
Face Recognition Vendor Test (FRVT)

Performance of Face Identification Algorithms

NIST Interagency Report 8009

Patrick Grother Mei Ngan

Information Access Division
National Institute of Standards and Technology



May 26, 2014

Acknowledgements

The authors would like to thank the sponsors of this activity. These are the Criminal Justice Information Systems (CJIS) division and the Biometric Center of Excellence (BCOE) of the Federal Bureau of Investigation, the Science and Technology (S&T) Directorate in the Department of Homeland Security (DHS), and the Office of Biometric Identity Management (OBIM) office, also in DHS.

Disclaimer

Specific hardware and software products identified in this report were used in order to perform the evaluations described in this document. In no case does identification of any commercial product, trade name, or vendor, 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.

Executive Summary

Purpose and scope: This report documents the performance of one-to-many face recognition algorithms, and compares it with that measured in 2010. Performance in this context refers to recognition accuracy and computational resource usage as measured by executing those algorithms on massive sequestered datasets. These are: reasonable quality law enforcement mugshot images; poor quality webcam images collected in similar detention operations; and moderate quality visa application images. The mugshot and visa images are used to approximate performance obtainable using high quality ISO standardized images collected in passport, visa and driving license duplicate detection operations. These applications constitute the largest revenue segment in the face recognition marketplace. The webcam images are included to show how recognition accuracy degrades in non-ideal poorly-controlled situations - results will mimic recognition in adverse e.g. surveillance situations to the extent that those applications produce images similar to the webcam set.

Out of scope: Not within the scope of this report are: performance of live transactional systems like automated border control gates; human recognition accuracy as used in forensic applications; and recognition of persons in video sequences (which NIST is evaluating separately and will report on later). Some of those applications are likely to share technologies that *are* tested in this report.

Participation: The report includes performance figures for prototype algorithms from the research laboratories of many of the major commercial suppliers of face recognition technologies. It thereby facilitates robust comparative evaluation. However, while participation in the test was open to any organization worldwide, neither social media companies nor most academic institutions elected to submit algorithms, and this report therefore does not capture their technical capabilities except to the extent that those technologies have been adopted or licensed by FRVT participants.

Background: Face recognition error rates have declined massively in the two decades since initial commercialization of the various technologies. NIST has tracked that improvement and its conduct of regular independent, free, open, and public evaluations has fostered improvements in the state of the art. One-to-many face identification systems are mostly used in conjunction with trained human facial reviewers. The systems are configured to operate in two regimes: first, with a low threshold that necessitates adjudication of many false positive candidates by a reviewer specifically employed to do so; second, with a high threshold, in which false positive outcomes are rare and human intervention is only needed infrequently. Low false positive rates are accompanied by higher false negative rates - this report includes extensive quantification of this tradeoff.

Results summary: Since NIST's last evaluation was published in August 2010, the algorithms from NEC remain the most accurate followed by those of Morpho which merged its algorithms with those acquired from L1 Identity Solutions in 2011. Thereafter Toshiba, Cognitec Systems, and 3M/Cogent constitute the leading commercial suppliers. Algorithms with lesser levels of capability are those from Neurotechnology, Zhuhai Yisheng, HP, Decatur, and Ayonix. Importantly, however, performance is not single-faceted and any ranking of performance across algorithms must be weighed by application-specific requirements. For example, some algorithms are more suited to recognition of difficult webcam images; and the search speed of some algorithms increases only slowly with enrolled population size.

The headline results follow in the Technical Summary.

A = 3M/Cogent	B = Cognitec	C = Neurotechnology	D = Safran Morpho	E = NEC	F = Tsinghua U.	FNIR(N,R,T,L) "Miss rate"
G = Hisign	H = CAS-IA	I = CAS-ICT	J = Toshiba	L = Tsinghua U. II	M = HP	FPIR(N,T,L) "False alarm rate"
P = Zhuhai-Yisheng	Q = JunYu	S = Decatur	T = Ayonix			

Technical Summary

Absolute accuracy: When **mugshot images** from 1.6 million individuals are enrolled by the most accurate algorithm, 4.1% of subsequent mated searches fail to yield the correct mate in rank one position. Practically this result assumes that a human reviewer will be employed to adjudicate the candidate identities. If, on the other hand, a threshold is elevated to limit false positive outcomes to only 1 in 500 searches (0.2%), the failure to find the mate rises to 7.5%. Using **poorly constrained webcam images**, which exhibit serious departures from most quality-related clauses of published image standards, identification miss rates are typically between two and five times higher such that the correct mate is not found at rank one in 20-60% of searches. Exceptionally, however, the most accurate algorithm fails in only 11.3% of searches. This latter result is notable in that it indicates that face recognition can work on non-ideal images. Thus, while the webcam images were collected from nominally cooperative subjects, the use of inferior equipment and procedures means, for example, that images from non-cooperative bank ATM machine and surveillance camera deployments will be recognizable in a useful number of cases. Such outcomes have been reported operationally [19] and in the laboratory [17].



Figure 1: Examples of images used in this report.

Accuracy across commercial providers: Recognition accuracy is very strongly dependent on the algorithm and, more specifically, on the developer of the algorithm. Recognition error rates in a particular scenario range from a few percent up to beyond fifty percent. Among the most accurate developers, the rank one miss rates for recognition in a population size of 1.6 million are 4.1% (NEC), 9.1% (Morpho), 10.7% (Toshiba), 13.6% (Cognitec), 17.2% (3M) and 20.5% (Neurotechnology). For webcam images, this sequence is 11.3% (NEC), 23.7% (Toshiba), 29.8% (Morpho), 36.4% (3M), 57.6% (Cognitec) and 66.9% (Neurotechnology). While results for up to six algorithms from each developer are reported here, the intra-provider accuracy variations are usually smaller than the inter-provider variations. That said, some developers submitted different, less accurate but computationally lightweight algorithms.

Accuracy under increasing population size: As more identities are enrolled into a biometric system, the possibility of a false positive increases due to lookalike faces that yield extreme values in the tail of the nonmate score distribution. However these scores are lower than most mate scores such that when an identification is configured with a threshold of zero, and where human adjudication is always necessary, rank-one identification miss rates scale very favorably with population size, N , growing approximately as a power law, aN^b . Depending on the algorithm, the exponent b for mugshot searches is low, on the range $[0.08, 0.16]$ meaning that a large, 10-fold, increase in N , yields only a 1.2 to 1.4 fold increase in the rate at which mated searches do not yield the correct mate. Thus the rank one mugshot miss rates, for search into a database of just 160,000, are only modestly improved over those at 1.6 million: 4.1% to 3.4% (NEC), 9.1% to 7.6% (Morpho), 10.7% to 7.9% (Toshiba), 13.6% to 10.9% (Cognitec), 17.2% to 13.3% (3M) and 20.5% to 16.9% (Neurotechnology). While extrapolation to larger populations is technically problematic, face identification systems will be useful, if imperfect, in nation-state population sizes.

Utility of adjudicating long candidate lists: In the regime where a system is configured with a threshold of zero, and where human adjudication is always necessary, the reviewer will find mates on candidate lists at ranks far from one. The accuracy benefits of traversing such lists are usually substantial: For example, in a population of 1.6 million, the rank one and rank fifty miss rates are 4.1% to 2.6% (NEC), 9.1% to 7.1% (Morpho), 10.7% to 5.7% (Toshiba), 13.6% to 8.4% (Cognitec), 17.2% to 9.8% (3M) and 20.5% to 13.6% (Neurotechnology). These are diminishing returns, however, with miss rates for some algorithms growing as power-law, aR^c , in the number of candidates a reviewer is willing to consider, R . The exponent c for mugshot searches is typically on the range $[-0.2, -0.1]$ indicating that rank 1 miss rates are reduced by from 20% to 40% if 10 candidates are available for inspection, but only by 30% to 50% if 50 are considered.

Human adjudication workload: Human reviewers will typically only need to search the first few (highest scoring) candidates returned in a search. The expected number of comparisons constitutes a workload measure which can be used to both compare algorithms and to inform operational labor requirements. If a score threshold is applied to reduce the

A = 3M/Cogent	B = Cognitec	C = Neurotechnology	D = Safran Morpho	E = NEC	F = Tsinghua U.	FNIR(N,R,T,L) "Miss rate"
G = Hisign	H = CAS-IA	I = CAS-ICT	J = Toshiba	L = Tsinghua U. II	M = HP	FPIR(N,T,L) "False alarm rate"
P = Zhuhai-Yisheng	Q = JunYu	S = Decatur	T = Ayonix			

length of candidate lists, substantial reviewer workload reductions can be realized but at the expense of increased miss rates. In the best case, application of a threshold to candidates from a Morpho algorithm reduced workload by 60% with only a factor of 1.05 more misses than without a threshold.

Accuracy by age group: Identification accuracy is strongly dependent on subject age. For all algorithms, older individuals are both easier to recognize as themselves, and easy to tell apart from each other. The opposite is true in children: both false negative *and* false positive rates are much higher, with infants being very hard to identify. Moreover, the trends are progressive throughout adulthood, with young adults being identified with worse accuracy than older. These results are derived over images taken, on average, 2-3 years apart so the effects of ageing (craniofacial shape change) are influential only on the results for younger individuals.

Accuracy gains 2010-2013: For the four developers who submitted algorithms to NIST in 2010 and 2013, accuracy has improved in all cases. Rank one miss rates have reduced by about 10% for Cognitec, Neurotechnology, and Morpho, and by about 28% for NEC (from 8.9% to 6.4%). More substantial reductions have been realized when a threshold is applied to permit false positive outcomes in only 1 in 500 nonmate searches: There Morpho realizes a 21% reduction in misses (from 24.8% to 19.5%), while for NEC, the figure is 60% (26.9% to 10.8%). Finally the accuracy gains when reviewers examine up to 50 candidates are more modest, ranging from 3% (Cognitec) to 12% (Neurotechnology).

Sketch recognition: By searching a non-operational set of sketch images against photographs seeded into a population of 640,000 nonmated mugshots, the most accurate algorithms produce the mated photograph only infrequently: The mate is not among the top 50 candidates at the following rates: 73.3% (3M/Cogent), 73.8% (NEC), 78.5% (Toshiba), 80.3% (Morpho), and 81.5% (Neurotechnology). While these error rates are very high, they are nevertheless valuable in developing investigative leads in cases which are otherwise cold. An important caveat is that sketch-identification was never declared to be part of FRVT and better algorithms may be available from the providers. That said, face recognition algorithms *are* being used to recognize sketches operationally. Further accuracy will clearly be dependent on eye-witness recall, artist interpretation if any, and software interfaces.

Conclusions: As with other biometrics, accuracy of facial recognition implementations varies greatly across the industry. Absent other performance or economic parameters, users should prefer the most accurate algorithm. Note, however, that the results of this section are entirely rank-based befitting use of face recognition in the investigational mode in which an reviewer is willing to traverse candidate lists looking for mates. Subsequent investigations in this report consider threshold-based metrics appropriate for identification mode applications. Note that the absolute values of identification accuracy will always depend on the dataset used, specifically to the properties of the images in use. In particular, the main dataset used here includes some images that are not perfectly frontal, such that conformance to the appearance-related requirements of the ISO/IEC 19794-5 “gold” standard is imperfect.

Improving face recognition: On the basis of the results in this report, we identify the following drivers of overall recognition accuracy, and quantify their relative influence relative on false negative miss rates.

- ▷ **Quality:** Improvement of image quality is the largest contributing factor to recognition accuracy. Results in this report note a four fold reduction in miss rates using mugshots vs. webcam images. Further improvements in accuracy can be obtained by enhancing conformance to the ISO/IEC 19794-5 “gold” standard [26], particularly using two complementary approaches: *by design* via improved optical and photographic aspects, and by careful application of compression algorithms; and *by detection* of non-conformant (for example, blurred or non-frontal) images at the time they are collected. A new program has been established to evaluate algorithms capable of detecting defective images [9, 10]. The best practice should be to collect and retain forensic quality photographs (i.e. subject acquisition profile 50/51 instances of the ANSI/NIST Type 10 standard [27]), from which ISO/IEC 19794-5 images should be prepared for automated recognition.
- ▷ **Human adjudication of candidate lists:** Recognizing that a human is often involved in examining candidates produced in an automated one-to-many search, it is imperative that the enrolled reference “exemplar” image be of high quality. Given ready availability of high resolution digital cameras, the ANSI/NIST standard advocates collection of forensic quality images for which the interocular distance is around 800 pixels. Such images are not used directly for automated face recognition, so the default guidance is to collect and retain the forensic image, and to prepare from it the ISO standard image (120 pixels interocular distance) for automated face recognition.

Accuracy can be improved by supplementing automated facial identification with human adjudication of the can-

A = 3M/Cogent	B = Cognitec	C = Neurotechnology	D = Safran Morpho	E = NEC	F = Tsinghua U.	FNIR(N,R,T,L) “Miss rate”
G = Hisign	H = CAS-IA	I = CAS-ICT	J = Toshiba	L = Tsinghua U. II	M = HP	FPIR(N,T,L) “False alarm rate”
P = Zhuhai-Yisheng	Q = JunYu	S = Decatur	T = Ayonix			

didate photographs. Particularly by traversing long candidate lists (for example, up to length 50) miss rates can readily be reduced by a factor of two (i.e. 50%) An important caveat here is that the accuracy with which human reviewers can reliably adjudicate the most-similar faces returned in a large-population one-to-many search remains poorly quantified.

- ▷ **Algorithm selection:** Algorithm performance varies substantially - even among the leading providers a factor of two reduction in identification miss rates can be realized by replacing one algorithm with a better one. Such a step should be undertaken only while paying due attention to other performance related factors, such as computational expense and impostor distribution stability.

Release Notes

- ▷ **FRVT Tracks:** NIST initiated FRVT in the second half of 2012. We invited participation in five tracks.
 - *Class A:* Accuracy of algorithms executing one-to-one verification comparisons to determine if two samples originate from the same person or not.
 - *Class B:* Accuracy of algorithms executing one-to-one verification but with an enrollment database present. This track was discontinued after the 2010 evaluation. Accuracy gains over class A are available.
 - *Class C:* Accuracy of algorithms executing one-to-many identification searches to determine either that the person is not enrolled, or to determine the identity of the person.
 - *Class D:* Accuracy of algorithms tasked with determining the sex or age of a person in one or more input images. A separate class D track tasked algorithms to determine whether a face in an image is frontal to the camera or not.
 - *Class F:* Effectiveness of algorithms that take one or more (non-frontal) input images of a person, and produce one (or more) frontally posed images of that person.
 - *Class V:* Effectiveness of algorithms that execute one-to-many identification of persons whose faces appear in frames extracted from video surveillance sequences.

This report details results only for class C algorithms.

- ▷ **FRVT Reports:** The results of the FRVT appear as a series of NIST Interagency Reports. The reports were developed separately and released on different schedules. In prior years NIST has mostly reported FRVT results as a single report; this had the disadvantage that results from completed sub-studies were not published until all other studies were complete.

All reports are linked from <http://face.nist.gov/frvt> and its sub-pages.

- ▷ **Appendices:** This report is accompanied by a number of appendices which present exhaustive results on a per-algorithm basis. These are machine-generated and are included because the authors believe that visualization of such data is broadly informative and vital to understanding the context of the report.
- ▷ **Typesetting:** Virtually all of the tabulated content in this report was produced automatically. This involved the use of scripting tools to generate directly type-settable L^AT_EX content. This improves timeliness, flexibility, maintainability, and reduces transcription errors.
- ▷ **Graphics:** Many of the Figures in this report were produced using both Deepayan Sarkar's Lattice package [22] and Hadley Wickham's ggplot2 package running under R, the capabilities of which extend beyond those evident in this document.
- ▷ **Contact:** Correspondence regarding this report should be directed to PGROTHER at NIST dot GOV.

Contents

ACKNOWLEDGEMENTS	1
DISCLAIMER	1
EXECUTIVE SUMMARY	2
TECHNICAL SUMMARY	3
RELEASE NOTES	6
1 INTRODUCTION	11
2 PARTICIPATION	11
3 EVALUATION DATASETS	12
3.1 IMAGE TYPES	12
3.2 ENROLMENT TYPES	14
3.3 LIMITATIONS OF THE DATA	15
4 PERFORMANCE METRICS	16
4.1 QUANTIFYING FALSE ALARMS	16
4.2 QUANTIFYING HITS AND MISSES	17
4.3 BEST PRACTICE TESTING REQUIRES EXECUTION OF SEARCHES WITH AND WITHOUT MATES	18
4.4 FAILURE TO EXTRACT	18
4.5 TIMING MEASUREMENT	19
4.6 UNCERTAINTY ESTIMATION	19
5 RESULTS	20
5.1 COMPARATIVE ACCURACY	20
5.2 ACCURACY DEPENDENCE ON RANK	26
5.3 EFFECT OF POPULATION SIZE	27
5.4 EFFECT OF ENROLLING ALL HISTORICAL IMAGES	34
5.5 ACCURACY DEPENDENCE ON SUBJECT AGE	36
5.6 ACCURACY OF SKETCH RECOGNITION	39
5.7 HUMAN WORKLOAD FOR CANDIDATE LIST ADJUDICATION	41
5.8 IMPOSTOR DISTRIBUTION STABILITY	43
5.9 COMPUTATIONAL EXPENSE	45
5.10 TEMPLATE SIZE	48
5.11 EXPLOITING MULTIPLE CORES	51
A BIOMETRIC ERROR RATE TRADEOFF CHARACTERISTICS	54
A ALGORITHM REPORT CARDS	58
A ALGORITHM ACCURACY BY AGE GROUP	98

List of Tables

1 PARTICIPANTS	11
2 IMAGE PROPERTIES	12
3 ENROLMENT AND SEARCH SETS	15
4 PERFORMANCE WITH N = 160,000	21
5 PERFORMANCE WITH N = 640,000	22
6 PERFORMANCE WITH N = 1,600,000	23
7 ACCURACY GAINS 2010 TO 2013	25
8 ENROLMENT AND SEARCH SETS	26
9 MUGSHOT VS. WEBCAM INTEROPERABILITY	26

10	MISS RATE VS. POPULATION SIZE AT RANK 1	30
11	MISS RATE VS. POPULATION SIZE AT RANK 50	31
12	POWER-LAW MODELS OF CUMULATIVE MATCH CHARACTERISTICS	35
13	AGE GROUPS	36
14	SKETCH VS. PHOTO IDENTIFICATION ACCURACY	40
15	EFFECT OF POPULATION SIZE	46

List of Figures

1	IMAGE QUALITY MATTERS	3
2	IMAGES OF THE LEO DATASET	13
3	FERET IMAGE AND CORRESPONDING SKETCH	13
4	MULTIPLE-IMAGE ENROLLMENT TYPES	14
5	SELECTIVITY VS. FALSE POSITIVE IDENTIFICATION RATE	17
6	ALGORITHM COMPARISON: DET CHARACTERISTICS	24
7	MUGSHOT VS. WEBCAM INTEROPERABILITY CMC	27
8	EFFECT OF RANK	28
9	MISS RATES VS. POPULATION SIZE	29
10	MISS RATE VS. POPULATION SIZE	32
11	EFFECT OF AGE GROUP	38
12	REVIEWER WORKLOAD REDUCTIONS WITH THRESHOLDING	44
13	PERFORMANCE TRADESPACES	50
14	DET PROPERTIES AND INTERPRETATION :: ERROR RATES	55
15	DET PROPERTIES AND INTERPRETATION :: NON-IDEAL TESTS	56
16	DET PROPERTIES AND INTERPRETATION :: ALGORITHMS USED IN COMBINATION	57
17	KEY TO REPORT CARD FIGURES	58
18	PERFORMANCE REPORT A20C	59
19	PERFORMANCE REPORT A30C	60
20	PERFORMANCE REPORT A31C	61
21	PERFORMANCE REPORT A32C	62
22	PERFORMANCE REPORT B30C	63
23	PERFORMANCE REPORT B31C	64
24	PERFORMANCE REPORT B32C	65
25	PERFORMANCE REPORT B33C	66
26	PERFORMANCE REPORT C20C	67
27	PERFORMANCE REPORT C30C	68
28	PERFORMANCE REPORT C31C	69
29	PERFORMANCE REPORT C32C	70
30	PERFORMANCE REPORT D20C	71
31	PERFORMANCE REPORT D30C	72
32	PERFORMANCE REPORT D31C	73
33	PERFORMANCE REPORT D32C	74
34	PERFORMANCE REPORT D33C	75
35	PERFORMANCE REPORT D34C	76
36	PERFORMANCE REPORT E20C	77
37	PERFORMANCE REPORT E21C	78
38	PERFORMANCE REPORT E30C	79
39	PERFORMANCE REPORT E31C	80
40	PERFORMANCE REPORT F20C	81
41	PERFORMANCE REPORT F30C	82
42	PERFORMANCE REPORT F31C	83
43	PERFORMANCE REPORT G30C	84
44	PERFORMANCE REPORT G31C	85
45	PERFORMANCE REPORT H30C	86
46	PERFORMANCE REPORT J20C	87
47	PERFORMANCE REPORT J30C	88
48	PERFORMANCE REPORT J31C	89
49	PERFORMANCE REPORT J32C	90
50	PERFORMANCE REPORT J33C	91
51	PERFORMANCE REPORT L30C	92
52	PERFORMANCE REPORT L31C	93
53	PERFORMANCE REPORT M30C	94
54	PERFORMANCE REPORT P30C	95
55	PERFORMANCE REPORT Q30C	96

56	PERFORMANCE REPORT T30C	97
57	EFFECT OF AGE A20C	99
58	EFFECT OF AGE A30C	100
59	EFFECT OF AGE A31C	101
60	EFFECT OF AGE A32C	102
61	EFFECT OF AGE B30C	103
62	EFFECT OF AGE B31C	104
63	EFFECT OF AGE B32C	105
64	EFFECT OF AGE B33C	106
65	EFFECT OF AGE C20C	107
66	EFFECT OF AGE C30C	108
67	EFFECT OF AGE C31C	109
68	EFFECT OF AGE C32C	110
69	EFFECT OF AGE D20C	111
70	EFFECT OF AGE D30C	112
71	EFFECT OF AGE D31C	113
72	EFFECT OF AGE D32C	114
73	EFFECT OF AGE D33C	115
74	EFFECT OF AGE D34C	116
75	EFFECT OF AGE E20C	117
76	EFFECT OF AGE E21C	118
77	EFFECT OF AGE E30C	119
78	EFFECT OF AGE E31C	120
79	EFFECT OF AGE F20C	121
80	EFFECT OF AGE F30C	122
81	EFFECT OF AGE F31C	123
82	EFFECT OF AGE G30C	124
83	EFFECT OF AGE G31C	125
84	EFFECT OF AGE H30C	126
85	EFFECT OF AGE J20C	127
86	EFFECT OF AGE J30C	128
87	EFFECT OF AGE J31C	129
88	EFFECT OF AGE J32C	130
89	EFFECT OF AGE J33C	131
90	EFFECT OF AGE L30C	132
91	EFFECT OF AGE L31C	133
92	EFFECT OF AGE M30C	134
93	EFFECT OF AGE Q30C	135
94	EFFECT OF AGE S20C	136
95	EFFECT OF AGE T30C	137

Provider Name	Letter Code	Number of algorithms submitted			
		Phase 1	Phase 2	Phase 3	Total
3M/Cogent	A	3	1	3	7
Cognitec	B	1	0	4	5
Neurotechnology	C	3	1	3	7
Safran Morpho	D	1	1	5	7
NEC	E	1	2	2	5
Tsinghua University (EE - Prof. Wen)	F	2	3	2	7
Beijing Ivsign Technology Co. Ltd.	G	2	1	2	5
Chinese Academy of Sci. - Inst. Automation (Prof. Liu)	H	1	0	1	2
Chinese Academy of Sci. - Inst. Computing Technology (Prof. Shan)	I	2	1	0	3
Toshiba Corporation	J	2	2	3	7
Tsinghua University (EE - Prof. Su)	L	3	2	2	7
HP / Virage	M	1	2	1	4
Zhuhai Yisheng Electronics Tech. Co. Ltd.	P	2	0	1	3
JunYu Technology Co. Ltd.	Q	3	1	1	5
Decatur Industries Inc.	S	0	1	0	1
Ayonix Inc. (JP)	T	0	0	1	1

Table 1: *Algorithm submissions by provider* The number of different algorithms (i.e. libraries and ancillary files) submitted to FRVT by Phase, and in total. Phase 1 closed in October 2012, Phase 2 in March 2013, and Phase 3 in October 2013. Algorithms are identified in this report by the letter code in column 2, the phase it was submitted, and a sequence number starting from zero. This report does not include results for Phase 1 algorithms. NIST limited the total number of algorithms to 7.

1 Introduction

One-to-many identification represents the largest market for face recognition technology. Algorithms are used across the world in a diverse range of biometric applications: detection of duplicates in databases, detection of fraudulent applications for credentials such as passports and driving licenses, token-less access control, surveillance, social media tagging, lookalike discovery, criminal investigation, and forensic clustering.

FRVT tested only open-set identification algorithms because real-world applications are almost always “open-set”. This means that some searches have an enrolled mate, but some do not. For example, some subjects have truly not been issued a visa or drivers license before; some law enforcement searches are from first-time arrestees¹. In an “open-set” application, algorithms make no prior assumption about whether or not to return a high-scoring result, and for a mated search, the ideal behaviour is that the search produces the correct mate at high score and first rank. For a nonmate search, the ideal behavior is that the search produces zero high-scoring candidates.

In most applications, the core accuracy of a facial recognition algorithm is the most important performance variable. Resource consumption will be important also as it drives the amount of hardware, power, and cooling necessary to accommodate workflows. Algorithms consume processing time, they require computer memory, and their static template data requires storage space. This report documents all of these variables.

2 Participation

The organizations listed in Table 1 elected to submit algorithm to the FRVT evaluation.

¹Operationally closed-set applications are rare because it is usually not the case that all searches have an enrolled mate. One counter-example, however, is a cruise ship in which all passengers are enrolled and all searches should produce one, and only one, identity. Another example is forensic identification of dental records from an aircraft crash.

Property	LEO/Mugshot	LEO/Webcam	VISA
Collection Environment	Law enforcement booking	Border detainee booking	Visa application process
Collection Era	~1960s-2008	~2006-2010	~2006-2010
Camera	Digital, live	Webcam, live	Scanned paper, some live
Documentation	NIST Special Db 32 Vol 1	NIST Special Db 32 Vol 1	http://travel.state.gov/visa/visa_1750.html
Image dimensions	Various, 480x640, 240x240, 768x960	240x240	Mostly 252x300
Compression	JPEG ~20:1, mean-size: 48kB	JPEG, mean size: 5.7kB	JPEG, mean size: 9.2kB
Eye to eye distance	Mean = 107 pixels, SD = 40 pixels	Mean = 45 pixels, SD = 12	Mean = 67 pixels, SD = 6
Frontal pose	Moderate control	Poor control	Well controlled
Full frontal geometry	Mostly, but varying torso visibility	Rarely, cluttered, large, bright backgrounds	Good but cropped more closely than ISO full-frontal
Parent	Operational data	Operational data	Operational data
Population	US, adult	Central America, adult	Global, adult + children

Table 2: **Image sources and properties.** The table summarizes the key characteristics of the images used in this study.

3 Evaluation datasets

3.1 Image types

LEO images: As in our last evaluation, this report primarily reports results for an operational dataset of law enforcement images, referred to as LEO. Additionally it employs a new dataset of VISA images. The properties of these sets is summarized in Table 2. As shown in Figure 2, the LEO set contains images of two distinct types:

- ▷ **Mugshots:** Comprising about 86% of the LEO database, are mugshots having reasonable compliance with the ANSI/NIST ITL 1-2011 Type 10 standard’s subject acquisition profiles levels 10-20 for frontal images [27]. The major departure from the standard’s requirements is the presence of mild pose variations around frontal - the images of Figure 2 are typical. The images vary in size, with many being 480x600 pixels with JPEG compression applied to produce filesizes of between 18 and 36KB with many images outside this range, implying that about 1.25 bits are being encoded per color pixel.
- ▷ **Webcam images:** The remaining 14% of the images were collected using an inexpensive webcam attached to a flexible operator-directed mount. These images are all of size 240x240 pixels, that are in considerable violation of most quality-related clauses of all face recognition standards. As evident in the figure, the most common defects are non-frontal pose (associated with the rotational degrees of freedom of the camera mount), low contrast (due to varying and intense background lights), and poor spatial resolution (due to inexpensive camera optics). The images are overly JPEG compressed, to between 4 and 7KB, implying that only 0.5 to 1 bits are being encoded per color pixel.

In our 2010 report [12], accuracy was only stated for the aggregate LEO dataset. Here, for the first time, we report accuracy separately for the mugshots and the webcams images.

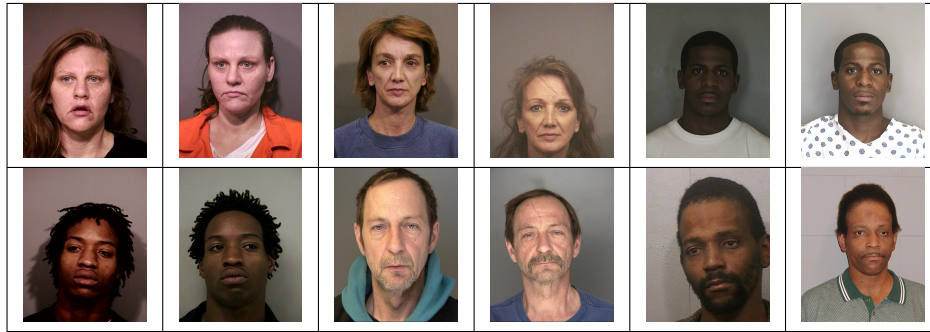
Visa images: In addition, we utilize a smaller visa database that has not been used in prior NIST evaluations. It consists of very well controlled frontal photographs of adults and children born worldwide. However, some of the images are scanned from paper photographs and therefore exhibit a reduced optical resolution. Further, as noted in Table 2, the images have a low mean interocular distances (IOD) of 67 pixels and compressed sizes of 9.2KB. These values are below the preferred minimums for the “Token” image format indicated in the ISO/IEC 19794-5:2005 standard for modern credentials, namely 120 pixels interocular distance, filesize of at least 15KB, and an image size of 480x640 pixels.

FERET Sketch images: The FERET database was collected in the 1990s and has been very widely studied. As such it is not used for evaluation here. However the City University of Hong Kong employed an artist to produce, for each person,

A = 3M/Cogent	B = Cognitec	C = Neurotechnology	D = Safran Morpho	E = NEC	F = Tsinghua U.	FNIR(N,R,T,L) “Miss rate”
G = Hisign	H = CAS-IA	I = CAS-ICT	J = Toshiba	L = Tsinghua U. II	M = HP	FPIR(N,T,L) “False alarm rate”
P = Zhuhai-Yisheng	Q = JunYu	S = Decatur	T = Ayonix			



(a) Webcam



(b) Mugshot

Figure 2: Examples of the two kinds of image that comprise the LEO database.

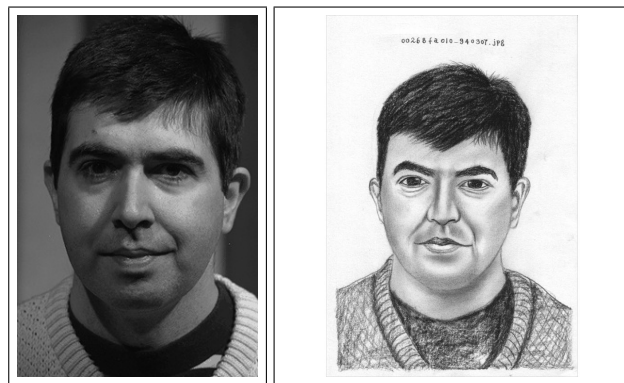


Figure 3: *Sketch realism*: A pair of images of an individual in the FERET database. At left is the grayscale version of the original 1994 frontal image. At right is the artists sketch of the grayscale image. This production of a sketch is atypical operationally: it is unusual for an artist to have access to an image of the individual.


Image				
Encounter	1	...	$K_i - 1$	K_i
Capture Time	T_1	...	T_{K_i-1}	T_{K_i}
Role RECENT	Not used	Not used	Enrolled	Search
Role LIFETIME	Enrolled	Enrolled	Enrolled	Search

Figure 4: Depiction of the “recent” and “lifetime” enrollment types.

“a face photo with lighting variation and a sketch with shape exaggeration drawn by an artist when viewing this photo.” The data was collected to support research into synthesis and recognition of sketches [25, 28]. An example is shown in Figure 3.

3.2 Enrolment types

Many operational applications include collection and enrolment of biometric data from subjects on more than one occasion. This might be done on a regular basis, as might occur in credential (re-)issuance, or irregularly, as might happen in a criminal recidivist situation [3]. The number of images per person will depend on the application area: In civil identity credentialing (e.g. passports, driving licenses), the images will be acquired approximately uniformly over time (e.g. ten years for a German passport). While the distribution of dates for such images of a person might be assumed uniform, a number of factors might undermine this assumption². In criminal applications, the number of images would depend on the number of arrests. The distribution of dates for arrest records for a person (i.e. the recidivism distribution) has been modeled using the exponential distribution but is recognized to be more complicated³.

In any case, the 2010 NIST evaluation of face recognition showed that considerable accuracy benefits accrue with retention and use of *all* historical images [12].

To this end, the FRVT API document provides $K \geq 1$ images of an individual to the enrolment software. The software is tasked with producing a single proprietary undocumented “black-box” template⁴ from the K images. This affords the algorithm an ability to generate a *model* of the individual, rather than to simply extract features from each image on a sequential basis.

As depicted in Figure 4, the i -th individual in the LEO dataset has K_i images. These are labelled x_k for $k = 1 \dots K_i$. To measure the utility of having multiple enrolment images, this report evaluates two kinds of enrolment:

- ▷ **Recent:** Only the second most recent image, x_{K_i-1} is enrolled. This type of enrolment mimics the operational policy of retaining the imagery from the most recent encounter. This might be done operationally to ameliorate the effects of face ageing.
- ▷ **Lifetime:** All except the last image are enrolled, $x_1 \dots x_{K_i-1}$. This strategy might be adopted if quality variations exist where an older image might be more suitable for matching, despite ageing.

²For example, a person might skip applying for a passport for one cycle, letting it expire. In addition, a person might submit identical images (from the same photography session) to consecutive passport applications at five year intervals.

³A number of distributions have been considered to model recidivism, see for example [2].

⁴There are no formal face template standards. Template standards only exist for fingerprint minutiae - see ISO/IEC 19794-2:2011.

	Enrolment					Search			
	Type	Num. IDs	Num. Images			Num. Images and IDs			
	See					Mate		Nonmate	
	See sec. 3.2	N	Webcam	Mugshot	Total	Webcam	Mugshot	Webcam	Mugshot
1	RECENT	20,000	0	20,000	20,000	0	20,000	28,936	171,066
2	RECENT	160,000	22,886	137,114	160,000	10,660	50,000	28,936	171,066
3	RECENT	640,000	81,221	558,779	640,000	10,660	50,000	28,936	171,066
4	RECENT	1,600,000	196,885	1,403,115	1,600,000	10,660	50,000	28,936	171,066
5	LIFETIME	20,000	0	32,948	32,948	0	20,000	28,936	171,066
6	LIFETIME	160,000	25,726	199,859	225,585	10,660	50,000	28,936	171,066
7	LIFETIME	640,000	87,864	769,070	856,934	10,660	50,000	28,936	171,066
8	LIFETIME	1,600,000	210,693	1,907,057	2,117,750	10,660	50,000	28,936	171,066

	Image Type	Enrol Type	Num. IDs	Num. enrol images	Num. mate searches	Num. nonmate searches
12	VISA	RECENT	19,972	19,972	19,972	203,082

Table 3: **Enrolment and search sets.** Each row summarizes one identification trial. The upper table concerns use of the LEO images; the lower, the VISA images. The column labeled “Num. IDs” gives the number of enrolled identities. This precedes the numbers of images, and then the number of mate, and nonmate, searches. Rows 1-8 describe trials in which webcam and mugshot images are enrolled in the natural proportions. Row 12 refers to the visa database that, uniquely, contains images of children.

In all cases, the most recent image, x_{K_i} , is reserved as the search image. For the 1.6 million subject enrolment partition of the LEO data, $1 \leq K_i \leq 33$ with $K_i = 1$ in 80.1% of the individuals, $K_i = 2$ in 13.4%, $K_i = 3$ in 3.7%, $K_i = 4$ in 1.4%, $K_i = 5$ in 0.6%, $K_i = 6$ in 0.3%, and $K_i > 6$ is 0.2% for everyone else. This distribution is substantially dependent on United States recidivism rates.

We did not evaluate the case of retaining only the highest quality image, since automated quality assessment is out of scope for this report. We do not anticipate that such strategies will prove beneficial when the quality assessment apparatus is imperfect and unvalidated.

Finally, we did not evaluate the case where $K_i > 1$ images from the same person are enrolled under different identifiers. This very common circumstance arises in so-called “event-based” applications where no attempt is made to consolidate images of a person into a single identity. Searches against such systems are likely to yield more than one image of a person in the top ranks. We do not test this kind of enrolment. Instead, we use our consolidated identity design to: a) realize accuracy gains by affording the algorithm the opportunity to build a holistic model of identity (as is common in speaker recognition systems); and b) to simplify accuracy measurement.

3.3 Limitations of the data

Neither the mugshots nor the visa images have ideal properties. The former, particularly, has too much pose variation, and the latter is degraded by its acquisition process (scan-from-paper) and too much JPEG compression. A prior NIST report [21] documents the dependence of recognition accuracy on compression and spatial sampling rate (eye-to-eye distance). While results for the LEO/Mugshot images are the best indicators of the accuracy that can be expected in contemporary deployments of one-to-many identification, it should be understood that accuracy increases can be realized by improving conformance to the appearance-related requirements of the ISO/IEC 19794-5 standard [11]⁵. That said, conformance to that standard requires both deliberate system design (capture equipment, capture environment, image preparation processes), and capture-time monitoring (to check for aberrant subject behavior). This latter aspect will in turn require conformance checks, either automated or human, even if the photographic design is perfect. Commercial quality-conformance checking tools are available, and recently have become subject to independent testing [9,10]. Human-

⁵ The ISO/IEC 19794-5:2005 Face Image standard [11], was amended in 2007 to guide photography [6] and, in 2009, to add 3D data. These additions and other improvements resulted in a second edition in 2011 [26]. Passport guidelines(e.g. [24]) are derived from this standard.

A = 3M/Cogent	B = Cognitec	C = Neurotechnology	D = Safran Morpho	E = NEC	F = Tsinghua U.	FNIR(N,R,T,L) “Miss rate”
G = Hisign	H = CAS-IA	I = CAS-ICT	J = Toshiba	L = Tsinghua U. II	M = HP	FPIR(N,T,L) “False alarm rate”
P = Zhuhai-Yisheng	Q = JunYu	S = Decatur	T = Ayonix			

mediated checks on image quality are common in passport and visa issuance procedures where photographers and, later, immigration officials, check adherence to published quality requirements.

Results for the webcam partition have some applicability to cases where image quality is significantly degraded e.g. bank ATM machine, surveillance, although the data is acquired from (mostly) cooperating subjects.

4 Performance metrics

This section gives specific definitions for accuracy and timing metrics. Tests of open-set biometric algorithms must quantify frequency of two error conditions:

- ▷ **False alarms:** Type I errors occur when search data from a person who has never been seen before is incorrectly associated with one or more enrollees' data.
- ▷ **Misses:** Type II errors arise when a search of an enrolled person's biometric does not return the correct identity.

Many practitioners prefer to talk about "hit rates" instead of "miss rates" - the first is simply one minus the other as detailed below. Sections 4.1 and 4.2 define metrics for the Type I and Type II performance variables.

Additionally, because recognition algorithms sometimes fail to produce a template from an image, or fail to execute a one-to-many search, the occurrence of such events must be recorded. Further because algorithms might elect to not produce a template from, for example, a poor quality image, these failure rates must be combined with the recognition error rates to support algorithm comparison. This is addressed in section 4.4.

Finally, section 4.5 discusses measurement of computation duration, and section 4.6 addresses the uncertainty associated with various measurements. Template size measurement is included with the results.

4.1 Quantifying false alarms

It is typical for a search to be conducted into an enrolled population of N identities, and for the algorithm to be configured to return the closest L candidate identities. These candidates are ranked by their score, in descending order. A human analyst might examine either all L candidates, or just the top $R \leq L$ identities, or only those with score greater than threshold, T . The workload associated with such examination is discussed later, in 5.7.

False alarm performance is quantified in two related ways. These express how many searches produces false positives, and then, how many false positives are produced in a search.

False positive identification rate: The first quantity, FPIR, is the proportion of nonmate searches that produce an adverse outcome:

$$\text{FPIR}(N, T, L) = \frac{\text{Num. nonmate searches where one or more enrolled candidates are returned at or above threshold, } T}{\text{Num. nonmate searches attempted.}} \quad (1)$$

Under this definition, FPIR can be computed from the highest nonmate candidate produced in a search - it is not necessary to consider candidates at rank 2 and above. FPIR is the primary measure of Type I errors in this report.

Selectivity: However, note that in any given search, more than one nonmate may be returned above threshold. In order to quantify such events, a second quantity, selectivity (SEL), is defined as the *number* of nonmates returned on a candidate

A = 3M/Cogent	B = Cognitec	C = Neurotechnology	D = Safran Morpho	E = NEC	F = Tsinghua U.	FNIR(N,R,T,L) "Miss rate"
G = Hisign	H = CAS-IA	I = CAS-ICT	J = Toshiba	L = Tsinghua U. II	M = HP	FPIR(N,T,L) "False alarm rate"
P = Zhuhai-Yisheng	Q = JunYu	S = Decatur	T = Ayonix			

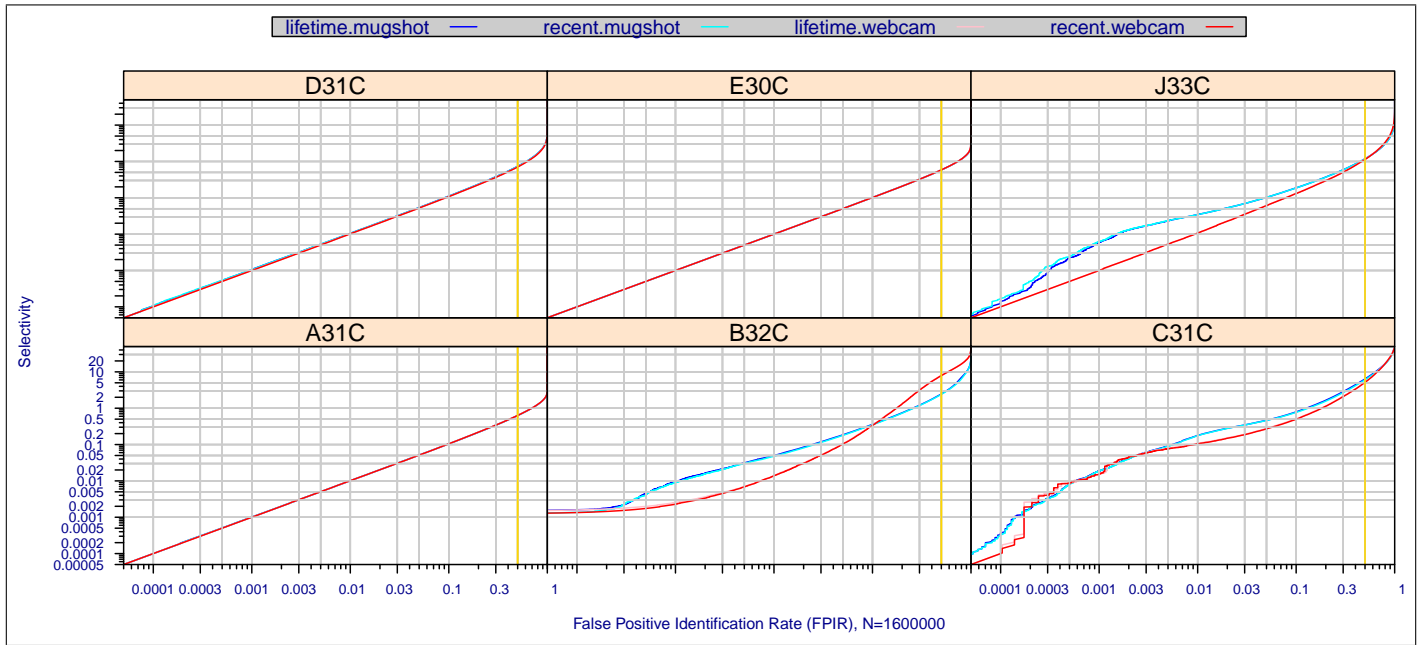


Figure 5: **Relationship of SEL and FPIR.** For six algorithms from the six most accurate suppliers, the figures plot $SEL(N, T, L)$ vs. $FPIR(N, T, L)$ parametrically with threshold, T . Selectivity is always greater than or equal to FPIR. The two are not equal when false positives are concentrated in candidate lists rather than being distributed across searches. The population size is $N = 1,600,000$. Algorithms were tasked with reporting $L = 50$ candidates. Each panel corresponds to one algorithm. The red lines correspond to searching poor quality webcam images. The blue lines correspond to searching better quality mugshot images. Analogous plots for other algorithms appear in the “report cards” of Appendix A.

list, averaged over all searches.

$$SEL(N, T, L) = \frac{\text{Num. nonmate enrolled candidates returned at or above threshold, } T}{\text{Num. nonmate searches attempted.}} \quad (2)$$

Both of these metrics are useful operationally. FPIR is useful for targeting how often an adverse false positive outcome can occur, while SEL as a number is related to workload associated with adjudicating candidate lists. The relationship between the two quantities is complicated - it depends on whether an algorithm concentrates the false alarms in the results of a few searches or whether it disburses them across many. Figure 5 plots $SEL(T)$ vs. $FPIR(T)$ for several algorithms. Note that $SEL \geq FPIR$, by definition. It is clear that some algorithms exhibit an identity relationship, $SEL = FPIR$, except at high values, but that some algorithms give selectivity values with complicated dependence on FPIR. It is not clear that the linear behavior is more, or less, desirable. On the one hand, the concentration of false matches in particular searches would lead to reduced adjudication labor in the long run. But on the other hand, the presence of several false matches on a candidate list is peculiar in that it is unlikely that several individuals would be biometrically similar to a search image. Such occurrences must be due to some property common to the pair of the images (for example, similar pose, the presence of moles, or heavy-framed glasses) and of algorithm sensitivity to it.

4.2 Quantifying hits and misses

If L candidates are returned in a search, a shorter candidate list can be prepared by taking the top $R \leq L$ candidates for which $T \geq 0$. This reduction of the candidate list is done because thresholds may be applied, and only short lists might be reviewed (according to policy or labor availability, for example). It is useful then to state accuracy in terms of R and T , so

A = 3M/Cogent	B = Cognitec	C = Neurotechnology	D = Safran Morpho	E = NEC	F = Tsinghua U.	FNIR(N,R,T,L) “Miss rate”
G = Hisign	H = CAS-IA	I = CAS-ICT	J = Toshiba	L = Tsinghua U. II	M = HP	FPIR(N,T,L) “False alarm rate”
P = Zhuhai-Yisheng	Q = JunYu	S = Decatur	T = Ayonix			

we define a “miss rate” with the general name false negative identification rate (FNIR), as follows:

$$\text{FNIR}(N, R, T, L) = \frac{\text{Num. mate searches with enrolled mate found outside top } R \text{ ranks or score below threshold, } T}{\text{Num. mate searches attempted.}} \quad (3)$$

This formulation is simple for evaluation in that it does not distinguish between causes of misses. Thus a mate that is not reported on a candidate list is treated the the same as a miss arising from face finding failure, algorithm intolerance of poor quality, or software crashes. Thus if the algorithm fails to produce a candidate list, either because the search failed, or because a search template was not made, the result is regarded as miss, adding to FNIR.

Hit rates, and true positive identification rates: While FNIR states the “miss rate” as how often the correct candidate is either not above threshold or not at good rank, many communities prefer to talk of “hit rates”. This is simply the **true positive identification rate**(TPIR) which is the complement of FNIR giving a positive statement of how often mated searches are successful:

$$\text{TPIR}(N, R, T, L) = 1 - \text{FNIR}(N, R, T, L) \quad (4)$$

Reliability and **sensitivity** are corresponding terms, typically being identical to TPIR. This quantity is often cited in automated fingerprint identification system (AFIS) evaluations.

An important special case is the **cumulative match characteristic**(CMC) which summarizes accuracy of mated-searches only. It ignores similarity scores by relaxing the threshold requirement, and just reports the fraction of mated searches returning the mate at rank R or better.

$$\text{CMC}(N, R, L) = 1 - \text{FNIR}(N, R, 0, L) \quad (5)$$

We primarily cite the complement of this quantity $1 - \text{FNIR}$, the fraction of mates *not* in the top R ranks.

The **rank one hit rate** is the fraction of mated searches yielding the correct candidate at best rank, i.e. $\text{CMC}(N, 1, L)$. While this quantity is the most common summary indicator of an algorithms’ efficacy, it is not dependent on similarity scores, so it does not distinguish between strong (high scoring) and weak hits. It also ignores that an adjudicating reviewer is often willing to look at many candidates.

4.3 Best practice testing requires execution of searches with and without mates

FRVT embedded 1:N searches of two kinds: Those for which there is an enrolled mate, and those for which there is not. The respective numbers for these types of searches appear in Table 3. However, it is common to conduct only mated searches. This is bad practice because if the information that a mate always exists is revealed to a test participant, or can be reasonably assumed, then unrealistic gaming of the test is possible. The cumulative match characteristic is computed from candidate lists produced in mated searches. Even if the CMC is the only metric of interest, the actual trials executed in a test should nevertheless include searches for which no mate exists. As detailed in Table 3 the FRVT reserved disjoint populations of subjects for executing true nonmate searches.

4.4 Failure to extract

Template generation can fail either during enrolment, or ahead of one-to-many search. This report quantifies only failure-to-enrol. This quantity is denoted FTE. It is computed as the proportion of images that do not produce an output template. The corresponding search-phase number, FTX, is not reported because we consider the FTE number adequate. This is

A = 3M/Cogent G = Hisign P = Zhuhai-Yisheng	B = Cognitec H = CAS-IA Q = JunYu	C = Neurotechnology I = CAS-ICT S = Decatur	D = Safran Morpho J = Toshiba T = Ayonix	E = NEC L = Tsinghua U. II	F = Tsinghua U. M = HP	FNIR(N,R,T,L) “Miss rate” FPIR(N,T,L) “False alarm rate”
---	---	---	--	-------------------------------	---------------------------	---

supported by a) the observed error rates being very low, usually zero; b) we assume, that the same underlying algorithm is used⁶; and c) the enrolment sets are much larger.

Failure to extract rates are incorporated into FNIR and FPIR measurements as follows.

- ▷ **Enrolment templates:** Any failed enrolment is regarded as producing a zero length template. Algorithms are required by the API [14] to transparently process zero length templates. The effect of template generation failure on search accuracy depends on whether subsequent searches are mated, or nonmated: Mated searches will fail giving elevated FNIR; Nonmated searches will not produce false positives, so FPIR will be reduced by (to first order) a factor of $1 - \text{FTE}$.
- ▷ **Search templates and 1:N search:** In cases where the algorithm fails to produce a search template from input imagery, the result is taken to be a candidate list whose entries have no hypothesized identities and zero score. The effect of template generation failure on search accuracy depends on whether searches are mated, or nonmated: Mated searches will fail giving elevated FNIR; Nonmated searches will not produce false positives, so FPIR will be reduced.

$$\text{FNIR}^\dagger = \text{FTX} + (1 - \text{FTX})\text{FNIR} \quad (6)$$

$$\text{FPIR}^\dagger = (1 - \text{FTX})\text{FPIR} \quad (7)$$

This approach is the correct treatment for positive-identification applications such as access control where cooperative users are enrolled and make attempts at recognition. This approach is not appropriate to negative identification applications, such as visa fraud detection, in which hostile individuals may attempt to evade detection by submitting poor quality samples. In those cases, template generation failure should be investigated as though a false alarm had occurred.

4.5 Timing measurement

Algorithms were submitted to NIST as implementations of the application programming interface(API) specified by NIST in the Evaluation Plan [14]. The API includes functions for initialization, template generation, finalization, and search. Two template generation functions are required, one for the preparation of an enrolment template, and one for a search template.

In NIST's test harness, all functions were wrapped by calls to the function *get_time_of_day()* which enables duration measurements with microsecond resolution. Timing was measured on a dedicated computer equipped with 192GB of main memory. The computer was not running any other processes except those back-grounded as part of the operating system. Timing measurements do not include disk access unless the algorithm under test elected to access enrolment or configuration data during a search (something that is not necessary because the API supported initialization prior to searching).

The FRVT test plan formally stated the durations limits on the core elemental functions of the algorithms. The times were stated as 90-th percentiles.

4.6 Uncertainty estimation

This study leverages operational datasets for measurement of recognition error rates. This affords several advantages. First, large numbers of searches are conducted (see Table 3) giving precision to the measurements. Moreover, these do

⁶While this assumption is violated for at least Morpho's algorithms which use larger search templates than enrolment templates, the failure to make a template rates are very close to zero.

A = 3M/Cogent G = Hisign P = Zhuhai-Yisheng	B = Cognitec H = CAS-IA Q = JunYu	C = Neurotechnology I = CAS-ICT S = Decatur	D = Safran Morpho J = Toshiba T = Ayonix	E = NEC L = Tsinghua U. II	F = Tsinghua U. M = HP	FNIR(N,R,T,L) "Miss rate" FPIR(N,T,L) "False alarm rate"
---	---	---	--	-------------------------------	---------------------------	---

not involve reuse of individuals, and thus, binomial statistics can be expected to apply to recognition error counts. In that case, an observed count of a particular recognition outcome (i.e. a false negative or false positive) in M trials will sustain 95% confidence that the actual error rate is no larger than some value. As an example, the minimum number of searches conducted in this report is $M = 10,660$, and the observed FNIR is never below 0.01 so the measurement supports a conclusion that the actual FNIR is no higher than 0.0113 at 95% confidence level, and 0.0133 at the 99.9% level. On the false positive side, we tabulate FNIR at FPIR values as low as 0.002. Given estimates based on 171,066 nonmate trials, the actual FPIR values will be below 0.00214 (at 95%) and 0.00235 (at 99.9% confidence). The point is that large scale evaluation, without reuse of subjects, supports tight uncertainty bounds on the measured error rates.

Bootstrapping is an empirical method of measuring the variability of a statistic, often employed when the variability cannot be determined analytically. In the context of this evaluation, bootstrapping is sometimes used to measure the distribution of error statistics (i.e. FNIR or FPIR) at a fixed threshold. Each bootstrap iteration samples with replacement from the original set of comparisons. The statistic of interest is then computed over the sampled data. This process is repeated for a large number of bootstrap iterations to produce a distribution of the measured statistic. Bootstrapping relies on several assumptions, including that the sample data is independent and identically distributed. However, when different comparisons involve the same individual, the comparisons are likely to be correlated due to the existence of Doddington’s zoo [7]. Thus, the independence assumption is violated. Determining the effect this has on the bootstrapped distributions is beyond the scope of this evaluation, but the likely result is an underestimation of the variability of FNIR and/or FPIR in some cases. The experimental design here avoids this issue.

