NISTIR 7775

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

M. Indovina
R. A. Hicklin
G. I. Kiebuzinski

National Institute of
Standards and Technology
U.S. Department of Commerce

(This page intentionally left blank.)

NISTIR 7775

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

M. Indovina
Image Group / Information Access Division
Information Technology Laboratory

R. A. Hicklin
Noblis, Inc.

G. I. Kiebuzinski
Noblis, Inc.

March 2011

U.S. Department of Commerce
Gary Locke, Secretary

National Institute of Standards and Technology
Patrick D. Gallagher, Director

Abstract

The National Institute of Standards and Technology (NIST) Evaluation of Latent Fingerprint Technologies - Extended Feature Sets (ELFT-EFS) consists of multiple ongoing latent algorithm evaluations. This report describes the design, process, results, and conclusions of ELFT-EFS Evaluation #1; an accuracy test of latent fingerprint searches using features marked by experienced human latent fingerprint examiners, in addition to automatic feature extraction and matching (AFEM). There has never previously been an evaluation of latent fingerprint matchers of this scale in which systems from different vendors used a common, standardized feature set. The results show that searches using images plus manually marked Extended Features (EFS) demonstrated effectiveness as an interoperable feature set. The four most accurate matchers demonstrated benefit from manually marked features when provided along with the latent image. The latent image itself was shown to be the single most effective search component for improving accuracy, and was superior to features alone in most cases. For most matchers, the addition of new EFS features provided an improvement in accuracy. In several cases, some of the algorithms provided counterintuitive results that may be indicative of implementation issues; therefore, these results are preliminary, and broad conclusions on the efficacy of these features in improving performance should await the subsequent results from Evaluation #2, in which some known software issues are being corrected by the participants. The accuracy when searching with EFS features is promising considering the results are derived from early-development, first-generation matchers. Further studies using next-generation matchers are warranted (and underway) to determine the performance gains possible with EFS.

Acknowledgements

The authors would like to thank the Department of Homeland Security's Science and Technology Directorate and the Federal Bureau of Investigation's Criminal Justice Information Services Division and Biometric Center of Excellence for sponsoring this work.

Disclaimer

In no case does identification of any commercial product, trade name, or vendor, used in order to perform the evaluations described in this document, imply recommendation or endorsement by the National Institute of Standards and Technology, nor does it imply that the products and equipment identified are necessarily the best available for the purpose.

Technology Providers

This table lists the technology providers who participated in this study. The letter keys listed down the first column are used throughout the report to identify results from specific algorithms. The authors wish to thank the technology providers for their voluntary participation and contribution.

Table ES-1: SDK letter keys and the corresponding technology provider

Key	Technology Provider Name
A	Sagem Securite
B	NEC Corporation
C	Cogent, Inc.
D	Sonda Technologies, Ltd.
E	Warwick Warp, Ltd.

Executive Summary

Introduction

The National Institute of Standards and Technology (NIST) Evaluation of Latent Fingerprint Technologies - Extended Feature Sets (ELFT-EFS) consists of multiple ongoing latent algorithm evaluations. This report describes the design, process, results, and conclusions of ELFT-EFS Evaluation #1; an accuracy test of latent fingerprint searches using features marked by experienced human latent fingerprint examiners, in addition to automatic feature extraction and matching (AFEM).

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 sets of features marked by experienced latent print examiners. The feature sets include different subsets of the Extended Feature Set (EFS) [1] defined by the Committee to Define an Extended Fingerprint Feature Set (CDEFFS) [2], which is being incorporated into ANSI/NIST ITL-1 2011 "Data Format for the Interchange of Fingerprint, Facial, & Other Biometric Information" (forthcoming) and is a superset of the latent fingerprint feature set used by the FBI's Integrated Automated Fingerprint Identification System (IAFIS). EFS was developed to serve as an interoperable interchange format for automated fingerprint or palmprint systems, as well as a means of data interchange among latent print examiners. One of the purposes of ELFT-EFS is to determine the extent to which 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 other features is appropriate. ELFT-EFS is not a test of automatic EFS extraction (i.e. conformance to the standard), but rather a test of data interoperability and how potentially useful such human marked features are when processed by the matcher.

The ELFT-EFS evaluations are open to both the commercial and academic community. In Evaluation #1, participants included five commercial vendors of Automated Fingerprint Identification Systems (AFIS). The five participants each submitted three Software Development Kits (SDKs) which respectively contained (i) a latent fingerprint feature extraction algorithm; (ii) algorithms for ten-print feature extraction and gallery creation, and (iii) a 1-to-many match algorithm that returns a candidate list report. The fingerprint features automatically extracted by (i) and (ii) were proprietary, at the discretion of the technology provider, and could include the original input image(s). Evaluations were run at NIST on commodity NIST hardware.

Many of the results from Evaluation #1 were made public in a Preliminary Report on the CDEFFS website (2) in January 2010. The ongoing Evaluation #2 is providing an opportunity for the testing of algorithms that were revised in light of the preliminary results.

Fingerprint and Feature Data

The test dataset contained 1,114 latent fingerprint images from 837 subjects. The gallery was comprised of (mated) exemplar sets from all 837 latent subjects, as well as (non-mated) exemplar sets from 99,163 other subjects chosen at random from an FBI provided and de-identified dataset. Each subject in the gallery had two associated exemplar sets: one set of ten rolled impression fingerprint images, and one set of ten plain impression fingerprint images (plains).

In addition to fingerprint images, each latent had an associated set of hand-marked features. The features were marked by twenty-one International Association for Identification Certified Latent Print Examiners (IAI CLPE) using guidelines developed specifically for this process [3]. 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. The various subsets of latent features are summarized in Table ES-2. The additional extended features in subsets LE and LF included ridge quality maps, creases, dots, incipient ridges, ridge edge protrusions, and pores — in addition to the minutiae, ridge counts, cores & deltas, and pattern class included in

MATCHER KEY	**A** = *Sagem*	**B**=*NEC*	**C**=*Cogent*	**D**=*Sonda*	**E**=*Warwick*

FEATURE SUBSET KEY	**LA**=*Image*	**LB**=*Image +ROI*	**LC**=*Image +ROI +Quality map +Pattern class*	**LD**=*Image +ROI +Minutiae +Ridge counts*	**LE**=*Image +Full EFS (no Skeleton)*	**LF**=*Image +Full EFS with Skeleton*	**LG**=*Minutiae +Ridge counts (comp to IAFIS)*	**Page 7**

other subsets. Latent examiners made determinations of Value, Limited Value (latents of value for exclusion only), or No Value at the time of markup, in addition to informal quality assessments of "Excellent", "Good", "Bad", "Ugly", and "No Value". A subset of the latents had skeletons marked (including associated ridge flow maps). Features were marked in latent images without reference to exemplars, with the exception of a subset of 458 latent images that included an additional Ground Truth (GT) markup based on the latent and all available exemplars; GT markup provides a measure of ideal (but operationally infeasible) performance when compared to the original examiner markup.

Table ES-2: Latent feature subsets

Subset	Image	ROI	Pattern class	Quality map	Minutiae + ridge counts	Additional extended features	Skeleton
LA	■						
LB	■	■					
LC	■	■	■	■			
LD	■	■	■		■		
LE	■	■	■	■	■	■	
LF	■	■	■	■	■	■	■
LG		■	■		■		

Primary Findings

There has never previously been an evaluation of latent fingerprint matchers of this scale in which systems from different vendors used a common, standardized feature set. The results show that searches using images plus manually marked Extended Features (EFS) demonstrated effectiveness as an interoperable feature set. The four most accurate matchers demonstrated benefit from manually marked features when provided along with the latent image. The latent image itself was shown to be the single most effective search component for improving accuracy, and was superior to features alone in most cases. For most matchers, the addition of new EFS features provided an improvement in accuracy. In several cases, some of the algorithms provided counterintuitive results that may be indicative of implementation issues; therefore, these results are preliminary, and broad conclusions on the efficacy of these features should await the subsequent results from Evaluation #2, in which some known software issues are being corrected by the participants. The accuracy when searching with EFS features is promising considering the results are derived from early-development, first-generation implementation of the new standard. Further studies using next-generation matchers are warranted (and underway) to determine the performance gains possible with EFS. After the Evaluation #2 results, further data-specific analyses will be conducted to determine the cases in which specific EFS features provide improvements in accuracy.

Results and Conclusions

1. The highest accuracy for all participants was observed for searches that included examiner-marked features in addition to the latent images.
2. Image-only searches were more accurate than feature-only searches for most matchers.
3. Since score-based results are more scalable than rank-based results, they provide a better indication of how accuracy would be affected by an increase in database size. This capability could provide operational benefits such as reduced or variable size candidate lists, or for reverse latent searches (searches of databases containing unsolved latents) where using a score threshold is used to limit candidate list size. Comparing rank-1 and score-based results at different values of False Positive Identification Rate (FPIR) showed that image-only and image + full EFS[a] searches are less sensitive to differences in FPIR, whereas minutiae-only searches are most sensitive.

[a] *"full EFS" refers to all EFS features used in this study (ridge quality maps, creases, dots, incipient ridges, ridge edge protrusions, and pores) in addition to the EFS subset used by IAFIS (minutiae, ridge counts, cores & deltas, and pattern class).*

MATCHER KEY	**A** = *Sagem*	**B**=*NEC*	**C**=*Cogent*	**D**=*Sonda*	**E**=*Warwick*

FEATURE SUBSET KEY	**LA**=*Image*	**LB**=*Image +ROI*	**LC**=*Image +ROI +Quality map +Pattern class*	**LD**=*Image +ROI +Minutiae +Ridge counts*	**LE**=*Image +Full EFS (no Skeleton)*	**LF**=*Image +Full EFS with Skeleton*	**LG**=*Minutiae +Ridge counts (comp to IAFIS)*	**Page 9**

4. The effect of the use of EFS features other than minutiae was mixed, as shown in Table ES-3. Bolded results indicate best performance for each matcher. In most cases, additional features resulted in accuracy improvement, highlighted in green; cases where accuracy declined may be indicative of implementation issues, highlighted in yellow. When using score-based results, the highest overall accuracy is achieved through the use of full EFS with skeletons.

Table ES-3: Rank-1 identification rates for latent feature subsets

(Baseline-QA dataset, 458 latents — 100,000 rolled+plain 10-finger exemplar sets)

		Latent Feature Subset						
		LA	**LB**	**LC**	**LD**	**LE**	**LF**	**LG**
Matcher	**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[b]	**22.3**	*n/a*	*n/a*	14.0	*n/a*	14.4	10.9
	E	42.6	44.3	46.5	44.5	**47.2**	31.2	24.2

5. Ground truth (GT) markup yielded an increase in performance over the original examiner markup of about 5 to 8 percentage points for image + full EFS searches, and about 11 to 15 percentage points for minutiae-only searches. This shows that matcher accuracy is highly affected by the precision of latent examiner markup.
6. Latent orientation (angle) has an impact on matcher accuracy. When the orientation of latents was unknown, the rank-1 identification rates were 17.8 to 18 percentage points lower than the overall average.
7. Matcher accuracy is very clearly related to the examiners' latent print value determinations, with much greater accuracy for latents determined a priori to be of value. The matching algorithms demonstrated an unexpected ability to identify low feature content latents: Sagem's rank-1 accuracy for No Value latents was 20% on image-only searches, and 26.2% on Limited Value latents.
8. The performance of all matchers decreased consistently as lower quality latents were searched, with respect to the informal scale of "Excellent", "Good", "Bad", or "Ugly".
9. Analysis showed that the greatest percentage of the misses were for latents with low minutiae count, and those assessed by examiners as poor quality ("Ugly") or "No Value". Algorithm accuracy for all participants was highly correlated to the number of minutiae. For latents with more than 10 minutiae, minutiae count was the most important factor for successful identification with examiner-assessed quality being secondary. For latents with fewer than 10 minutiae, examiner-assessed quality was a better predictor of match accuracy than minutiae count.
10. Approximately 22% of the latents in the test were missed by all matchers at rank 1, more than half of which could be individualized by a certified latent examiner. The initial or reviewing examiners determined that 14% of the latents in the test were of No Value, of value for exclusion only, or resulted in an inconclusive determination; about one third of these could be matched by one or more matchers at rank 1.
11. The highest measured accuracy achieved by any matcher at rank-1 on any latent feature subset was 66.7%, even though approximately 78% of the latents in the test were matched by one or more matchers at rank-1. This indicates a potential for additional accuracy improvement through improved algorithms, or through the use of data based fusion (e.g. search using image-only and again search using image+features). The differences in which latents were identified by the various matchers also points to a potential accuracy improvement by using algorithm fusion.
12. The use of both rolled and plain impressions in the gallery resulted in higher accuracy than the use of either rolled or plain impressions separately for most matchers. Use of plain impressions in the gallery as compared to rolled impressions resulted in a drop in accuracy for most matchers.

[b] *Participant D's performance was erratic and therefore is not generally included in the summary conclusions.*

MATCHER KEY	**A** = *Sagem*	**B**=*NEC*	**C**=*Cogent*	**D**=*Sonda*	**E**=*Warwick*

FEATURE SUBSET KEY	**LA**=*Image*	**LB**=*Image +ROI*	**LC**=*Image +ROI +Quality map +Pattern class*	**LD**=*Image +ROI +Minutiae +Ridge counts*	**LE**=*Image +Full EFS (no Skeleton)*	**LF**=*Image +Full EFS with Skeleton*	**LG**=*Minutiae +Ridge counts (comp to IAFIS)*	**Page 11**

13. The results obtained during ELFT Phase II (NISTIR 7577 [8], April 2009) show substantially higher accuracy than ELFT-EFS results for image-only matching for three of the participants who participated in both tests. This is an expected result because the ELFT-EFS test data included a greater proportion of poor-quality latents, had higher throughput requirements, and larger gallery size.

Caveats

In evaluating results, note that this was an evaluation of data compliant with an emerging draft 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, and did not permit time for extensive research and development and software debugging. Therefore, the evaluation may not be indicative of potential future improvements. The ongoing ELFT-EFS Evaluation #2 is being conducted to provide an opportunity for participants to incorporate lessons learned from the preliminary results of this evaluation into further improvement of the algorithms.

The performance impact of any specific feature as measured in this test may be limited because the participants have not yet developed approaches to take advantage of such features; the participants already take advantage of the underlying information context through automated processing of the latent image (AFEM), and there is limited or no additional gain from explicitly marking the features; there was limited opportunity for improvement due to the limited presence of such features in the data; or explicit human markup of such features is ineffective for automated matching.

The results may not be applicable to other datasets and operational systems with different processing constraints: specifically, the relative performance of image-based and feature-based matching may be affected by differences in systems resources. The cost in computer resources may be a significant factor in determining the practicality of image-only or image+feature searches. ELFT-EFS did not measure or evaluate resource requirements for the different types of matching. It should be noted that as the cost of the hardware continues to fall in accordance with Moore's Law, greater use of image searches is likely to be more acceptable.

Table of Contents

Terms and Definitions

This table provides ELFT-specific definitions to various words and acronyms found in this report.

Table 1: Terminology and acronyms

Term	Definition
AFEM	Automated Feature Extraction and Matching
AFIS	Automated Fingerprint Identification System: generic term for large-scale fingerprint matching system
ANSI/NIST	The ANSI/NIST-ITL biometric data standard, used here as the file format containing fingerprint images and features. The current version is ANSI/NIST-ITL 1-2007. ANSI/NIST-ITL 1-2011 is under development and will include the EFS definitions used here. [4]
API	Application Programming Interface
CDEFFS	Committee to Define an Extended Fingerprint Feature Set [2]
CMC	Cumulative Match Characteristic
DET	Detection Error Tradeoff characteristic
EBTS	Electronic Biometric Transmission Specification is the standard currently used by FBI for IAFIS, and which will be used for NGI. EBTS is an implementation of the ANSI/NIST-ITL standard. EBTS version 8.0 is currently implemented; version 9.2 is in review and will contain the EFS fields specified in ANSI/NIST-ITL 1-2011. Note there was a name change from Electronic Fingerprint Transmission Specification (EFTS) to EBTS in October 2007. [5]
EFS	Extended Feature Set (proposed extension to ANSI/NIST standards by CDEFFS)
Exemplar	Fingerprint image deliberately acquired during an enrollment process; the known mate of a latent fingerprint
FNIR	False Negative Identification Rate (also called miss rate or false non-match rate)
FPIR	False Positive Identification Rate (also called false-match rate)
Fusion	A method of combining biometric information to increase accuracy
Gallery	A set of enrolled ten-prints; synonymous with "database." An ELFT Gallery is composed of foreground and background ten-prints.
Ground truth	Definitive association of a print and exemplar at the source attribution ***or*** at the feature level: • ground-truth source attribution is the definitive association of a fingerprint with a specific finger from a subject • feature-level ground truth is the feature-by-feature validation that all marked features (e.g. minutiae) definitively correspond to features present in the exemplar(s).
Hit/hit-rate	A "hit" results when the correct mate is placed on the candidate list; the "hit rate" is the fraction of times a hit occurs, assuming a mate is in the gallery.
IAFIS	The FBI's Integrated Automated Fingerprint Identification System, operational since 1999; for latent prints, IAFIS is scheduled to be replaced by NGI in 2013.
Latent	A fingerprint image inadvertently left on a surface touched by an individual
Matcher	Software functionality which produces one or more plausible candidates matching a search print
Mate	An exemplar fingerprint corresponding to a latent
NGI	The FBI's Next Generation Identification system, which will replace the current IAFIS
NIST	National Institute of Standards and Technology
ROC	Receiver Operator Characteristic

Term	Definition
ROI	Region of Interest
Rolled print	A fingerprint image acquired by rolling a finger from side to side on the capture surface
Plain print	A fingerprint image acquired by pressing a finger flat on the capture surface. The plain images used in this test were segmented from slap fingerprints.
Slap print	An image containing multiple plain fingerprints collected simultaneously

1 Introduction

The National Institute of Standards & Technology (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 part of NIST's larger ELFT testing program. The ELFT evaluations to date, notably ELFT Phase II [8], have focused solely on automated feature extraction and matching (AFEM, or "image-only" matching). The ELFT-EFS evaluations address the accuracy of latent matching using features marked by experienced human latent fingerprint examiners, which is the standard procedure for most operational uses of AFIS.

The purpose of this test is to evaluate the current state of the art in latent feature-based matching and to assess the accuracy of searches using images alone with searches using different feature sets. A key objective of the evaluations 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 feature sets include different subsets of the Extended Feature Set (EFS) features [1] defined by CDEFFS [2]. One of these feature subsets corresponds directly to the current latent fingerprint feature search used by IAFIS. EFS was developed to serve as an interoperable interchange format for automated fingerprint or palmprint systems, as well as a means for annotation and interchange among latent print examiners. It was designed in collaboration with the vendor community and other subject matter experts to be a superset of IAFIS and vendor-specific feature sets incorporated in the Universal Latent Workstation (ULW) [6]. The use of a standardized interchange format is intended to eliminate the vendor specific training of latent examiners currently required to obtain optimal search results for each type of AFIS. EFS is intended to be a uniform set of markup rules that can be used by AFIS vendors to achieve high levels of performance without requiring that the markup be vendor specific.

ELFT-EFS is not a test of automatic EFS extraction (i.e. conformance), but rather a test of data interoperability and how potentially useful such human features are when processed by a matcher. When images 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 the manually specified EFS features): which features are automatically extracted from latent or exemplar images are at the sole discretion of the participant and are not examined in this study.

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 was intended for those purposes. The participants in the ELFT-EFS Public Challenge 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. Participation in the public challenge was a prerequisite for participation in Evaluation #1. Many of the results from Evaluation #1 were made publicly available in a Preliminary Report on the CDEFFS website in January 2010.

ELFT-EFS Evaluation #2

ELFT-EFS Evaluation #2 is currently in progress. The test data, API, and test protocols in Evaluation #2 will be similar but not identical to those in Evaluation #1. ELFT-EFS Evaluation #2 is being conducted with the expectation that participants from Evaluation #1 may incorporate lessons learned and/or miss analysis conducted at NIST, with the addition of new participants.

Subsequent Evaluations

Subsequent ELFT-EFS Evaluations are expected to be conducted to evaluate different aspects of latent matching, respond to lessons learned, and track ongoing progress.

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

2 Participants

The ELFT-EFS evaluations are open to both the commercial and academic community. In Evaluation #1, participants included five commercial vendors of Automated Fingerprint Identification Systems (AFIS): Sagem Securite, NEC, Cogent, Sonda, and Warwick (see Table ES-1). Participation in Evaluation #1 was predicated on participation in the ELFT-EFS Public Challenge.

3 Evaluation Data

The test dataset contained 1,114 latent fingerprint images from 837 subjects. The gallery was comprised of (mated) exemplar sets from all 837 latent subjects, as well as (non-mated) exemplar sets from 99,163 other subjects chosen at random from an FBI provided and de-identified dataset. Each subject in the gallery had two associated exemplar sets: one set of ten rolled impression fingerprint images, and one set of ten plain impression fingerprint images (plains, segmented from slap images).

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 2. Each of the initial data sources included a larger number of latent images. From these sources, examiners provided assessments of the quality of the latents using an informal "Excellent", "Good", "Bad", "Ugly" and "No Value" scale. The latents selected for subsequent markup included approximately equal proportions of the Good, Bad, and Ugly categories, with less than 2% of each of the Excellent and No Value categories. (See Section 3.1.3) In none of the cases were the mates selected through the use of automated fingerprint matchers (possibly producing AFIS bias), as was true in the ELFT Phase 2 evaluation. Each latent image was from a distinct finger.

Table 2: Sources of latent images *(Baseline dataset)*

Name	# Latents	Description	Minutiae count	
			Mean	St dev
Casework 1	372	Operational casework images	20	12
Casework 2	165	Operational casework images	18	9
WVU	446	Laboratory collected images	27	20
FLDS	93	Laboratory collected images	20	17
MLDS	38	Laboratory collected images (small set of publicly releasable images for examples in reports)	18	15
Total (Baseline)	1114		22	16

A combination of prints from both operational casework and laboratory collected images was included to provide a diversity of data types, including a broad range of quality, deposition, and substrate/background attributes. All of the prints in the casework datasets were considered of value by the original examiners who determined the individualizations.

For an evaluation such as this, source attribution, or "ground truth"[c] association of the latent with a specific finger of a known subject, is critical. The source attribution of the datasets was accomplished differently for the casework and laboratory-collected sources. The laboratory-collected images were acquired under carefully controlled conditions so that the source attribution could not be in doubt. In the casework datasets, the prints used were selected from cases in which multiple additional corroborating latents and exemplars were used in the individualization, so that the latent-exemplar relation was not made solely through the use of these latents.

Late in the analysis process, during miss analysis, it was determined that due to administrative errors, 10 of the 1114 latents did not have mates in the gallery. The results are not corrected for this, with the result that measures of accuracy may be underestimated by as much as 0.9%.

The technology providers had no knowledge of, or access to, the test datasets prior to or during the tests, other than a small set of public fingerprints provided to serve as examples – the Multi-Latent Dataset (MLDS) test set consisting of 38 latents.

3.1.2 Latent features

In addition to fingerprint images, each latent had an associated set of hand-marked features. The features were marked by twenty-one International Association for Identification Certified Latent Print Examiners (IAI CLPE). Latent features were included in ANSI/NIST files 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), which is being incorporated into ANSI/NIST ITL-1 2011 "Data Format for the Interchange of Fingerprint, Facial, & Other Biometric Information" (forthcoming, [4]) and is a superset of the latent fingerprint feature set used by the FBI's Integrated Automated Fingerprint Identification System (IAFIS). In marking the latent features, examiners followed the specific guidelines defined in [3]; no vendor specific rules for feature encoding were used. The test evaluated different combinations of EFS fields. The different subsets of EFS features included in the latent files (Subsets LA-LG) are defined in Table 3. The specific EFS fields included in each subset are listed in Appendix A (ELFT-EFS Test Plan), Section 7. The additional EFS features in subsets LE and LF included ridge quality maps, creases, dots, incipient ridges, ridge edge protrusions, and pores — in addition to minutiae, ridge counts, cores & deltas, and pattern class. Latent examiners made determinations of Value, Limited Value (latents of value for exclusion only), or No Value at the time of markup, in addition to informal quality assessments of "Excellent", "Good", "Bad", "Ugly", and "No Value".

[c] *Note that the term "ground truth" is commonly used in two contexts: source attribution (the definitive association of a fingerprint with a specific finger from a subject), or feature-level ground truth (the feature-by-feature validation that all marked features (e.g. minutiae) definitively correspond to features present in the exemplar(s)).*

MATCHER KEY	**A** = *Sagem*	**B**=*NEC*	**C**=*Cogent*	**D**=*Sonda*	**E**=*Warwick*

FEATURE SUBSET KEY	**LA**=*Image*	**LB**=*Image +ROI*	**LC**=*Image +ROI +Quality map +Pattern class*	**LD**=*Image +ROI +Minutiae +Ridge counts*	**LE**=*Image +Full EFS (no Skeleton)*	**LF**=*Image +Full EFS with Skeleton*	**LG**=*Minutiae +Ridge counts (comp to IAFIS)*	**Page 18**

Table 3: Latent feature subsets

Subset	Image	ROI	Pattern class	Quality map	Minutiae + ridge counts	Additional extended features	Skeleton
LA	■						
LB	■	■					
LC	■	■	■	■			
LD	■	■	■		■		
LE	■	■	■	■	■	■	
LF	■	■	■	■	■	■	■
LG		■	■		■		

Features were marked in latent images without reference to exemplars, with the sole exception of a subset of the latent images that included an additional Ground Truth (GT) markup based on the latent and all available exemplars discussed in section 9. GT markup provides a measure of ideal (but operationally infeasible) performance when compared to the original examiner markup. GT data eliminates the variability in feature identification introduced by the latent examiner. GT results represent the upper performance limit of what the matching algorithm can achieve.

Note that conformance testing of automatic extraction of EFS 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. The extent of use of the EFS features for the test was solely decided by the participants. However, the EFS feature set was presented to all vendors to use as they saw fit. 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 image based information to match against a corresponding area that would otherwise have many "false" minutiae and poor ridge flow.

All latent markup was not available for all of the latents: skeletons were only available for a subset due to the extensive time required for manual markup of skeletons. The full 1,114 dataset is named here as the "Baseline" dataset. The 458 latent subset of Baseline that includes skeletons is named here as the "Baseline-QA" dataset. Because skeletons were only available for Baseline-QA, the Baseline-QA dataset was used to compare all of the latent feature subsets (LA-LG). Baseline results were only used to compare LA, LE, and LG.

In addition to the standard markup approach described throughout this document, the 458 latents in Baseline-QA were marked up using two additional markup methods to assess the effects of a more conservative markup approach ("AFIS") and ideal (but operationally infeasible) Ground Truth markup ("GT") – see Table 16.

A quality assurance review was performed by additional latent examiners to verify that the markup for each latent was acceptable. When the reviewing examiner did not believe that the markup was accurate (due to factors such as missing minutiae or incorrect pattern class), the marked up latent was returned to the original examiner with instructions to review and correct feature markup, without specifically referring to individual features of concern; when the original examiner was not available, latents with known markup issues were removed from the dataset.[d]

Examiners were instructed to mark all features present in each latent (with the exception of skeletons, which were only marked in a specified subset of the latents). Each latent had a non-zero number of minutiae, and ridge quality maps were marked in all latents. The prevalence of the other features is summarized in Table 4, which shows the

[d] *To be explicit: the only data removed from the dataset was done on the basis of inaccurate markup by examiners, not due to any attributes of the images.*

proportion of the latents in the Baseline dataset with any of the specified features marked; distributions are detailed in Appendix C.

Table 4: Presence of additional extended features

Feature type	% of Baseline with features present	Feature count (when present)	
		Mean	St dev
Dots	21%	3.1	5.1
Ridge edge features	4%	3.7	7.1
Distinctive features (e.g. scars)	3%	1.1	0.4
Incipient ridges	19%	5.7	10.0
Pores	70%	137.3	177.2
Local quality issues	61%	1.8	1.3
Creases	31%	4.8	5.8

3.1.3 Value and quality assessments of latents

When latent images were selected for markup, the latents were assessed by an examiner using an informal quality scale of "Excellent", "Good", "Bad", "Ugly", and "No Value". Subsequently, the (different) examiners who marked the data made value determinations at the time of markup, using the categories defined in EFS [Field 9.353, Examiner value assessment]:

- 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.
- Limited: The impression is of limited, marginal, value and may be appropriate for exclusion only.
- No value: The impression is of no value, is not appropriate for further analysis, and has no use for potential comparison.

Table 5 shows the counts and proportion of each of the value and quality determinations in the Baseline dataset. Note that 2.2% of the latents in Baseline were marked as No Value, and an additional 11% were marked as Limited value. The prints of Limited or No Value were included so that the capabilities of matchers could be tested on very poor-quality prints; all prints included at least one marked minutia. Since different examiners made the quality and value assessments, there is variation, most notably in the two No Value categories, and in the fact that some latents were assessed as both Good and Limited or No Value.

Table 5: Comparison of quality and value determinations ***(Baseline dataset)***

		Value assessment					
		Of Value	Limited Value	No Value	*n/a*	Total	%
Quality assessment	**Excellent**	12				12	1.1%
	Good	328	4	1	*6*	339	30.4%
	Bad	347	18	2	*3*	370	33.2%
	Ugly	243	94	6	*2*	345	31.0%
	No Value	1	4	14		19	1.7%
	n/a	*25*	*2*	*2*		*29*	*2.6%*
	Total	956	122	25	*11*	1114	
	%	85.8%	11.0%	2.2%	*1.0%*		

Table 6 shows the relationship between quality/value determination and minutiae count. The minutiae quality was determined by the EFS ridge quality map field, which differentiates between areas with definitive and debatable

minutiae. As expected, both quality and value determination are highly correlated with minutiae count, as well as with the count and proportion of definitive minutiae. The relationships between minutiae count, value determination, and quality with respect to accuracy are reported in Sections 12-14.

Table 6: Minutiae count statistics by quality and value determination ***(Baseline dataset)***

	Definitive minutiae (mean)	Debatable minutiae (mean)	Definitive minutiae as % of total minutiae
All (Baseline)	13.6	8.9	60%
Excellent	51.6	11.8	81%
Good	24.4	9.3	72%
Bad	11.1	9.7	53%
Ugly	5.2	7.7	40%
No Value	0.0	3.9	0%
Of Value	15.4	9.5	62%
Limited Value	2.2	5.4	29%
No Value	0.3	3.3	8%

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. The relationship between latent orientation and accuracy is reported in Section 11.

Table 7: Orientation of latents ***(Baseline dataset)***

Orientation (degrees from upright)	% 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 Galleries

The galleries against which the latent datasets were searched included the combinations of rolled and plain impressions specified in Table 8. Unless otherwise noted, all results in this report are against the rolled + plain (E1) subset of exemplars. Exemplars came from optical livescan and inked paper sources. An exemplar record for one subject always included all ten fingers. The relationship between exemplar impression type and accuracy is reported in Section 8.

Table 8: Gallery subsets

Exemplar subset	# subjects	Description
rolled + plain (E1)	100,000	10 rolled & 10 plain impressions each
rolled (E2)	10,000	10 rolled impressions each
plain (E3)	10,000	10 plain impressions each

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. Exemplar images were retained in the same orientation as they were captured, including the segmented slap images; this was conveyed to the participants in the Test Plan.

Of the non-mated exemplars, approximately 54% were from live-scan sources, and 46% were from ink. Of the mated exemplars, approximately 48% were from live-scan sources, and 52% were from ink.

No feature markup data was provided for the exemplar images.

The number of subjects in the gallery was selected to be as large as possible given finite testing resources and throughput requirements (see Section 4.4).

3.3 Data Format

All images and EFS feature markup data were contained in ANSI/NIST files as described in Appendix A.

All images were 8-bit grayscale. All latent images were 1000 pixels per inch (39.37 pixels per millimeter (ppmm)), uncompressed. Exemplar images were 500 pixels per inch (19.69 pixels per millimeter (ppmm)), compressed using Wavelet Scalar Quantization (WSQ) [7].

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

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

The five participants each submitted feature extraction and matching software contained in Software Development Kits (SDKs). Technology providers were encouraged to submit research algorithms in this study: there was no requirement for the SDKs to be in operational use or commercially available. NIST performed a pre-test of the SDKs to ensure all functional capabilities were working. Evaluations were run at NIST on NIST hardware: after validation of the SDKs, the technology providers were no longer involved in the testing. NIST performed the same tests on all SDKs.

4.1 Feature extraction and matching software

Each participant submitted a set of SDKs (Software Development Kits) that provided the interfaces defined by the ELFT-EFS-1 API (Application Programmer Interface) specified in Appendix A. The ELFT-EFS API was modeled after the API from ELFT Phase 2.

Each participant submitted

i. one SDK for exemplar feature extraction and exemplar enrollment in the gallery
ii. one SDK for latent feature extraction
iii. one SDK for a 1-to-many match algorithm that returns a candidate list report

The fingerprint features automatically extracted by (i) and (ii) were at the discretion of the technology provider and could include the original input image(s).

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 configurations. Processors were dual 2.8 GHz/1MB Cache, Xeon (dual-core), with 800 MHz Front Side Bus for PE 1855, with 16GB RAM (15GB available to 64-bit applications; 3GB available to 32-bit applications), and 300GB 15K RPM Ultra SCSI Hard drives. Operating systems were RedHat Linux 3.1 64-bit or Microsoft Windows 2008 Server (64-bit or 32 bit), as requested by the participants.

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 SDK output

Searches were required to return a candidate list with a fixed length of one hundred (100) candidates. The candidate list consisted of

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

Note that the result is for a finger from a single subject, not a specific image, so that a single candidate could be a fused result combining matches against both a rolled and a plain impression.

Although the test plan provided for additional optional quality metrics or minutiae counts in the candidate lists, not enough of the participants returned the optional fields to be useful in analysis.

4.4 Timing requirements

The ELFT-EFS Evaluation placed limits on the processing time of the major operations involving feature extraction and enrollment (exemplars and latents) and searching, based on input from the participants. The throughput rates were relatively demanding, selected to place the participants on an equal footing, and to be comparable to some large-scale operational scenarios. It is well understood that matching accuracy is typically an inverse function of the time permitted.

Average search time requirements were specified for Subtests LC-LG (see Table 9). Throughput on Subtest LA (image-only) and LB (image+ROI) were permitted 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 seconds/10-finger exemplar set (rolled or pre-segmented slap)
Latent feature extraction	120 seconds/latent
Search	0.05 seconds/exemplar set (20 exemplar sets/seconds, per latent, assuming an exemplar set consists of 10 rolled and 10 segmented slap fingerprints)

5 Methods of analysis

Analyses of the accuracy of 1:N identification searches returning candidate lists can be with respect to rank or score.

5.1 Rank-based analyses

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.

Overall accuracy results for rank-based metrics are presented 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 (also known as "hit rate") vs. recognition rank.

Rank-based analyses are specific to the gallery size used in the test, and cannot be assumed to scale to substantially larger gallery sizes.

5.2 Score-based analyses

As defined for ELFT Phase II,

- The True Positive Identification Rate (TPIR) indicates the fraction of searches where an enrolled mate exists in the gallery in which enrolled mates appear in the top candidate list position (i.e. rank 1) with a score greater than the threshold. (Note that the False Negative Identification Rate (FNIR = 1-TPIR) indicates the fraction of searches 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 candidate list position with a score greater than the threshold.

Score-based 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. As a rule of thumb, the TPIR at FPIR = 0.01 provides a rough projection of accuracy for an increase in database size of 100x.
- 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 potential effectiveness of reverse matching.

In theory, analysis could use a combination of score and rank, in which scores are filtered based on rank. In practice, score-based results at rank 1 and at rank 100 were not notably different, so results presented are for scores at rank 1.

For score-based results, derivations of Receiver Operating Characteristic (ROC) curves were plotted using the methodology defined in ELFT Phase II ([8], 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)[e]. A horizontal line in an ROC indicates no degradation in accuracy when non-mates are excluded.

Note also that when FPIR=1.0, the score-based TPIR is the same as the rank-1 identification rate shown in the rank-based (CMC) analyses shown in Section 6.

[e] *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.*

In each case, participants returned a raw score and a normalized score estimating the probability (1-100) of a match. The normalized scores provided equal or better results than the raw scores and therefore only the normalized results are reported here.

6 Rank-based results

The following tables summarize the rank-1 identification rates for each of the matchers for each of the latent subsets as searched against exemplar set E1. Table 10 shows summary results for the Baseline-QA dataset, the (458 latent) subset of the (1,114 latent) Baseline dataset for which skeletons were available; for that reason, the Baseline-QA dataset is used to compare all of the latent subsets. Bolded results indicate best performance for each matcher. Cases in which additional features resulted in accuracy improvement are highlighted in green; cases where accuracy declined are highlighted in yellow.[f]

Table 10: Rank-1 identification rates for latent feature subsets
(Baseline-QA dataset, 458 latents — 100,000 rolled+plain 10-finger exemplar sets)

		Latent Feature Subset						
		LA	**LB**	**LC**	**LD**	**LE**	**LF**	**LG**
Matcher	**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[g]	**22.3**	*n/a*	*n/a*	14.0	*n/a*	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. Bolded results indicate best performance for each matcher.

While Baseline-QA was a randomly selected subset of Baseline, the performance for the complete Baseline dataset is better than Baseline-QA by about 3-5% in most cases; this is an incidental artifact of the process by which Baseline-QA was selected, which resulted in a greater proportion of low-quality latents. Therefore, Baseline results should not be compared directly against Baseline-QA results.

[f] *This is based on the following superset relationships (see also Table 3):*

Latent subset	*Is a superset of the features in*
LA	-
LB	*LA*
LC	*LB*
LD	*LB (or LG)*
LE	*LD (or LC)*
LF	*LE*
LG	-

[g] *Participant D's performance was erratic and therefore is not generally included in observations or conclusions.*

Table 11: Rank-1 identification rates for latent feature subsets
(Baseline dataset, 1114 latents — 100,000 rolled+plain 10-finger exemplar sets)

		Latent Feature Subset		
		LA	**LE**	**LG**
Matcher	**A**	62.2	**66.7**	44.0
	B	61.2	**63.3**	48.3
	C	48.3	**62.0**	47.8
	D	**25.1**	16.6[h]	11.9
	E	47.2	**50.3**	29.4

Image-only searches (LA) were far more accurate for matchers A/B/E than minutiae-only searches (LG). This is a notable result because the performance of 1990s AFIS would have resulted in expectations that the reverse would have been true. It should be noted that relative performance of image-based and feature-based matching may be affected by differences in systems resources, and therefore may differ among evaluations.

In general, matchers showed a net improvement when features were added, most notably between LA and LD|LE. These differences were limited in most cases; further analysis will be conducted after the Evaluation #2 results.

Overall accuracy is affected by the quality distribution of the latent and exemplar prints (see Sections 13-14); as discussed above, 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.

The following graphs show the complete CMCs for the rank-1 through rank-100 results.

Observations (all CMC graphs):

- The CMC curves for the different matchers are generally quite parallel, so that the difference in performance between two matchers or subsets does not differ substantially with respect to rank.
- The CMCs are relatively flat after rank 20: the number of additional candidates in excess of 20 is very small for the top performing matchers.
- The proportion of the total identifications made by matchers that were recorded at rank 1 ($IR_{rank\text{-}1}$ / $IR_{rank\text{-}100}$) is an indication of scalability of performance because identifications at higher ranks are less likely as gallery size increases. For matchers A/B/C, 87-92% of identifications are recorded at rank 1 for subset LA; 88-93% of identifications are recorded at rank 1 for subset LE; 78-86% of identifications are recorded at rank 1 for subset LG. (These results are reported in Appendix C).
- For matchers A/B/C/E, substantial improvement (10-20%) is shown for LE over the legacy minutiae-only feature set (LG).
- Matcher A performance for LA searches did not show substantially different performance for the Baseline and Baseline-QA datasets, unlike the other matchers. Since the Baseline-QA dataset had a greater number of poor quality prints, this result may indicate that A is more robust when presented with poor quality image data (see also Section 13).

