latent fingerprint matching at higher throughput (to be more representative of operations)
  - Latent to latent matching
  - Reverse latent search (unsolved latent file)

- Latent Interoperability Transmission Specification (LITS)
  - proposed standardized latent search transactions based on EFS
- Best practices
  - Search and encoding strategies for latent examiners
  - When to use image only searches, other search profiles
  - What are best strategies for encoding palmprint latents
- Latent quality metrics

For More Information

Web ➔ fingerprint.nist.gov/latent

Email ➔ latent-efs@nist.gov