5 Results

This section details performance of the algorithms submitted to the Class-C identification algorithm track of the FRVT evaluation. Performance metrics were described in section 4. The following subsections address specific aspects of performance. Appendix A includes a report for each algorithm, where multiple plots are assembled side-by-side.

5.1 Comparative accuracy

Methods: LEO/Mughots and LEO/Webcam images are enrolled in their natural distributions using both the “lifetime” and “recent” enrolment types of section 3.2. Four different population sizes are used. The numbers of enrolled individuals are 20000, 160000, 640000, and 1600000. For the recent enrolment type, the number of images is identical to the number of individuals. For the lifetime enrolment type, the numbers of enrolled images are: 32948, 225585, 856934, and 2117750 as documented in Table 3. Each successive enrolled population includes the subjects and images of the smaller population.

Two nonmate search sets are used, one containing 10660 webcam images, and another containing 171066 mugshots.

Results: Table 4 presents FNIR values for all algorithms submitted to FRVT in 2013⁷ applied to an enrolled population size of $N = 160,000$. Table 5 presents the same data for $N = 640,000$ but it excludes algorithms that did not demonstrate high accuracy in the $N = 160K$ trial. Table 6 likewise presents the same data for $N = 1.6$ million enrolled identities.

Figure 6 shows detection error tradeoff (DET) characteristics for algorithms from the most capable algorithms. These summarize the tradeoff of FNIR for FPIR for searches conducted into enrolled galleries containing 1.6 million people.

Discussion: As in the 2010 test, the algorithms from the NEC corporation give broadly the lowest error rates on all

⁷The table excludes entries that did not execute to completion or that did not meet minimum speed criteria. It additionally excludes algorithms submitted to phase 1 of the FRVT in August 2012.

A = 3M/Cogent	B = Cognitec	C = Neurotechnology	D = Safran Morpho	E = NEC	F = Tsinghua U.	FNIR(N,R,T,L) “Miss rate”
G = Hisign	H = CAS-IA	I = CAS-ICT	J = Toshiba	L = Tsinghua U. II	M = HP	FPIR(N,T,L) “False alarm rate”
P = Zhuhai-Yisheng	Q = JunYu	S = Decatur	T = Ayonix			

ENROL N=160K ALG	ALL LIFETIME PHOTOS					MOST RECENT PHOTO					TEMPLATE		SEARCH					
	MUGSHOTS					MUGSHOTS					SIZE (BYTES)		TIME (MSEC)					
	FPIR=0.002	FPIR=0.02	R=50	R=10	R=1	BOTH	FTE	FPIR=0.02	R=1	WORK	FPIR=0.02	R=1	WORK	ENROL	SEARCH	TEMPLATE	SEARCH	TIME (MSEC)
A20C	²³ 0.247	0.193	²² 0.063	0.083	²² 0.127	0.000		0.207	0.145	4.575	0.496	¹⁷ 0.355	10.975	4106	4106	³⁰ 522	³¹ 1235	
A30C	²¹ 0.240	0.185	¹⁷ 0.055	0.074	²⁰ 0.116	0.000		0.199	0.133	4.021	0.426	¹⁶ 0.291	8.281	4622	4622	¹⁹ 315	¹⁶ 243	
A31C	²² 0.241	0.183	¹⁸ 0.055	0.074	¹⁹ 0.116	0.000		0.198	0.133	4.021	0.425	¹⁵ 0.291	8.281	4622	4622	²¹ 328	¹⁷ 245	
A32C	²⁵ 0.261	0.200	²³ 0.068	0.087	²⁴ 0.132	0.000		0.247	0.173	5.501	0.536	²⁵ 0.394	12.574	1046	1046	⁸ 194	⁵ 102	
B30C	¹⁴ 0.183	0.144	¹⁹ 0.058	0.072	¹⁷ 0.100	0.002		0.171	0.120	4.016	0.840	³⁷ 0.604	21.956	4796	4796	¹⁸ 295	²³ 756	
B31C	¹⁶ 0.191	0.143	²⁰ 0.058	0.072	¹⁶ 0.100	0.002		0.170	0.120	4.016	0.854	³⁶ 0.604	21.955	4796	4796	¹⁷ 280	²⁴ 760	
B32C	¹⁷ 0.193	0.137	¹⁶ 0.050	0.063	¹⁵ 0.091	0.002		0.164	0.109	3.508	0.795	³⁰ 0.502	16.961	9932	9932	²⁷ 441	³⁷ 2021	
B33C	¹⁹ 0.204	0.139	¹⁵ 0.050	0.063	¹⁴ 0.091	0.002		0.165	0.109	3.508	0.792	²⁹ 0.502	16.961	9932	9932	²⁵ 386	³⁸ 2022	
C20C	²⁰ 0.543	0.281	³³ 0.115	0.139	²⁸ 0.177	0.000		0.319	0.208	7.651	0.906	⁴³ 0.784	33.648	35994	35994	¹⁴ 225	³⁰ 1108	
C30C	³¹ 0.566	0.267	²⁵ 0.082	0.106	²⁶ 0.146	0.000		0.304	0.173	5.683	0.817	³⁵ 0.603	21.964	37008	37008	²⁴ 366	⁴² 7117	
C31C	²⁷ 0.518	0.254	²⁶ 0.084	0.106	²⁵ 0.143	0.000		0.291	0.169	5.762	0.827	³⁹ 0.640	25.118	37008	37008	²³ 361	³² 1256	
C32C	³⁰ 0.558	0.276	²⁷ 0.087	0.113	²⁷ 0.155	0.000		0.314	0.183	6.017	0.819	³⁸ 0.612	22.490	5040	37008	²² 360	²⁰ 1004	
D20C	⁹ 0.128	0.095	¹⁴ 0.044	0.050	¹⁰ 0.068	0.000		0.114	0.081	2.856	0.369	¹⁰ 0.252	7.728	8005	20489	⁴² 713	²² 510	
D30C	⁸ 0.122	0.095	¹¹ 0.042	0.048	⁸ 0.067	0.000		0.109	0.078	2.716	0.369	¹¹ 0.254	7.688	8562	21046	⁴⁰ 709	²⁶ 806	
D31C	⁷ 0.122	0.091	¹⁰ 0.042	0.048	⁵ 0.065	0.000		0.107	0.077	2.691	0.379	¹⁴ 0.260	7.754	12247	24731	³⁸ 708	³⁵ 1275	
D32C	²⁰ 0.236	0.182	²⁴ 0.076	0.094	²³ 0.128	0.000		0.207	0.150	5.125	0.512	²⁴ 0.389	13.751	857	857	²⁸ 469	⁸ 87	
D33C	¹¹ 0.144	0.100	¹² 0.042	0.048	⁷ 0.066	0.000		0.106	0.076	2.706	0.374	¹² 0.258	7.802	8005	20489	⁴¹ 710	²⁷ 809	
D34C	¹⁰ 0.142	0.094	⁷ 0.038	0.046	⁶ 0.066	0.000		0.106	0.076	2.571	0.379	¹³ 0.259	7.659	8005	20489	³⁹ 708	²⁵ 805	
E20C	³ 0.066	0.048	² 0.018	0.023	⁴ 0.032	0.000		0.057	0.037	1.202	0.145	² 0.081	2.088	2465	2465	¹⁰ 205	⁸ 120	
E21C	⁴ 0.073	0.052	¹ 0.018	0.023	³ 0.032	0.000		0.060	0.037	1.202	0.169	¹ 0.081	2.088	2465	2465	⁹ 204	¹⁰ 139	
E30C	¹ 0.052	0.040	³ 0.019	0.022	¹ 0.029	0.000		0.047	0.034	1.179	0.136	⁴ 0.084	2.196	2529	2529	¹¹ 205	⁷ 118	
E31C	² 0.054	0.042	⁴ 0.021	0.023	² 0.031	0.000		0.047	0.035	1.289	0.135	³ 0.082	2.397	2529	2529	¹³ 213	¹ 152	
F20C	³³ 0.656	0.494	³⁵ 0.155	0.202	³⁵ 0.282	0.006		0.538	0.312	10.181	0.713	²⁸ 0.474	15.197	6760	6760	² 100	¹⁸ 341	
F30C	³⁸ 0.749	0.459	²⁹ 0.108	0.143	³⁰ 0.214	0.001		0.490	0.247	7.650	0.639	¹⁸ 0.373	11.338	6440	6440	¹⁵ 257	³⁵ 1525	
F31C	³⁸ 0.791	0.460	³¹ 0.108	0.142	³² 0.214	0.002		0.493	0.248	7.615	0.643	²³ 0.382	11.451	7484	7484	¹⁶ 265	³⁸ 1940	
G30C	³⁷ 0.709	0.502	³⁴ 0.119	0.162	³⁴ 0.245	0.010		0.538	0.285	8.439	0.718	²⁶ 0.444	12.736	3484	3484	³ 102	²⁸ 811	
G31C	²⁸ 0.536	0.429	³² 0.113	0.150	³³ 0.219	0.010		0.477	0.254	7.898	0.602	²¹ 0.382	11.830	2240	2240	⁵ 170	⁷ 84	
H30C	³² 0.614	0.523	³⁶ 0.183	0.229	³⁶ 0.309	0.000		0.566	0.353	12.221	0.705	²⁷ 0.468	15.692	4420	4420	¹⁴ 233	¹ 43	
J20C	²⁴ 0.254	0.176	¹⁴ 0.044	0.063	¹⁸ 0.102	0.000		0.216	0.121	3.341	0.433	⁹ 0.250	6.933	6206	6206	²⁹ 490	¹¹ 148	
J30C	¹⁵ 0.189	0.129	⁸ 0.039	0.051	¹² 0.079	0.000		0.150	0.091	2.628	0.339	⁷ 0.203	5.407	4158	4158	³¹ 526	⁶ 110	
J31C	¹⁸ 0.197	0.135	⁹ 0.039	0.053	¹³ 0.082	0.000		0.156	0.094	2.655	0.347	⁸ 0.207	5.444	4158	4158	³² 530	⁹ 130	
J32C	¹³ 0.179	0.117	⁶ 0.034	0.044	¹¹ 0.069	0.000		0.138	0.081	2.335	0.314	⁶ 0.184	4.639	8254	8254	³⁴ 545	¹⁵ 236	
J33C	¹² 0.171	0.114	⁵ 0.033	0.043	⁹ 0.067	0.000		0.133	0.079	2.255	0.305	⁵ 0.181	4.516	8254	8254	³⁵ 548	¹³ 188	
L30C	⁴¹ 0.850	0.712	⁴³ 0.293	0.351	⁴² 0.442	0.003		0.743	0.490	18.525	0.932	⁴² 0.708	27.775	7328	7328	³³ 539	⁴¹ 4031	
L31C	⁴⁴ 1.000	0.639	⁴¹ 0.253	0.310	⁴¹ 0.398	0.003		0.679	0.444	16.272	0.903	⁴⁰ 0.658	25.050	7328	7328	³⁷ 635	⁴⁰ 4000	
M20C	³⁵ 0.678	0.592	⁴⁰ 0.221	0.278	⁴⁰ 0.371	0.024		0.627	0.412	14.088	0.770	³⁴ 0.587	21.009	5608	5608	⁴³	⁴³	
M21C	³⁸ 0.678	0.586	³⁹ 0.211	0.265	³⁸ 0.355	0.024		0.618	0.390	13.282	0.753	³² 0.563	19.754	5608	5608	⁴⁴	⁴⁴	
M30C	³⁴ 0.667	0.576	³⁷ 0.198	0.252	³⁷ 0.342	0.024		0.618	0.390	13.279	0.753	³¹ 0.562	19.768	5608	5608	⁴ 104	¹² 171	
P30C	²⁶ 0.421	0.291	²¹ 0.059	0.080	²¹ 0.123	0.005		0.293	0.151	4.664	0.601	¹⁹ 0.375	11.687	3713	3713	³⁶ 559	³⁹ 2930	
Q30C	⁴³ 0.915	0.595	³⁸ 0.206	0.265	³⁹ 0.361	0.004		0.632	0.405	13.897	0.770	³³ 0.567	18.766	5600	5600	⁷ 191	³⁴ 1390	
S20C	⁴² 0.851	0.773	⁴⁴ 0.373	0.433	⁴⁴ 0.523	0.033		0.800	0.569	22.745	0.928	⁴⁴ 0.810	34.726	800	800	⁶ 171	²⁰ 444	
T30C	⁴⁰ 0.836	0.753	⁴² 0.288	0.350	⁴³ 0.449	0.000		0.779	0.495	18.360	0.838	⁴¹ 0.664	25.564	1936	1936	¹² 207	⁴ 94	

Table 4: **Accuracy, speed, template sizes at N = 160K.** For LEO mugshots and webcam images, the values are FNIR “miss rates” except at right where template sizes and search durations are shown, and the column FTE indicating rate of template generation failure. Five different accuracy criteria appear: At left are values for high-threshold, low false positive rates relevant to more “lights-out” applications with minimal human adjudication. The next columns apply to no-threshold rank-based criteria relevant to investigative applications where a reviewer is permanently employed to adjudicate candidate lists of length 50, 10 or 1. The column “WORK” gives the workload statistic of eq. 13, with recidivism rate 100% ($\beta = 1$) gives the expected number of candidate reviews a reviewer must make. From left to right, the colored columns are likely the most important metrics for, respectively, the “lights-out”, the “investigational” and the “poor image” communities. The blue superscripts give column-wise algorithm rankings.

ENROL N=640K ALG	ALL LIFETIME PHOTOS					BOTH			MOST RECENT PHOTO			TEMPLATE		SEARCH		
	MUGSHOTS					MUGSHOTS			WEBCAM			SIZE (BYTES)		TIME (MSEC)		
	FPIR=0.002	FPIR=0.02	R=50	R=10	R=1	FTE	FPIR=0.02	R=1	WORK	FPIR=0.02	R=1	WORK	ENROL	SEARCH	TEMPLATE	SEARCH
A20C	²¹ 0.274	0.219	²² 0.080	0.104	²³ 0.150	0.000	0.231	0.167	5.579	0.534	¹⁷ 0.396	13.237	4106	4106	²⁶ 524	³⁰ 4921
A30C	²⁰ 0.267	0.212	¹⁹ 0.071	0.092	¹⁹ 0.139	0.000	0.225	0.156	4.963	0.466	¹⁵ 0.333	10.329	4622	4622	¹⁷ 344	²⁰ 979
A31C	¹⁹ 0.267	0.211	¹⁹ 0.071	0.092	²⁰ 0.139	0.000	0.223	0.156	4.963	0.465	¹⁶ 0.334	10.328	4622	4622	¹⁶ 356	¹⁶ 956
A32C	²² 0.288	0.230	²³ 0.084	0.107	²⁴ 0.156	0.000	0.272	0.200	6.641	0.576	²¹ 0.439	15.117	1046	1046	⁶ 188	⁵ 419
B30C	¹¹ 0.208	0.161	¹⁸ 0.069	0.085	¹⁷ 0.113	0.002	0.190	0.136	4.770	0.872	³¹ 0.649	25.371	4796	4796	¹⁵ 284	²² 2975
B31C	¹³ 0.222	0.163	¹⁷ 0.069	0.085	¹⁶ 0.113	0.002	0.192	0.137	4.770	0.886	³⁰ 0.649	25.372	4796	4796	¹⁴ 280	²³ 2985
B32C	¹⁵ 0.231	0.156	¹⁶ 0.061	0.075	¹⁵ 0.104	0.002	0.186	0.125	4.232	0.840	²⁵ 0.546	20.215	9932	9932	²² 381	³² 8068
B33C	¹⁷ 0.248	0.160	¹⁵ 0.061	0.075	¹⁴ 0.104	0.002	0.189	0.125	4.232	0.836	²⁴ 0.546	20.215	9932	9932	²¹ 378	³¹ 8058
C20C	²⁶ 0.617	0.307	²⁵ 0.134	0.157	²⁵ 0.194	0.000	0.348	0.229	8.763	0.921	³⁷ 0.807	35.913	35994	35994	¹¹ 240	²⁸ 4361
C30C	³⁶		²⁵ 0.102	0.126	²⁶ 0.165	0.000	0.195	0.195	6.879	0.855	²⁹ 0.643	24.988	37008	37008	²⁰ 373	³⁴ 28399
C31C	²⁵ 0.605	0.283	²⁶ 0.103	0.124	²⁵ 0.162	0.000	0.323	0.191	6.865	0.855	³³ 0.676	27.597	37008	37008	¹⁸ 362	²⁸ 4766
C32C	²⁵ 0.663	0.310	²⁷ 0.109	0.133	²⁷ 0.175	0.000	0.351	0.207	7.293	0.853	³² 0.654	25.540	5040	5040	¹⁹ 363	²⁷ 4052
D20C	⁷ 0.131	0.100	¹³ 0.053	0.056	⁹ 0.075	0.000	0.119	0.089	3.378	0.391	⁹ 0.278	9.527	8005	20489	⁴⁴	⁴⁴
D30C	⁶ 0.129	0.101	¹² 0.051	0.056	⁸ 0.073	0.000	0.116	0.087	3.233	0.397	¹⁰ 0.282	9.528	8562	21046	³¹ 705	¹⁷ 956
D31C	⁵ 0.127	0.096	¹⁰ 0.051	0.055	⁵ 0.071	0.000	0.115	0.084	3.185	0.408	¹⁴ 0.287	9.597	12247	24731	³³ 709	²¹ 1425
D32C	¹⁸ 0.260	0.202	²⁴ 0.090	0.109	²¹ 0.144	0.000	0.229	0.168	6.033	0.550	¹⁹ 0.430	15.765	857	857	²³ 472	²⁶
D33C	⁹ 0.156	0.106	¹¹ 0.051	0.056	⁶ 0.072	0.000	0.112	0.085	3.228	0.398	¹¹ 0.284	9.632	8005	20489	³⁴ 711	¹⁹ 965
D34C	⁸ 0.153	0.101	⁷ 0.046	0.053	⁷ 0.073	0.000	0.114	0.086	3.037	0.406	¹² 0.286	9.147	8005	20489	³² 708	¹⁸ 958
E20C	³ 0.076	0.057	³ 0.022	0.027	⁴ 0.038	0.000	0.067	0.044	1.438	0.173	³ 0.100	2.702	2465	2465	¹⁰ 216	⁸ 522
E21C	⁴ 0.090	0.062	² 0.022	0.027	³ 0.038	0.000	0.074	0.044	1.438	0.204	² 0.100	2.702	2465	2465	⁷ 205	⁷ 475
E30C	¹ 0.059	0.044	¹ 0.021	0.024	¹ 0.032	0.000	0.052	0.037	1.301	0.160	⁴ 0.100	2.766	2529	2529	⁸ 207	⁶ 458
E31C	² 0.062	0.046	⁴ 0.024	0.026	² 0.034	0.000	0.053	0.038	1.461	0.157	¹ 0.098	3.033	2529	2529	⁹ 208	¹ 168
J20C	²³ 0.294	0.205	¹⁴ 0.058	0.078	¹⁸ 0.119	0.000	0.248	0.143	4.244	0.498	¹³ 0.287	8.729	6206	6206	²⁴ 482	¹¹ 572
J30C	¹⁴ 0.227	0.154	⁸ 0.050	0.063	¹² 0.093	0.000	0.179	0.109	3.332	0.389	⁷ 0.241	7.179	4158	4158	²⁷ 530	¹⁰ 555
J31C	¹⁶ 0.238	0.165	⁹ 0.050	0.065	¹³ 0.099	0.000	0.187	0.113	3.369	0.400	⁸ 0.247	7.252	4158	4158	²⁵ 523	⁹ 554
J32C	¹² 0.218	0.141	⁶ 0.044	0.055	¹¹ 0.083	0.000	0.164	0.097	2.980	0.359	⁶ 0.219	6.259	8254	8254	²⁸ 541	¹³ 766
J33C	¹⁰ 0.208	0.135	⁵ 0.042	0.053	¹⁰ 0.081	0.000	0.159	0.095	2.863	0.352	⁵ 0.214	6.042	8254	8254	²⁹ 552	¹⁴ 812
P30C	²⁴ 0.470	0.344	²¹ 0.076	0.100	²² 0.147	0.005	0.332	0.172	5.784	0.644	¹⁸ 0.421	14.182	3713	3713	³⁰ 599	³³ 11620

Table 5: **Accuracy, speed, template sizes at N = 640K.** For LEO mugshots and webcam images, the values are FNIR “miss rates” except at right where template sizes and search durations are shown, and the column FTE indicating rate of template generation failure. Five different accuracy criteria appear: At left are values for high-threshold, low false positive rates relevant to more “lights-out” applications with minimal human adjudication. The next columns apply to no-threshold rank-based criteria relevant to investigative applications where a reviewer is permanently employed to adjudicate candidate lists of length 1, 10 or 50. The workload statistic (equation 13), with recidivism rate 100% ($\beta = 1$) gives the expected number of candidate reviews a reviewer must make. From left to right, the colored columns are likely the most important metrics for, respectively, the “lights-out”, the “investigational” and the “poor image” communities. The blue superscripts give column-wise algorithm rankings.

ENROL N=1600K ALG	ALL LIFETIME PHOTOS					MOST RECENT PHOTO					TEMPLATE		SEARCH	
	MUGSHOTS					MUGSHOTS					SIZE (BYTES)		TIME (MSEC)	
	FPIR=0.002	FPIR=0.02	R=50	R=10	R=1	FTE	FPIR=0.02	R=1	WORK	FPIR=0.02	R=1	WORK	ENROL	SEARCH
A20C	²¹ 0.294	0.236	²¹ 0.093	0.119	²⁷ 0.166	0.000	0.247	0.182	6.285	0.559	¹⁷ 0.427	14.965	4106	4106
A30C	²⁰ 0.288	0.229	¹⁹ 0.083	0.107	²¹ 0.155	0.000	0.240	0.172	5.643	0.498	¹⁸ 0.364	11.870	4622	4622
A31C	¹⁹ 0.287	0.228	²⁰ 0.083	0.107	²⁰ 0.155	0.000	0.239	0.172	5.643	0.497	¹⁶ 0.364	11.870	4622	4622
A32C	²² 0.312	0.248	²² 0.096	0.122	²⁸ 0.172	0.000	0.291	0.216	7.489	0.604	¹⁹ 0.474	16.920	1046	1046
B30C	¹⁰ 0.231	0.173	¹⁸ 0.078	0.094	¹⁷ 0.123	0.002	0.205	0.147	5.291	0.891	²⁴ 0.673	27.447	4796	4796
B31C	¹² 0.249	0.175	¹⁷ 0.078	0.094	¹⁹ 0.123	0.002	0.207	0.147	5.291	0.903	²³ 0.673	27.446	4796	4796
B32C	¹⁵ 0.268	0.170	¹⁶ 0.069	0.084	¹⁸ 0.113	0.002	0.201	0.136	4.741	0.865	²⁰ 0.576	22.235	9932	9932
B33C	¹⁸ 0.286	0.174	¹⁵ 0.069	0.084	¹⁴ 0.113	0.002	0.206	0.136	4.748	0.861	²¹ 0.576	22.244	9932	9932
C20C	²⁵ 0.655	0.325	²⁷ 0.147	0.169	²⁷ 0.206	0.000	0.367	0.243	9.475	0.928	²⁷ 0.823	37.275	35994	35994
C30C	³³		²⁵ 0.115	0.139	²⁵ 0.179	0.000		0.211	7.662		²² 0.669	27.038	37008	37008
C31C	²⁴ 0.653	0.303	²⁴ 0.114	0.135	²⁴ 0.173	0.000	0.344	0.205	7.567	0.872	²⁶ 0.702	29.369	37008	37008
C32C	²⁶ 0.705	0.333	²⁶ 0.122	0.146	²⁶ 0.190	0.000	0.375	0.223	8.097	0.876	²⁵ 0.682	27.611	5040	5040
D20C	⁷ 0.137	0.104	¹³ 0.061	0.066	⁹ 0.080	0.000	0.124	0.096	3.792	0.408	⁹ 0.298	10.839	8005	8005
D30C	⁶ 0.136	0.107	¹² 0.059	0.063	⁸ 0.080	0.000	0.123	0.093	3.678	0.418	¹⁰ 0.302	10.777	8562	8562
D31C	⁵ 0.134	0.102	¹⁰ 0.059	0.063	⁵ 0.077	0.000	0.120	0.091	3.620	0.427	¹³ 0.307	10.913	12247	12247
D32C	¹⁷ 0.283	0.219	²³ 0.101	0.120	¹⁹ 0.155	0.000	0.246	0.180	6.647	0.582	¹⁸ 0.462	17.177	857	857
D33C	⁸ 0.161	0.111	¹¹ 0.059	0.063	⁶ 0.078	0.000	0.119	0.091	3.673	0.418	¹¹ 0.303	10.877	8005	8005
D34C	⁹ 0.167	0.107	⁷ 0.054	0.059	⁷ 0.079	0.001	0.121	0.092	3.429	0.425	¹² 0.307	10.206	8005	8005
E20C	³ 0.089	0.064	³ 0.025	0.030	⁴ 0.042	0.000	0.075	0.049	1.603	0.192	⁴ 0.115	3.198	2465	2465
E21C	⁴ 0.105	0.071	² 0.025	0.030	³ 0.042	0.000	0.084	0.049	1.603	0.229	³ 0.115	3.198	2465	2465
E30C	¹ 0.068	0.049	¹ 0.023	0.026	¹ 0.035	0.000	0.057	0.041	1.445	0.176	² 0.113	3.201	2529	2529
E31C	² 0.070	0.050	⁴ 0.026	0.029	² 0.037	0.000	0.058	0.042	1.602	0.176	¹ 0.113	3.520	2529	2529
J20C	²³ 0.339	0.227	¹⁴ 0.068	0.090	¹⁸ 0.133	0.000	0.275	0.161	4.940	0.534	¹⁴ 0.318	10.162	6206	6206
J30C	¹⁴ 0.258	0.173	⁸ 0.058	0.071	¹² 0.104	0.000	0.200	0.122	3.831	0.422	⁷ 0.270	8.371	4158	4158
J31C	¹⁶ 0.275	0.184	⁹ 0.058	0.074	¹³ 0.110	0.000	0.208	0.127	3.886	0.432	⁸ 0.274	8.452	4158	4158
J32C	¹³ 0.256	0.158	⁶ 0.051	0.063	¹¹ 0.093	0.000	0.185	0.108	3.448	0.394	⁶ 0.246	7.427	8254	8254
J33C	¹¹ 0.237	0.153	⁵ 0.049	0.061	¹⁰ 0.091	0.000	0.179	0.107	3.332	0.382	⁵ 0.237	7.142	8254	8254
P30C	²⁷ 0.731	0.652	²⁸ 0.530	0.540	²⁸ 0.564				50.000		³³	50.000		

Table 6: **Accuracy, speed, template sizes at N = 1.6M:** For LEO mugshots and webcam images, the values are FNIR “miss rates” except at right where template sizes and search durations are shown, and the column FTE indicating rate of template generation failure. Five different accuracy criteria appear: At left are values for high-threshold, low false positive rates relevant to more “lights-out” applications with minimal human adjudication. The next columns apply to no-threshold rank-based criteria relevant to investigative applications where a reviewer is permanently employed to adjudicate candidate lists of length 1, 10 or 50. The workload statistic (equation 13), with recidivism rate 100% ($\beta = 1$) gives the expected number of candidate reviews a reviewer must make. From left to right, the colored columns are likely the most important metrics for, respectively, the “lights-out”, the “investigational” and the “poor image” communities. The blue superscripts give column-wise algorithm rankings.

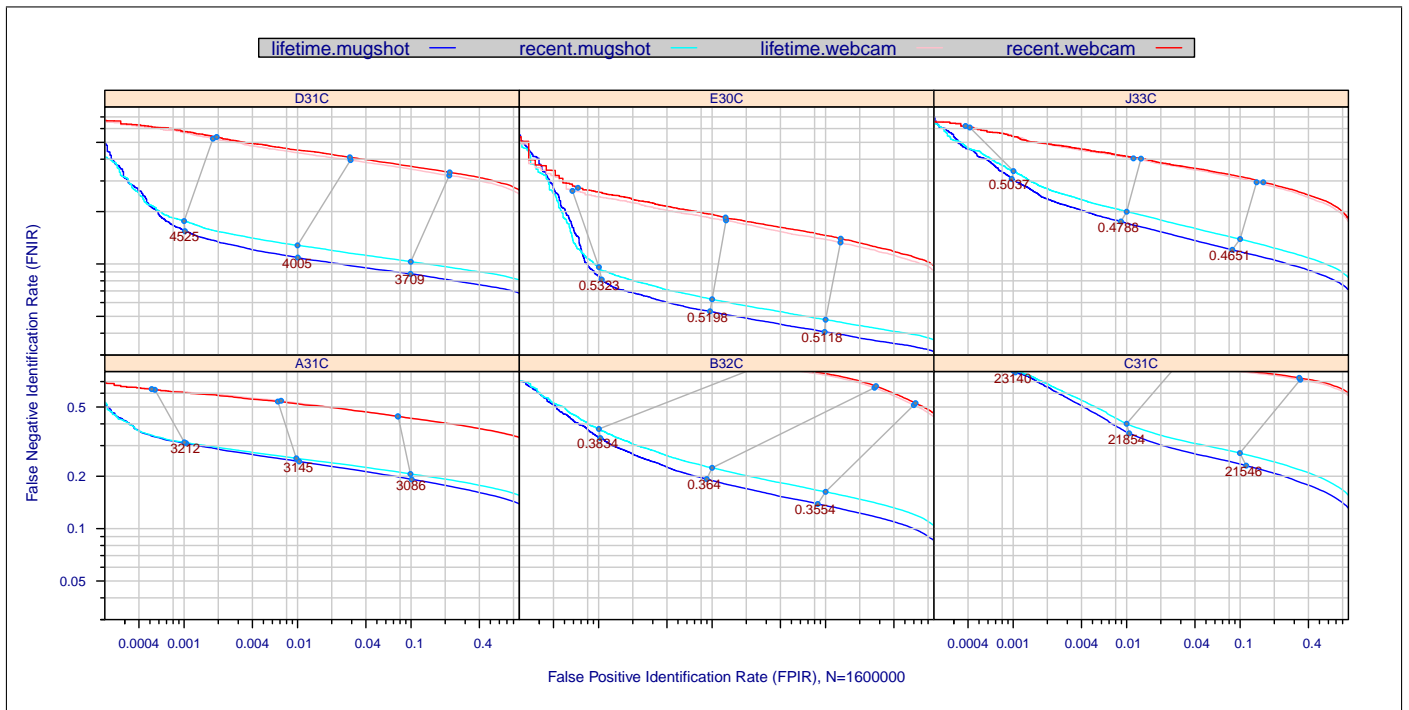


Figure 6: Identification error tradeoff for most accurate developers. For mugshot and webcam images searched against images of 1.6 million individuals enrolled in the “recent” and “lifetime” types of section 3.2, each panel plots FNIR (miss rate) vs. FPIR (false alarm rate) on log-log axes. Each panel presents data for the leading algorithm from the providers of the six most accurate algorithms. The (upper) red lines correspond to searching poor quality webcam images. The (lower) blue lines correspond to searching better quality mugshot images. The short grey lines link points of fixed threshold; they show how the error rates change with image and enrolment types. The threshold itself is shown in dark red. Non vertical lines indicate FPIR variation that a system operator would need to plan for. Performance data is tabulated in Tables 10 and 4-6. DETs for other algorithms appear in the “report cards” of Appendix A.

datasets. This applies across image types, mugshots and webcam, and enrolment types (sec: 3.2). FNIR values for NEC are a factor of two or more lower than for the closest competitors, Morpho and Toshiba. For example, when 1.6 million individuals’ lifetime mugshots are enrolled, the NEC’s rank one miss rate is 0.035 vs. Morpho’s 0.077. At rank 50, the NEC result is 0.023 with Toshiba at 0.049 and Morpho at 0.054. Similarly the margin for webcam images is larger still: Table 6 shows rank one miss rate for NEC is 0.11 vs. 0.24 for Toshiba and 0.30 for Morpho. Indeed the NEC algorithms yield webcam recognition accuracy better than many algorithms’ accuracy on mugshot images.

At the other end of the DET plot, however, the NEC algorithm exhibits rapid elevation in error rates at $FPIR < 0.001$. There the lifetime mugshot DETs of Figure 6 show that Morpho’s algorithm gains a modest advantage: at $FPIR = 0.0003$, the NEC miss rate is 0.374 while the Morpho value is 0.311. While the shape of the DETs is persuasive, this result is at the limits of statistical significance since the number of false positives, 51, supports a statement only that FPIR is, with 99.9% confidence, below 0.00044 and there, Morpho’s error rate is virtually tied with NECs: 0.248 vs. 0.252 respectively. Refinement of this comparison would require more nonmate searches to improve FPIR estimates.

For mugshots, Toshiba’s algorithms give slightly fewer misses than Cognitec’s which, in turn, produce error rates below those of 3M/Cogent. Finally Neurotechnology’s algorithms are beaten by those of Zhuhai Yisheng at $N < 640,000$ but not for $N = 1,600,000$ where the P30C algorithm fails substantially.

All algorithms are intolerant of webcam images, giving elevated miss rates, a factor of around three higher than for mugshots. This is particularly true for algorithms from Neurotechnology which give miss rates in excess of 0.6. As discussed earlier, results for webcam images have only niche operational relevance because most contemporary systems

A = 3M/Cogent	B = Cognitec	C = Neurotechnology	D = Safran Morpho	E = NEC	F = Tsinghua U.	FNIR(N,R,T,L) “Miss rate”
G = Hisign	H = CAS-IA	I = CAS-ICT	J = Toshiba	L = Tsinghua U. II	M = HP	FPIR(N,T,L) “False alarm rate”
P = Zhuhai-Yisheng	Q = JunYu	S = Decatur	T = Ayonix			

ENROLLED N=1600000	RANK ONE ACCURACY					
	SDK	YYYY-MM	FNIR(R=1)	FNIR(R=10)	FNIR(R=50)	FNIR(FPIR=0.002)
Cognitec	X21	2010-02	0.187	0.148	0.125	0.342
	B10C	2012-08	0.213	0.174	0.151	0.349
	B32C	2013-10	0.170	0.140	0.121	0.421
Neurotechnology	Z22	2010-02	0.258	0.219	0.194	0.537
	C11C	2012-08	0.268	0.232	0.208	0.952
	C20C	2013-03	0.267	0.229	0.206	0.684
	C31C	2013-10	0.231	0.193	0.170	0.685
Safran Morpho	W22	2010-02	0.135	0.113	0.103	0.248
	D10C	2012-08	0.145	0.126	0.120	0.233
	D20C	2013-03	0.124	0.106	0.100	0.194
	D31C	2013-10	0.121	0.103	0.099	0.195
NEC	V21	2010-02	0.089	0.063	0.050	0.269
	E10C	2012-08	0.164	0.121	0.098	0.410
	E20C	2013-03	0.072	0.055	0.047	0.132
	E30C	2013-10	0.064	0.051	0.046	0.108

Table 7: **Accuracy gains and losses 2010 to 2013:** For the four providers whose algorithms successfully completed the largest identification trial in the 2010 MBE evaluation [12], the table shows how accuracy has changed in the three and half years from early 2010 to late 2013. This is done for a fixed dataset of $N = 1.6$ million individuals each of whom is enrolled with $K \geq 1$ historical images. The number of mated searches is 40,000. The number of nonmate searches is also 40,000. The images are a subset of the LEO dataset drawn randomly in their natural mixture of approximately of 86% mugshots and 14% webcams.

target the ISO/IEC 19794-5 standard’s appearance requirements. That said the algorithm providers were aware that such images had been used in the 2010 test, and have access to images of this type in the MEDS Special Database 32 [5].

5.1.1 Face accuracy 2013 vs. 2010

Methods: Algorithms from providers who participated in both 2010 and 2013 were used to identify LEO images, repeating almost exactly the $N = 1.6$ million trial reported in MBE 2010 [12]. A small percentage of duplicate same-image pairs that were removed from the 2010 results were not removed in this study. The number of candidates was restricted to $L = 50$.

Results: Table 7 shows modest reductions in miss rates from 2010 to 2013.

Discussion: At rank 1, there will be approximately 10% fewer misses for the most accurate Cognitec, Neurotechnology and Morpho algorithms and nearly 30% fewer with NEC’s. At rank 50, the gains are more modest, 10% for Neurotechnology and NEC, and below 4% for Cognitec and Morpho. At the other end of the DET curve, where thresholds are applied to produce false positive outcomes in one in every 500 searches ($FPIR = 0.002$) the situation is mixed. While NEC realizes a full 60% reduction in miss rate, and Morpho a 20% reduction, both Neurotechnology and Cognitec appear to have *worse* accuracy. These observations ignore speed differences - particularly, the Cognitec algorithms in 2013 are considerably faster than those evaluated in 2010.

5.1.2 Cross-acquisition recognition

The prior subsection gave exhaustive tabulations of results for the general LEO population. There, mated pairs could be either mugshot-mugshot, webcam-webcam, or mugshot-webcam, and accuracy was only reported by whether the search image was a mugshot or a webcam. This section reports specifically on the accuracy of the various combinations. The experimental design is given in Table 8 and the results appear in Table 9.

For all the algorithms tested, the best recognition is obtained, as expected, by searching mugshots against enrolled mugshots. The worst recognition results, however, are not obtained by searching webcams against webcams, but instead from searching webcams against mugshots. This result is unfortunately contrary to the aspiration of frontal image

A = 3M/Cogent	B = Cognitec	C = Neurotechnology	D = Safran Morpho	E = NEC	F = Tsinghua U.	FNIR(N,R,T,L) “Miss rate”
G = Hisign	H = CAS-IA	I = CAS-ICT	J = Toshiba	L = Tsinghua U. II	M = HP	FPIR(N,T,L) “False alarm rate”
P = Zhuhai-Yisheng	Q = JunYu	S = Decatur	T = Ayonix			

	Enrolment					Search			
	Type	Num. IDs	Num. Images			Num. Images and IDs			
	See					Mate		Nonmate	
	See sec. 3.2	N	Webcam	Mugshot	Total	Webcam	Mugshot	Webcam	Mugshot
9	RECENT	0	10,660	0	10,660	10,660	0	28,936	171,066
10	RECENT	10,660	0	10,660	10,660	0	10,660	28,936	171,066
11	RECENT	10,660	0	10,660	10,660	10,660	0	28,936	171,066

Table 8: **Enrolment and search sets.** Each row summarizes one identification trial. The column labeled “Num. IDs” gives the number of enrolled identities. This precedes the numbers of images, and then the number of mate, and nonmate, searches. Rows 9-11 are dedicated to mugshot-mugshot, webcam-webcam, and mugshot-webcam recognition - the number of individuals here is limited to 10660 which corresponds to the number of individuals who were captured with webcams and traditional mugshot cameras.

ENROL	RANK 1			RANK 50		
	MUGSHOT	MUGSHOT	WEBCAM	MUGSHOT	MUGSHOT	WEBCAM
SEARCH	MUGSHOT	WEBCAM	WEBCAM	MUGSHOT	WEBCAM	WEBCAM
A31C	⁵ 0.092	³ 0.467	⁴ 0.264	⁵ 0.042	⁴ 0.187	⁴ 0.107
B32C	⁴ 0.082	⁶ 0.591	⁸ 0.476	⁴ 0.040	⁷ 0.344	⁸ 0.240
C31C	⁷ 0.125	⁹ 0.782	¹⁰ 0.617	⁷ 0.063	⁹ 0.472	¹¹ 0.428
D31C	³ 0.061	⁴ 0.489	³ 0.230	³ 0.036	³ 0.184	³ 0.105
E30C	¹ 0.025	¹ 0.104	¹ 0.066	¹ 0.018	¹ 0.031	¹ 0.026
G31C	⁸ 0.182	¹²	⁵ 0.343	⁸ 0.075	¹²	⁵ 0.153
H30C	⁹ 0.271	⁸ 0.762	⁷ 0.428	⁹ 0.136	⁸ 0.389	⁷ 0.223
J32C	² 0.052	² 0.236	² 0.153	² 0.023	² 0.067	² 0.051
L31C	¹¹ 0.358	¹⁰ 0.913	¹¹ 0.629	¹¹ 0.188	¹⁰ 0.714	¹⁰ 0.389
M30C	¹⁰ 0.298	⁷ 0.670	⁹ 0.521	¹⁰ 0.142	⁶ 0.341	⁹ 0.280
P30C	⁶ 0.107	⁵ 0.530	⁶ 0.344	⁶ 0.045	⁵ 0.242	⁶ 0.161
S20C	¹² 0.492	¹¹ 0.968	¹² 0.789	¹² 0.306	¹¹ 0.890	¹² 0.604

Table 9: **Mugshots, webcams, and interoperable recognition:** The figures are miss rates, $FNIR(N, R, T, L)$ with $N = 10,660$, $R = \{1, 50\}$, $L = 50$, $T = 0$, for enrolment using the image type given on row 2 and search image type given on row 3. The blue superscripts indicate algorithm rankings, by column. Missing values are due to software crashes.

standards (i.e. ISO/IEC 19794-5), which aims to support best accuracy by establishing full-frontal geometry requirements for all images. Here recognition accuracy is better if either both images are full-frontal (i.e. mugshots) or both are not (i.e. webcams). As soon as there is heterogeneity (i.e. mugshot-webcam) recognition accuracy degrades. The specific cause of this result is that while the webcam images are inferior, any given pair can share the same non-frontal pose (typically an adverse pitch angle). Further study of this issue might quantify the relative roles of pose angle, optical resolution and illumination.

5.2 Accuracy dependence on rank

Identification algorithms yield candidate lists where the mate is sometimes not at rank 1 because it has a low score relative to some nonmates. This occurs typically because the face is of poor quality or has some difference to the enrolment image. While rank-1 performance is a primary indicator for comparison of algorithms, a more relevant metric for applications where a human reviewer is retained to adjudicate an L element candidate list, is the frequency at which mates can be found at ranks up to rank L . In this report we quantify this as the miss rate $FNIR(N, R, 0, L)$ for $R \rightarrow L$, and we fix $L = 50$. The workload associated with reviewer adjudication of candidate lists is discussed further in 5.7.

Method: The benefit of considering higher ranks is quantified by re-analyzing results from the same LEO images and enrolment and search sets given in rows 1-8 of Table 3 i.e. for N up to 1.6 million identities.

Results: Previously, Tables 4 - 6 had given some $FNIR$ values for, respectively, $N = 160K$, $640K$ and $1.6M$. A visualization of those results for $N = 640K$ appears in Figure 8 which shows $FNIR$ vs. R on log-log axes.

A = 3M/Cogent	B = Cognitec	C = Neurotechnology	D = Safran Morpho	E = NEC	F = Tsinghua U.	FNIR(N,R,T,L) “Miss rate”
G = Hisign	H = CAS-IA	I = CAS-ICT	J = Toshiba	L = Tsinghua U. II	M = HP	FPIR(N,T,L) “False alarm rate”
P = Zhuhai-Yisheng	Q = JunYu	S = Decatur	T = Ayonix			

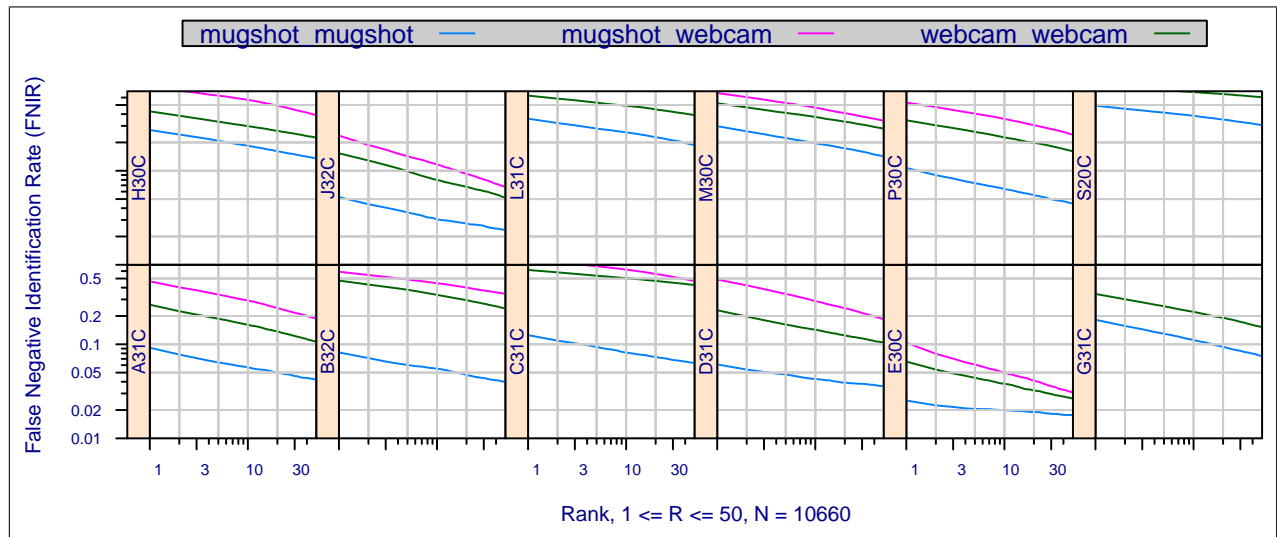


Figure 7: **Identification accuracy by acquisition type.** The figures plot miss rate vs. rank i.e. $FNIR(N, R, 0, L)$ with $N = 10,660$ and $L = 50$. Each panel corresponds to one algorithm. The three traces correspond to the use of the identified acquisition type for the enrolment and search image respectively. Notably, cross-type recognition gives the worst accuracy.

Discussion: The following observations are notable.

- ▷ **Magnitude of accuracy gains:** By definition, miss rates decrease monotonically with R , and the size of the effect can readily extend to a factor of two reduction in FNIR from $R = 1$ to $R = 50$.
- ▷ **Differences across algorithms:** For enrolment of $N = 640K$ “lifetime” identities, the A31C algorithm gives a miss rate of 0.071 at rank 50 vs. 0.139 at rank 1 for a miss rate reduction factor of 2.0. For B32C, the values are 0.061 vs. 0.104 for a gain of 1.7. For C31C, the values are 0.103 vs. 0.162 implying a gain of 1.6. For D31C, the factor is 1.4 from 0.051 vs. 0.071. For the most accurate algorithm overall E30C, FNIR reduces to 0.021 from 0.032 for a gain of 1.5. Finally for J33C, the gain is 1.9 as FNIR dropped to 0.042 from 0.081. These trends are evident in Figure 8. These ratios broadly apply to other algorithms from the respective suppliers. The more accurate algorithms typically realize fewer gains at rank 50 because they place the mate at rank 1 more frequently. The result, however, is that the J33C algorithm is inferior to the D-series algorithms at rank 1, but superior at rank 50.
- ▷ **Power-law dependency:** The observation that many of the plots in Figure 8 are straight lines is discussed later in section 5.3.1. Note the plots for the D-series and F-series algorithms are not straight.

5.3 Effect of population size

How algorithmic performance degrades with increasing population size is a primary challenge in any application where individuals are enrolled at a greater rate than they are un-enrolled. This scalability issue is examined in this section by re-analyzing results from the same LEO images and enrolment and search sets given in rows 1-8 of Table 3 i.e. for N up to 1.6 million identities.

The results are presented in four ways as follows. This includes results for both mugshot and webcam searches, and both recent and lifetime enrolment types.

A = 3M/Cogent	B = Cognitec	C = Neurotechnology	D = Safran Morpho	E = NEC	F = Tsinghua U.	FNIR(N,R,T,L) “Miss rate”
G = Hisign	H = CAS-IA	I = CAS-ICT	J = Toshiba	L = Tsinghua U. II	M = HP	FPIR(N,T,L) “False alarm rate”
P = Zhuhai-Yisheng	Q = JunYu	S = Decatur	T = Ayonix			

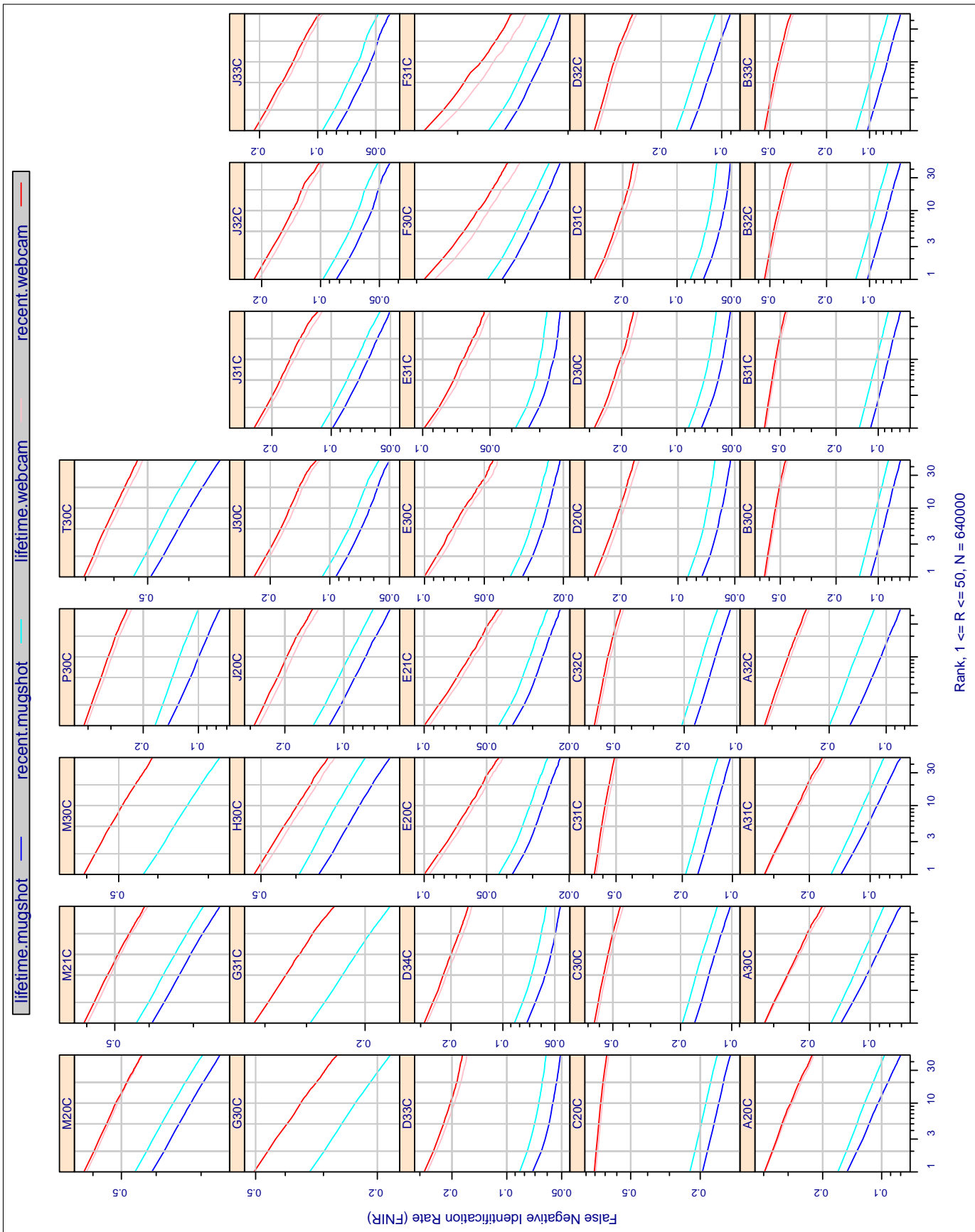
Rank, $1 \leq R \leq 50$, $N = 640000$

Figure 8: **Identification accuracy gains with human search for mates at poor rank.** The figures plot miss rate vs. rank i.e. $FNIR(N, R, 0, L)$ with $N = 640,000$, and $L = 50$. Each panel corresponds to one algorithm. Note that because the vertical scales vary for each algorithm and the scale is logarithmic, the visually perceived gradients of the lines are not comparable (see Tables 4-6 for tabulated values). The (upper) red lines correspond to searching poor quality webcam images. The (lower) blue lines correspond to searching better quality mugshot images.

A = 3M/Cogent
G = Hisign
P = Zhuhai-Yisheng

B = Cognitec
H = CAS-IA
Q = JunYu

C = Neurotechnology
I = CAS-ICT
S = Decatur

D = Safran Morpho
J = Toshiba
T = Ayonix

E = NEC
L = Tsinghua U. II

F = Tsinghua U.
M = HP

$FNIR(N, R, T, L)$ "Miss rate"
 $FPFR(N, T, L)$ "False alarm rate"

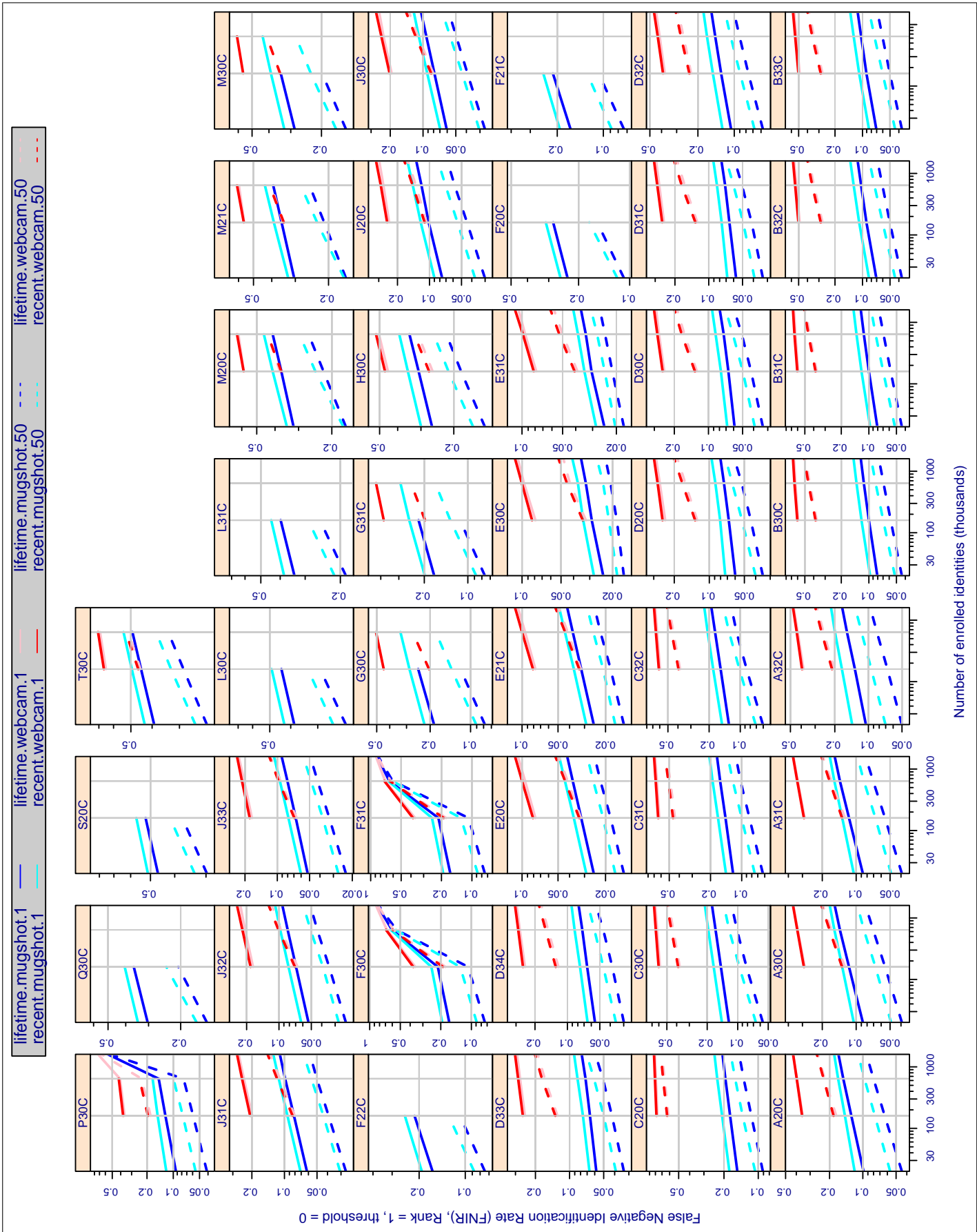


Figure 9: **Identification accuracy declines with the number of enrolled subjects.** The figures plot miss rate vs. population size, N , i.e. $FNIR(N, R, 0, L)$ with $R = 50$ (dotted lines) and $R = 1$ (solid). Note the vertical scales vary for each algorithm. Each panel corresponds to one algorithm. The (upper) red lines correspond to searching better quality mugshot images. The (lower) blue lines correspond to searching better quality webcam images.