[h] *Baseline-QA results for subset LD were used as proxy for Baseline-QA LE for the purpose of scoring Baseline LE*

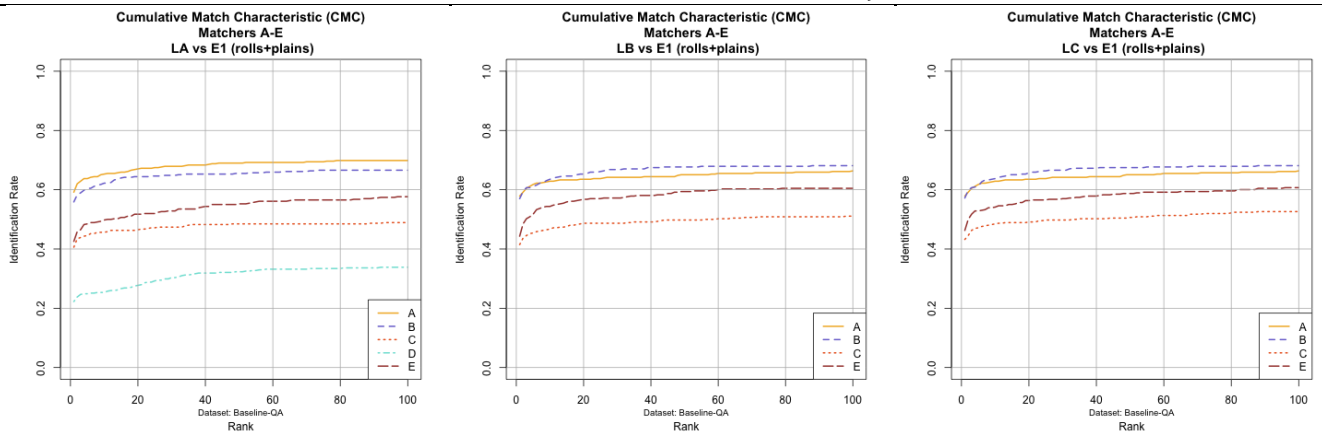


Figure 1: Rank-based comparison of matchers for latent subsets LA-LC
(Baseline-QA dataset, 458 latents — 100,000 rolled+plain 10-finger exemplar sets)

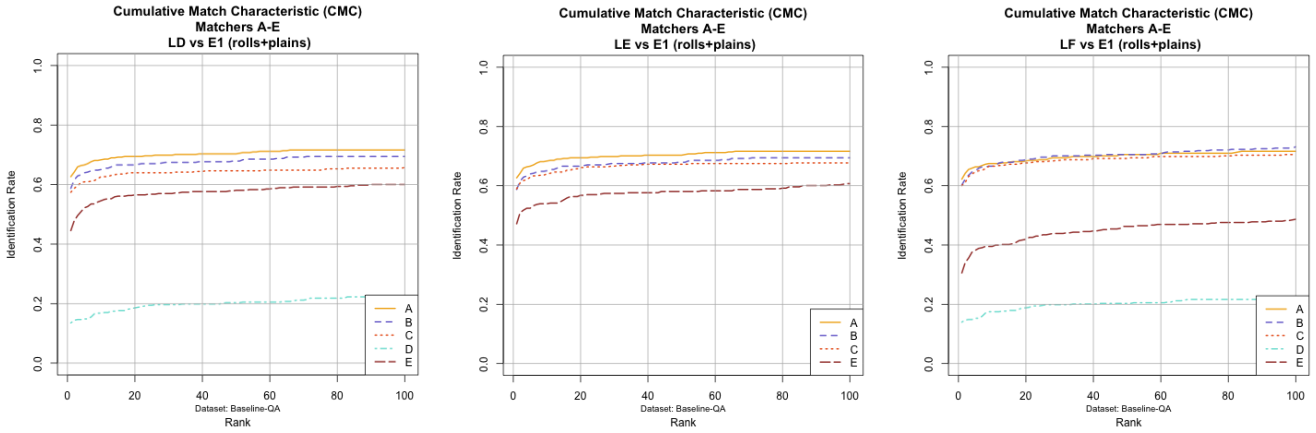


Figure 2: Rank-based comparison of matchers for latent subsets LD-LF
(Baseline-QA dataset, 458 latents — 100,000 rolled+plain 10-finger exemplar sets)

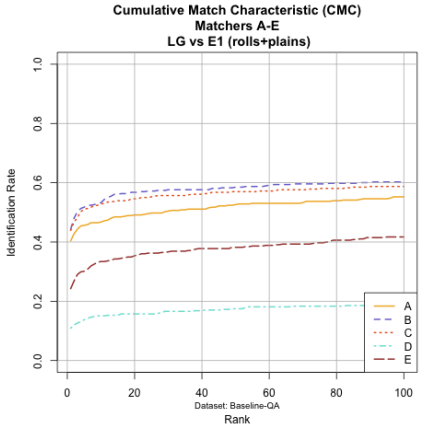


Figure 3: Rank-based comparison of matchers for latent subset LG
(Baseline-QA dataset, 458 latents — 100,000 rolled+plain 10-finger exemplar sets)

MATCHER KEY	**A** = *Sagem*	**B**=*NEC*	**C**=*Cogent*	**D**=*Sonda*	**E**=*Warwick*

FEATURE SUBSET KEY	**LA**=*Image*	**LB**=*Image +ROI*	**LC**=*Image +ROI +Quality map +Pattern class*	**LD**=*Image +ROI +Minutiae +Ridge counts*	**LE**=*Image +Full EFS (no Skeleton)*	**LF**=*Image +Full EFS with Skeleton*	**LG**=*Minutiae +Ridge counts (comp to IAFIS)*	**Page 27**

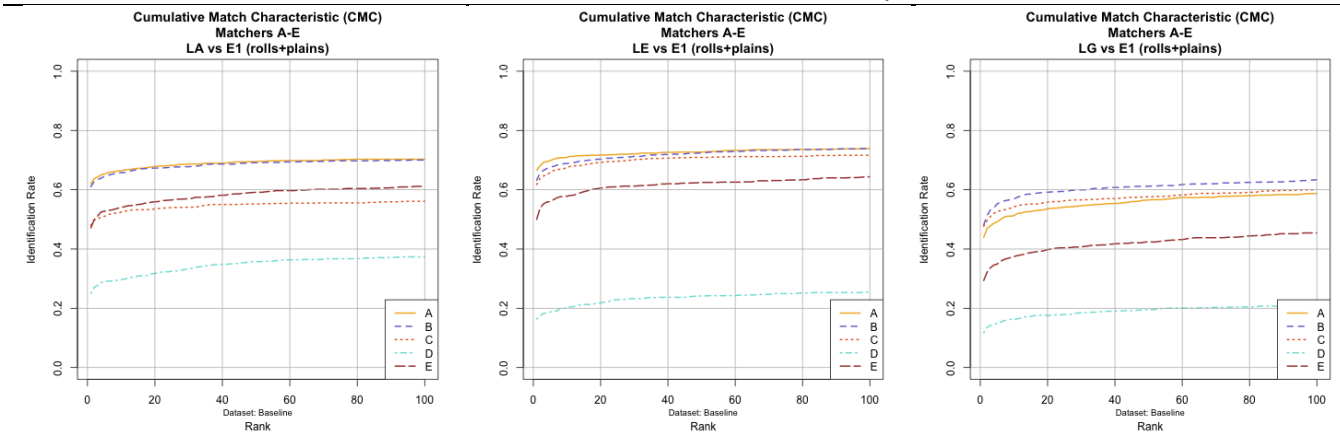


Figure 4: Rank-based comparison of matchers for latent subsets LA, LE, LG
(Baseline dataset, 1,114 latents — 100,000 rolled+plain 10-finger exemplar sets)

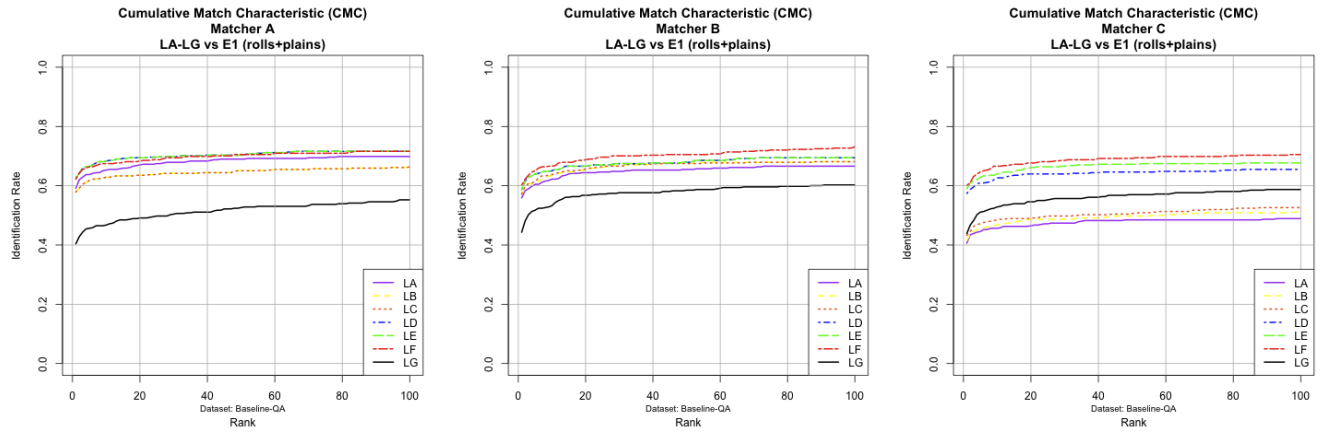


Figure 5: Rank-based comparison of all latent subsets for matchers A-C
(Baseline-QA dataset, 458 latents — 100,000 rolled+plain 10-finger exemplar sets)

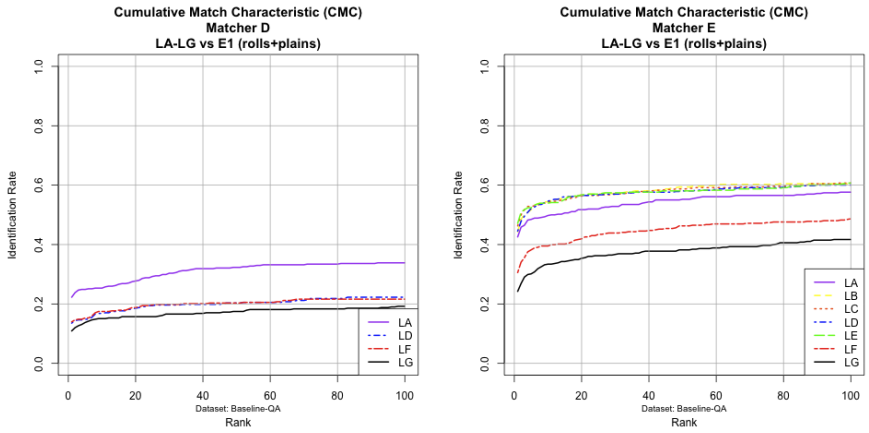


Figure 6: Rank-based comparison of all latent subsets for matchers D-E
(Baseline-QA dataset, 458 latents — 100,000 rolled+plain 10-finger exemplar sets)

MATCHER KEY	**A** = *Sagem*	**B**=*NEC*	**C**=*Cogent*	**D**=*Sonda*	**E**=*Warwick*

FEATURE SUBSET KEY	**LA**=*Image*	**LB**=*Image +ROI*	**LC**=*Image +ROI +Quality map +Pattern class*	**LD**=*Image +ROI +Minutiae +Ridge counts*	**LE**=*Image +Full EFS (no Skeleton)*	**LF**=*Image +Full EFS with Skeleton*	**LG**=*Minutiae +Ridge counts (comp to IAFIS)*	**Page 28**

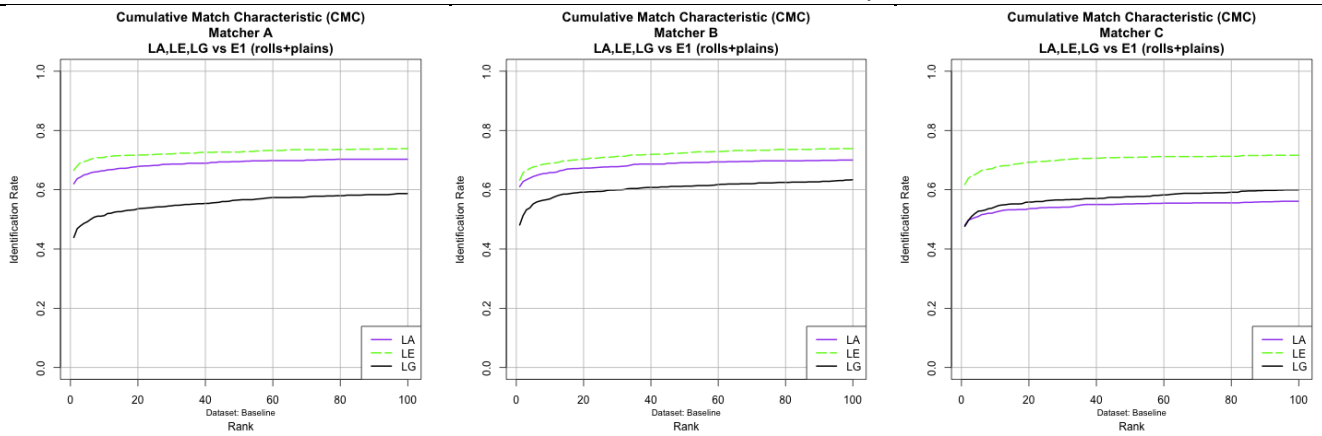


Figure 7: Rank-based comparison of latent subsets LA,LE,LG for matchers A-C
(Baseline dataset, 1114 latents — 100,000 rolled+plain 10-finger exemplar sets)

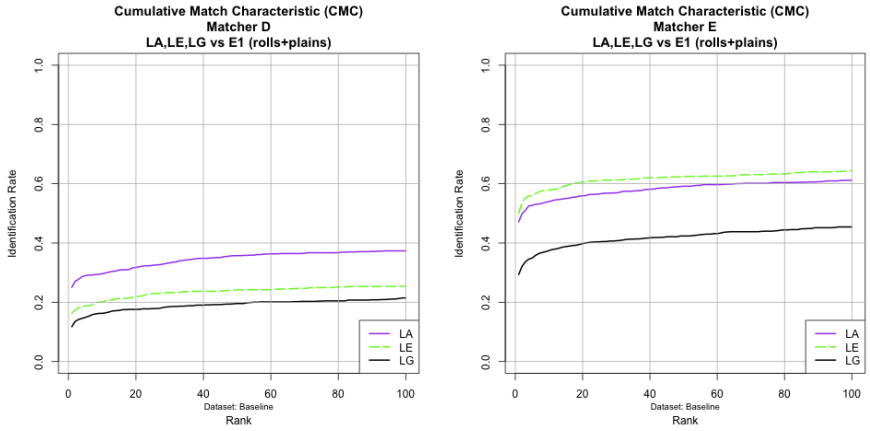


Figure 8: Rank-based comparison of latent subsets LA,LE,LG for matchers D-E
(Baseline dataset, 1114 latents — 100,000 rolled+plain 10-finger exemplar sets)

7 Score-based results

The ROC charts in Figure 9-Figure 12 show the effect of automatically filtering candidates based on score. For example, Table 10 shows that matcher A has a rank-1 IR of 0.59 (latent subset LA, Baseline-QA). In the ROCs below, the rightmost point of each curve is identical to the rank-1 result for the corresponding CMC in Section 6. From Figure 9 below we see that the TPIR 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 accuracy. If the score threshold is set to eliminate 99% of the candidate lists (FPIR=0.01), the TPIR 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. The matchers with the higher TPIR at the lower values of FPIR can also be expected to maintain a higher identification rate as gallery size increases. In general, the flatter the curve, the less sensitivity to increases in the gallery size.

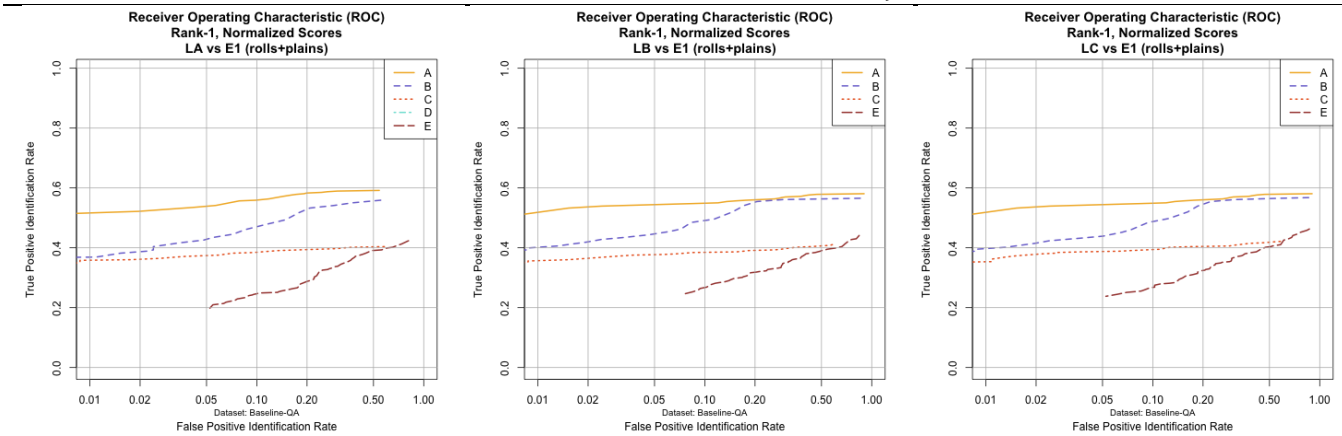


Figure 9: Score-based comparison of matchers for latent subsets LA-LC
(Baseline-QA dataset, 458 latents — 100,000 rolled+plain 10-finger exemplar sets)

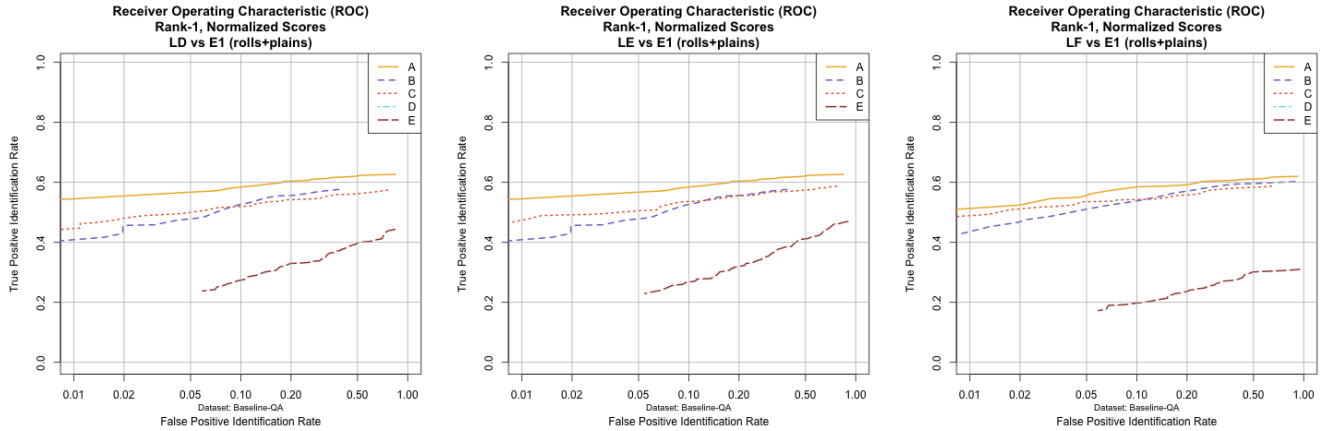


Figure 10: Score-based comparison of matchers for latent subsets LD-LF
(Baseline-QA dataset, 458 latents — 100,000 rolled+plain 10-finger exemplar sets)

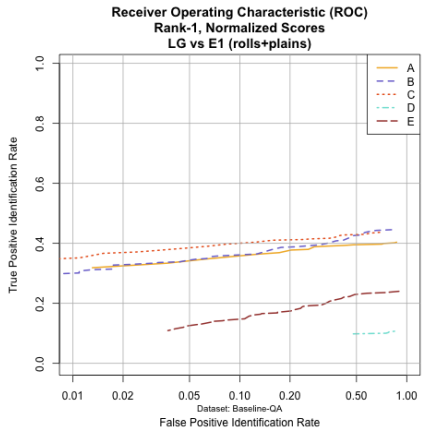


Figure 11: Score-based comparison of matchers for latent subset LG
(Baseline-QA dataset, 458 latents — 100,000 rolled+plain 10-finger exemplar sets)

MATCHER KEY	**A** = *Sagem*	**B**=*NEC*	**C**=*Cogent*	**D**=*Sonda*	**E**=*Warwick*

FEATURE SUBSET KEY	**LA**=*Image*	**LB**=*Image* *+ROI*	**LC**=*Image +ROI* *+Quality map* *+Pattern class*	**LD**=*Image +ROI* *+Minutiae* *+Ridge counts*	**LE**=*Image* *+Full EFS* *(no Skeleton)*	**LF**=*Image* *+Full EFS* *with Skeleton*	**LG**=*Minutiae* *+Ridge counts* *(comp to IAFIS)*	**Page** **30**

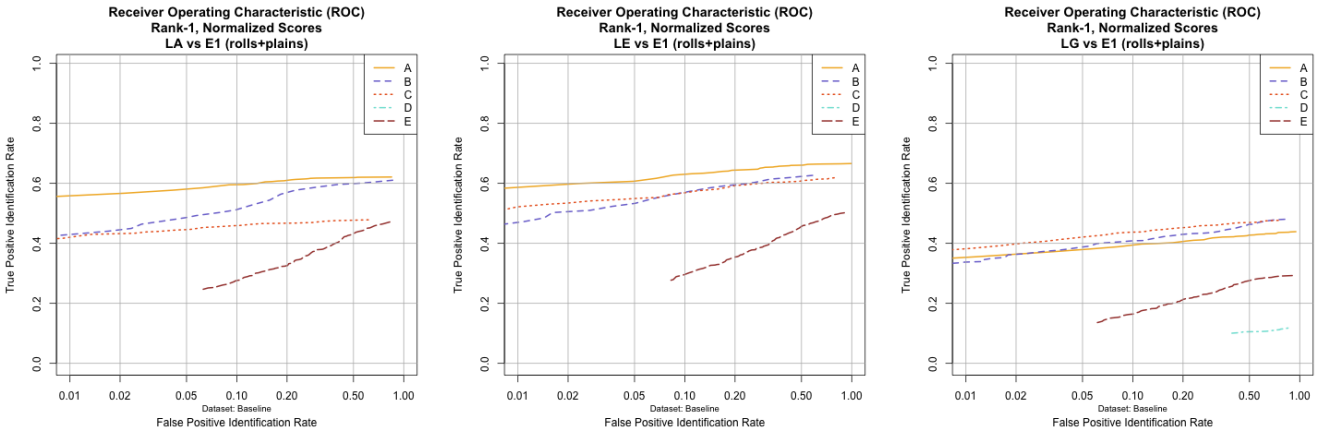


Figure 12: Score-based comparison of matchers for latent subsets LA,LE,LG
(Baseline dataset, 1114 latents — 100,000 rolled+plain 10-finger exemplar sets)

The order of the matchers with respect to accuracy varies depending on FPIR: while the CMC curves were generally parallel, the ROC curves often cross. Note that two matchers can have similar performance at a high FPIR, but show substantial differences in accuracy as FPIR approaches 0.01. For example, compare matchers A/B in subsets LA-LC, or matchers A/B/C in subsets LD-LG.

The highest overall accuracy for Baseline-QA at FPIR=0.01 was by matcher A using the LE feature subset. Matcher C had the highest performance for LG based searches. Matcher A shows the best performance at low values of FPIR (0.01) for LA-LF.

The following tables summarize the results from the ROCs above. Table 12 shows the TPIR (rank 1, normalized score) at an FPIR of 0.01 for the Baseline-QA dataset; Table 13 shows the corresponding results for the Baseline dataset. For cases in which the TPIR could not be measured at FPIR=0.01 for the specified matcher-subset combination, the closest point is indicated in gray. Highlights follow the pattern used in Table 10.

Table 12: TPIR at FPIR=0.01, for all Latent Feature Subsets
(Baseline-QA dataset, 458 latents — 100,000 rolled+plain 10-finger exemplar sets)

		Latent Feature Subset						
		LA	**LB**	**LC**	**LD**	**LE**	**LF**	**LG**
Matcher	**A**	51.3	51.9	51.9	53.9	53.9	51.1	31.9
	B	36.9	40.0	39.5	41.2	41.2	42.8	30.1
	C	35.8	36.2	35.4	46.3	47.8	48.7	35.2
	D	*n/a*	*n/a*[i]	*n/a*	*n/a*	*n/a*	*n/a*	*9.8@ FPIR=47.6*
	E	*19.2@ FPIR=0.05*	*24.5@ FPIR=0.07*	*23.8@ FPIR=0.05*	*23.8@ FPIR=0.06*	*22.9@ FPIR=0.06*	*17.2@ FPIR=0.06*	*10.9@ FPIR=0.04*

[i] *Participant D informed NIST that their software did not utilize the additional features in subsets LC - LF, and therefore subsets LB-LF were expected to have identical results. For Baseline-QA NIST ran SDK D on subsets LA, LD, LF and LG. Despite identical extracted features, results for subset LF differed slightly from LF suggesting non-deterministic behavior.*

MATCHER KEY	**A** = *Sagem*	**B**=*NEC*	**C**=*Cogent*	**D**=*Sonda*	**E**=*Warwick*

FEATURE SUBSET KEY	**LA**=*Image*	**LB**=*Image +ROI*	**LC**=*Image +ROI +Quality map +Pattern class*	**LD**=*Image +ROI +Minutiae +Ridge counts*	**LE**=*Image +Full EFS (no Skeleton)*	**LF**=*Image +Full EFS with Skeleton*	**LG**=*Minutiae +Ridge counts (comp to IAFIS)*	**Page 31**

Table 13: TPIR at FPIR=0.01, for Latent Feature Subsets LA, LE, LG

(Baseline dataset, 1114 latents — 100,000 rolled+plain 10-finger exemplar sets)

		Latent Feature Subset		
		LA	**LE**	**LG**
Matcher	**A**	55.5	58.3	34.2
	B	42.3	46.0	33.8
	C	41.7	52.2	38.3
	D	*n/a*	*n/a*	*10.0 @ FPIR=37.3*
	E	*24.0 @ FPIR=0.06*	*27.3 @ FPIR=0.08*	*13.6 @ FPIR=0.06*

8 Effect of Rolled and/or Plain Exemplars

The results discussed so far have been based on searches of gallery E1 (described in Table 8) which consists of ten rolled and ten plain impressions for each subject. Searches were also conducted against two other galleries, E2 and E3 which consist, respectively, of only ten rolled impressions per subject, and only ten plain impressions per subject, in order to measure the effect of exemplar impression type on identification performance. Table 14 compares the rank-1 accuracy for the different exemplar types. A general expectation would be that rolled+plain (E1) accuracy would always be higher than either rolled (E2) or plain (E3) separately, and that accuracy for rolled (E2) would generally be higher than plain (E3); cases that do not follow this expectation are highlighted.

Table 14: Rank-1 accuracy by exemplar type: rolled+plain (E1); rolled (E2); plain (E3)

(Baseline dataset, 1114 latents —10-finger exemplar sets: E1: 100,000 rolled+plain; E2: 10,000 rolled; E3: 10,000 plain)

		Latent Feature Subset / Exemplar Type								
		LA			LE			LG		
		E1	E2	E3	E1	E2	E3	E1	E2	E3
Matcher	A	62.2	57.1	50.1	66.7	61.0	52.2	44.0	39.7	32.9
	B	61.2	57.1	49.5	63.3	60.8	52.8	48.3	55.6	47.8
	C	48.3	46.7	41.2	62.0	57.1	49.5	47.8	43.2	44.9
	D	25.1	19.6	20.4	16.6	14.2	11.8	11.9	11.8	11.6
	E	47.2	41.1	36.2	50.3	44.2	38.4	29.4	25.4	21.4

Table 15 shows the data from Table 14 in terms of relative gains in accuracy. In most cases rolled+plain is more accurate than rolled alone (E1-E2), and is much more accurate than plain alone (E1-E3); in most cases, rolled alone is more accurate than plain alone (E2-E3).

Table 15: Differences in rank-1 accuracy by exemplar type:
rolled+plain vs. rolled (E1-E2); rolled+plain vs. plain (E1-E3); rolled vs. plain (E2-E3)
(Baseline dataset, 1114 latents —10-finger exemplar sets: E1: 100,000 rolled+plain; E2: 10,000 rolled; E3: 10,000 plain)

		E1-E2			E1-E3			E2-E3		
		LA	LE	LG	LA	LE	LG	LA	LE	LG
Matcher	A	5.1	5.7	4.3	12.1	14.5	11.1	7.0	8.8	6.8
	B	4.1	2.5	-7.3	11.7	10.5	0.5	7.6	8.0	7.8
	C	1.6	4.9	4.6	7.1	12.5	2.9	5.5	7.6	-1.7
	D	5.5	2.4	0.1	4.7	4.8	0.3	-0.8	2.4	0.2
	E	6.1	6.1	4.0	11.0	11.9	8.0	4.9	5.8	4.0

CMCs comparing searches of E1, E2 and E3 are provided in Appendix C.

9 Effect of Examiner Markup Method

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 "Ground Truth" 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 (e.g. NIST SD27).

In every other section of this report, all latent images were marked by examining the latent in isolation. 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 16.

Table 16: Methods of examiner markup

Markup method	Description
(standard)	Features were marked in latent images without reference to exemplars. Used for Baseline and Baseline-QA.
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.[j]
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[k] 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.

[j] *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.*

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

These results depict the difference in accuracy between theoretically ideal markup and operational human markup. Table 17 and Table 18 summarize rank-1 performance data for searches using the different feature markup sets. In each case, the same set of latents is used, varying only by the markup. Groundtruth is indicated with a "G" superscript, and AFIS is indicated with an "A" superscript (e.g. LE^G, or LE^A).

Table 17: Comparison of rank-1 IR for standard, groundtruth, and AFIS markup methods (LE and LG)

(Baseline-QA dataset, 458 latents — 100,000 rolled+plain 10-finger exemplar sets)

	LA		LE	LE^G	LE^A		LG	LG^G	LG^A
A	59.4		65.6	70.9	64.0		43.2	54.3	42.3
B	56.3		62.4	70.4	62.1		47.6	62.6	48.5
C	40.6		61.4	69.3	60.7		46.0	61.0	47.8
D	22.3		15.2	15.5	16.2		11.6	13.4	10.4
E	42.6		50.1	48.7	49.4		25.6	38.3	24.3

Table 18: Differences in rank-1 IR for standard, groundtruth, and AFIS markup methods (LE and LG)

(Baseline-QA dataset, 458 latents — 100,000 rolled+plain 10-finger exemplar sets)

	Effect of Ground truth markup		Effect of AFIS markup	
	LE^G - LE	LG^G – LG	LE^A – LE	LG^A-LG
A	5.3	11.1	-1.6	-0.9
B	8.0	15.0	-0.3	0.9
C	7.9	15.0	-0.7	1.8
D	0.3	1.8	1.0	-1.2
E	-1.4	12.7	-0.7	-1.3

The CMCs below show the effect on rank-1 accuracy of the different markup methods, comparing the latent subsets LE and LG to image-only searches (LA).

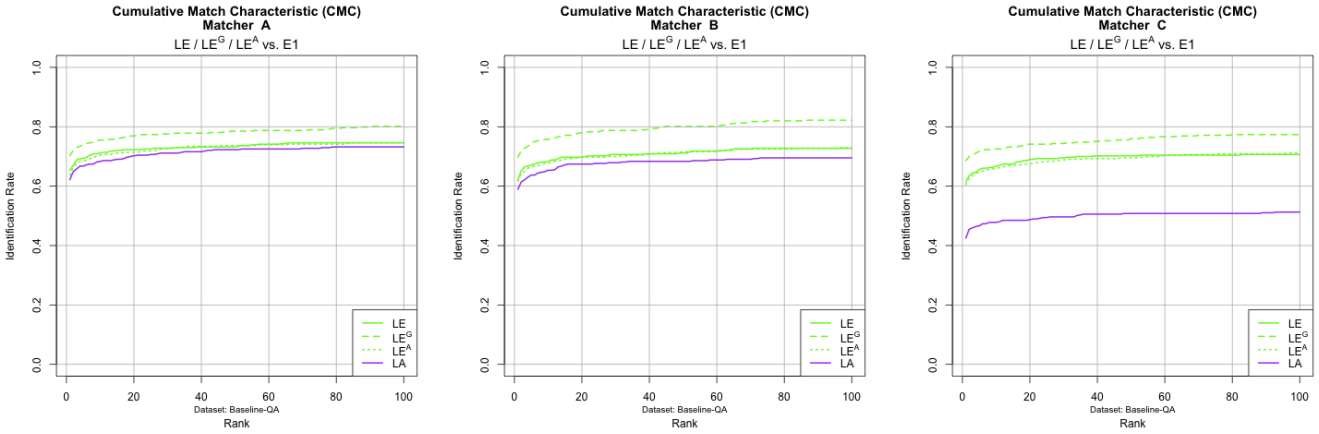


Figure 13: Rank-based comparison of standard, groundtruth, and AFIS markup methods for latent subset LE (matchers A,B,C)

(Baseline-QA dataset, 458 latents — 100,000 rolled+plain 10-finger exemplar sets)

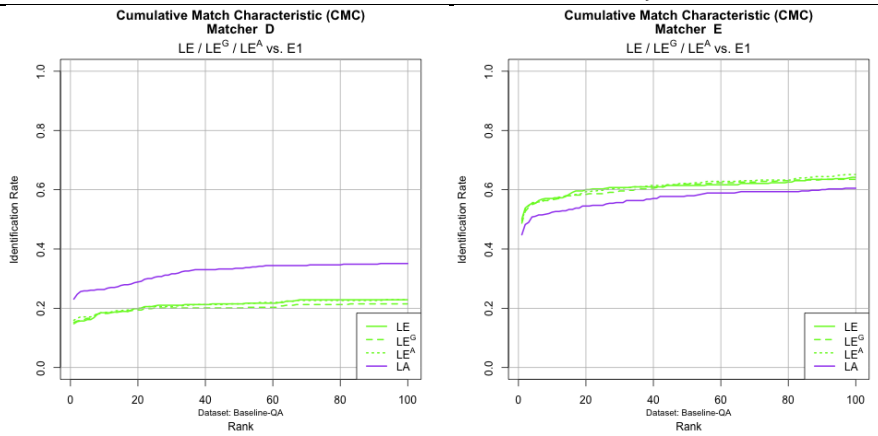


Figure 14: Rank-based comparison of standard, groundtruth, and AFIS markup methods for latent subset LE (matchers D,E)

(Baseline-QA dataset, 458 latents — 100,000 rolled+plain 10-finger exemplar sets)

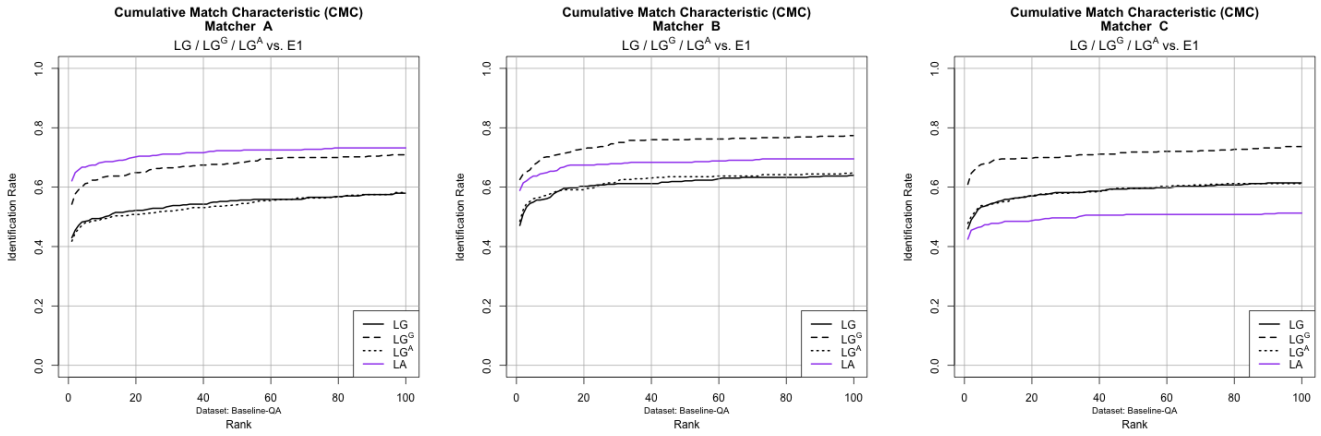


Figure 15: Rank-based comparison of standard, groundtruth, and AFIS markup methods for latent subset LG (matchers A,B,C)

(Baseline-QA dataset, 458 latents — 100,000 rolled+plain 10-finger exemplar sets)

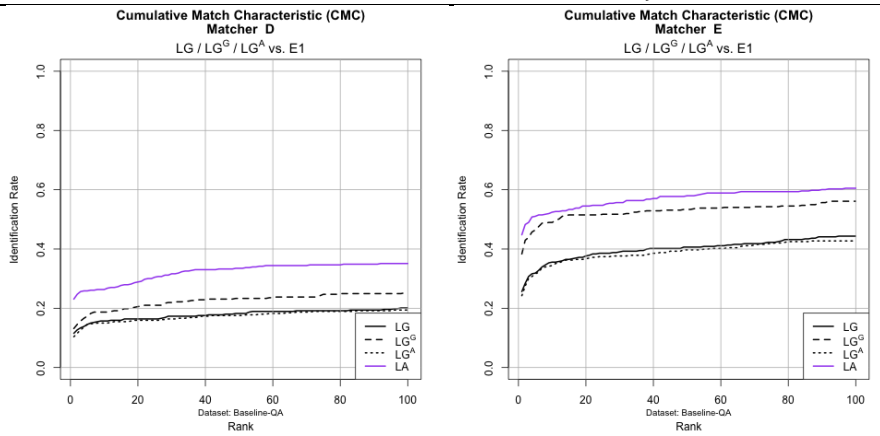


Figure 16: Rank-based comparison of standard, groundtruth, and AFIS markup methods for latent subset LG (matchers A,B,C)

(Baseline-QA dataset, 458 latents — 100,000 rolled+plain 10-finger exemplar sets)

The groundtruth (GT) results were beneficial using latent subset LE (matchers A/B/C), but were dramatically beneficial for using latent subset LG (matchers A/B/C/E). In practice this means that for minutiae-only markup (LG), there is about 11-15% difference in rank-1 accuracy between ideal and ordinary examiner markup; this difference is pronounced since the matcher had no recourse to the image. When the image and features are included (LE), the difference drops to about 5-8%. Matchers B/C derived the most benefit overall from "GT". markup. Note that for matchers A/D/E, image-only matching is more accurate than minutiae-only matching even when using groundtruth features.

The "AFIS" markup approach provided was counterproductive in most cases, and marginally beneficial in a few others. This result indicates that the matchers are relatively robust when processing debatable minutiae.

10 Effect of Latent Data Source

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

All matchers were shown to be sensitive to dataset characteristics that are created as part of the dataset capture process. The difference in rank-1 IR between the WVU and the FLDS sources was over 20% for matchers A and B using latent feature subset LA, and was even higher for participants C and E. For latent feature subset LE, a similar difference was noted for participants A, B, C, and E. The differences in rank-1 identification rate across the data sources were even greater when latent feature subset LG was used. The WVU data, which contains the greatest average number of minutiae, was approximately 30% more accurate for matchers A, B, C, and E when compared with Casework 2.

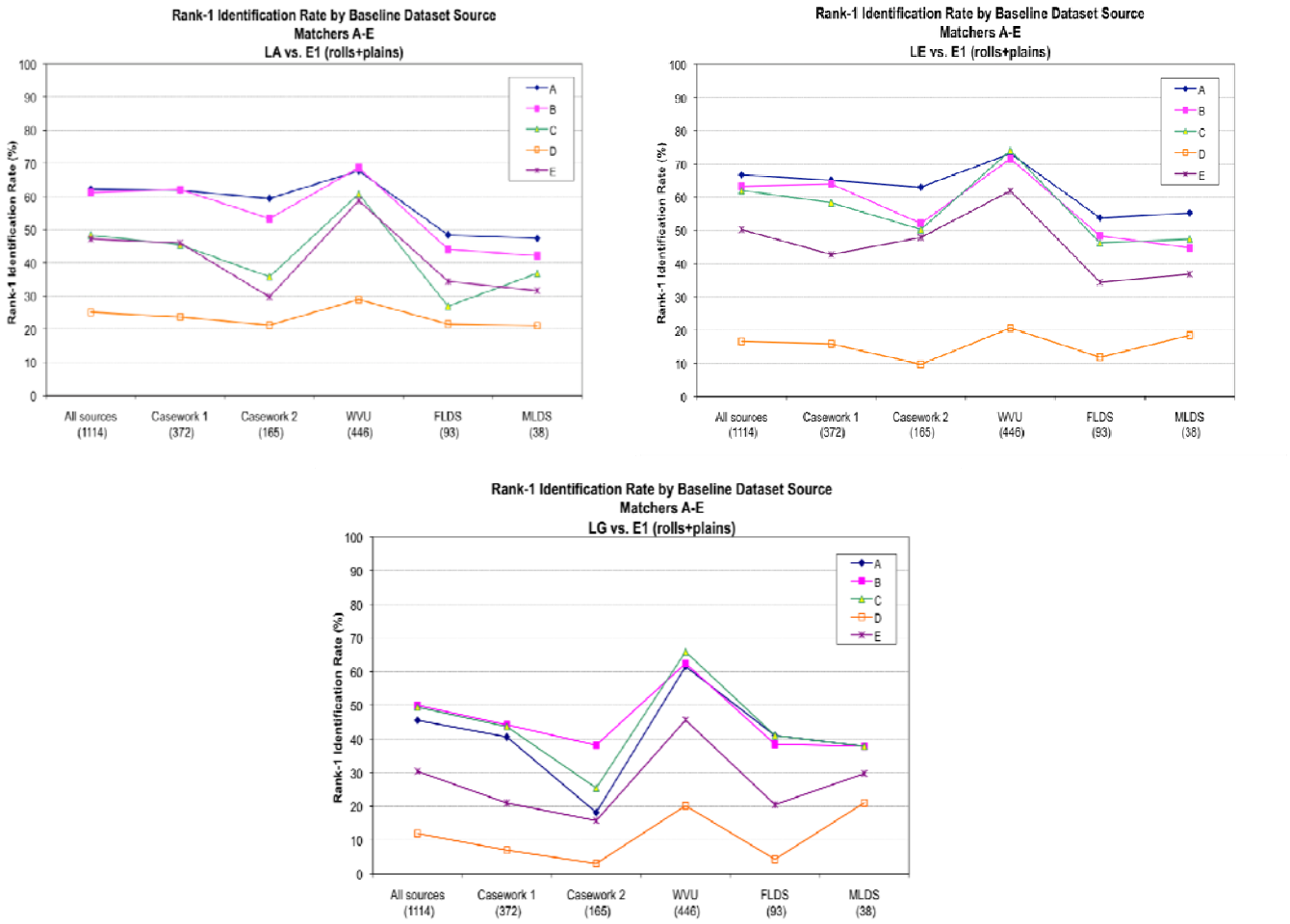


Figure 17: Comparison of rank-1 IR by latent data source
(Baseline dataset, 1114 latents — 100,000 rolled+plain 10-finger exemplar sets)

MATCHER KEY	**A** = *Sagem*	**B**=*NEC*	**C**=*Cogent*	**D**=*Sonda*	**E**=*Warwick*

FEATURE SUBSET KEY	**LA**=*Image*	**LB**=*Image* *+ROI*	**LC**=*Image +ROI* *+Quality map* *+Pattern class*	**LD**=*Image +ROI* *+Minutiae* *+Ridge counts*	**LE**=*Image* *+Full EFS* *(no Skeleton)*	**LF**=*Image* *+Full EFS* *with Skeleton*	**LG**=*Minutiae* *+Ridge counts* *(comp to IAFIS)*	**Page 37**

11 Effect of Latent Orientation

This analysis shows the effect of orientation on rank-1 identification rate (see Section 3.1.4). The latents from the Baseline dataset were grouped by orientation (rotation from upright). Note that the first two bins are aggregates of the other bins: 'All' contains all latents in the Baseline dataset, regardless of orientation; '0-45 degrees' is shown because it illustrates a range typical for operational searches. The latents in bin '0-45 degrees' comprise 87.6 % of the latents in the Baseline dataset. LE and LG searches included orientation information (when known); for LA searches, the orientation was never known to the participants.

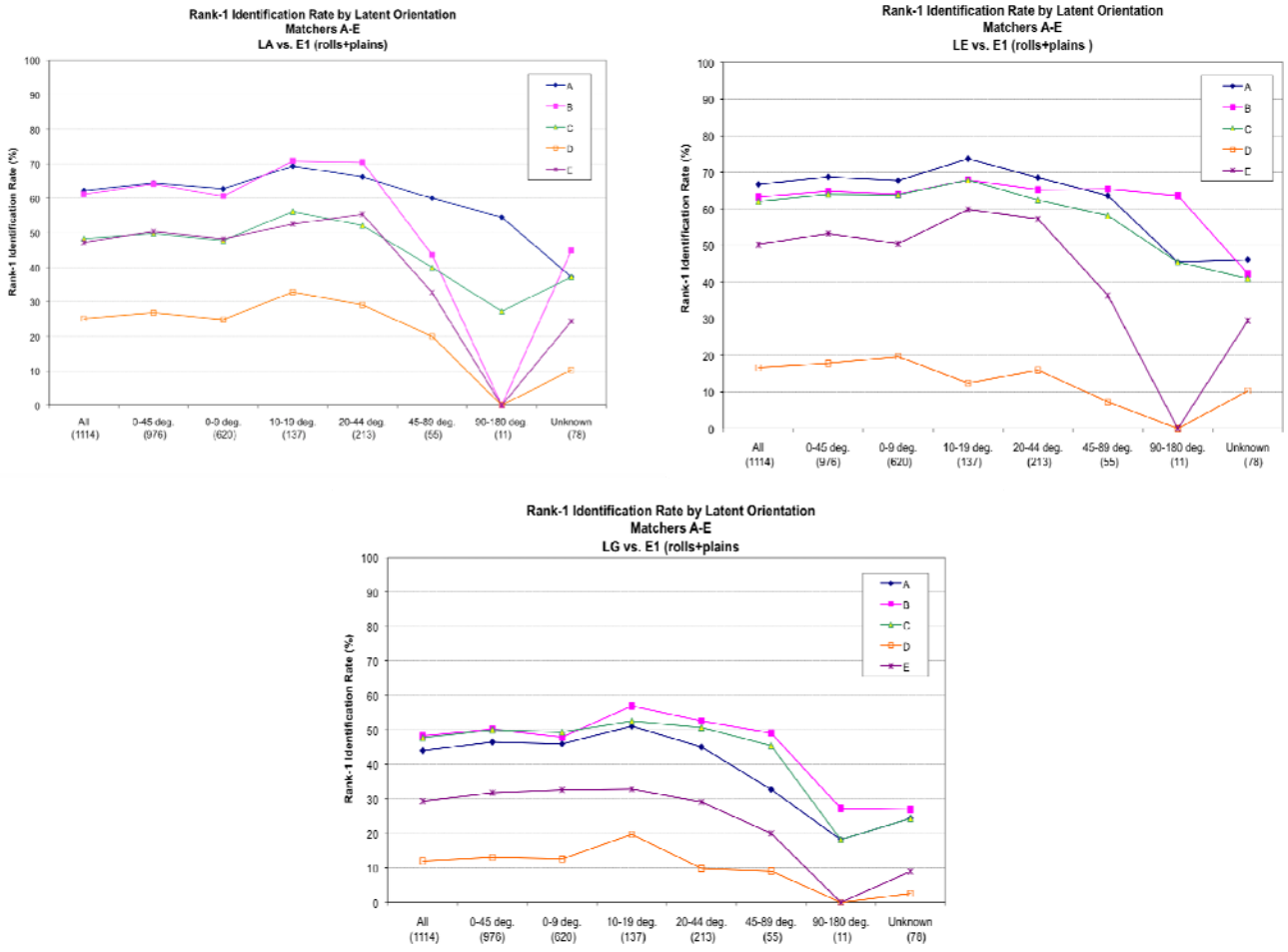


Figure 18: Comparison of rank-1 IR by orientation of latent image
(Baseline dataset, 1114 latents — 100,000 rolled+plain 10-finger exemplar sets)

For all matchers on all latent feature subsets, searches of latents with a known orientation in the range 0 to 45 degrees are on average more accurate than for latents that are 45-90 degrees from vertical, and are usually much more accurate than latents rotated more than 90 degrees.

12 Effect of Latent Minutiae Count

The following charts show rank-1 identification rates broken into bins by minutiae count, for latent subsets LA, LE, and LG. True positive identification rates (TPIR) at rank-1 and FPIR=0.01, where available, are also shown for comparison.

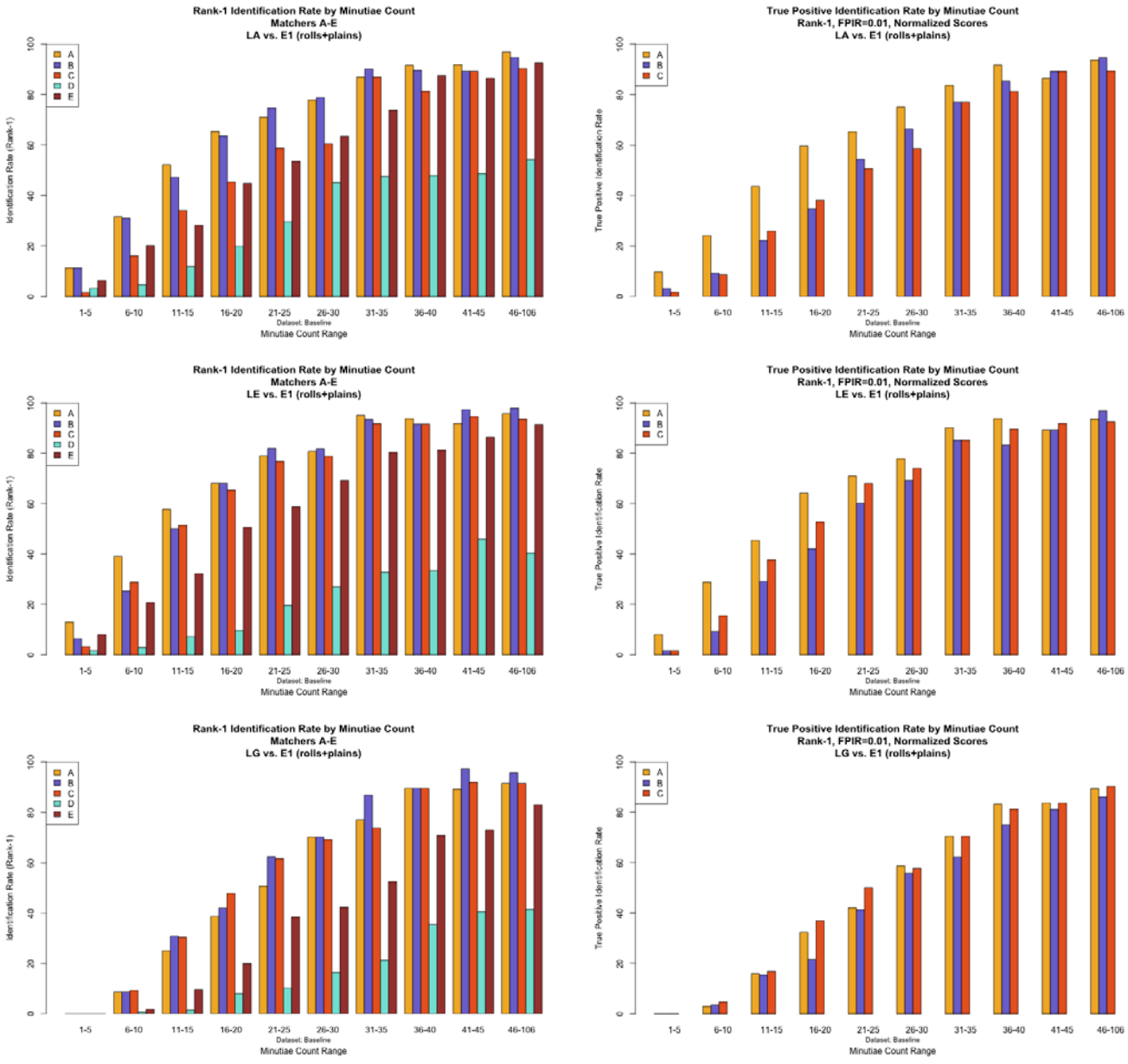


Figure 19: Rank and score-based accuracy by number of minutiae
(Baseline dataset, 1114 latents — 100,000 rolled+plain 10-finger exemplar sets)

The following chart shows rank-1 identification rate change (difference) for searches of latents in the Baseline dataset, between latent subset LA (image-only) and latent subset LE (image+EFS), broken out by minutiae count

range. A positive difference indicates an increase in rank-1 identification rate performance when image + EFS (LE) searches are used instead of image-only (LA) searches.

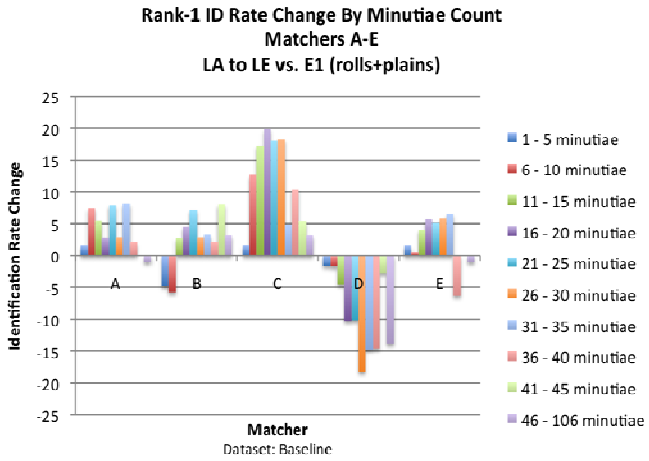


Figure 20: Difference in rank-1 accuracy between LA and LE, by number of minutiae
(Baseline dataset, 1114 latents — 100,000 rolled+plain 10-finger exemplar sets)

Performance improves linearly as a function of minutiae count, however, the performance levels off significantly for all matchers at about 30 minutiae.

Matchers A/B achieve a 10%+ identification rate even on latents with 1-5 minutiae, for subset LA; matcher A's corresponding TPIR drops to 8%+ at FPIR=0.01. 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. Matcher B unexpectedly exhibited performance loss to zero gain for image + feature searches (LE or LD) compared with image-only (LA) searches in the 1-10 minutiae count range.

For latents with more than 30 minutiae, the collective miss rates for the leading matchers were from 3.8% to 4.6%, depending upon the latent feature subset. Only three latents with over 30 minutiae were missed by all matchers[l]; see Section 15.

Most matchers showed performance increases for image+feature searches (LE or LD) compared with image-only (LA) searches across all minutiae count ranges. This performance increase grew in magnitude as minutiae count increased, with the maximum gains for most matchers occurring in the 21 to 25 minutiae count range. Beyond this number of minutiae, the increases tend to level out and/or decrease in magnitude.

Overall, there is an indication that some matchers may not benefit from manual markups over automated feature extraction for certain levels of minutiae. However, no definitive level for when markup is required can be defined. This level is highly dependent on the extent of the overlap of the latent with the exemplar. Since this degree of overlap cannot be known a priori, it may not be possible to establish a markup/no markup threshold.

13 Effect of Latent Value Determination

A 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 (see Table 6 for minutiae counts by value determination). Table 19 and

[l] *Omitting latents without mates in the gallery.*

Table 20 show the relationship between value determination and rank-based and score-based accuracy. Value determination assessments are only in reference to the latent, and do not consider exemplar quality, which will be investigated in the analysis of Evaluation #2 results.

Table 19: Rank-1 identification rate by value determination

(Baseline dataset, 1114 latents — 100,000 rolled+plain 10-finger exemplar sets)

		All	No Value	Limited Value	Value
Count		1114[m]	25	122	956
LA	A	62.2%	20.0%	26.2%	67.9%
	B	61.2%	4.0%	19.7%	67.9%
	C	48.3%	0.0%	14.8%	53.9%
	D	25.1%	0.0%	3.3%	28.7%
	E	47.2%	0.0%	13.9%	52.5%
LE	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%
	D	16.6%	0.0%	2.5%	18.9%
	E	50.3%	8.0%	12.3%	56.3%
LG	A	44.0%	0.0%	4.9%	50.0%
	B	48.3%	4.0%	3.3%	55.2%
	C	47.8%	4.0%	7.4%	54.0%
	D	11.9%	0.0%	1.6%	13.6%
	E	29.4%	0.0%	1.6%	33.8%

[m] *Note: 11 latents (out of 1114) in Baseline, and 5 latents (out of 458) in Baseline-QA did not have value determinations.*

Table 20: Rank-1 TPIR at FPIR=1% by value determination
(Baseline dataset, 1114 latents — 100,000 rolled+plain 10-finger exemplar sets)

		All	No Value	Limited Value	Value
Count		1114[n]	25	122	956
LA	A	55.5%	8.0%	20.5%	62.7%
	B	42.3%	4.0%	5.7%	49.1%
	C	41.7%	0.0%	9.8%	48.0%
	D	-	-	-	-
	E	-	-	-	-
LE	A	58.3%	8.0%	25.4%	65.8%
	B	46.0%	0.0%	7.4%	53.7%
	C	52.2%	4.0%	13.9%	59.2%
	D	-	-	-	-
	E	-	-	-	-
LG	A	34.2%	0.0%	0.8%	42.8%
	B	33.8%	0.0%	1.6%	39.0%
	C	38.3%	0.0%	3.3%	44.8%
	D	-	-	-	-
	E	-	-	-	-

Figure 21 shows rank-1 identification rate change (difference) for searches of latents in the Baseline-QA dataset, between various latent subsets, broken out by latent value determination. For example in the first chart below, "LA to LD" means that the first set of searches used the feature subset LA, and the second set used the feature subset LD, thus the difference in ID rate is a measure of "benefit" from supplying minutiae and ridge count information to the matcher in addition to the image. In some latent value determination classes the difference in ID rate is not always positive: a negative ID rate differences indicates that the matcher actually performed worse when additional features were added.

[n] *Note: 11 latents (out of 1114) in Baseline, and 5 latents (out of 458) in Baseline-QA did not have value determinations.*

MATCHER KEY	**A** = *Sagem*	**B**=*NEC*	**C**=*Cogent*	**D**=*Sonda*	**E**=*Warwick*

FEATURE SUBSET KEY	**LA**=*Image*	**LB**=*Image +ROI*	**LC**=*Image +ROI +Quality map +Pattern class*	**LD**=*Image +ROI +Minutiae +Ridge counts*	**LE**=*Image +Full EFS (no Skeleton)*	**LF**=*Image +Full EFS with Skeleton*	**LG**=*Minutiae +Ridge counts (comp to IAFIS)*	**Page 42**

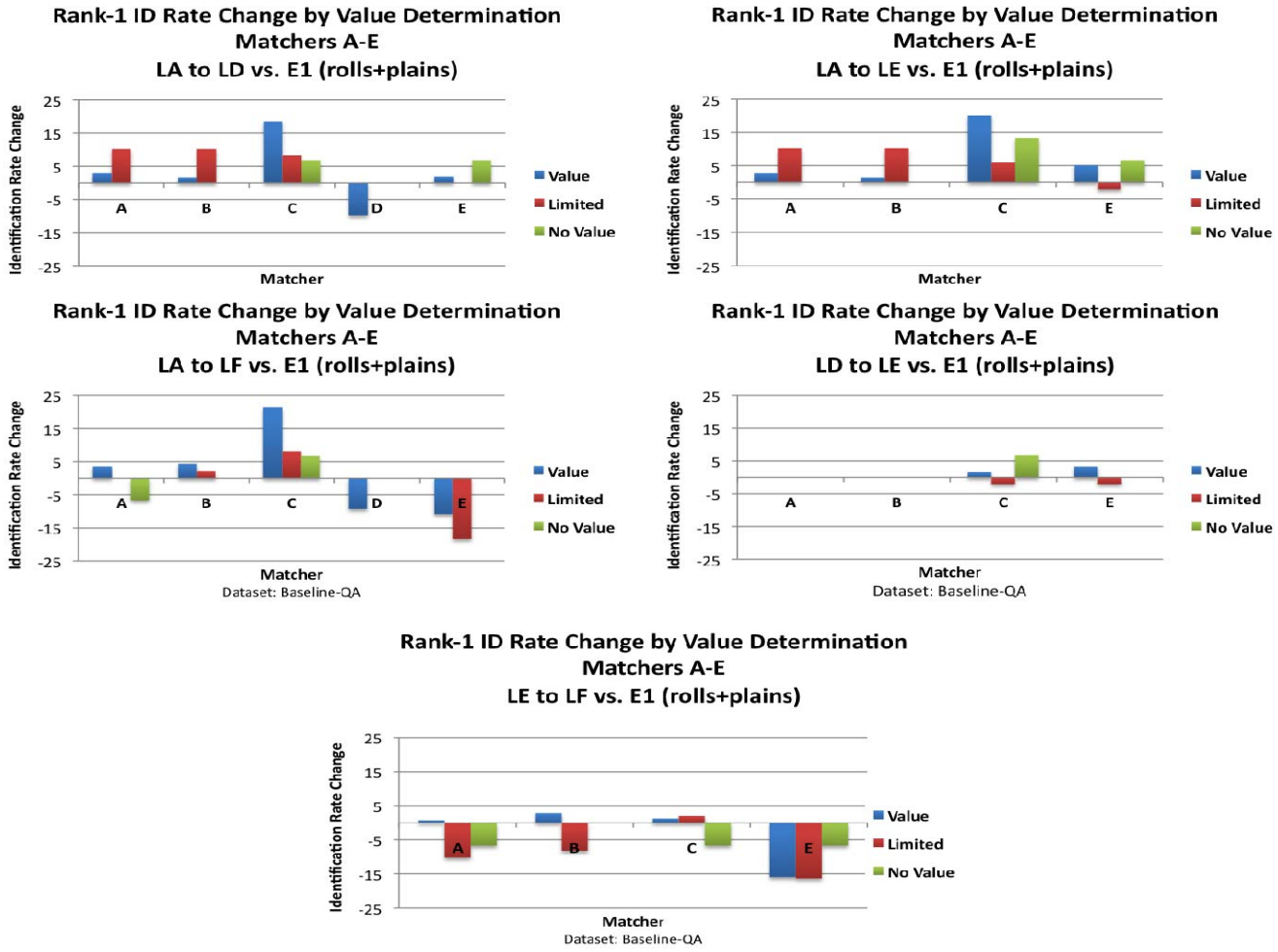


Figure 21: Difference in rank-1 identification rate between latent subsets, by latent value
(Baseline-QA dataset, 458 latents — 100,000 rolled+plain 10-finger exemplar sets)

As expected, accuracy is very clearly related to latent value determination, with much greater accuracy for the latents determined a priori to be of value.

The notable and surprising result is that matcher A's rank-1 identification rate for No Value latents is 20% on subsets LA/LE, and even higher (26.2%) on Limited Value latents. Their corresponding TPIR rates decrease at FPIR=0.01, though they are still notable. These results echo the performance on latents with few minutiae, as discussed in Section 12. The results for participants A/B/C/E show that matching may be practical even for Limited Value or No Value latents, given lower expectations of accuracy (latents considered to have no value may not be subject to individualization or elimination, but they could be used as an investigative tool). The decision to include No Value and Limited Value latents in the test has been justified. The ability to match these low value prints requires the image; none of the No-Value images could be matched using minutiae alone at rank 1 or FPIR=0.01.

The ordering of matchers by rank-1 identification rate tended to remain the same across different latent value determinations, though matcher A performs relatively better as quality worsens. The identification rates for matchers A/B are equal for Value latents, but for Limited and No Value latents matcher A's identification rate becomes increasingly greater than matcher B's.

MATCHER KEY	**A** = *Sagem*	**B**=*NEC*	**C**=*Cogent*	**D**=*Sonda*	**E**=*Warwick*

FEATURE SUBSET KEY	**LA**=*Image*	**LB**=*Image +ROI*	**LC**=*Image +ROI +Quality map +Pattern class*	**LD**=*Image +ROI +Minutiae +Ridge counts*	**LE**=*Image +Full EFS (no Skeleton)*	**LF**=*Image +Full EFS with Skeleton*	**LG**=*Minutiae +Ridge counts (comp to IAFIS)*	**Page** **43**

Matcher A and B's identification rate gains from adding feature markup (minutiae or EFS) to the search, versus image-only, were greater for Limited Value latents than for Value latents. The opposite was true for matcher's C and E. Matcher C's identification rate gain from minutiae markup always decreases as quality decreases (the opposite of the general trend).

14 Effect of Latent Good / Bad / Ugly Quality Classifications

As discussed in Section 3.1.3, a set of latent examiners categorized the latents according to an informal Excellent, Good, Bad, Ugly, No Value quality scale[o]; these examiners were different from those providing feature markup or value assessments. See Table 6 for minutiae counts by value determination; see Table 5 for comparison of value and quality assessments.

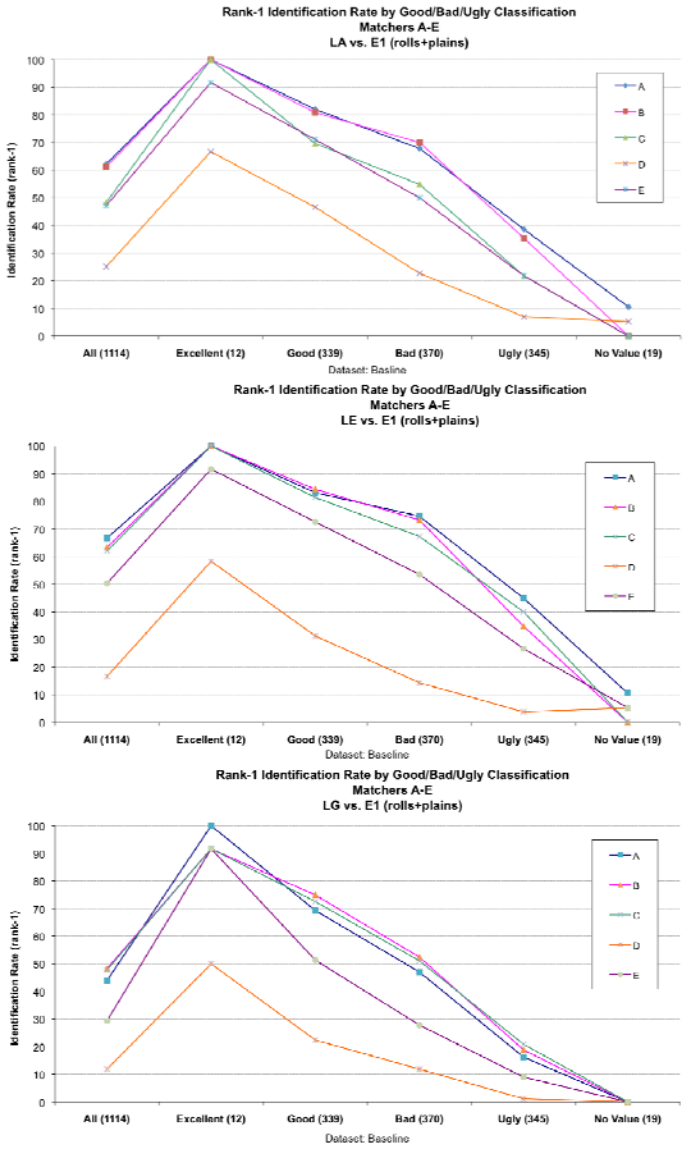


Figure 22: Rank-1 identification rate by Good/Bad/Ugly quality assessment
(Baseline dataset, 1114 latents — 100,000 rolled+plain 10-finger exemplar sets)

[o] *29 of the Baseline latents, including 9 of the Baseline-QA latents, did not have the Good-Bad-Ugly quality assessments.*

MATCHER KEY	**A** = *Sagem*	**B**=*NEC*	**C**=*Cogent*	**D**=*Sonda*	**E**=*Warwick*

FEATURE SUBSET KEY	**LA**=*Image*	**LB**=*Image +ROI*	**LC**=*Image +ROI +Quality map +Pattern class*	**LD**=*Image +ROI +Minutiae +Ridge counts*	**LE**=*Image +Full EFS (no Skeleton)*	**LF**=*Image +Full EFS with Skeleton*	**LG**=*Minutiae +Ridge counts (comp to IAFIS)*	**Page 44**

The following charts show rank-1 identification rate change (difference) for latents in the Baseline-QA dataset between different searches of latent feature subsets, broken out by Good / Bad / Ugly classification. A positive difference indicates an increase in rank-1 ID rate performance between successive sets of searches of the same latents where the second set of searches supplies additional feature information to the matcher. For example in the first chart below, "LA to LD" means that the first set of searches used the feature subset LA, and the second set used the feature subset LD, thus the difference in ID rate is a measure of "benefit" from supplying minutiae and ridge count information to the matcher in addition to the image. In some Good / Bad / Ugly classes the difference in ID rate is not always positive, a negative ID rate differences indicates that the matcher actually performed worse when additional features were added.

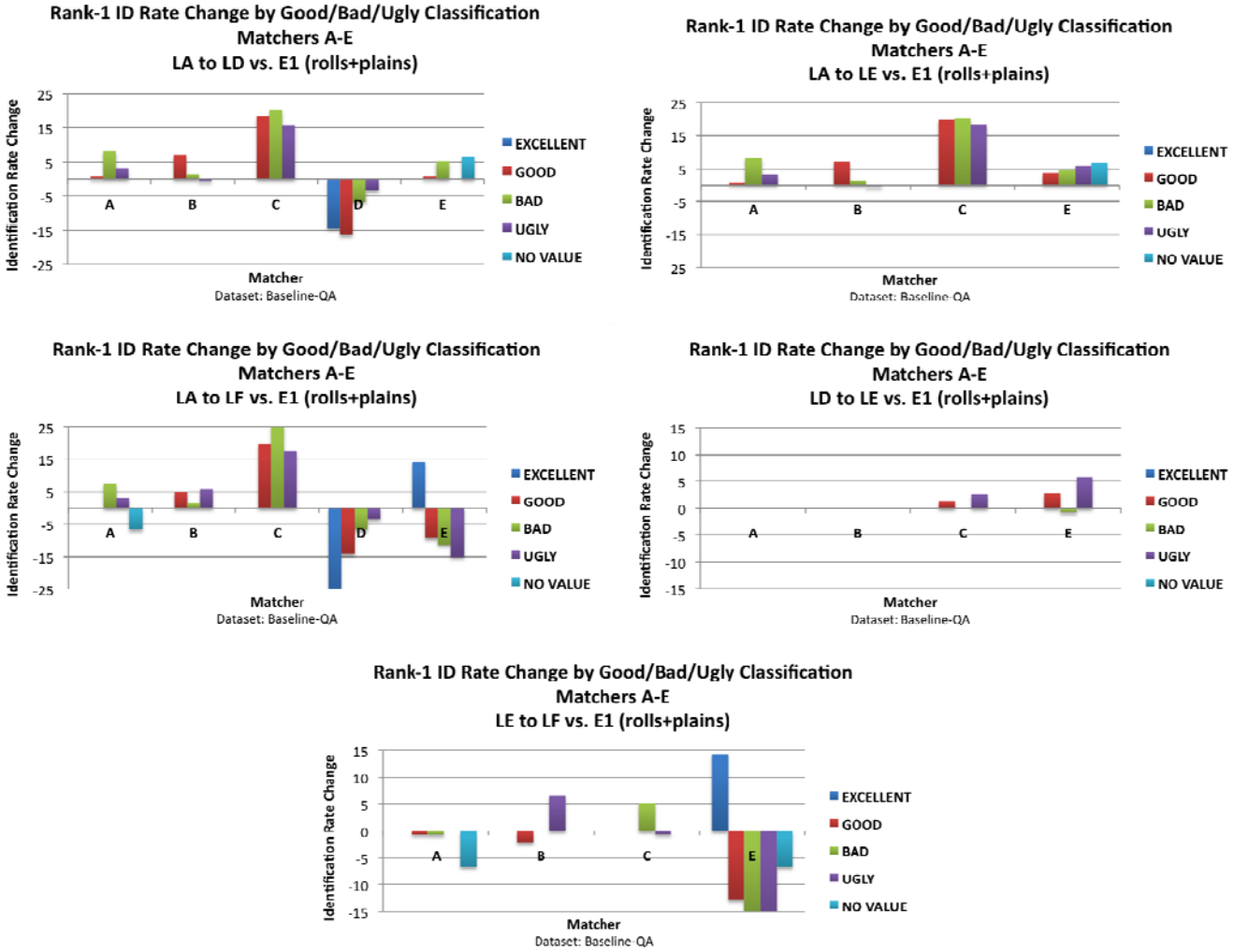


Figure 23: Difference in rank-1 identification rate between latent subsets, by Good/Bad/Ugly quality assessment
(Baseline-QA dataset, 458 latents — 100,000 rolled+plain 10-finger exemplar sets)

For each feature subset, matcher performance is closely correlated with the Good/Bad/Ugly quality assessment, as would be expected from the relationship between quality assessment and minutiae. Most matchers achieve 90-100% accuracy for the Excellent latents, and all matchers show clear decreases in accuracy with respect to quality. Results for the Baseline set were approximately the same as for the Baseline-QA set.

15 Hit / Miss / Loss and Gain Analysis

15.1 Miss Analysis

Following completion of the test, the results were reviewed by a certified latent fingerprint examiner in order to assess the difficulty of human-based matching for the low performing latent mate pairs, as well as to detect any issues with the latent mate associations used to score the results. Based on the results, latent mate pairs were selected for review if for any latent subset (i) none of the matchers placed a mate exemplar on a candidate list at rank 1; (ii) any matcher placed a mate on a candidate list from the correct subject exemplar set, but wrong finger position; (iii) all matchers placed a low ranking and/or low scoring mate exemplar on a candidate list (iv); or any matcher placed a high ranking and/or high scoring non-mate exemplar. This resulted in a subsequent review of 311 cases. The examiner could individualize all of these cases, with these exceptions:

- The examiner determined that 9 of the latents were associated with the correct subject but the wrong finger position was indicated. These ***have been corrected*** for the results shown in this report.
- For 10 of the latents, it was determined that due to administrative error, the correct mate was not in the gallery: the examiner was able to exclude all of the fingers from the exemplars for the specified subject; all of these cases were limited to the Baseline dataset. Because these cases were found late in the reporting process, the data ***are not corrected*** for this, with the result that measures of accuracy in this report may be underestimated by as much as 0.9%.
- The examiner could not make a conclusive determination regarding 112 of the latents (10.1% of the Baseline dataset), however these latents were tested and as was found with the No Value latents, a portion of these inconclusive determination were successfully matched by at least one matcher: 8 were matched at rank-1 and 31 were matched at rank-100 (or higher) by at least one matcher. The inconclusive results by the reviewing examiner did not necessarily coincide with the previous value determinations: of the No Value latents, 3 were individualized during review; of the Value for Exclusion Only latents, 40 were individualized during review; 51 of the latents previously determined to be Of Value resulted in inconclusives, most of which were due to poor-quality exemplars.

Figure 24 compares the ability of the matchers to correctly identify latents with the latent examiners' determinations. The figure shows examiner determinations of No Value or value for exclusion only (Limited Value), combined with inconclusive determinations from the subsequent examiner review, indicating cases in which either the initial or reviewing examiner determined individualizations were not appropriate. Approximately 22% of the latents in the test were missed by all matchers at rank 1, more than half of which could be individualized by a certified latent examiner. The initial or reviewing examiners determined that 14% of the latents in the test were determined by examiners to be of No Value, of value for exclusion only ("Limited"), or resulted in an inconclusive determination; about one third of these could be matched by one or more matchers at rank 1. If the No/Limited Value and inconclusive prints had not been included in the test, 14% of the resulting latents in the test would have been missed by all matchers at rank 1.

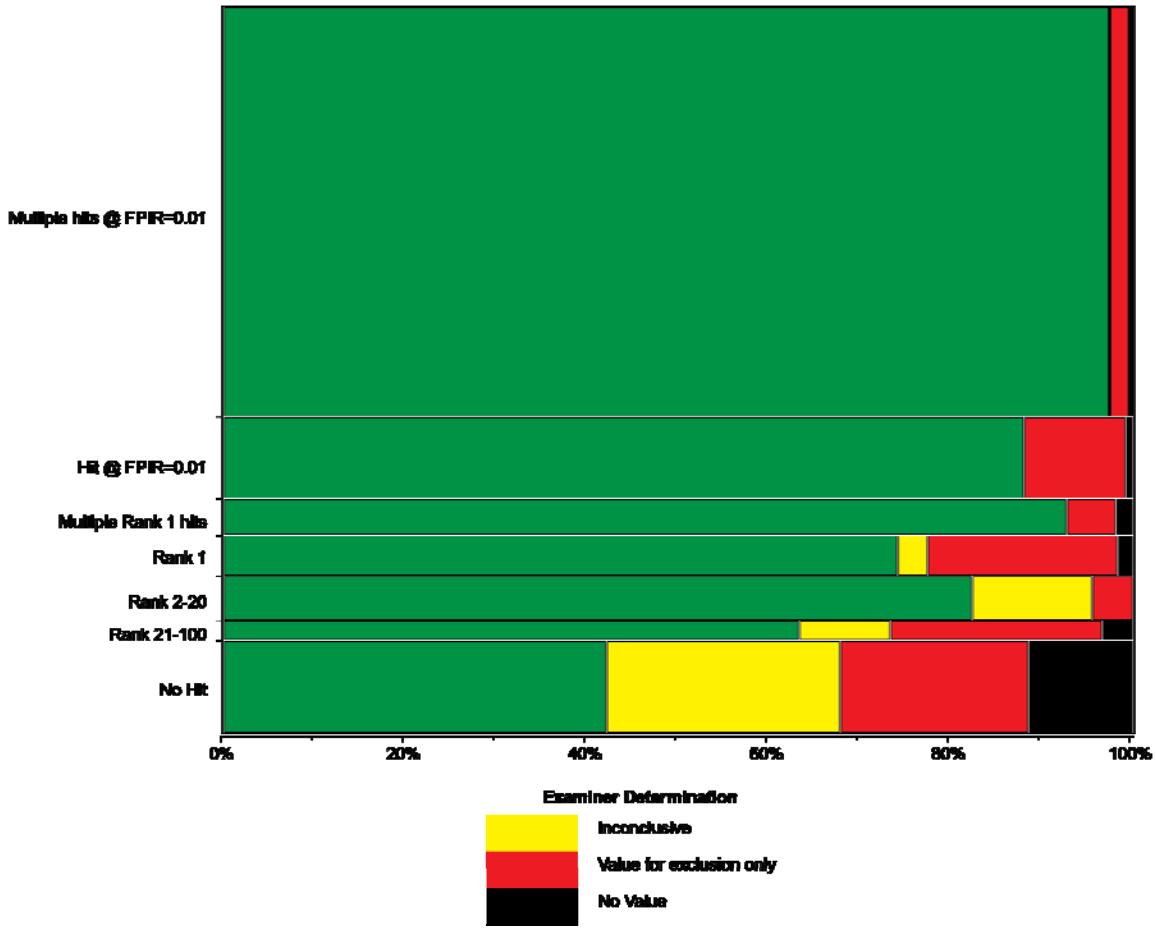


Figure 24: Comparison of matcher hit rates and examiner inconclusive/value determinations, as percentages of Baseline dataset, across all latent subsets[p]

(Baseline dataset with 10 latents without mates excluded, 1104 latents — 100,000 rolled+plain 10-finger exemplar sets)

The identification by the matchers of latents that are No Value or that result in an inconclusive determination precludes the latent examiner's ability to make an individualization. However, it is possible that these algorithmic identifications could yet be of value as an investigative tool.

Three latents with more than 30 minutiae were missed by all matchers (for all feature subsets, ignoring latents with no mates). Of these, two were found to be inconclusive after review by a latent examiner performing a manual comparison of the latent with the exemplar, due to the quality of the exemplar and/or lack of overlap between the latent and exemplar. Thus, only one latent with more than 30 minutiae that was judged to be capable of being individualized was missed by all matchers. This miss analysis underscores the fact that common misses drop nearly to zero for latents with 30+ minutiae.

Table 21 shows the proportion of each latent subset that was missed by all matchers at rank 1. For both Baseline and Baseline-QA, the LG subset had by far the greatest proportion of missed-by-all latents; the other subsets were similar to each other, with LE having the lowest proportion. Table 21 also shows the average minutiae count of the

[p] *Inconclusive cases of no/limited value are indicated by the value determination.*

MATCHER KEY	**A** = *Sagem*	**B**=*NEC*	**C**=*Cogent*	**D**=*Sonda*	**E**=*Warwick*

FEATURE SUBSET KEY	**LA**=*Image*	**LB**=*Image +ROI*	**LC**=*Image +ROI +Quality map +Pattern class*	**LD**=*Image +ROI +Minutiae +Ridge counts*	**LE**=*Image +Full EFS (no Skeleton)*	**LF**=*Image +Full EFS with Skeleton*	**LG**=*Minutiae +Ridge counts (comp to IAFIS)*	**Page 47**

missed-by-all cases: these are relatively consistent, and contrast to the overall average minutiae count for Baseline of 22.5.

Table 21: Proportion of data missed by all matchers at rank 1, with average minutiae
(100,000 rolled+plain 10-finger exemplar sets)

	% of data missed by all	Avg. minutiae
Baseline		
LA	28.4	13.2
LE	24.7	12.4
LG	44.9	13.5
All latent subsets	22.3	12.5
Baseline-QA		
LA	30.8	12.2
LB	31.2	12.1
LC	30.6	12.0
LD	27.5	11.7
LE	26.9	11.6
LF	28.6	11.8
LG	47.8	12.8
All latent subsets	20.7	12.3

The distribution of orientation for the missed-by-all cases was not notably different from the overall distribution.

As expected, the latents missed by all matchers include a greater proportion of low-quality prints (Ugly, Limited, or No Value) than the overall distribution, as shown in Figure 25. However, the latents missed by all are not solely limited to these cases, and do include Good prints (though none assessed as Excellent).

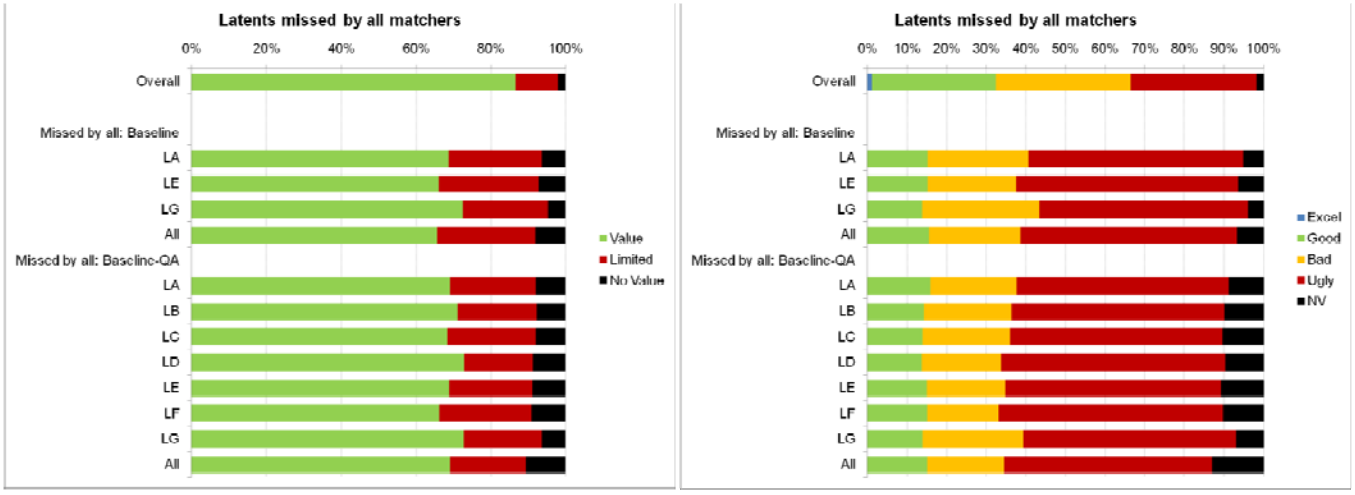


Figure 25: Distribution of value and quality assessments among latents missed by all matchers at rank 1
(100,000 rolled+plain 10-finger exemplar sets)

15.2 Hit Analysis

Table 22 and Table 23 show the collective rank-1 "hits" (identifications) made by any matcher (i.e. the union of all latents hit) broken down by minutiae count with respect to examiner-assessed value and quality. For all latents in the complete Baseline dataset, as well as for each quality category, the collective rank-1 hit rates are broken down into search subsets LA, LE, and LG, as well as "any search subset" (i.e. the union of all search subset hits). Adjacent cells representing the search subsets LA and LE are highlighted in green for cases where LE exceeded LA by more than 5%, and highlighted in yellow for cases where LA exceeded LE.

MATCHER KEY	**A** = *Sagem*	**B**=*NEC*	**C**=*Cogent*	**D**=*Sonda*	**E**=*Warwick*

FEATURE SUBSET KEY	**LA**=*Image*	**LB**=*Image +ROI*	**LC**=*Image +ROI +Quality map +Pattern class*	**LD**=*Image +ROI +Minutiae +Ridge counts*	**LE**=*Image +Full EFS (no Skeleton)*	**LF**=*Image +Full EFS with Skeleton*	**LG**=*Minutiae +Ridge counts (comp to IAFIS)*	**Page 48**

The lowest percentage of "hits" were for latents with low minutiae count and poor image quality. For latents with more than 10 minutiae, minutiae count was the most important factor determining the identification rate, with image quality being secondary. For all of the latent feature subsets, 100% of the latents rated as "Excellent" (all with more than 35 minutiae) were identified by one or more matchers.

For low minutiae counts image quality was a significant factor. Examiner markup had no effect or degraded performance for very poor quality (Limited Value, Ugly, or No Value) very low minutiae count latents (1-5 minutiae). In these cases image-only searches (LA) produced the same or better results.

The greatest gains from adding examiner markup to the search (LD or LE), versus the image alone (LA), were for poor quality (Limited Value and Ugly) latents having 6-30 minutiae.

Table 22: Rank-1 identification rates by any matcher, by latent value and minutiae count

(Baseline dataset, 1114 latents — 100,000 rolled+plain 10-finger exemplar sets)

		All Latents				Value				Limited Value				No Value			
# Min	% data	any subset	LA	LE	LG	any subset	LA	LE	LG	any subset	LA	LE	LG	any subset	LA	LE	LG
1-5	5.7	22.6	21.0	16.1	0.0	0.0	0.0	0.0	0.0	28.9	26.3	18.4	0.0	14.3	14.3	14.3	0.0
6-10	15.6	55.7	45.4	51.1	14.9	58.5	50.9	52.8	18.9	49.2	34.9	46.0	7.9	50.0	50.0	50.0	0.0
11-15	19.7	75.0	64.1	70.0	40.5	75.4	64.5	70.9	41.9	76.9	61.6	61.5	23.1	50.0	50.0	50.0	50.0
16-20	15.8	79.5	76.7	79.0	53.4	80.2	77.8	79.6	55.1	62.5	50.0	62.5	12.5	-	-	-	-
21-25	12.4	89.9	80.4	88.4	69.6	89.7	80.1	88.2	69.9	-	-	-	-	-	-	-	-
26-30	9.3	88.5	85.6	87.5	79.8	88.1	85.1	87.1	79.2	-	-	-	-	-	-	-	-
31-35	5.5	98.4	95.1	98.4	88.5	98.4	95.1	98.4	88.5	-	-	-	-	-	-	-	-
36-40	4.3	93.8	93.8	93.8	93.8	93.8	93.8	93.8	93.8	-	-	-	-	-	-	-	-
41-45	3.3	100.0	95.6	100.0	97.3	100.0	95.6	100.0	97.3	-	-	-	-	-	-	-	-
46-106	8.4	97.9	97.9	97.9	96.8	97.9	97.9	97.9	96.8	-	-	-	-	-	-	-	-

Table 23: Rank-1 identification rates by any matcher, by latent quality and minutiae count

(Baseline dataset, 1114 latents — 100,000 rolled+plain 10-finger exemplar sets)

	Excellent				Good				Bad				Ugly				No Value			
# Min	any subset	LA	LE	LG	any subset	LA	LE	LG	any subset	LA	LE	LG	any subset	LA	LE	LG	any subset	LA	LE	LG
1-5	-	-	-	-	50.0	50.0	50.0	0.0	33.3	16.7	33.3	0.0	23.7	23.7	15.8	0.0	14.3	14.3	7.1	0.0
6-10	-	-	-	-	60.0	46.7	53.3	33.3	69.2	57.7	65.4	19.2	50.0	40.0	45.0	11.0	25.0	25.0	25.0	0.0
11-15	-	-	-	-	78.6	78.6	75.0	53.6	79.5	66.7	76.9	51.3	72.9	59.8	66.4	30.8	0.0	-	0.0	0.0
16-20	-	-	-	-	73.2	70.7	73.2	58.5	89.2	87.8	89.2	62.2	71.2	66.1	69.5	37.3	-	-	-	-
21-25	-	-	-	-	93.2	84.1	93.2	79.5	89.1	85.9	89.1	68.8	87.0	56.5	78.3	47.8	-	-	-	-
26-30	-	-	-	-	90.2	88.2	88.2	84.3	88.6	85.7	88.6	80.0	86.7	80.0	86.7	66.7	-	-	-	-
31-35	-	-	-	-	96.7	96.7	96.7	93.3	100.0	96.0	100.0	80.0	100.0	100.0	100.0	50.0	-	-	-	-
36-40	100.0	100.0	100.0	100.0	93.1	93.1	93.1	93.1	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	-	-	-	-
41-45	100.0	100.0	100.0	100.0	100.0	96.4	100.0	96.4	100.0	87.5	100.0	100.0	-	-	-	-	-	-	-	-
46-106	100.0	100.0	100.0	100.0	97.2	97.2	97.2	95.8	100.0	100.0	100.0	100.0	-	-	-	-	-	-	-	-

15.3 Loss/Gain Analysis

When comparing searches based on different latent subsets, a net improvement in accuracy does not necessarily mean that each separate search increased the likelihood of an identification. Figure 26 shows, for successive search runs of the same latents: the number of latents ***not*** identified at rank-1 ("losses"); the number of latents correctly identified at rank-1 ("gains"), and the sum of the two (i.e. "net change" in the number of rank-1 identifications) when additional information (image or feature data) is used for one of the search runs being compared.