ENROL	ALL LIFETIME PHOTOS							MOST RECENT PHOTO				
RANK=1	WEBCAM			MUGSHOT				VISA	MUGSHOT			
ALG	N=160K	N=640K	N=1600K	N=20K	N=160K	N=640K	N=1600K	N=20K	N=20K	N=160K	N=640K	N=1600K
A20C	0.351	0.395	¹⁷ 0.429	0.099	0.127	0.150	²² 0.166	¹⁶ 0.101	0.119	0.145	0.167	²² 0.182
A30C	0.286	0.331	¹⁶ 0.362	0.087	0.116	0.139	²¹ 0.155	¹¹ 0.084	0.105	0.133	0.156	²⁰ 0.172
A31C	0.287	0.331	¹⁵ 0.362	0.087	0.116	0.139	²⁰ 0.155	¹² 0.084	0.105	0.133	0.156	¹⁹ 0.172
A32C	0.382	0.431	¹⁹ 0.467	0.105	0.132	0.156	²³ 0.172	²⁴ 0.124	0.140	0.173	0.200	²⁵ 0.216
B30C	0.592	0.634	²⁴ 0.658	0.081	0.100	0.113	¹⁷ 0.123	²³ 0.123	0.097	0.120	0.136	¹⁷ 0.147
B31C	0.592	0.634	²³ 0.658	0.081	0.100	0.113	¹⁶ 0.123	²² 0.123	0.097	0.120	0.137	¹⁶ 0.147
B32C	0.488	0.531	²¹ 0.560	0.071	0.091	0.104	¹⁵ 0.113	¹⁸ 0.106	0.085	0.109	0.125	¹⁴ 0.136
B33C	0.488	0.531	²⁰ 0.560	0.071	0.091	0.104	¹⁴ 0.113	¹⁷ 0.106	0.085	0.109	0.125	¹⁵ 0.136
C20C	0.768	0.793	²⁸ 0.810	0.153	0.177	0.194	²⁷ 0.206	³³ 0.244	0.179	0.208	0.229	²⁷ 0.243
C30C	0.585	0.624	²⁷ 0.651	0.122	0.146	0.165	²⁵ 0.179	²⁶ 0.213	0.143	0.173	0.195	²⁴ 0.211
C31C	0.624	0.659	²⁶ 0.685	0.121	0.143	0.162	²⁴ 0.173	²⁵ 0.200	0.142	0.169	0.191	²³ 0.205
C32C	0.596	0.637	²⁵ 0.664	0.129	0.155	0.175	²⁶ 0.190	²⁸ 0.221	0.150	0.183	0.207	²⁶ 0.223
D20C	0.239	0.267	⁹ 0.287	0.060	0.068	0.075	⁹ 0.080	⁹ 0.066	0.073	0.081	0.089	⁹ 0.096
D30C	0.245	0.273	¹⁰ 0.290	0.059	0.067	0.073	⁸ 0.080	⁵ 0.061	0.068	0.078	0.087	⁸ 0.093
D31C	0.251	0.276	¹² 0.295	0.057	0.065	0.071	⁵ 0.077	⁷ 0.064	0.068	0.077	0.084	⁵ 0.091
D32C	0.376	0.415	¹⁸ 0.445	0.106	0.128	0.144	¹⁹ 0.155	²⁰ 0.115	0.125	0.150	0.168	²¹ 0.180
D33C	0.248	0.274	¹¹ 0.291	0.058	0.066	0.072	⁶ 0.078	⁶ 0.062	0.067	0.076	0.085	⁶ 0.091
D34C	0.248	0.277	¹³ 0.296	0.055	0.066	0.073	⁷ 0.079	⁸ 0.064	0.065	0.076	0.086	⁷ 0.092
E20C	0.077	0.094	⁴ 0.108	0.025	0.032	0.038	⁴ 0.042	⁴ 0.026	0.029	0.037	0.044	⁴ 0.049
E21C	0.077	0.094	³ 0.108	0.025	0.032	0.038	³ 0.042	³ 0.026	0.029	0.037	0.044	³ 0.049
E30C	0.079	0.097	² 0.108	0.024	0.029	0.032	¹ 0.035	¹ 0.017	0.028	0.034	0.037	¹ 0.041
E31C	0.079	0.095	¹ 0.106	0.025	0.031	0.034	² 0.037	² 0.023	0.028	0.035	0.038	² 0.042
F20C	0.465		³⁷	0.232	0.282		³⁷	³⁶ 0.278	0.256	0.312		³⁶
F30C	0.356	0.678	²⁹ 0.867	0.165	0.214	0.603	²⁹ 0.843	²⁹ 0.232	0.194	0.247	0.618	³⁴
F31C	0.365	0.720	³⁰ 0.886	0.163	0.214	0.656	³⁰ 0.864	³² 0.239	0.192	0.248	0.670	⁴³
G30C	0.423		⁴⁴	0.187	0.245		⁴⁴	³⁹ 0.308	0.220	0.285	0.333	⁴⁴
G31C	0.369		³⁸	0.171	0.219		³⁸	²⁷ 0.214	0.198	0.254	0.293	³⁷
H30C	0.451	0.504	³⁵	0.262	0.309	0.346	³⁵	³⁸ 0.285	0.300	0.353	0.392	³²
J20C	0.240	0.272	¹⁴ 0.297	0.076	0.102	0.119	¹⁸ 0.133	¹⁹ 0.114	0.090	0.121	0.143	¹⁸ 0.161
J30C	0.194	0.231	⁷ 0.253	0.061	0.079	0.093	¹² 0.104	¹⁵ 0.097	0.070	0.091	0.109	¹² 0.122
J31C	0.199	0.237	⁸ 0.265	0.062	0.082	0.099	¹³ 0.110	¹⁴ 0.092	0.071	0.094	0.113	¹³ 0.127
J32C	0.174	0.207	⁵ 0.229	0.053	0.069	0.083	¹¹ 0.093	¹³ 0.088	0.062	0.081	0.097	¹¹ 0.108
J33C	0.173	0.204	⁶ 0.229	0.052	0.067	0.081	¹⁰ 0.091	¹⁰ 0.082	0.060	0.079	0.095	¹⁰ 0.107
L30C	0.691		⁴¹	0.385	0.442		⁴¹	⁴² 0.348	0.428	0.490		⁴⁰
L31C	0.640		⁴⁰	0.339	0.398		⁴⁰	⁴¹ 0.332	0.382	0.444		³⁹
M20C	0.586	0.625	³²	0.319	0.371	0.412	³²	³⁷ 0.279	0.346	0.412	0.458	²⁹
M21C	0.561	0.598	³⁹	0.304	0.355	0.393	³⁹	³⁴ 0.273	0.327	0.390	0.435	³⁸
M30C	0.547		³³	0.285	0.342		³³	³⁵ 0.273	0.327	0.390	0.435	³⁰
P30C	0.361	0.408	²⁷ 0.728	0.093	0.123	0.147	²⁸ 0.564	²¹ 0.117	0.120	0.151	0.172	³³
Q30C	0.551		⁴²	0.303	0.361		⁴²	⁴⁰ 0.309	0.343	0.405		⁴¹
S20C	0.796		⁴³	0.467	0.523		⁴³	⁴⁴ 0.420	0.512	0.569		⁴²
T30C	0.650	0.692	³⁴	0.391	0.449	0.492	³⁴	⁴³ 0.351	0.432	0.495	0.540	³¹

Table 10: **Rank-1 miss rates:** The table gives FNIR values for different enrolled population sizes, three different image types, and two enrolment strategies. At left are seven columns indicating accuracy for searches of legacy webcam images and mugshots; these results apply where all but the most recent image of an individual is enrolled, and the last image is searched. On the right side, only the penultimate image is enrolled, and the last image is searched (see section 3.2). The column in yellow is highlighted because it applies to well posed visa images of individuals including children. All other columns apply to images drawn from the LEO set. The blue superscripts are column-wise algorithm rankings. Empty cells indicate that the run was not attempted, or was unsuccessful.

ENROL	ALL LIFETIME PHOTOS							MOST RECENT PHOTO				
RANK=50	WEBCAM			MUGSHOT				VISA	MUGSHOT			
ALG	N=160K	N=640K	N=1600K	N=20K	N=160K	N=640K	N=1600K	N=20K	N=20K	N=160K	N=640K	N=1600K
A20C	0.174	0.222	¹⁷ 0.255	0.044	0.063	0.080	²¹ 0.093	¹⁶ 0.033	0.057	0.078	0.097	²¹ 0.110
A30C	0.126	0.167	¹⁵ 0.198	0.039	0.055	0.071	¹⁹ 0.083	⁵ 0.019	0.049	0.067	0.085	¹⁹ 0.098
A31C	0.126	0.167	¹⁶ 0.198	0.039	0.055	0.071	²⁰ 0.083	⁶ 0.019	0.049	0.067	0.086	²⁰ 0.098
A32C	0.203	0.253	¹⁸ 0.288	0.050	0.068	0.084	²² 0.096	¹⁸ 0.039	0.070	0.094	0.115	²³ 0.131
B30C	0.371	0.443	²⁵ 0.491	0.044	0.058	0.069	¹⁸ 0.078	²⁴ 0.053	0.052	0.070	0.084	¹⁸ 0.094
B31C	0.371	0.443	²⁴ 0.491	0.044	0.058	0.069	¹⁷ 0.078	²³ 0.053	0.052	0.070	0.084	¹⁷ 0.094
B32C	0.274	0.341	²¹ 0.383	0.037	0.050	0.061	¹⁶ 0.069	²⁰ 0.042	0.044	0.060	0.074	¹⁵ 0.084
B33C	0.274	0.341	²⁰ 0.383	0.037	0.050	0.061	¹⁵ 0.069	¹⁹ 0.042	0.044	0.060	0.074	¹⁶ 0.084
C20C	0.615	0.667	²⁸ 0.700	0.091	0.115	0.134	²⁷ 0.147	³⁹ 0.143	0.107	0.137	0.160	²⁷ 0.174
C30C	0.373	0.433	²² 0.478	0.057	0.082	0.102	²⁵ 0.115	³¹ 0.106	0.066	0.096	0.122	²⁵ 0.137
C31C	0.441	0.490	²⁶ 0.528	0.063	0.084	0.103	²⁴ 0.114	³⁰ 0.102	0.075	0.100	0.122	²⁴ 0.136
C32C	0.382	0.444	²³ 0.489	0.060	0.087	0.109	²⁶ 0.122	³³ 0.109	0.071	0.102	0.128	²⁶ 0.145
D20C	0.127	0.161	¹¹ 0.188	0.035	0.044	0.053	¹³ 0.061	¹¹ 0.021	0.040	0.052	0.064	¹³ 0.072
D30C	0.127	0.163	¹² 0.189	0.033	0.042	0.051	¹² 0.059	⁸ 0.020	0.039	0.050	0.061	¹¹ 0.071
D31C	0.127	0.164	¹⁴ 0.191	0.033	0.042	0.051	¹⁰ 0.059	⁷ 0.020	0.039	0.049	0.060	¹⁰ 0.070
D32C	0.228	0.264	¹⁹ 0.292	0.058	0.076	0.090	²³ 0.101	²² 0.049	0.068	0.089	0.107	²² 0.119
D33C	0.128	0.166	¹³ 0.191	0.033	0.042	0.051	¹¹ 0.059	⁹ 0.020	0.039	0.050	0.061	¹² 0.071
D34C	0.121	0.151	¹⁰ 0.174	0.030	0.038	0.046	⁷ 0.054	¹⁰ 0.020	0.035	0.046	0.056	⁷ 0.064
E20C	0.032	0.042	³ 0.049	0.014	0.018	0.022	³ 0.025	³ 0.007	0.017	0.021	0.025	³ 0.028
E21C	0.032	0.042	² 0.049	0.014	0.018	0.022	² 0.025	² 0.007	0.017	0.021	0.025	² 0.028
E30C	0.032	0.043	¹ 0.048	0.016	0.019	0.021	¹ 0.023	¹ 0.006	0.018	0.021	0.024	¹ 0.026
E31C	0.039	0.050	⁴ 0.058	0.018	0.021	0.024	⁴ 0.026	⁴ 0.007	0.020	0.024	0.028	⁴ 0.031
F20C	0.251		³⁷	0.108	0.155		³⁷	³⁷ 0.130	0.118	0.173		³⁶
F30C	0.180	0.584	²⁹ 0.831	0.074	0.108	0.545	²⁹ 0.820	²⁵ 0.082	0.089	0.129	0.555	³⁴
F31C	0.180	0.636	³⁰ 0.855	0.074	0.108	0.606	³⁰ 0.844	²⁸ 0.085	0.088	0.128	0.616	⁴³
G30C	0.196		⁴⁴	0.078	0.119		⁴⁴	³⁶ 0.123	0.091	0.139	0.182	⁴⁴
G31C	0.187		³⁸	0.076	0.113		³⁸	²⁹ 0.087	0.088	0.132	0.168	³⁷
H30C	0.260	0.312	³⁵	0.136	0.183	0.219	³⁵	³⁸ 0.132	0.158	0.214	0.257	³²
J20C	0.107	0.134	⁹ 0.160	0.030	0.044	0.058	¹⁴ 0.068	¹⁷ 0.034	0.037	0.055	0.071	¹⁴ 0.083
J30C	0.080	0.111	⁷ 0.135	0.027	0.039	0.050	⁸ 0.058	¹⁵ 0.029	0.031	0.044	0.056	⁸ 0.067
J31C	0.080	0.110	⁸ 0.135	0.027	0.039	0.050	⁹ 0.058	¹³ 0.023	0.031	0.044	0.056	⁹ 0.067
J32C	0.067	0.096	⁵ 0.117	0.024	0.034	0.044	⁶ 0.051	¹⁴ 0.024	0.028	0.039	0.051	⁶ 0.059
J33C	0.066	0.094	⁶ 0.117	0.023	0.033	0.042	⁵ 0.049	¹² 0.021	0.027	0.038	0.048	⁵ 0.057
L30C	0.489		⁴¹	0.222	0.293		⁴¹	⁴³ 0.173	0.254	0.334		⁴⁰
L31C	0.433		⁴⁰	0.188	0.253		⁴⁰	⁴¹ 0.161	0.214	0.290		³⁹
M20C	0.377	0.436	³²	0.169	0.221	0.266	³²	³² 0.108	0.174	0.245	0.297	²⁹
M21C	0.349	0.403	³⁹	0.162	0.211	0.252	³⁹	³⁴ 0.109	0.165	0.230	0.281	³⁸
M30C	0.333		³³	0.145	0.198		³³	³⁵ 0.109	0.165	0.230	0.281	³⁰
P30C	0.182	0.230	²⁷ 0.663	0.041	0.059	0.076	²⁸ 0.530	²¹ 0.043	0.056	0.078	0.100	³³
Q30C	0.305		⁴²	0.144	0.206		⁴²	⁴⁰ 0.148	0.164	0.240		⁴¹
S20C	0.639		⁴³	0.298	0.373		⁴³	⁴⁴ 0.237	0.333	0.418		⁴²
T30C	0.452	0.514	³⁴	0.225	0.288	0.338	³⁴	⁴² 0.163	0.254	0.328	0.384	³¹

Table 11: **Rank-50 miss rates:** The table gives FNIR values for different enrolled population sizes, three different image types, and two enrolment strategies. At left are seven columns indicating accuracy for searches of legacy webcam images and mugshots; these results apply where all but the most recent image of an individual is enrolled, and the last image is searched. On the right side, only the penultimate image is enrolled, and the last image is searched (see section 3.2). The column in yellow is highlighted because it applies to well posed visa images of individuals including children. All other columns apply to images drawn from the LEO set. The blue superscripts are column-wise algorithm rankings. Empty cells indicate that the run was not attempted, or was unsuccessful.

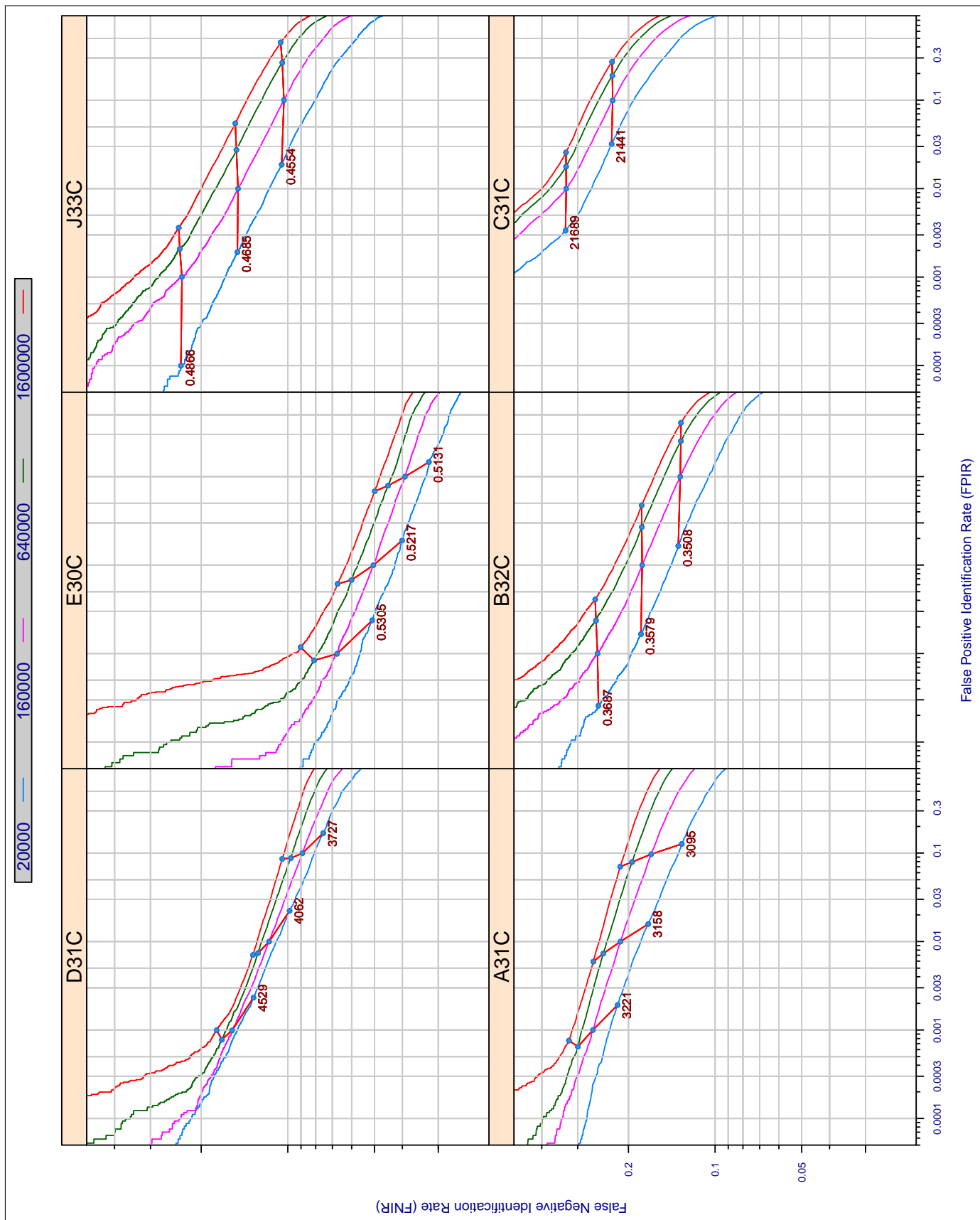


Figure 10: **Identification error tradeoff for increasing numbers of enrolled subjects.** For mugshots searched against the most recent image of an individual, the figures plot FNIR (miss rate) vs. FPIR (false alarm rate) for four population sizes, $N=20K, 160K, 640K, 1.6M$. Each panel corresponds to one algorithm. The short upward lines link points of fixed threshold; they show how the error rates change with N . Conventional (binomial) models of identification, which regard a $1:N$ search as N $1:1$ comparisons, indicate that only FPIR increases with N (e.g. J33C, B32C). However, some algorithms attempt to stabilize FPIR against increases in N (e.g. D31C, E30C) - and this relieves a system operator of the need to elevate threshold with increasing N .

- ▷ **Tabulation:** $\text{FNIR}(N, R, 0, L)$ is tabulated exhaustively for four population sizes and, for $R = 1, 50$, in Tables 10 and 11.
- ▷ **Graphs:** This same data is plotted as $\text{FNIR}(N, R, 0, L)$ against N in Figure 9.
- ▷ **DETs:** Detection error tradeoff characteristics appear for the more accurate algorithms in Figure 10. These plot $\text{FNIR}(N, L, T, L)$ vs. $\text{FPIR}(N, T, L)$ for the four population sizes, and all thresholds.
- ▷ **Models:** The rank-based $\text{FNIR}(N, R, 0, L)$ data is modeled using illustrative power-law models.

Discussion: It is clear from Figure 9 that mated search miss rates increase with enrolled population size and decrease with rank. Most importantly, miss rates usually only increase slowly with population size. While this aspect is common to other biometric modalities, it alone is largely responsible for the operational utility of face identification algorithms. Specifically, with a ten-fold population size increase, from 160,000 to 1,600,000, the rank one miss rates only increase by small factors, 1.2 for E30C (0.029 to 0.035, Table 10), 1.1 for D31C (0.065 to 0.071), and 1.4 for J33C (0.067 to 0.091). This result is modeled in the next section.

The F and P algorithms have problems at the largest population size; while this may be a property of the underlying mathematics, it may also be a (software) implementation issue.

When algorithms are configured with high thresholds to curtail false positives, the dependence on N is mostly similar. However, Figure 10 shows that, at $\text{FPIR} = 0.002$, there is reduced sensitivity of FNIR on N . For example for the D31C algorithm, as population size increases from 160,000 to 1,600,000, FNIR increases from 0.122 to 0.134, a factor of 1.1 (lifetime mugshots, Tables 4 and 6).

Figure 10 also shows lines connecting DETs corresponding to fixed thresholds. These reveal various dependencies of FPIR on N ; this is discussed further in section 5.3.2.

5.3.1 Models of FNIR dependence on N and R

Many plots in Figure 9 are approximately linear on log-log axes. Moreover, Figure 8 shows a similar linear reduction in FNIR with rank, again on log-log axes. These observations motivate development here of FNIR models i.e. empirical formulae giving miss rate as a function of population size, N , and rank, R . The goal here is *not* to produce a definitive and usable model of accuracy but instead to highlight sublinear scalability effects, and to support reasoning about how accuracy varies.

For any given dataset, the coarse empirical observation is that $\log \text{FNIR}$ is approximately linear in $\log N$ and $\log R$. This corresponds to a power-law model:

$$\text{FNIR}(N, R, 0, L) = aN^bR^c \quad (8)$$

where the coefficient a is an implied recognition rate for $N = R = 1$, b is a scalability exponent, and c quantifies the occurrence of mates at higher (poorer) rank. Note that this model applies to the investigational application where the threshold is relaxed $T = 0^8$, and the requested number of candidates, L , is assumed to not affect the candidates appearing at rank $R = 1 \dots L$.

This model does not exhibit the correct behavior for the case $N \rightarrow R$ and, more seriously, does not capture the behavior seen (e.g. for algorithm D20C) showing that searching to rank 50 is less worthwhile relative to rank 1 at large population

⁸More advanced models of how accuracy scales with N have been proposed. These include threshold as an independent variable - the goal being to assist in setting thresholds as the enrolled population increases. These models [4, 13, 15, 16, 23] are beyond our purpose here.

A = 3M/Cogent	B = Cognitec	C = Neurotechnology	D = Safran Morpho	E = NEC	F = Tsinghua U.	$\text{FNIR}(N, R, T, L)$ "Miss rate"
G = Hisign	H = CAS-IA	I = CAS-ICT	J = Toshiba	L = Tsinghua U. II	M = HP	$\text{FPIR}(N, T, L)$ "False alarm rate"
P = Zhuhai-Yisheng	Q = JunYu	S = Decatur	T = Ayonix			

sizes. This implies that the scaling exponent b should itself be dependent on R .

$$\text{FNIR}(N, R, 0, L) = a(N - R)^{b-d \log R} R^c \quad (9)$$

The models of equations 8 and 9 are fit to the available empirical FNIR data for each algorithm, for the “recent” enrolment type, and for mugshot and webcam acquisition types. The exponents of the fits are tabulated in Table 12. The result show the scalability exponents of equation 8 typically has values $0.08 < b < 0.16$ indicating very sublinear growth with N . The models also reveal that human adjudication of long candidate lists (to rank R) offers diminishing returns - the exponent c in eq. 8 has values $-0.18 < c < -0.09$. We caution against using any of these values to compare algorithms since the models are inexact and lack theoretical support.

5.3.2 Dependence of FPIR on N

Figure 10 shows lines connecting DETs corresponding to fixed thresholds. These reveal that algorithms exhibit one of two broad dependencies of FPIR on N :

- ▷ **Linear:** Some algorithms exhibit an almost linear dependence of FPIR, while others have some invariance to population size. In classical biometric theory, $\text{FPIR}(N, T) = N \text{FMR}(T)$ where FMR is the one-to-one false match probability. This is an approximation to the Binomial model of independent failures $\text{FPIR} = 1 - (1 - \text{FMR})^N$ with small FMR. The linear dependence of FPIR on N approximately holds for algorithms from Cognitec, Neurotechnology and Toshiba. Implementations that implement 1:N search as N 1:1 comparisons would exhibit this behavior.
- ▷ **Constant:** Other algorithms, including those from 3M/Cogent, Morpho and NEC, exhibit $\text{FPIR}(N, T)$ having a complicated dependence on N , with FPIR being approximately constant (independent of N), or in some cases even reducing at larger population sizes. Such behavior can be achieved via score-normalization [15] and by implementing results from extreme value theory - the statistics of the largest of N random samples.

In any case, owners of systems that employ non-zero thresholds will need to understand the dependence of false positive rates on population size - the goal being to have *predictable* FPIR. Either approach is viable given adequate documentation and planning.

5.4 Effect of enrolling all historical images

Methods: The LIFETIME enrolment type is defined by enrolment of K_i images per person as described in section 3.2. By executing the standard set of searches, FNIR can be computed over searches where the mate was enrolled with $K_i = 1, 2, \dots$ enrolled images.

Results: Substantially reduced miss rates are measured for all algorithms as functions of K .

Discussion: The overall FNIR values reported previously for the LIFETIME enrolments represent the maximum realizable accuracy given this data. The relevance of the results in this section is on system design policy: operators should plan to store, enrol, and use the full lifetime history of images from an individual. The results here do not, however, answer the question of how old an image should be before it is retired from use. While, the default guidance here is to retain *all* images regardless of capture date, an ageing study is indicated to quantify un-enrolment schedules. Such a study would need to quantify false positive consequences of retaining too many images.

A = 3M/Cogent	B = Cognitec	C = Neurotechnology	D = Safran Morpho	E = NEC	F = Tsinghua U.	FNIR(N,R,T,L) “Miss rate”
G = Hisign	H = CAS-IA	I = CAS-ICT	J = Toshiba	L = Tsinghua U. II	M = HP	FPIR(N,T,L) “False alarm rate”
P = Zhuhai-Yisheng	Q = JunYu	S = Decatur	T = Ayonix			

ALGORITHM	SIMPLE MODEL OF EQUATION 8		EXTENDED MODEL OF EQUATION 9	
	RECENT MUGSHOT	RECENT WEBCAM	RECENT MUGSHOT	RECENT WEBCAM
A20C	$0.034 N^{0.120} R^{-0.145}$	$0.092 N^{0.109} R^{-0.140}$	$0.044 (N - R)^{0.099+0.014 \log R} R^{-0.330}$	$0.137 (N - R)^{0.080+0.019 \log R} R^{-0.392}$
A30C	$0.027 N^{0.132} R^{-0.158}$	$0.060 N^{0.129} R^{-0.165}$	$0.034 (N - R)^{0.114+0.012 \log R} R^{-0.315}$	$0.090 (N - R)^{0.098+0.021 \log R} R^{-0.445}$
A31C	$0.027 N^{0.132} R^{-0.158}$	$0.060 N^{0.129} R^{-0.165}$	$0.034 (N - R)^{0.114+0.012 \log R} R^{-0.315}$	$0.090 (N - R)^{0.098+0.021 \log R} R^{-0.444}$
A32C	$0.041 N^{0.118} R^{-0.142}$	$0.105 N^{0.107} R^{-0.130}$	$0.052 (N - R)^{0.101+0.011 \log R} R^{-0.290}$	$0.152 (N - R)^{0.080+0.017 \log R} R^{-0.362}$
B30C	$0.031 N^{0.110} R^{-0.126}$	$0.237 N^{0.076} R^{-0.089}$	$0.038 (N - R)^{0.095+0.010 \log R} R^{-0.251}$	$0.359 (N - R)^{0.045+0.018 \log R} R^{-0.329}$
B31C	$0.031 N^{0.110} R^{-0.126}$	$0.237 N^{0.076} R^{-0.089}$	$0.038 (N - R)^{0.095+0.010 \log R} R^{-0.251}$	$0.359 (N - R)^{0.045+0.018 \log R} R^{-0.329}$
B32C	$0.024 N^{0.122} R^{-0.137}$	$0.164 N^{0.091} R^{-0.109}$	$0.030 (N - R)^{0.107+0.010 \log R} R^{-0.262}$	$0.261 (N - R)^{0.056+0.021 \log R} R^{-0.392}$
B33C	$0.024 N^{0.122} R^{-0.137}$	$0.163 N^{0.091} R^{-0.109}$	$0.030 (N - R)^{0.107+0.010 \log R} R^{-0.263}$	$0.261 (N - R)^{0.056+0.021 \log R} R^{-0.393}$
C20C	$0.072 N^{0.088} R^{-0.105}$	$0.504 N^{0.036} R^{-0.043}$	$0.090 (N - R)^{0.070+0.011 \log R} R^{-0.237}$	$0.625 (N - R)^{0.019+0.009 \log R} R^{-0.157}$
C30C	$0.042 N^{0.116} R^{-0.131}$	$0.248 N^{0.072} R^{-0.090}$	$0.061 (N - R)^{0.087+0.019 \log R} R^{-0.377}$	$0.358 (N - R)^{0.044+0.016 \log R} R^{-0.302}$
C31C	$0.047 N^{0.106} R^{-0.121}$	$0.319 N^{0.056} R^{-0.071}$	$0.062 (N - R)^{0.083+0.014 \log R} R^{-0.302}$	$0.403 (N - R)^{0.039+0.010 \log R} R^{-0.200}$
C32C	$0.044 N^{0.116} R^{-0.130}$	$0.255 N^{0.071} R^{-0.088}$	$0.063 (N - R)^{0.088+0.018 \log R} R^{-0.363}$	$0.363 (N - R)^{0.045+0.015 \log R} R^{-0.293}$
D20C	$0.025 N^{0.095} R^{-0.096}$	$0.061 N^{0.113} R^{-0.127}$	$0.038 (N - R)^{0.063+0.019 \log R} R^{-0.340}$	$0.107 (N - R)^{0.071+0.027 \log R} R^{-0.485}$
D30C	$0.023 N^{0.098} R^{-0.098}$	$0.063 N^{0.110} R^{-0.130}$	$0.033 (N - R)^{0.069+0.017 \log R} R^{-0.319}$	$0.112 (N - R)^{0.068+0.027 \log R} R^{-0.493}$
D31C	$0.024 N^{0.095} R^{-0.096}$	$0.064 N^{0.110} R^{-0.134}$	$0.035 (N - R)^{0.064+0.018 \log R} R^{-0.328}$	$0.119 (N - R)^{0.064+0.029 \log R} R^{-0.528}$
D32C	$0.042 N^{0.104} R^{-0.119}$	$0.135 N^{0.087} R^{-0.110}$	$0.053 (N - R)^{0.085+0.012 \log R} R^{-0.268}$	$0.162 (N - R)^{0.074+0.008 \log R} R^{-0.220}$
D33C	$0.022 N^{0.100} R^{-0.092}$	$0.067 N^{0.106} R^{-0.130}$	$0.032 (N - R)^{0.071+0.017 \log R} R^{-0.308}$	$0.123 (N - R)^{0.061+0.029 \log R} R^{-0.512}$
D34C	$0.022 N^{0.100} R^{-0.115}$	$0.077 N^{0.098} R^{-0.152}$	$0.030 (N - R)^{0.077+0.015 \log R} R^{-0.303}$	$0.113 (N - R)^{0.069+0.019 \log R} R^{-0.408}$
E20C	$0.008 N^{0.124} R^{-0.140}$	$0.010 N^{0.171} R^{-0.210}$	$0.009 (N - R)^{0.120+0.002 \log R} R^{-0.169}$	$0.012 (N - R)^{0.156+0.011 \log R} R^{-0.359}$
E21C	$0.008 N^{0.124} R^{-0.140}$	$0.010 N^{0.171} R^{-0.210}$	$0.009 (N - R)^{0.120+0.002 \log R} R^{-0.169}$	$0.012 (N - R)^{0.156+0.011 \log R} R^{-0.359}$
E30C	$0.012 N^{0.083} R^{-0.112}$	$0.014 N^{0.145} R^{-0.206}$	$0.012 (N - R)^{0.085+-0.001 \log R} R^{-0.094}$	$0.018 (N - R)^{0.128+0.013 \log R} R^{-0.377}$
E31C	$0.011 N^{0.089} R^{-0.083}$	$0.013 N^{0.149} R^{-0.163}$	$0.012 (N - R)^{0.087+0.001 \log R} R^{-0.101}$	$0.017 (N - R)^{0.131+0.012 \log R} R^{-0.324}$
F20C	$0.069 N^{0.128} R^{-0.164}$	-	$0.099 (N - R)^{0.096+0.022 \log R} R^{-0.409}$	-
F21C	$0.046 N^{0.142} R^{-0.177}$	-	$0.057 (N - R)^{0.122+0.014 \log R} R^{-0.334}$	-
F22C	$0.046 N^{0.142} R^{-0.178}$	-	$0.056 (N - R)^{0.123+0.013 \log R} R^{-0.326}$	-
F30C	$0.000 N^{0.653} R^{-0.046}$	$0.000 N^{0.614} R^{-0.058}$	$0.001 (N - R)^{0.453+0.116 \log R} R^{-1.574}$	$0.002 (N - R)^{0.445+0.096 \log R} R^{-1.321}$
F31C	$0.000 N^{0.733} R^{-0.037}$	$0.000 N^{0.654} R^{-0.052}$	$0.001 (N - R)^{0.509+0.127 \log R} R^{-1.710}$	$0.001 (N - R)^{0.469+0.105 \log R} R^{-1.429}$
G30C	$0.048 N^{0.148} R^{-0.175}$	$0.087 N^{0.133} R^{-0.172}$	$0.068 (N - R)^{0.119+0.020 \log R} R^{-0.427}$	$0.139 (N - R)^{0.097+0.025 \log R} R^{-0.495}$
G31C	$0.047 N^{0.139} R^{-0.160}$	$0.092 N^{0.117} R^{-0.150}$	$0.065 (N - R)^{0.112+0.018 \log R} R^{-0.385}$	$0.134 (N - R)^{0.088+0.019 \log R} R^{-0.394}$
H30C	$0.103 N^{0.102} R^{-0.125}$	$0.138 N^{0.101} R^{-0.127}$	$0.140 (N - R)^{0.077+0.016 \log R} R^{-0.315}$	$0.178 (N - R)^{0.081+0.012 \log R} R^{-0.286}$
J20C	$0.019 N^{0.151} R^{-0.184}$	$0.046 N^{0.137} R^{-0.172}$	$0.024 (N - R)^{0.133+0.013 \log R} R^{-0.357}$	$0.069 (N - R)^{0.107+0.021 \log R} R^{-0.457}$
J30C	$0.015 N^{0.147} R^{-0.172}$	$0.031 N^{0.153} R^{-0.181}$	$0.019 (N - R)^{0.128+0.014 \log R} R^{-0.352}$	$0.048 (N - R)^{0.120+0.024 \log R} R^{-0.499}$
J31C	$0.015 N^{0.151} R^{-0.180}$	$0.032 N^{0.153} R^{-0.184}$	$0.019 (N - R)^{0.133+0.013 \log R} R^{-0.345}$	$0.051 (N - R)^{0.118+0.025 \log R} R^{-0.518}$
J32C	$0.013 N^{0.147} R^{-0.169}$	$0.026 N^{0.159} R^{-0.193}$	$0.017 (N - R)^{0.128+0.013 \log R} R^{-0.343}$	$0.044 (N - R)^{0.120+0.029 \log R} R^{-0.585}$
J33C	$0.013 N^{0.149} R^{-0.174}$	$0.027 N^{0.155} R^{-0.197}$	$0.016 (N - R)^{0.132+0.012 \log R} R^{-0.333}$	$0.045 (N - R)^{0.116+0.029 \log R} R^{-0.585}$
L30C	$0.169 N^{0.091} R^{-0.110}$	-	$0.228 (N - R)^{0.064+0.016 \log R} R^{-0.291}$	-
L31C	$0.140 N^{0.099} R^{-0.122}$	-	$0.192 (N - R)^{0.070+0.018 \log R} R^{-0.318}$	-
M20C	$0.112 N^{0.108} R^{-0.130}$	$0.214 N^{0.083} R^{-0.104}$	$0.156 (N - R)^{0.081+0.018 \log R} R^{-0.343}$	$0.299 (N - R)^{0.057+0.016 \log R} R^{-0.305}$
M21C	$0.105 N^{0.109} R^{-0.131}$	$0.198 N^{0.086} R^{-0.110}$	$0.147 (N - R)^{0.081+0.017 \log R} R^{-0.343}$	$0.280 (N - R)^{0.059+0.016 \log R} R^{-0.319}$
M30C	$0.105 N^{0.109} R^{-0.131}$	$0.199 N^{0.085} R^{-0.110}$	$0.147 (N - R)^{0.081+0.017 \log R} R^{-0.343}$	$0.283 (N - R)^{0.058+0.017 \log R} R^{-0.322}$
P30C	$0.031 N^{0.131} R^{-0.159}$	$0.096 N^{0.112} R^{-0.148}$	$0.042 (N - R)^{0.105+0.017 \log R} R^{-0.370}$	$0.139 (N - R)^{0.083+0.019 \log R} R^{-0.393}$
Q30C	$0.106 N^{0.115} R^{-0.151}$	-	$0.161 (N - R)^{0.078+0.025 \log R} R^{-0.430}$	-
S20C	$0.246 N^{0.072} R^{-0.090}$	-	$0.321 (N - R)^{0.048+0.014 \log R} R^{-0.244}$	-
T30C	$0.176 N^{0.086} R^{-0.103}$	$0.302 N^{0.065} R^{-0.082}$	$0.231 (N - R)^{0.064+0.013 \log R} R^{-0.266}$	$0.389 (N - R)^{0.045+0.011 \log R} R^{-0.227}$

Table 12: **Approximate models of FNIR.** For each algorithm, the table gives the power-law estimates (equations 8 and 9) as estimated using non-linear least squares. The formulae apply to LEO images searched against enrolment sets comprised of the most recent image of $N = 20,000 \dots 1,600,000$ individuals. The power-law functional form is chosen on the basis of the approximately straight lines observed in Figures 9 and 8. **Caution:** These formulae represent empirical models that lack theoretical support. It would be an abuse to evaluate these formulae at values $N > 1.6$ million. In addition, the formulae only approximate the actual measured values.

5.5 Accuracy dependence on subject age

Methods: The VISA dataset includes age information. As shown in the second entry of Table 3, one image of each of the $N = 19972$ individuals is enrolled. Thereafter, one mated search is conducted per person to allow FNIR estimation. Finally, 203,082 nonmated searches are run to support FPIR measurement.

We compute accuracy by age group. We define seven age groups corresponding to stages in life in which facial appearance is similar, and between which there is consensus that appearance is usually different⁹. We attach informal labels to these, as shown in Table 13. The age of the individual at the time the search image was collected is used to determine the age group bin. The enrolled image is acquired some time before that - statistics on the time elapsed between mated pairs appears in Table 13. FNIR is computed over those pairs via the normal definition of equation (3). The elapsed times are too short for the study to quantify longitudinal ageing effects. The results for young subjects are affected by ageing because the elapsed times, of 1-4 years are considerable in an infant.

Group No.	Group Label	Age Range	Search Age Mean	Mated Time Lapse Mean	Mated Count
1	<i>baby</i>	[0, 3)	2.3	1.6	57
2	<i>kid</i>	[3, 8)	5.7	2.8	340
3	<i>pre</i>	[8, 13)	10.7	3.7	533
4	<i>teen</i>	[13, 19)	17.0	2.5	1447
5	<i>young</i>	[19, 30)	25.4	2.0	5930
6	<i>parents</i>	[30, 55)	40.5	2.1	8293
7	<i>older</i>	[55, 101)	63.6	2.2	2709

Table 13: **Age groups:** Labels and sizes of age groups to which search images are assigned. Values in columns 3 to 5 are in years.

Nonmate searches produce top scoring enrolled candidates that vary by algorithm. FPIR is computed using equation (1), applied for each search-image age group. The distribution of the age difference between the enrolment images and whatever non-mates are returned in the search is algorithm-specific, and is not documented in this report.

Results: The results appear in two places: For six, more accurate, algorithms, Figure 11 shows detection error tradeoff characteristics for the seven age groups. Appendix A shows identical graphs for all algorithms.

Discussion: Regarding the figures, the notable points are:

- **Recognition is progressively easier with advancing age:** All algorithms exhibit a strong dependence of FNIR on age. This effect is very large, spanning a factor of ten from infant to senior, and a factor of around five from teen to senior. Miss rates for *older* persons are very low: at a fixed FPIR of 0.005, the most accurate algorithm, E30C, gives FNIR of 0.008 for persons over age 55, 0.027 for *young* 20-somethings, and 0.057 for teenagers. For younger persons, the miss rates climb rapidly to 0.29 for *pre*-teens, 0.4 for *kids*, to 0.7 for *babies*. This progression is common to all algorithms.
- **Young children are more difficult to recognize:** Identification miss rates (FNIR) ascend rapidly for *pre*-teens, *kids* and the youngest individuals. For the *baby* group, 0 to about 3 years old, identification fails more often than it succeeds, i.e. FNIR is above 50%. While the sample size is small (57 subjects), error rates are so high that the result remains significant. This result applies for image pairs collected on average 1.6 years apart (Table 13) and will be in considerable part due to the craniofacial shape change associated with rapid growth. The extent to which smooth “feature-less” skin texture affects FNIR is unknown. Likewise the pose variations inherent in photographing children have not been quantified.
- **Young children are more difficult to discriminate:** All of the algorithms exhibit higher false positive identification rates for younger subjects. The grey lines in Figure 11, which link points of equal threshold, slope upwards to the

⁹We are prevented from using Shakespeare’s seven ages of man from the *All The World’s a Stage* passage in *As You Like It*.

right, indicating simultaneously that younger subjects are less easy to recognize as themselves but also less easy to tell apart. This indicates that younger individuals are more difficult to discriminate from other individuals.

- **False positive identification rate excursions:** For algorithms C31C and J33C in particular, the reduction of FPIR at a fixed threshold continues progressively throughout adulthood. For example, using algorithm C31C configured with threshold 21539, FPIR reduces from 0.2 in infants, to 0.05 in teenagers, to below 0.002 in seniors.

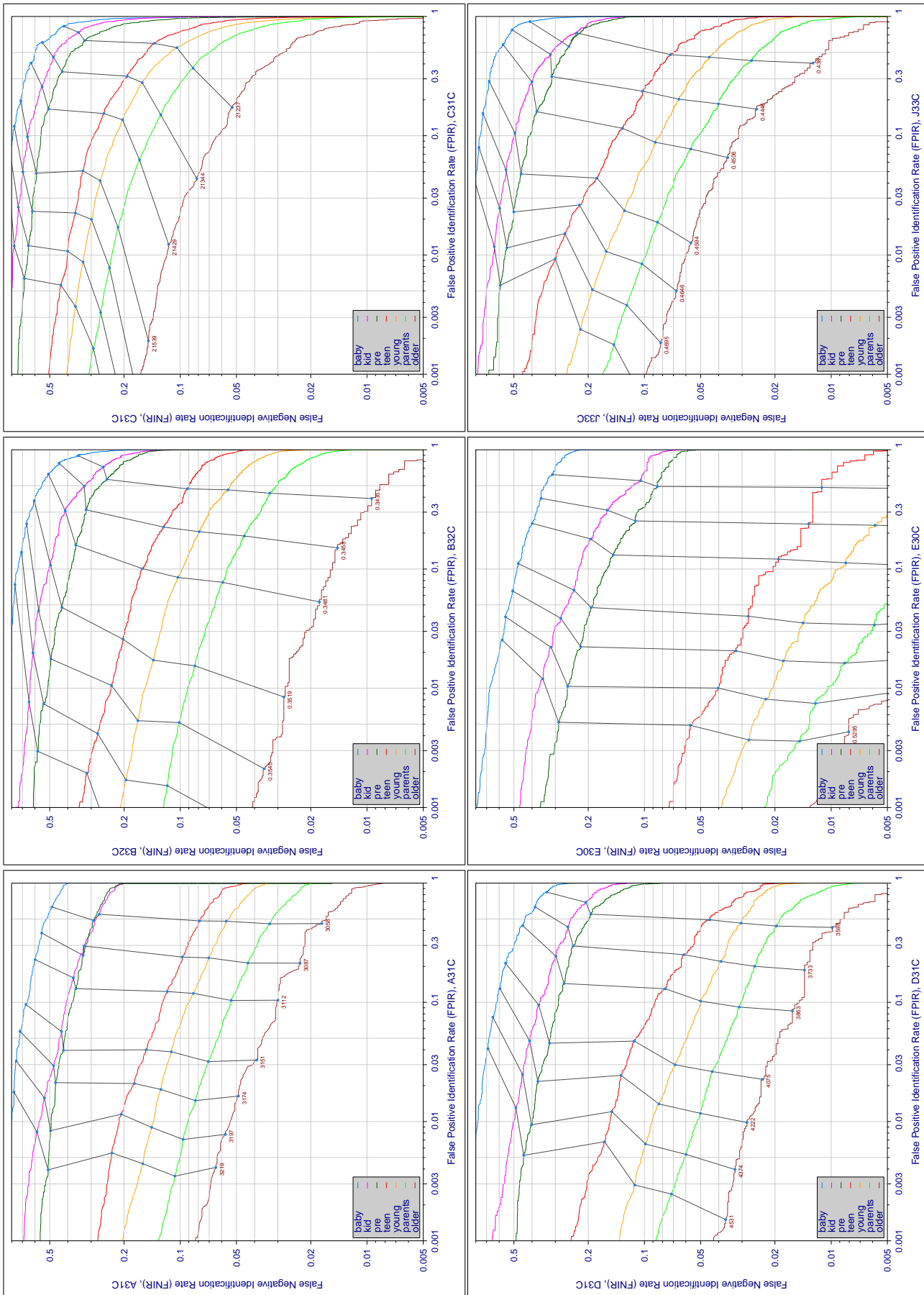


Figure 11: **Older subjects are more easily identified:** For six of the more accurate algorithms operating on VISA images, each panel includes seven traces which partition the population by subject age at the time of collection of the search image. The age bin labels correspond to age ranges of: [0, 3), [3, 8), [8, 13), [13, 19), [19, 30), [30, 55), [55, 101), respectively. Both FNIR and FPIR appear on log axes. The enrolled population size is $N = 19972$. The grey lines link points of equal threshold. Traces toward the bottom indicate lower FNIR - subjects are more easy to recognize as themselves. Traces toward the left indicate lower FPIR - subjects are more easy to discriminate from others.

5.6 Accuracy of sketch recognition

Sketches have long been used in criminal investigations. Historically the most common occurrence is for a forensic artist to interview an eye-witness and, iteratively, produce a likeness of the individual recollected by the witness. The result, a forensic sketch, is essentially a pencil drawing. Nowadays, commercial software is typically used to the same ends, and this results in a composite sketch.

This section addresses whether such sketches can be matched, in one-to-many mode, against photographs resident in a mugshot database.

Methods: We leverage photos from the very commonly used FERET database and a set of sketches of those FERET images published by the City University of Hong Kong (CUHK). A pair is shown in Figure 3. Given photographs of $n = 840$ sketched individuals and $N = 640,000$ mugshots of different individuals (from the LEO dataset, we enrolled $n + N$ photographs. We then searched a) n different frontal photographs from the FERET database; and b) the n CUHK sketches of the enrolled FERET images. Searches produced $L = 50$ candidates. This experimental design is similar to that used in prior algorithm evaluation of sketch identification algorithms [18]. That study enrolled $n = 75$ plus $N = 10,000$ mugshots and searched both forensic and composite sketches with $L = 200$.

Results: Table 14 shows miss rates for the photo and sketch searches. The values are $\text{FNIR}(n + N, R, 0, L)$ for $R = \{1, 10, 50\}$, and $L = 50$.

Discussion: Sketch identification accuracy is far inferior to that documented for mugshots elsewhere in this report, particularly Table 5, for $N = 640K$.

- ▷ **Use of non-sequestered, public-domain, images:** The FERET images and sketches are all in the public domain so, in principle, the results could be manipulated either via training, or outright memorization. This is unlikely, however, since the algorithm providers had no prior reason to suspect that FERET images or sketches would be part of the FRVT evaluation.
- ▷ **Use of algorithms for tasks they were not designed:** As sketch-identification was never declared to be part of the study, the algorithms are being used in a manner not expressly intended by the providers. Such “off label” usage¹⁰ is in fact operational reality - automated algorithms *are* being used to recognize sketches in many police departments.
- ▷ **High miss rates:** The rank-1 identification miss rates are usually above 90%. The most accurate result, for A20C from 3M/Cogent, misses the mate 89.7% of the time. At rank 50, the best result, for that same algorithm, is 73.3% of mates are missed, i.e. only 26.7% are hit. Note the most accurate mugshot algorithms are not the most accurate sketch identification algorithms.
- ▷ **High miss rates may nevertheless be useful:** The poor accuracy figures have to be compared with not having an automated search capability *at all*. In such cases, the miss rate is 100% and investigative leads have to be developed without automated face recognition. As such, the 26.7% hit rate represents a very useful resource in otherwise cold cases. The workload associated with human adjudication of long candidate lists is considered, for mugshots, in section 5.7
- ▷ **An upper bound on accuracy:** The fact that the sketches were prepared by an artist viewing the exemplar photograph probably means that the accuracy measurements here represent a “best case” upper bound on accuracy. That said, the artist did not attempt to produce a photo-realistic image of the person but instead introduced a “shape

¹⁰The term “off label” is used in the medical community to indicate that a drug developed for one condition is prescribed by a doctor for a different, often unrelated, condition.

A = 3M/Cogent G = Hisign P = Zhuhai-Yisheng	B = Cognitec H = CAS-IA Q = JunYu	C = Neurotechnology I = CAS-ICT S = Decatur	D = Safran Morpho J = Toshiba T = Ayonix	E = NEC L = Tsinghua U. II	F = Tsinghua U. M = HP	$\text{FNIR}(N, R, T, L)$ “Miss rate” $\text{FPIR}(N, T, L)$ “False alarm rate”
---	---	---	--	-------------------------------	---------------------------	--

Enrolled N=640000	INVESTIGATION FNIR					
	PHOTOGRAPH			SKETCH		
	RANK=1	RANK=10	RANK=50	RANK=1	RANK=10	RANK=50
A20C	0.014	0.001	0.001	0.897	0.803	0.729
A30C	0.014	0.001	0.001	0.900	0.821	0.750
A31C	0.014	0.001	0.001	0.900	0.821	0.750
A32C	0.016	0.001	0.001	0.959	0.912	0.859
B30C	0.017	0.005	0.001	0.994	0.983	0.964
B31C	0.017	0.005	0.001	0.994	0.983	0.964
B32C	0.017	0.002	0.001	0.987	0.967	0.948
B33C	0.017	0.002	0.001	0.987	0.967	0.948
C20C	0.024	0.007	0.003	0.999	0.993	0.985
C30C	0.020	0.003	0.003	0.948	0.873	0.815
C31C	0.021	0.005	0.002	0.964	0.922	0.890
C32C	0.022	0.003	0.002	0.957	0.899	0.842
D20C	0.014	0.001	0.001	0.950	0.883	0.808
D30C	0.014	0.001	0.001	0.969	0.903	0.823
D31C	0.014	0.001	0.001	0.952	0.879	0.806
D32C	0.017	0.002	0.002	0.981	0.960	0.929
D33C	0.014	0.001	0.001	0.963	0.899	0.831
D34C	0.014	0.001	0.001	0.960	0.899	0.809
E20C	0.014	0.001	0.001	0.903	0.823	0.738
E21C	0.014	0.001	0.001	0.903	0.823	0.738
E30C	0.015	0.001	0.001	0.923	0.855	0.763
E31C	0.014	0.001	0.001	0.920	0.863	0.803
F20C	0.506	0.494	0.494	0.953	0.902	0.845
F30C	0.495	0.485	0.480	0.951	0.894	0.850
F31C	0.563	0.555	0.550	0.957	0.901	0.858
G20C	0.090	0.050	0.033	0.978	0.941	0.901
G30C	0.051	0.016	0.009	0.973	0.935	0.852
G31C	0.079	0.034	0.017	0.959	0.898	0.844
H30C	0.070	0.041	0.027	0.963	0.907	0.833
I20C	0.022	0.008	0.005	0.942	0.892	0.820
J20C	0.017	0.001	0.001	0.936	0.865	0.787
J30C	0.020	0.002	0.001	0.974	0.945	0.909
J31C	0.020	0.002	0.001	0.977	0.948	0.909
J32C	0.016	0.002	0.001	0.957	0.907	0.865
J33C	0.016	0.002	0.001	0.957	0.907	0.865
M30C	0.058	0.037	0.031	0.967	0.929	0.867
P30C	0.019	0.006	0.005	0.957	0.921	0.872
Q20C	0.466	0.087	0.049	1.000	0.993	0.973
S20C	0.169	0.105	0.070	0.994	0.984	0.971
T30C	0.088	0.056	0.051	0.979	0.937	0.888

Table 14: **Sketch searches mostly fail:** FNIR “miss rates” for two kinds of searches made into an enrolled dataset of $N = 640,864$ identities made up of 864 FERET photographs, and 640,000 background LEO mugshots. The first set is comprised of 864 searches of different frontal FERET photographs; the second set is of 864 CUHK sketches of those FERET photographs. The accuracy results for photo searches give very low error rates befitting good quality photographs. The sketch results are very poor - green shading is used to show miss rates below 90%, 85% and 80% at rank 1, 10 and 50 respectively. These thresholds are arbitrary.

exaggeration” to the face. While this does not go as far as a caricature artist would go, the change in appearance may or may not be representative of the situation in contemporary law enforcement processes where software is increasingly capable, but eye-witness recollections remain variable.