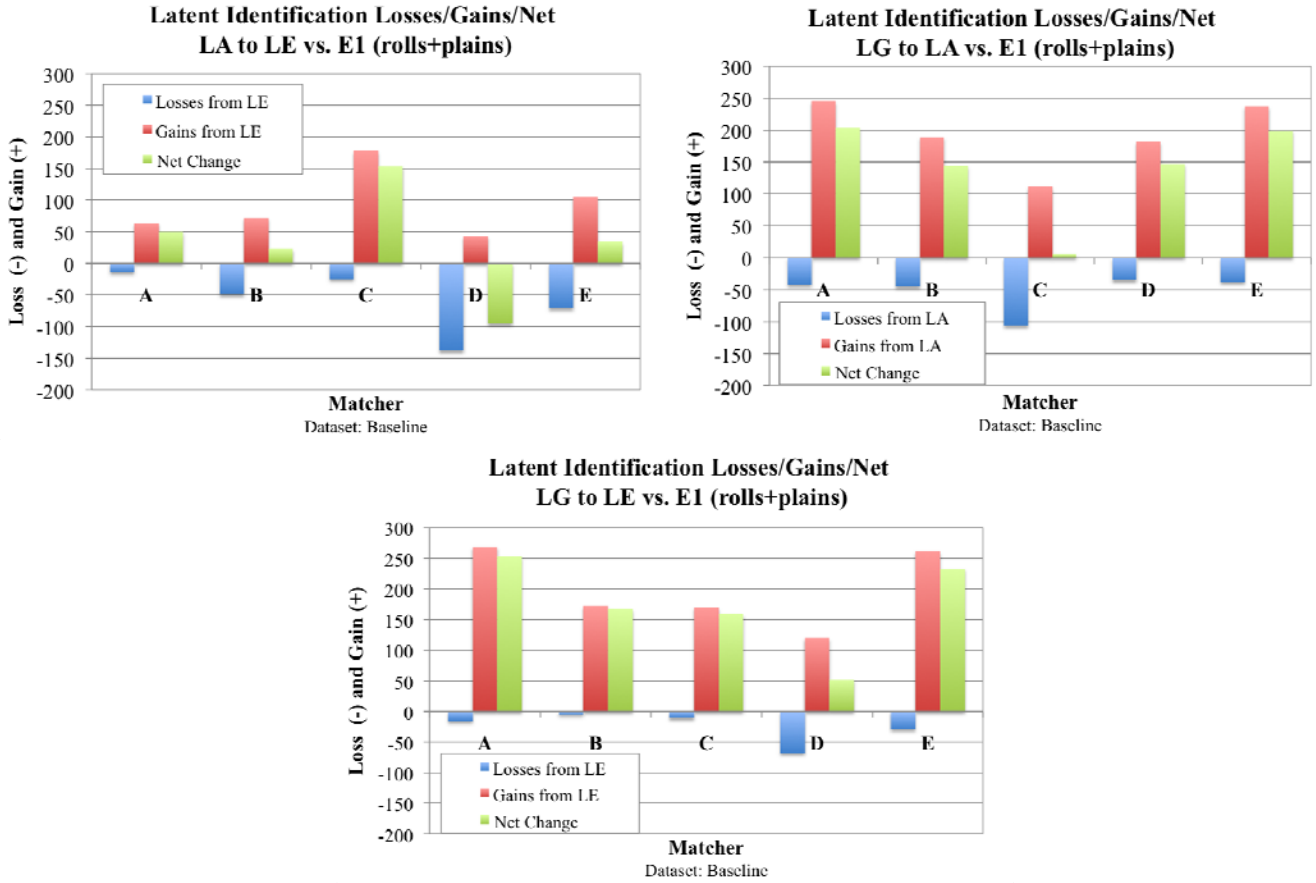


Figure 26: Counts of rank-1 latent searches gained and lost for specified subsets
(Baseline dataset, 1114 latents — 100,000 rolled+plain 10-finger exemplar sets)

In most cases additional image and/or feature data resulted in a net gain in the number of latents identified which is reflected in the identification rate increases discussed in previous sections. However, it is notable that in almost every case there are some latents which were ***only*** identified by supplying less data, rather than more, to the matcher. This unexpected behavior for a portion of the data warrants further investigation.

Little to no commonality exists amongst the losses for differing feature subsets between matchers.

The participants may consider improvements to their algorithms that address the variations in performance levels due to different encoding possibilities.

Even though LA typically far outperforms LG (except for matcher C), there are still a number of latents which are matched with LG but not with LA. LG has the least total information, however, when compared with other datasets still produces some hits that are lost with additional information.

MATCHER KEY	**A** = *Sagem*	**B**=*NEC*	**C**=*Cogent*	**D**=*Sonda*	**E**=*Warwick*

FEATURE SUBSET KEY	**LA**=*Image*	**LB**=*Image +ROI*	**LC**=*Image +ROI +Quality map +Pattern class*	**LD**=*Image +ROI +Minutiae +Ridge counts*	**LE**=*Image +Full EFS (no Skeleton)*	**LF**=*Image +Full EFS with Skeleton*	**LG**=*Minutiae +Ridge counts (comp to IAFIS)*	**Page 51**

If this behavior persists, it may be worthwhile for latents of high importance to be submitted as an image-only and image plus feature set.

16 Timing results

NIST attempted to check adherence to the timing requirements during validation on a small set of images having differing characteristics from the actual test data. In many cases, however, it was not possible for NIST to extrapolate from the validation set to the actual test data sets due to data dependent behavior (e.g. nonlinearity of latent search times with respect to gallery size). Given the complexity and time required to calibrate the SDKs to the test data with respect to the timing requirements, the processing time requirements were not enforced unless they precluded completion of the test within the schedule time frame.

The target thresholds are indicated below:

- Exemplar enrollment: 100 seconds/10-finger exemplar set
- Latent feature extraction: 120 seconds/latent
- Search (LA,LB): 10,000 seconds (0.1 seconds * 100,000 exemplar sets)
- Search (LC-LG): 5,000 seconds (0.05 seconds * 100,000 exemplar sets)

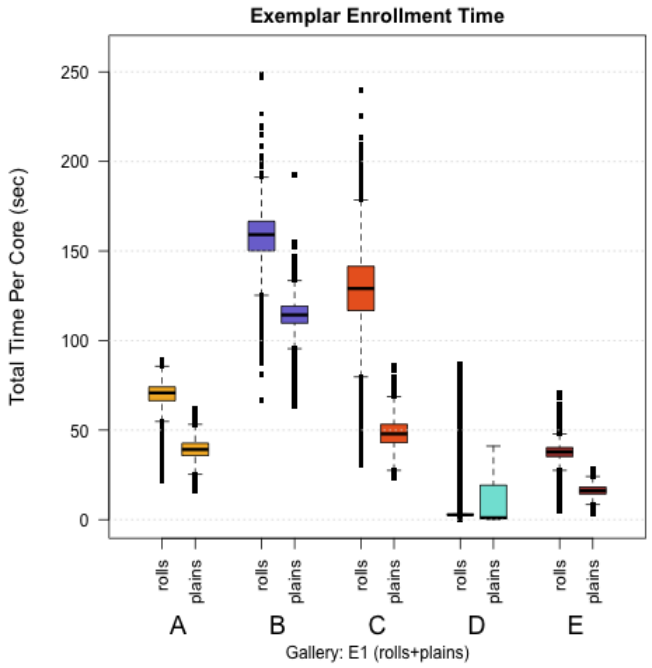


Figure 27: Exemplar enrollment time per 10-finger exemplar set (threshold=100 seconds)

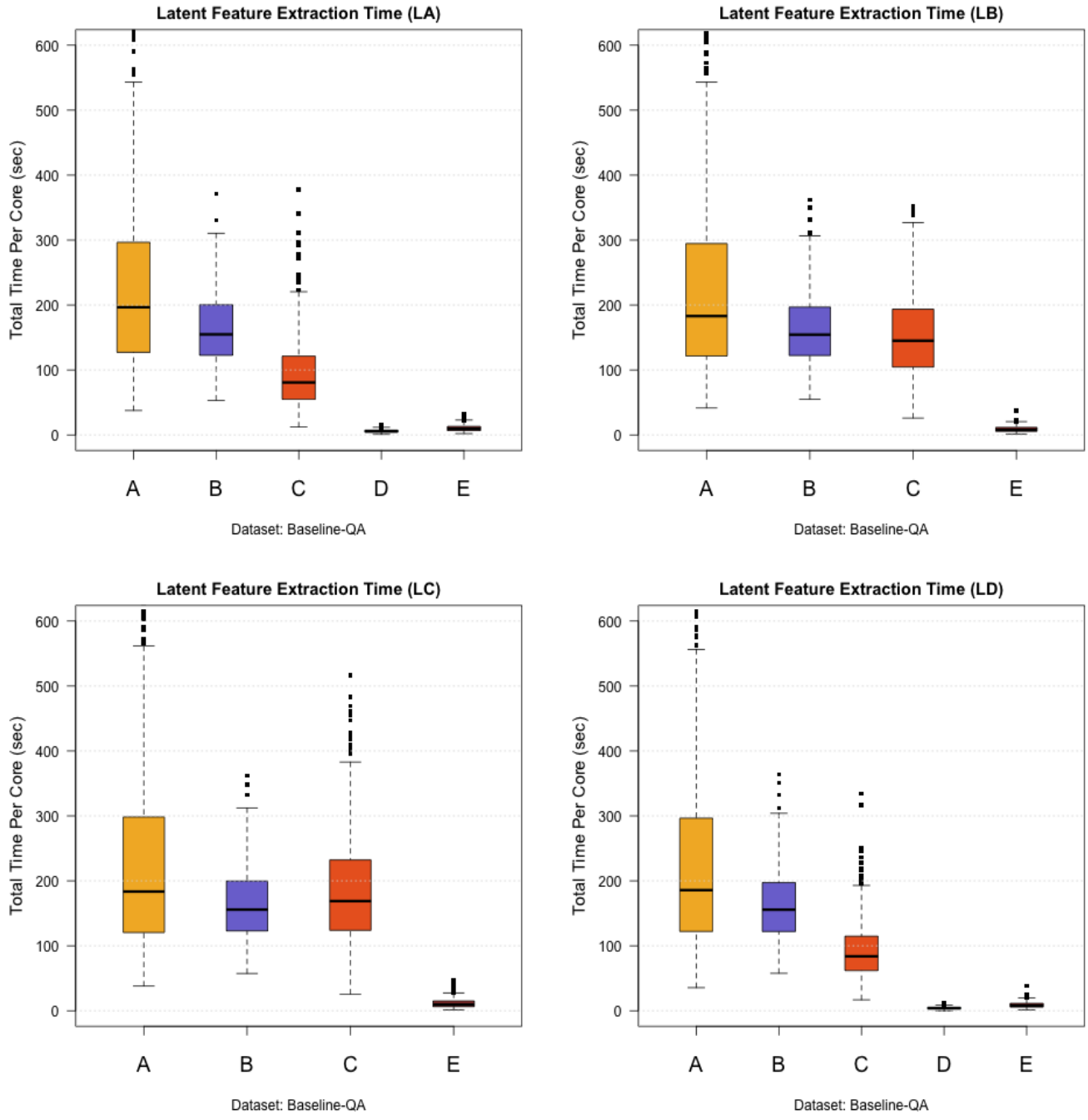


Figure 28: Feature extraction time for latent subsets LA-LD (threshold =120 seconds)

MATCHER KEY	**A** = *Sagem*	**B**=*NEC*	**C**=*Cogent*	**D**=*Sonda*	**E**=*Warwick*

FEATURE SUBSET KEY	**LA**=*Image*	**LB**=*Image* *+ROI*	**LC**=*Image +ROI* *+Quality map* *+Pattern class*	**LD**=*Image +ROI* *+Minutiae* *+Ridge counts*	**LE**=*Image* *+Full EFS* *(no Skeleton)*	**LF**=*Image* *+Full EFS* *with Skeleton*	**LG**=*Minutiae* *+Ridge counts* *(comp to IAFIS)*	**Page** **53**

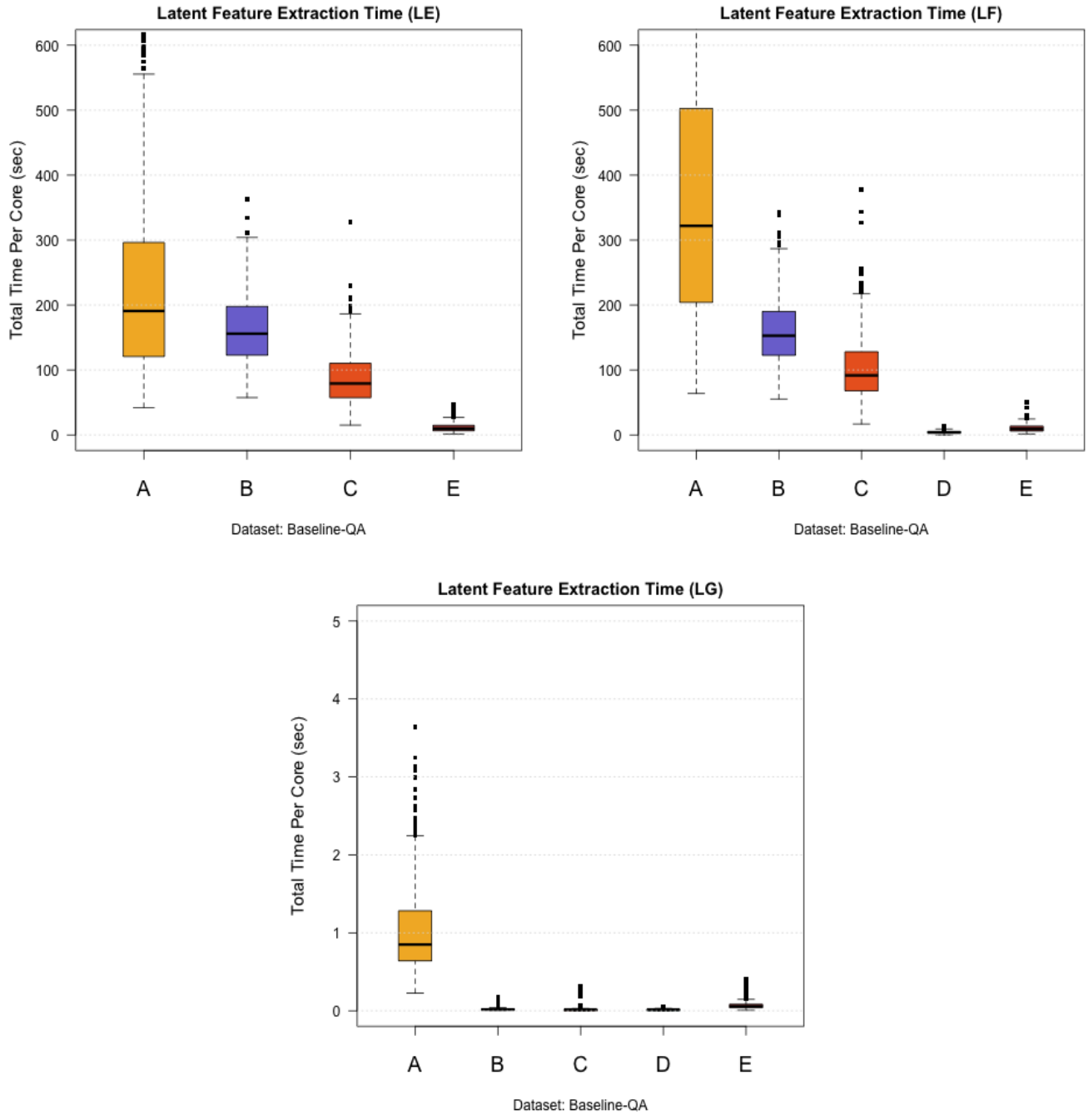


Figure 29: Feature extraction time for latent subsets LE-LG (threshold =120 seconds). Note the different scale for subset LG.

MATCHER KEY	**A** = *Sagem*	**B**=*NEC*	**C**=*Cogent*	**D**=*Sonda*	**E**=*Warwick*

FEATURE SUBSET KEY	**LA**=*Image*	**LB**=*Image +ROI*	**LC**=*Image +ROI +Quality map +Pattern class*	**LD**=*Image +ROI +Minutiae +Ridge counts*	**LE**=*Image +Full EFS (no Skeleton)*	**LF**=*Image +Full EFS with Skeleton*	**LG**=*Minutiae +Ridge counts (comp to IAFIS)*	**Page 54**

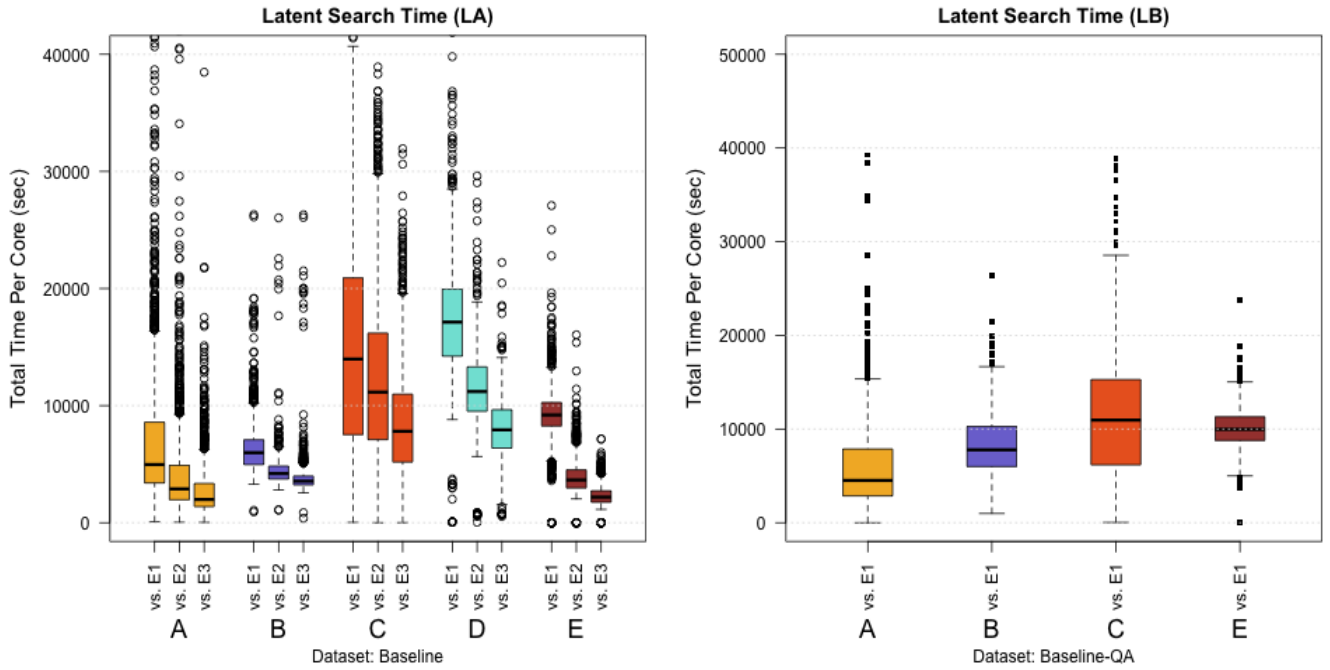


Figure 30: Latent Search time for latent subsets LA-LB (threshold=10,000 seconds)

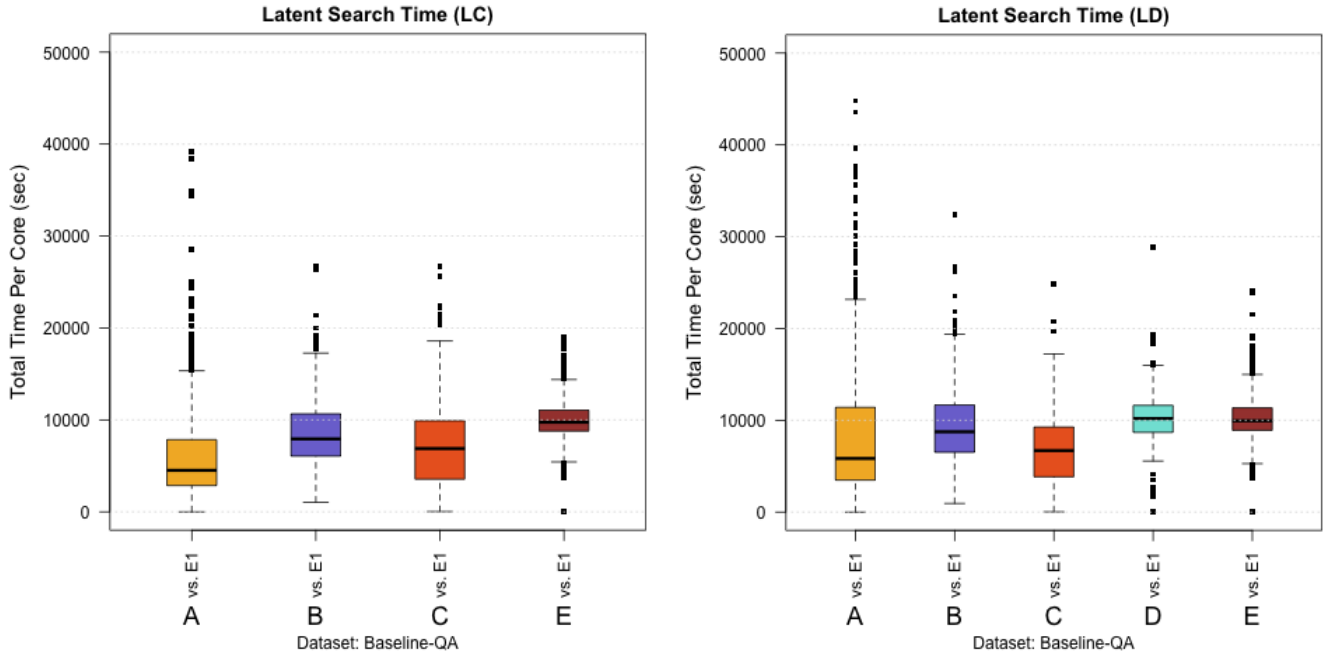


Figure 31: Latent Search time for latent subsets LC-LD (threshold=5,000 seconds)

MATCHER KEY	**A** = *Sagem*	**B**=*NEC*	**C**=*Cogent*	**D**=*Sonda*	**E**=*Warwick*

FEATURE SUBSET KEY	**LA**=*Image*	**LB**=*Image +ROI*	**LC**=*Image +ROI +Quality map +Pattern class*	**LD**=*Image +ROI +Minutiae +Ridge counts*	**LE**=*Image +Full EFS (no Skeleton)*	**LF**=*Image +Full EFS with Skeleton*	**LG**=*Minutiae +Ridge counts (comp to IAFIS)*	**Page 55**

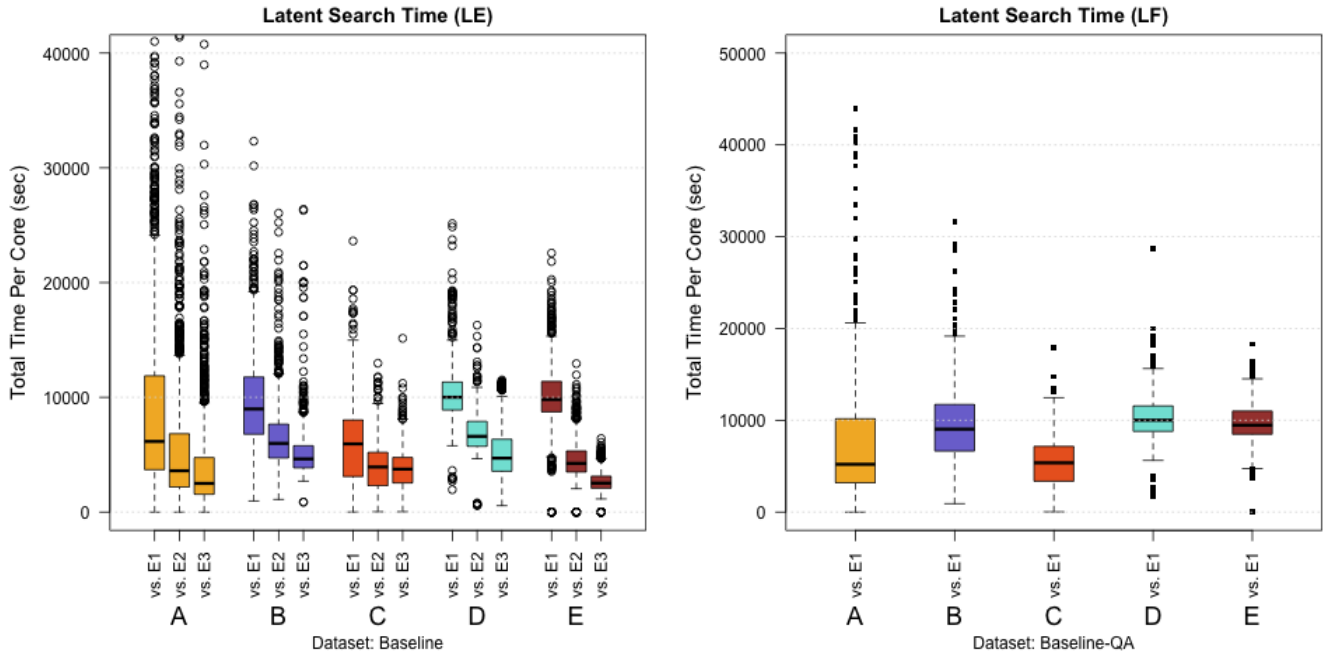


Figure 32: Latent Search time for latent subsets LE-LF (threshold=5,000 seconds)

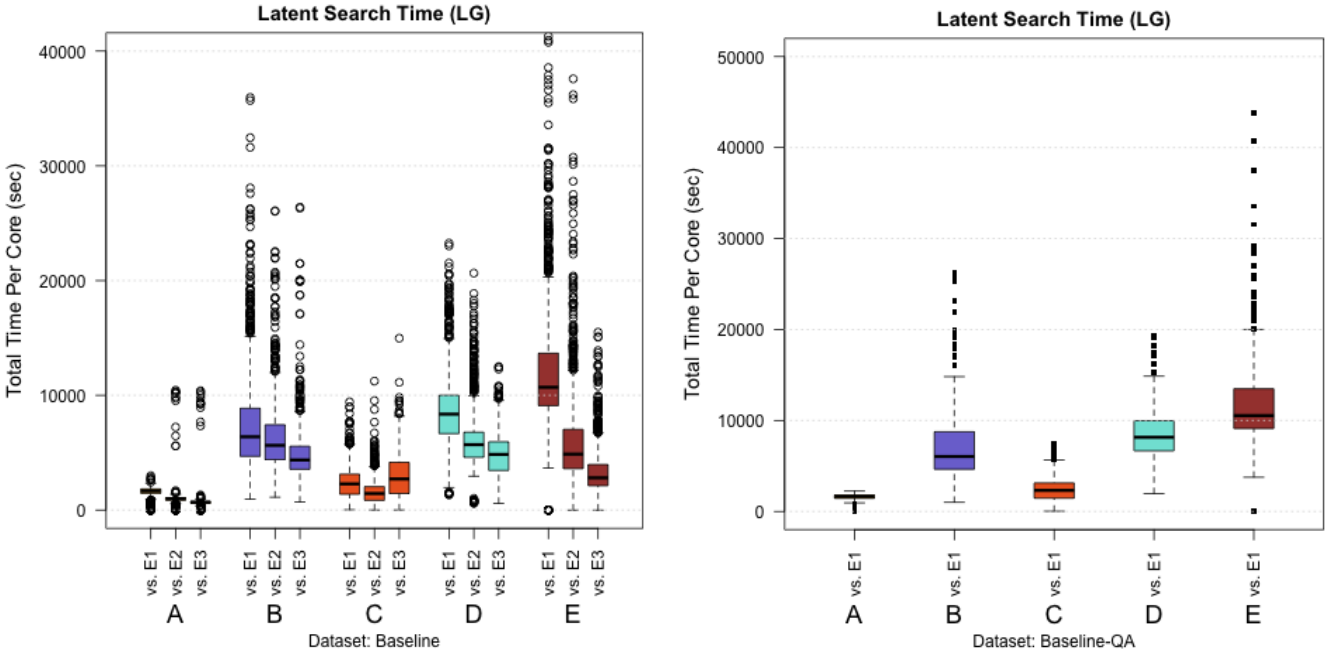


Figure 33: Latent Search time for latent subset LG (threshold=5,000 seconds)

17 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 [8] in April 2009. As shown in Table 24, 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.

MATCHER KEY	**A** = *Sagem*	**B**=*NEC*	**C**=*Cogent*	**D**=*Sonda*	**E**=*Warwick*

FEATURE SUBSET KEY	**LA**=*Image*	**LB**=*Image +ROI*	**LC**=*Image +ROI +Quality map +Pattern class*	**LD**=*Image +ROI +Minutiae +Ridge counts*	**LE**=*Image +Full EFS (no Skeleton)*	**LF**=*Image +Full EFS with Skeleton*	**LG**=*Minutiae +Ridge counts (comp to IAFIS)*	**Page 57**

Table 24: Comparison of ELFT-EFS and ELFT Phase II rank-1 results

ELFT-EFS	
LA - Image-only	
Rank 1 against 1M fingerprints	
Sagem (A)	62.2
NEC (B)	61.2
Cogent (C)	48.3
Warwick (E)	47.2
Sonda (D)	25.1

ELFT Phase II	
Rank 1 against 100K fingerprints	
NEC (M1)	97.2
Cogent (P1)	87.8
SPEX (O1)	80.0
Motorola (K1)	79.3
L1 Identity Solutions (Q1)	78.8
Peoplespot (N1)	67.9
Sonda (L1)	28.5
BioMG (R1)	27.5

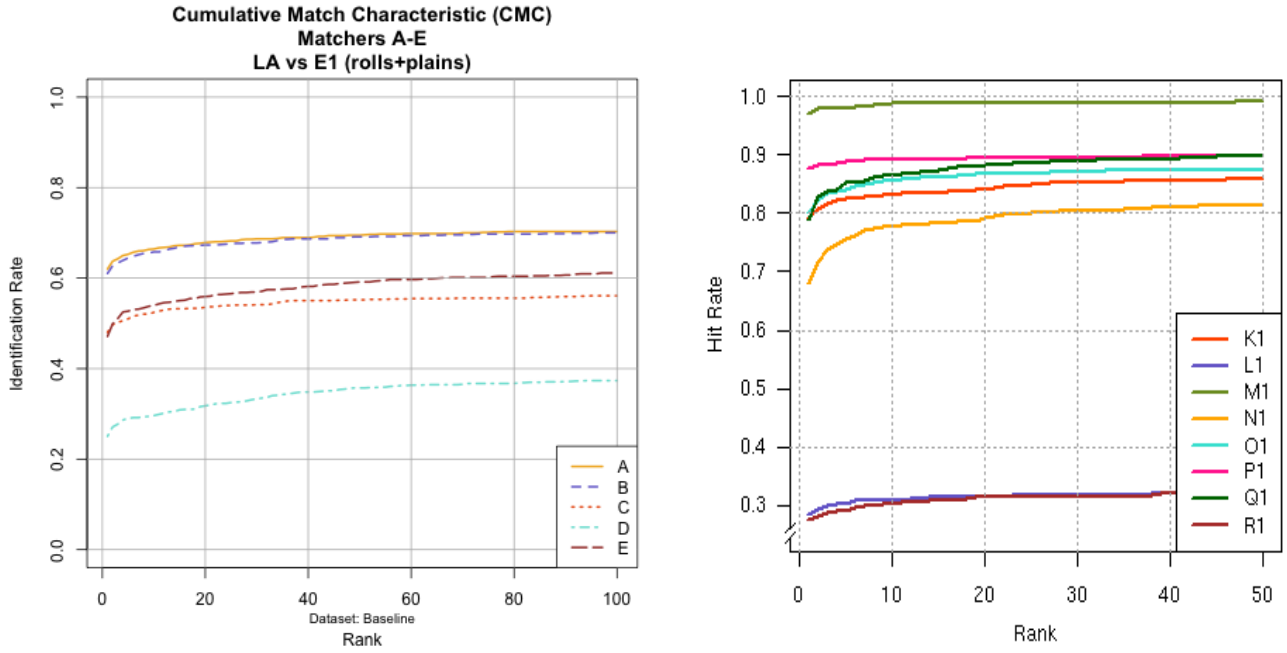


Figure 34: Comparison of ELFT-EFS and ELFT Phase II CMC results. Key to participants is included in Table 24. Note differences in Y axes.

MATCHER KEY	**A** = *Sagem*	**B**=*NEC*	**C**=*Cogent*	**D**=*Sonda*	**E**=*Warwick*

FEATURE SUBSET KEY	**LA**=*Image*	**LB**=*Image* *+ROI*	**LC**=*Image +ROI* *+Quality map* *+Pattern class*	**LD**=*Image +ROI* *+Minutiae* *+Ridge counts*	**LE**=*Image* *+Full EFS* *(no Skeleton)*	**LF**=*Image* *+Full EFS* *with Skeleton*	**LG**=*Minutiae* *+Ridge counts* *(comp to IAFIS)*	**Page 58**

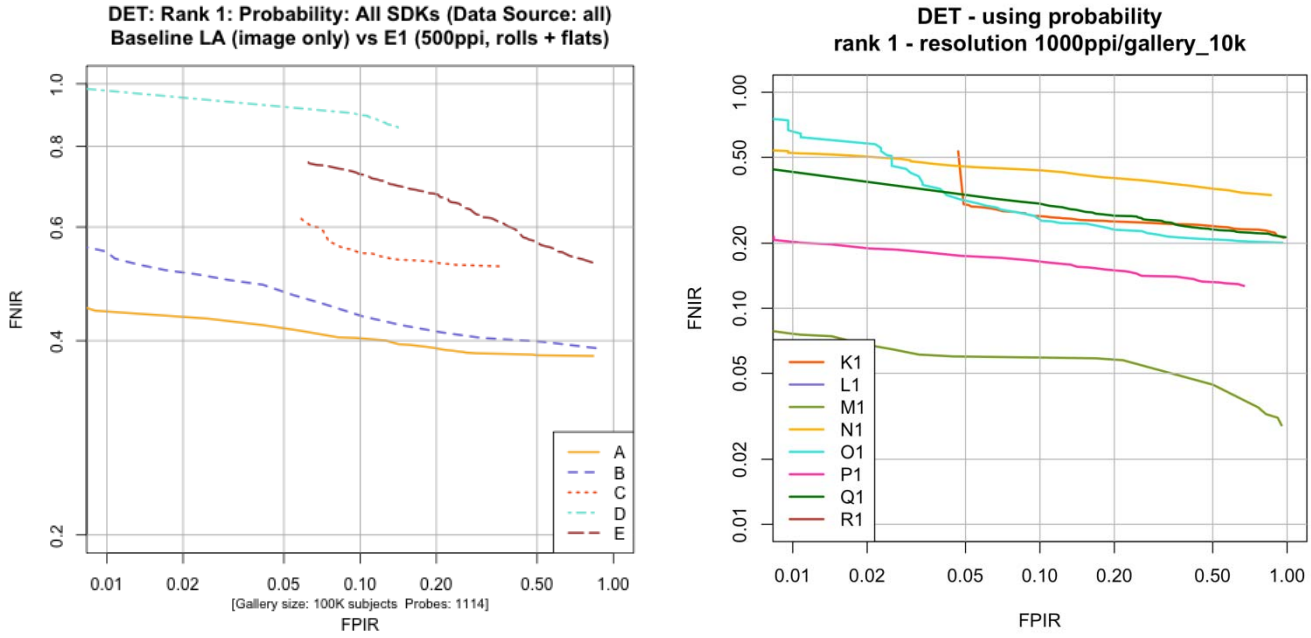


Figure 35: Comparison of ELFT-EFS and ELFT Phase II DET results.[9] . Key to participants is included in Table 24. Note differences in Y axes.

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:

- 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.
- 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.

18 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.

[9] *The DETs show the same data as an ROC, but with the Y axis inverted (FNIR = 1-TPIR) and shown in log scale.*

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.

19 Results and Conclusions

1. The highest accuracy for all participants was observed for searches that included examiner-marked features in addition to the latent images.
2. Image-only searches were more accurate than feature-only searches for most participants.
3. Since score-based results are more scalable than rank-based results, they provide a better indication of how accuracy would be affected by an increase in database size. This capability could provide operational benefits such as reduced or variable size candidate lists, or for reverse latent searches (searches of databases containing unsolved latents) where using a score threshold is used to limit candidate list size. Comparing rank-1 and score-based results at different values of False Positive Identification Rate (FPIR) showed that image-only and image + full EFS searches are less sensitive to differences in FPIR, whereas minutiae-only searches are most sensitive.
4. The effect of the use of EFS features other than minutiae was mixed. In most cases, additional features resulted in accuracy improvement; cases where accuracy declined may be indicative of implementation issues. When using score-based results, the highest overall accuracy is achieved through the use of full EFS with skeletons.
5. Ground truth (GT) markup yielded an increase in performance over the original examiner markup of about 5 to 8 percentage points for image + full EFS searches, and about 11 to 15 percentage points for minutiae-only searches. This shows that matcher accuracy is highly affected by the precision of latent examiner markup.
6. Latent orientation has an impact on matcher accuracy. When the orientation of latents was unknown, the rank-1 identification rates were 17.8 to 18 percentage points lower than the general case.
7. Matcher accuracy is very clearly related to the examiners' latent print value determinations, with much greater accuracy for latents determined a priori to be of value. The matching algorithms demonstrated an unexpected ability to identify low feature content latents: Sagem's rank-1 accuracy for No Value latents was 20% on image-only searches, and 26.2% on Limited Value latents.
8. The performance of all matchers decreased consistently as lower quality latents were searched, with respect to the informal scale of "Excellent", "Good", "Bad", or "Ugly".
9. Analysis showed that the greatest percentage of the misses were for latents with low minutiae count, and those assessed by examiners as poor quality ("Ugly") or "No Value". Algorithm performance for all participants was highly correlated to the number of minutiae. For latents with more than 10 minutiae, minutiae count was the most important factor for successful identification with examiner-assessed quality being secondary. For latents with fewer than 10 minutiae, examiner-assessed quality was a better predictor of match accuracy than minutiae count.
10. Approximately 22% of the latents in the test were missed by all matchers at rank 1, more than half of which could be individualized by a certified latent examiner. The initial or reviewing examiners determined that

14% of the latents in the test were of no value, of value for exclusion only, or resulted in an inconclusive determination; about one third of these could be matched by one or more matchers at rank 1.

11. The highest measured accuracy achieved by any matcher at rank-1 on any latent feature subset was 66.7%, even though approximately 78% of the latents in the test were matched by one or more matchers at rank-1. This indicates a potential for additional accuracy improvement through improved algorithms, or through the use of data based fusion (e.g. search using image-only and again search using image+features). The differences in which latents were identified by the various matchers also points to a potential accuracy improvement by using algorithm fusion.
12. The use of both rolled and plain impressions in the gallery resulted in higher accuracy than the use of either rolled or plain impressions separately for most matchers. Use of plain impressions in the gallery as compared to rolled impressions resulted in a drop in accuracy for most matchers.
13. The results obtained during ELFT Phase II (NISTIR 7577 [8], April 2009) show substantially higher accuracy than ELFT-EFS results for image-only matching for three of the participants who participated in both tests. This is an expected result because the ELFT-EFS test data included a greater proportion of poor-quality latents, had higher throughput requirements, and larger gallery size.

References

1 "Data Format for the Interchange of Extended Friction Ridge Features"; Proposed Addendum/Revision to *ANSI/NIST-ITL Data Format for the Interchange of Fingerprint, Facial, & Other Biometric Information*; Draft Version 0.4; 12 June 2009. (http://fingerprint.nist.gov/standard/cdeffs) [Note: Relatively minor revisions to this specification have been made since the version used in the test, but these did not affect the fields used in ELFT-EFS. The specification is in the process of being incorporated into ANSI/NIST-ITL 1-2011 (forthcoming)]

2 Committee to Define an Extended Fingerprint Feature Set (CDEFFS) (http://fingerprint.nist.gov/standard/cdeffs)

3 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)

4 *ANSI/NIST-ITL 1-2007 Data Format for the Interchange of Fingerprint, Facial, & Other Biometric Information*; (http://www.nist.gov/itl/iad/ig/ansi_standard.cfm) [ANSI/NIST-ITL 1-2011 (forthcoming)]

5 Federal Bureau of Investigation Criminal Justice Information Services; Electronic Biometric Transmission Specification (EBTS) (https://www.fbibiospecs.org/ebts.html)

6 Universal Latent Workstation (ULW) software (https://www.fbibiospecs.org/Latent/learnAboutULW.aspx)

7 Wavelet Scalar Quantization (WSQ) Gray-Scale Fingerprint Image Compression Specification (https://www.fbibiospecs.org/docs/WSQ_Gray-scale_Specification_Version_3_1.pdf)

8 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

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

Test Plan

Contents

A-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[2] (pixels per inch) images provided with human markup of EFS features.
- Exemplars for the gallery will be images only. Exemplars will be 500ppi[3] (pixels per inch).
- 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.

A-2 Participation

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

---

[1] CDEFFS is the ANSI/NIST Committee to Define an Extended Fingerprint Feature Set. The current working draft of the Extended Fingerprint and Palmprint Features document can be found at http://fingerprint.nist.gov/standard/cdeffs/.

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

[3] 500 pixels per inch (ppi) is equivalent to 19.69 pixels per millimeter (ppmm)

All systems must comply with the API outlined in Section A-5.4. Anonymous participation will not be permitted. The Application form[4] includes details regarding application and qualification.

A-3 Data

A-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 Validation 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.

A-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. 500 pixels per inch (19.69 pixels per millimeter) exemplar images will be compressed using WSQ.

A-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 two primary 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 A-7.

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

[4] *The Application form can be found at http://fingerprint.nist.gov/latent/elft-efs/*

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.

A-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).

A-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 from latent subtests LB-LG will include the orientation direction and uncertainty fields (9.301). Images from latent subtest LA will not.

A-3.6 Exemplar types

All exemplars will include rolled 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, and are therefore referred to here as a 10-finger exemplar set (also commonly called a ten print set).

Note that a 10-finger exemplar set will consist of either ten rolled prints, or ten plain prints.

In some cases, multiple sets of 10-finger exemplar sets 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.

A-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.

A-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.

A-4 Evaluation Criteria

A-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[5] 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 100, as 100 is the (maximum) candidate list size specified in the API.

- Matcher score metrics are evaluated in terms of DET/ROC[6] 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 A-5.6) will be used for DET analysis.

A-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 A-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
  - LF – image + baseline EFS + Skeleton
  - LG – IAFIS/EFTS equivalent features only
- Exemplar Subtests

[5] *Cumulative Match Characteristic*

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

- E1 – 100,000 subjects; 1 set of 10 rolled and 1 set of 10 plain impressions each; 500ppi
- E2 – 10,000 subjects; 1 set of 10 rolled impressions each; 500ppi
- E3 – 10,000 subjects; 1 set of 10 plain impressions each; 500ppi
- E4 – 10,000 subjects; 2 sets of 10 rolled impressions each; 500ppi
- E5 – 10,000 subjects; 3 sets of 10 rolled impressions each; 500ppi
- E6 – 10,000 subjects; 4 sets of 10 rolled impressions each; 500ppi

A-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.

A-5 Latent Matching Software

A-5.1 Overview

Participants shall submit a set of SDKs (Software Development Kits) that provide the interfaces defined by the ELFT-EFS 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 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.

A-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 Enterprise Linux Server 5.1 (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.

A-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 to submit multithreaded or 64-bit SDKs, we support other modes of operation as outlined below.

A-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).

A-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.

A-5.4 API

A-5.4.1 Test Interface Description

Participants shall submit an SDK which provides the interfaces defined in section 5.4.4. Section 5.4.3 defines the interfaces to functions provided by NIST for use by the SDK. Sections 5.4.2 and 5.4.5 specify the declaration of constants, error codes, data-types and functions used by both.

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 (SDKs) and NIST supplied (image extraction library and test driver).

A-5.4.2 Declarations

The following are declarations of data types and functions used in the Latent Fingerprint SDK testing interface:

```
////////////////////////////////////////////////////
// Declarations of constants                      //
////////////////////////////////////////////////////

// Impression type codes
#define IMPTYPE_LP      0     // Live-scan plain
#define IMPTYPE_LR      1     // Live-scan rolled
#define IMPTYPE_NP      2     // Nonlive-scan plain
#define IMPTYPE_NR      3     // Nonlive-scan rolled

// Finger position codes
#define FINGPOS_UK      0     // Unknown finger
#define FINGPOS_RT      1     // Right thumb
#define FINGPOS_RI      2     // Right index finger
#define FINGPOS_RM      3     // Right middle finger
#define FINGPOS_RR      4     // Right ring finger
#define FINGPOS_RL      5     // Right little finger
#define FINGPOS_LT      6     // Left thumb
#define FINGPOS_LI      7     // Left index finger
#define FINGPOS_LM      8     // Left middle finger
#define FINGPOS_LR      9     // Left ring finger
#define FINGPOS_LL      10    // Left little finger


////////////////////////////////////////////////////////////
// Declarations for the NIST provided library functions   //
////////////////////////////////////////////////////////////

// Structure to hold a single fingerprint record (image+metadata)
struct finger_record
{
      BYTE    impression_type;
      UINT16  resolution;     // Image resolution in pixels/cm
      BYTE    finger_position;
      UINT16  height;         // Image height in pixels
      UINT16  width;          // Image width in pixels
      BYTE    *image_data;    // 8-bit grayscale image data
};
typedef struct finger_record   FINGER_REC;


// Extracts 10 fingerprint records from a ten-print (AN2K) file
INT32 extract_image_data(     const    char    *tenprint_filename,
                              FINGER_REC **finger_recs);
```

```
// De-allocates the memory holding 10 fingerprint records
void free_image_data(FINGER_REC *finger_recs);

////////////////////////////////////////////////////////////
// Declarations for the SDK provided library functions    //
////////////////////////////////////////////////////////////

// Extracts features from exemplar
INT32 extract_exemplar( const char *exemplarFilename,
                        const char *outputDir);

// Creates a gallery from set of extracted exemplar features
INT32 create_gallery(   const INT32 numExemplars,
                        const char **exemplarFeatFilenames,
                        const char *galleryDir);

// Selects the current gallery for latent searching
INT32 set_gallery(      const char *galleryDir);

// Extracts features from latent file
INT32 extract_latent(   const char *latentFilename,
                        const char *outputDir);

// Searches for the latent in the gallery
INT32 latent_search(    const char *latentFeatFilename,
                        const char *outputDir);
```

A-5.4.3 NIST Provided Functions

A-5.4.3a Extract Image Data

INT32

```
extract_image_data(const char  *tenprint_filename,
                   FINGER_REC  **finger_recs);
```

Description

This function extracts ten fingerprint image records from a single (AN2K formatted) ten-print record file. The caller shall pass ***tenprint_filename*** as a

pointer to the fully qualified pathname of an AN2K formatted ten-print record file, and ***finger_recs*** as the address of a pointer of type `FINGER_REC` (see 5.4.2 above).

Upon return ***finger_recs*** will contain a pointer to an array of ten `FINGER_REC` structures ordered by finger position from 1 (right thumb) to 10 (left little finger). For any fingers that are missing from the original ten-print record file, the *image_data* field in the respective `FINGER_REC` will be a NULL pointer.

Example

```
// Example of processing a ten-print record

FINGER_REC *finger_recs;

INT32 status=extract_image_data("E000123.an2", &finger_recs);

if (status == 0) {
        for (i=0;i<10;i++) {
                if (finger_recs[i].image_data != NULL)
                        process_valid_finger(finger_recs[i]);
                else
                        process_missing_finger(finger_recs[i]);
        }
        free_image_data(finger_recs); // see 5.4.3b below
}
```

Parameters

tenprint_filename (input): A pointer to a ten-print record filename.

finger_recs (output): The address of a FINGER_REC pointer.

Return Value

This function returns *zero* on success or a documented *non-zero* error code otherwise.

A-5.4.3b Free Image Data

```
void
free_image_data(FINGER_REC *finger_recs);
```

Description

De-allocates all memory used by the array of `FINGER_REC` structures specified by ***finger_recs*** which was allocated during a call to `extract_image_data()`.

Parameters

finger_recs (input): A pointer to an array of FINGER_REC structures.

Return Value

None.

A-5.4.4 SDK Provided Functions

A-5.4.4a Exemplar Feature Extraction

```
INT32
extract_exemplar( const char  *exemplarFilename,
                  const char   *outputDir);
```

Description

This function produces a single proprietary formatted feature set file from a 10-print exemplar set. The output from multiple calls to this function (i.e. multiple proprietary feature set files) will be used to construct a gallery (see section 5.4.4b) that is searchable by `latent_search()`.

The 10-print exemplar set will be contained in an ANSI/NIST file with pathname specified by ***exemplarFilename*** (e.g. "/mnt1/input/E1/E199999_1.an2"), and that file will contain either 10 rolled or 10 segmented slap fingerprint images. The directory to which the proprietary feature set file shall be written is specified by the pathname pointed to by ***outputDir*** (e.g. "/mnt/output/feats/E1/").

The format of all pathnames will be canonical Unix style pathnames using forward slash directory separators. The maximum total pathname length is 255 characters.

A single proprietary feature set file shall be written to the directory specified by ***outputDir***. No files other than the feature set file may be written. The filename of the output feature set file is defined here as the base filename of ***exemplarFilename*** with the extension ".an2" replaced by ".feat" (no quotes). For example, if ***exemplarFilename*** ="/mnt1/input/E1/E199999_1.an2" and ***outputDir*** = "/mnt/output/feats/E1/", the proprietary feature set file shall be written as "/mnt/output/feats/E1/E199999_1.feat".

No format is prescribed for the output feature data. For example if desired it may contain images from the 10-print exemplar set. A feature file shall always be output, regardless of any internal failures such as a failure of automated feature extraction. The contents of the directory pointed to by ***outputDir*** (structure and other contents) are not relevant. Pre-computation of feature data avoids reprocessing of the original images upon subsequent calls to `latent_search()`.

The SDK shall use the function `extract_image_data()` (see 5.4.3a above) provided by NIST to extract the raw grayscale image and metadata from the 10-print exemplar set file specified by ***exemplarFilename***. Note that each call to `extract_image_data()` allocates memory to hold the extracted image and metadata, so this memory should be de-allocated using the NIST provided `free_image_data()` (see 5.4.3b above) function when no longer needed.

Return Value

This function returns *zero* on success or a documented *non-zero* error code otherwise.

A-5.4.4b Gallery Creation

```
INT32
create_gallery(   const INT32  numExemplars,
                  const char   **exemplarFeatFilenames,
                  const char   *galleryDir);
```

Description

This function writes a proprietary enrolled gallery to ***galleryDir*** (e.g. "/mnt/output/gallery/E1/"), based on a list of exemplar feature set file pathnames specified by ***exemplarFeatFilenames***. The gallery shall be usable in read-only mode by subsequent calls to `latent_search()`, and shall associate all exemplar feature sets having the same subject ID (see below). The format of the gallery is at the discretion of the SDK provider. Subdirectories and multiple files may be created within ***galleryDir***. All data produced by the SDK during the execution of this function shall be stored exclusively to the directory specified by ***galleryDir.***

The list of exemplar feature set file pathnames is contained in ***exemplarFeatFilenames***, which is an array of pointers having length ***numExemplars*** + 1, where each element of the array is a pointer to an exemplar feature set file pathname. The last element of the array will be equal to 0 (i.e. a NULL pointer).

The format of all pathnames will be canonical Unix style pathnames using forward slash directory separators. The maximum total pathname length is 255 characters.

Each exemplar feature set file pathname will be formatted *dirPath*"E"*subjectID* "_" *instance* ".feat" (no quotes or spaces), where *dirPath* is the full directory path of the file, *subjectID* is a 6-digit numeric ID (with leading zeros) uniquely identifying the subject, and *instance* is a 1-digit arbitrary numeric index to

differentiate between multiple exemplar sets belonging to the same subject. For example, "/mnt/output/feats/E1/E199999_1.feat"

Return Value

This function returns *zero* on success or a documented *non-zero* error code otherwise.

A-5.4.4c Set Gallery

```
INT32
set_gallery(const char  *galleryDir);
```

Description

This function selects the gallery which shall be used by all subsequent calls to `latent_search()`. The directory pathname specified by ***galleryDir*** (e.g. "/mnt/output/gallery/E1/") shall contain the gallery produced by a prior call to `create_gallery()`.

The format of the pathname will be canonical Unix style pathnames using forward slash directory separators. The maximum total pathname length is 255 characters.

Return Value

This function returns *zero* on success or a documented *non-zero* error code otherwise.

A-5.4.4d Latent Feature Extraction

```
INT32
extract_latent(   const char *latentFilename,
                  const char *outputDir);
```

Description

This function produces a single proprietary formatted feature set file from an ANSI/NIST file containing a set of 0 or more manually extracted features and a latent fingerprint image (except for subtest LG, see section 7). The proprietary formatted feature set file output by this function will be used as input to `latent_search()`.

The ANSI/NIST file will be specified by a pathname pointed to by ***latentFilename*** (e.g. "/mnt1/input/L3/L12ABC.an2"). The directory to which the proprietary feature set file shall be written is specified by the pathname pointed to by ***outputDir*** (e.g. "/mnt/output/feats/L3/").

The format of all pathnames will be canonical Unix style pathnames using forward slash directory separators. The maximum total pathname length is 255 characters.

A single proprietary formatted feature set file shall be written to the directory specified by ***outputDir***. No format is prescribed for the feature data. The feature data may include any or all manually extracted features already present in the ANSI/NIST file (e.g. it may encode them in a proprietary format). For example if desired it may contain the latent fingerprint image. No files other than the feature set file may be written. A feature file shall always be output, regardless of any internal failures such as a failure of automated feature extraction. The filename of the output feature set file is defined here as the base filename of ***latentFilename*** with the extension ".an2" replaced by ".feat" (no quotes). For example, if ***latentFilename*** = "/mnt1/input/L3/L12ABC.an2" and ***outputDir*** = "/mnt/output/feats/L3/", the proprietary feature set file shall be written as "/mnt/output/feats/L3/L12ABC.feat".

Return Value

This function returns *zero* on success or a documented *non-zero* error code on failure.

A-5.4.4e Latent Search

```
INT32
latent_search(    const BYTE *latentFeatFilename,
                  const char *outputDir);
```

Description

This function searches the current gallery (as selected by `set_gallery()`) for zero or more candidates matching the input latent feature set (created by `extract_latent()`) whose pathname is specified by ***latentFeatFilename***, and outputs a candidate list to the directory specified by ***outputDir***. The format of the candidate list is specified in section 5.6.

The selection of features on which to match is entirely at the discretion of the SDK. Note that during the call to this function the directory containing the current gallery and its contents are read-only.

The format of all pathnames will be canonical Unix style pathnames using forward slash directory separators. The maximum total pathname length is 255 characters.

One candidate list file (per call to this function) shall be written to the directory specified by ***outputDir***. A candidate list file shall always be output, regardless of any internal software failures. The filename of the candidate list file is defined here as the base filename of ***latentFeatFilename*** with the extension ".feat" replaced by ".CL" (no quotes). For example, if ***latentFeatFilename*** = "/mnt1/output/feats/L3/L12ABC.feat" and ***outputDir*** = "/mnt/output/clists/L3/", the candidate list file shall be written as "/mnt/output/clists/L3/L12ABC.CL".

Note 1: Since it may not be possible to keep all gallery data in memory, it might be necessary for the software to repeatedly retrieve the data from disk, and this extra fetch time will be included in the execution time measurement.

Note 2: The candidate list shall only depend on the inputs to this function and the currently selected gallery (not on any previous results from this function). Thus, identical latent feature inputs and gallery data shall produce identical candidate lists independent of all prior calls to this function.

Return Value

This function returns *zero* on success or a documented *non-zero* error code on failure.

A-5.4.5 Error Codes and Handling

The participant shall provide documentation of all (non-zero) error or warning return codes (see section A-5.4.8, Documentation).

The application should include error/exception handling so that in the case of a fatal error, the return code is still provided to the calling application.

All messages which convey errors, warnings or other information shall be suppressed, except where they may provide additional information not conveyable by the defined error codes alone (such as listing a specific file related to the error).

At minimum the following return codes shall be used.

Return code	Function	Explanation
0	All	Success
-1	extract_image_data()	unable to open file
-2	extract_image_data()	Incorrect file format
-3	extract_image_data()	error parsing ten-print file
-4	extract_image_data()	error decompressing image

-5	extract_image_data()	insufficient memory error
-6	extract_image_data()	unspecified error
100	extract_exemplar()	exemplar file not found
101	extract_exemplar()	output directory not found
102	extract_exemplar()	unable to write feature data
103	extract_exemplar()	error from extract_image_data (write to stdout)
201	create_gallery()	feature file not found (write filename to stdout)
202	create_gallery()	output directory not found
203	create_gallery()	unable to write gallery enrollment data
204	create_gallery()	insufficient memory available
301	extract_latent ()	latent file not found
302	extract_latent()	output directory not found
303	extract_latent()	unable to write feature data
401	set_gallery()	gallery directory not found
501	latent_search()	gallery directory not set
502	latent_search()	insufficient memory available
503	latent_search()	feature file not found
504	latent_search()	candidate list directory not found
505	latent_search()	unable to write candidate list

A-5.4.6 SDK Library and Platform Requirements

Participants shall provide NIST with binary code only (i.e. no source code) − supporting files such as header (".h") files notwithstanding.

Note that dependencies on external dynamic/shared libraries such as compiler-specific development environment libraries are discouraged. If absolutely necessary, external libraries must be provided to NIST upon prior approval by the Test Liaison.

The SDK will be tested in non-interactive "batch" mode (i.e. without terminal support). Thus, the library code provided shall not use any interactive functions such as graphical user interface (GUI) calls, or any other calls which require terminal interaction.

The use of multi-threading by the SDK is encouraged as the NIST test platform includes dual-processor dual-core support. The SDK need not be "thread safe" as the NIST test driver itself is single threaded. If multi-threading is utilized by the SDK is shall be documented.

NIST will link the provided library file(s) to a C language test driver application (developed by NIST) using the GCC compiler (*for Windows platforms Cygwin/GCC version 3.3.1 will be used; for Linux platforms GCC version 4.1.2 and GNU ld 2.17.50.0.6-5.el5 will be used. All GCC compilers use Libc 6*). For example,

```
gcc –o latenttest latenttest.c -L. –lelftEfsSDK
```

Participants are required to provide their library in a format that is linkable using GCC with the NIST test driver, which is compiled with GCC. All compilation and testing will be performed on x86 platforms running either Windows 2008 Server or Red Hat Enterprise Linux Server release

5.1 "Tikanga" (kernel 2.6.18-53 or higher) dependent upon the operating system requirements of the SDK. Thus, participants are strongly advised to verify library-level compatibility with GCC (on an equivalent platform) prior to submitting their software to NIST to avoid linkage problems later on (e.g. symbol name and calling convention mismatches, incorrect binary file formats, etc.).

A-5.4.7 Installation and Usage

The SDK must install easily (i.e. one installation step with no participant interaction required) to be tested, and shall be executable on any number of machines without requiring additional machine-specific license control procedures or activation.

The SDK's usage shall be unlimited. No usage controls or limits based on licenses, execution date/time, number of executions, etc. shall be enforced by the SDK.

It is requested that the SDK be installable using simple file copy methods, and not require the use of a separate installation program. Contact the Test Liaison for prior approval if an installation program is absolutely necessary.

A-5.4.8 Documentation

Complete documentation of the SDK shall be provided, and shall detail any additional functionality or behavior beyond what is specified in this document. The documentation must define all error and warning codes.

A-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

A-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	073141	2	3513	93	18	12

2	199999	2	605	5	18	5
3	004334	3	513	4	18	5
…						
100	920792	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. 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 6-digit *subjectID* (see section 5.4.4b) part of the exemplar filename specified to the create_gallery() function. E.g. if "/mnt/output/feats/E1/E199999_1.feat" was enrolled to the gallery being searched, the Mate ID shall be "199999", without quotes). Note also that the candidate list refers to a subject and finger position, not a specific exemplar impression.

A-5.7 Validation

As discussed in Section A-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.

A-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 A-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.

Proposed 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	0.025 sec/10-finger exemplar set Rate of 40 exemplar sets/sec, per latent (exemplar set = 10 all rolled or all plain prints)

Table 2: Timing requirements

A-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

A-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							

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
  - 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[2] rolled exemplars
  - E3 - 1000ppi plain exemplars (unsegmented slaps)
  - 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 k^{th} 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.

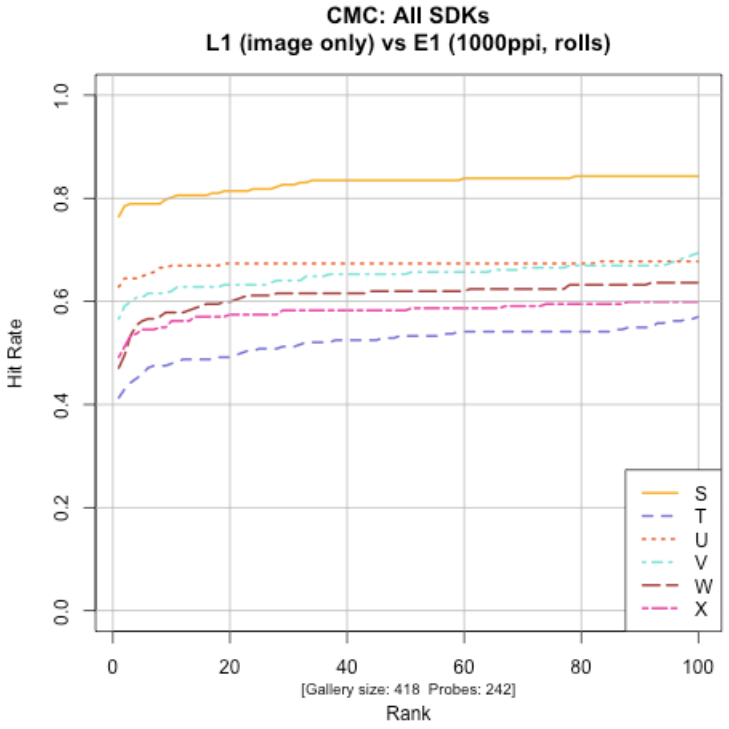
Note that not all participants returned results for all tests.

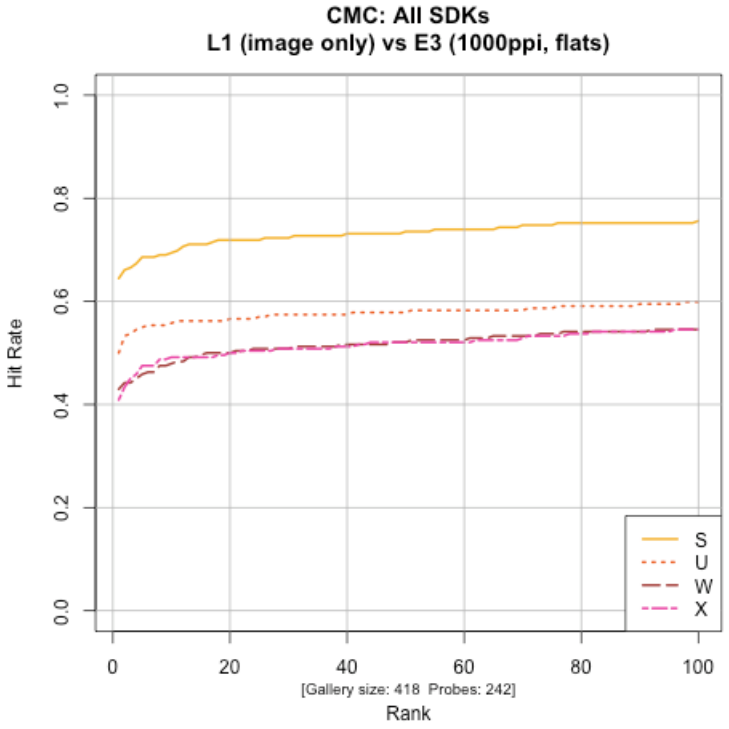
Table 1: Summary of rank-1 identification rates

		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	-

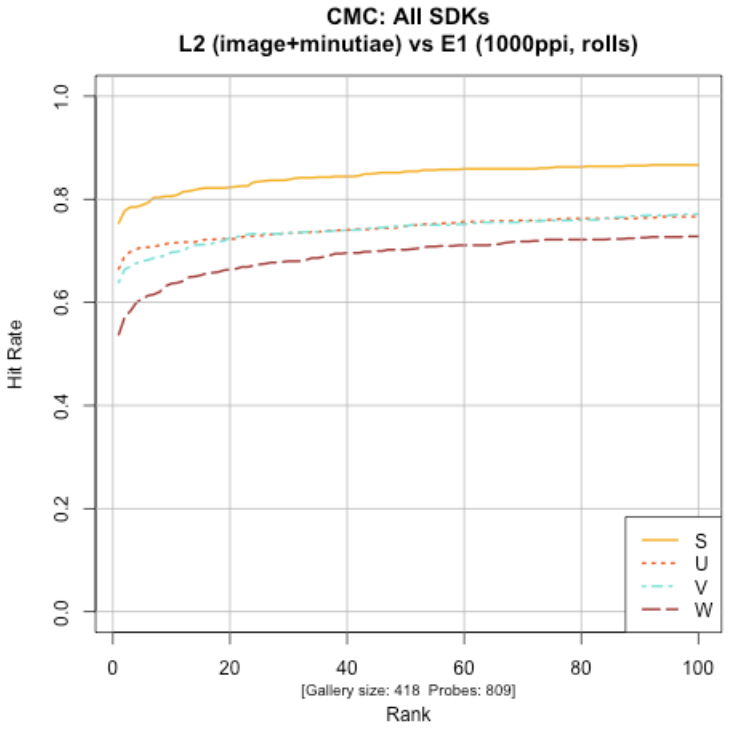
The gallery size was 418 for subtests E1 and E3, and 857 for subtests E2 and E4 (including multiple exemplar sets per subject). There were 242 distinct latent images with multiple markups per image, so there were 242 probes for the image-only subtest (L1), and 809 probes for the other subtests.

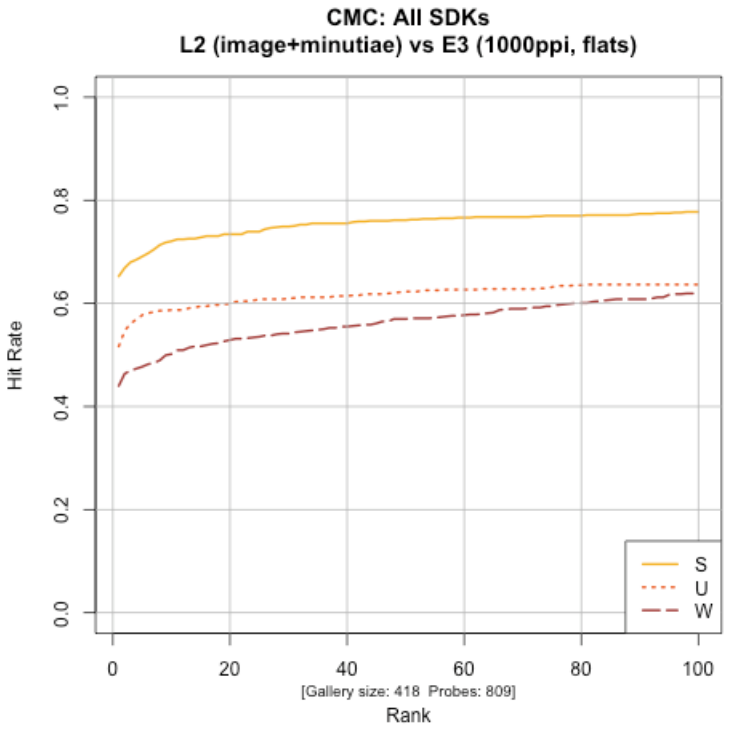
B-5.1 L1 – Image only



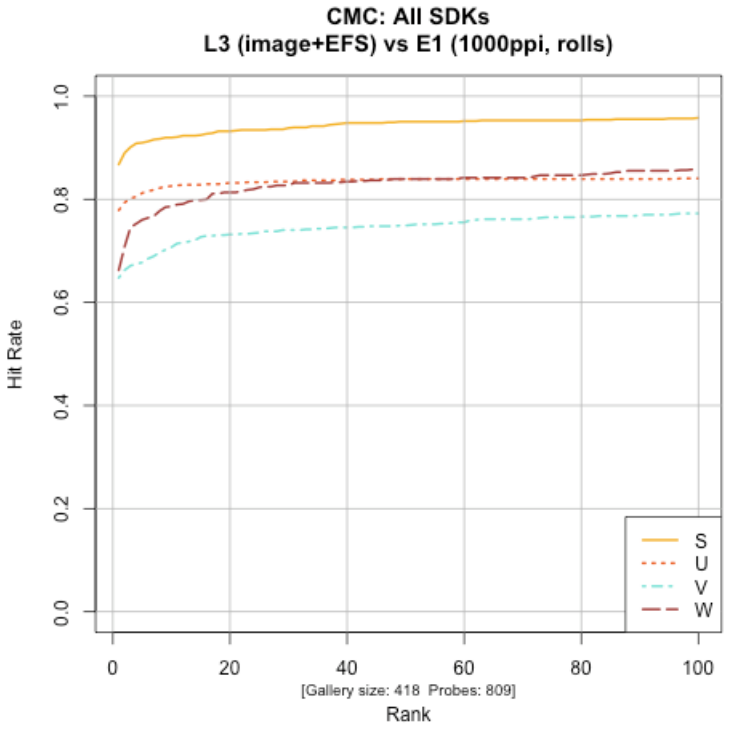


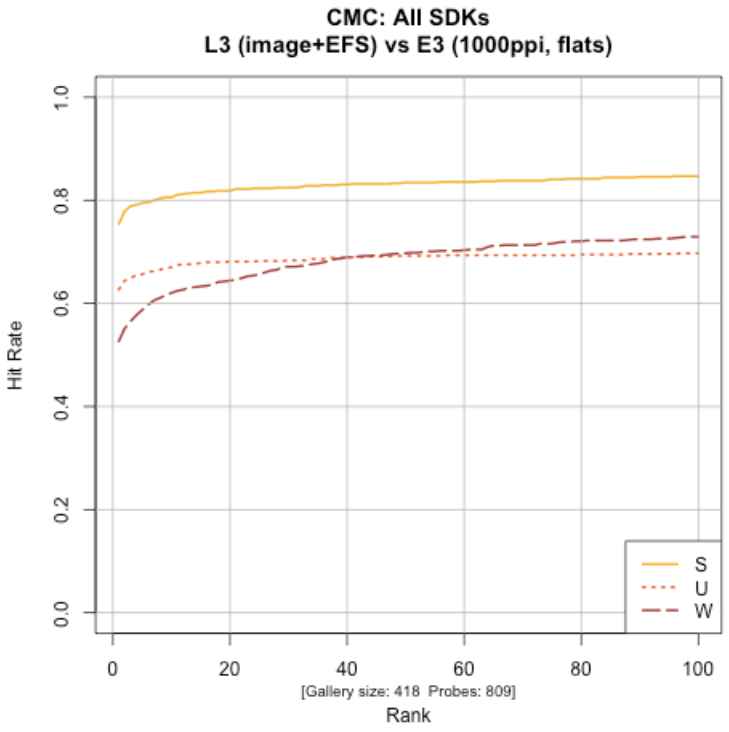
B-5.2 L2 – Image + IAFIS LFFS markup



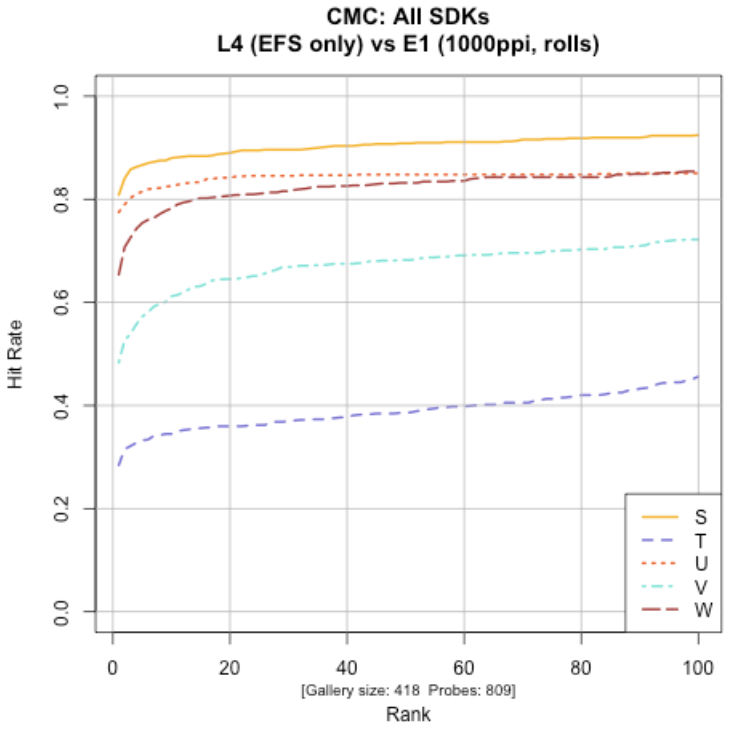


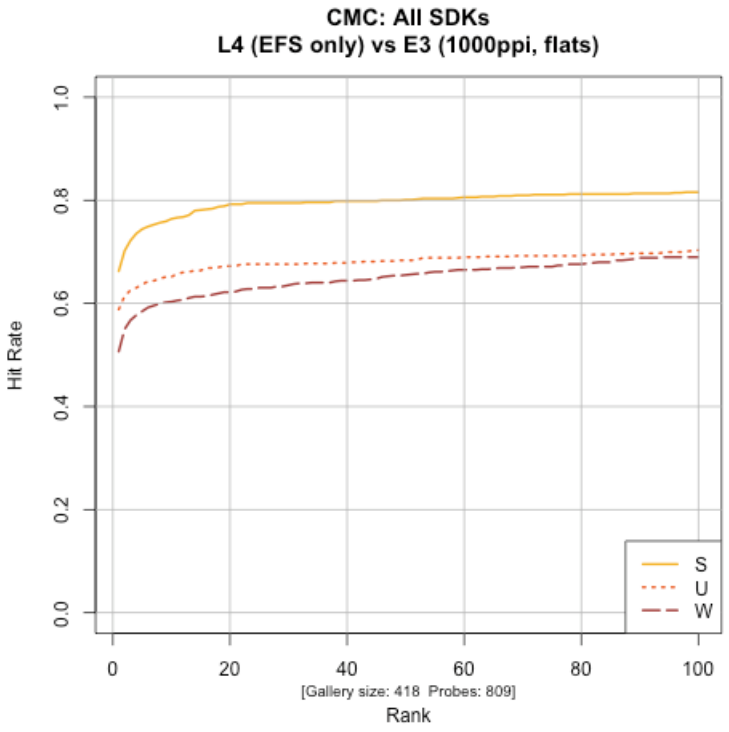
B-5.3 L3 – Image + Extended Feature Set markup



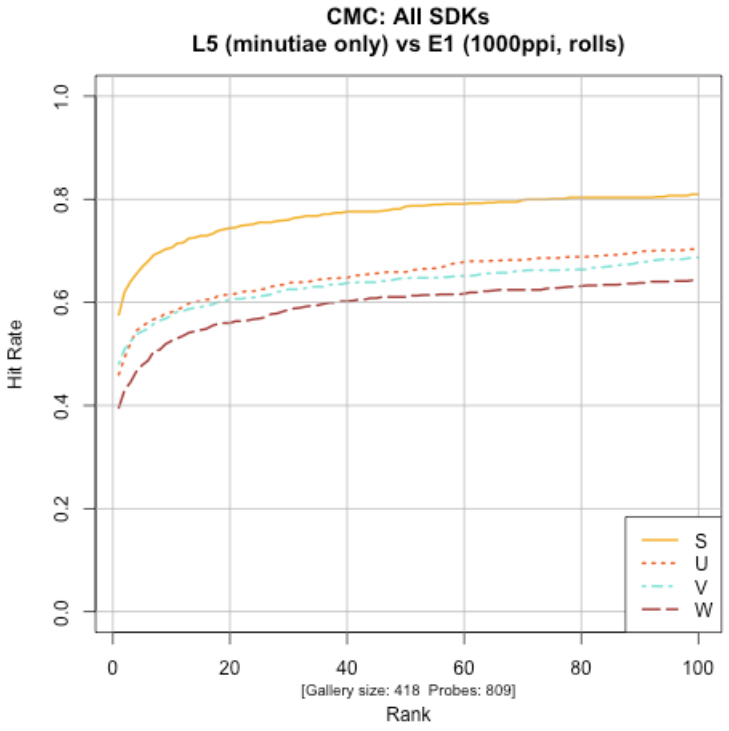


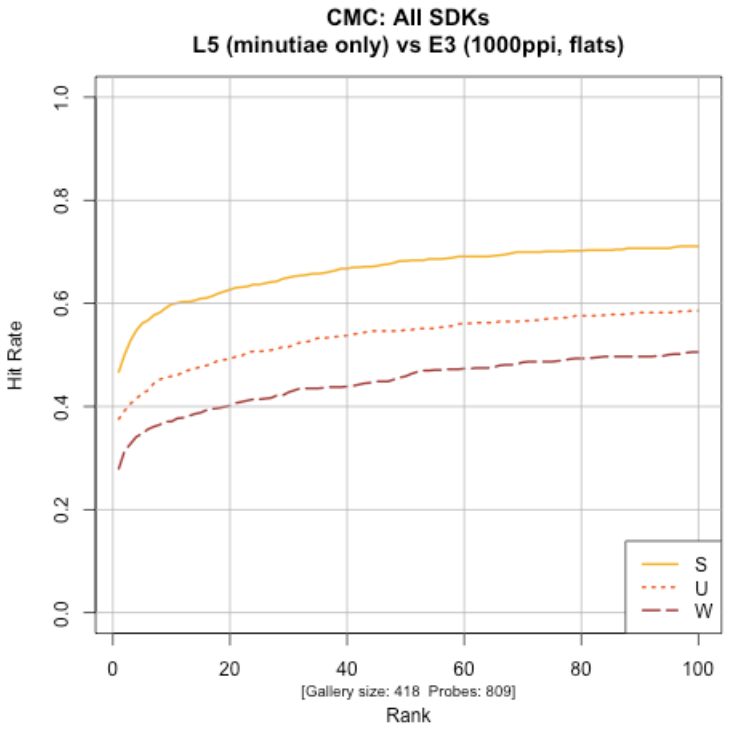
B-5.4 L4 – Extended Feature Set markup (no image)





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

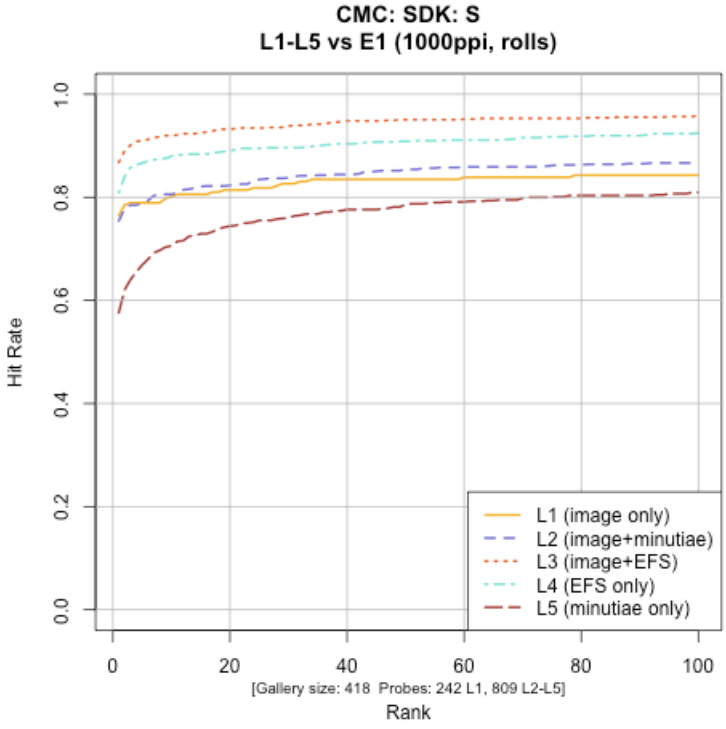


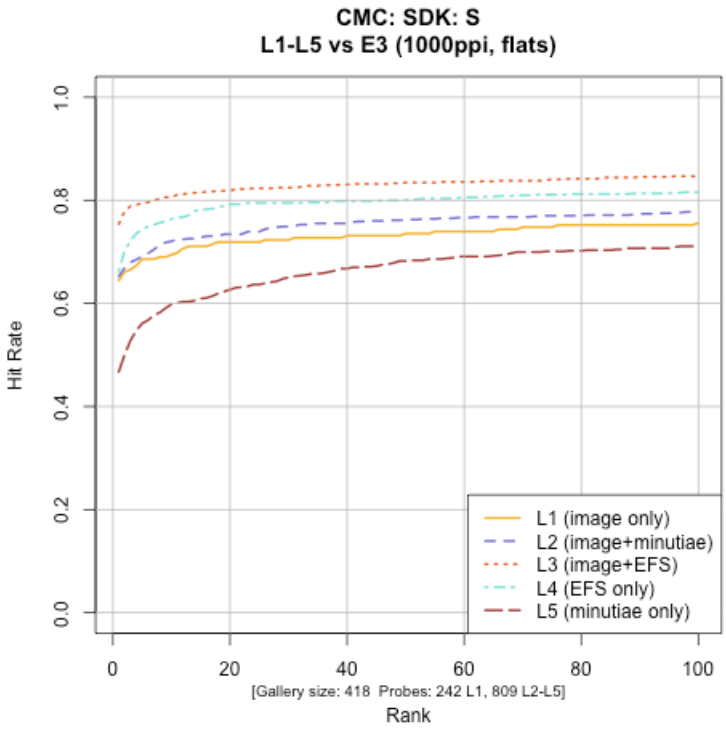


B-6 Results by participant

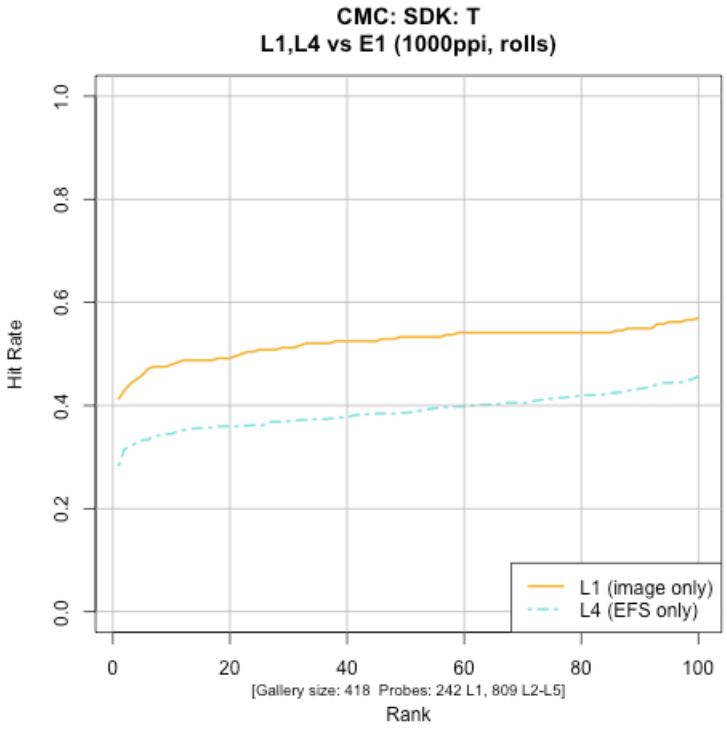
This section reports the same results as the previous section, but with charts grouped by participant.