- ▷ **Algorithm comparison:** Nevertheless, our use of this image set probably does reveal differences in algorithmic capability. Variation will in part depend on which facial information is represented. Prior work in this area [18] compared algorithms with and without landmark-based representations, the former outperforming the latter. Parties interested in establishing a sketch identification facility should inquire with their providers for algorithms specifically developed for sketch identification.
- ▷ **FERET photo recognition is very good:** Many of the recognition algorithms give very low miss rates on the FERET photographs. This occurs because the FERET fa-fb frontal mated pairs are of good quality, better than that for the LEO images studied earlier in this report.

5.7 Human workload for candidate list adjudication

In a law enforcement scenario, for example, a human reviewer is usually employed to review the candidates returned from an identification search. Typically, the reviewer inspects the search image, and compares each candidate with that image, usually proceeding in the order of descending similarity score, and stopping when he is able to positively confirm a mate. The length of the candidate list may be fixed for all searches, or variable, depending on system configuration. The reviewer sometimes has the option to request a number of candidates, up to a certain limit. Variable length lists arise as a result of applying a threshold.

The following subsections advance models for the workload and costs associated with reviewer-led identification searches.

Assumptions: This workload model assumes the following:

- ▷ The candidates are reviewed serially, not all at once in a large screen GUI, for example.
- ▷ The candidates are searched in decreasing order of similarity score, whether or not the reviewer is presented with the score.
- ▷ Reviewers will stop after confirming a mate.
- ▷ The database is correctly consolidated such that the number of mates is zero or one.
- ▷ Reviewers always find a mate if it is present, and reviewers do not incorrectly associate a search with a nonmate candidate. Particularly, reviewer success is independent of the natural prior occurrence of a mate. Effects of fatigue and boredom have been reported when this quantity is very low [20].
- ▷ The time taken to confirm or exclude a candidate is independent of the rank of the candidate.
- ▷ The time taken to confirm or exclude a candidate is independent of population size. This is potentially incorrect

5.7.1 Fixed length candidate lists, threshold independent workload

For now, assume also that the reviewer is not provided with, or ignores, similarity scores, and thresholds are not applied. Suppose an automated face identification algorithm returns L candidates, and a human reviewer is retained to examine up to R candidates, where $R \leq L$ might be set by policy, preference or labor availability. Given the algorithm typically places

A = 3M/Cogent G = Hisign P = Zhuhai-Yisheng	B = Cognitec H = CAS-IA Q = JunYu	C = Neurotechnology I = CAS-ICT S = Decatur	D = Safran Morpho J = Toshiba T = Ayonix	E = NEC L = Tsinghua U. II	F = Tsinghua U. M = HP	FNIR(N,R,T,L) “Miss rate” FPIR(N,T,L) “False alarm rate”
---	---	---	--	-------------------------------	---------------------------	---

mates at low (good) ranks, the number of candidates a reviewer can be expected to review can be derived as follows. Note that the reviewer will:

- ▷ Always inspect the first ranked image Frac. reviewed = 1
- ▷ Then inspect those candidates where mate not confirmed at rank 1 Frac. reviewed = 1-CMC(1)
- ▷ Then inspect those candidates where mate not confirmed at rank 1 or 2 Frac. reviewed = 1-CMC(2)

etc. Thus if the reviewer will stop at after a maximum of R candidates, the expected number of candidate reviews is

$$M(R) = 1 + (1 - CMC(1)) + (1 - CMC(2)) + \dots + (1 - CMC(R - 1)) \quad (10)$$

$$= R - \sum_{r=1}^{R-1} CMC(r) \quad (11)$$

A recognition algorithm that front-loads the cumulative match characteristic will offer reduced workload for the reviewer. This workload is defined only over the searches for which a mate exists. In the cases where there truly is no mate, the reviewer would review all R candidates. Thus, if the proportion of searches for which a mate does exist is β , which in the law enforcement context would be the recidivism rate [2], the full expression for workload becomes:

$$M(R) = \beta \left(R - \sum_{r=1}^{R-1} CMC(r) \right) + (1 - \beta)R \quad (12)$$

$$= R - \beta \sum_{r=1}^{R-1} CMC(r) \quad (13)$$

Results: Tables 4, 5 and 6 include values for this expression in columns labelled WORK.

Importantly, we restrict the analysis to the case of equation (11) where there is always a mate, i.e. $\beta = 1$. This is done because the goal is to compare algorithms. Note that if $\beta < 1$, reviewers will have to review more candidates than are plotted here. Indeed when $\beta = 0$ all candidates will be reviewed regardless of which algorithm is used.

The tables show that if a reviewer is willing to review, $R = 50$ candidates, then the expected number of candidates actually needing review will often be fewer than 10. For mugshot searches into an enrolled database of 1.6 million identities, enrolled with their most recent image, the NEC algorithms necessitate reviewers adjudicate around 1.6 candidates.

Cost implications: The above expressions for reviewer workload could be multiplied by suitable time and salary factors to estimate cost. Such a cost formulation should be extended to capture the cost of missing a mate altogether - this is a societal cost of failing to find a mate in the first L candidates.

Conclusions: The use of more accurate face recognition algorithms implies decreased workload for human reviewers retained to adjudicate candidate lists. The expected number of candidates before a mate is found is a useful performance metric for identification systems.

5.7.2 Workload with thresholded variable length candidate lists

In this section, the effect of thresholding candidate lists is considered. This has the potential to reduce workload further but at the expense of increased FNIR. If a score threshold is applied to a candidate list, either by the system or by the reviewer, the number of candidates that remain will be a random variable. In the previous section, that number was fixed at R ; here it becomes $R_i(T)$ i.e. the number of candidates on the i -th candidate list that have score greater than or equal to T . The workload associated with adjudication of that list is given by equation 11 as $M(R_i(T))$. An accuracy loss arises, however, because low-scoring mates that are present on the full candidate list will not be available to the reviewer. This is stated by $FNIR(N, R_i(T), T, L)$.

Results: Figure 12 shows this workload reduction, accuracy loss, tradeoff by plotting the two quantities parametrically with threshold, T .

Discussion: The notable observations are:

- ▷ **Best result:** For the D31C algorithm with an enrolled population of $N = 1.6$ million, workload can be reduced by 60% (to a factor of 0.4 times the baseline level) if a 5% increase in miss rates is tolerable.
- ▷ **Algorithm differences:** For other algorithms the cost benefit position is worse. For the A31C, B32C, and C31C implementations the same 60% workload reduction gives around a 20% increase in FNIR. The position with E30C is more complicated: The algorithm is very accurate, and baseline workload is low, so the available benefits are limited.
- ▷ **Effect of population size:** The tradeoff associated with thresholding is most pronounced at large population size. This occurs because the baseline workload is higher at large N (because mates are more likely to be displaced from rank-1 position when N is large). Thus for workload to be reduced to 0.4 times baseline, FNIR values are increased to as high as 1.4 times the baseline when $N = 20,000$.
- ▷ **Magnitude of the effect:** The workload benefits are substantial - a reduction in workload by a factor of 0.5 corresponds to half the labor requirement, at least in applications where high volumes are sustained. The FNIR increases are relatively low, compared to the multipliers associated with a) the use of webcam images and b) the use of inferior recognition algorithms. They are comparable with the use of “recent” vs. “lifetime” enrolment types.

It is clear that candidate list reduction via thresholding can reduce the amount of work a reviewer does, but it will also reduce the number of hits found, and beneficially reduce the number of false alarms.

5.8 Impostor distribution stability

Section 4.1 defines false positive identification rate (FPIR) as the fraction of searches that yield one or more false matches. Likewise, selectivity (SEL) is defined as the expected number of false matches produced in a search. Both of these quantities are a function of threshold, with higher thresholds giving reduced values for FPIR and SEL.

As some systems are configured with thresholds that target a known (often low) FPIR [1], the operational question arises of how stable FPIR is with changes to the properties of images or individuals that are used with the system. This issue is determined by the stability of the nonmate distribution i.e. whether FPIR is independent of image quality, population demographics and ethnicity.

Most of the academic literature addresses improvement of Type 1 error rates such as better hit rates. The primary performance metrics are 1:1 FNMR at fixed FMR, and closed-set CMC. The importance of a stable impostor distribution has

A = 3M/Cogent G = Hisign P = Zhuhai-Yisheng	B = Cognitec H = CAS-IA Q = JunYu	C = Neurotechnology I = CAS-ICT S = Decatur	D = Safran Morpho J = Toshiba T = Ayonix	E = NEC L = Tsinghua U. II	F = Tsinghua U. M = HP	FNIR(N,R,T,L) “Miss rate” FPIR(N,T,L) “False alarm rate”
---	---	---	--	-------------------------------	---------------------------	---

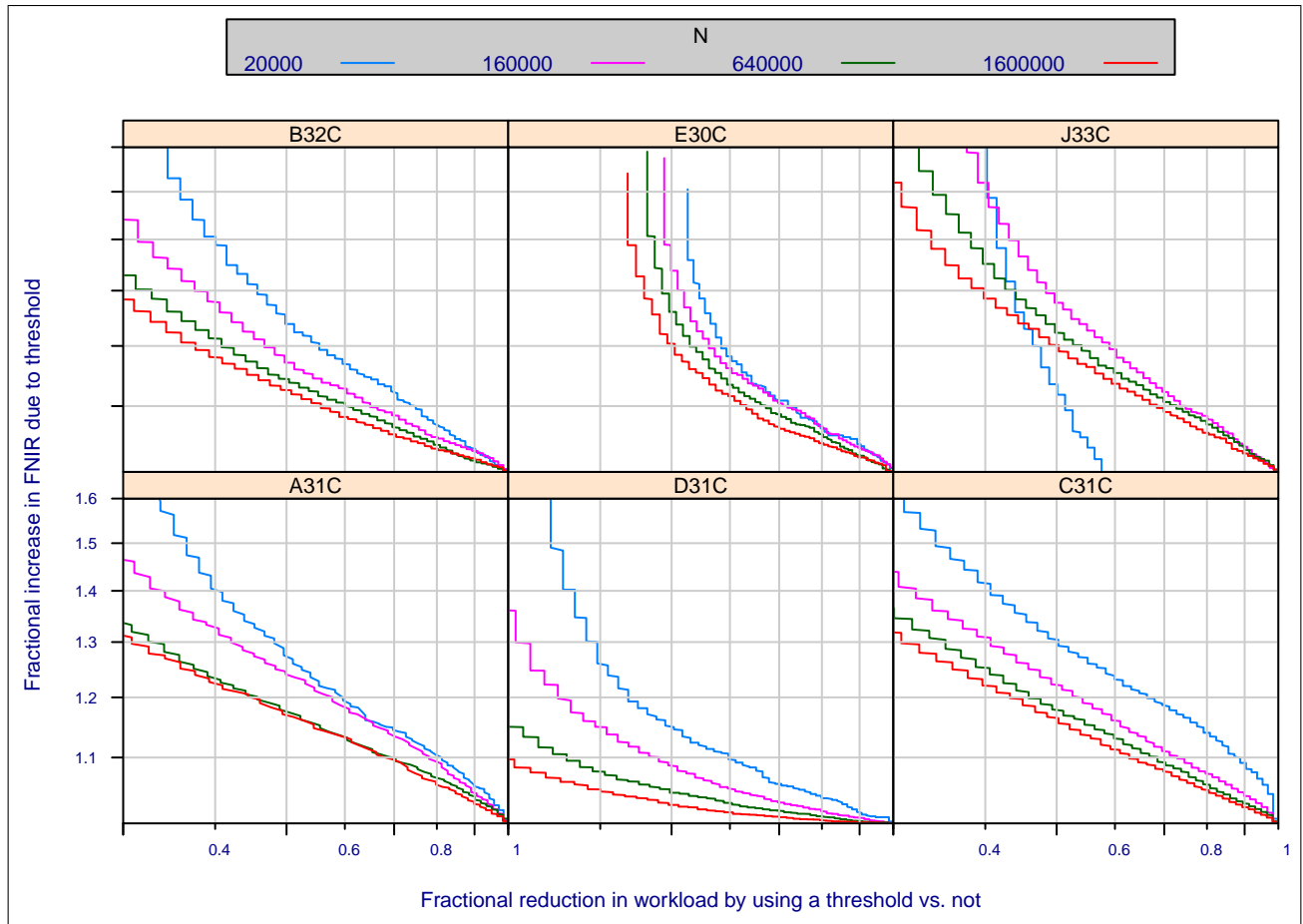


Figure 12: **Reviewer workload reductions with thresholding:** With recent LEO mugshots enrolled, the graphs plot on the x-axis the reduction in the number of candidates (workload eq. (11)) that would require human review if the candidate list were reduced in length by thresholding vs. if the full $L = 50$ candidates were eligible for review. On the y-axis, is the reduction in accuracy i.e. the fractional increase in miss rates incurred by thresholding vs. considering all $L = 50$.

received little attention in the academic literature [15]. However it is critical to systems configured with a fixed threshold that are exposed to different kinds of images.

Methods: Webcam and mugshot images are used in 1:N searches, and scores from both mate and nonmate searches are retained and used in computation of the DET characteristics plotting FNIR vs. FPIR (see equations (3) and 1).

Section 5.5 details the effect on both FNIR(T) and FPIR(T) when individuals in different age groups are identified.

Results: Figure 6 shows DETs for six more accurate algorithms. Appendix A includes DETs for each algorithm for mugshots and webcam images. Similarly, Figure 11 gives DETs for identification by age group. In both cases, the plots include lines connecting points of fixed threshold.

Discussion: The horizontal displacement of the grey lines is an indicator that the impostor distribution is not stable under condition-change. The 3M/Cogent (A31C) and NEC (E30C) algorithms exhibit less than a factor of two change in FPIR when processing webcams vs. mugshots (Fig. 6), and when processing young persons vs. old (Fig. 11). The Neurotechnology (C31C) algorithm, on the other hand, gives more than a 100-fold increase in FPIR when processing infants vs. seniors. The Cognitec algorithm gives a roughly 20-fold increase in FPIR when processing webcam images. These results will manifest themselves operationally as longer candidate lists

5.9 Computational expense

5.9.1 Search duration times

Background: In most deployments, the enrolled population increases over time. This may be a continuous process or the result of merging separate datasets. If the database doubles in size, so does the search time. This has major implications for capital expenditure on computing hardware and ancillary equipment.

In our 2010 evaluation of face recognition algorithms [12], we showed that search speed scales sublinearly with enrolled population size. That aside, there is little documentation of search speed in operational biometric systems. There is a large and mature literature on fast search algorithms, although much of this is outside of the biometric arena. The term fast refers to algorithms for which average search time increases better-than-linearly with population size N , for example as $\log N$.

Methods: See section 4.5. For identification trials, the algorithm was permitted to use the available hardware as it saw fit. It could elect to start any number of threads [1,16] and this could be varied dynamically and as a function of N . Only one provider, F, elected to use threading.

For timing measurement, both mate and nonmate searches were conducted in random order. However, the duration measurements given here are computed over only nonmate searches. Mate searches are not used because some algorithms implement so-called 1:FIRST semantics where the search is terminated once any sufficiently high scoring hit has been found. This may be operationally valuable, but it confuses measurement of underlying algorithmic efficiency.

The candidate list length was fixed at $L = 50$. Production of longer candidate lists may take longer - this is untested in FRVT. The estimates reported below are median values estimated over 2000 searches for which a mate exists, and 2000 searches for which a mate does not exist.

Results: Timing results are tabulated in Table 15. Graphs of search durations vs. N appear in the algorithm summaries of Appendix A.

There is wide variation in search speed. For $N = 1,600,000$, the most accurate algorithm (E30C) takes 1.093 seconds to

A = 3M/Cogent G = Hisign P = Zhuhai-Yisheng	B = Cognitec H = CAS-IA Q = JunYu	C = Neurotechnology I = CAS-ICT S = Decatur	D = Safran Morpho J = Toshiba T = Ayonix	E = NEC L = Tsinghua U. II	F = Tsinghua U. M = HP	FNIR(N,R,T,L) "Miss rate" FPIR(N,T,L) "False alarm rate"
---	---	---	--	-------------------------------	---------------------------	---

ALG	ENROLLMENT		FINALIZATION			SEARCH						
	E _{TIME}	E _{SIZE}	F _{EXP}	F _{TIME}	F _{POW}	S _{SIZE}	S _{TIME}	T _s (20)k	T _s (160)k	T _s (640)k	T _s (1600)k	T _s = aN ^b
A20C	520	4106	1.0	1.8	1.7	4106	519	155 ± 1	1235 ± 0	4953 ± 1	12568 ± 97	0.0080N ^{1.00}
A30C	327	4622	1.0	8.0	1.0	4622	324	29 ± 0	242 ± 0	979 ± 1	2454 ± 0	0.0018N ^{0.99}
A31C	324	4622	1.0	12.2	1.6	4622	324	28 ± 0	243 ± 0	982 ± 2	2452 ± 0	0.0017N ^{0.99}
A32C	296	1046	1.0	0.6	1.6	1046	189	11 ± 0	102 ± 0	447 ± 3	1080 ± 0	0.0017N ^{0.94}
B30C	282	4796	1.0	3.6	1.0	4796	284	96 ± 0	756 ± 0	2972 ± 0	7494 ± 18	0.0059N ^{0.98}
B31C	287	4796	1.0	3.8	0.4	4796	282	95 ± 0	756 ± 0	2982 ± 0	7498 ± 2	0.0053N ^{0.99}
B32C	382	9932	1.0	25.3	1.1	9932	388	255 ± 0	2018 ± 0	8076 ± 1	20239 ± 33	0.0117N ^{1.01}
B33C	385	9932	1.0	235.5	0.9	9932	382	255 ± 0	2019 ± 0	8064 ± 1	20285 ± 4	0.0119N ^{1.00}
C20C	263	35994	0.0	0.0	0.9	35994	222	164 ± 3	1143 ± 15	4518 ± 65	11149 ± 170	0.0079N ^{0.99}
C30C	363	37008	0.0	0.0	1.0	37008	362	991 ± 2	7157 ± 20	28652 ± 79	72246 ± 195	0.0382N ^{1.01}
C31C	361	37008	0.0	0.0	1.0	37008	361	179 ± 2	1279 ± 17	4986 ± 65	12575 ± 154	0.0056N ^{1.02}
C32C	236	5040	0.0	0.0	0.9	37008	360	135 ± 0	1014 ± 1	4032 ± 6	10136 ± 14	0.0067N ^{0.99}
D20C	714	8005	1.0	45.5	1.0	20489	716	443 ± 0	510 ± 0	-	-	228N ^{0.07}
D30C	754	8562	1.4	76.8	1.0	21046	709	722 ± 0	804 ± 0	958 ± 0	1263 ± 1	193N ^{0.13}
D31C	729	12247	1.2	88.1	1.0	24731	712	1218 ± 0	1278 ± 0	1422 ± 0	1761 ± 0	499N ^{0.08}
D32C	493	857	5.2	5.0	1.0	857	471	35 ± 0	87 ± 1	260 ± 0	605 ± 14	0.0026N ^{0.86}
D33C	733	8005	1.5	57.3	1.0	20489	712	736 ± 0	812 ± 0	954 ± 0	1271 ± 0	174N ^{0.13}
D34C	704	8005	1.3	46.7	1.0	20489	707	731 ± 0	813 ± 0	957 ± 0	1236 ± 0	182N ^{0.13}
E20C	216	2465	1.0	2.2	1.3	2465	204	15 ± 0	123 ± 0	513 ± 1	1143 ± 0	0.0058N ^{0.85}
E21C	196	2465	1.0	2.4	1.2	2465	204	17 ± 0	137 ± 0	481 ± 0	1245 ± 2	0.0004N ^{1.04}
E30C	229	2529	1.0	2.5	1.3	2529	208	22 ± 0	118 ± 0	450 ± 0	1093 ± 0	0.0021N ^{0.92}
E31C	205	2529	1.0	2.3	1.3	2529	210	13 ± 0	53 ± 0	168 ± 0	279 ± 1	0.0199N ^{0.67}
F20C	170	6760	1.0	3.2	-	6760	118	-	359 ± 9	-	-	0.0120N ^{0.87}
F30C	272	6440	1.0	3.0	1.4	6440	259	196 ± 0	1529 ± 0	3217 ± 5	-	1N ^{0.60}
F31C	262	7484	1.0	76.9	1.6	7484	263	242 ± 0	1901 ± 1	3510 ± 1	-	2N ^{0.54}
G30C	186	3484	1.0	1.3	1.7	3484	110	118 ± 2	843 ± 0	3376 ± 9	-	0.0043N ^{1.02}
G31C	145	2240	1.0	0.8	1.9	2240	192	14 ± 0	111 ± 0	418 ± 12	-	0.0003N ^{1.06}
H30C	81	4420	1.0	5.9	1.9	4420	65	55 ± 0	231 ± 2	901 ± 8	-	0.0051N ^{0.91}
J20C	466	6206	1.0	444.5	1.7	6206	485	14 ± 2	117 ± 0	568 ± 52	1137 ± 1	0.0071N ^{0.84}
J30C	532	4158	1.1	1167.7	0.9	4158	528	13 ± 0	134 ± 0	552 ± 0	1110 ± 1	0.0018N ^{0.94}
J31C	526	4158	1.1	4191.1	0.8	4158	527	13 ± 0	130 ± 0	454 ± 47	1194 ± 2	0.0012N ^{0.97}
J32C	490	8254	1.0	70.5	1.0	8254	546	23 ± 0	244 ± 1	829 ± 2	2449 ± 0	0.0005N ^{1.07}
J33C	550	8254	1.1	8403.1	0.6	8254	546	30 ± 0	191 ± 0	798 ± 2	2086 ± 38	-
L30C	543	7328	1.0	151.7	1.0	7328	537	507 ± 0	4024 ± 1	-	-	0.0242N ^{1.00}
L31C	626	7328	1.0	9.2	0.8	7328	629	497 ± 0	3998 ± 1	-	-	0.0252N ^{1.00}
M20C	164	5608	1.0	0.1	0.0	5608	-	-	-	-	-	-
M21C	164	5608	1.0	0.1	0.0	5608	-	-	-	-	-	-
M30C	136	5608	1.0	0.1	0.9	5608	112	25 ± 0	179 ± 1	688 ± 0	-	0.0007N ^{1.03}
P30C	565	3713	1.0	11.0	1.0	3713	564	364 ± 0	2938 ± 1	11650 ± 8	-	0.0195N ^{0.99}
Q30C	170	5600	1.0	4.5	-	5600	170	78 ± 1	1417 ± 5	-	-	-
S20C	411	800	1.0	0.3	-	800	172	55 ± 0	443 ± 0	-	-	0.0027N ^{1.00}
T30C	283	1936	0.0	0.0	1.0	1936	189	12 ± 0	94 ± 0	348 ± 0	-	0.0022N ^{0.89}

Table 15: **Resource consumption:** From left to right: the duration of enrolment template generation; enrolment template size in bytes; the ratio of the size of post-finalized enrolment data on disk to N times E_{SIZE} ; the duration of the finalization call in seconds for $N = 1.6$ million; the dependence of finalization time on N i.e. b in N^b ; search template size in bytes; and its generation time; the time taken to execute searches of the given population, N ; the power-law model of search duration; All times are medians, unless stated otherwise, and apply to the execution of API functions on a single core of a c. 2011 server-class PC-architecture processor. One exception to this is that the 1:N search for F algorithms used up to 16 cores - those times have not been adjusted and are not therefore comparable.

execute a search. The slowest algorithm by far, C30C, takes 72.2 seconds.

As shown in Appendix A, most search durations increase linearly on a log-log plot.

$$\log T = a \log N + b, \quad (14)$$

This observation corresponds to a power-law form

$$T = cN^a \quad (15)$$

where the constant $b = \log c$ determines the intercept on the observed plot, and the constant a is the slope. The parameters were estimated using nonlinear least squares regression without log transformation. Uncertainty estimates are from bootstrapping.

Discussion: By referencing the a exponents in Table 15, it is evident that with three notable exceptions, all providers' algorithms exhibit a linear dependence of search duration on the enrolled population size. The exceptions are Morpho, where the exponent a is typically around 0.1. This implies that a 10 fold increase in N gives a 1.25 fold increase in search duration. While this is an attractive proposition for very large deployments (several United States' driving license databases have sizes above 10^7), the practical relevance of this will depend on the transaction volume demands and the number of available computers. The second exception is NEC, where the exponent a varies from 0.92(E30C) down to 0.67(E31C) for which a ten fold increase in N leads to a 4.6 fold increase in duration. Finally the Tsinghua (F) algorithms exhibit sublinear speed.

Note that if K processing cores are available on a computer, K searches cannot typically be conducted simultaneously without some loss of speed due to memory bus bandwidth constraints.

Conclusions: Search durations scale approximately as a power of the database size. The exponents are dependent on the algorithm. There is approximately an order of magnitude difference in the search durations measured for the four most accurate algorithm providers. The most accurate algorithms are among the fastest.

5.9.2 Template creation times

Background: How long does it take to extract features from an image and make a template? Does this depend on the width and height of the input image? Does it depend on whether the template is used for enrolment, verification, or identification?

Drivers: Template generation time is often a large component of a 1:N identification search, obviously depending on N . If multiple images are to be searched, e.g. frames from a video sequence, then template generation time can be important, and can dominate an overall transaction time. Additionally, template generation time will be important if an existing image corpus is going to be re-enrolled by a new provider. For example, re-enrolment of an 18M person driving license database takes $1 \times 18 \times 106 / 64 / 3600 = 156$ hours if a one second template generation were sustained on a 32 core blade installation. This does not include de-duplication searches.

Methods: See section 4.5.

Results: Timing results are tabulated in Table 15. Template generation durations are graphed alongside search duration in the algorithm summaries of Appendix A.

Discussion: Template generation times start below 100 milliseconds (H30C, Q30C) and range from 200-750 milliseconds for the most accurate algorithms. There is some industry-wide tradeoff of template generation speed with accuracy: Several of the more inaccurate algorithms are efficient in their production of templates (algorithms from Q, G, M, durations

A = 3M/Cogent G = Hisign P = Zhuhai-Yisheng	B = Cognitec H = CAS-IA Q = JunYu	C = Neurotechnology I = CAS-ICT S = Decatur	D = Safran Morpho J = Toshiba T = Ayonix	E = NEC L = Tsinghua U. II	F = Tsinghua U. M = HP	FNIR(N,R,T,L) "Miss rate" FPIR(N,T,L) "False alarm rate"
---	---	---	--	-------------------------------	---------------------------	---

below 200 msec) whereas the providers of the second and third most accurate algorithms prepare templates more slowly, J in approximately 500 msec and D in around 700 milliseconds. This implication is contradicted however by the production of provider E of templates in less than 250 milliseconds.

Conclusions: Template creation times are independent of the target population size, suggesting that developers did not tailor their algorithmic representation to the size of the identification search.

5.10 Template size

Each implementation encodes information derived from the face image in a proprietary representation of the feature data. This information is generally a trade secret. It encodes one or more mathematical representations of the face shape, structure, or texture but could also, in principle, encode anything else (e.g. non-traditional information such as hair color, style, eye color).

Demand driver: Templates contain the mathematical representation of one or more images of a person. Biometric templates are proprietary, non-standard, and their content is protected as a trade-secret.

The size of the feature data is an important system-design parameter in most biometric applications. Template size is clearly influential on storage requirements, both on-disk and in-memory, on network transmission bandwidth requirements, and on machine throughput. In addition, a large template may be associated with computational complexity and computational expense of the matching algorithm.

Methods: The FRVT Evaluation Plan and API [14] explicitly supported measurement and reporting of facial recognition template size. Two direct measurements of template size are made: one for enrolment templates, and one for search templates. The API supports passage of $K \geq 1$ images to the template generation function under test. KB bytes is pre-allocated, where maximum template size, B , was returned by an initialization function. For any given input, the exact template size was returned and used to save the template to disk.

Results: Table 15 shows template sizes, in bytes. The values also appear alongside the accuracy results in Tables 4 - 6.

Discussion:

- ▷ **Between-provider size variation:** Across all providers, template sizes vary from 0.8 to 37 kilobytes, with the more accurate algorithms having sizes between 2 and 10KB. Some providers submitted algorithms with notably small “lightweight” templates.
- ▷ **Within-provider size variation:** Some providers submitted algorithms with varying template sizes. Larger templates are assumed to contain richer, more discriminative, features that should afford better accuracy.

3M/Cogent’s most and least accurate algorithms used templates of size 4622 and 1046 bytes respectively. This more than four-fold size difference gave rank one miss rates of 0.133 and 0.173 respectively (Table 4, RECENT enrolment). This loss is less pronounced at low FPIR and for LIFETIME enrolment.

Cognitec submitted algorithms with sizes of 9932 versus 4796 bytes. This approximately two-fold variation gave rise to rank one miss rates of 0.109 vs. 0.120 respectively (Table 4, RECENT enrolment).

Morpho pushed this tradeoff still further by submitting one algorithm, D32C, with a template size of 857 bytes. The best algorithm, D31C, used enrolment templates of size 12247 bytes and search templates of about twice that size, 24731 bytes. This algorithm offers very small accuracy improvements over their second most accurate algorithm used enrolment templates of size 8005 bytes and search templates of size 20489 bytes. Thus comparing the smallest

A = 3M/Cogent G = Hisign P = Zhuhai-Yisheng	B = Cognitec H = CAS-IA Q = JunYu	C = Neurotechnology I = CAS-ICT S = Decatur	D = Safran Morpho J = Toshiba T = Ayonix	E = NEC L = Tsinghua U. II	F = Tsinghua U. M = HP	FNIR(N,R,T,L) “Miss rate” FPIR(N,T,L) “False alarm rate”
---	---	---	--	-------------------------------	---------------------------	---

template (D32C) with a second tier algorithm (D34C), the error rate approximately halves from 0.150 to 0.076 (Table 4, RECENT enrolment).

Toshiba explored this variation also submitting templates of size 4158 and 8254 bytes. These afforded error rates of 0.079 (J33C) and 0.091 (J30C).

In conclusion, the ability to reduce template size may be useful operationally, especially for example in speed and bandwidth sensitive applications. Error rate increases are often modest, but increase substantially with very small templates. Within-provider accuracy variations are usually smaller than between-provider - this is evident in the algorithm rankings of Tables 4 - 6.

- ▷ **No dependence on N:** In all cases, the size of a single enrolment template is independent of the size of the enrolled population. This shows that developers did not exploit the provision, via the API initialization call, of the integer number of subjects about to be enrolled. This information would have allowed the implementation to use larger, richer templates for larger N.
- ▷ **Linear in K:** Template sizes vary with the number of images input to the template generation function - see section 3.2. The API sends $K \geq 1$ images to the template generation function. For all algorithms, the size of the enrolment template grows linearly with the number of images that went into its creation. This indicates that algorithms are not integrating facial information across images.
- ▷ **Some asymmetric templates:** The API supported asymmetric or role-specific templates. This allows a template to be used only for enrolment, or only for search, but not vice-versa. Many template sizes are independent of role. The exceptions are algorithm C32C and all the D algorithms except D32C. Operationally, a verification template is usually not stored permanently as it exists only for the duration of a recognition transaction.
- ▷ **Within-memory representation:** Further when N images are enrolled, each producing a template of size, x , the notional enrolment database size will be Nx . This collection of templates is sent to a one-time *finalization* function provided by the algorithm. This prepares the data for subsequent searches. Specifically, it takes the Nx bytes of input and writes F bytes to disk. This size is recorded and used to compute the expansion factor F/Nx . Some algorithms simply copy the input data, whence the expansion factor is 1. Other algorithms re-arrange the input data for efficient in-memory search.

5.10.1 Finalization times

Background: Finalization is a processing step that is applied over a set of N templates from N individuals. It is executed once, before any searches are executed. It is included in the FRVT execution pipeline to allow algorithms to derive valuable information from the entirety of the data, i.e. information that cannot be derived during the image-to-template feature extraction operation.

Results: Finalization times are tabulated in Table 15.

Discussion: Most implementations execute finalization very quickly indicating only trivial data copying or re-arrangement. However, some algorithms execute more slowly, particularly those of Toshiba, indicating that the finalization step is a non-trivial mathematical operation. It is not known whether such processing could be executed in an operational context, where it is necessary to add and delete entries from the enrolment database on an ongoing basis. The assumption is that finalization is an operationally realistic process that could either be executed periodically during the lifetime of a biometric system, or on-demand after a batch enrolment.

A = 3M/Cogent	B = Cognitec	C = Neurotechnology	D = Safran Morpho	E = NEC	F = Tsinghua U.	FNIR(N,R,T,L) "Miss rate"
G = Hisign	H = CAS-IA	I = CAS-ICT	J = Toshiba	L = Tsinghua U. II	M = HP	FPIR(N,T,L) "False alarm rate"
P = Zhuhai-Yisheng	Q = JunYu	S = Decatur	T = Ayonix			

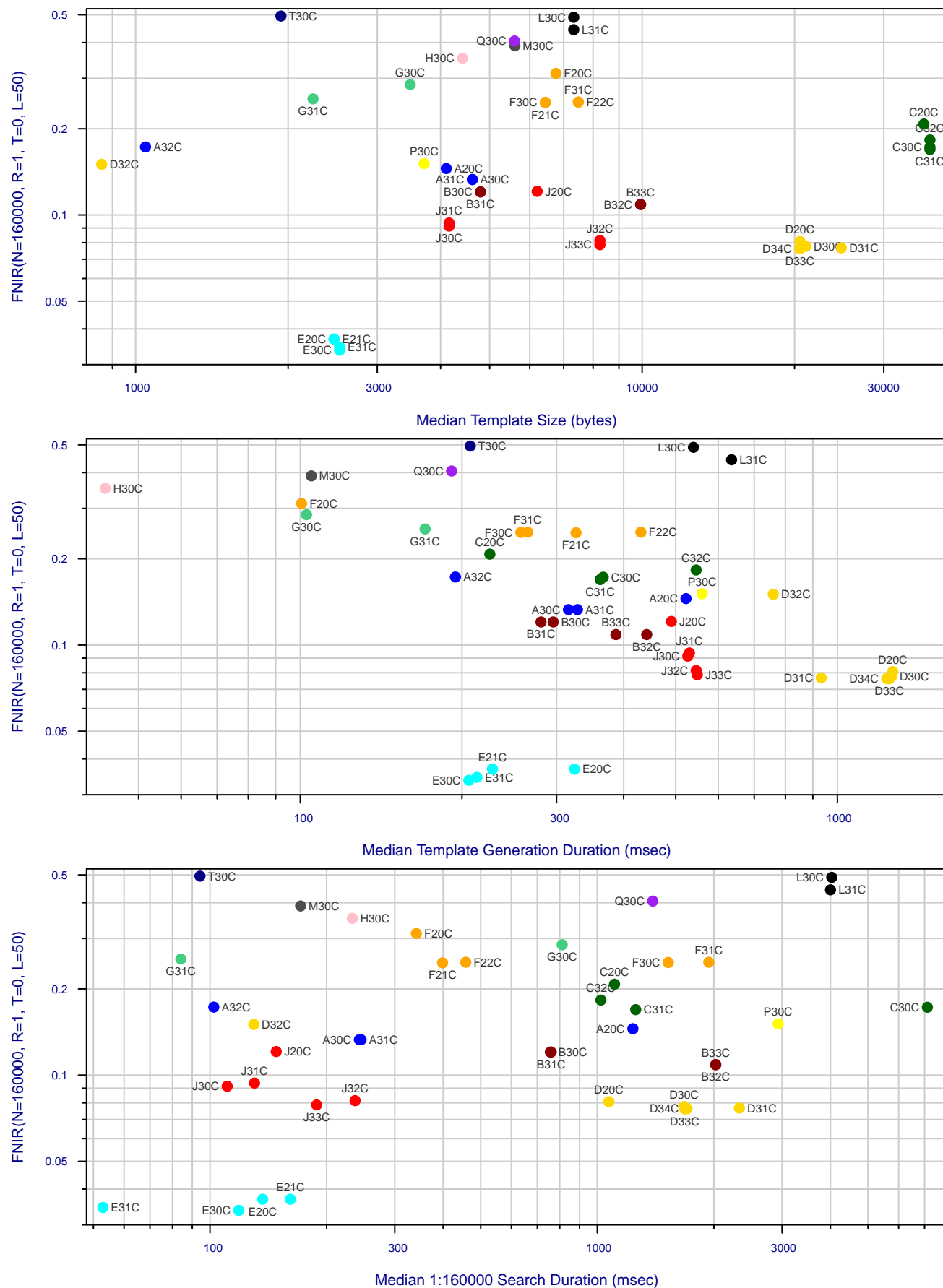


Figure 13: **Performance tradespaces:** The plots show rank one miss rates against three measures of resource consumption: feature size, feature extraction time, and nonmated search duration.

A = 3M/Cogent	B = Cognitec	C = Neurotechnology	D = Safran Morpho	E = NEC	F = Tsinghua U.	FNIR(N,R,T,L) "Miss rate"
G = Hisign	H = CAS-IA	I = CAS-ICT	J = Toshiba	L = Tsinghua U. II	M = HP	FPIR(N,T,L) "False alarm rate"
P = Zhuhai-Yisheng	Q = JunYu	S = Decatur	T = Ayonix			

5.11 Exploiting multiple cores

The FRVT Evaluation Plan and API document [14] did not support execution of a search across $B > 1$ blades because there was no need. The reasons are as follows:

- ▷ In all cases, for all population sizes, the entire enrolment database is small enough to fit in main memory.
- ▷ A blade equipped with $C > 1$ cores was fully utilized by running C searches simultaneously as separate processes. This facility was not used for threaded implementations. (Timing measurements were made with $C = 1$ process).
- ▷ When searching an enrolment database of size E on a blade with memory M , the number of copies of the enrolment data that can be made and kept in memory is $c = \lfloor M/E \rfloor$. This supports execution of $\min(c, C)$ completely independent processes, each running separate searches.
- ▷ However, we can avoid this memory limit by making only $c = 1$ copies of the enrolment database by using the LINUX *fork()* system call C times. While this spawns C entirely separate processes, the LINUX implementation of *fork()* uses *copy-on-write* semantics, which means that the enrolment data is not copied because, as a read-only element, it does not change.

References

- [1] Face recognition as a search tool foto-fahndung. Technical report, Bundeskriminalamt (BKA), Thaerstrasse 11, 65193, Wiesbaden, Germany, February 2007.
- [2] Blumstein, Cohen, Roth, and Visser, editors. *Random parameter stochastic models of criminal careers*. National Academy of Sciences Press, 1986.
- [3] Thomas P. Bonczar and Lauren E. Glaze. Probation and parole in the united statesm 2007, statistical tables. Technical report, Bureau of Justice Statistics, December 2008.
- [4] M. Brauckmann and C. Busch. Large scale database search. In Anil K. Jain and Stan Z. Li, editors, *Handbook of Face Recognition*, pages 639–654. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2011.
- [5] S. Curry, D. Founds, J. Marques, N. Orlans (Mitre), and C. Watson (NIST). Meds - multiple encounter deceased subject face database - nist special database 32. NIST Interagency Report 7679, National Institute of Standards and Technology, 2011. <http://www.nist.gov/itl/iad/ig/sd32.cfm>.
- [6] Working Group 3. Ed. D. D’Amato. *ISO/IEC 19794-5 Amendment 1 - Biometric data interchange formats - Part 5: Face image data - Conditions for taking photographs for face image data*. JTC1 :: SC37, 1 edition, 12 2007. <http://webstore.ansi.org>.
- [7] G. Doddington, W. Liggett, A. Martin, M. Przybicki, and D. Reynolds. Sheep, goats, lambs and wolves: A statistical analysis of speaker performance in the nist 1998 speaker recognition evaluation. In *Proceedings of 5th International Conference of Spoken Language Processing*, ICSLP 98, Sydney, Australia, 1998. Paper 608 on CD-ROM.
- [8] J. Egan. *Signal Detection Theory and Analysis*, pages 46–48. Academic Press, 1975.
- [9] Ferrara et al. Face image iso compliance verification. Technical report, Università di Bologna, 2 2014. <https://biolab.csr.unibo.it/fvcongoing/UI/Form/BenchmarkAreas/BenchmarkAreaFICV.aspx>.
- [10] Matteo Ferrara, Annalisa Franco, Dario Maio, and Davide Maltoni. Face image conformance to iso/icao standards in machine readable travel documents. *IEEE Transactions on Information Forensics and Security*, 7(4):1204–1213, 2012.
- [11] Working Group 3. Ed. P. Griffin. *ISO/IEC 19794-5 Information Technology - Biometric Data Interchange Formats - Part 5: Face image data*. JTC1 :: SC37, 1 edition, 2005. <http://webstore.ansi.org>.
- [12] P. Grother, G. W. Quinn, and P. J. Phillips. Evaluation of 2d still-image face recognition algorithms. NIST Interagency Report 7709, National Institute of Standards and Technology, 8 2010. <http://face.nist.gov/mbe> as MBE2010 FRVT2010.
- [13] P. J. Grother and P. J. Phillips. Models of large population recognition performance. In *IEEE Proceedings of the Computer Vision and Pattern Recognition (CVPR)*, June 2004. Main conference.
- [14] Patrick Grother, George W. Quinn, and Mei Ngan. Face recognition vendor test - still face image and video concept, evaluation plan and api. Technical report, National Institute of Standards and Technology, 7 2013. http://biometrics.nist.gov/cs.links/face/frvt/frvt2012/NIST.FRVT2012_api_Aug15.pdf.
- [15] J.P. Hube. Using biometric verification to estimate identification performance. In *Biometric Consortium Conference, 2006 Biometrics Symposium: Special Session on Research at the*, pages 1–6, 9 2006.
- [16] Herve Jarosz and Jean-Christophe Fondeur. Large scale identification system design. In Wayman, editor, *Biometric Systems*, chapter 9. Springer, 2005.
- [17] J. Klontz and A. Jain. A case study of automated face recognition: The boston marathon bombings suspects. *IEEE Computer Magazine*, 11 2013. <http://openbiometrics.org/publications/KlontzJain.IEEECOMPUTERNov2013.pdf>.
- [18] Scott Klum, Hu Han, Anil K. Jain, and Brendan Klare. Sketch based face recognition: Forensic vs. composite sketches. In *Proc. 6th IAPR International Conference on Biometrics (ICB)*, June 2013. Madrid, Spain.
- [19] Frank Main. Chicago police go high-tech with facial recognition software. *Chicago Sun-Times*, 7 2013. www.suntimes.com/21268770-761/chicago-police-go-high-tech-with-facial-recognition-software.html.

- [20] Expert Working Group on Human Factors in Latent Print Analysis. Latent print examination and human factors: Improving the practice through a systems approach. Technical report, U.S. Department of Commerce, National Institute of Standards and Technology, 2 2012.
- [21] G. W. Quinn and P. Grother. Performance of face recognition algorithms on compressed images. NIST Interagency Report 7830, National Institute of Standards and Technology, 12 2011. <http://face.nist.gov/mbe as MBE2010> FRVT2010.
- [22] Deepayan Sarkar. *Lattice: Multivariate Data Visualization with R*. Springer, New York, 2008. ISBN 978-0-387-75968-5.
- [23] Jamie Sherrah. False alarm rate: a critical performance measure for face recognition. In *FGR*, pages 189–194. IEEE Computer Society, 2004.
- [24] United States. *Passport photo guidelines*. Department of State, 2011. <http://travel.state.gov/content/passports/english/passports/photos/photos/examples.html>.
- [25] X. Wang and X. Tang. Face photo-sketch synthesis and recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 31(11):1955–1967, 2009.
- [26] Working Group 3. Ed. M. Werner. *ISO/IEC 19794-5 Information Technology - Biometric Data Interchange Formats - Part 5: Face image data*. JTC1 :: SC37, 2 edition, 2011. <http://webstore.ansi.org>.
- [27] Bradford Wing and R. Michael McCabe. Nist special publication 500-271: American national standard for information systems data format for the interchange of fingerprint, facial, and other biometric information part 1. Technical report, September 2011. ANSI/NIST ITL 1-2011.
- [28] W. Zhang, X. Wang, and X. Tang. Coupled information-theoretic encoding for face photo-sketch recognition. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011.

A Biometric Error Rate Tradeoff Characteristics

This Appendix is intended to give a biometric identification-specific overview of the Detection Error Tradeoff characteristic DET. More general and detailed information is given in the Egan's class book [8].

Accuracy Terms + Definitions

A **detection error tradeoff** (DET) characteristic represents the tradeoff between Type II and Type I classification errors. A **receiver operating characteristic** (ROC) is usually equivalent and the terms are synonymous. In biometrics, Type II errors occur when two samples of one person do not match – this is called a **false negative**. Correspondingly, Type I errors occur when samples from two persons do match – this is called a **false positive**. Matches are declared by a biometric system when the native comparison score from the recognition algorithm meets some **threshold**. Comparison scores can be either **similarity scores**, in which case higher values indicate that the samples are more likely to come from the same person, or **dissimilarity scores**, in which case higher values indicate different people. Similarity scores are traditionally computed by **fingerprint** and **face** recognition algorithms, while dissimilarities are used in **iris recognition**. In some cases, the dissimilarity score is a distance; this applies only when **metric** properties are obeyed. In any case, scores can be either **mate** scores, coming from a comparison of one person's samples, or **nonmate** scores, coming from comparison of different persons' samples. The words **genuine** or **authentic** are synonyms for mate, and the word **impostor** is used a synonym for nonmate. The words mate and nonmate are traditionally used in identification applications (such as law enforcement search, or background checks) while genuine and impostor are used in verification applications (such as access control).

For iris recognition, mate comparisons yielding dissimilarities greater than a threshold are false negatives. In identification these are called **misses** and contribute to the **false negative identification rate** (FNIR). Nonmate comparisons at or below a threshold are false positives; in identification these are sometime called **false alarms**, and they contribute to **false positive identification rate** (FPIR). The threshold can take on any real value, and it is conventional in biometrics testing to examine error rates as a function of the threshold. In many systems, the threshold can be varied continuously, while in other (production) systems, it may only take on a few settings.

Returning to the DET, it plots a function of FNIR against a function of FPIR. Here and in many other reports, the function is the logarithm function (log axes). However, a DET might also plot the **hit rate**, and the true positive identification rate, $\text{TPIR} = 1 - \text{FNIR}$ is plotted on a linear scale; this is often referred to as a ROC. More rarely, the function might be the inverse Gaussian function.

More detail and generality is provided in formal biometrics testing standards, see the various parts of [ISO/IEC 19795 Biometrics Testing and Reporting](#). More terms, including and beyond those to do with accuracy, see [ISO/IEC 2382-37 Information technology -- Vocabulary -- Part 37: Harmonized biometric vocabulary](#)

FNIR = False Negative Identification Rate

$\text{FNIR} = \text{FNIR}(N, T, L, R)$

FNIR is computed by executing mate searches into an enrolled population of size N. It is the proportion of mate searches for which the mate is

- EITHER not returned as any of L candidates,
- OR is present but has dissimilarity above threshold T
- OR is present at rank greater than R.

In IREX III, the rank criterion is not used for DET computations, i.e. $R \rightarrow \infty$, so FNIR is solely a function of population size, N and threshold, T. $\text{FNIR}(N, T)$.

FPIR = False Positive Identification Rate

$\text{FPIR} = \text{FPIR}(N, T, L)$

FPIR is computed by executing nonmate searches into an enrolled population of size N. It is the proportion of returned candidates which have dissimilarity at or below threshold T. If S searches are conducted, $S \times L$ candidates will be returned, and FPIR is the number at or below threshold, divided by $(S \times L)$.

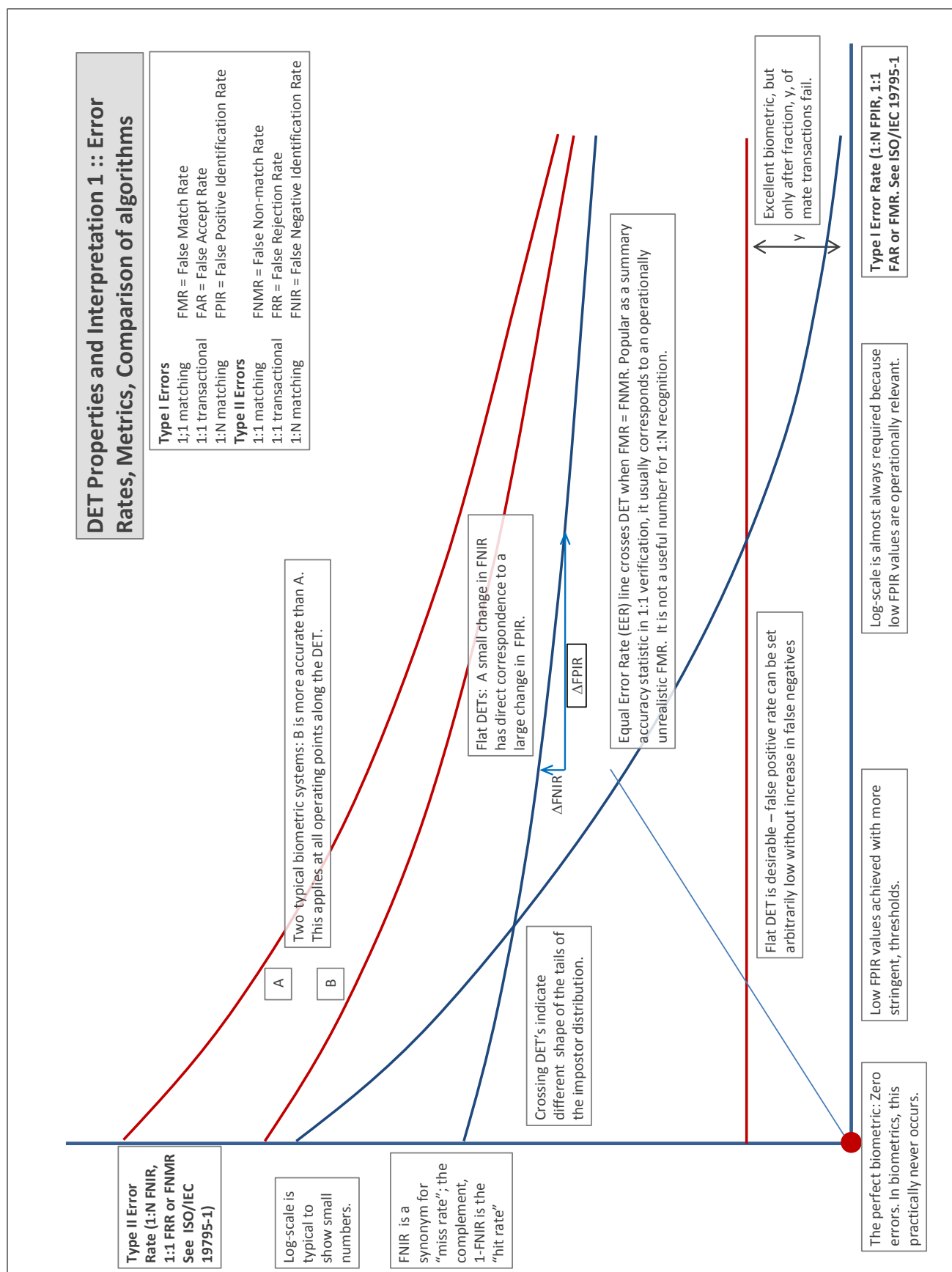


Figure 14

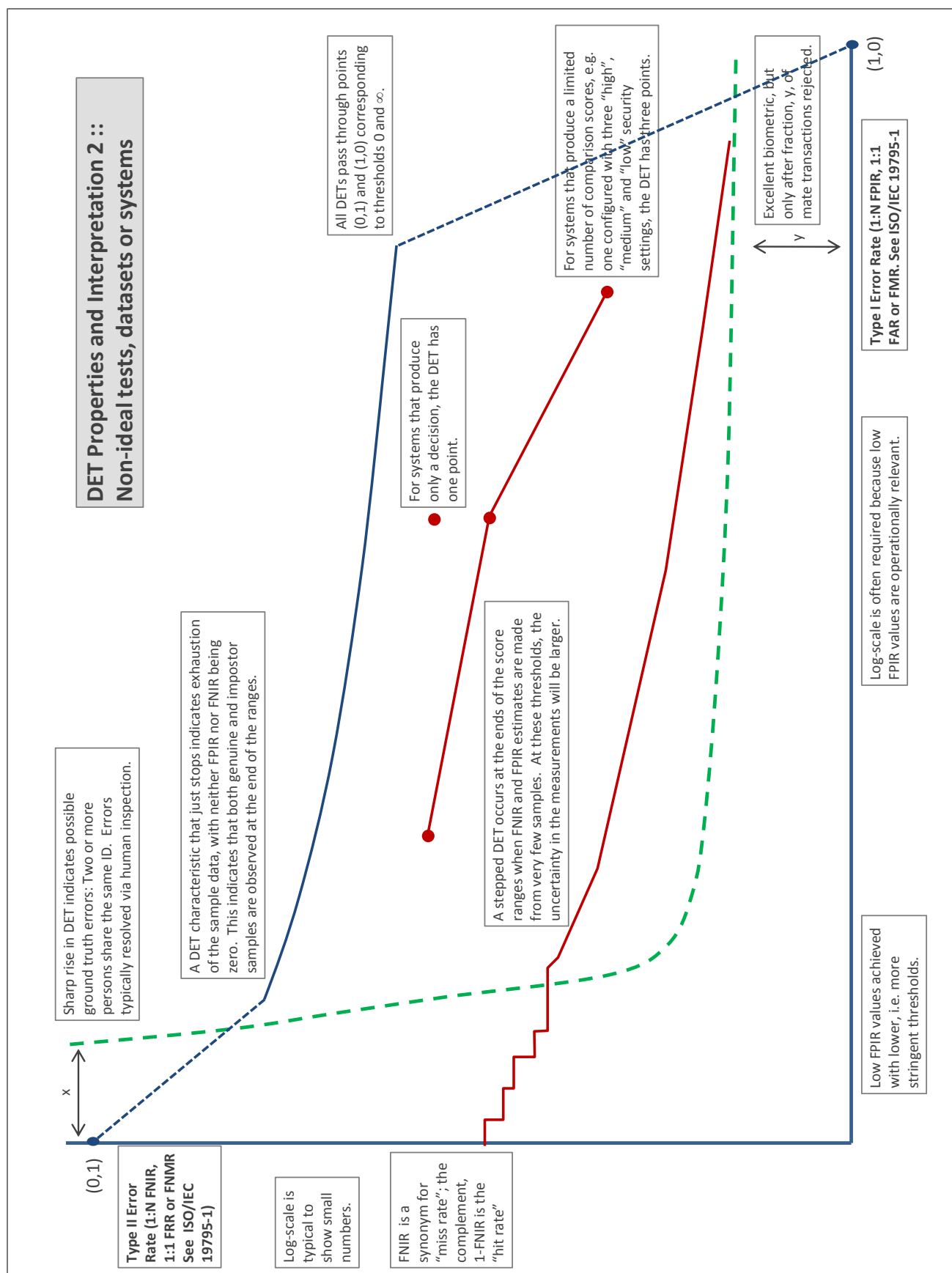


Figure 15

A = 3M/Cogent
G = Hisign
P = Zhuhai-Yisheng

B = Cognitec
H = CAS-IA
Q = JunYu

C = Neurotechnology
I = CAS-ICT
S = Decatur

D = Safran Morpho
J = Toshiba
T = Ayonix

E = NEC
L = Tsinghua U. II

F = Tsinghua U.
M = HP

FNIR(N,R,T,L) "Miss rate"
FPIR(N,T,L) "False alarm rate"

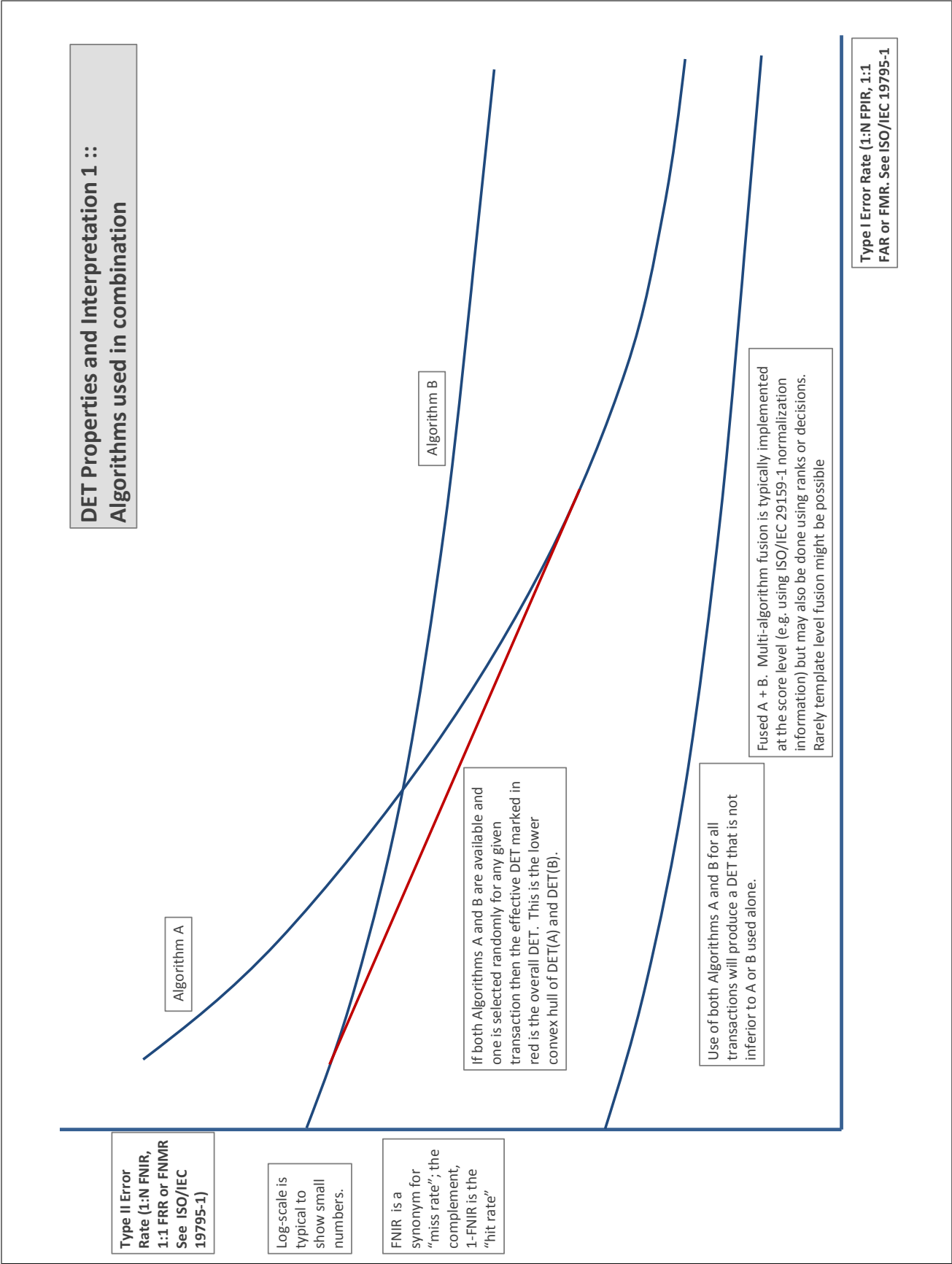


Figure 16

A Algorithm report cards

This section details individual algorithm performance by including eight graphs on a single page. Figure 17 is a key for those graphs, giving captions applicable to each report.