B-6.1 Participant S

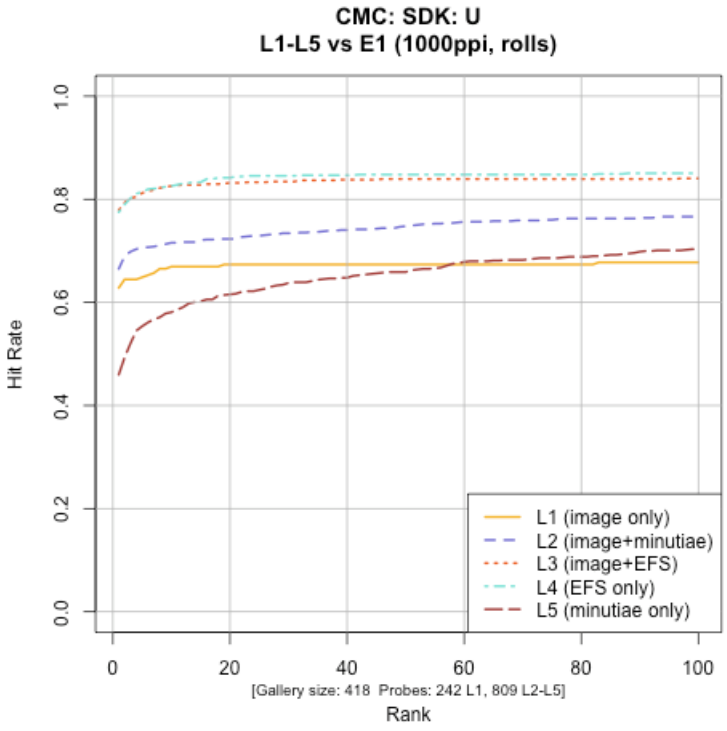


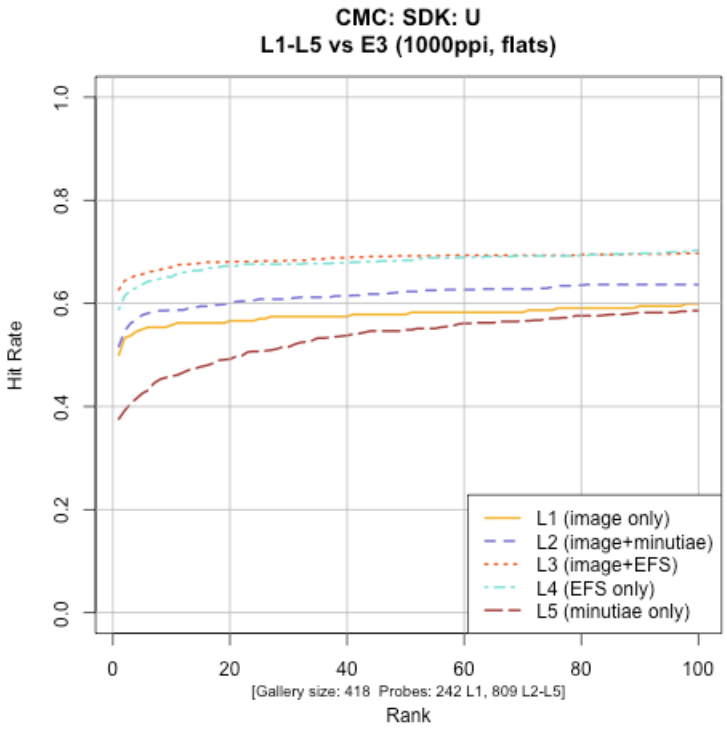


B-6.2 Participant T

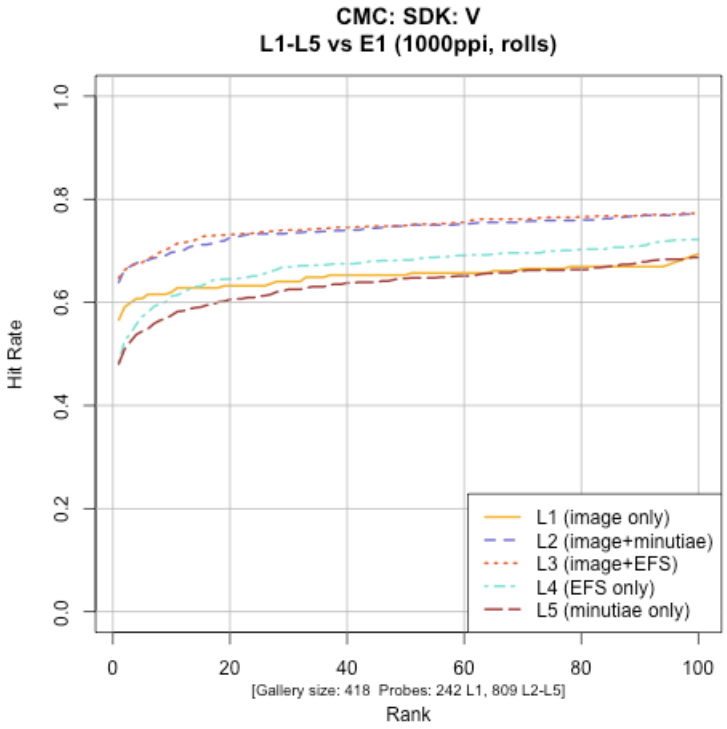


B-6.3 Participant U

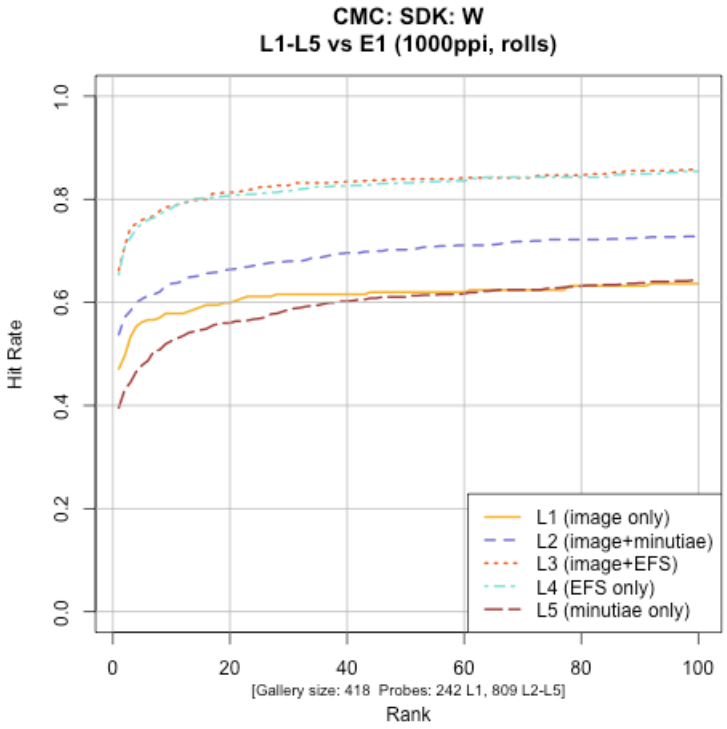


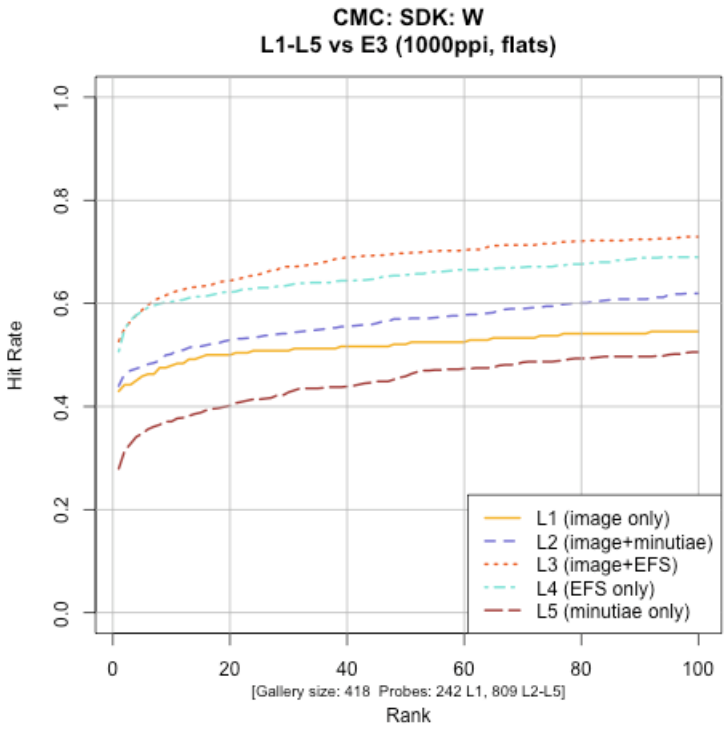


B-6.4 Participant V

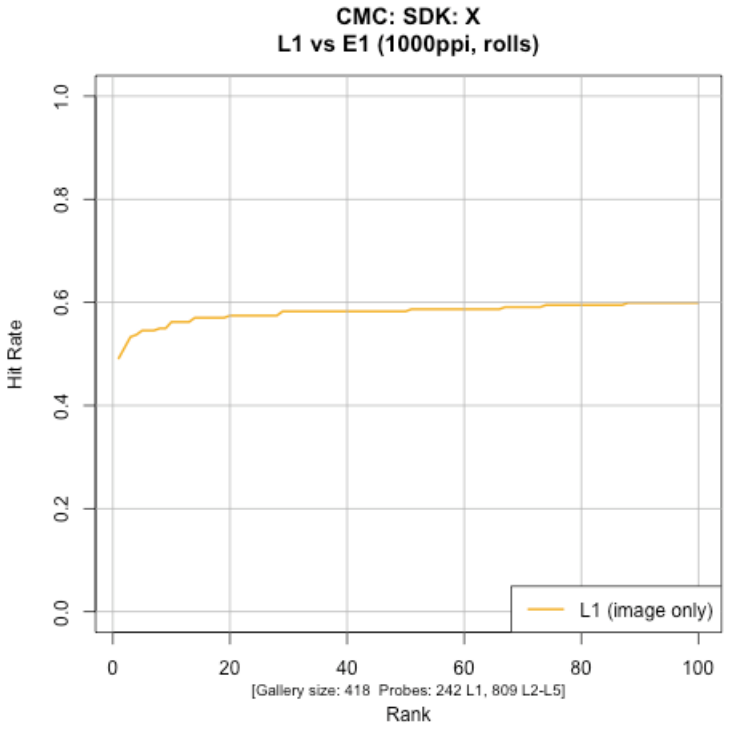


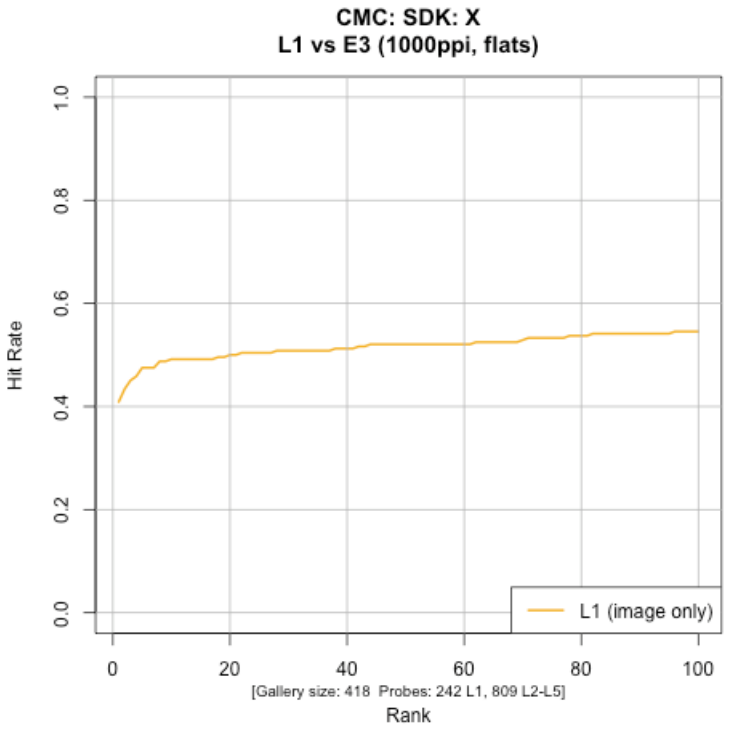
B-6.5 Participant W





B-6.6 Participant X





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.

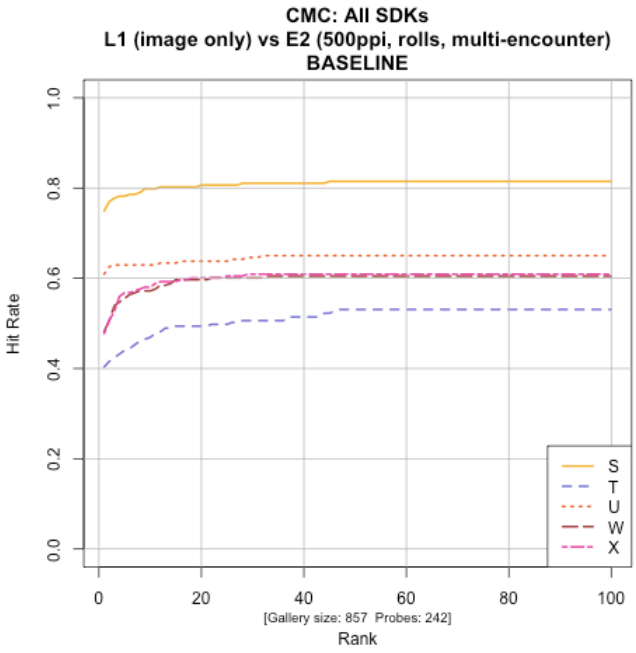
Best Rank

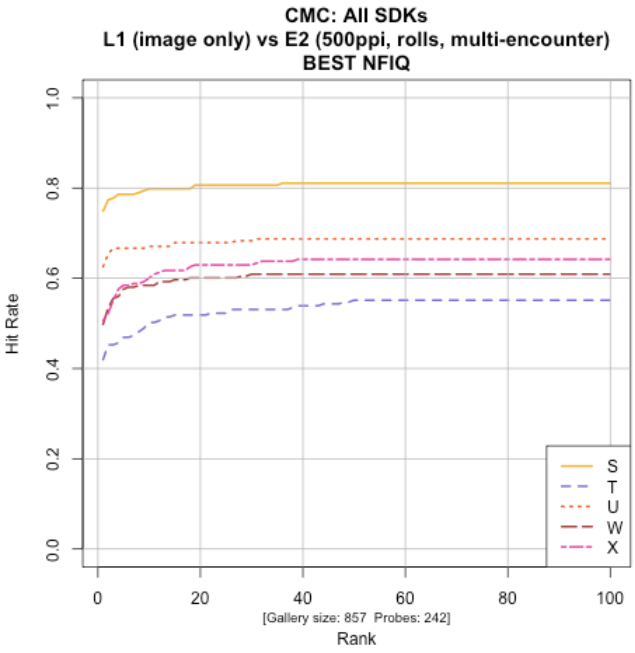
The selected exemplar is simply the highest-ranking result returned for each subject:finger combination. This shows the effect of retaining all exemplars in the gallery and using the highest-scoring results.

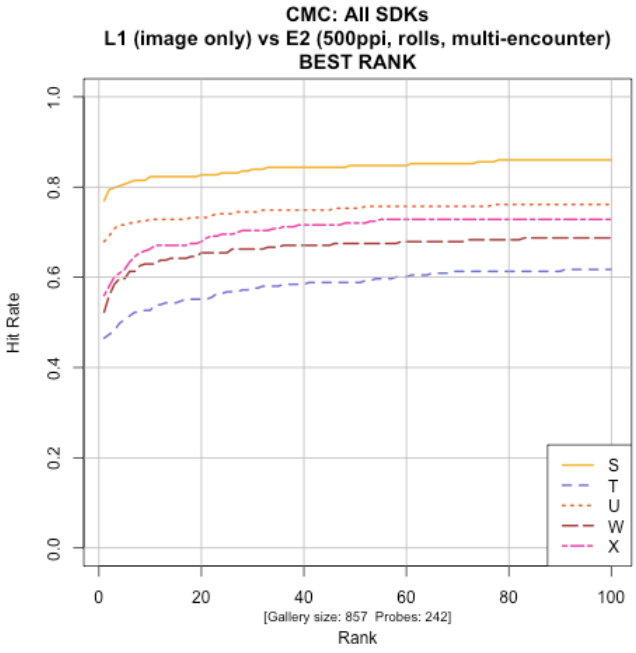
[4] *Ed. Note: the multi-exemplar 500ppi charts in this draft show curves that extend all the way to rank 100. Because of the pruning approach used to handle multi-exemplar data, the candidate lists were almost always reduced to fewer than 100 candidates (generally about 50). In subsequent reporting,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

[5] *NIST Fingerprint Image Quality*

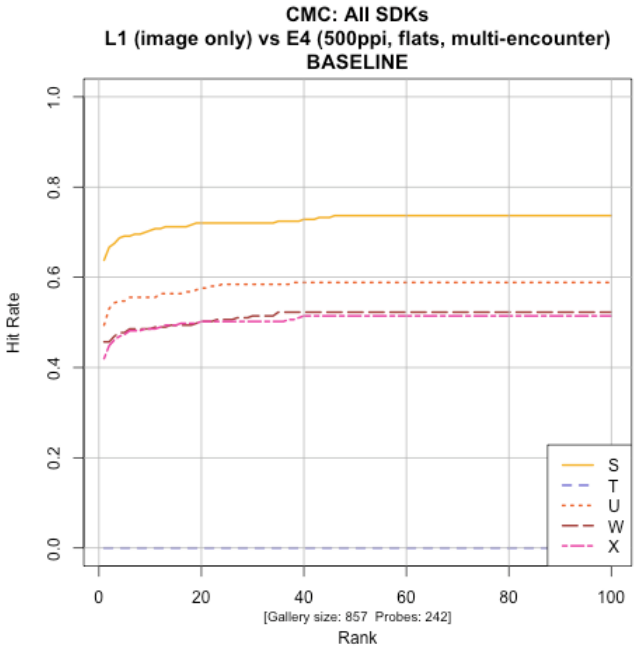
B-7.1 L1 (Image only) x E2 (500ppi rolls)

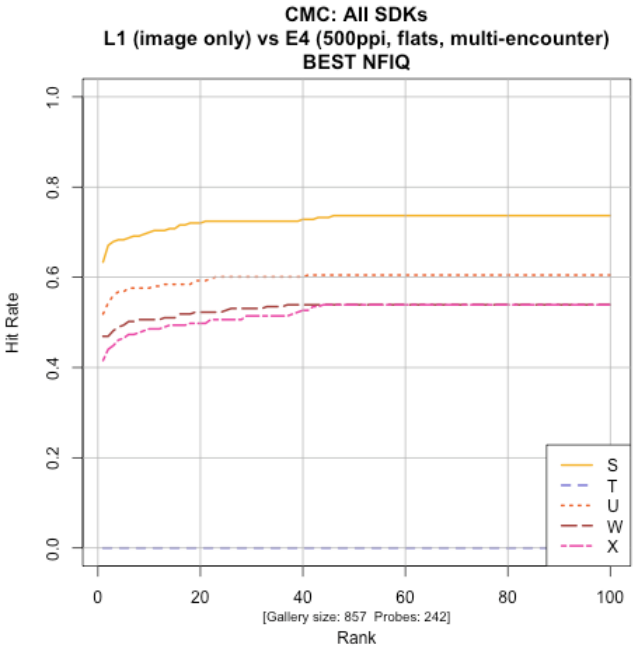


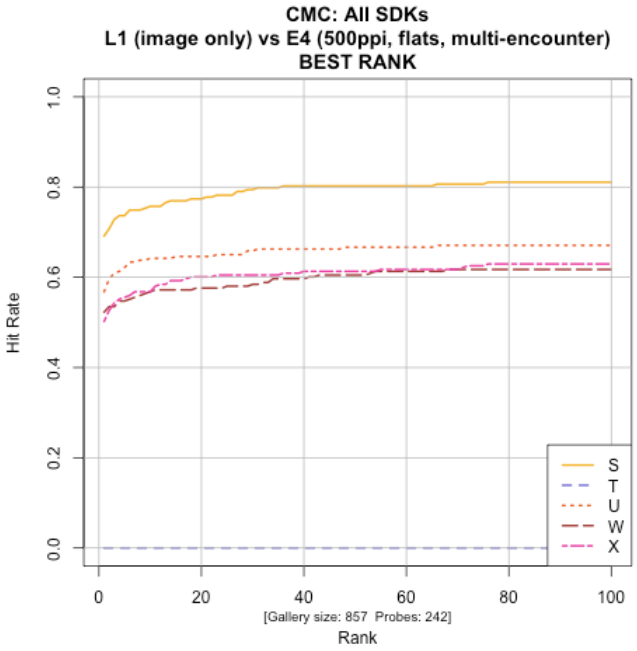




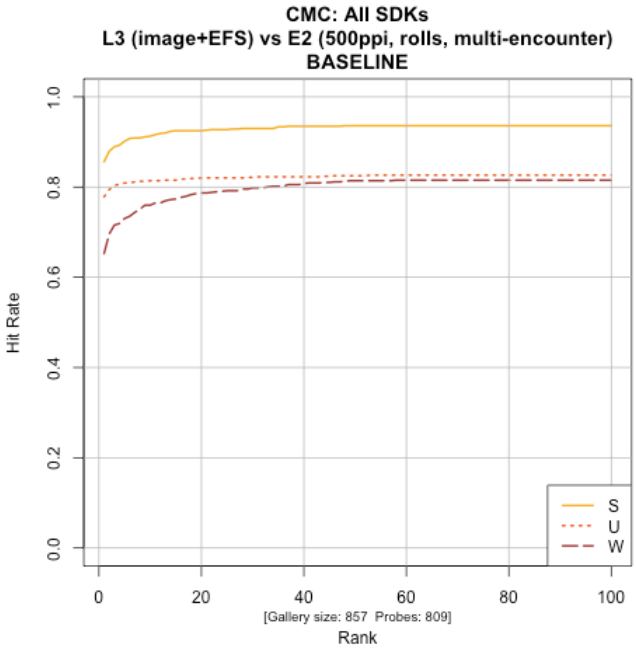
B-7.2 L1 (Image only) x E4 (500ppi flats)

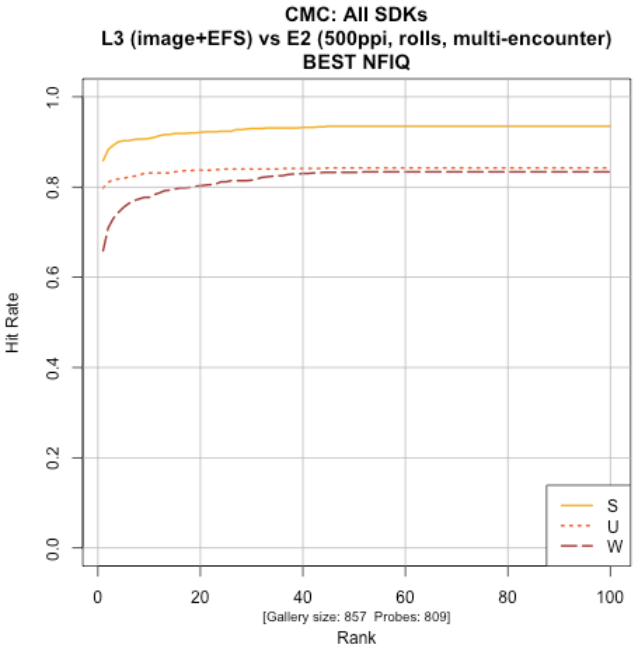


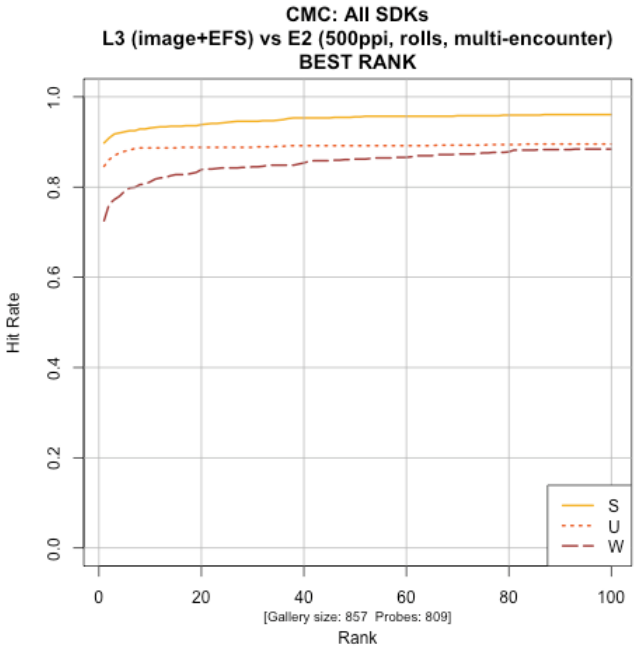




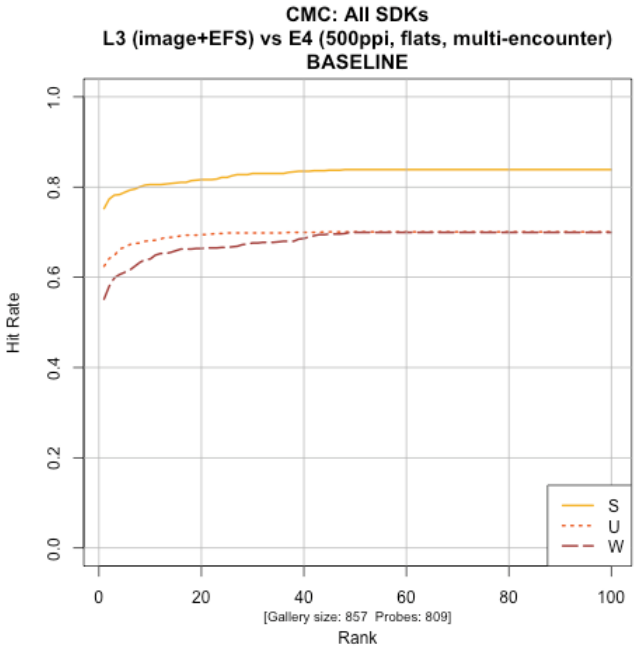
B-7.3 L3 (Image + Extended Feature Set markup) x E2 (500ppi rolls)

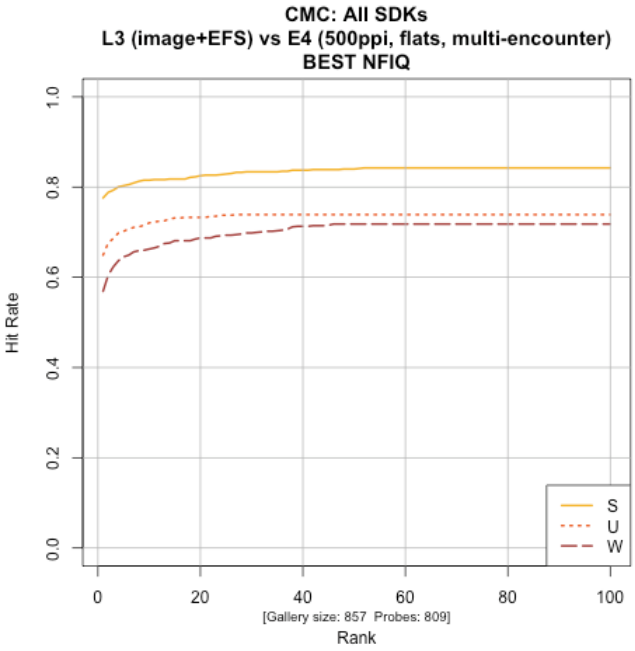


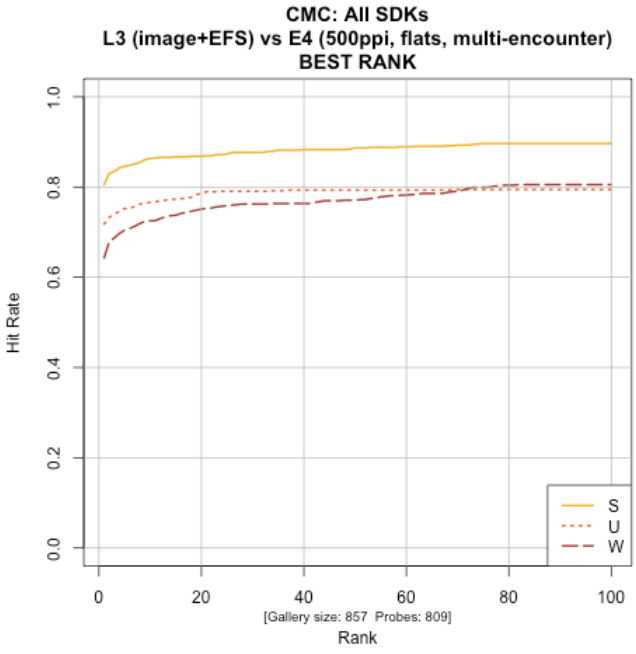




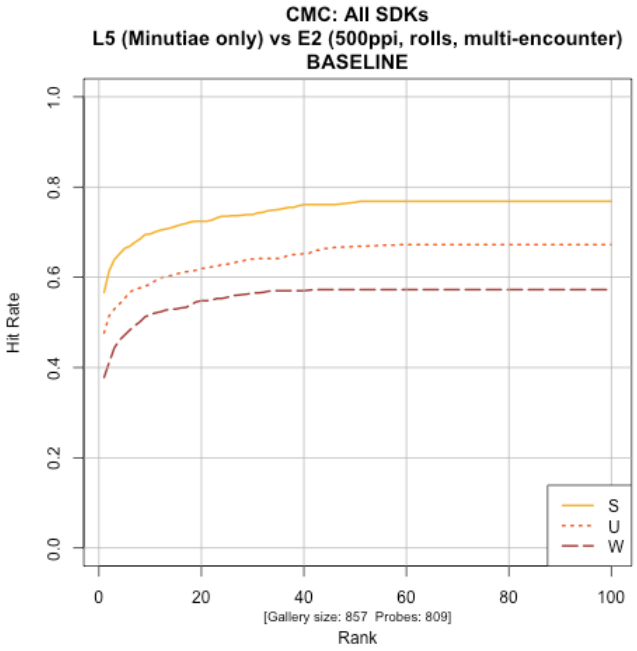
B-7.4 L3 (Image + Extended Feature Set markup) x E4 (500ppi flats)

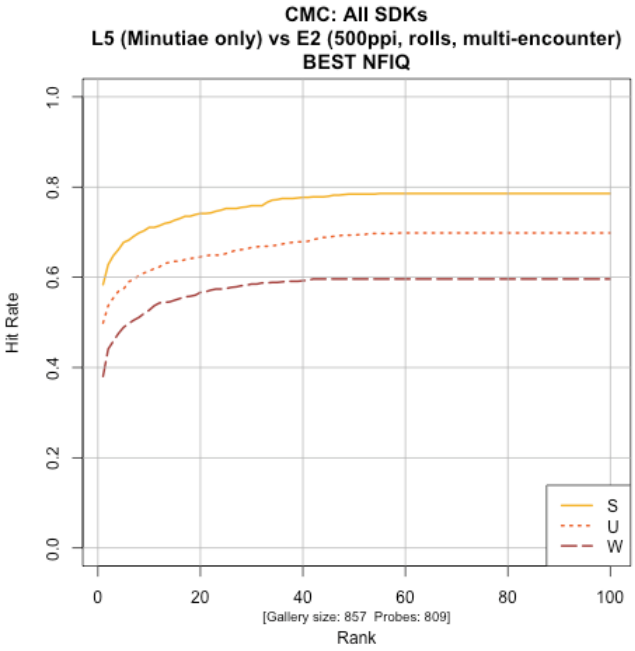


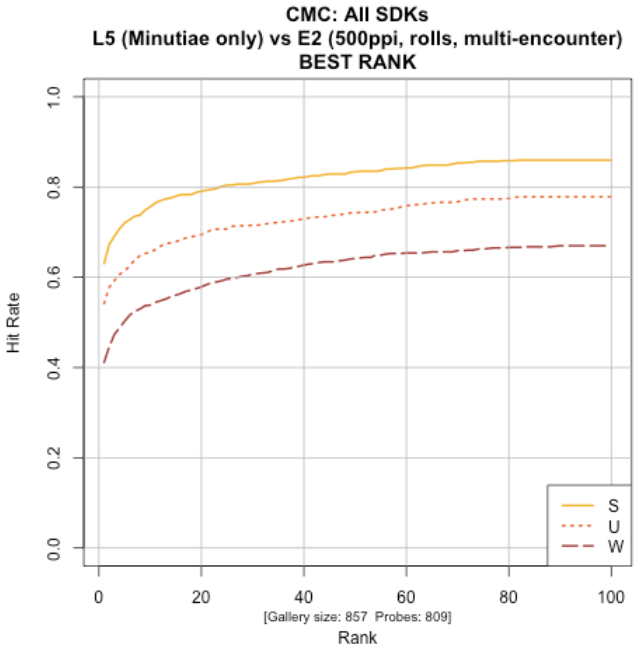




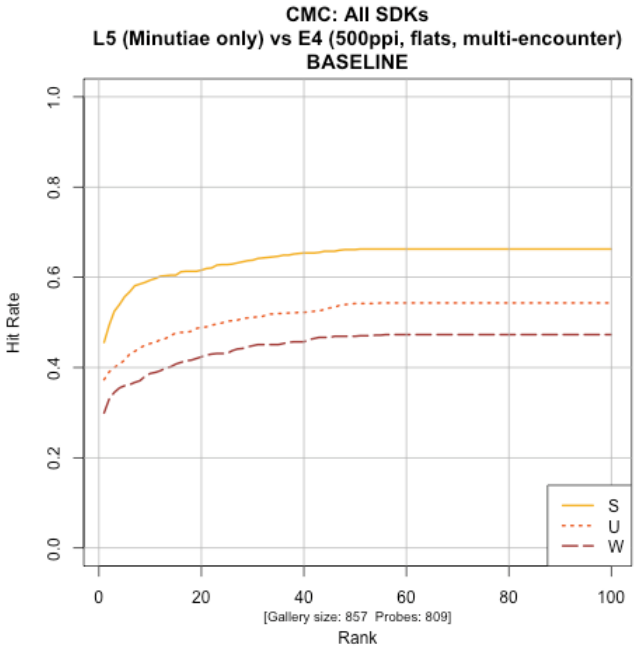
B-7.5 L5 (IAFIS LFFS markup, no image) x E2 (500ppi rolls)

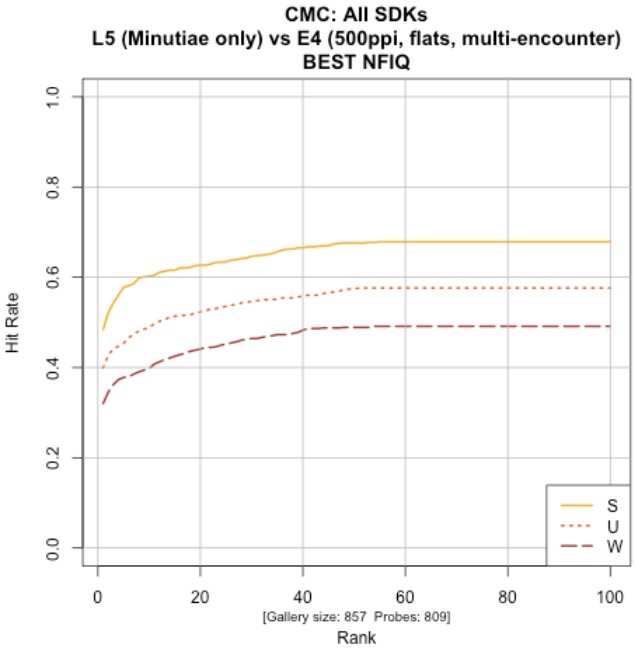


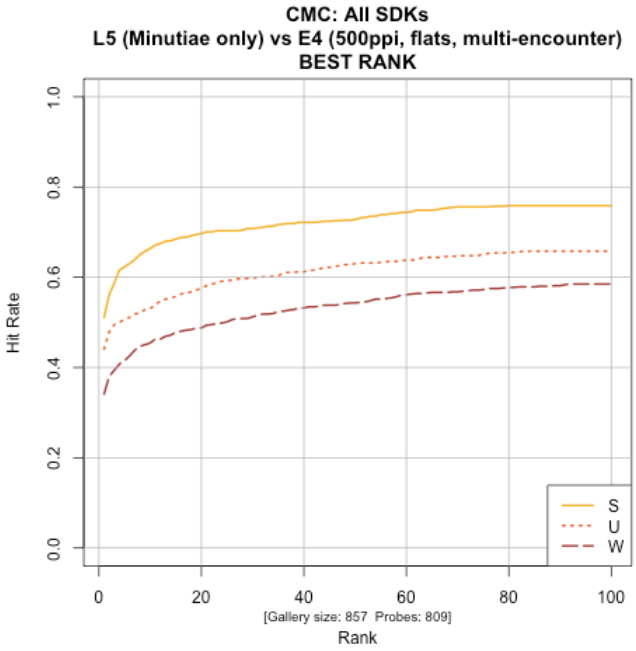




B-7.6 L5 (IAFIS LFFS markup, no image) x E4 (500ppi flats)



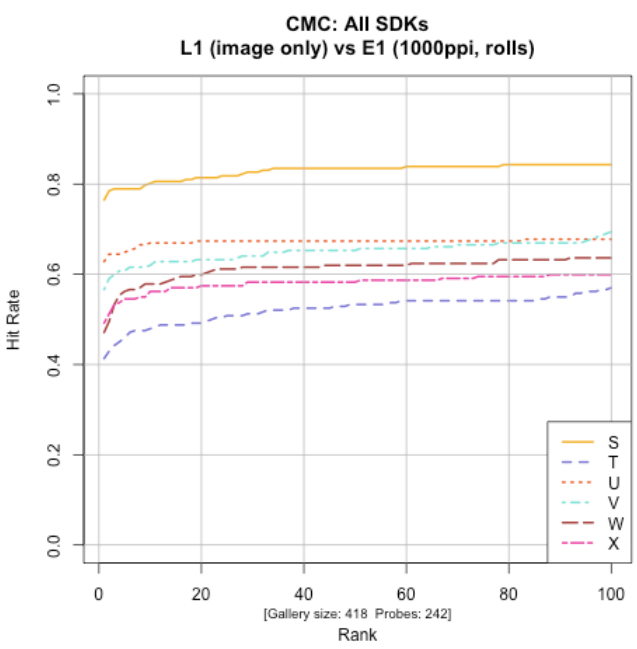


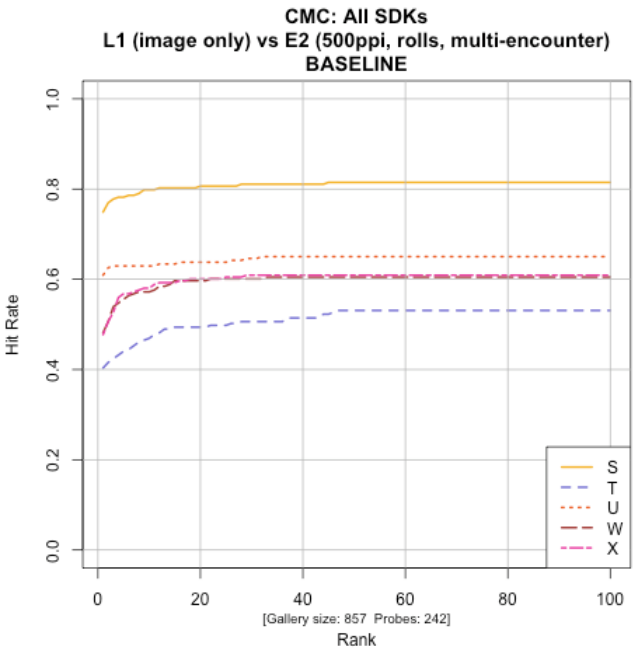


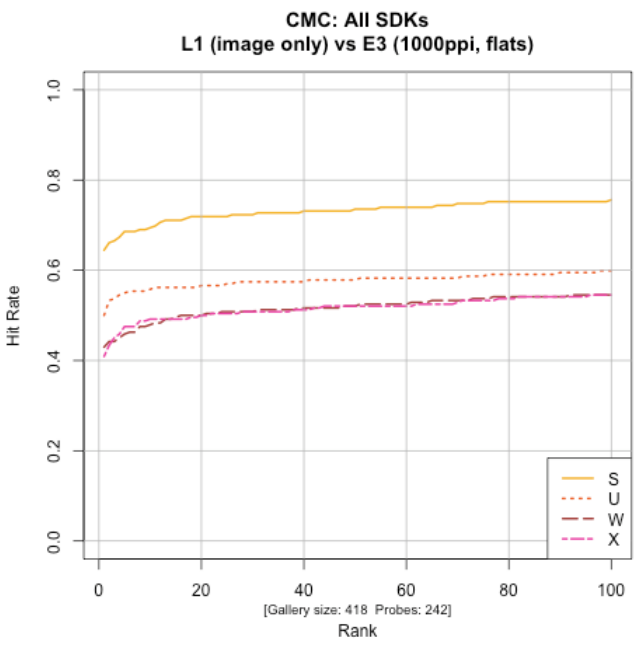
B-8 Resolution

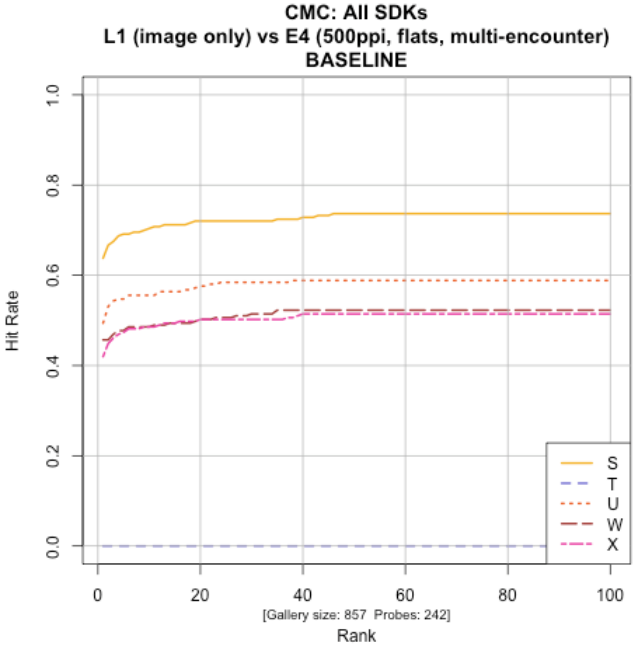
The following charts (repeated from elsewhere in this report) compare the effects for 1000ppi and 500ppi images.