Error tradeoff by camera, enrolment type: The figure shows the tradeoff between false negative and false positives by plotting FNIR(T) vs. FPIR(T). This is the detection error tradeoff characteristic (DET). The plots are plotted parametrically on threshold, T , which is swept across the entire range of genuine scores produced by the algorithm. Four traces are shown, one each of the two enrolment types, and one each for the two capture processes. The mostly vertical grey lines connect points of fixed threshold: vertical lines show that only FNIR changes across those four conditions; horizontal lines show that FPIR is affected. The grey text identifies the threshold values.

Effect of rank: The figure shows the effect of considering long candidate lists by plotting miss rate against rank, i.e. FNIR(R). This is closely related to a cumulative match characteristic, which conventionally plots 1 - FNIR. Four traces are shown, one each of the two enrolment types, and one each for the two capture processes.

Error tradeoff by population size: The figure shows the tradeoff between false negative and false positives by plotting FNIR(T) vs. FPIR(T). This is the detection error tradeoff characteristic (DET). The plots are plotted parametrically on threshold, T , which is swept across the entire range of genuine scores produced by the algorithm. Four traces are shown, one each population size. The mostly vertical grey lines connect points of fixed threshold: horizontal lines show that FPIR grows linearly with population size, as expected from classical binomial models of identification. The grey text identifies the threshold values.

Workload savings: The figure plots two quantities parametrically with threshold. On the x-axis is the factor by which workload is reduced when a threshold is applied to shorten candidate lists vs. when when all $L = 50$ candidates are retained for possible inspection. On the y-axis is the factor by which thresholding candidate lists increases miss rate over the case of a full candidate list, i.e. $\text{FNIR}(N, R, T, L) / \text{FNIR}(N, L, 0, L)$.

Miss rates increase with enrolled population sizes: The figure shows the growth of miss rates with N for two enrolment types, two cameras, and two rank values, 1 and 50. The threshold is set to zero, and the values are simply the proportion of mated searches that do not yield the mate in the top R ranks. The traces are typically straight lines on a log-log plot consistent with a power-law behavior (see 5.3.1).

Processing time: The figure shows three traces: a horizontal line indicating the time taken to produce a template from an image prior to search; an ascending line indicating the time taken to compare a template with data from N enrolled subjects; and the sum of these two durations - the total search time. The curves cross where N is large enough such that the 1: N search time exceeds the template generation time. All durations apply to processing on a single core of a c. 2011 server-class processor. All templates are resident in memory.

Value of enrolling historical images: The figure shows the reduction in miss rates with the number of available enrolment images, for four population sizes. As presented in section 3.2 when images from all historical encounters are retained and enrolled, accuracy can be improved vs. the case where only one image is retained. The error bars indicate confidence intervals from bootstrapping applied over searches (see section 4.6). The relevance of this result is discussed further in section 5.4.

Selectivity vs. FPIR: As presented in section 4.1, false alarm rates can be quantified by FPIR - the fraction of nonmate searches that produce *any* candidates at or above threshold - or by selectivity - the expected *number* of candidates at or above threshold. This figure shows SEL(T) vs. FPIR(T) plotted parametrically on threshold T . Selectivity is always greater than or equal to FPIR. The two are not equal when false positives are concentrated in candidate lists rather than being distributed across searches.

Figure 17: **Key to report card figures:** The boxed text of this figure describe the graphs that appear in the report cards of this Appendix. Each report card contains 8 graphs, with a one-to-one spatial correspondence with this Figure.

A = 3M/Cogent	B = Cognitec	C = Neurotechnology	D = Safran Morpho	E = NEC	F = Tsinghua U.	FNIR(N,R,T,L) "Miss rate"
G = Hisign	H = CAS-IA	I = CAS-ICT	J = Toshiba	L = Tsinghua U. II	M = HP	FPIR(N,T,L) "False alarm rate"
P = Zhuhai-Yisheng	Q = JunYu	S = Decatur	T = Ayonix			

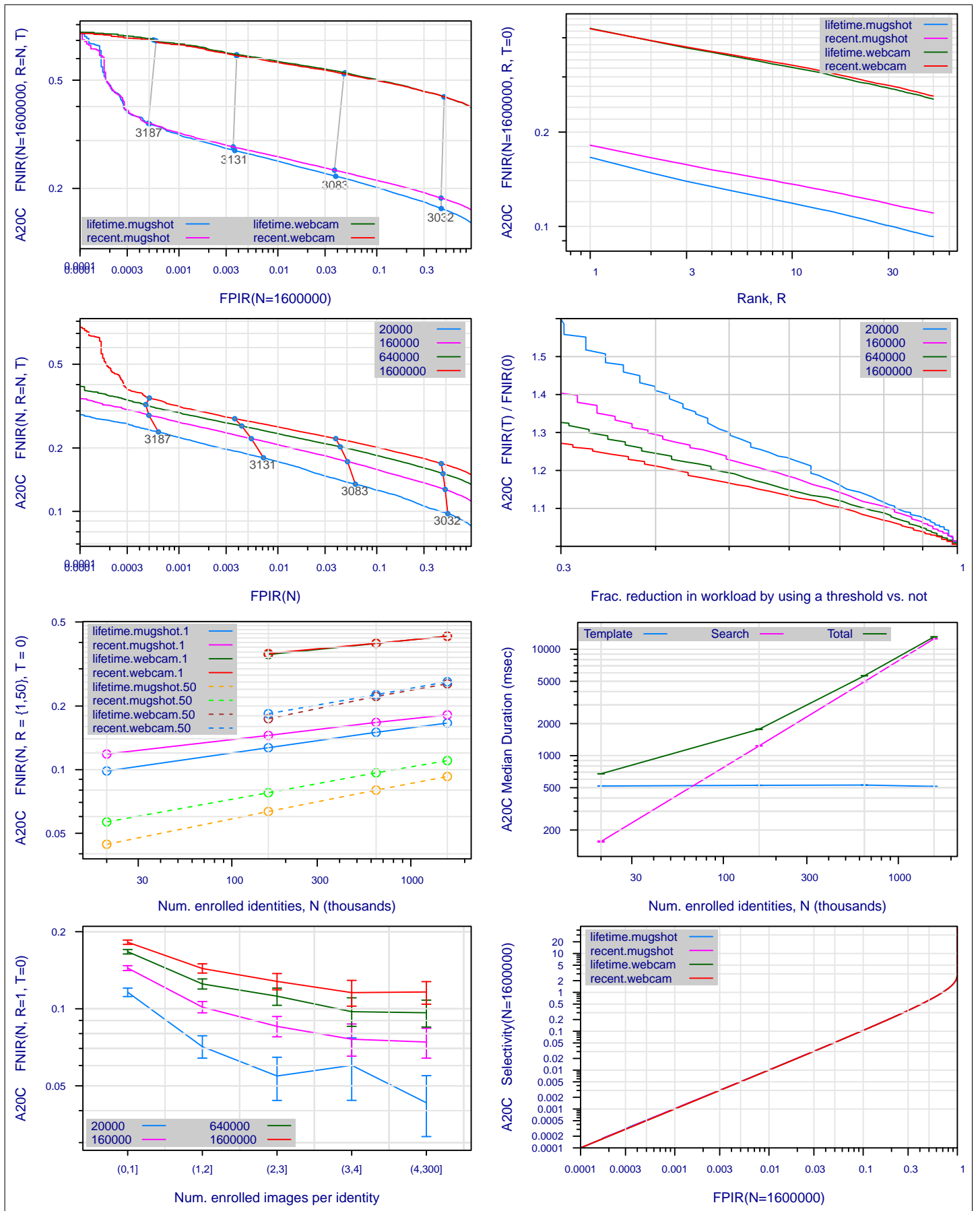


Figure 18: Collected performance reports for algorithm A20C. The figures are described at the beginning of this Appendix.

A = 3M/Cogent	B = Cognitec	C = Neurotechnology	D = Safran Morpho	E = NEC	F = Tsinghua U.	FNIR(N,R,T,L) "Miss rate"
G = Hisign	H = CAS-IA	I = CAS-ICT	J = Toshiba	L = Tsinghua U. II	M = HP	FPIR(N,T,L) "False alarm rate"
P = Zhuhai-Yisheng	Q = JunYu	S = Decatur	T = Ayonix			

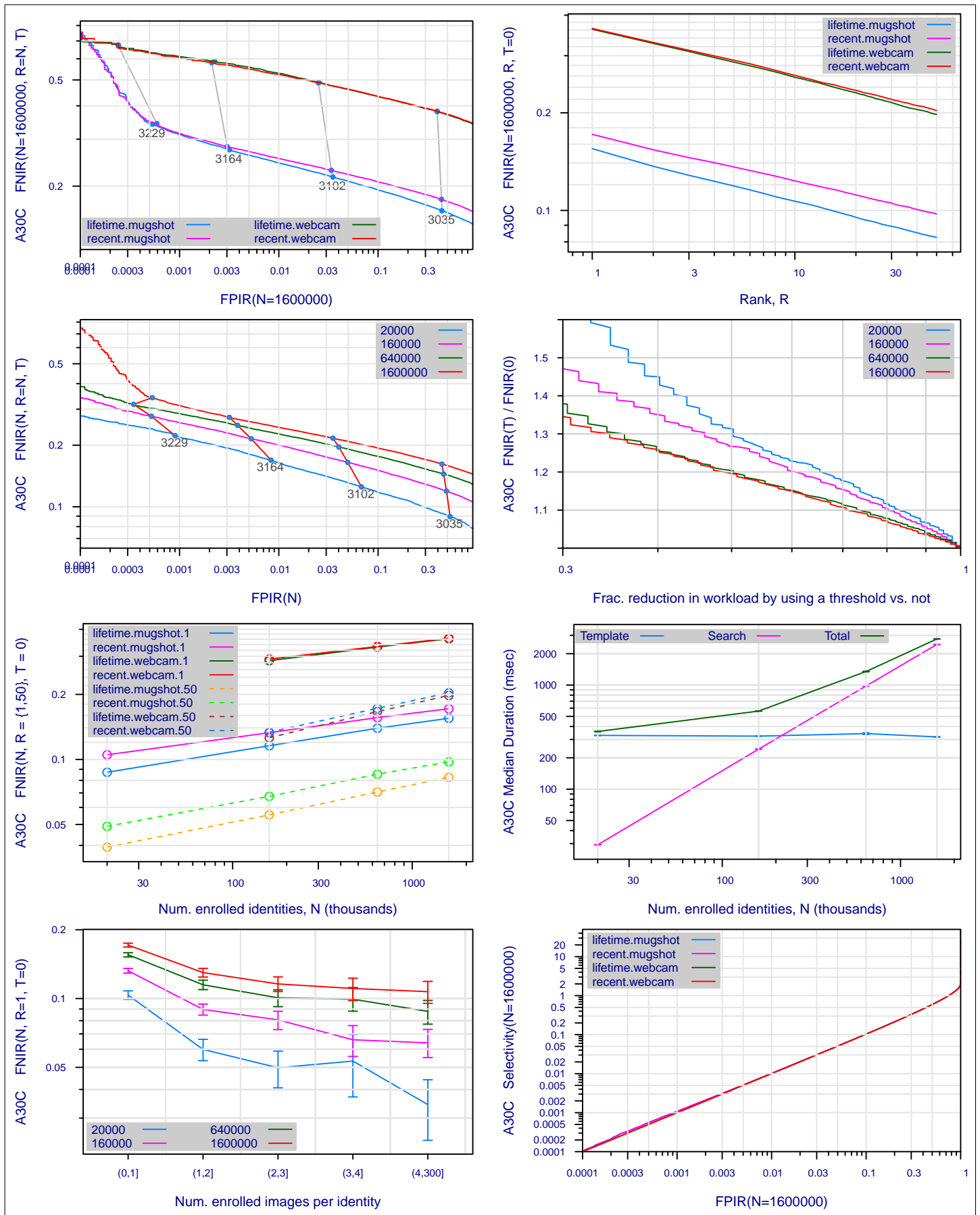


Figure 19: Collected performance reports for algorithm A30C. The figures are described at the beginning of this Appendix.

A = 3M/Cogent	B = Cognitec	C = Neurotechnology	D = Safran Morpho	E = NEC	F = Tsinghua U.	FNIR(N,R,T,L) "Miss rate"
G = Hisign	H = CAS-IA	I = CAS-ICT	J = Toshiba	L = Tsinghua U. II	M = HP	FPIR(N,T,L) "False alarm rate"
P = Zhuhai-Yisheng	Q = JunYu	S = Decatur	T = Ayonix			

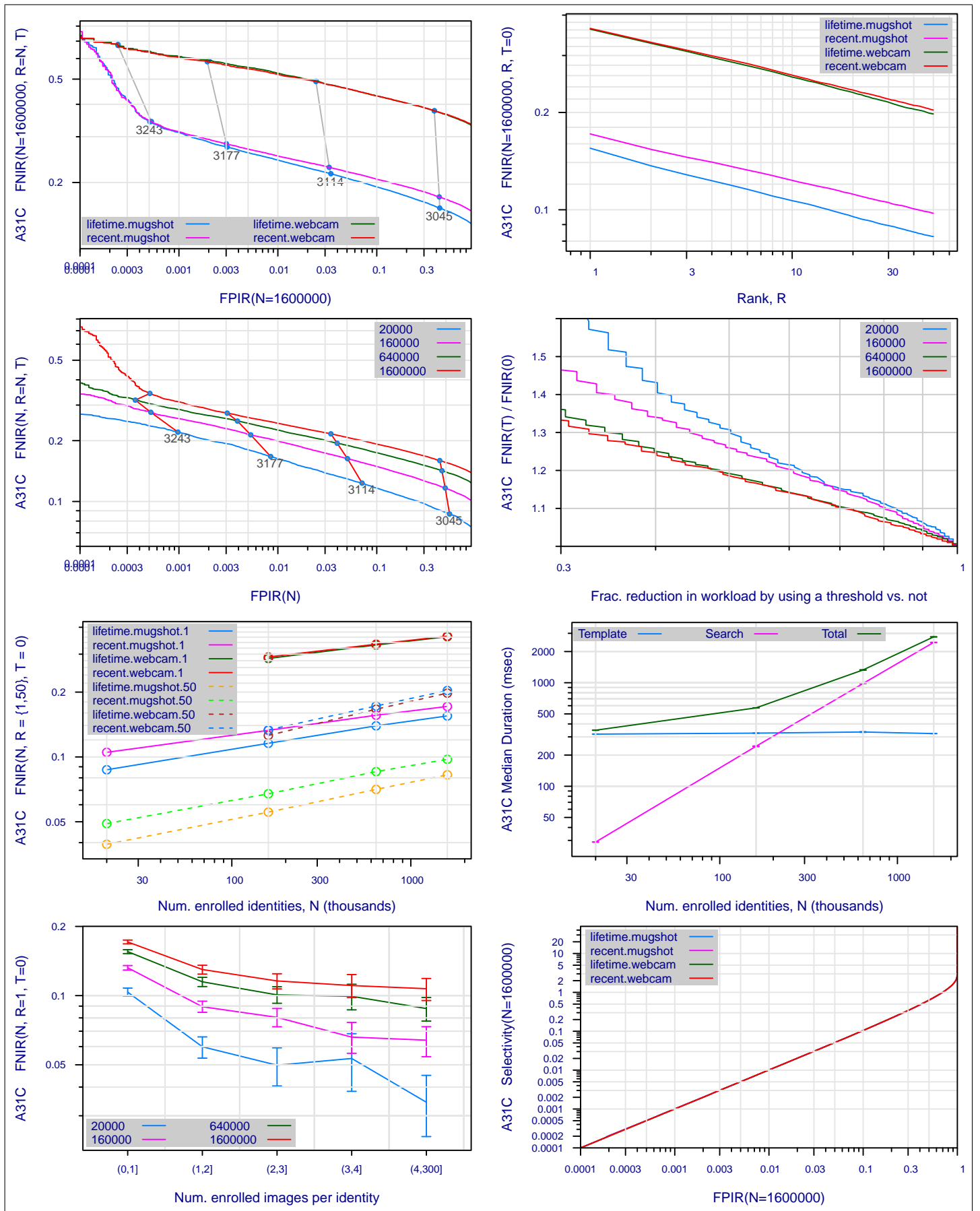


Figure 20: Collected performance reports for algorithm A31C. The figures are described at the beginning of this Appendix.

A = 3M/Cogent	B = Cognitec	C = Neurotechnology	D = Safran Morpho	E = NEC	F = Tsinghua U.	FNIR(N,R,T,L) "Miss rate"
G = Hisign	H = CAS-IA	I = CAS-ICT	J = Toshiba	L = Tsinghua U. II	M = HP	FPIR(N,T,L) "False alarm rate"
P = Zhuhai-Yisheng	Q = JunYu	S = Decatur	T = Ayonix			

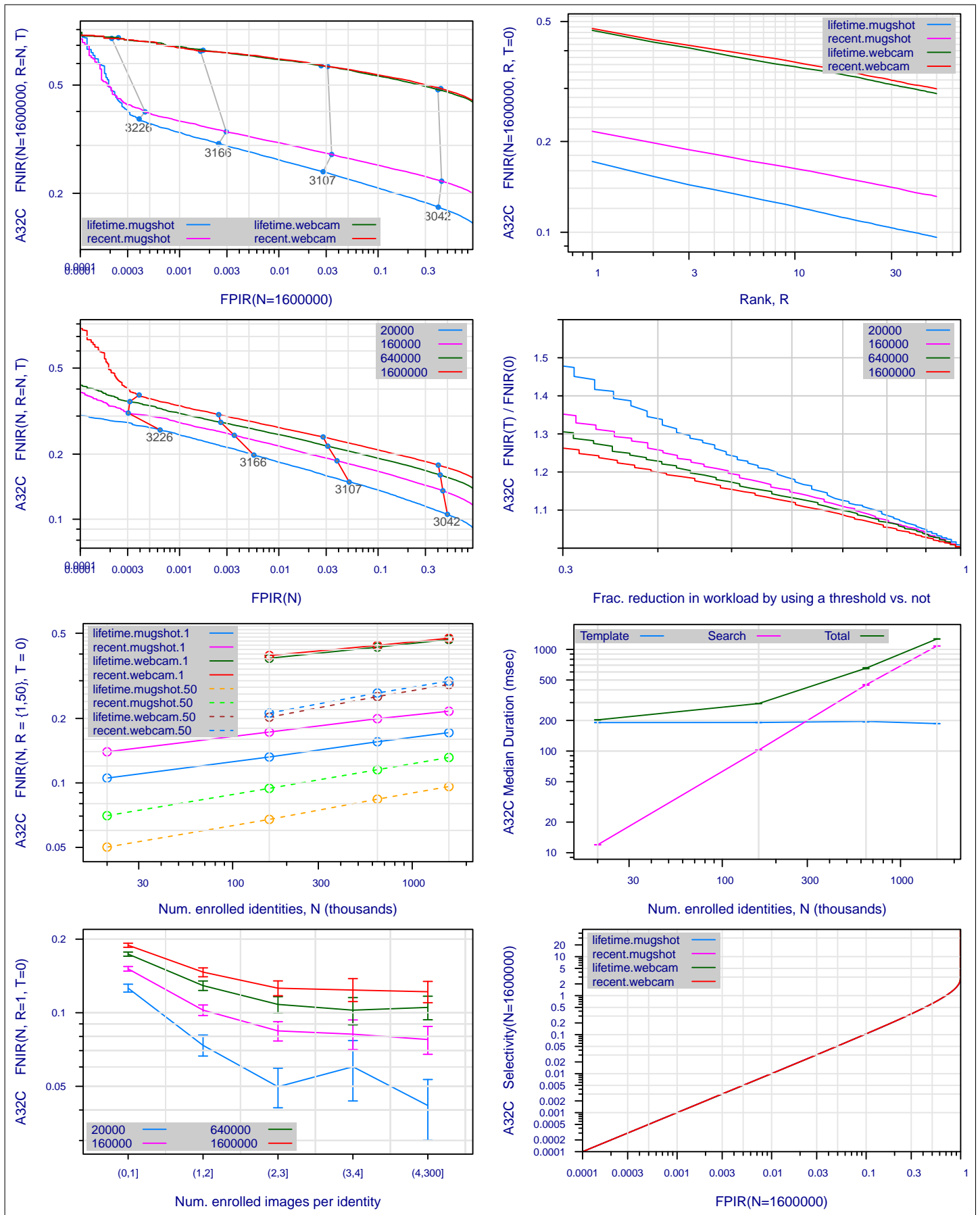


Figure 21: Collected performance reports for algorithm A32C. The figures are described at the beginning of this Appendix.

A = 3M/Cogent	B = Cognitec	C = Neurotechnology	D = Safran Morpho	E = NEC	F = Tsinghua U.	FNIR(N,R,T,L) "Miss rate"
G = Hisign	H = CAS-IA	I = CAS-ICT	J = Toshiba	L = Tsinghua U. II	M = HP	FPIR(N,T,L) "False alarm rate"
P = Zhuhai-Yisheng	Q = JunYu	S = Decatur	T = Ayonix			

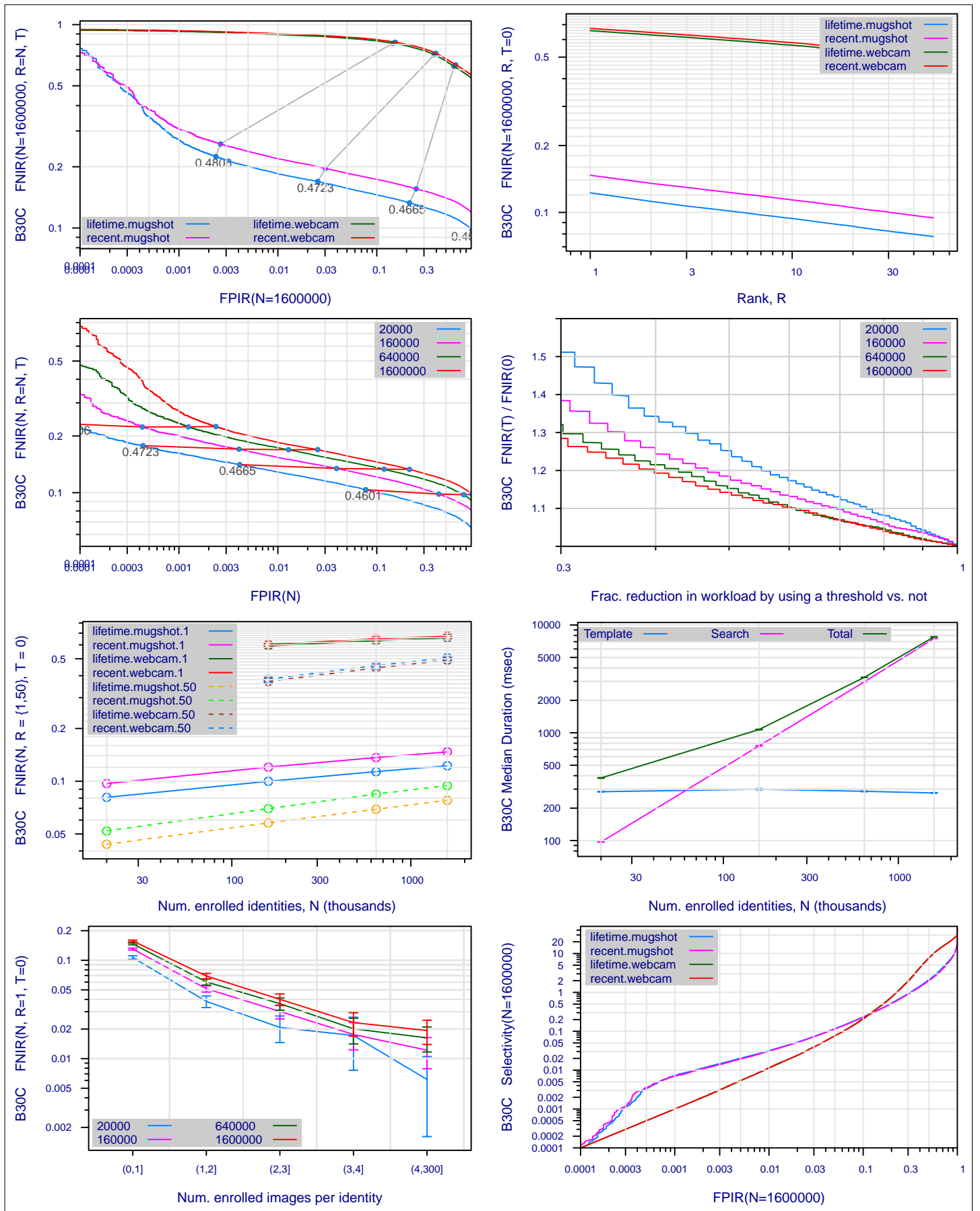


Figure 22: Collected performance reports for algorithm B30C. The figures are described at the beginning of this Appendix.

A = 3M/Cogent	B = Cognitec	C = Neurotechnology	D = Safran Morpho	E = NEC	F = Tsinghua U.	FNIR(N,R,T,L) "Miss rate"
G = Hisign	H = CAS-IA	I = CAS-ICT	J = Toshiba	L = Tsinghua U. II	M = HP	FPIR(N,T,L) "False alarm rate"
P = Zhuhai-Yisheng	Q = JunYu	S = Decatur	T = Ayonix			

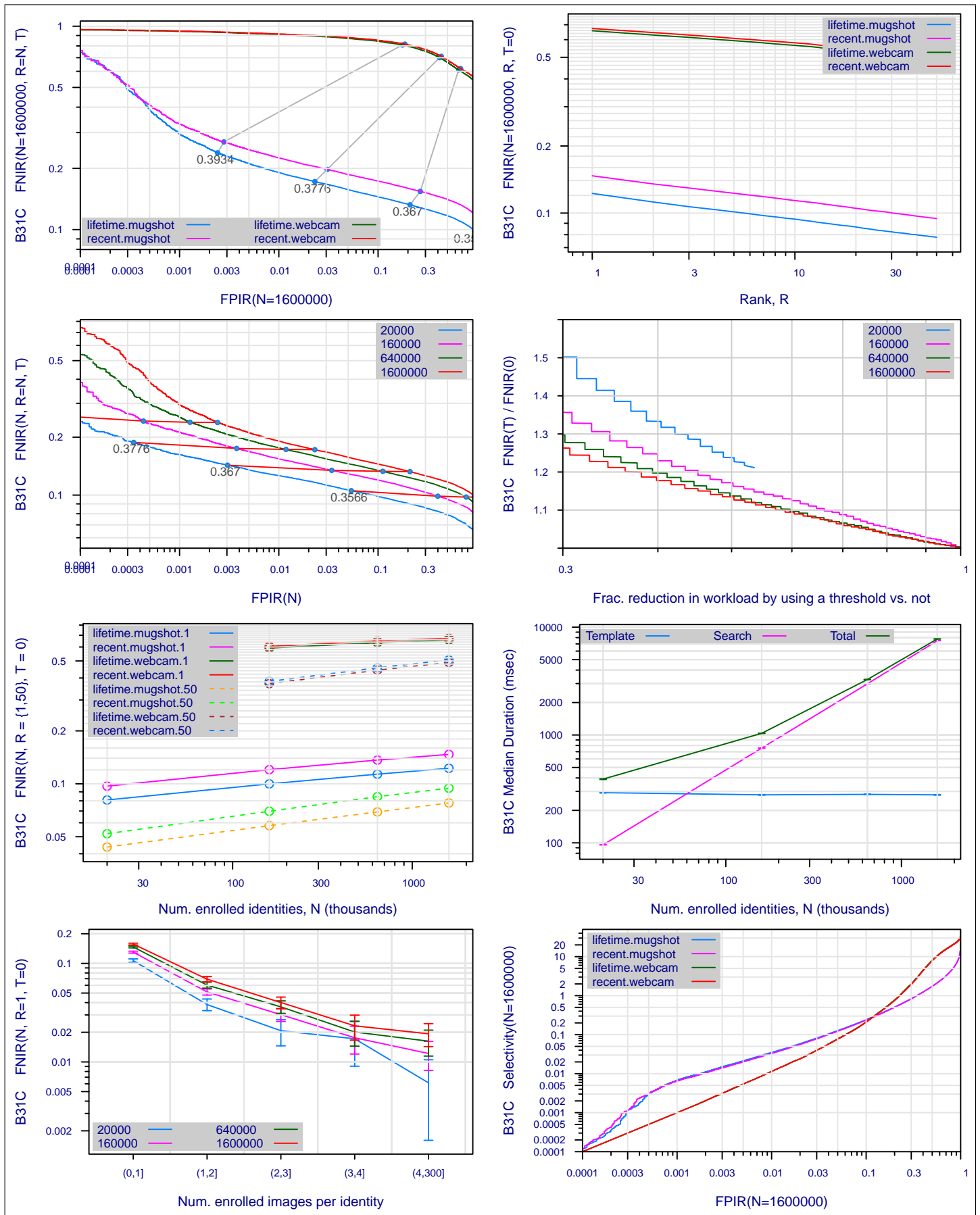


Figure 23: Collected performance reports for algorithm B31C. The figures are described at the beginning of this Appendix.

A = 3M/Cogent	B = Cognitec	C = Neurotechnology	D = Safran Morpho	E = NEC	F = Tsinghua U.	FNIR(N,R,T,L) "Miss rate"
G = Hisign	H = CAS-IA	I = CAS-ICT	J = Toshiba	L = Tsinghua U. II	M = HP	FPIR(N,T,L) "False alarm rate"
P = Zhuhai-Yisheng	Q = JunYu	S = Decatur	T = Ayonix			

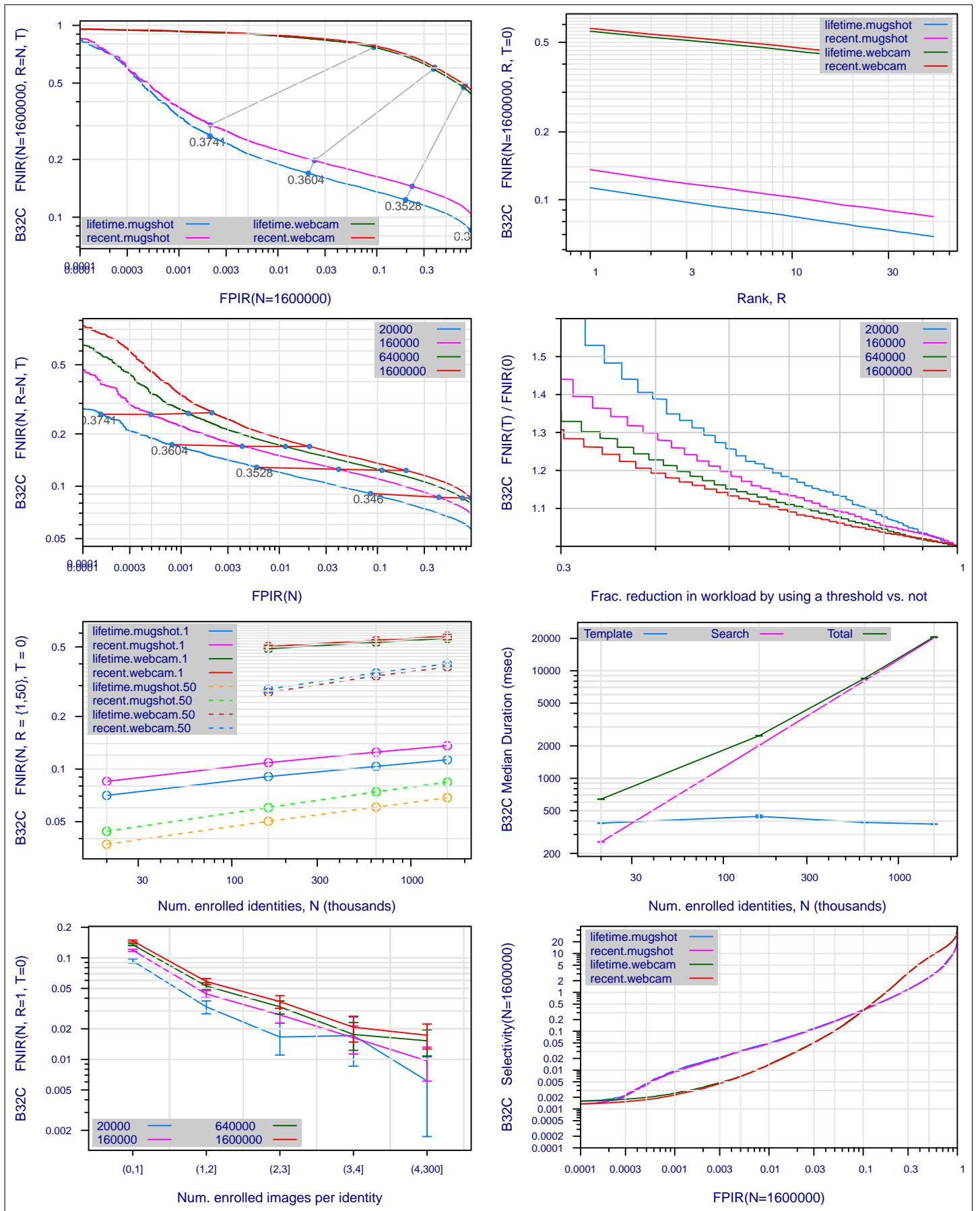


Figure 24: Collected performance reports for algorithm B32C. The figures are described at the beginning of this Appendix.

A = 3M/Cogent	B = Cognitec	C = Neurotechnology	D = Safran Morpho	E = NEC	F = Tsinghua U.	FNIR(N,R,T,L) "Miss rate"
G = Hisign	H = CAS-IA	I = CAS-ICT	J = Toshiba	L = Tsinghua U. II	M = HP	FPIR(N,T,L) "False alarm rate"
P = Zhuhai-Yisheng	Q = JunYu	S = Decatur	T = Ayonix			

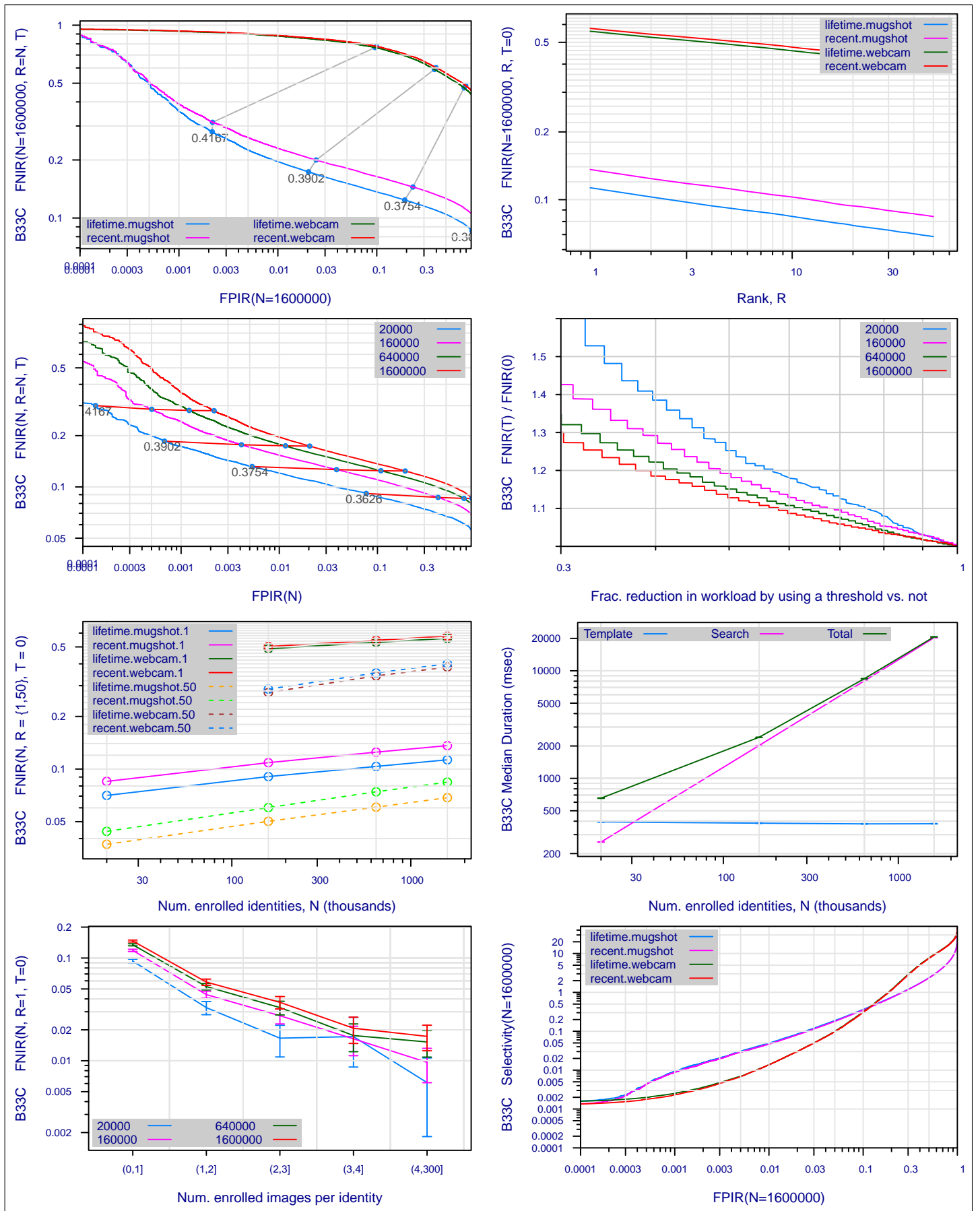


Figure 25: Collected performance reports for algorithm B33C. The figures are described at the beginning of this Appendix.

A = 3M/Cogent	B = Cognitec	C = Neurotechnology	D = Safran Morpho	E = NEC	F = Tsinghua U.	FNIR(N,R,T,L) "Miss rate"
G = Hisign	H = CAS-IA	I = CAS-ICT	J = Toshiba	L = Tsinghua U. II	M = HP	FPIR(N,T,L) "False alarm rate"
P = Zhuhai-Yisheng	Q = JunYu	S = Decatur	T = Ayonix			

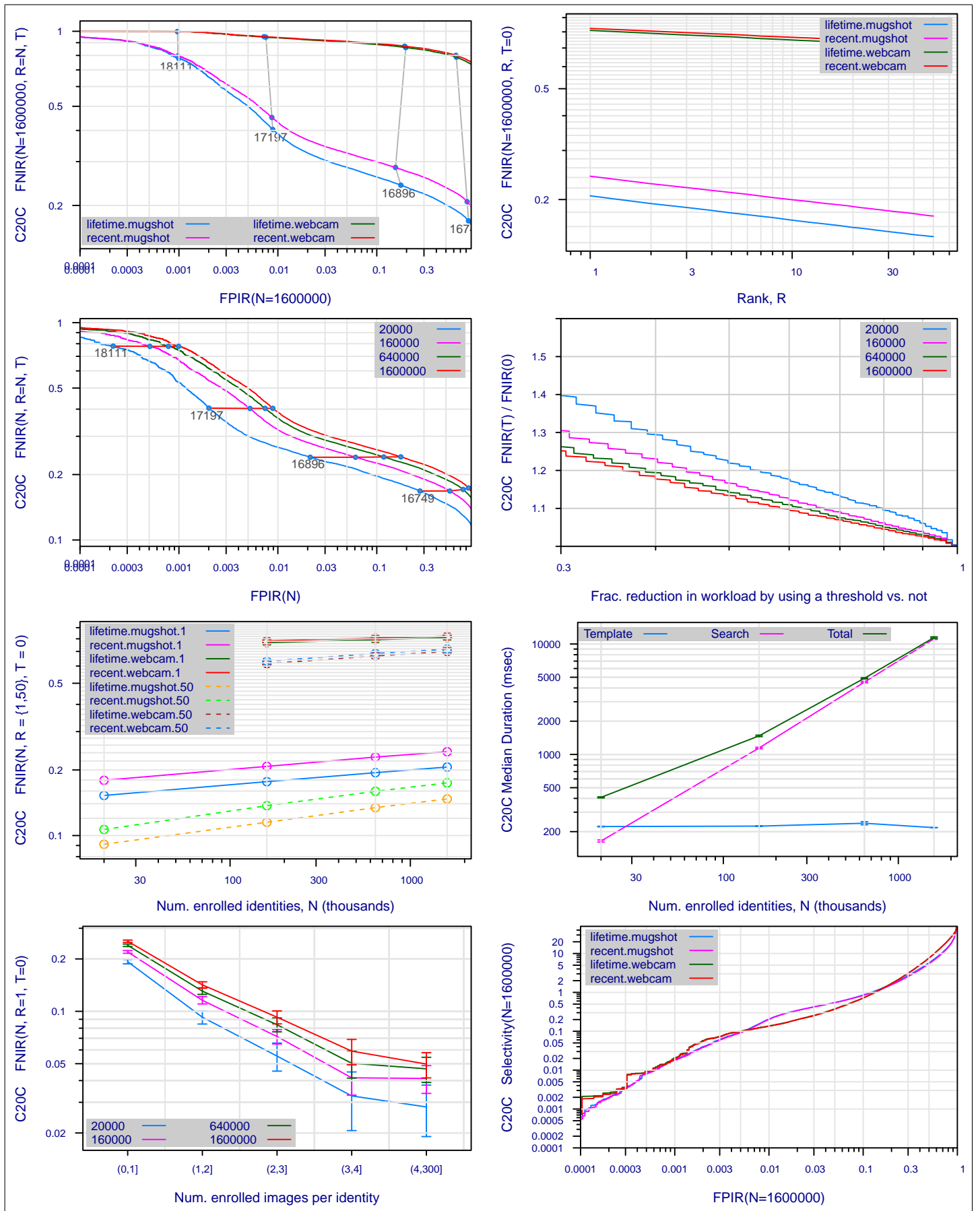


Figure 26: Collected performance reports for algorithm C20C. The figures are described at the beginning of this Appendix.

A = 3M/Cogent	B = Cognitec	C = Neurotechnology	D = Safran Morpho	E = NEC	F = Tsinghua U.	FNIR(N,R,T,L) "Miss rate"
G = Hisign	H = CAS-IA	I = CAS-ICT	J = Toshiba	L = Tsinghua U. II	M = HP	FPIR(N,T,L) "False alarm rate"
P = Zhuhai-Yisheng	Q = JunYu	S = Decatur	T = Ayonix			

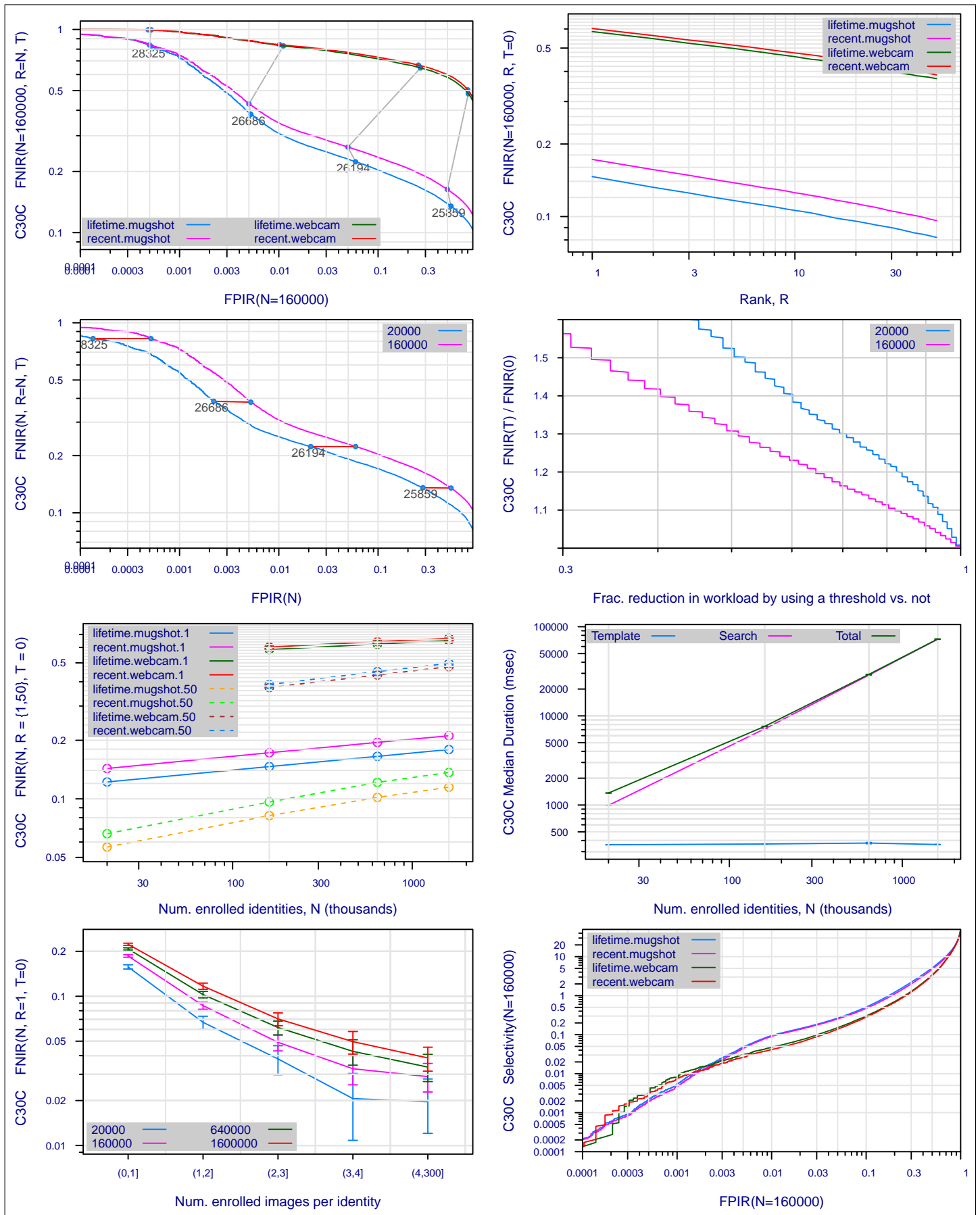


Figure 27: Collected performance reports for algorithm C30C. The figures are described at the beginning of this Appendix.

A = 3M/Cogent	B = Cognitec	C = Neurotechnology	D = Safran Morpho	E = NEC	F = Tsinghua U.	FNIR(N,R,T,L) "Miss rate"
G = Hisign	H = CAS-IA	I = CAS-ICT	J = Toshiba	L = Tsinghua U. II	M = HP	FPIR(N,T,L) "False alarm rate"
P = Zhuhai-Yisheng	Q = JunYu	S = Decatur	T = Ayonix			

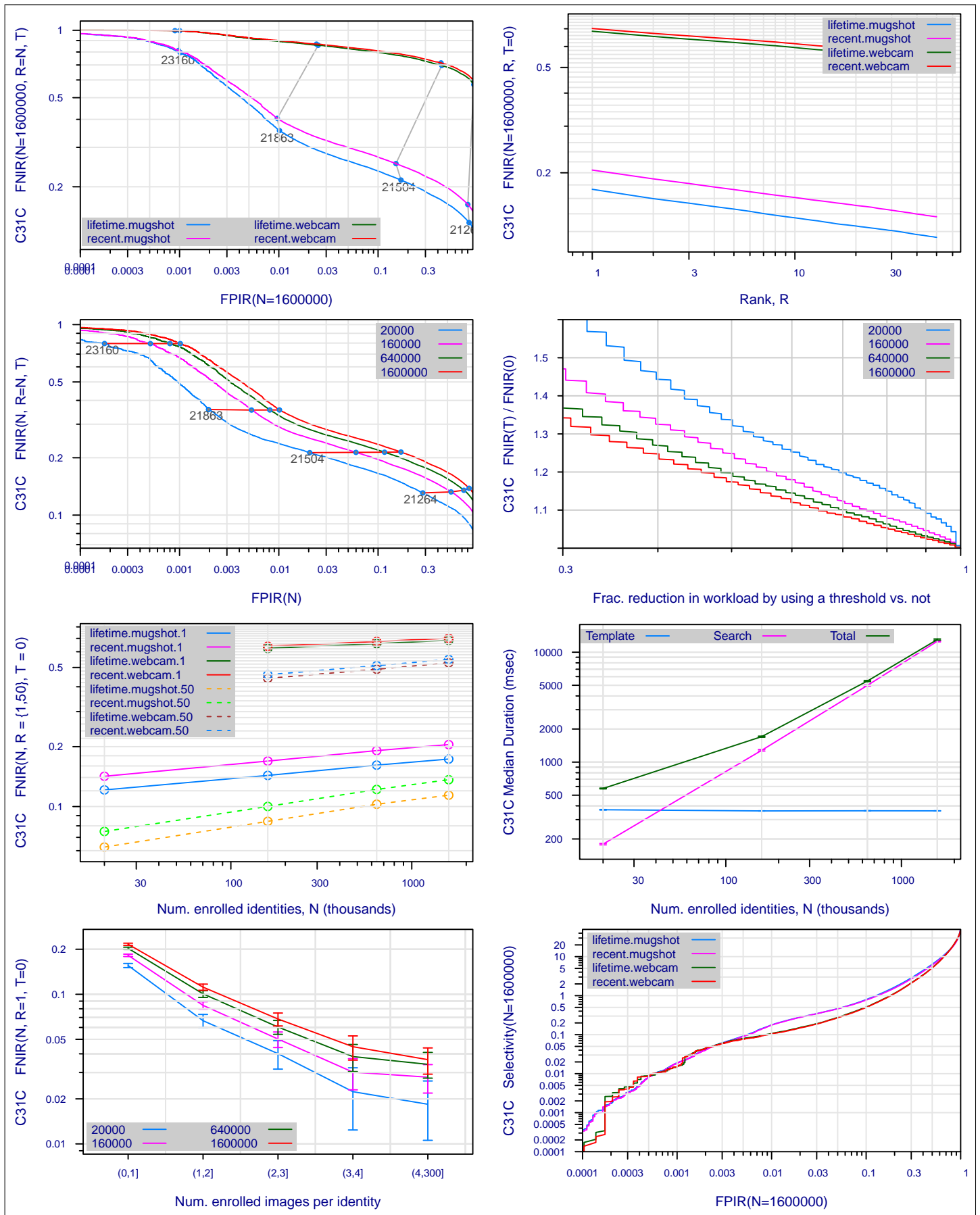


Figure 28: Collected performance reports for algorithm C31C. The figures are described at the beginning of this Appendix.

A = 3M/Cogent	B = Cognitec	C = Neurotechnology	D = Safran Morpho	E = NEC	F = Tsinghua U.	FNIR(N,R,T,L) "Miss rate"
G = Hisign	H = CAS-IA	I = CAS-ICT	J = Toshiba	L = Tsinghua U. II	M = HP	FPIR(N,T,L) "False alarm rate"
P = Zhuhai-Yisheng	Q = JunYu	S = Decatur	T = Ayonix			

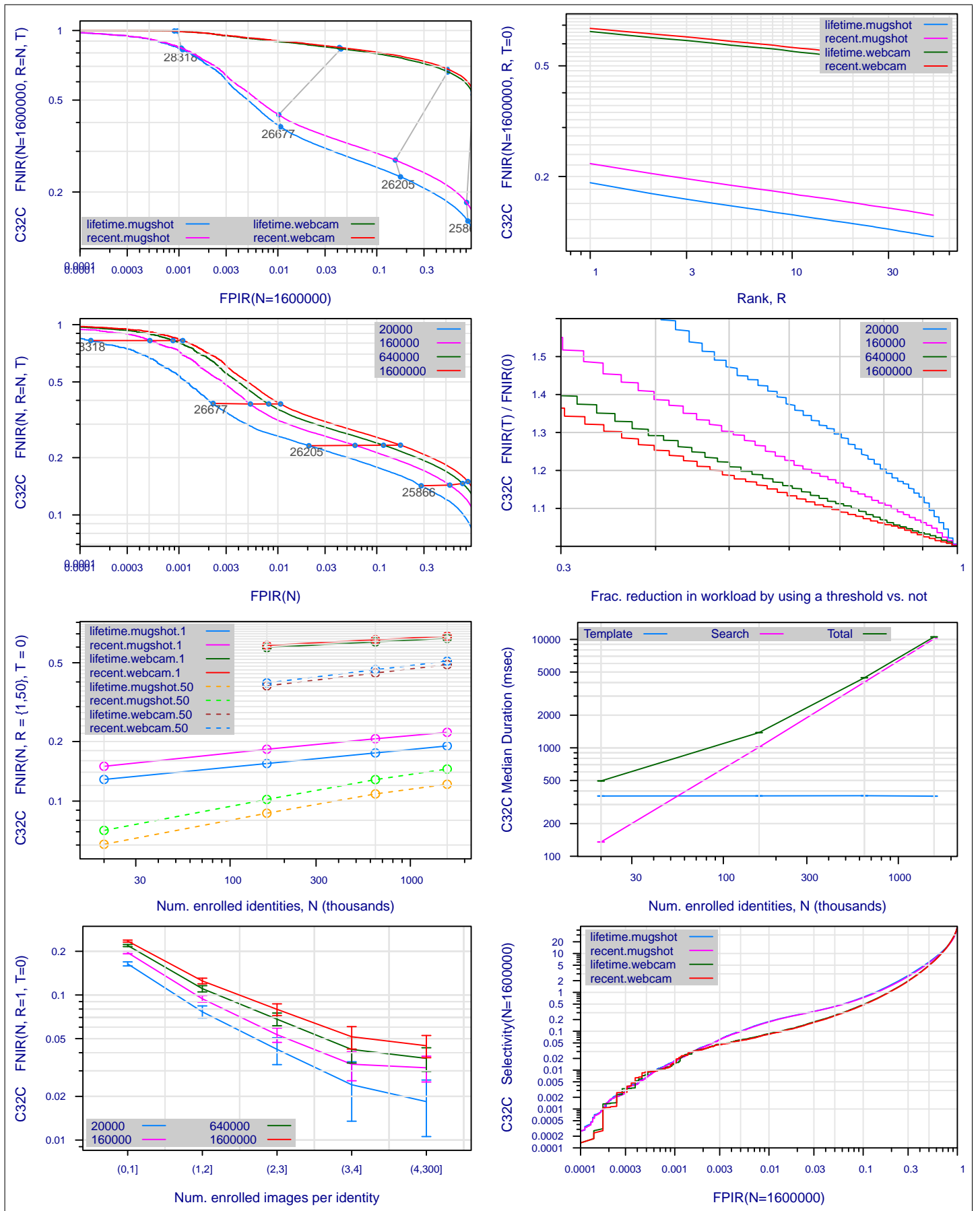


Figure 29: Collected performance reports for algorithm C32C. The figures are described at the beginning of this Appendix.

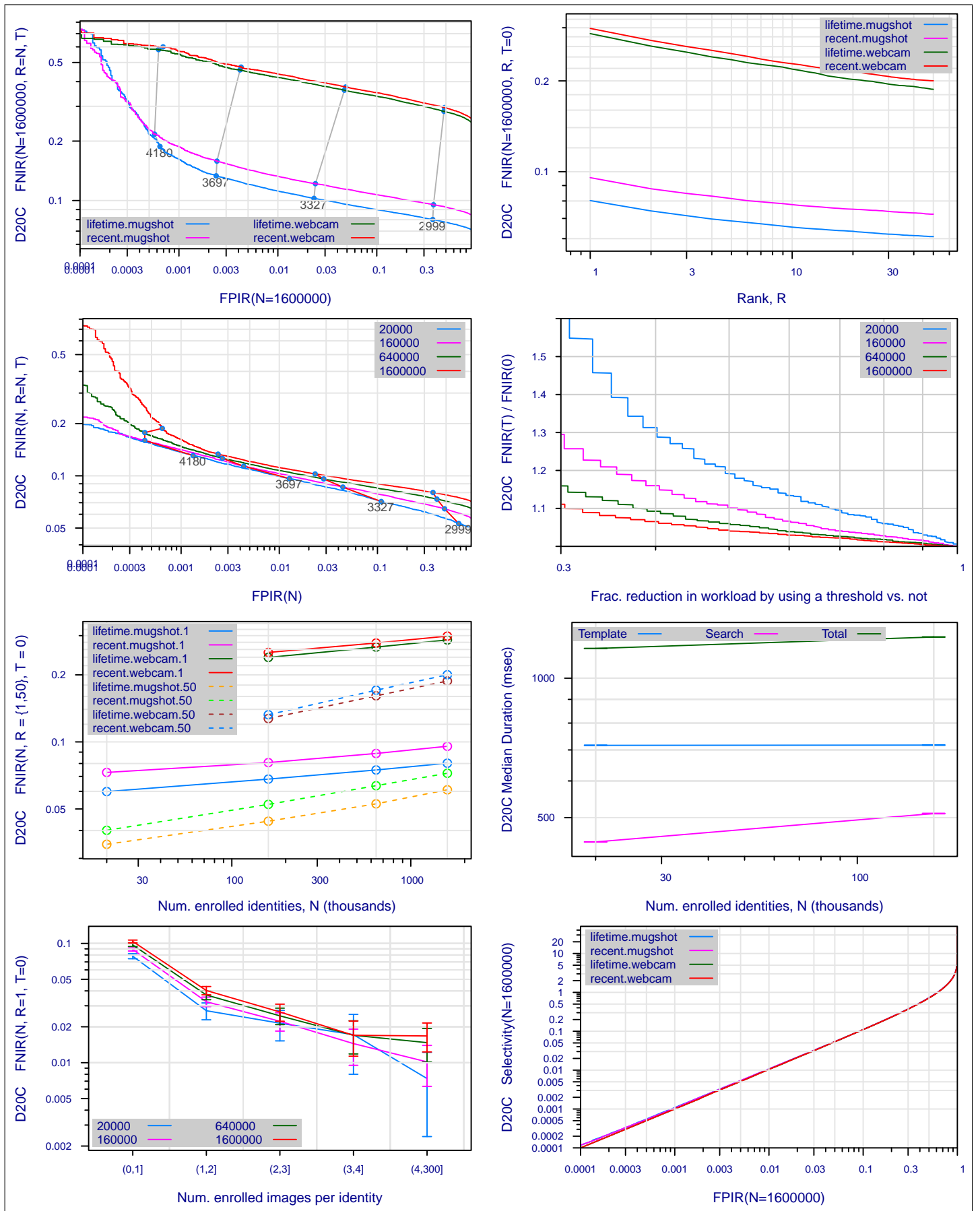


Figure 30: Collected performance reports for algorithm D20C. The figures are described at the beginning of this Appendix.

A = 3M/Cogent	B = Cognitec	C = Neurotechnology	D = Safran Morpho	E = NEC	F = Tsinghua U.	FNIR(N,R,T,L) "Miss rate"
G = Hisign	H = CAS-IA	I = CAS-ICT	J = Toshiba	L = Tsinghua U. II	M = HP	FPIR(N,T,L) "False alarm rate"
P = Zhuhai-Yisheng	Q = JunYu	S = Decatur	T = Ayonix			

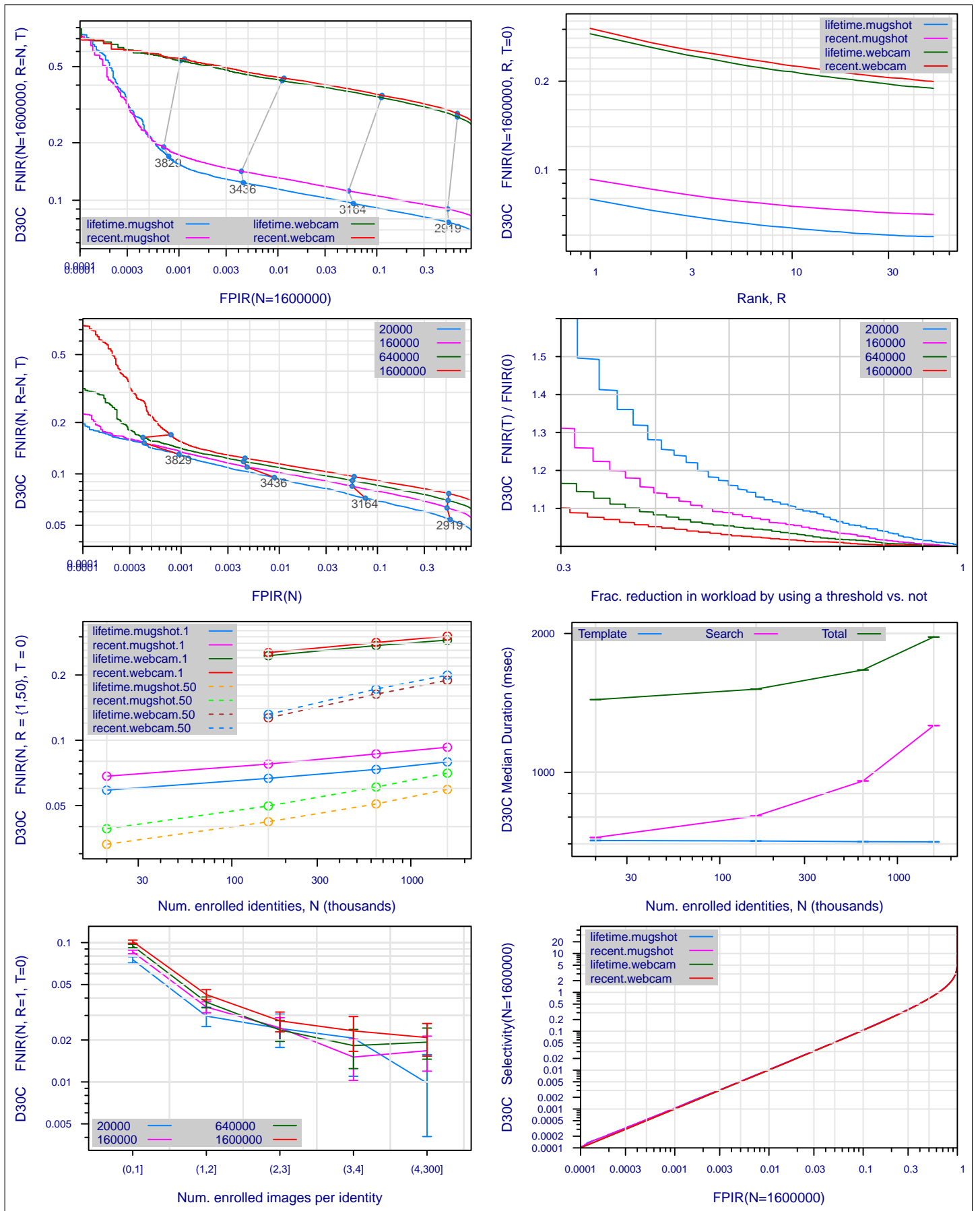


Figure 31: Collected performance reports for algorithm D30C. The figures are described at the beginning of this Appendix.

A = 3M/Cogent	B = Cognitec	C = Neurotechnology	D = Safran Morpho	E = NEC	F = Tsinghua U.	FNIR(N,R,T,L) "Miss rate"
G = Hisign	H = CAS-IA	I = CAS-ICT	J = Toshiba	L = Tsinghua U. II	M = HP	FPIR(N,T,L) "False alarm rate"
P = Zhuhai-Yisheng	Q = JunYu	S = Decatur	T = Ayonix			

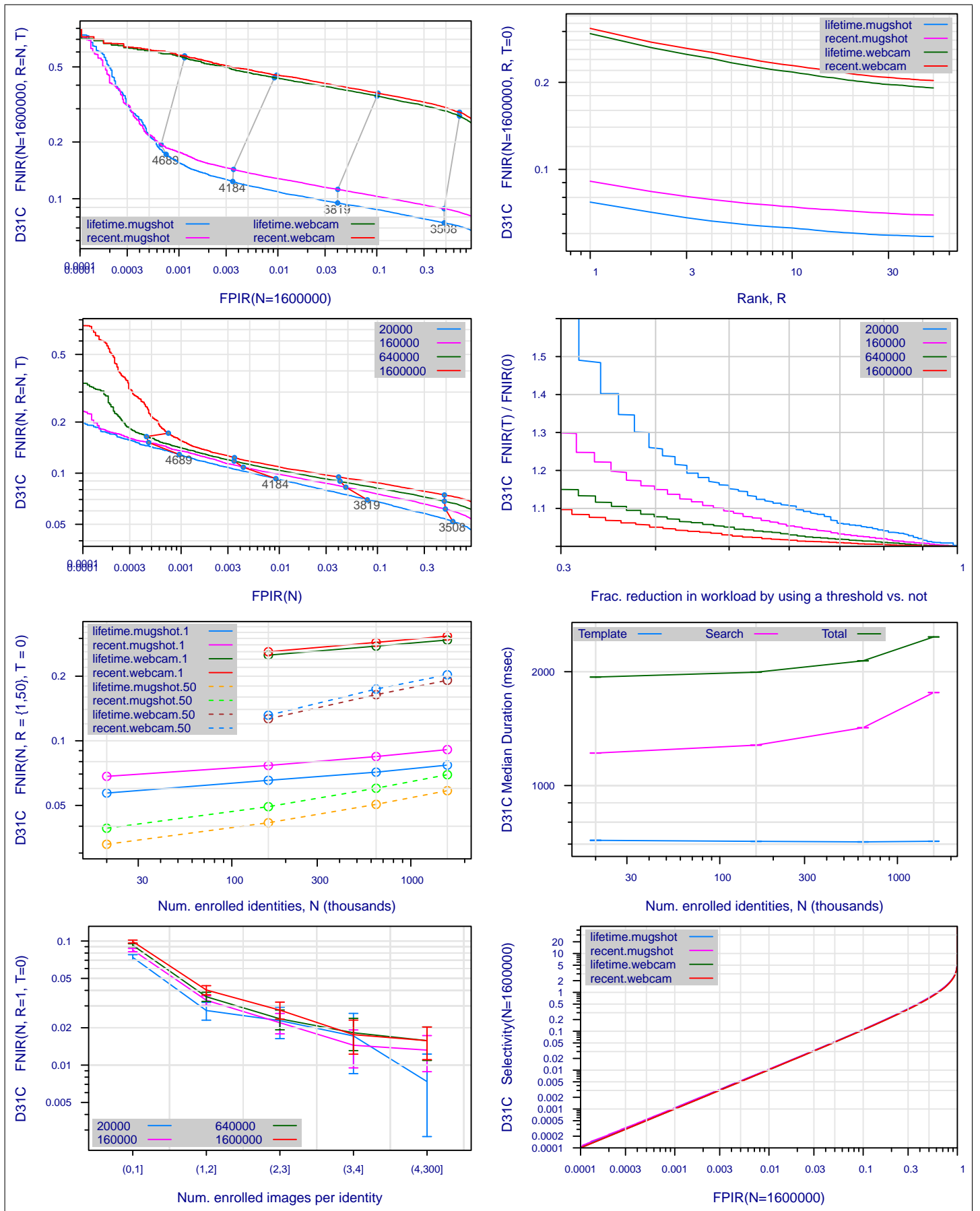


Figure 32: Collected performance reports for algorithm D31C. The figures are described at the beginning of this Appendix.

A = 3M/Cogent	B = Cognitec	C = Neurotechnology	D = Safran Morpho	E = NEC	F = Tsinghua U.	FNIR(N,R,T,L) "Miss rate"
G = Hisign	H = CAS-IA	I = CAS-ICT	J = Toshiba	L = Tsinghua U. II	M = HP	FPIR(N,T,L) "False alarm rate"
P = Zhuhai-Yisheng	Q = JunYu	S = Decatur	T = Ayonix			

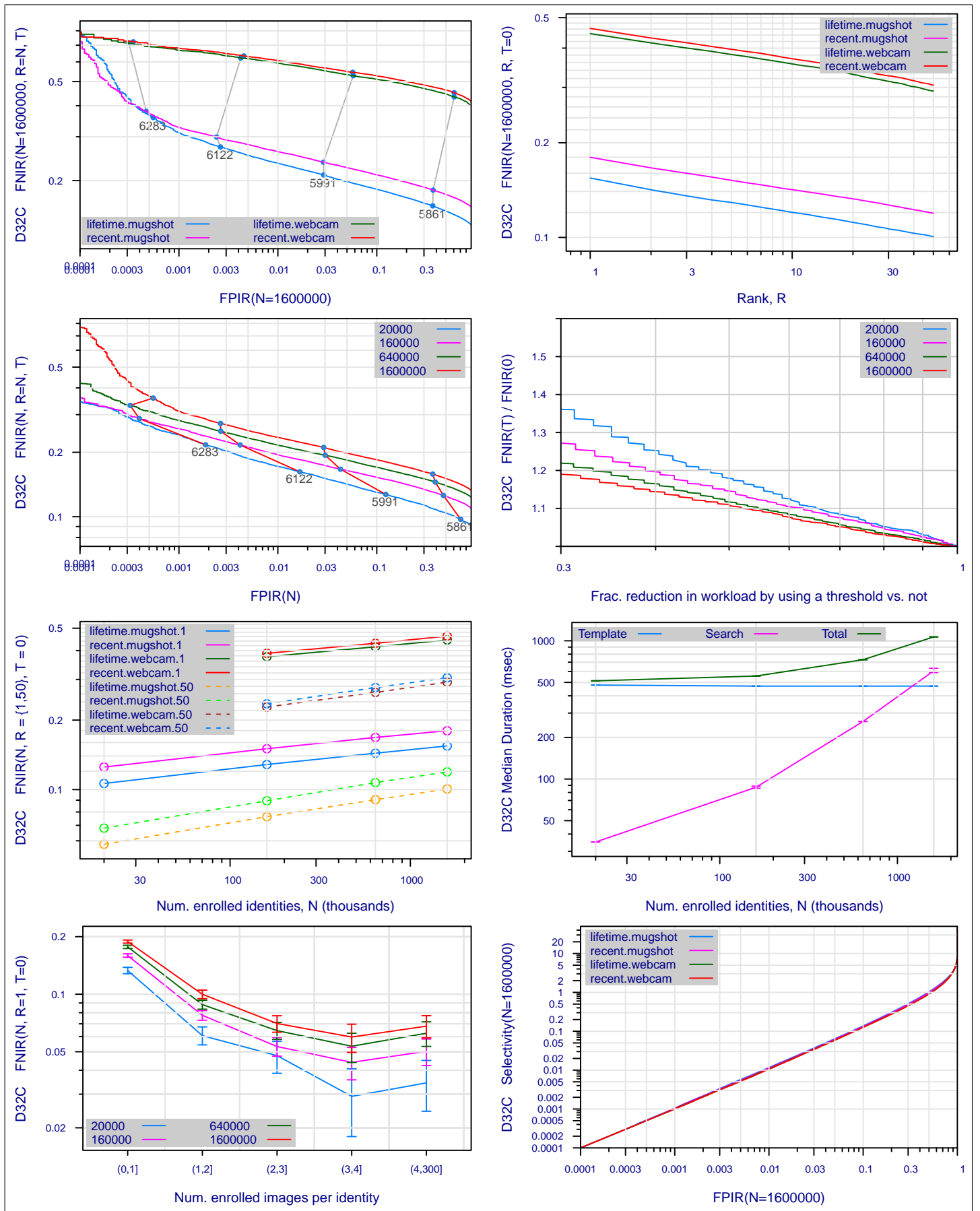


Figure 33: Collected performance reports for algorithm D32C. The figures are described at the beginning of this Appendix.

A = 3M/Cogent	B = Cognitec	C = Neurotechnology	D = Safran Morpho	E = NEC	F = Tsinghua U.	FNIR(N,R,T,L) "Miss rate"
G = Hisign	H = CAS-IA	I = CAS-ICT	J = Toshiba	L = Tsinghua U. II	M = HP	FPIR(N,T,L) "False alarm rate"
P = Zhuhai-Yisheng	Q = JunYu	S = Decatur	T = Ayonix			

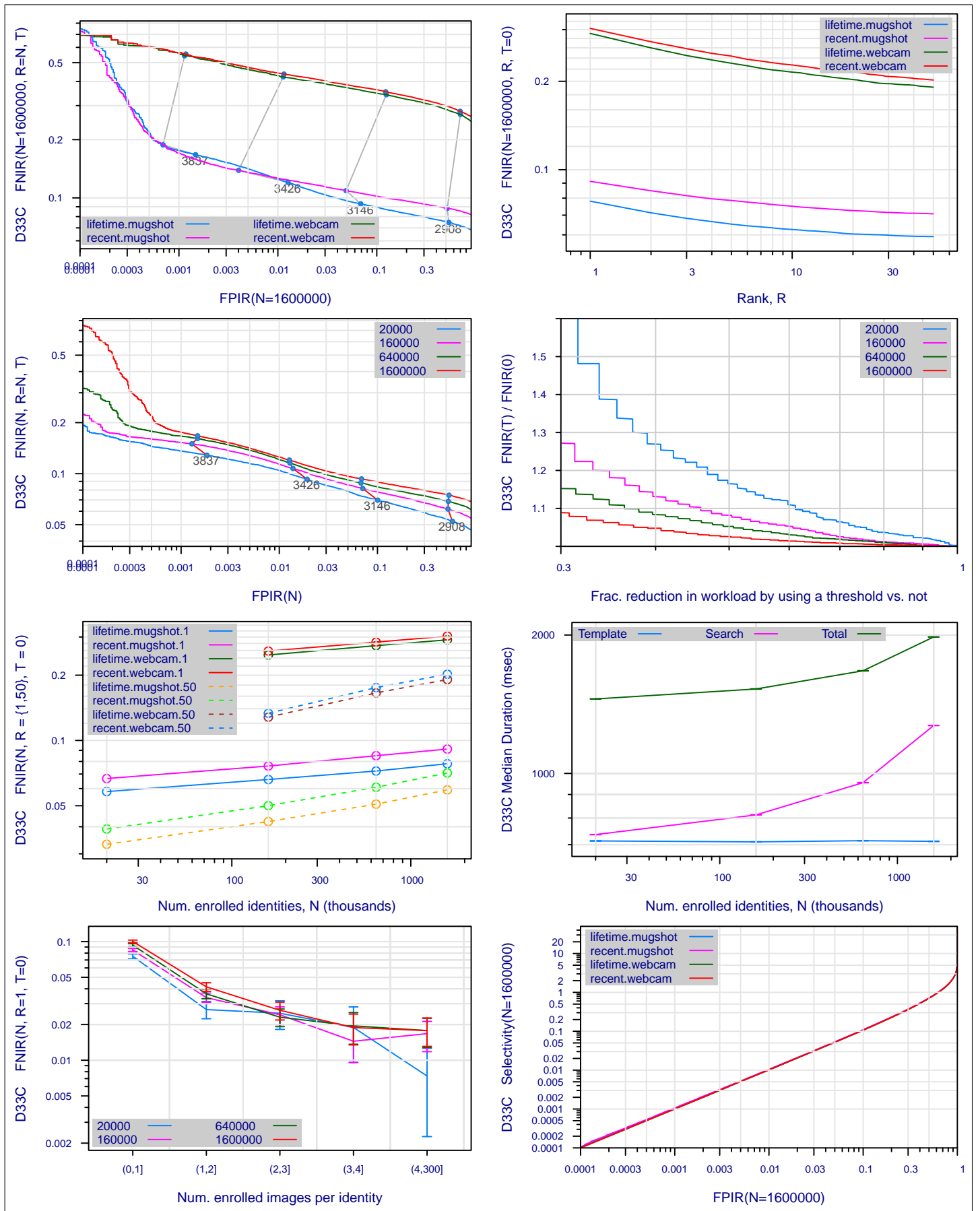


Figure 34: Collected performance reports for algorithm D33C. The figures are described at the beginning of this Appendix.

A = 3M/Cogent	B = Cognitec	C = Neurotechnology	D = Safran Morpho	E = NEC	F = Tsinghua U.	FNIR(N,R,T,L) "Miss rate"
G = Hisign	H = CAS-IA	I = CAS-ICT	J = Toshiba	L = Tsinghua U. II	M = HP	FPIR(N,T,L) "False alarm rate"
P = Zhuhai-Yisheng	Q = JunYu	S = Decatur	T = Ayonix			

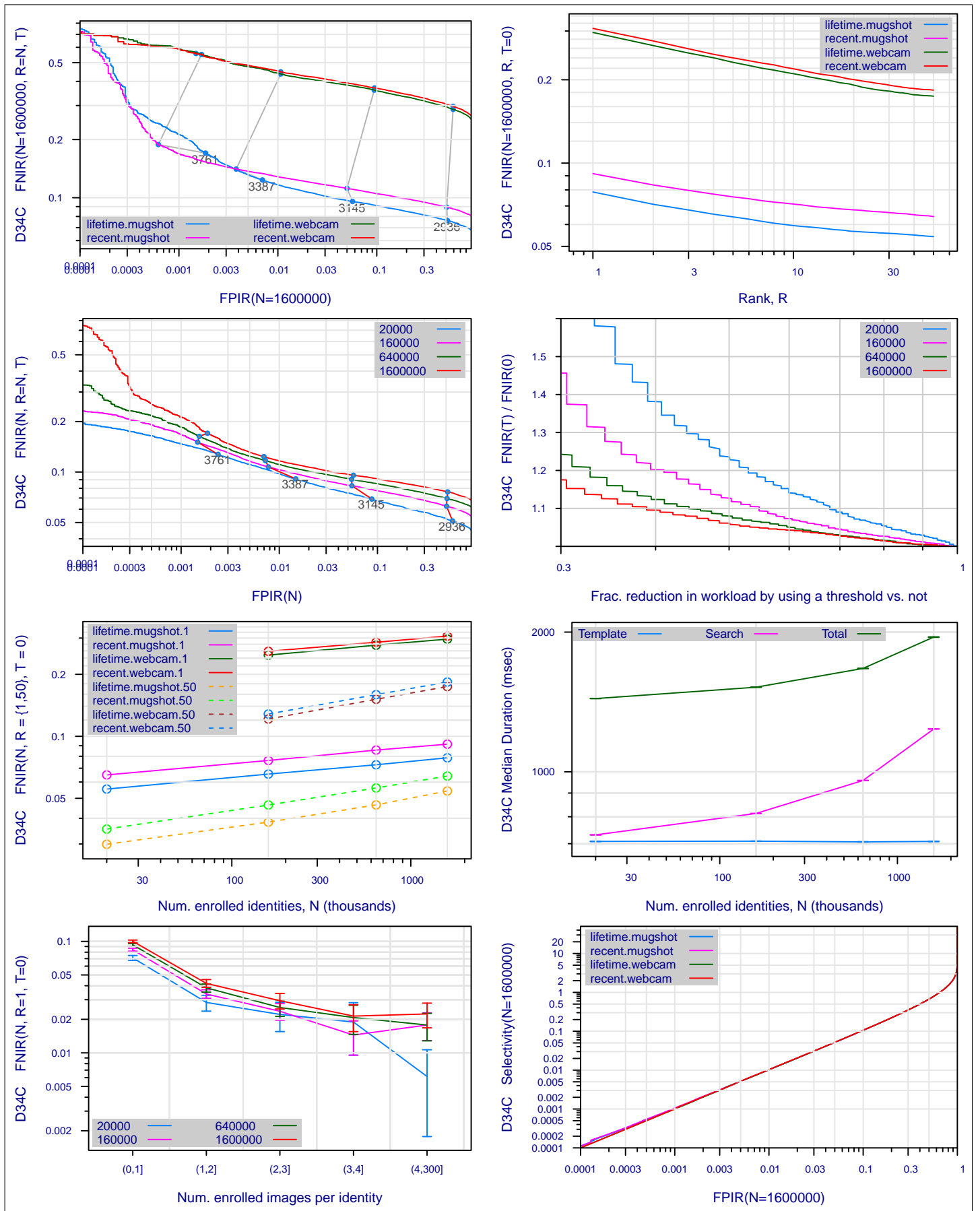


Figure 35: Collected performance reports for algorithm D34C. The figures are described at the beginning of this Appendix.

A = 3M/Cogent	B = Cognitec	C = Neurotechnology	D = Safran Morpho	E = NEC	F = Tsinghua U.	FNIR(N,R,T,L) "Miss rate"
G = Hisign	H = CAS-IA	I = CAS-ICT	J = Toshiba	L = Tsinghua U. II	M = HP	FPIR(N,T,L) "False alarm rate"
P = Zhuhai-Yisheng	Q = JunYu	S = Decatur	T = Ayonix			

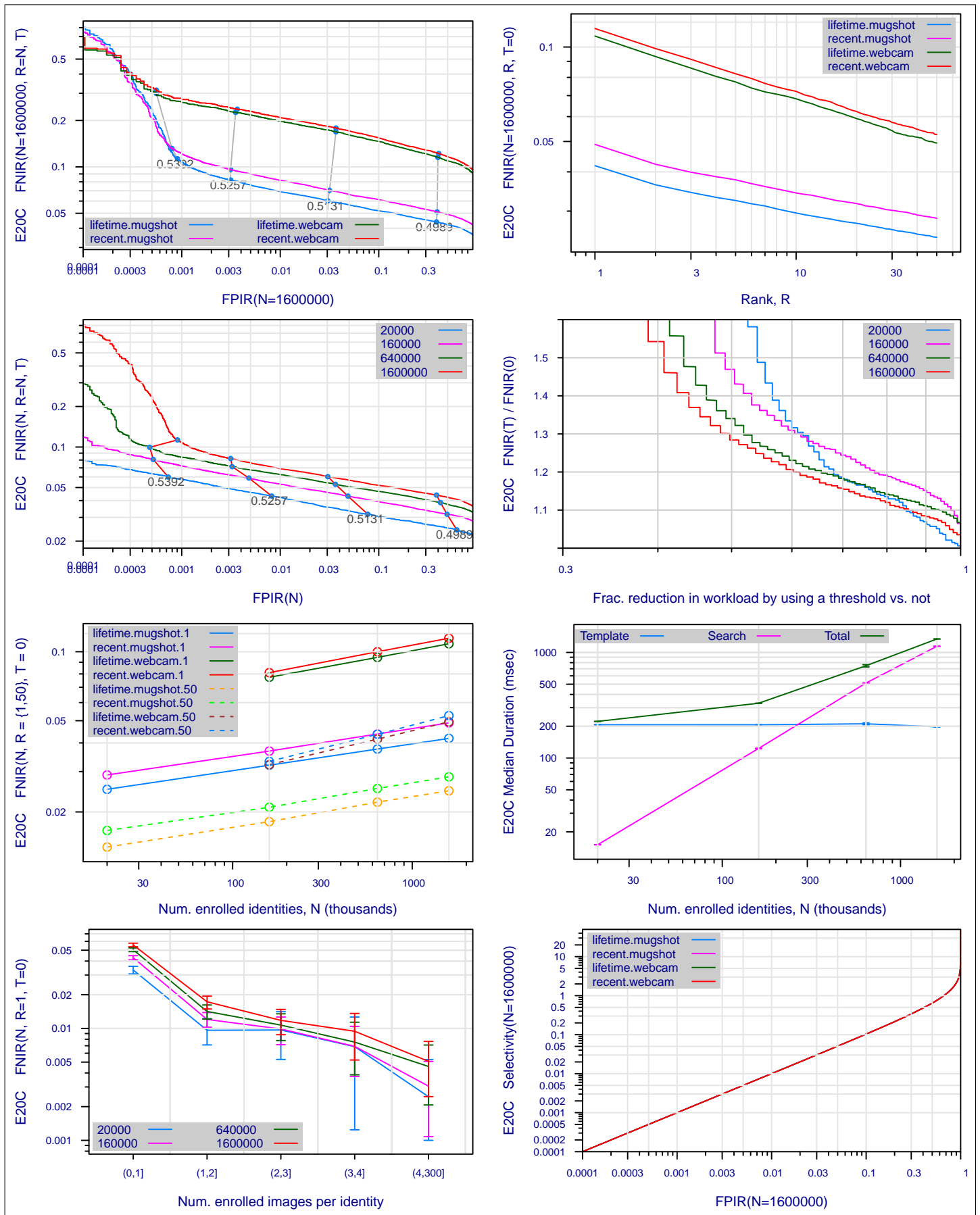


Figure 36: Collected performance reports for algorithm E20C. The figures are described at the beginning of this Appendix.

A = 3M/Cogent	B = Cognitec	C = Neurotechnology	D = Safran Morpho	E = NEC	F = Tsinghua U.	FNIR(N,R,T,L) "Miss rate"
G = Hisign	H = CAS-IA	I = CAS-ICT	J = Toshiba	L = Tsinghua U. II	M = HP	FPIR(N,T,L) "False alarm rate"
P = Zhuhai-Yisheng	Q = JunYu	S = Decatur	T = Ayonix			

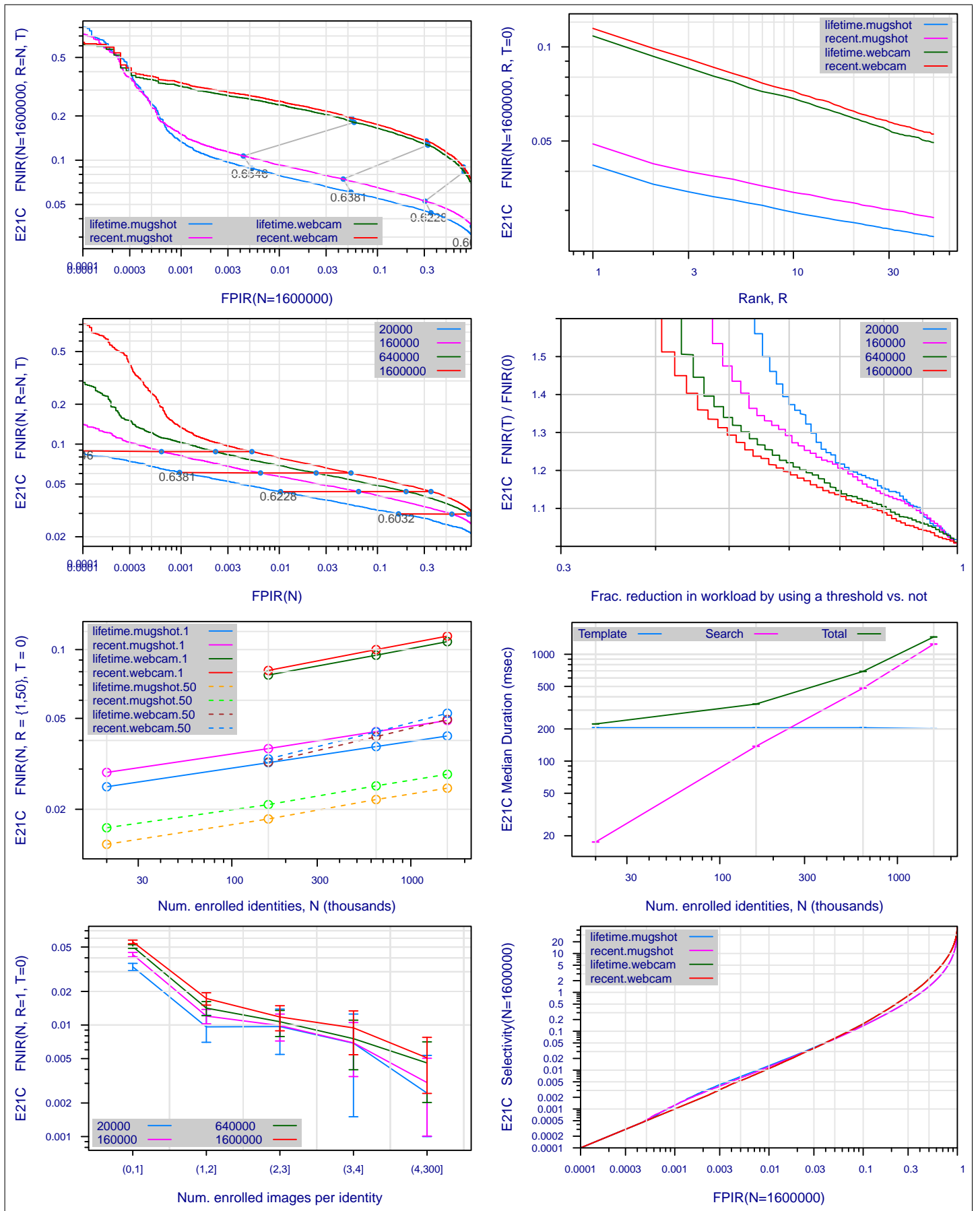


Figure 37: Collected performance reports for algorithm E21C. The figures are described at the beginning of this Appendix.

A = 3M/Cogent	B = Cognitec	C = Neurotechnology	D = Safran Morpho	E = NEC	F = Tsinghua U.	FNIR(N,R,T,L) "Miss rate"
G = Hisign	H = CAS-IA	I = CAS-ICT	J = Toshiba	L = Tsinghua U. II	M = HP	FPIR(N,T,L) "False alarm rate"
P = Zhuhai-Yisheng	Q = JunYu	S = Decatur	T = Ayonix			

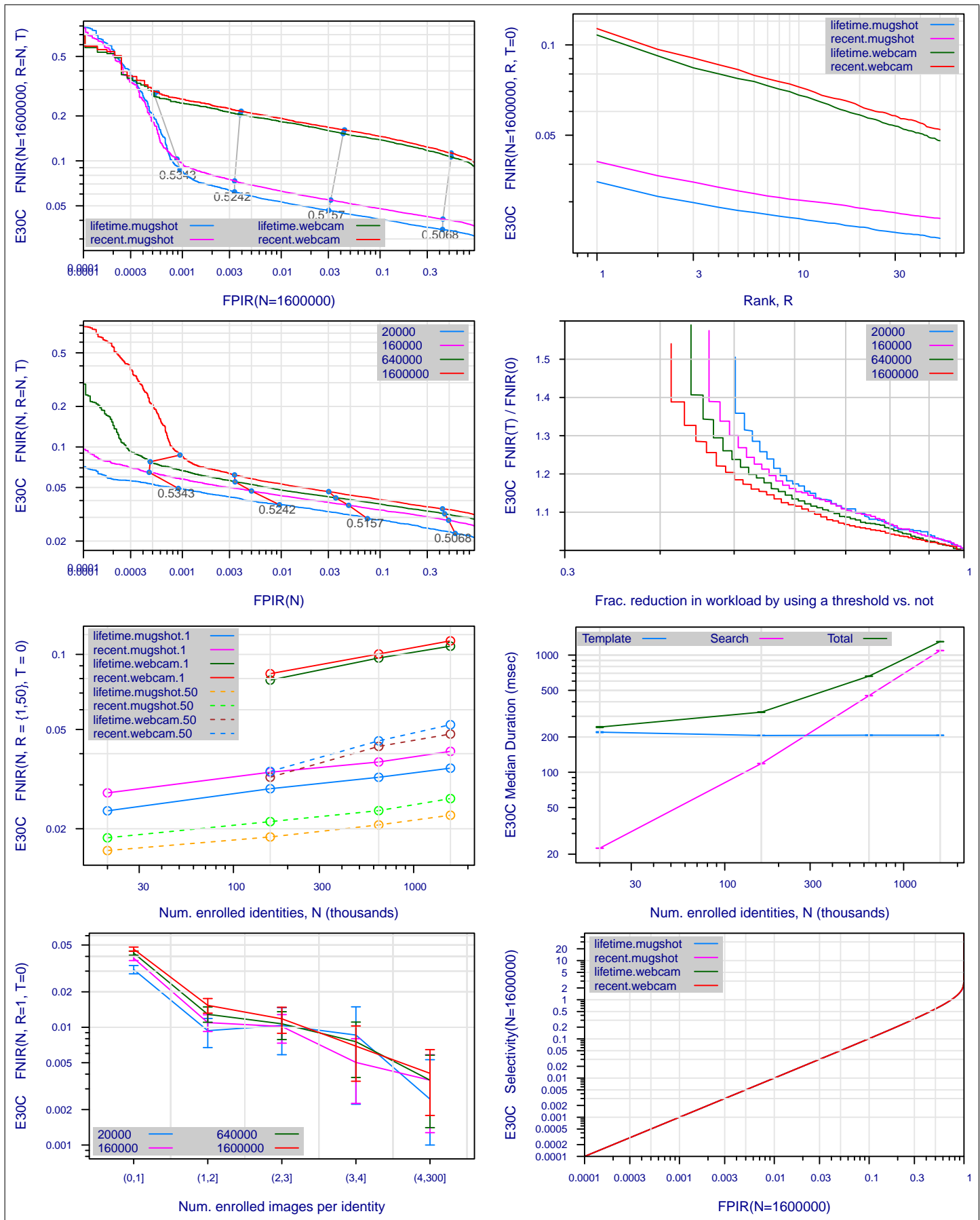


Figure 38: Collected performance reports for algorithm E30C. The figures are described at the beginning of this Appendix.

A = 3M/Cogent	B = Cognitec	C = Neurotechnology	D = Safran Morpho	E = NEC	F = Tsinghua U.	FNIR(N,R,T,L) "Miss rate"
G = Hisign	H = CAS-IA	I = CAS-ICT	J = Toshiba	L = Tsinghua U. II	M = HP	FPIR(N,T,L) "False alarm rate"
P = Zhuhai-Yisheng	Q = JunYu	S = Decatur	T = Ayonix			

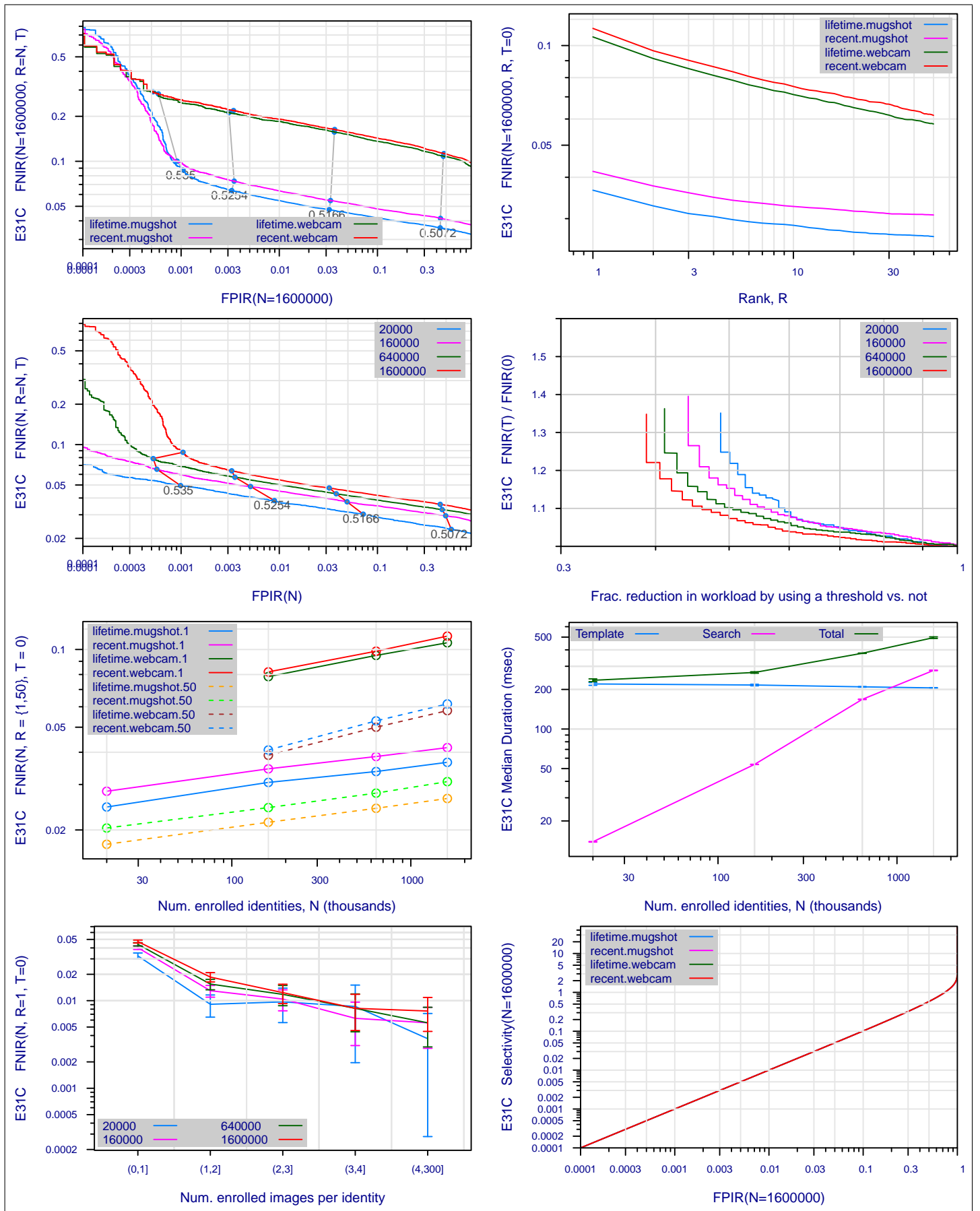


Figure 39: Collected performance reports for algorithm E31C. The figures are described at the beginning of this Appendix.

A = 3M/Cogent	B = Cognitec	C = Neurotechnology	D = Safran Morpho	E = NEC	F = Tsinghua U.	FNIR(N,R,T,L) "Miss rate"
G = Hisign	H = CAS-IA	I = CAS-ICT	J = Toshiba	L = Tsinghua U. II	M = HP	FPIR(N,T,L) "False alarm rate"
P = Zhuhai-Yisheng	Q = JunYu	S = Decatur	T = Ayonix			

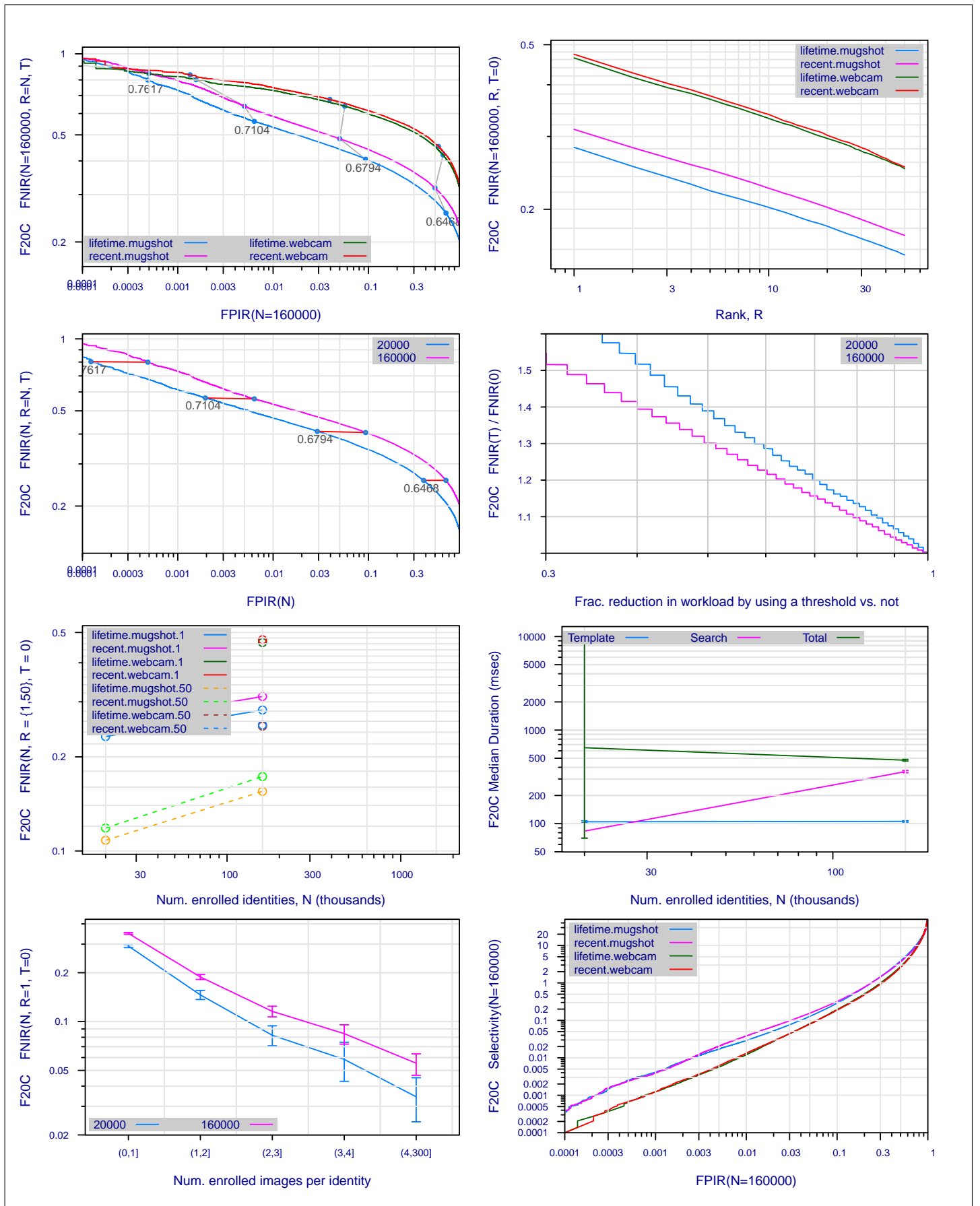


Figure 40: Collected performance reports for algorithm F20C. The figures are described at the beginning of this Appendix.

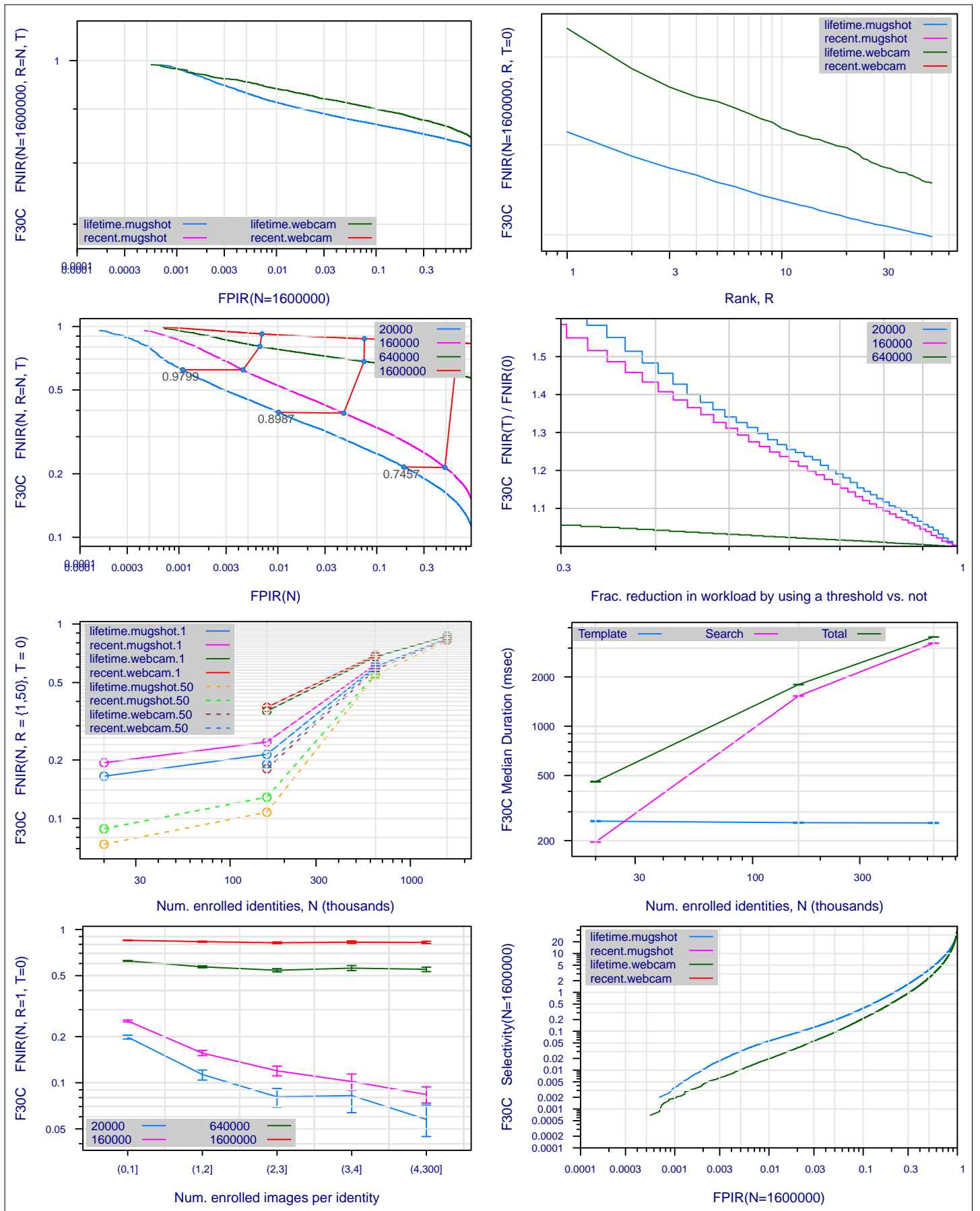


Figure 41: Collected performance reports for algorithm F30C. The figures are described at the beginning of this Appendix.