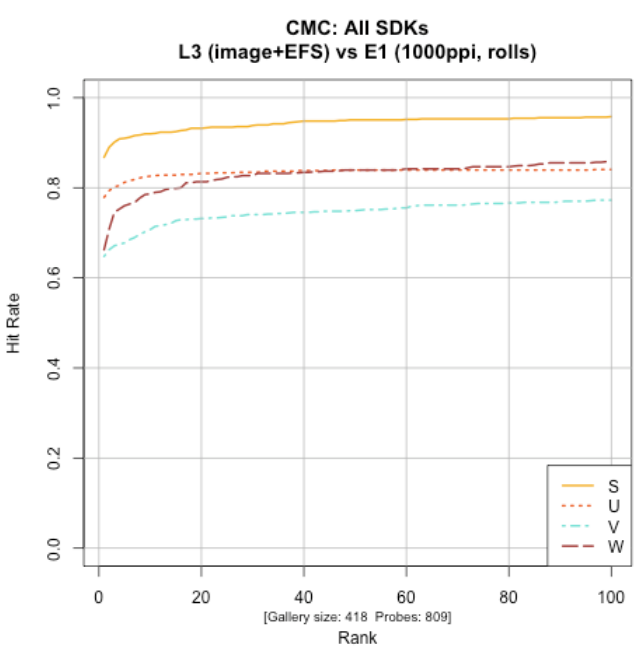
Ed. Note: the multi-exemplar 500ppi charts in this draft show curves that extend all the way to rank 100. Because of the pruning approach used to handle multi-exemplar data, the candidate lists were almost always reduced to fewer than 100 candidates (generally about 50). In subsequent reporting, this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

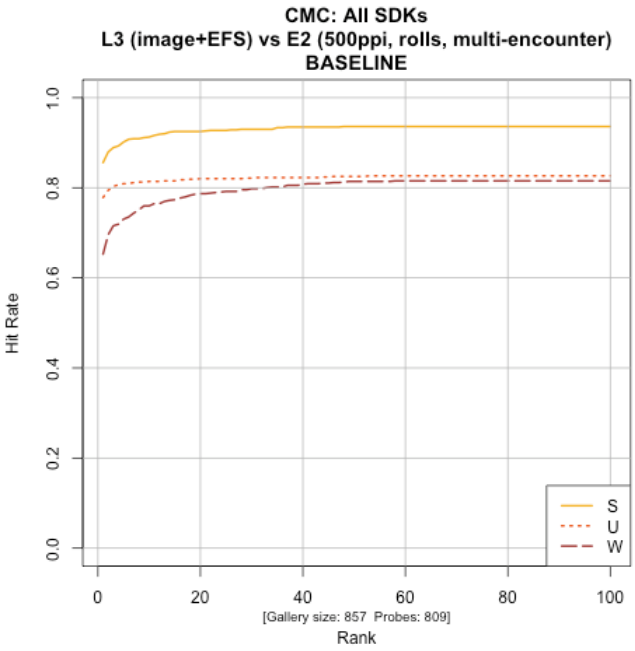


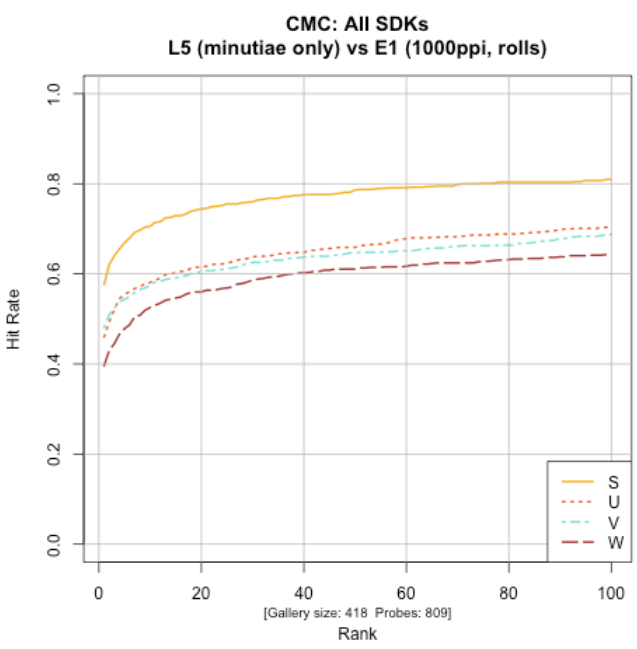


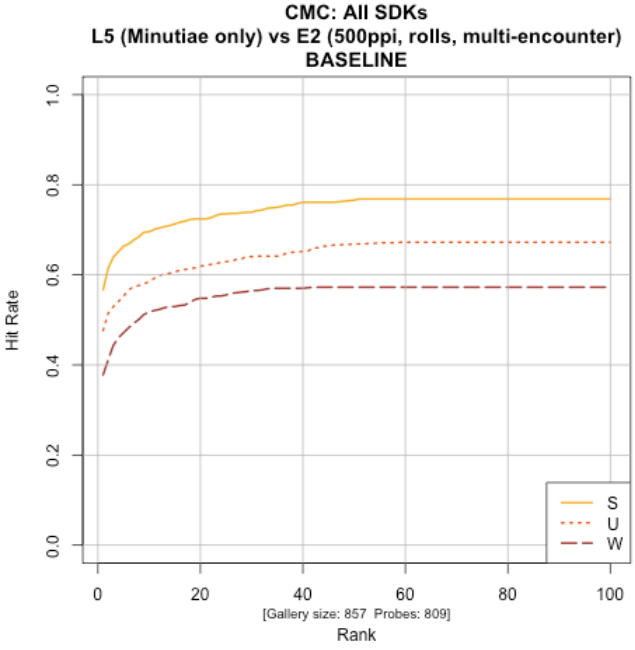












B-9 Score-based results

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.

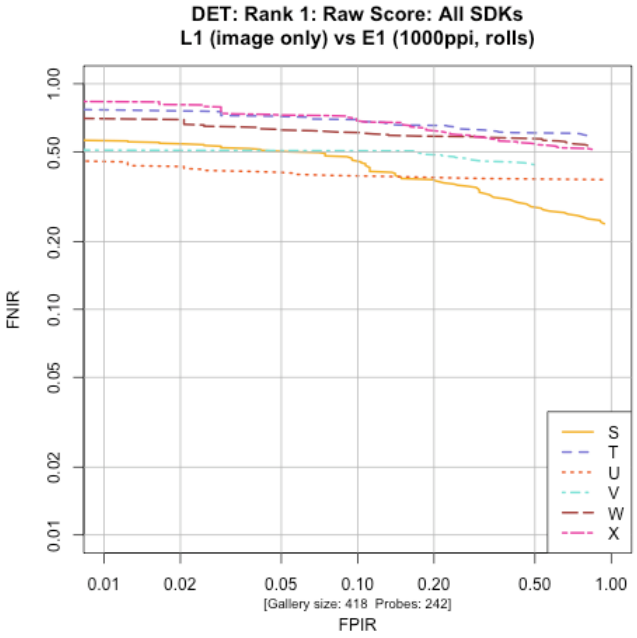
[6] *Indovina, et al; ELFT Phase II - An Evaluation of Automated Latent Fingerprint Identification Technologies; NISTIR 7577; Section 3.1.2 p 24.*

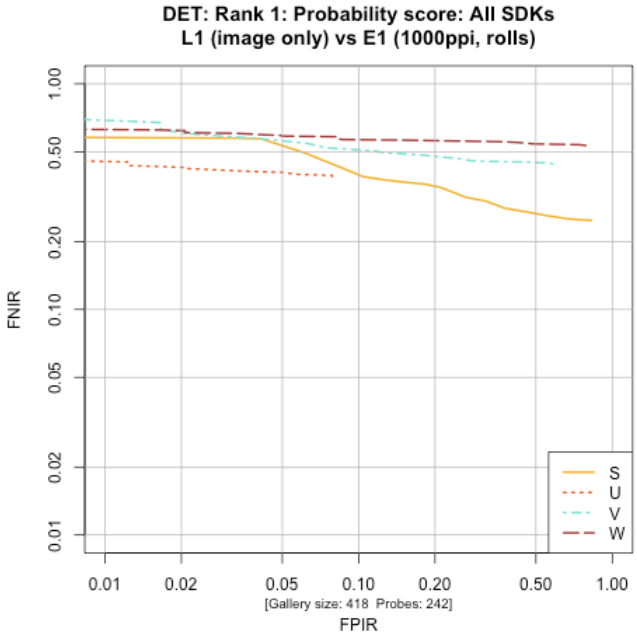
- False Positive Identification Rate (FPIR) indicates the fraction of candidate lists (without enrolled mates) that contain a non-mate entry in the top position with a score greater than the threshold.

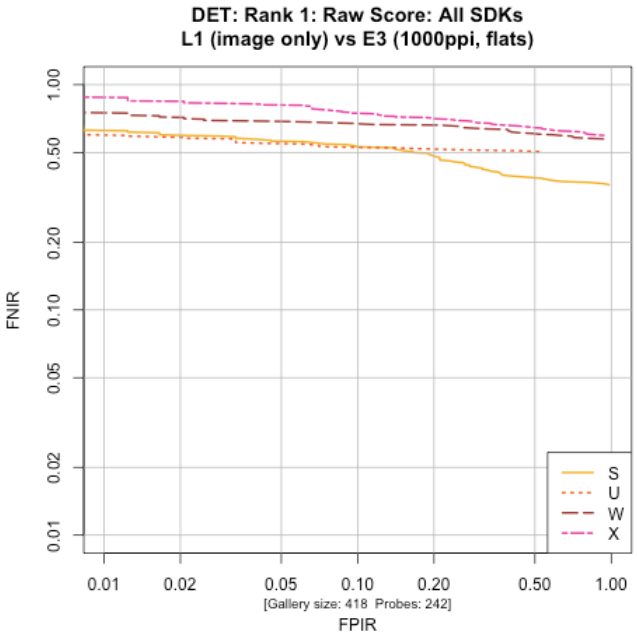
In practice, these charts show the effect of automatically eliminating candidates based on score. For example, in the first chart (Raw score DET for L1 vs E1), for participant S, the FNIR=0.236 @ FPIR=1.0, and reduces to FNIR=0.5 @ FPIR=0.05. What this means is that if a score threshold is set so that 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.

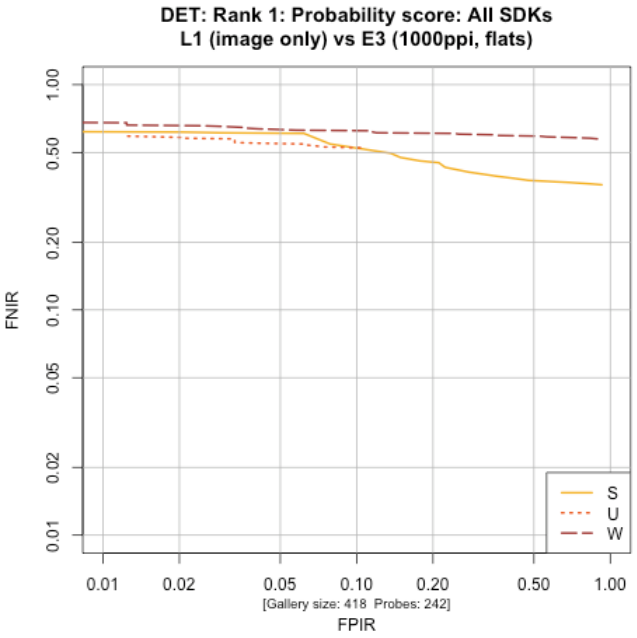
Note that a horizontal line is ideal, indicating no degradation in accuracy as non-mates are automatically excluded. Note also that when the FPIR=1.0, the raw score FNIR I the same as the rank-1 identification rate shown in Table 1 and the CMC analyses above.

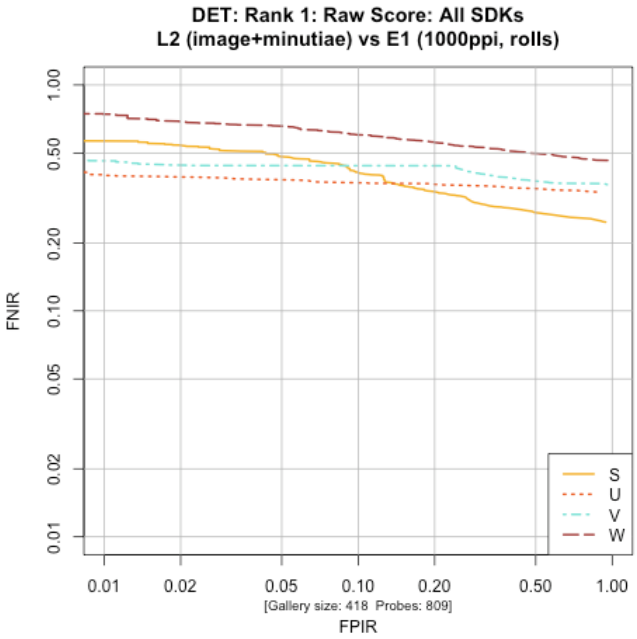
In each case, participants returned a raw score and a normalized score estimating the probability of a match. Not all participants returned probability scores.

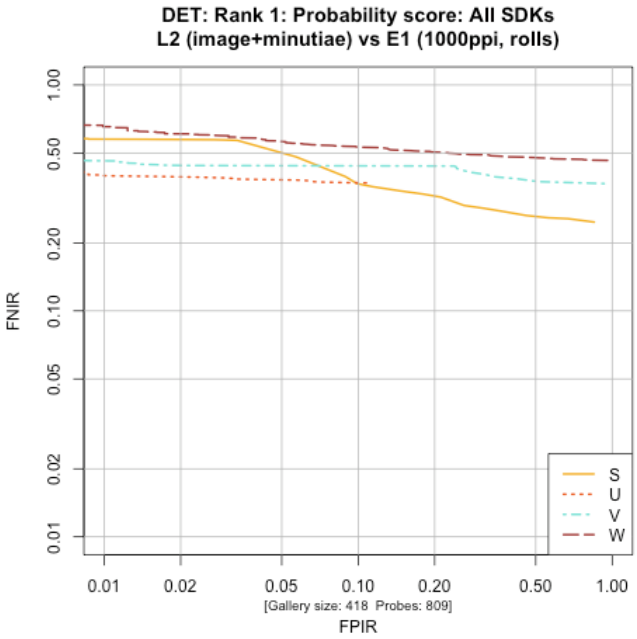


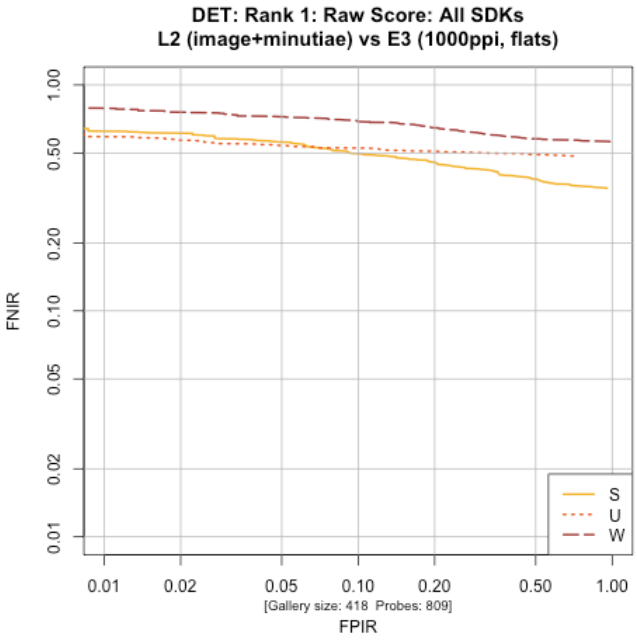


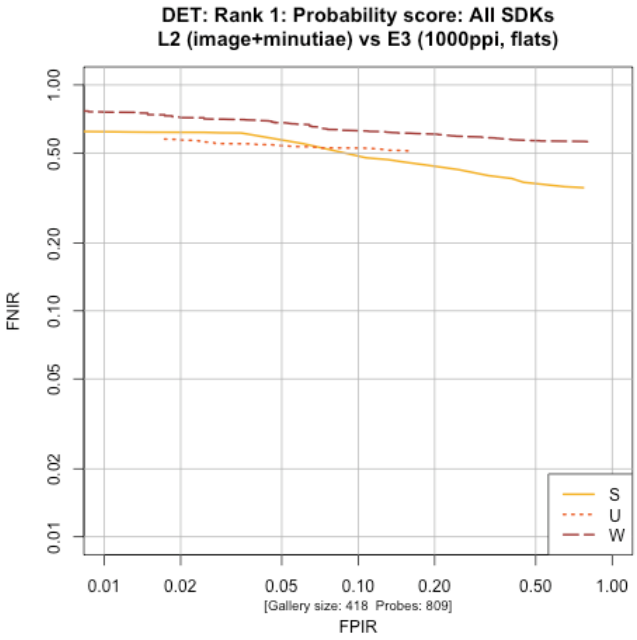

DET: Rank 1: Raw Score: All SDKs
L1 (image only) vs E3 (1000ppi, flats)
FNIR
FPIR
[Gallery size: 418 Probes: 242]
S
U
W
X

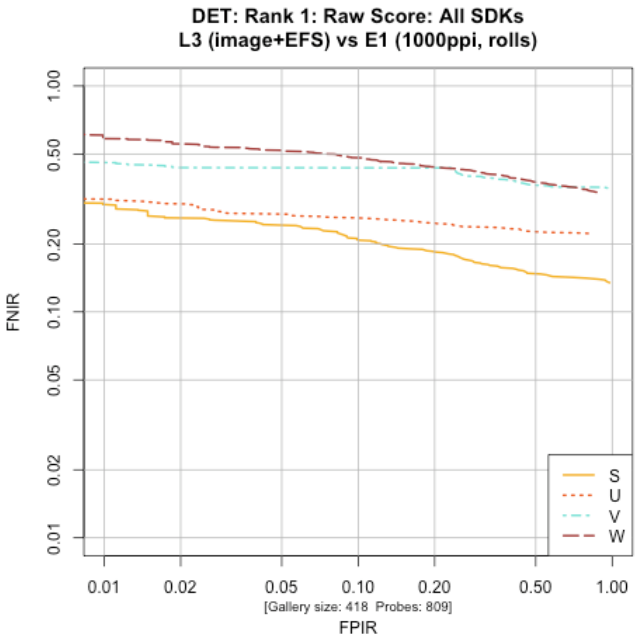

DET: Rank 1: Probability score: All SDKs
L1 (image only) vs E3 (1000ppi, flats)
FNIR
FPIR
[Gallery size: 418 Probes: 242]
S
U
W

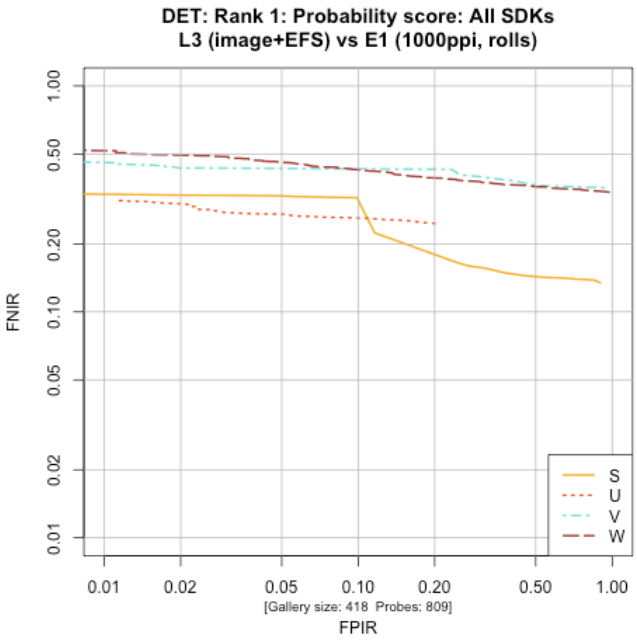

DET: Rank 1: Raw Score: All SDKs
L2 (image+minutiae) vs E1 (1000ppi, rolls)
FNIR
FPIR
[Gallery size: 418 Probes: 809]
S
U
V
W

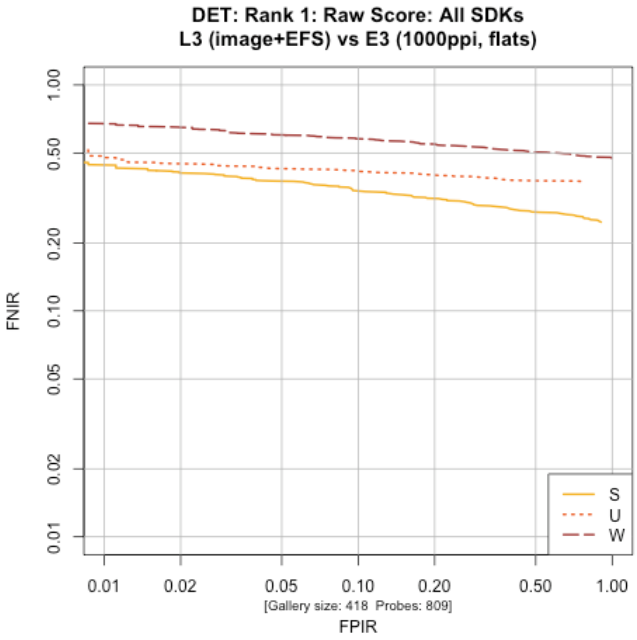

DET: Rank 1: Probability score: All SDKs
L2 (image+minutiae) vs E1 (1000ppi, rolls)
FNIR
FPIR
[Gallery size: 418 Probes: 809]
S
U
V
W

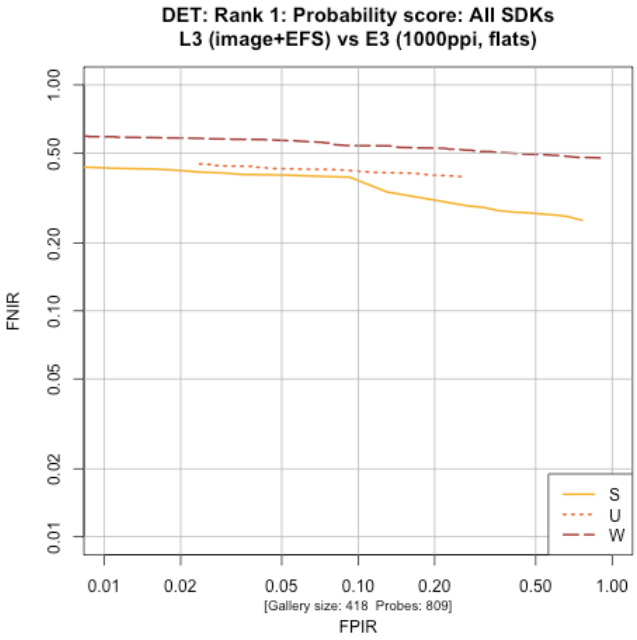

DET: Rank 1: Raw Score: All SDKs
L2 (image+minutiae) vs E3 (1000ppi, flats)
FNIR
FPIR
[Gallery size: 418 Probes: 809]
S
U
W

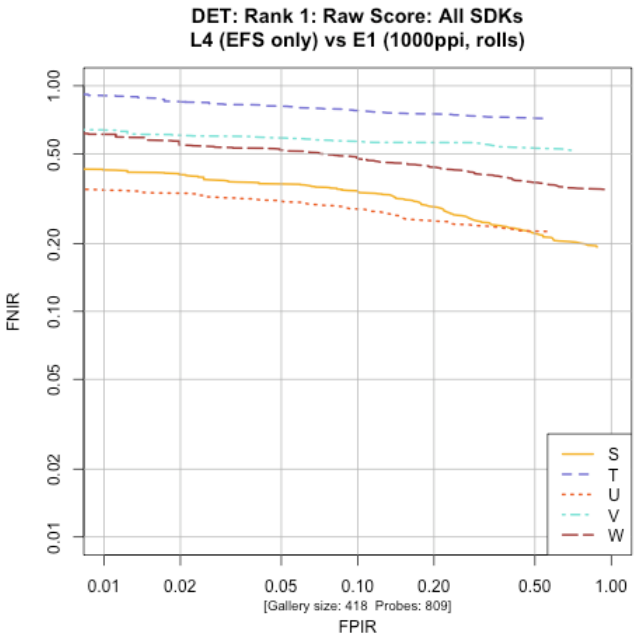

DET: Rank 1: Probability score: All SDKs
L2 (image+minutiae) vs E3 (1000ppi, flats)
FNIR
FPIR
[Gallery size: 418 Probes: 809]
S
U
W

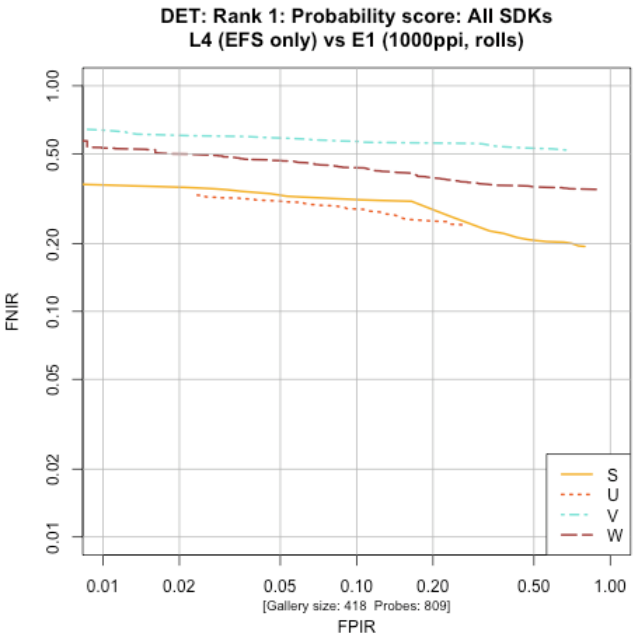

DET: Rank 1: Raw Score: All SDKs
L3 (image+EFS) vs E1 (1000ppi, rolls)
FNIR
FPIR
[Gallery size: 418 Probes: 809]
S
U
V
W

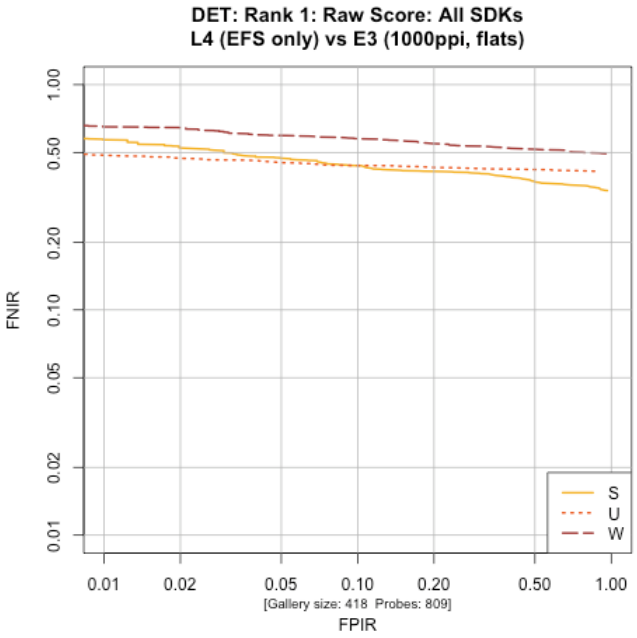

DET: Rank 1: Probability score: All SDKs
L3 (image+EFS) vs E1 (1000ppi, rolls)
FNIR
FPIR
[Gallery size: 418 Probes: 809]
S
U
V
W

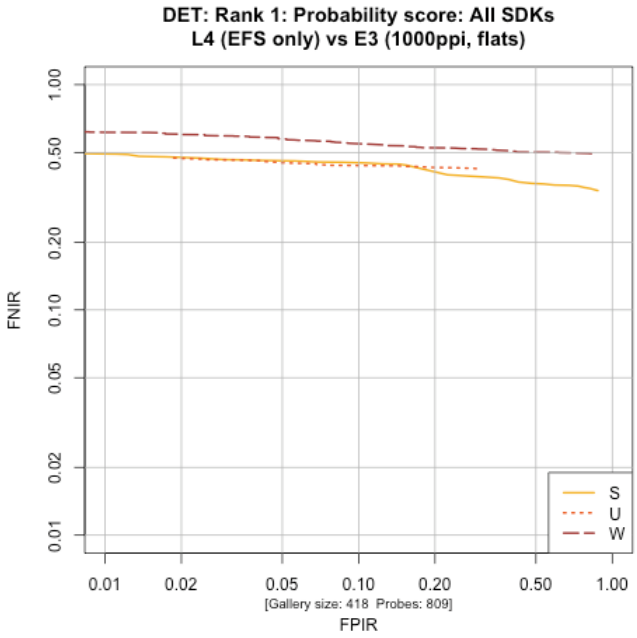

DET: Rank 1: Raw Score: All SDKs
L3 (image+EFS) vs E3 (1000ppi, flats)
FNIR
FPIR
[Gallery size: 418 Probes: 809]
S
U
W

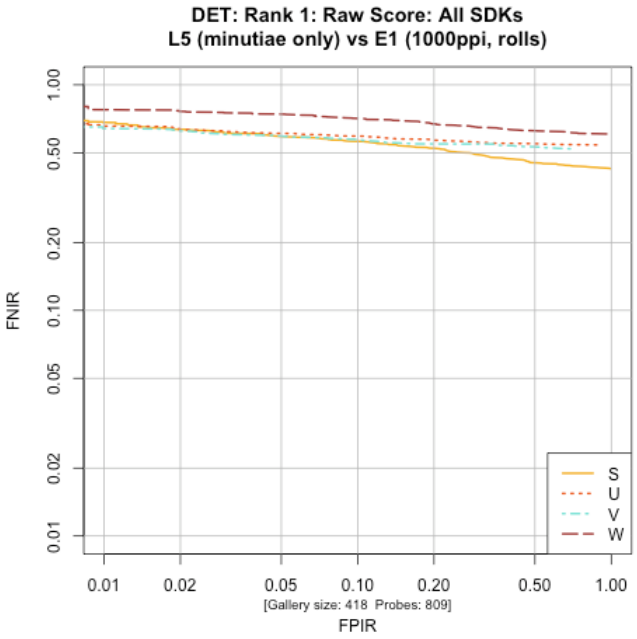

DET: Rank 1: Probability score: All SDKs
L3 (image+EFS) vs E3 (1000ppi, flats)
FNIR
FPIR
[Gallery size: 418 Probes: 809]
S
U
W

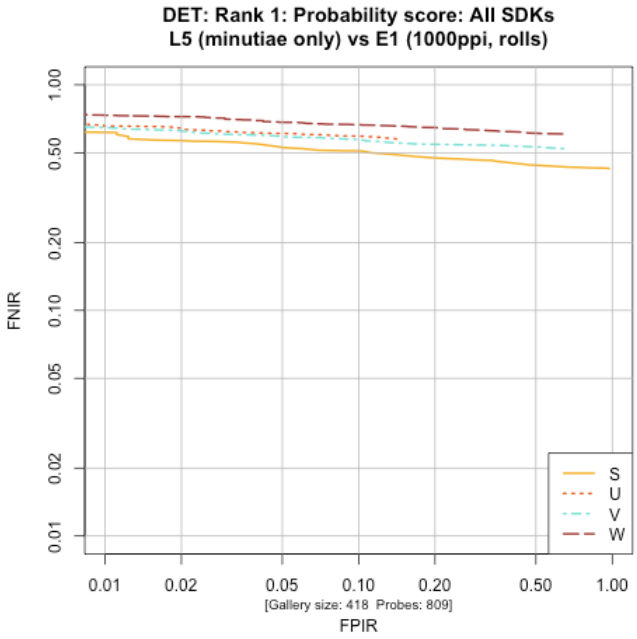

DET: Rank 1: Raw Score: All SDKs
L4 (EFS only) vs E1 (1000ppi, rolls)
FNIR
FPIR
[Gallery size: 418 Probes: 809]
S
T
U
V
W

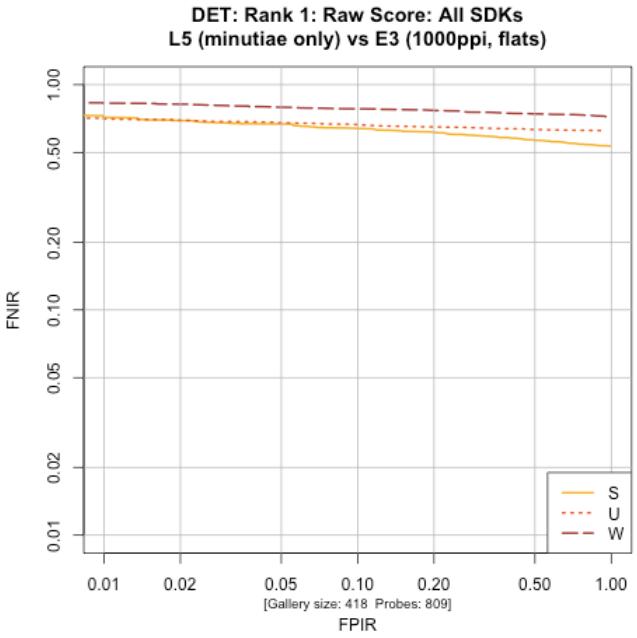

DET: Rank 1: Probability score: All SDKs
L4 (EFS only) vs E1 (1000ppi, rolls)
FNIR
FPIR
[Gallery size: 418 Probes: 809]
S
U
V
W

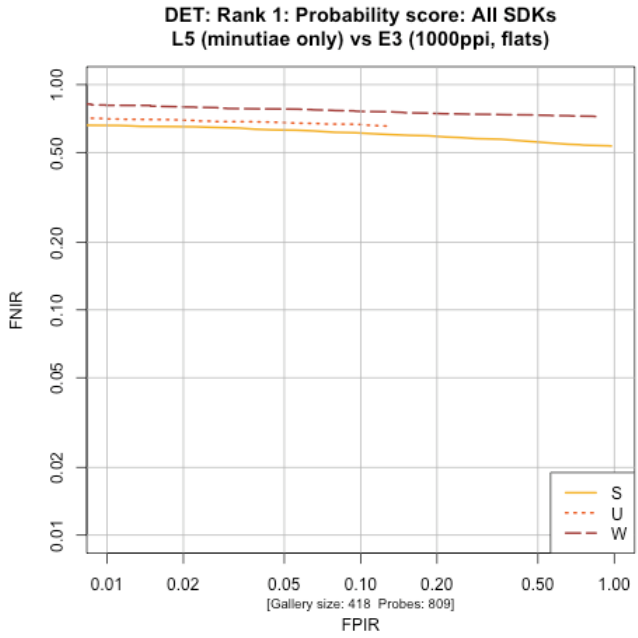

DET: Rank 1: Raw Score: All SDKs
L4 (EFS only) vs E3 (1000ppi, flats)
FNIR
FPIR
[Gallery size: 418 Probes: 809]
S
U
W


DET: Rank 1: Probability score: All SDKs
L4 (EFS only) vs E3 (1000ppi, flats)
FNIR
FPIR
[Gallery size: 418 Probes: 809]
S
U
W


DET: Rank 1: Raw Score: All SDKs
L5 (minutiae only) vs E1 (1000ppi, rolls)
FNIR
FPIR
[Gallery size: 418 Probes: 809]
S
U
V
W


DET: Rank 1: Probability score: All SDKs
L5 (minutiae only) vs E1 (1000ppi, rolls)
FNIR
FPIR
[Gallery size: 418 Probes: 809]
S
U
V
W





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	-

Appendix C

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

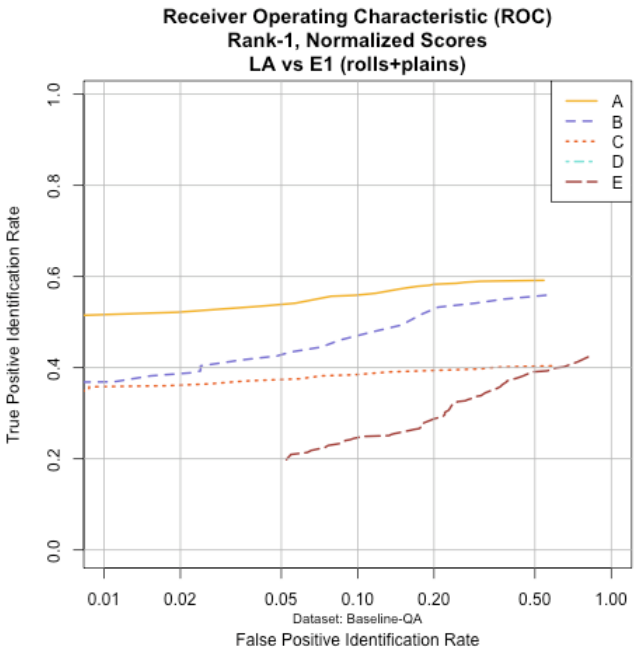
Additional Results

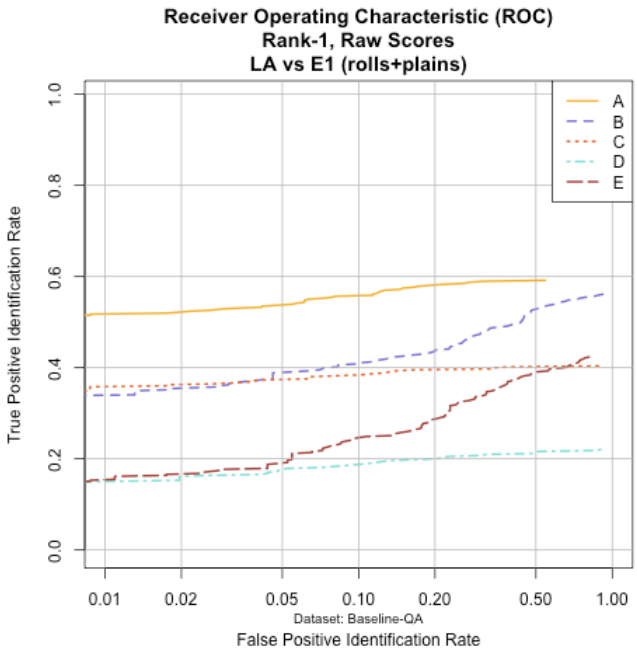
Contents

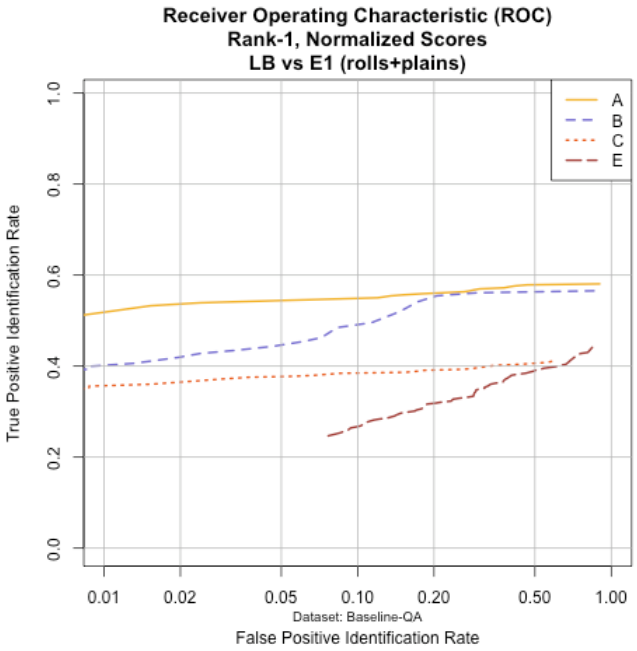
C-1 Raw score and probability score results

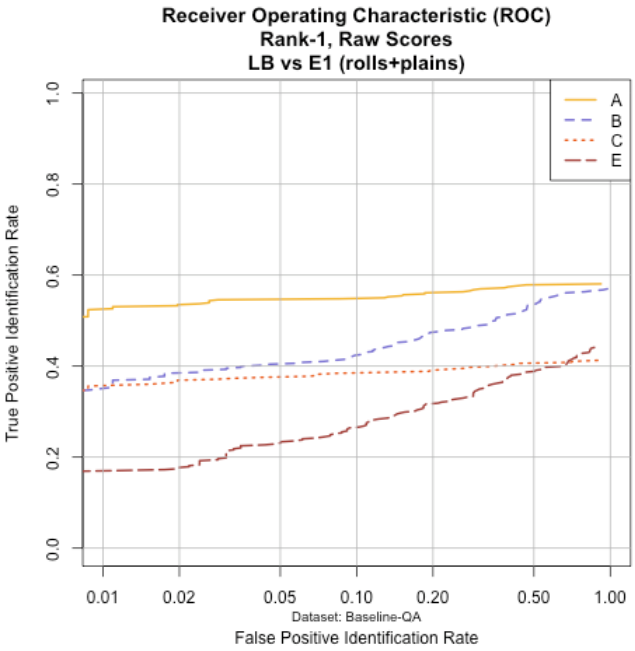
The following ROCs show the differences between the normalized "Probability" scores and the raw scores. In each case, participants returned a raw score and a normalized score estimating the probability of a match. In no cases did the raw scores provided better results than the normalized scores, though matcher D notably is better able to populate the range of FPIR values using raw scores. In each case, the probability score is on the left and the corresponding raw score on the right.