A = 3M/Cogent	B = Cognitec	C = Neurotechnology	D = Safran Morpho	E = NEC	F = Tsinghua U.	FNIR(N,R,T,L) "Miss rate"
G = Hisign	H = CAS-IA	I = CAS-ICT	J = Toshiba	L = Tsinghua U. II	M = HP	FPIR(N,T,L) "False alarm rate"
P = Zhuhai-Yisheng	Q = JunYu	S = Decatur	T = Ayonix			

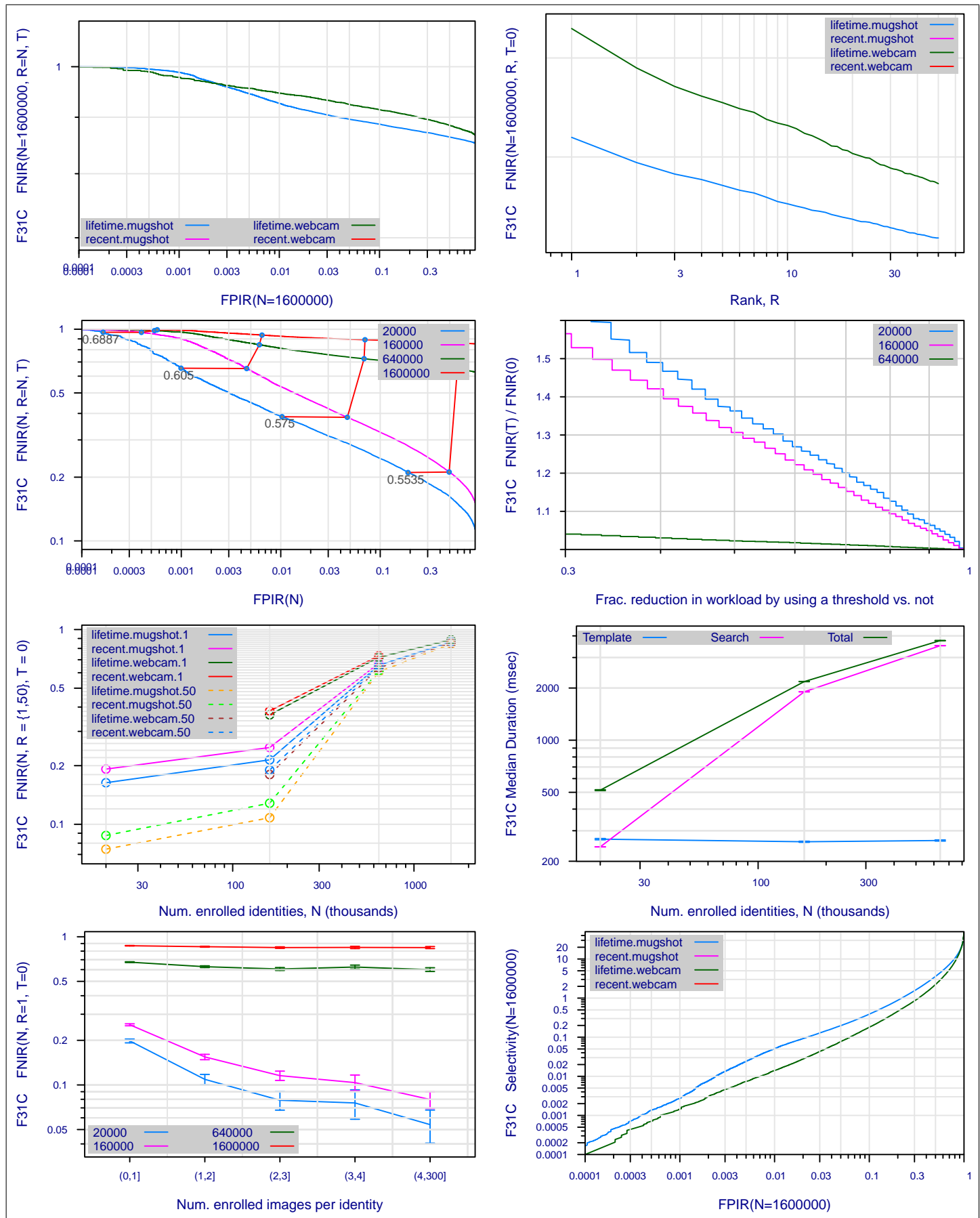


Figure 42: Collected performance reports for algorithm F31C. The figures are described at the beginning of this Appendix.

A = 3M/Cogent	B = Cognitec	C = Neurotechnology	D = Safran Morpho	E = NEC	F = Tsinghua U.	FNIR(N,R,T,L) "Miss rate"
G = Hisign	H = CAS-IA	I = CAS-ICT	J = Toshiba	L = Tsinghua U. II	M = HP	FPIR(N,T,L) "False alarm rate"
P = Zhuhai-Yisheng	Q = JunYu	S = Decatur	T = Ayonix			

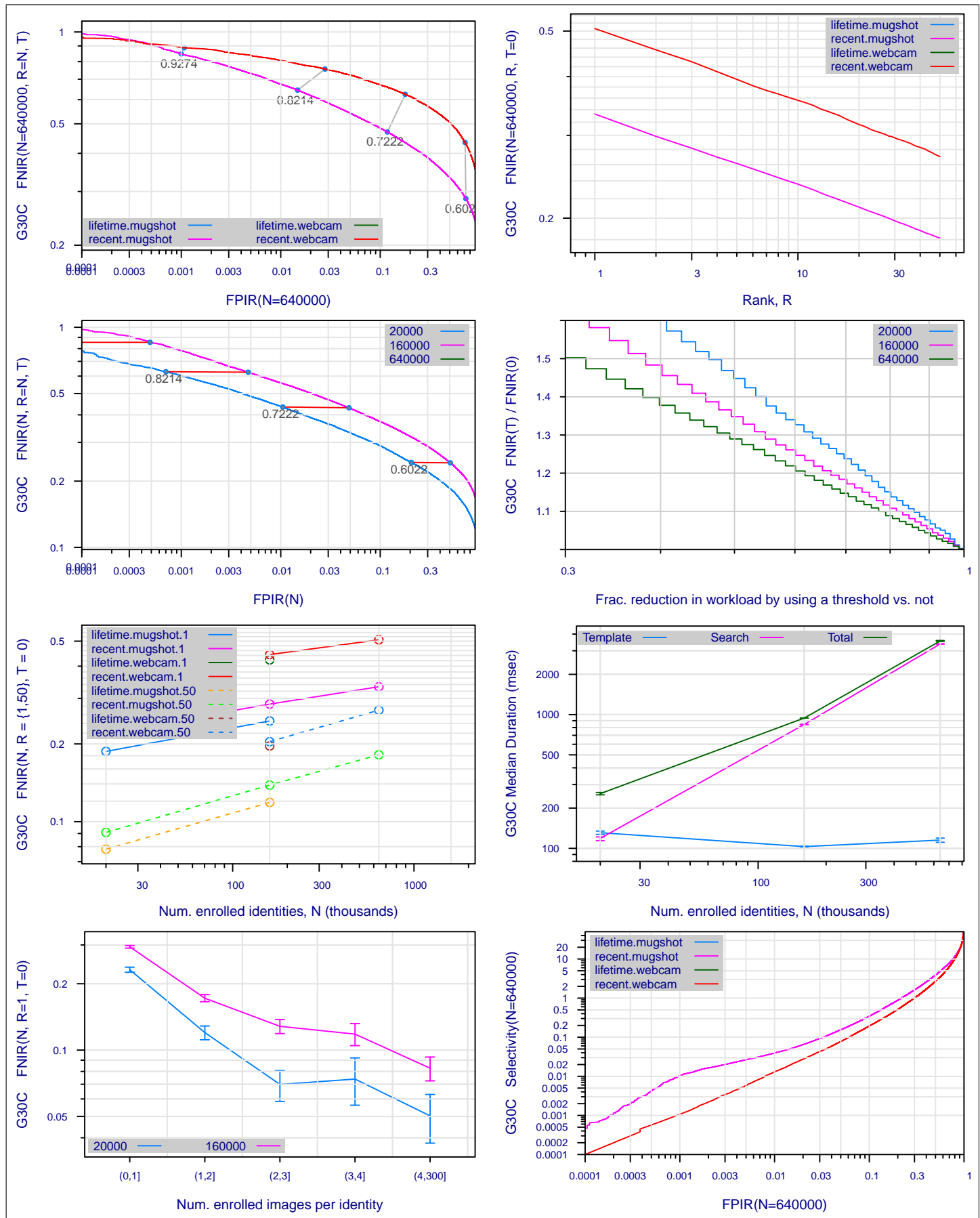


Figure 43: Collected performance reports for algorithm G30C. The figures are described at the beginning of this Appendix.

A = 3M/Cogent	B = Cognitec	C = Neurotechnology	D = Safran Morpho	E = NEC	F = Tsinghua U.	FNIR(N,R,T,L) "Miss rate"
G = Hisign	H = CAS-IA	I = CAS-ICT	J = Toshiba	L = Tsinghua U. II	M = HP	FPIR(N,T,L) "False alarm rate"
P = Zhuhai-Yisheng	Q = JunYu	S = Decatur	T = Ayonix			

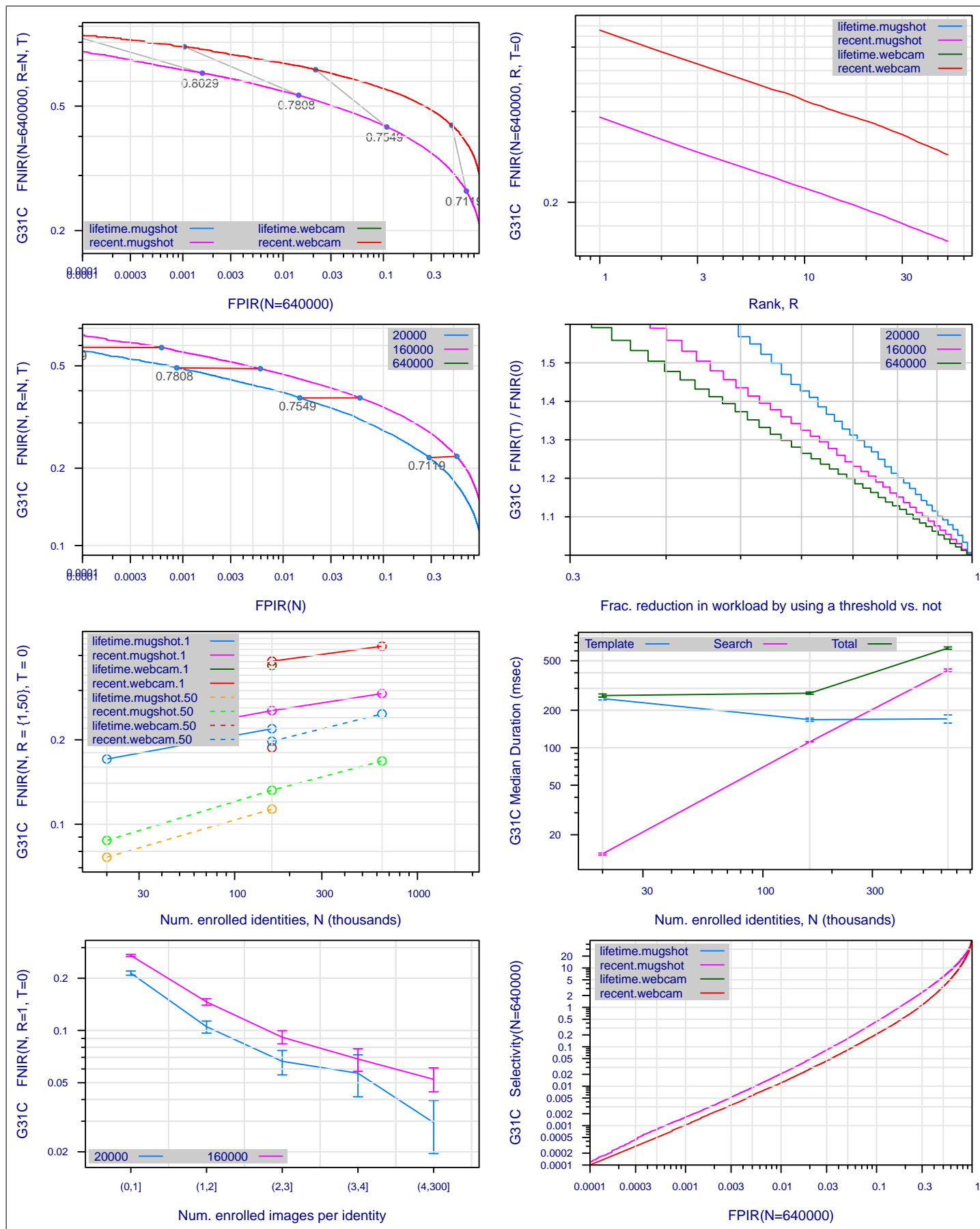


Figure 44: Collected performance reports for algorithm G31C. The figures are described at the beginning of this Appendix.

A = 3M/Cogent	B = Cognitec	C = Neurotechnology	D = Safran Morpho	E = NEC	F = Tsinghua U.	FNIR(N,R,T,L) "Miss rate"
G = Hisign	H = CAS-IA	I = CAS-ICT	J = Toshiba	L = Tsinghua U. II	M = HP	FPIR(N,T,L) "False alarm rate"
P = Zhuhai-Yisheng	Q = JunYu	S = Decatur	T = Ayonix			

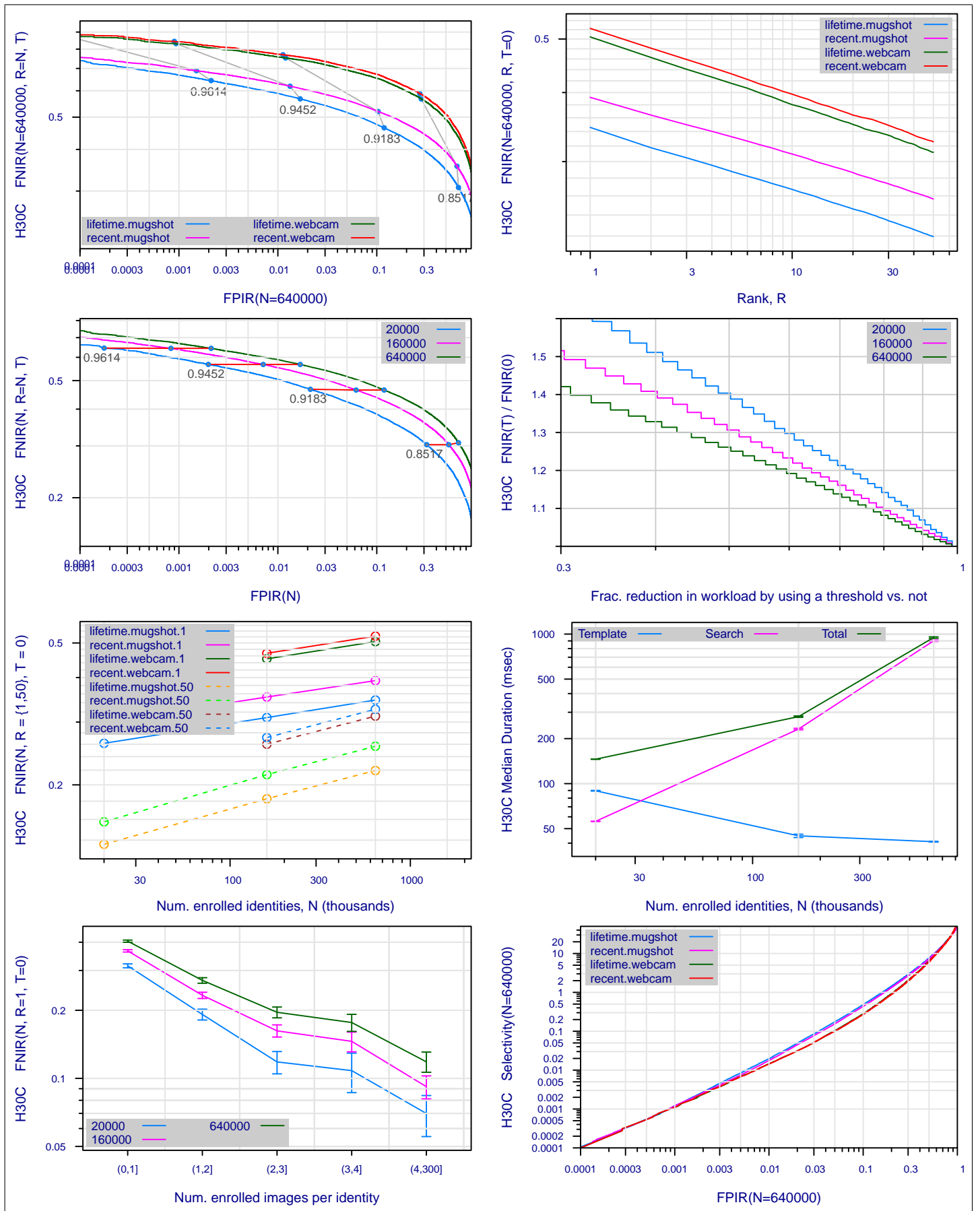


Figure 45: Collected performance reports for algorithm H30C. The figures are described at the beginning of this Appendix.

A = 3M/Cogent	B = Cognitec	C = Neurotechnology	D = Safran Morpho	E = NEC	F = Tsinghua U.	FNIR(N,R,T,L) "Miss rate"
G = Hisign	H = CAS-IA	I = CAS-ICT	J = Toshiba	L = Tsinghua U. II	M = HP	FPIR(N,T,L) "False alarm rate"
P = Zhuhai-Yisheng	Q = JunYu	S = Decatur	T = Ayonix			

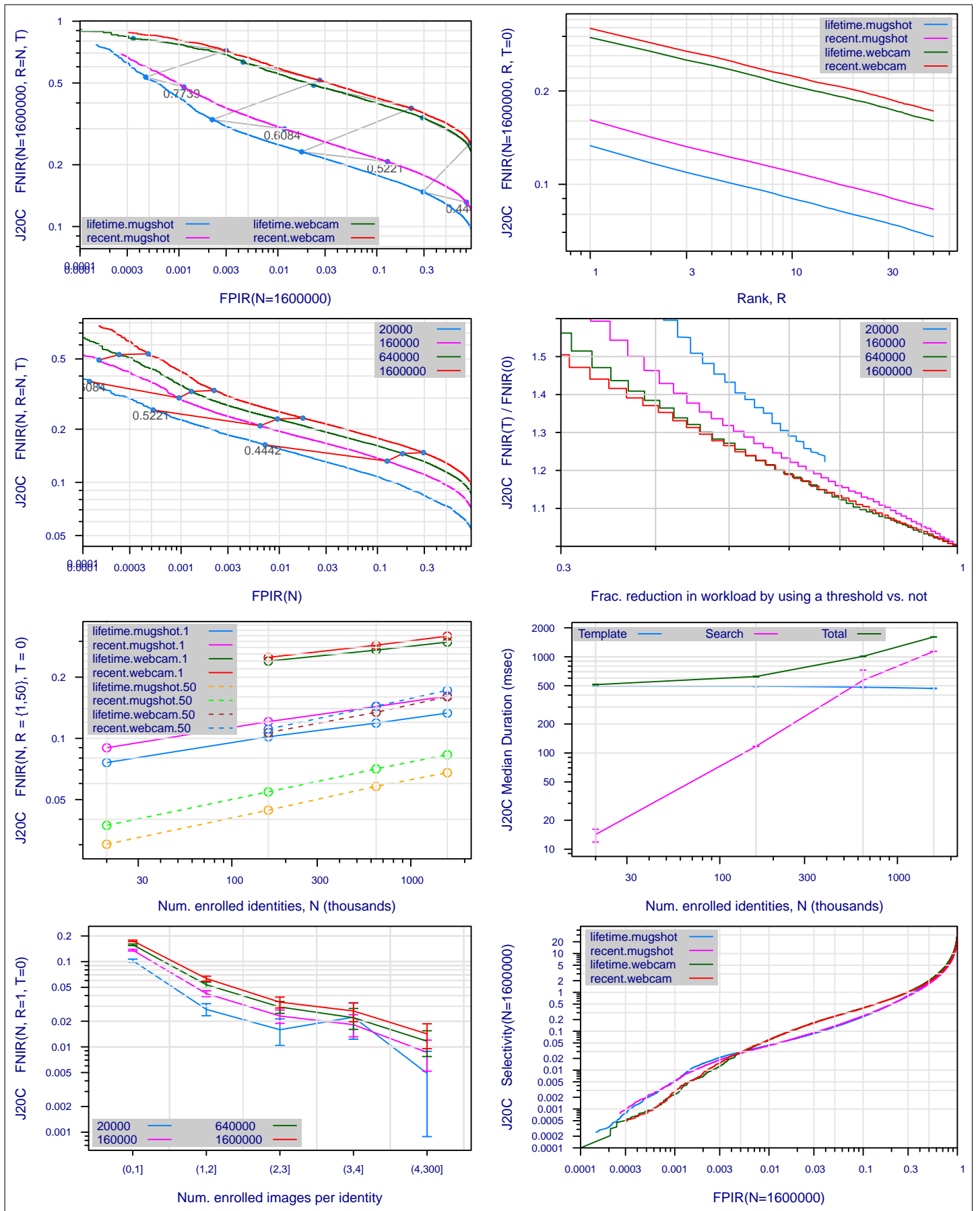


Figure 46: Collected performance reports for algorithm J20C. The figures are described at the beginning of this Appendix.

A = 3M/Cogent	B = Cognitec	C = Neurotechnology	D = Safran Morpho	E = NEC	F = Tsinghua U.	FNIR(N,R,T,L) "Miss rate"
G = Hisign	H = CAS-IA	I = CAS-ICT	J = Toshiba	L = Tsinghua U. II	M = HP	FPIR(N,T,L) "False alarm rate"
P = Zhuhai-Yisheng	Q = JunYu	S = Decatur	T = Ayonix			

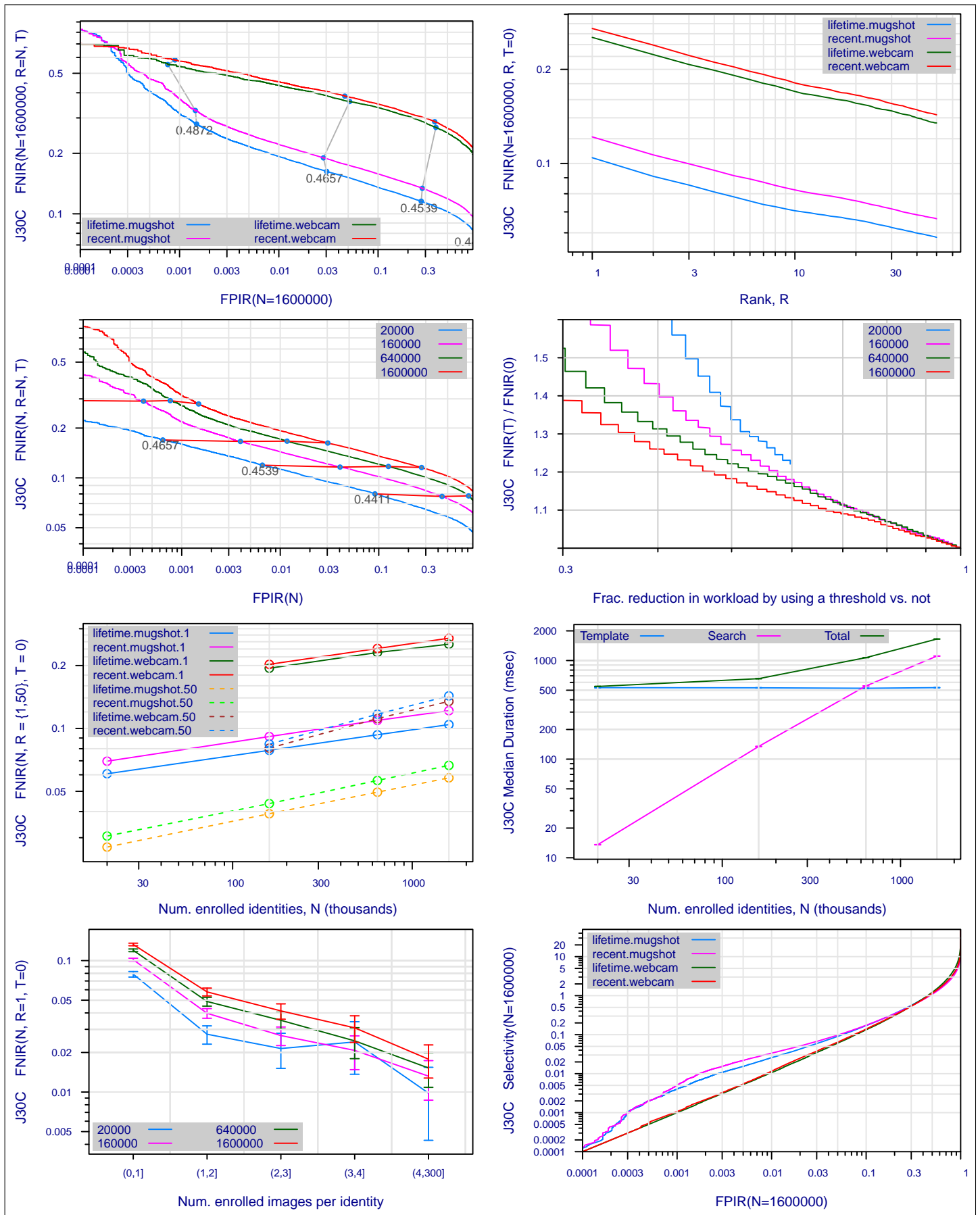


Figure 47: Collected performance reports for algorithm J30C. The figures are described at the beginning of this Appendix.

A = 3M/Cogent	B = Cognitec	C = Neurotechnology	D = Safran Morpho	E = NEC	F = Tsinghua U.	FNIR(N,R,T,L) "Miss rate"
G = Hisign	H = CAS-IA	I = CAS-ICT	J = Toshiba	L = Tsinghua U. II	M = HP	FPIR(N,T,L) "False alarm rate"
P = Zhuhai-Yisheng	Q = JunYu	S = Decatur	T = Ayonix			

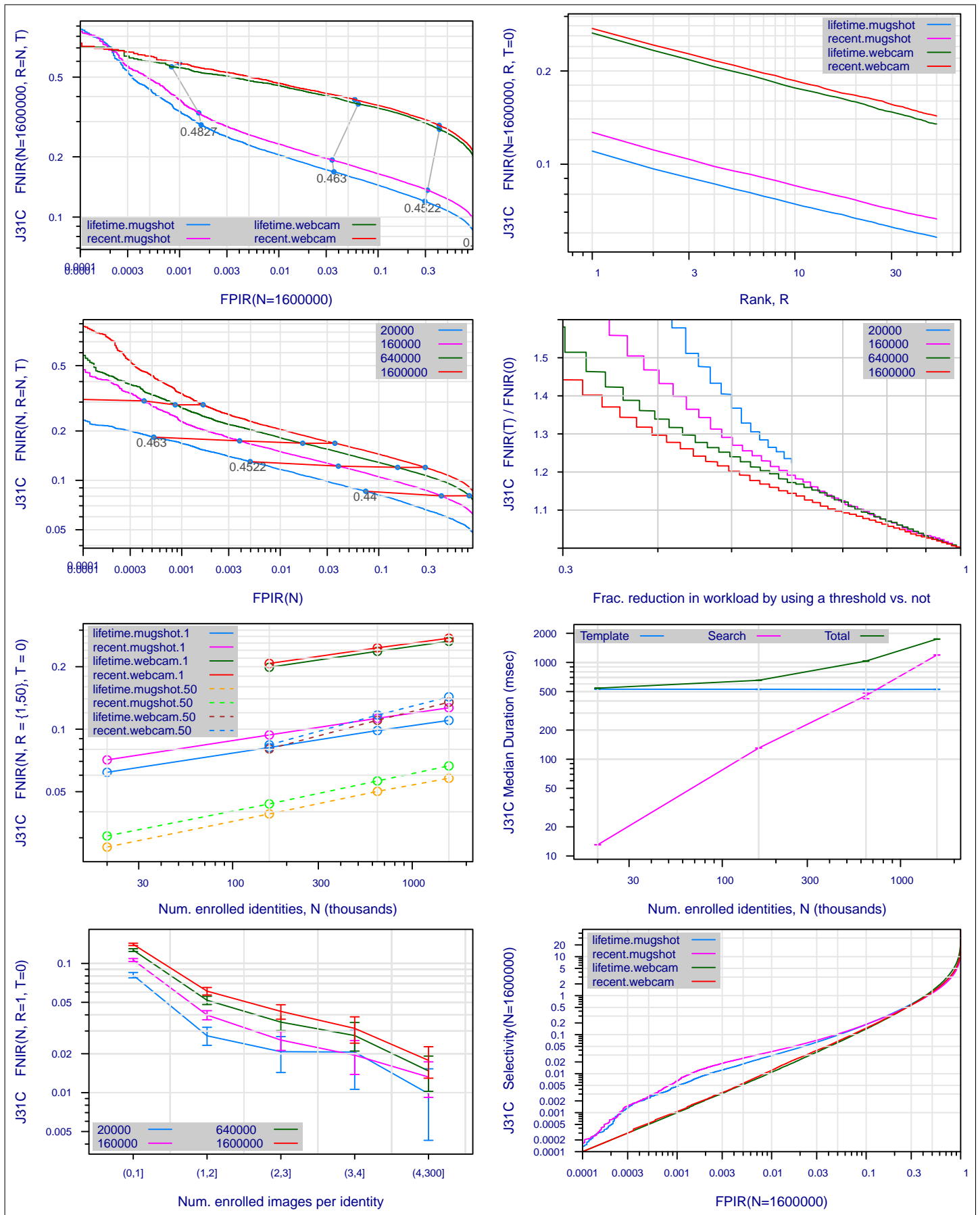


Figure 48: Collected performance reports for algorithm J31C. The figures are described at the beginning of this Appendix.

A = 3M/Cogent	B = Cognitec	C = Neurotechnology	D = Safran Morpho	E = NEC	F = Tsinghua U.	FNIR(N,R,T,L) "Miss rate"
G = Hisign	H = CAS-IA	I = CAS-ICT	J = Toshiba	L = Tsinghua U. II	M = HP	FPIR(N,T,L) "False alarm rate"
P = Zhuhai-Yisheng	Q = JunYu	S = Decatur	T = Ayonix			

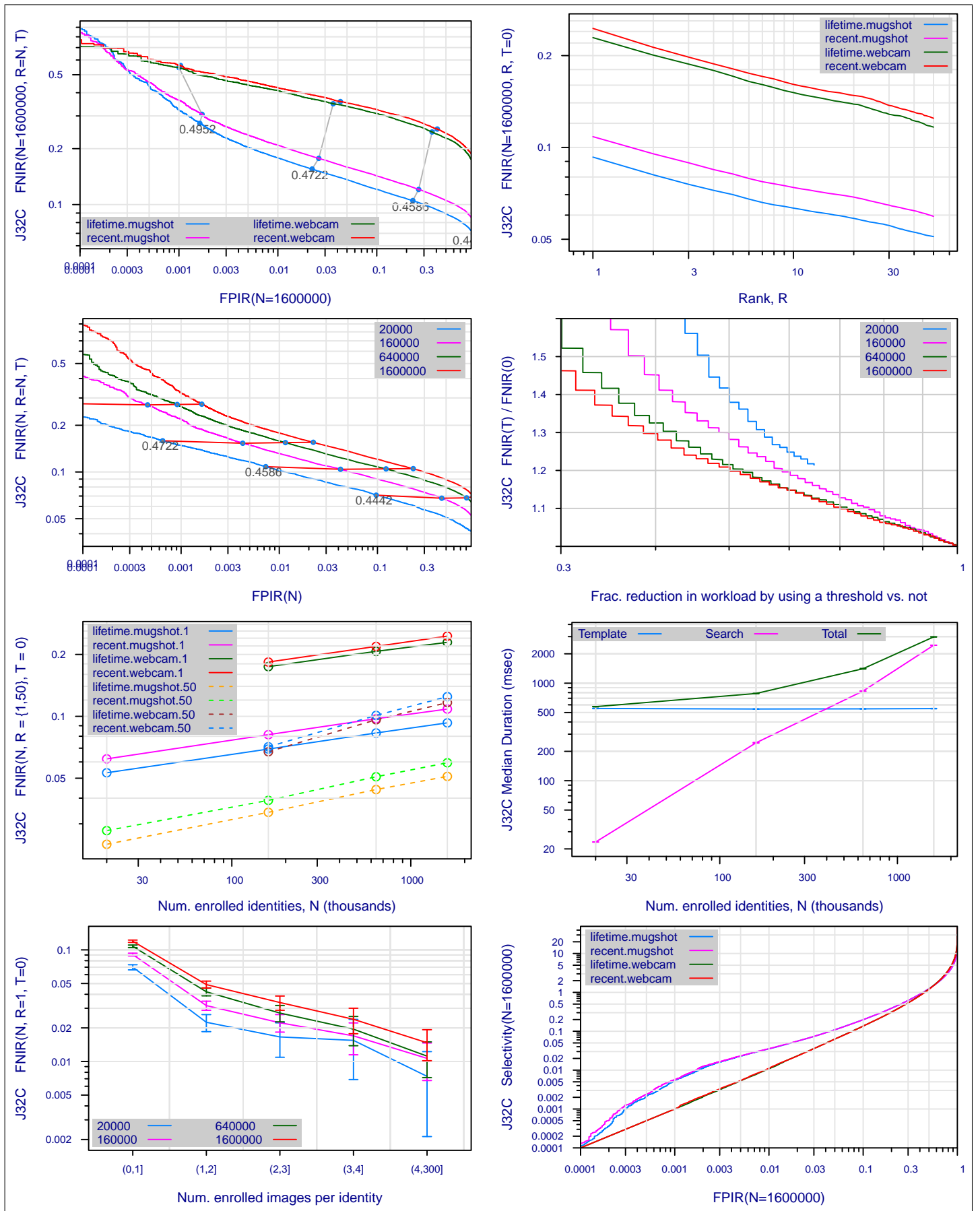


Figure 49: Collected performance reports for algorithm J32C. The figures are described at the beginning of this Appendix.

A = 3M/Cogent	B = Cognitec	C = Neurotechnology	D = Safran Morpho	E = NEC	F = Tsinghua U.	FNIR(N,R,T,L) "Miss rate"
G = Hisign	H = CAS-IA	I = CAS-ICT	J = Toshiba	L = Tsinghua U. II	M = HP	FPIR(N,T,L) "False alarm rate"
P = Zhuhai-Yisheng	Q = JunYu	S = Decatur	T = Ayonix			

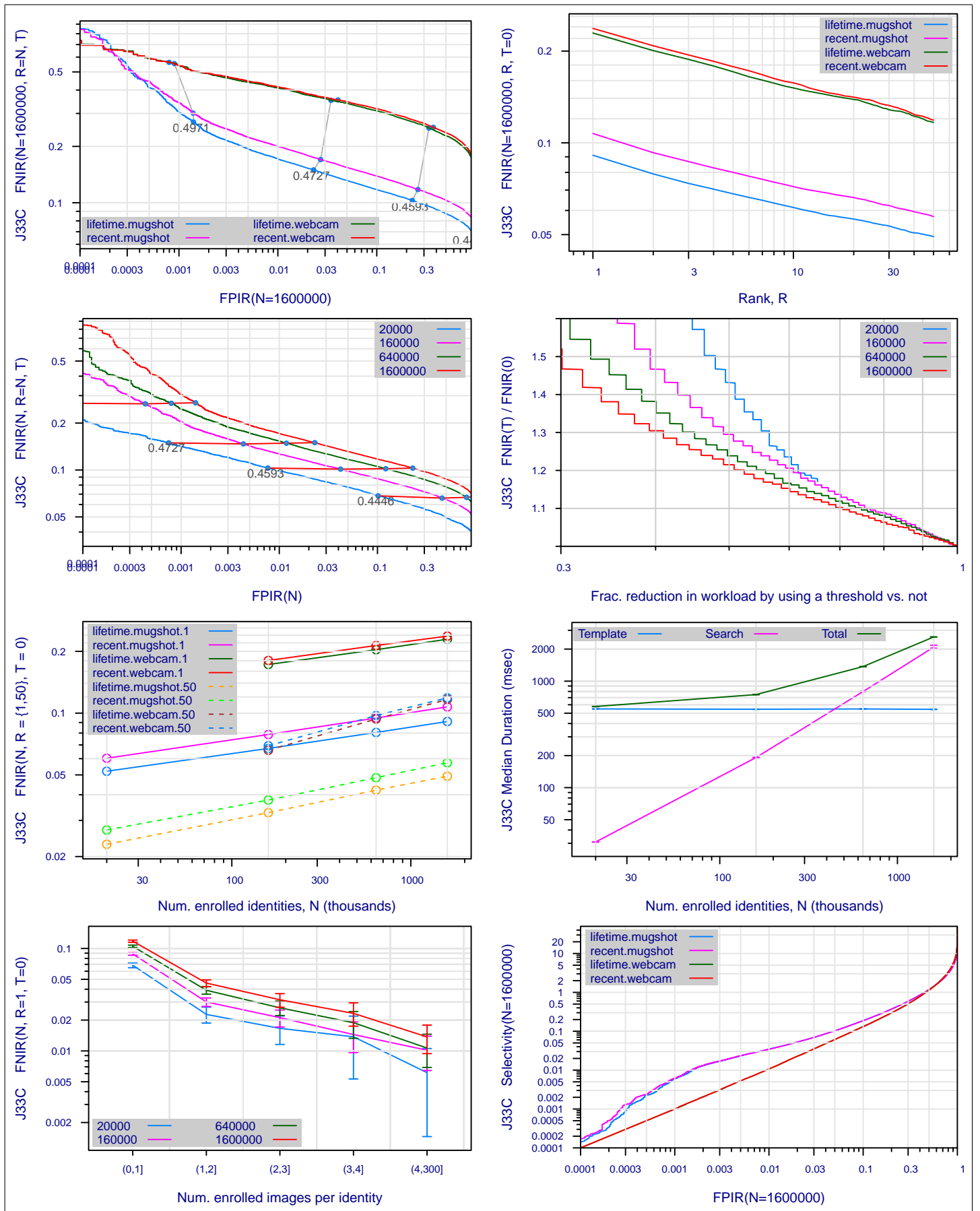


Figure 50: Collected performance reports for algorithm J33C. The figures are described at the beginning of this Appendix.

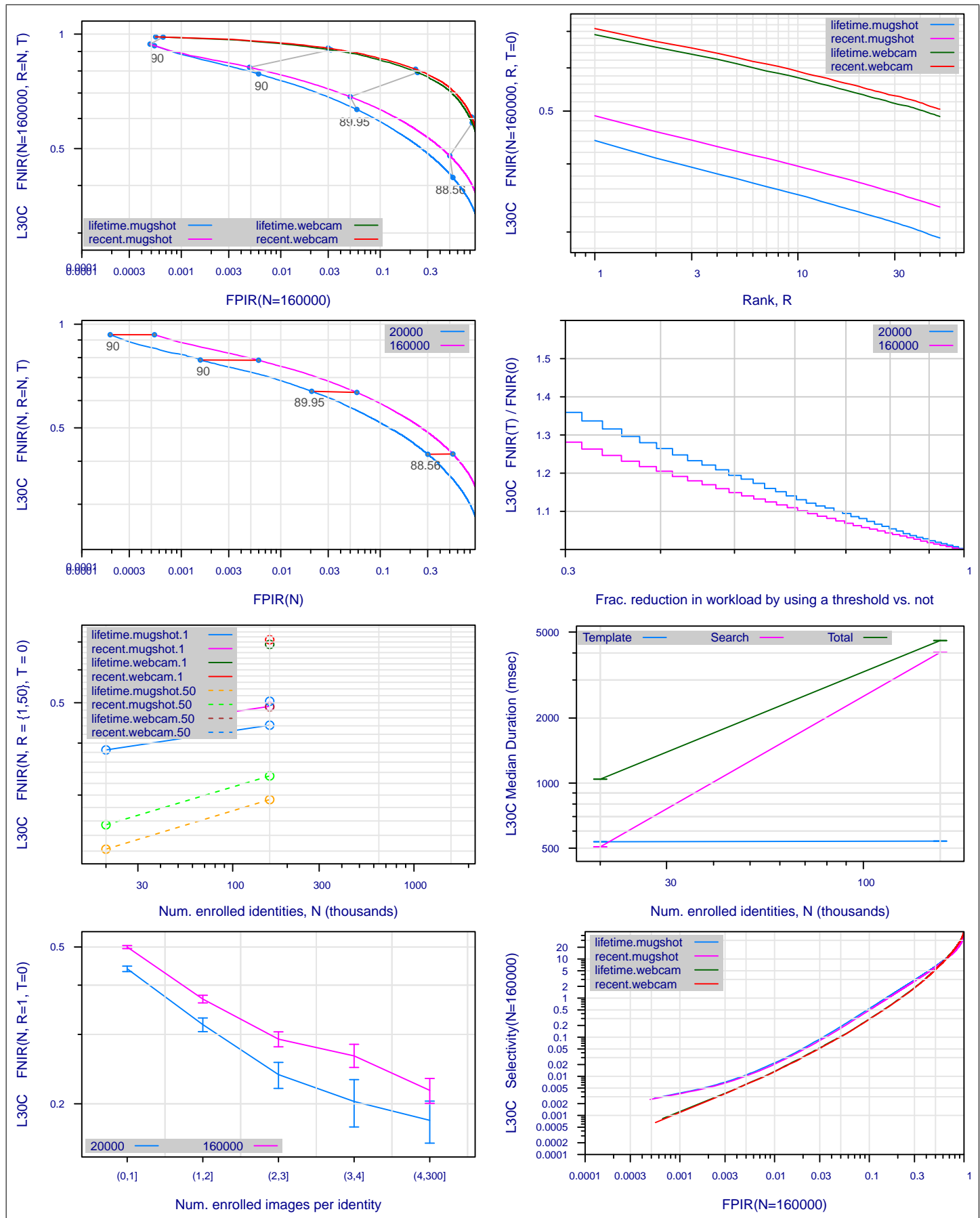


Figure 51: Collected performance reports for algorithm L30C. The figures are described at the beginning of this Appendix.

A = 3M/Cogent	B = Cognitec	C = Neurotechnology	D = Safran Morpho	E = NEC	F = Tsinghua U.	FNIR(N,R,T,L) "Miss rate"
G = Hisign	H = CAS-IA	I = CAS-ICT	J = Toshiba	L = Tsinghua U. II	M = HP	FPIR(N,T,L) "False alarm rate"
P = Zhuhai-Yisheng	Q = JunYu	S = Decatur	T = Ayonix			

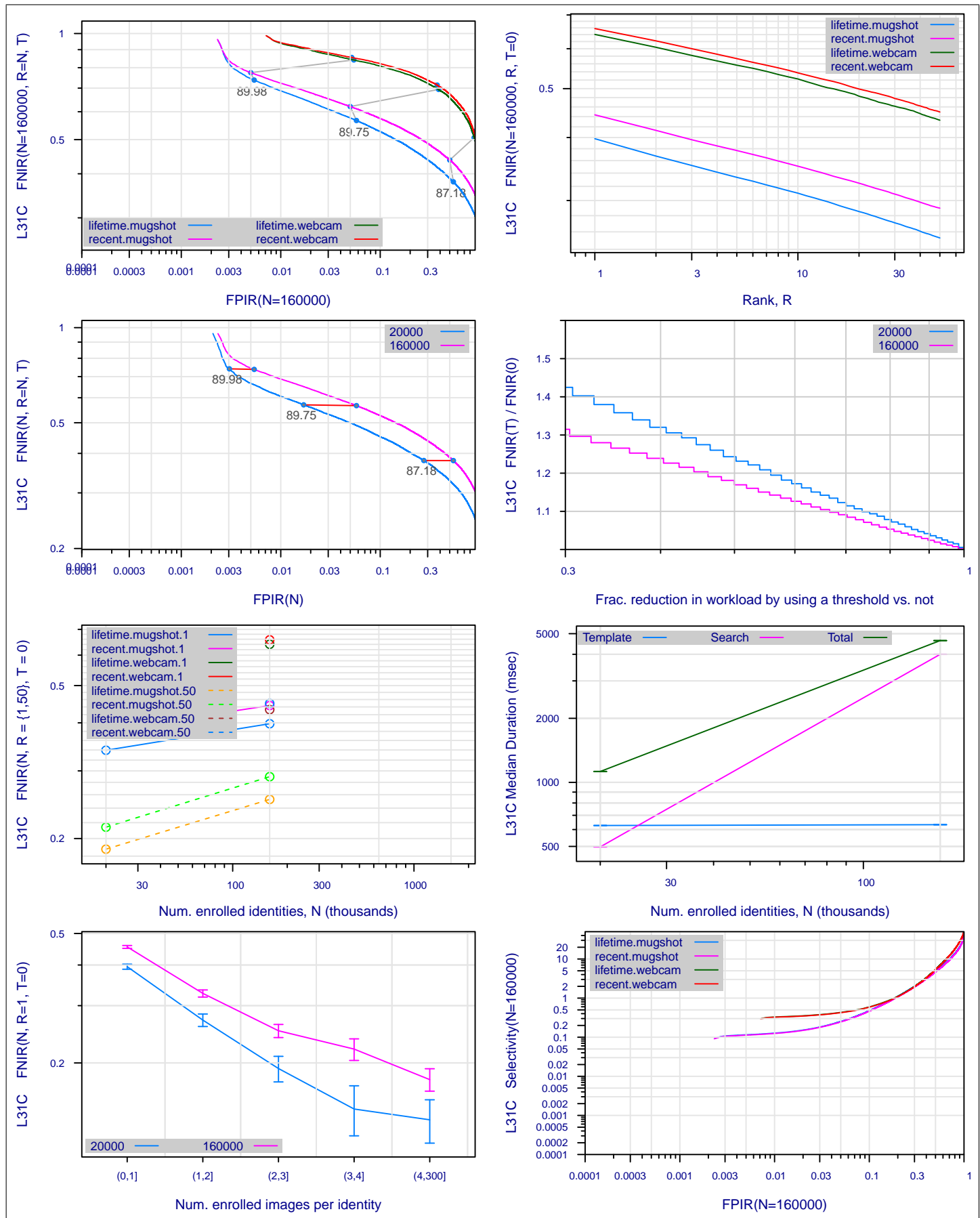


Figure 52: Collected performance reports for algorithm L31C. The figures are described at the beginning of this Appendix.

A = 3M/Cogent	B = Cognitec	C = Neurotechnology	D = Safran Morpho	E = NEC	F = Tsinghua U.	FNIR(N,R,T,L) "Miss rate"
G = Hisign	H = CAS-IA	I = CAS-ICT	J = Toshiba	L = Tsinghua U. II	M = HP	FPIR(N,T,L) "False alarm rate"
P = Zhuhai-Yisheng	Q = JunYu	S = Decatur	T = Ayonix			

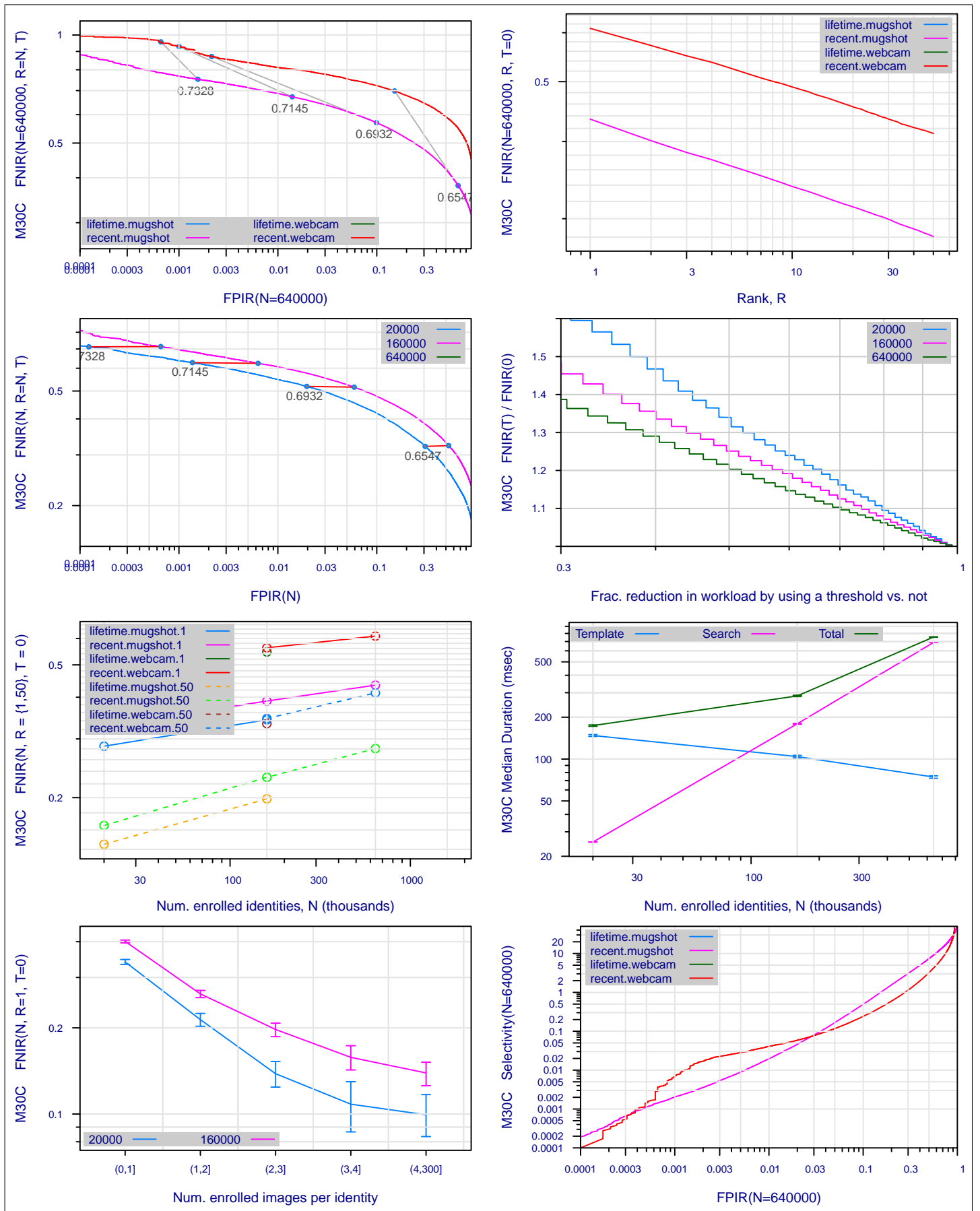


Figure 53: Collected performance reports for algorithm M30C. The figures are described at the beginning of this Appendix.

A = 3M/Cogent	B = Cognitec	C = Neurotechnology	D = Safran Morpho	E = NEC	F = Tsinghua U.	FNIR(N,R,T,L) "Miss rate"
G = Hisign	H = CAS-IA	I = CAS-ICT	J = Toshiba	L = Tsinghua U. II	M = HP	FPIR(N,T,L) "False alarm rate"
P = Zhuhai-Yisheng	Q = JunYu	S = Decatur	T = Ayonix			

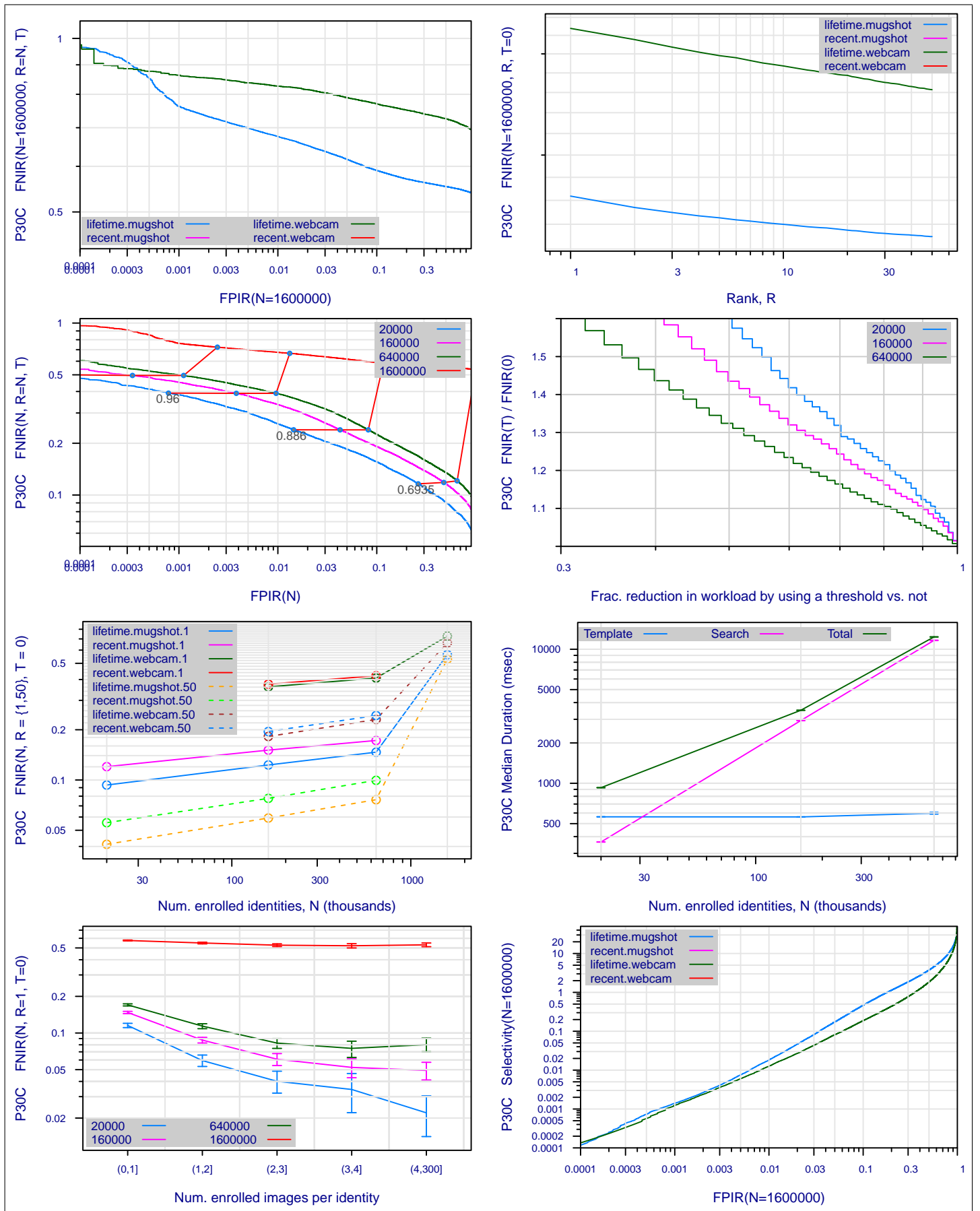


Figure 54: Collected performance reports for algorithm P30C. The figures are described at the beginning of this Appendix.