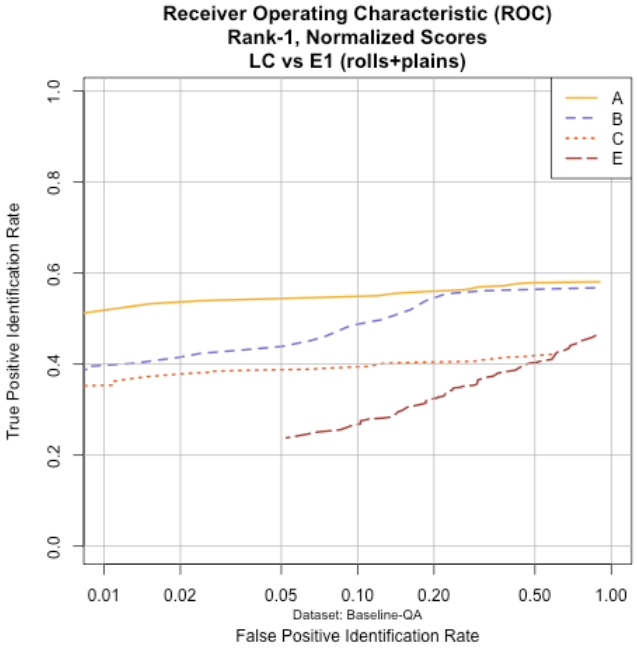
C-1.1 Baseline-QA dataset

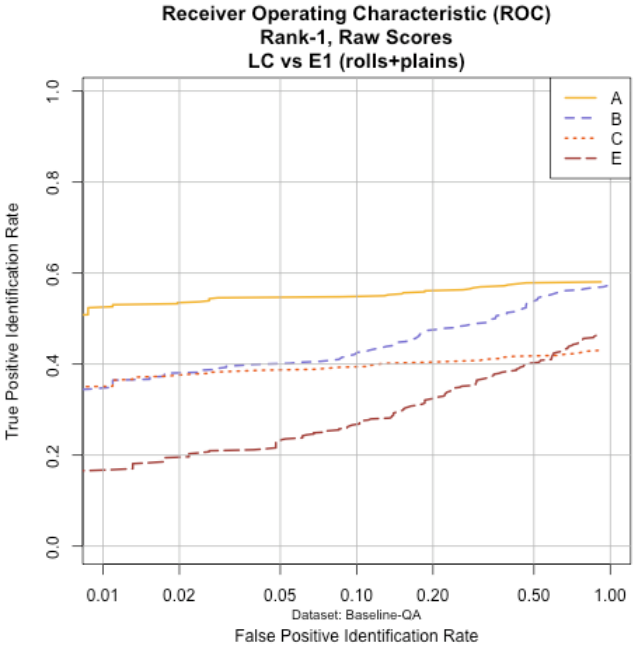


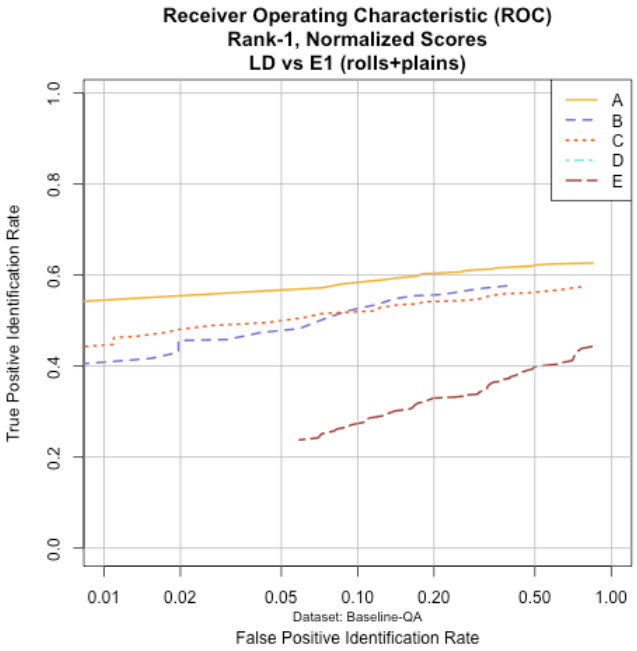


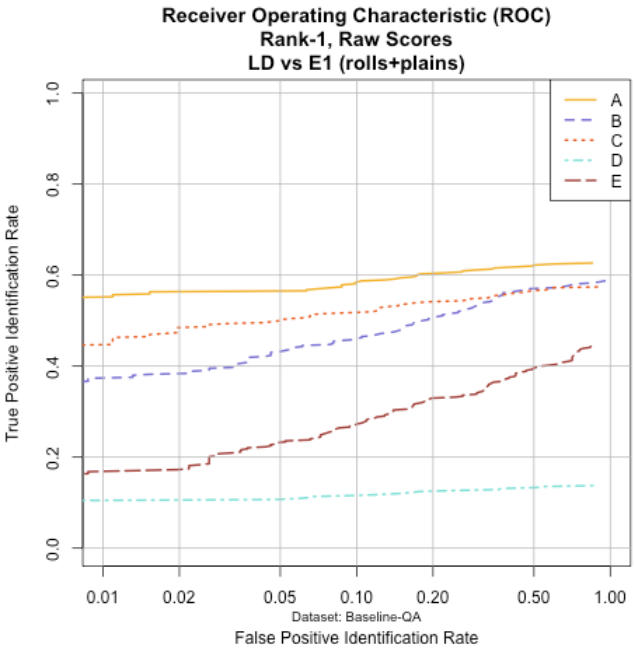

Receiver Operating Characteristic (ROC)
Rank-1, Normalized Scores
LB vs E1 (rolls+plains)
True Positive Identification Rate
False Positive Identification Rate
Dataset: Baseline-QA
A
B
C
E
0.0
0.2
0.4
0.6
0.8
1.0
0.01
0.02
0.05
0.10
0.20
0.50
1.00

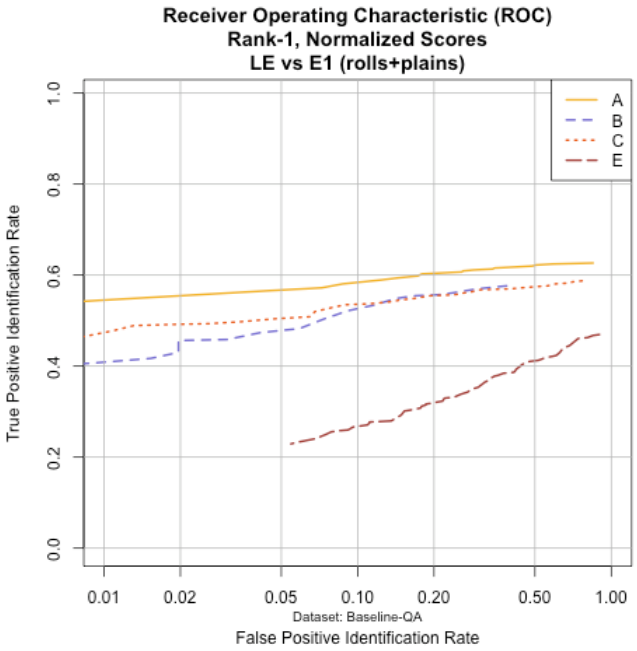

Receiver Operating Characteristic (ROC)
Rank-1, Raw Scores
LB vs E1 (rolls+plains)
True Positive Identification Rate
False Positive Identification Rate
Dataset: Baseline-QA
A
B
C
E
0.0
0.2
0.4
0.6
0.8
1.0
0.01
0.02
0.05
0.10
0.20
0.50
1.00

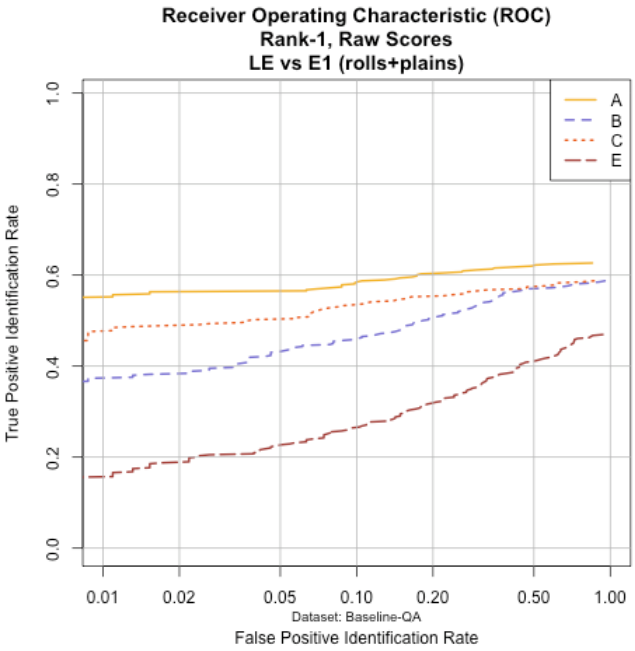

Receiver Operating Characteristic (ROC)
Rank-1, Normalized Scores
LC vs E1 (rolls+plains)
True Positive Identification Rate
False Positive Identification Rate
Dataset: Baseline-QA
A
B
C
E
0.0
0.2
0.4
0.6
0.8
1.0
0.01
0.02
0.05
0.10
0.20
0.50
1.00

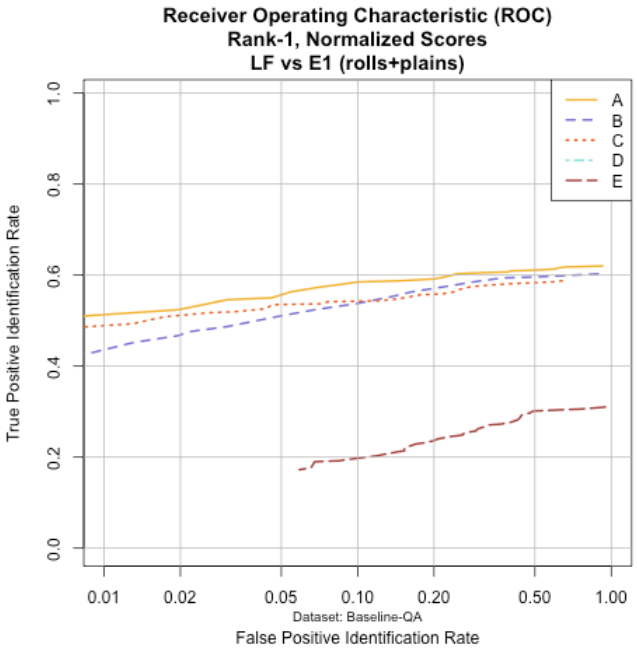

Receiver Operating Characteristic (ROC)
Rank-1, Raw Scores
LC vs E1 (rolls+plains)
True Positive Identification Rate
False Positive Identification Rate
Dataset: Baseline-QA
A
B
C
E
0.0
0.2
0.4
0.6
0.8
1.0
0.01
0.02
0.05
0.10
0.20
0.50
1.00

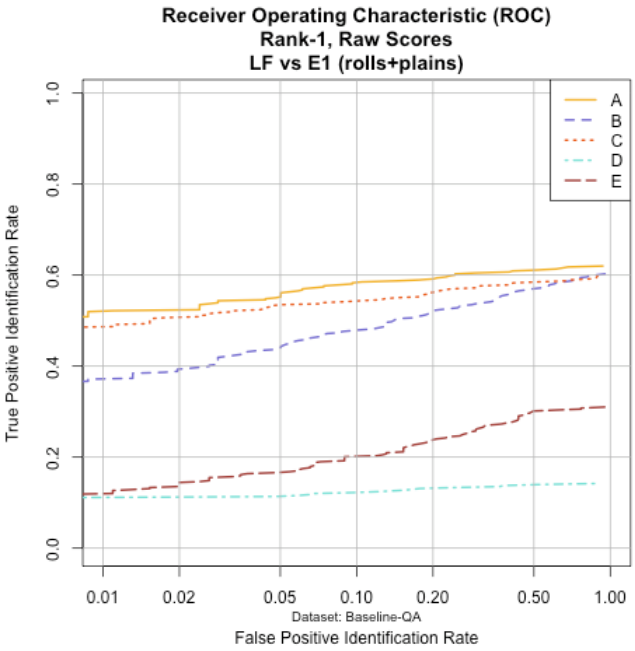

Receiver Operating Characteristic (ROC)
Rank-1, Normalized Scores
LD vs E1 (rolls+plains)
True Positive Identification Rate
False Positive Identification Rate
Dataset: Baseline-QA
A
B
C
D
E

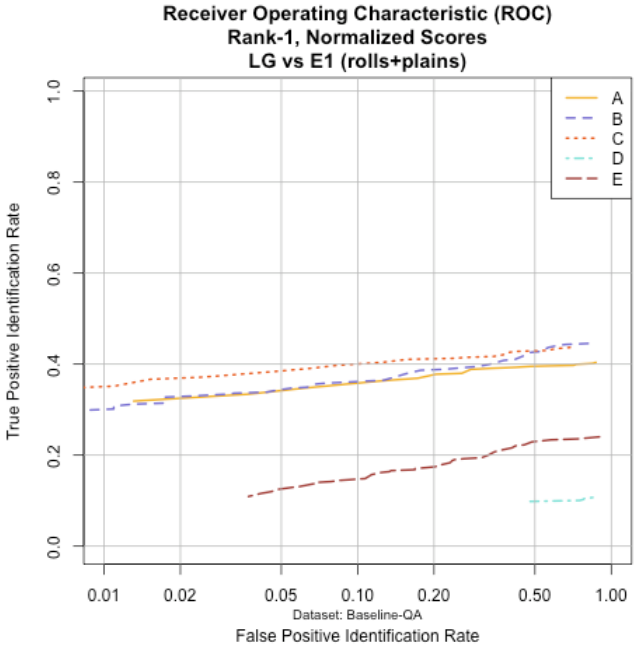

Receiver Operating Characteristic (ROC)
Rank-1, Raw Scores
LD vs E1 (rolls+plains)
True Positive Identification Rate
False Positive Identification Rate
Dataset: Baseline-QA
A
B
C
D
E

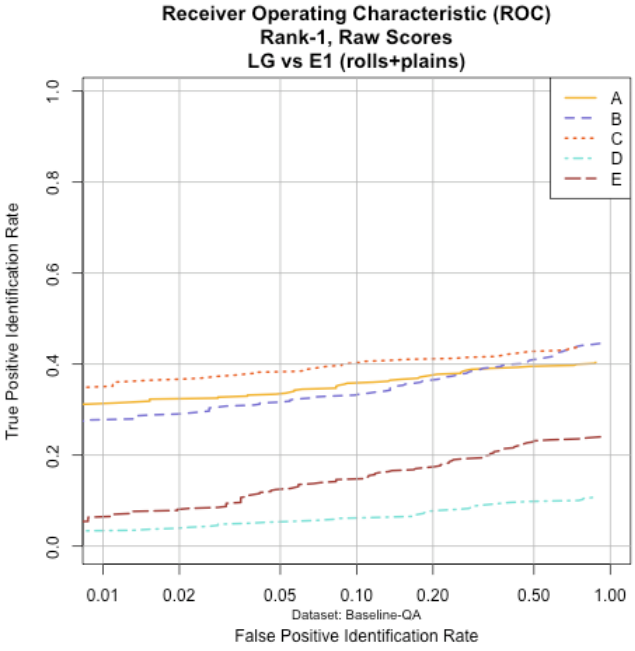

Receiver Operating Characteristic (ROC)
Rank-1, Normalized Scores
LE vs E1 (rolls+plains)
True Positive Identification Rate
False Positive Identification Rate
Dataset: Baseline-QA
A
B
C
E


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

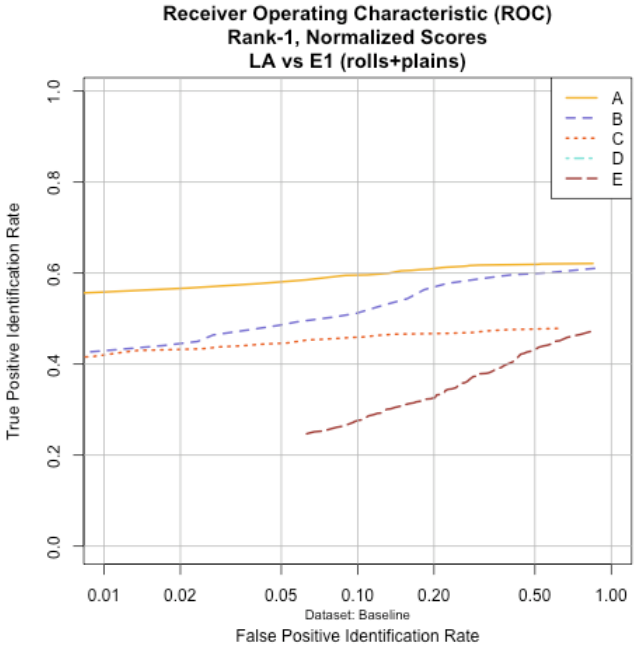

Receiver Operating Characteristic (ROC)
Rank-1, Normalized Scores
LF vs E1 (rolls+plains)
True Positive Identification Rate
False Positive Identification Rate
Dataset: Baseline-QA
A
B
C
D
E
0.0
0.2
0.4
0.6
0.8
1.0
0.01
0.02
0.05
0.10
0.20
0.50
1.00

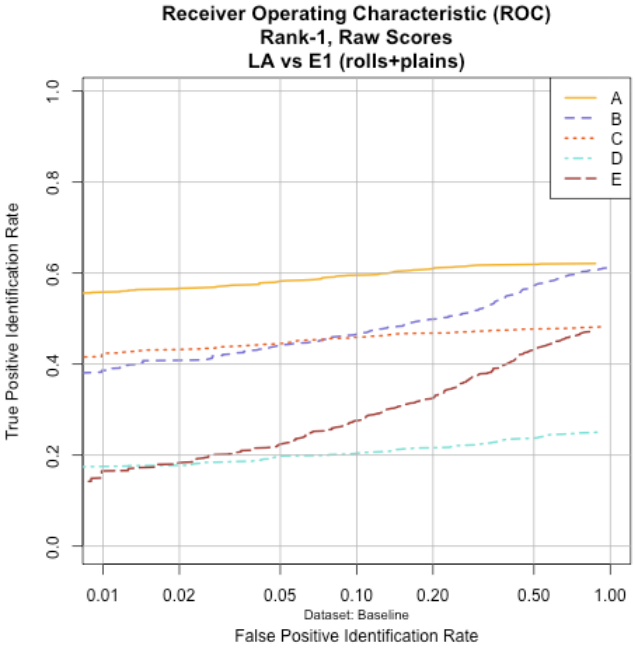

Receiver Operating Characteristic (ROC)
Rank-1, Raw Scores
LF vs E1 (rolls+plains)
True Positive Identification Rate
False Positive Identification Rate
Dataset: Baseline-QA
A
B
C
D
E
0.0
0.2
0.4
0.6
0.8
1.0
0.01
0.02
0.05
0.10
0.20
0.50
1.00

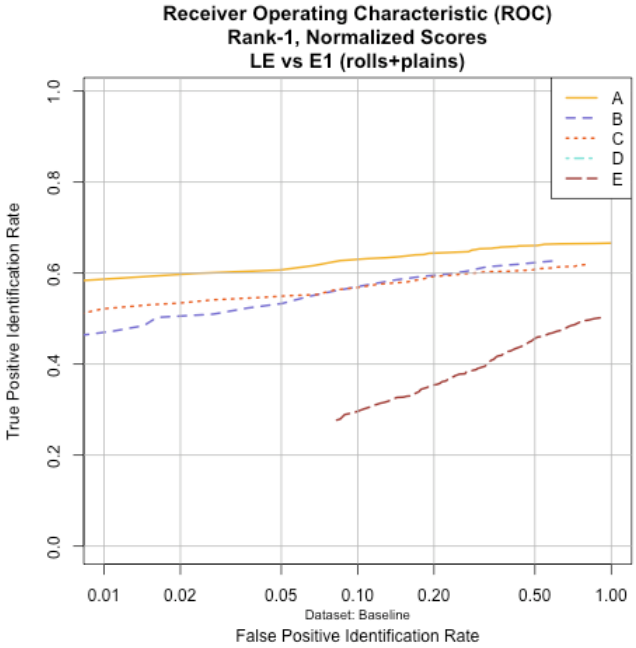

Receiver Operating Characteristic (ROC)
Rank-1, Normalized Scores
LG vs E1 (rolls+plains)
True Positive Identification Rate
False Positive Identification Rate
Dataset: Baseline-QA
A
B
C
D
E
0.0
0.2
0.4
0.6
0.8
1.0
0.01
0.02
0.05
0.10
0.20
0.50
1.00

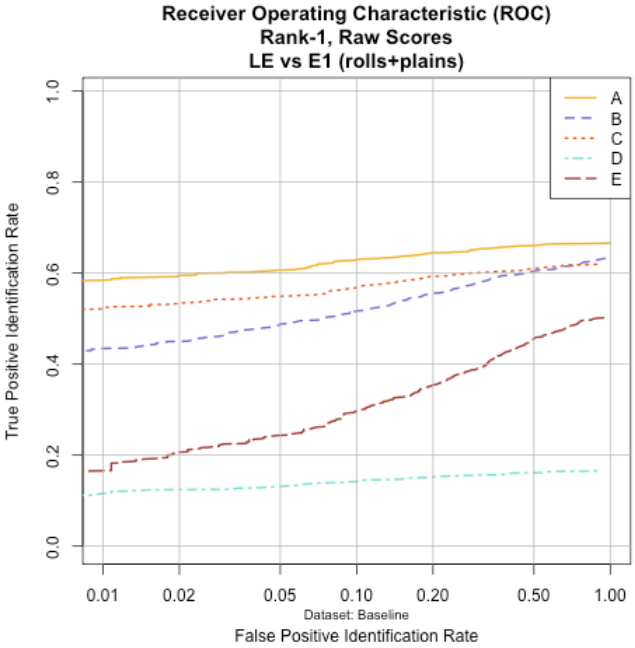

Receiver Operating Characteristic (ROC)
Rank-1, Raw Scores
LG vs E1 (rolls+plains)
True Positive Identification Rate
False Positive Identification Rate
Dataset: Baseline-QA
A
B
C
D
E
0.0
0.2
0.4
0.6
0.8
1.0
0.01
0.02
0.05
0.10
0.20
0.50
1.00

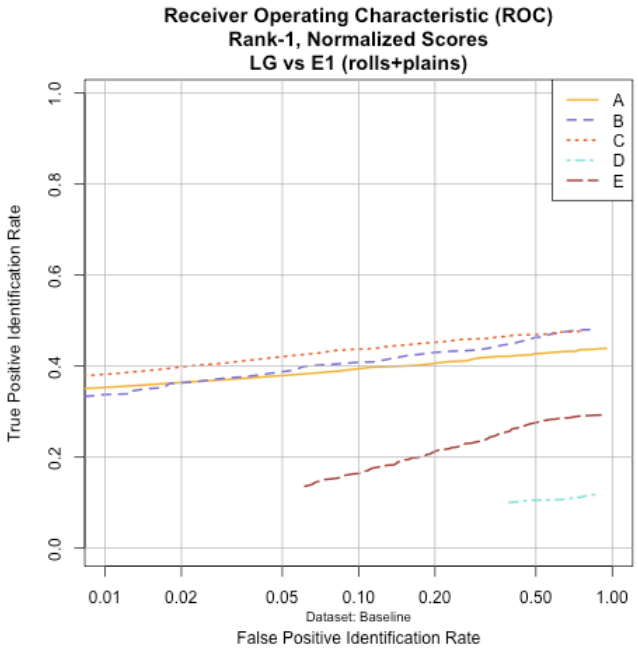
Baseline Dataset

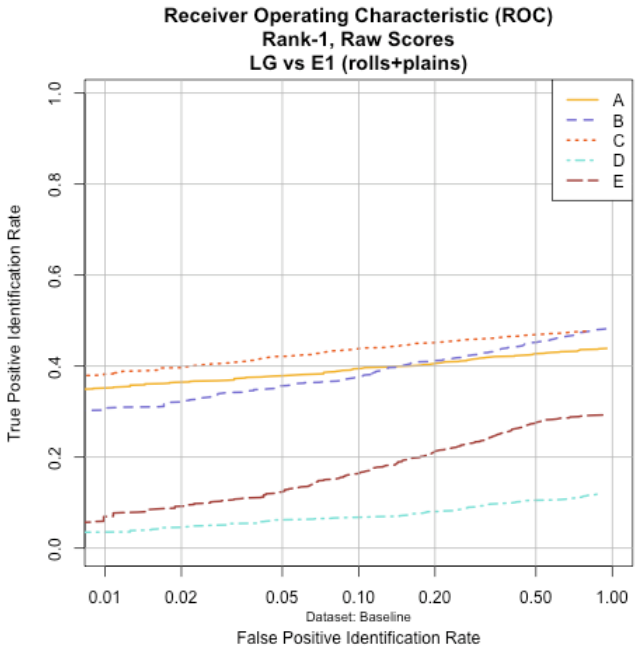






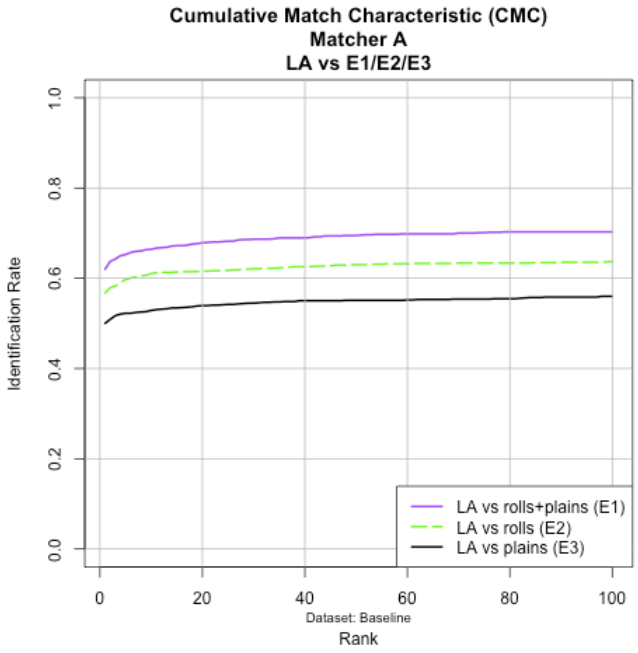


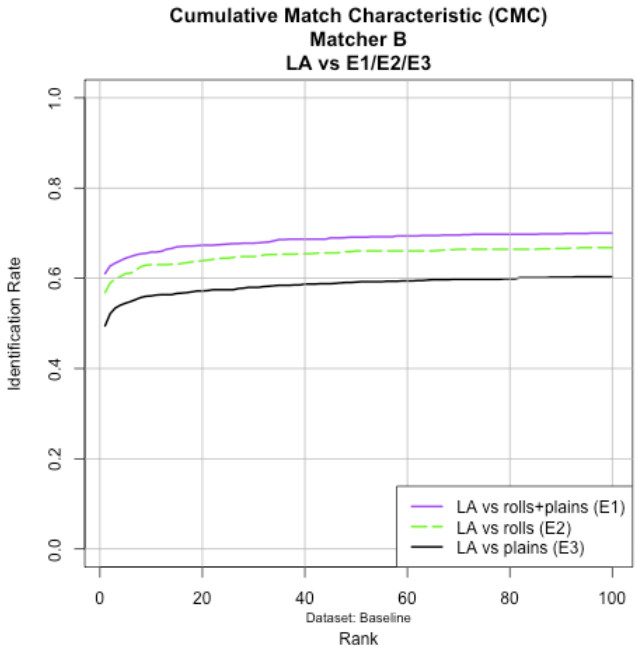


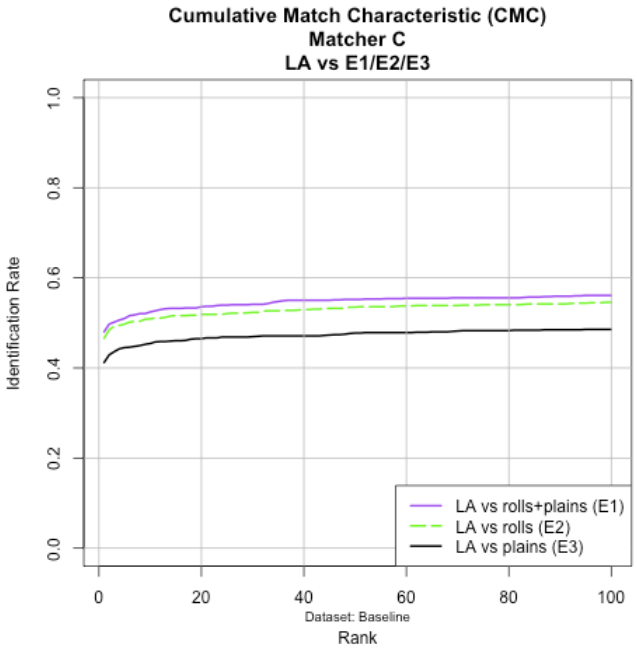


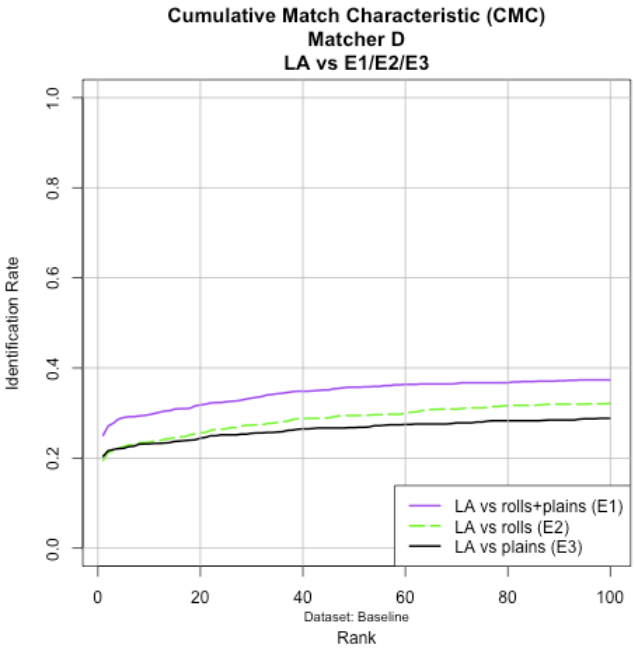
C-2 CMC results comparing searches of Galleries E1, E2 and E3

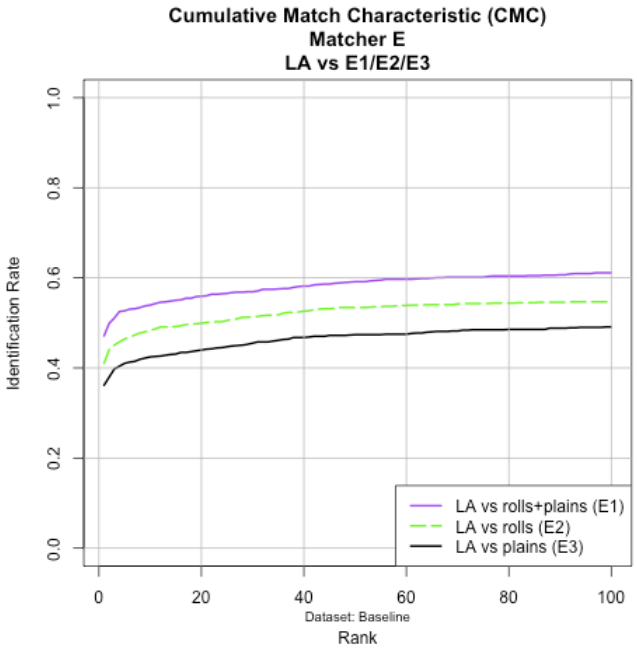
C-2.1 Image-only (LA) Results by Matcher, Baseline Dataset



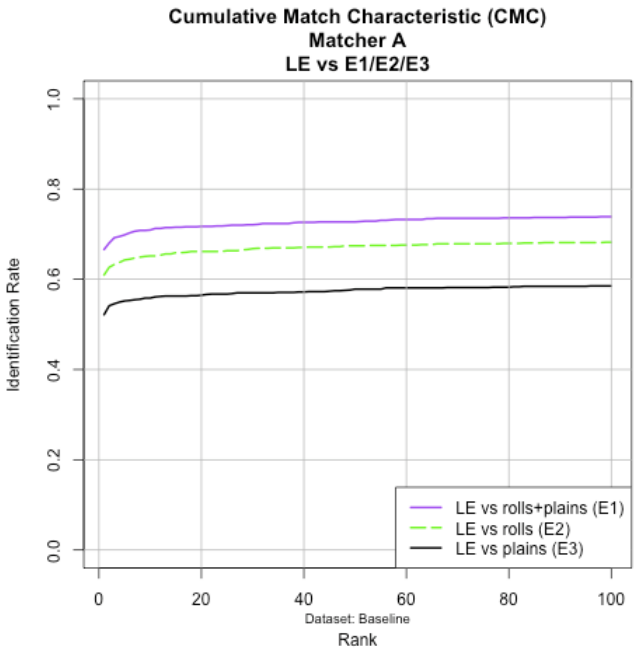


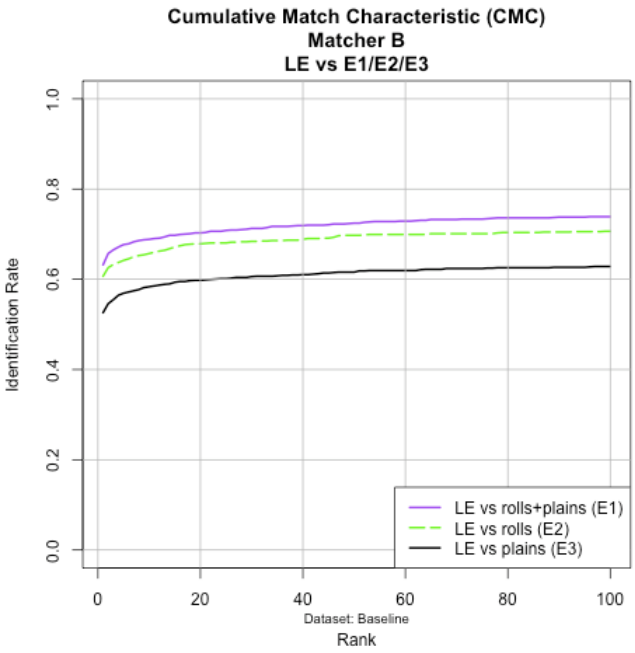

Cumulative Match Characteristic (CMC)
Matcher C
LA vs E1/E2/E3
Identification Rate
0.0
0.2
0.4
0.6
0.8
1.0
0
20
40
60
80
100
Dataset: Baseline
Rank
LA vs rolls+plains (E1)
LA vs rolls (E2)
LA vs plains (E3)

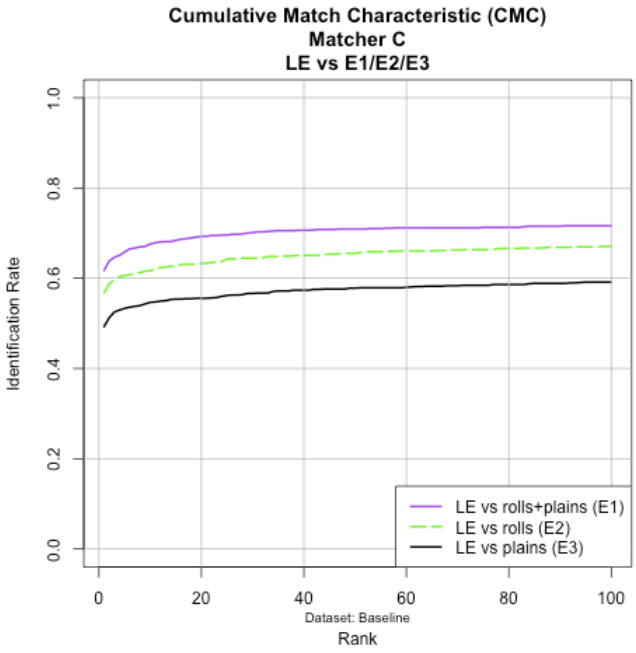

Cumulative Match Characteristic (CMC)
Matcher D
LA vs E1/E2/E3
Identification Rate
0.0
0.2
0.4
0.6
0.8
1.0
0
20
40
60
80
100
Dataset: Baseline
Rank
LA vs rolls+plains (E1)
LA vs rolls (E2)
LA vs plains (E3)

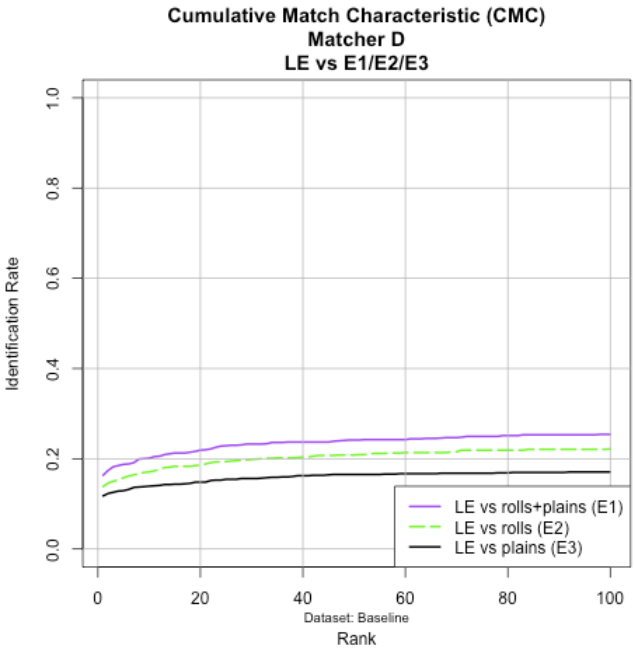

Cumulative Match Characteristic (CMC)
Matcher E
LA vs E1/E2/E3
Identification Rate
0.0
0.2
0.4
0.6
0.8
1.0
0
20
40
60
80
100
Dataset: Baseline
Rank
LA vs rolls+plains (E1)
LA vs rolls (E2)
LA vs plains (E3)

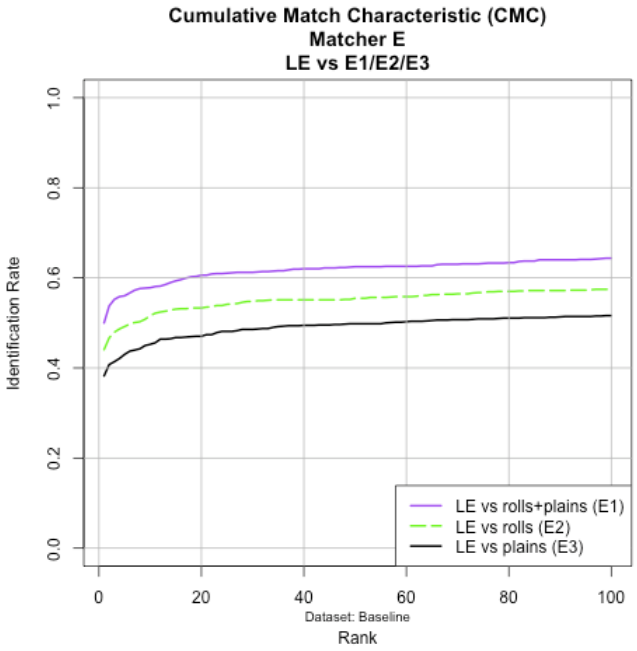
C-2.2 Image + full EFS (LE) Results by Matcher, Baseline Dataset



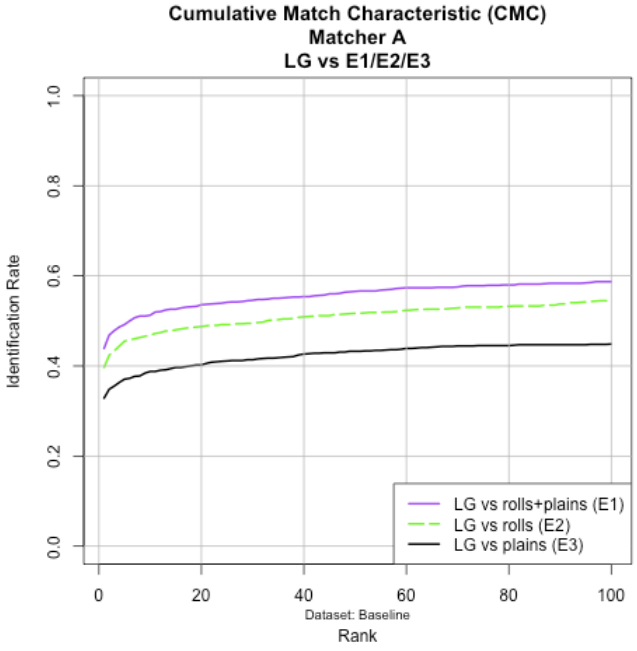


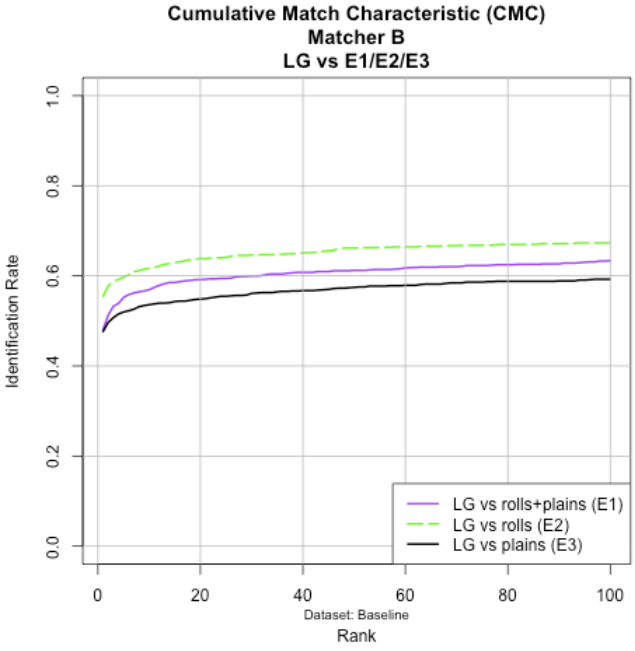


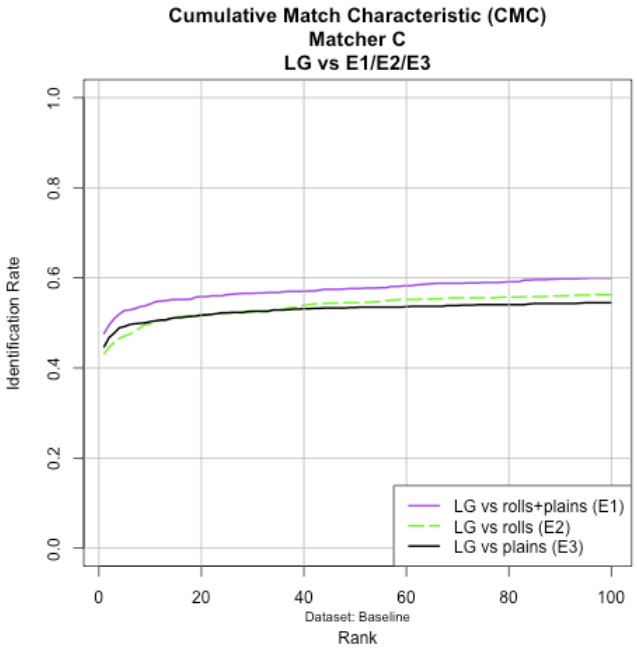


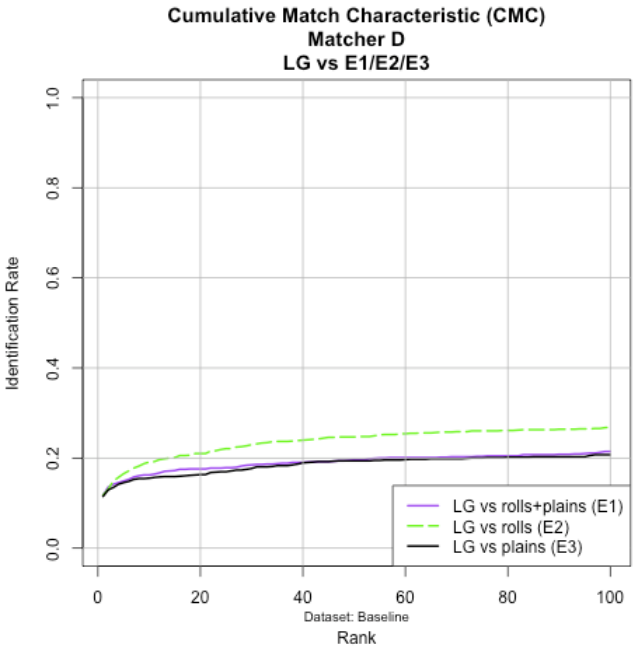


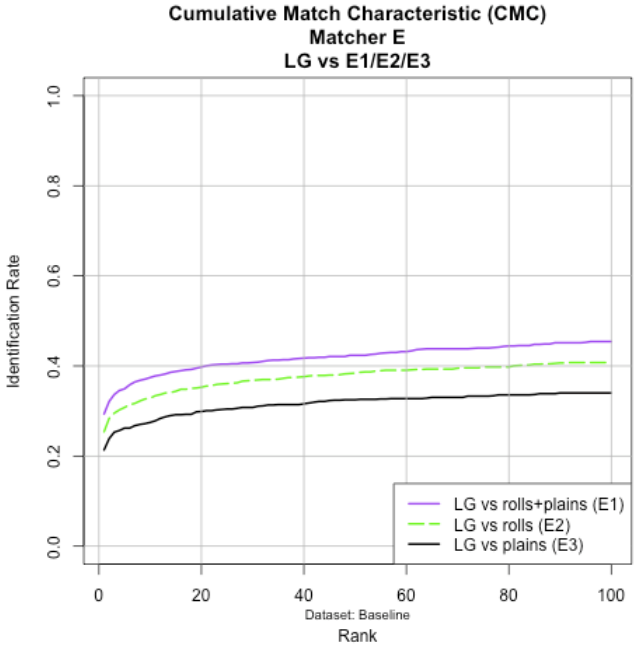
C-2.3 Minutiae Only (LG) Results by Matcher, Baseline Dataset






Cumulative Match Characteristic (CMC)
Matcher C
LG 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
LG vs rolls+plains (E1)
LG vs rolls (E2)
LG vs plains (E3)


Cumulative Match Characteristic (CMC)
Matcher D
LG 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
LG vs rolls+plains (E1)
LG vs rolls (E2)
LG vs plains (E3)


Cumulative Match Characteristic (CMC)
Matcher E
LG 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
LG vs rolls+plains (E1)
LG vs rolls (E2)
LG vs plains (E3)

C-3 Performance on latents removed due to markup issues

Summary of rank-1 hit rates for the latents removed due to markup issues (225 latents)

	Latent Subset		
	LA	**LE**	**LG**
	Image only	Image + EFS	Minutiae + Ridge Counts only (EBTS)
A	58.0	63.9	31.5
B	50.2	56.6	36.5
C	37.0	53.9	37.4
D	16.9	*11.4*[*]	7.76
E	37.0	45.7	22.4

C-4 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.[†]

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%	NA	NA	73%	NA	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%

[*] *Baseline-QA results for subset LD were used as proxy for Baseline-QA LE for the purpose of scoring Baseline LE*

[†] *This is sometimes known as the "Ray Moore statistic". AFIS pioneer Ray Moore observed that this tended to be about 83% at the time.*

B	90%	90%	81%
C	89%	89%	86%
D	78%	74%	69%
E	84%	83%	74%