A = 3M/Cogent	B = Cognitec	C = Neurotechnology	D = Safran Morpho	E = NEC	F = Tsinghua U.	FNIR(N,R,T,L) "Miss rate"
G = Hisign	H = CAS-IA	I = CAS-ICT	J = Toshiba	L = Tsinghua U. II	M = HP	FPIR(N,T,L) "False alarm rate"
P = Zhuhai-Yisheng	Q = JunYu	S = Decatur	T = Ayonix			

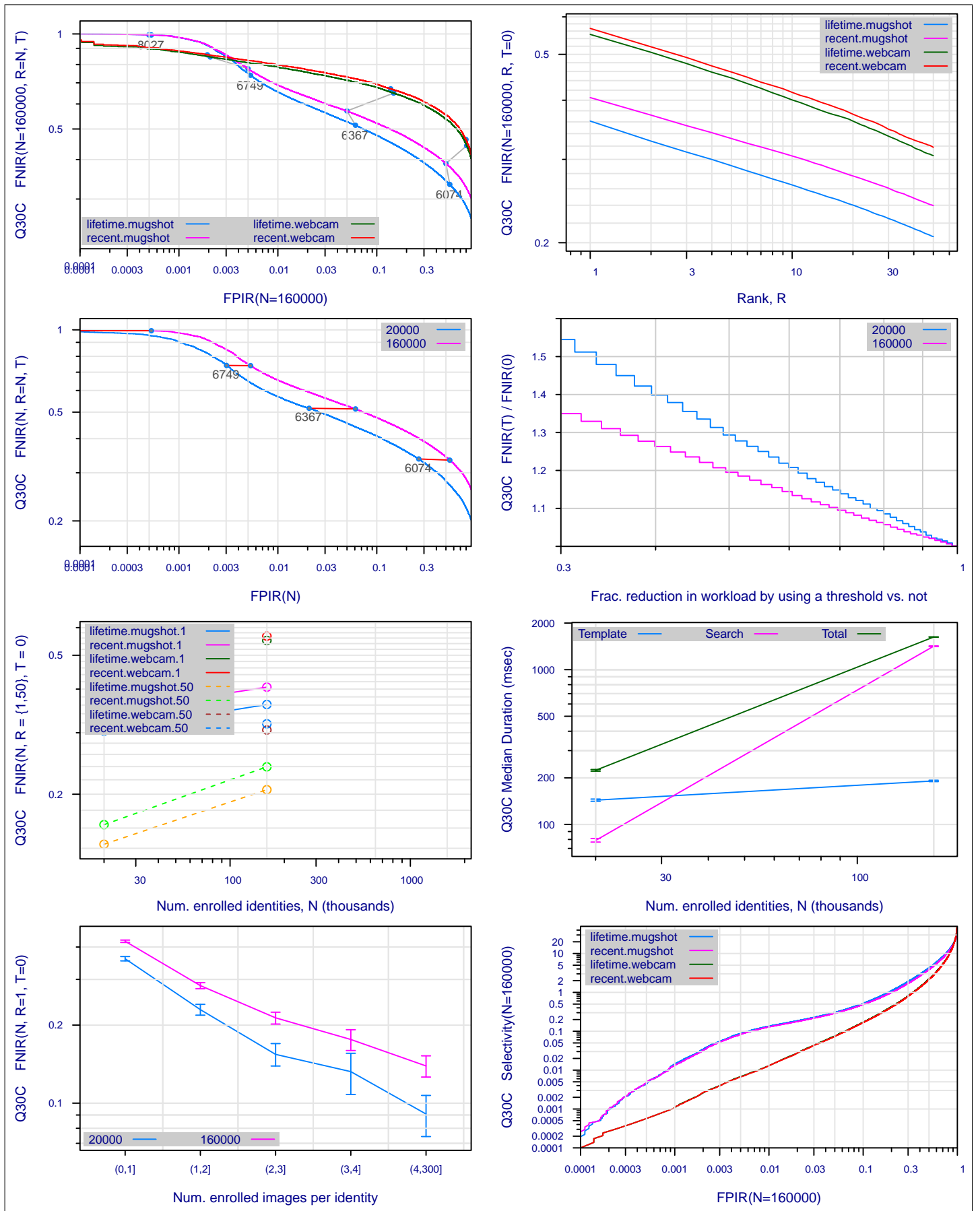


Figure 55: Collected performance reports for algorithm Q30C. The figures are described at the beginning of this Appendix.

A = 3M/Cogent	B = Cognitec	C = Neurotechnology	D = Safran Morpho	E = NEC	F = Tsinghua U.	FNIR(N,R,T,L) "Miss rate"
G = Hisign	H = CAS-IA	I = CAS-ICT	J = Toshiba	L = Tsinghua U. II	M = HP	FPIR(N,T,L) "False alarm rate"
P = Zhuhai-Yisheng	Q = JunYu	S = Decatur	T = Ayonix			

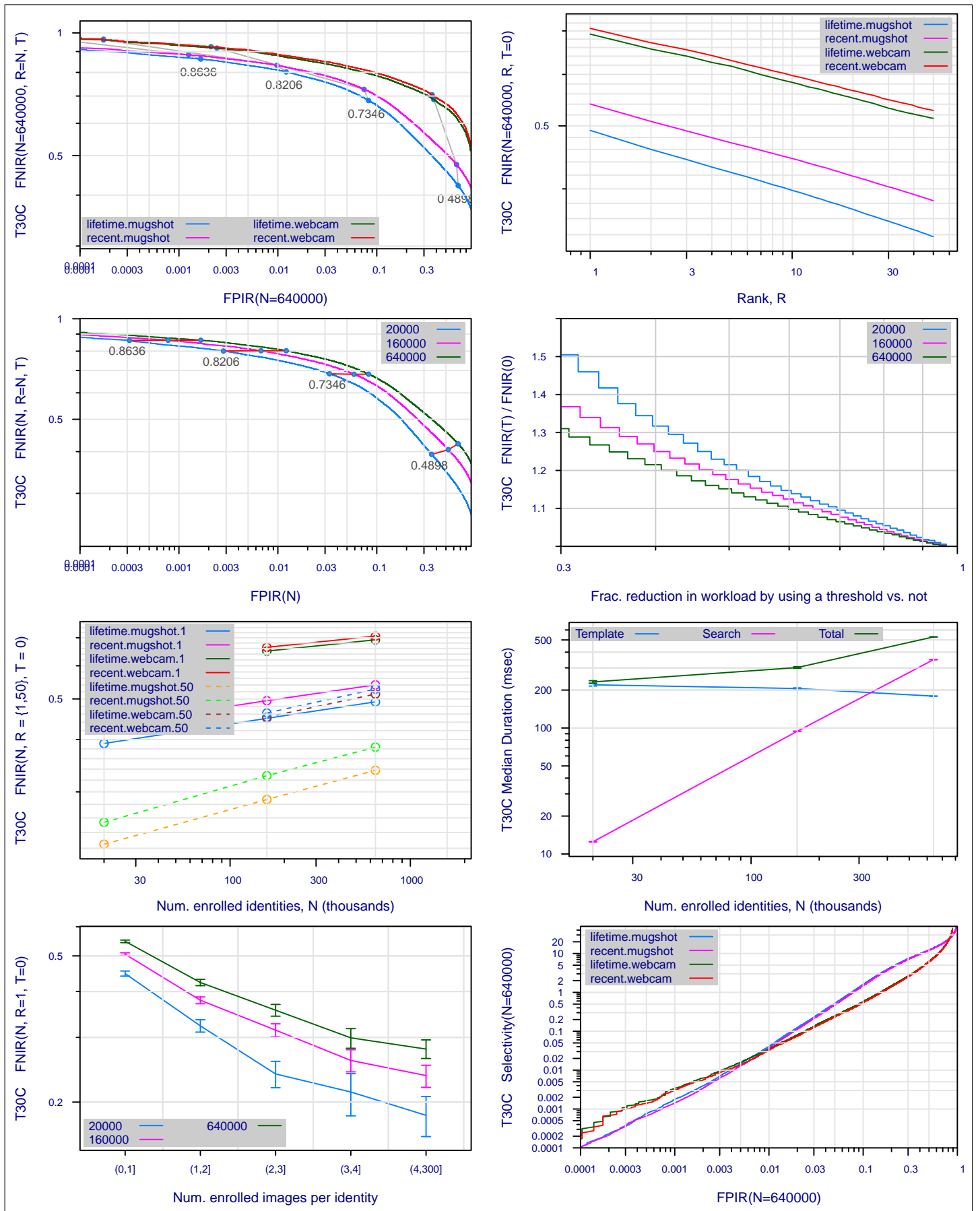


Figure 56: Collected performance reports for algorithm T30C. The figures are described at the beginning of this Appendix.

A = 3M/Cogent	B = Cognitec	C = Neurotechnology	D = Safran Morpho	E = NEC	F = Tsinghua U.	FNIR(N,R,T,L) "Miss rate"
G = Hisign	H = CAS-IA	I = CAS-ICT	J = Toshiba	L = Tsinghua U. II	M = HP	FPIR(N,T,L) "False alarm rate"
P = Zhuhai-Yisheng	Q = JunYu	S = Decatur	T = Ayonix			

A Algorithm accuracy by age group

This section details individual algorithm performance by age group as discussed in [section 5.5](#).

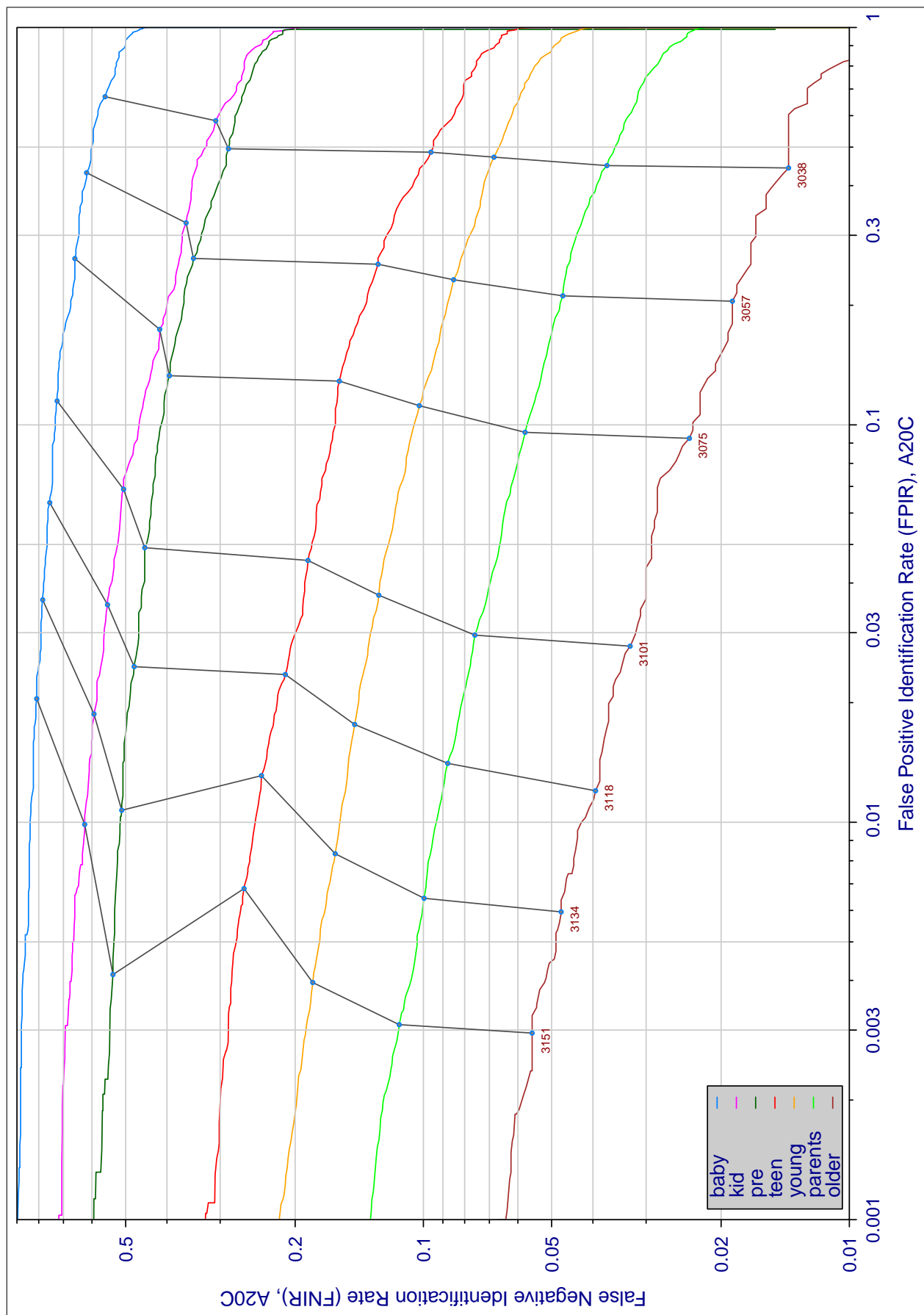


Figure 57: The effect of age on accuracy A20C. Each trace is a detection error tradeoff characteristic for individuals within one of seven age groups at the time their search image was collected. Upward vertical displacement indicates increased difficulty for persons to be recognized as themselves. Rightward displacement indicates that individuals are more often confused as somebody else (i.e. false positives).

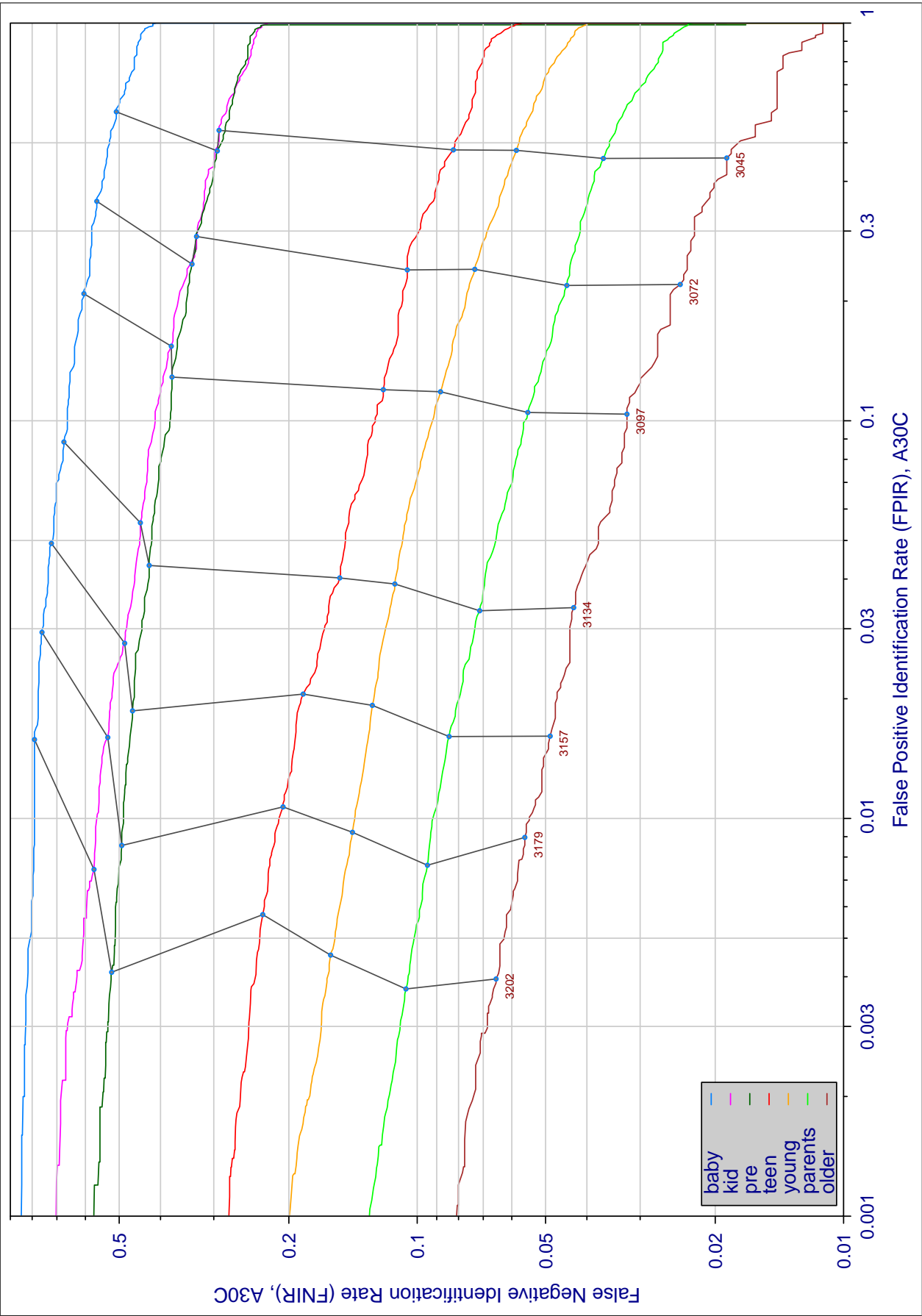


Figure 58: The effect of age on accuracy A30C. Each trace is a detection error tradeoff characteristic for individuals within one of seven age groups at the time their search image was collected. Upward vertical displacement indicates increased difficult for persons to be recognized as themselves. Rightward displacement indicates that individuals are more often confused a somebody else (i.e. false positives).

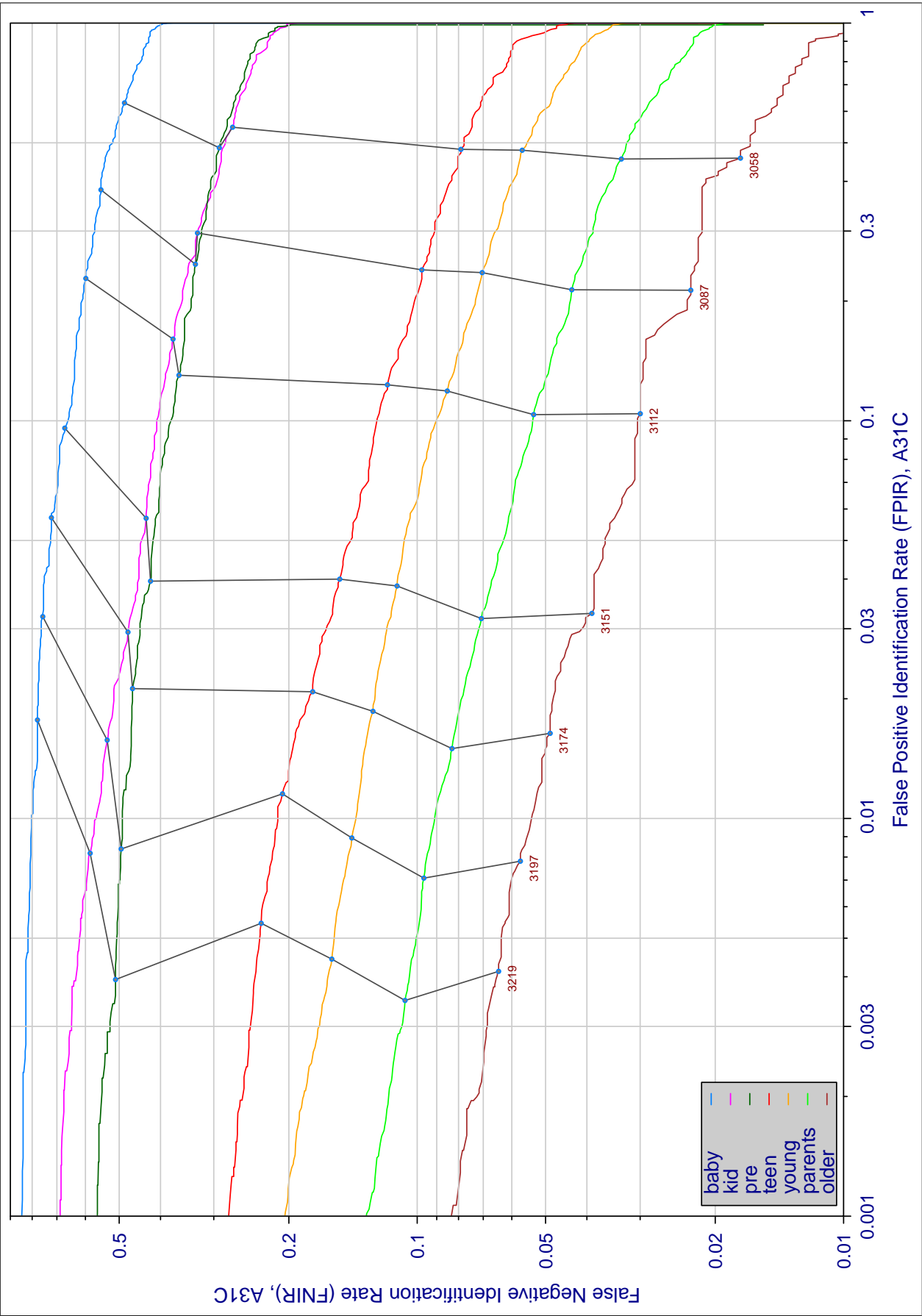


Figure 59: The effect of age on accuracy A31C. Each trace is a detection error tradeoff characteristic for individuals within one of seven age groups at the time their search image was collected. Upward vertical displacement indicates increased difficult for persons to be recognized as themselves. Rightward displacement indicates that individuals are more often confused a somebody else (i.e. false positives).

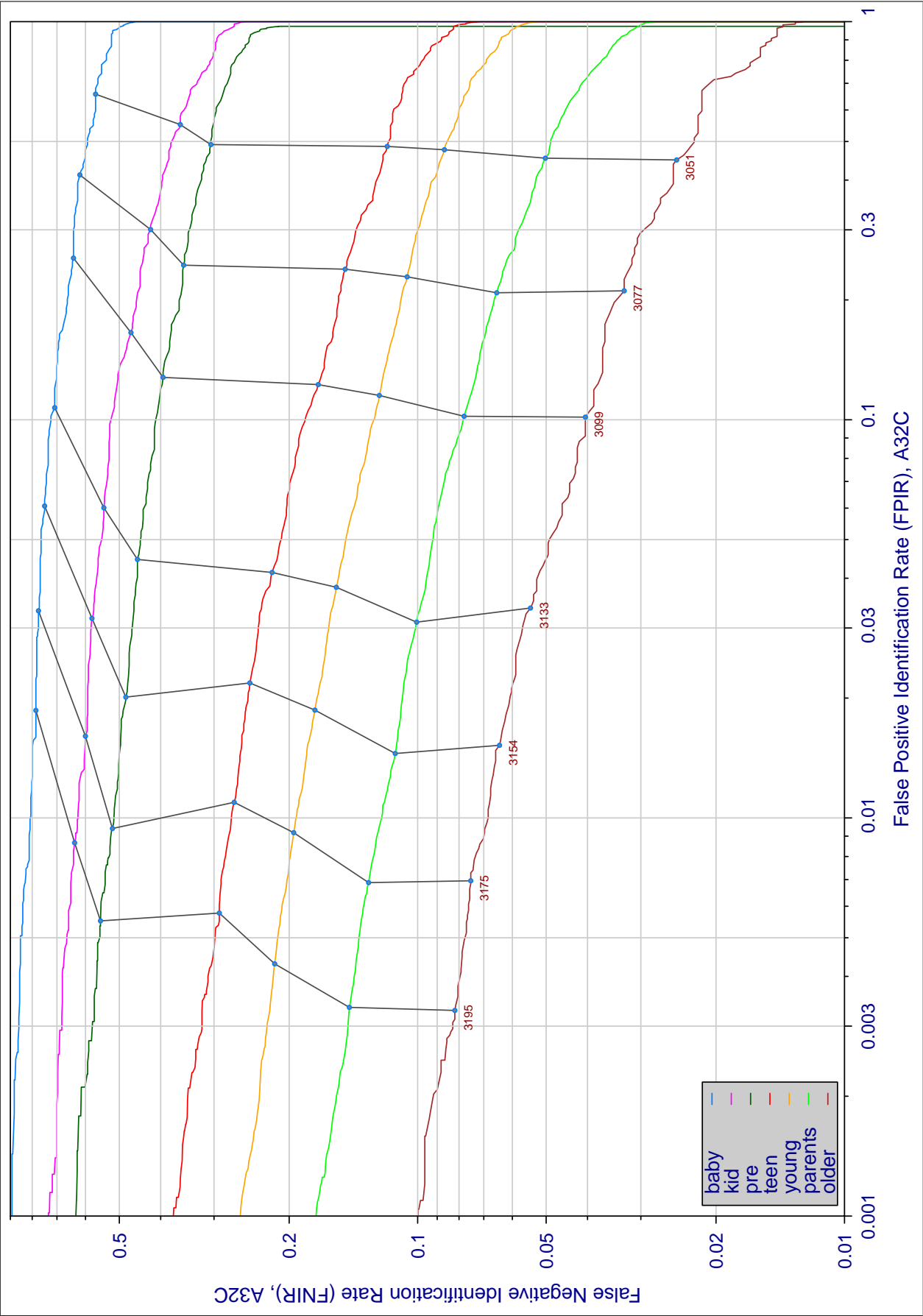


Figure 60: The effect of age on accuracy A32C. Each trace is a detection error tradeoff characteristic for individuals within one of seven age groups at the time their search image was collected. Upward vertical displacement indicates increased difficult for persons to be recognized as themselves. Rightward displacement indicates that individuals are more often confused a somebody else (i.e. false positives).

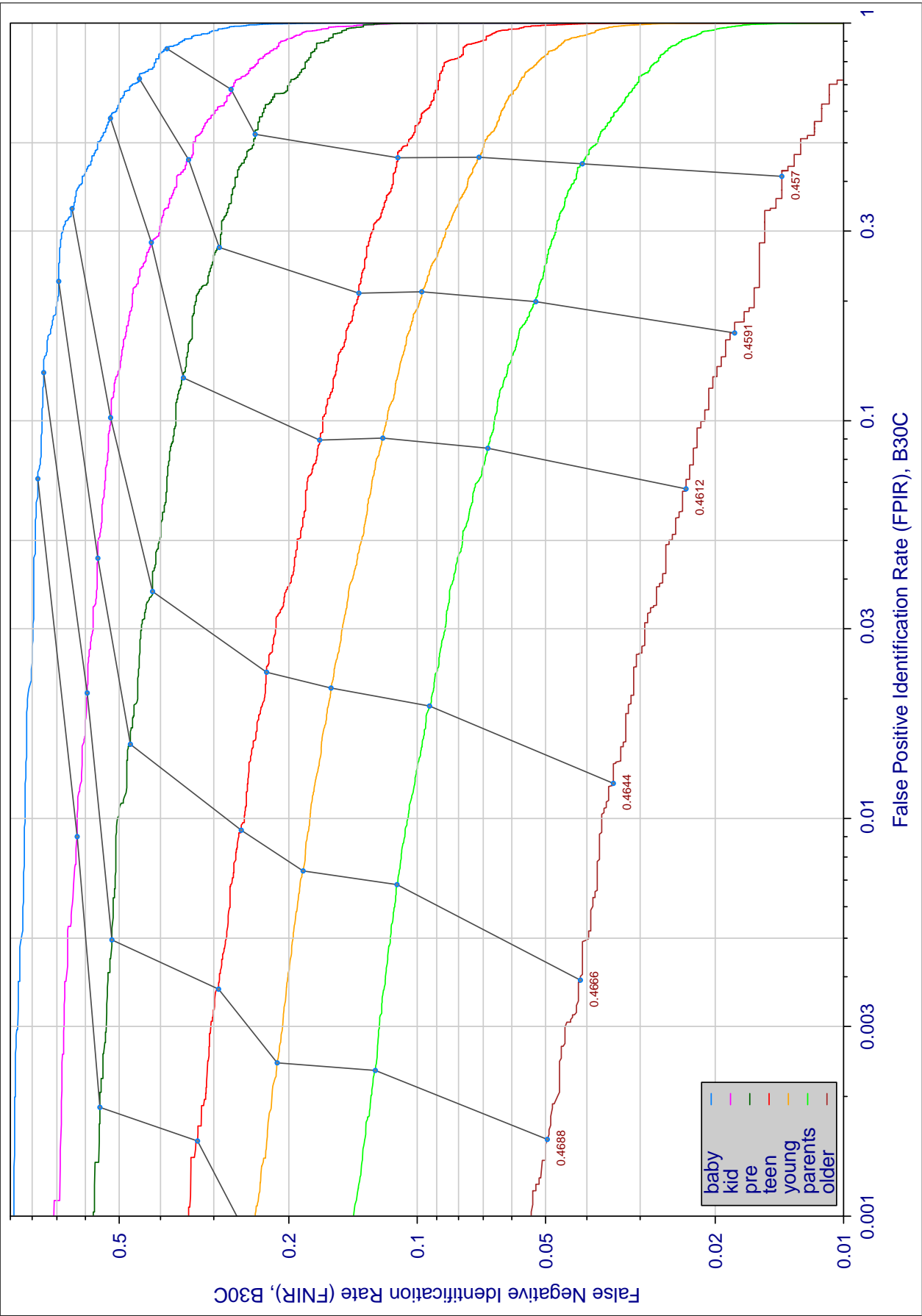


Figure 61: The effect of age on accuracy B30C. Each trace is a detection error tradeoff characteristic for individuals within one of seven age groups at the time their search image was collected. Upward vertical displacement indicates increased difficult for persons to be recognized as themselves. Rightward displacement indicates that individuals are more often confused a somebody else (i.e. false positives).

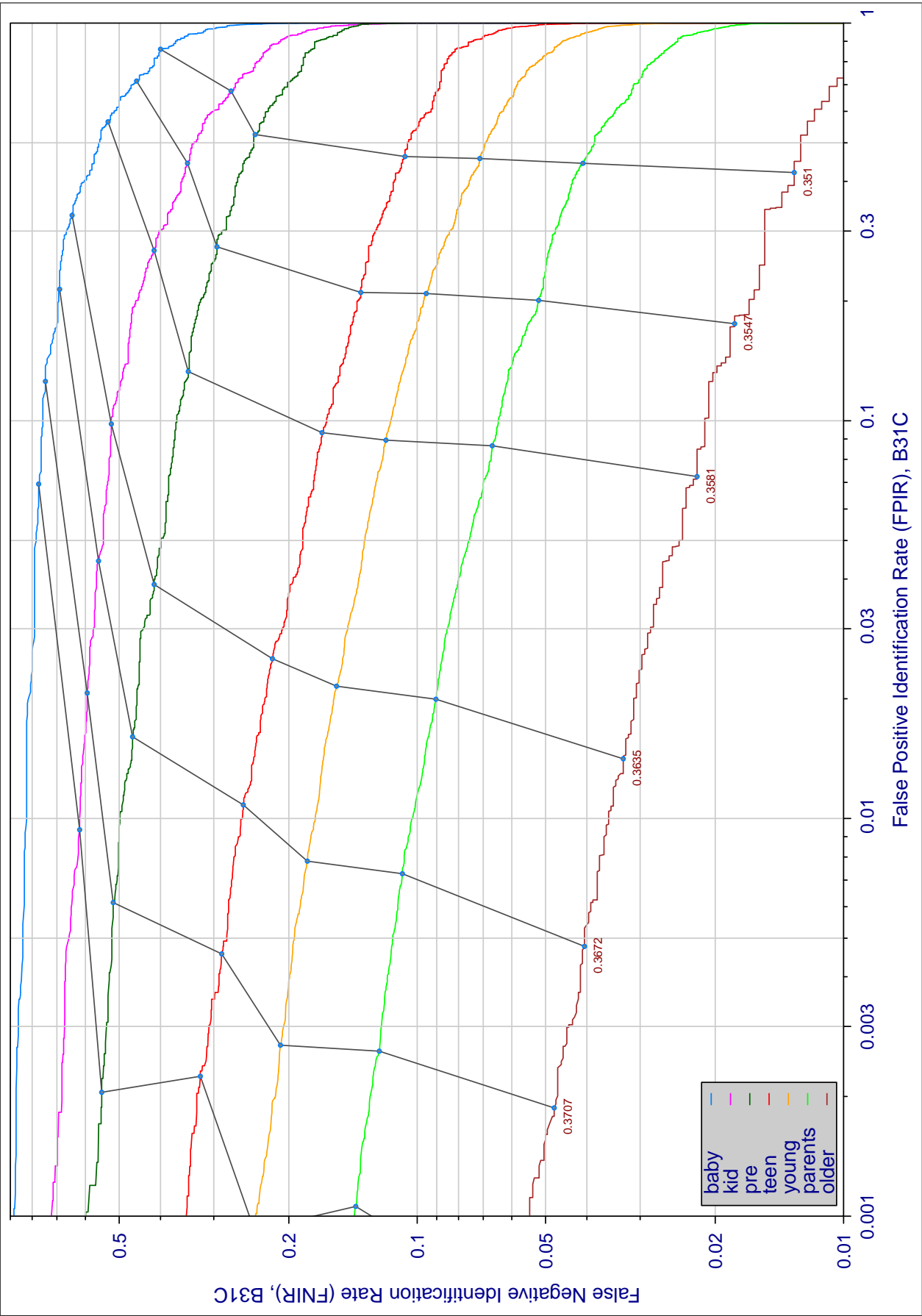


Figure 62: The effect of age on accuracy B31C. Each trace is a detection error tradeoff characteristic for individuals within one of seven age groups at the time their search image was collected. Upward vertical displacement indicates increased difficult for persons to be recognized as themselves. Rightward displacement indicates that individuals are more often confused a somebody else (i.e. false positives).

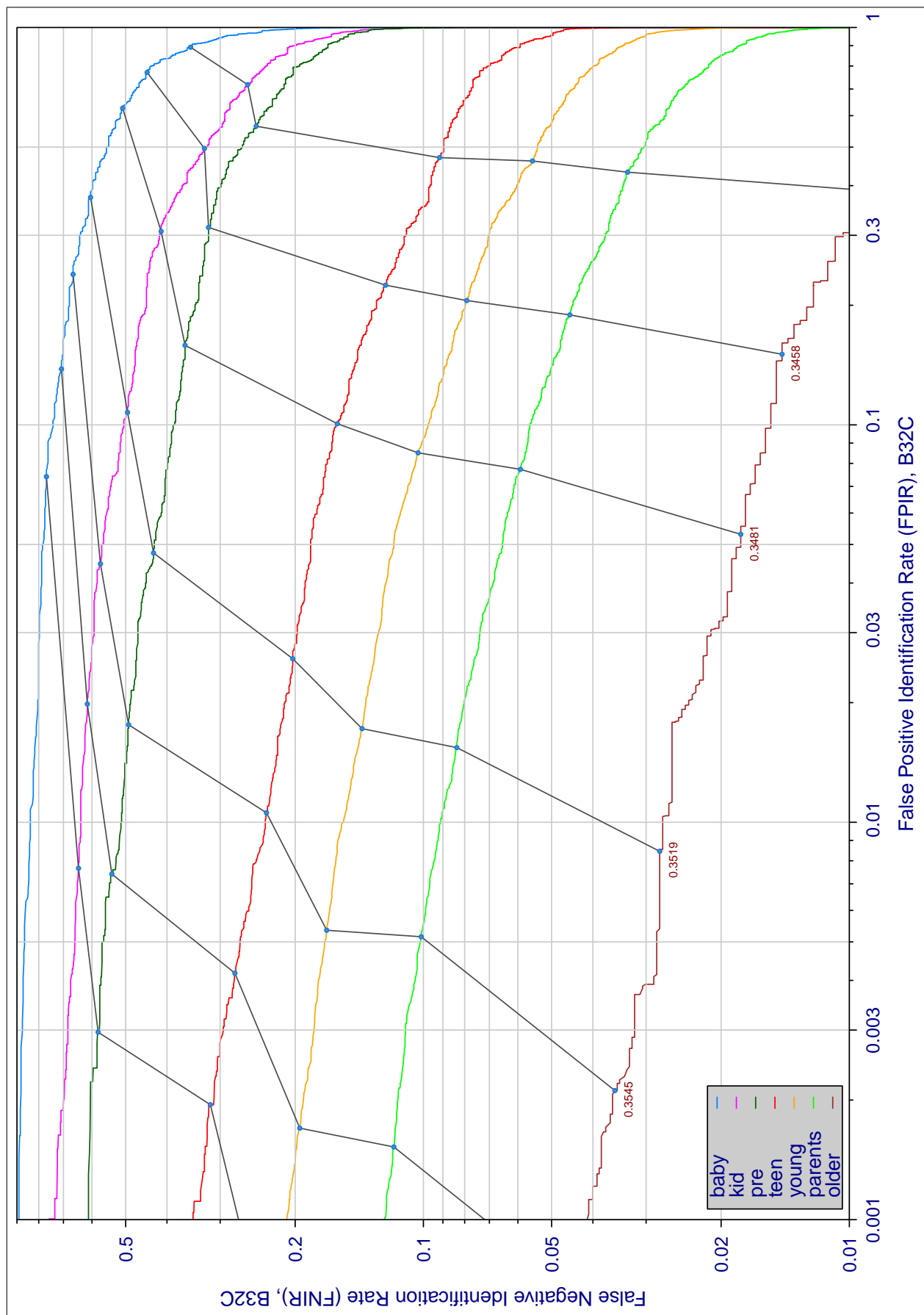


Figure 63: The effect of age on accuracy B32C. Each trace is a detection error tradeoff characteristic for individuals within one of seven age groups at the time their search image was collected. Upward vertical displacement indicates increased difficulty for persons to be recognized as themselves. Rightward displacement indicates that individuals are more often confused as somebody else (i.e. false positives).

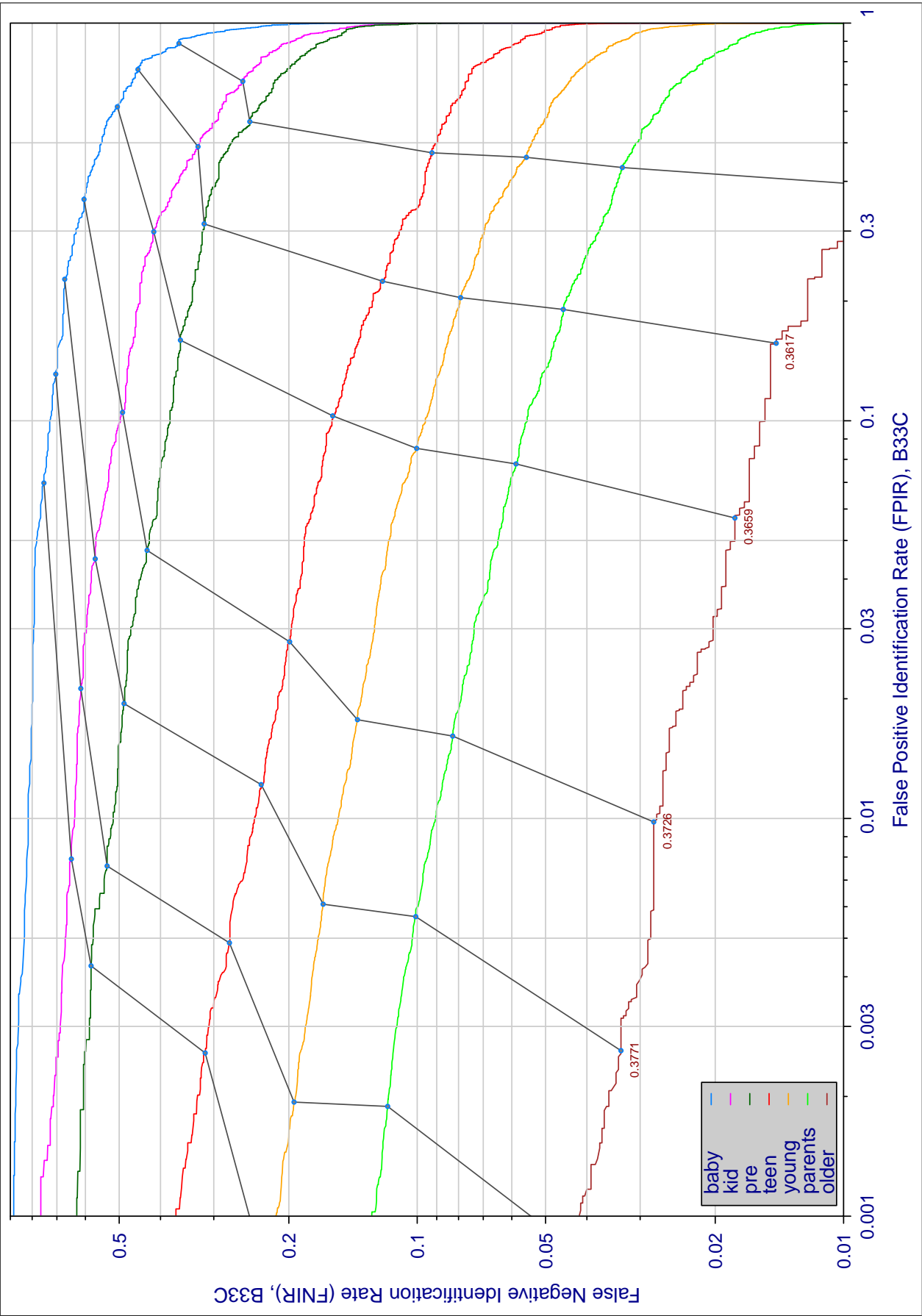


Figure 64: The effect of age on accuracy B33C. Each trace is a detection error tradeoff characteristic for individuals within one of seven age groups at the time their search image was collected. Upward vertical displacement indicates increased difficult for persons to be recognized as themselves. Rightward displacement indicates that individuals are more often confused a somebody else (i.e. false positives).

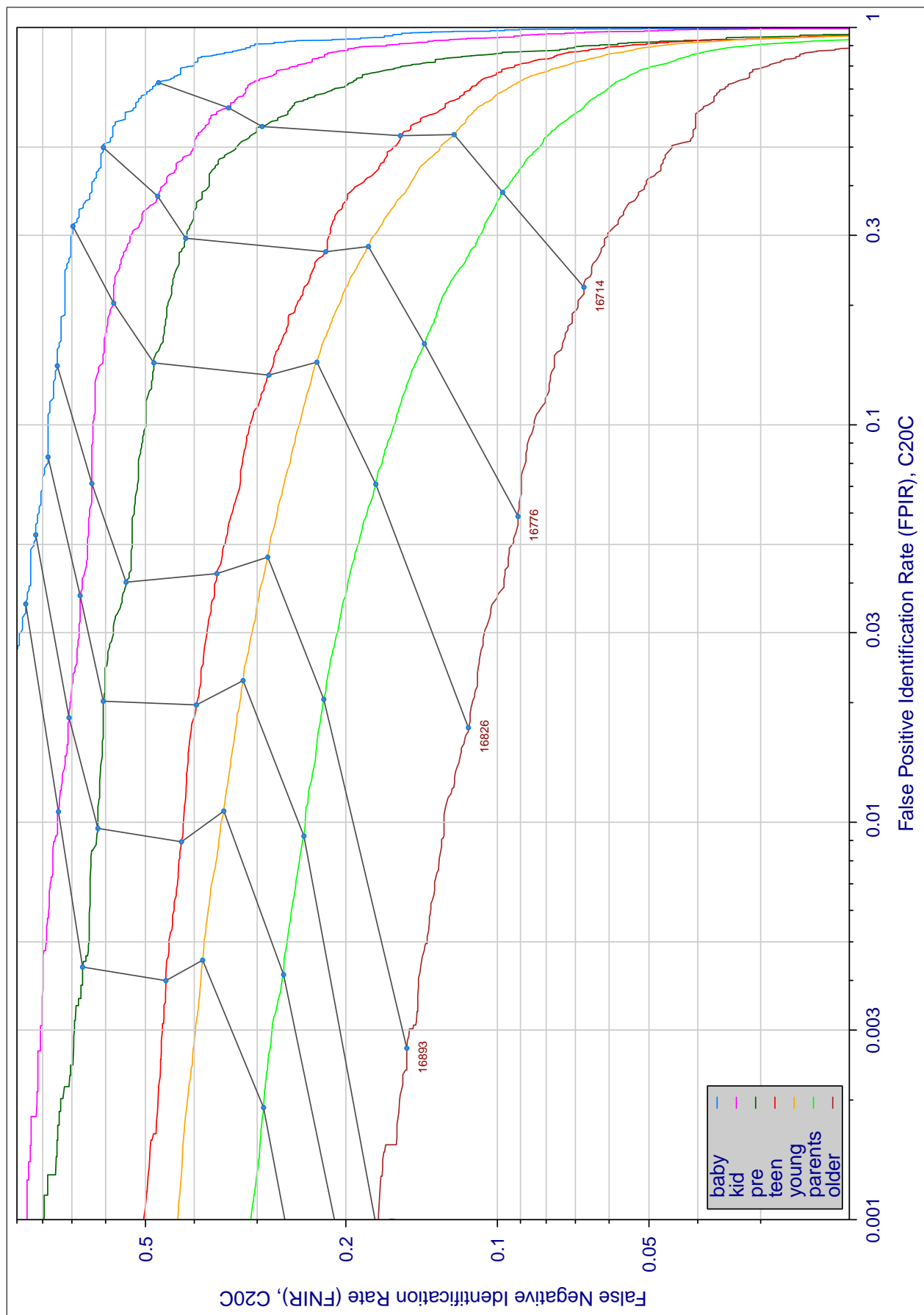


Figure 65: The effect of age on accuracy C20C. Each trace is a detection error tradeoff characteristic for individuals within one of seven age groups at the time their search image was collected. Upward vertical displacement indicates increased difficulty for persons to be recognized as themselves. Rightward displacement indicates that individuals are more often confused as somebody else (i.e. false positives).

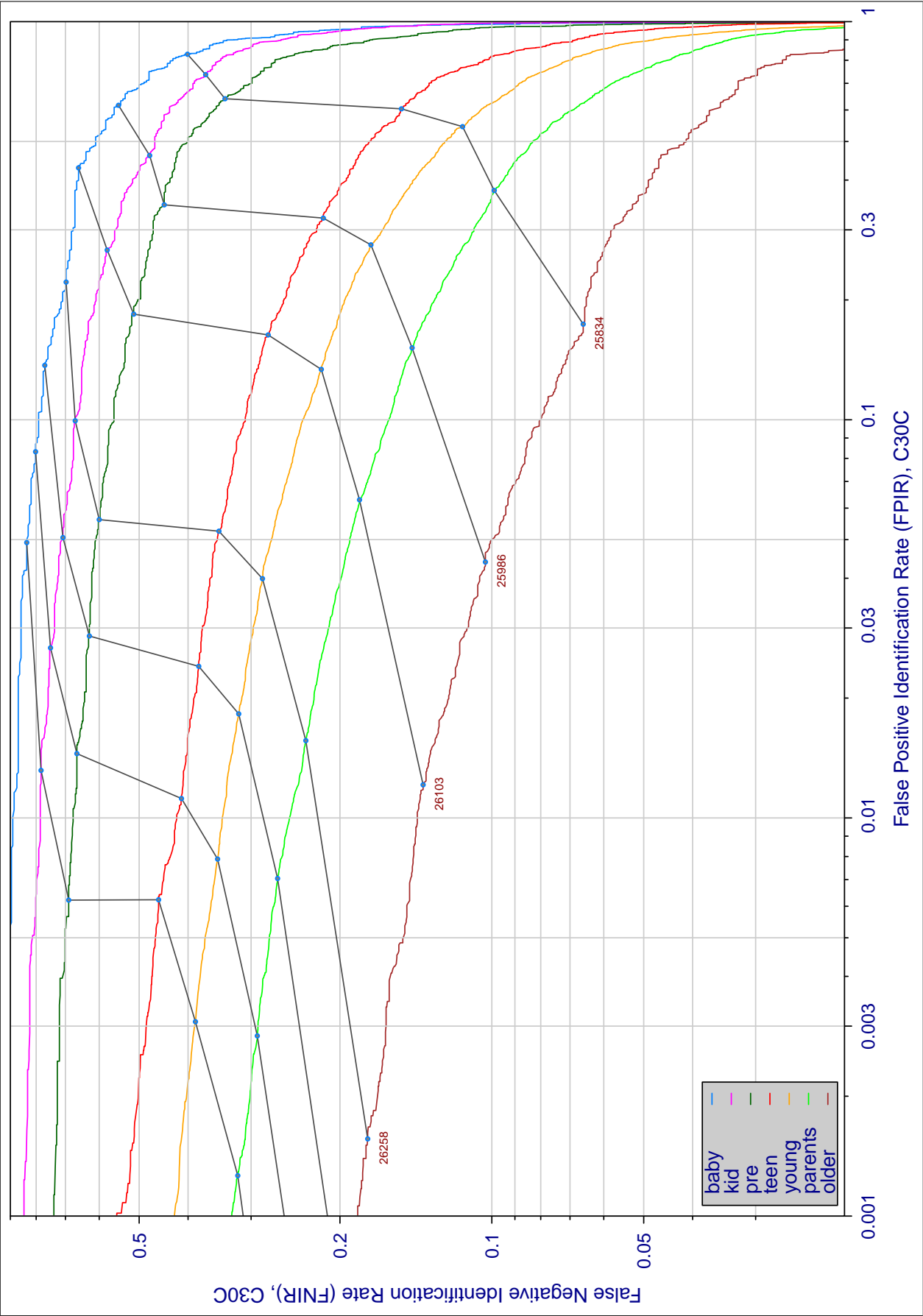


Figure 66: The effect of age on accuracy C30C. Each trace is a detection error tradeoff characteristic for individuals within one of seven age groups at the time their search image was collected. Upward vertical displacement indicates increased difficult for persons to be recognized as themselves. Rightward displacement indicates that individuals are more often confused a somebody else (i.e. false positives).

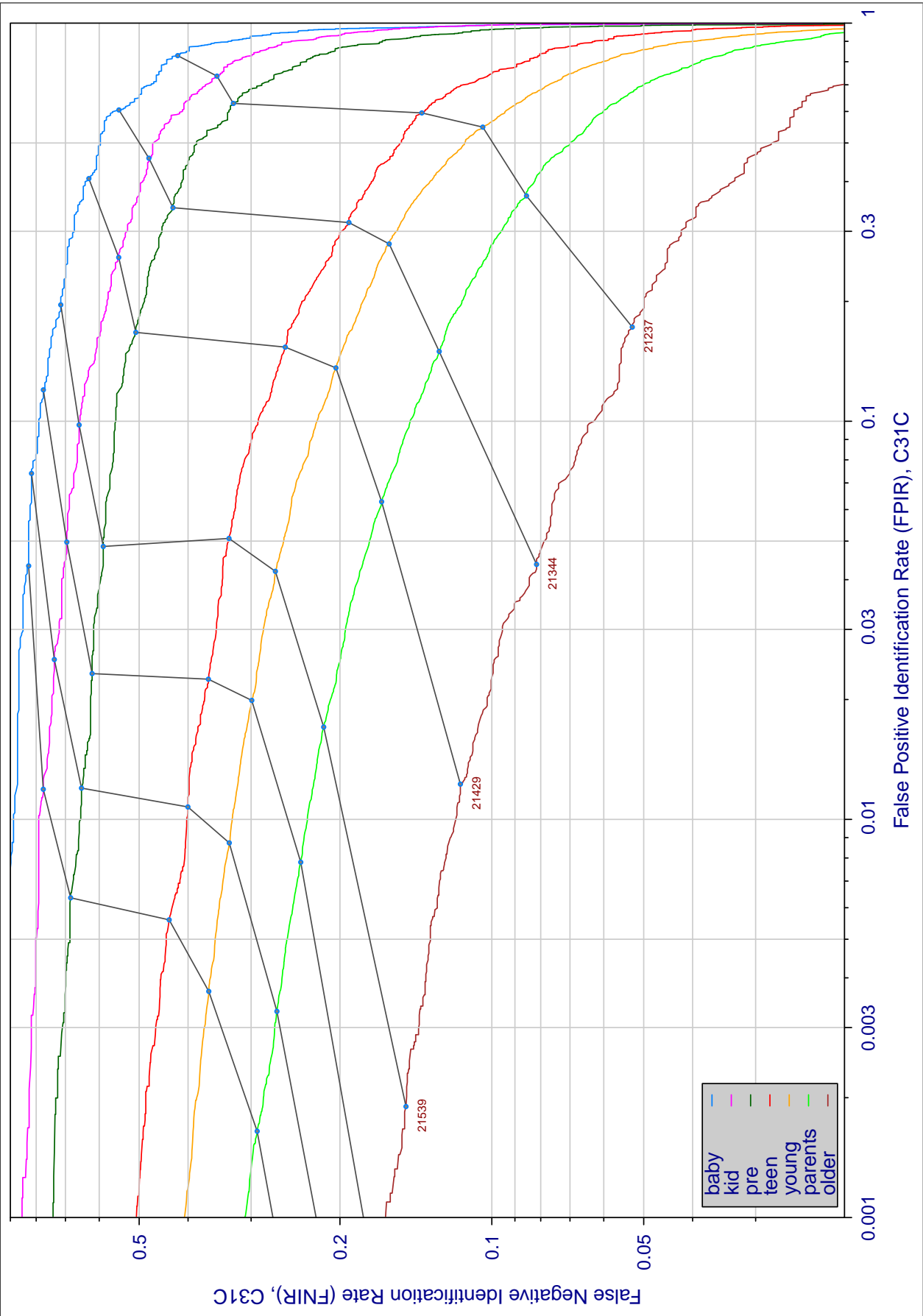


Figure 67: The effect of age on accuracy C31C. Each trace is a detection error tradeoff characteristic for individuals within one of seven age groups at the time their search image was collected. Upward vertical displacement indicates increased difficult for persons to be recognized as themselves. Rightward displacement indicates that individuals are more often confused a somebody else (i.e. false positives).

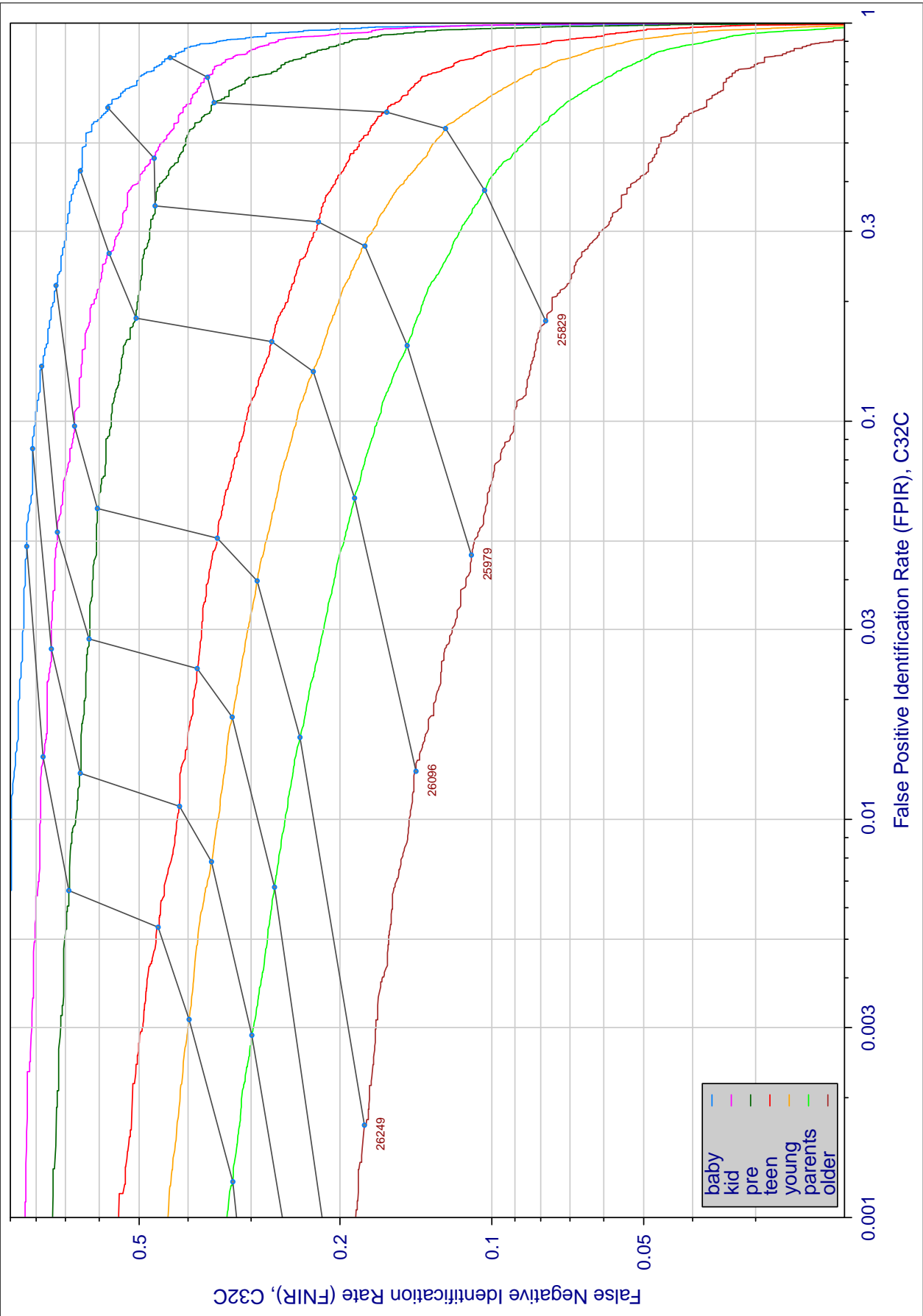


Figure 68: The effect of age on accuracy C32C. Each trace is a detection error tradeoff characteristic for individuals within one of seven age groups at the time their search image was collected. Upward vertical displacement indicates increased difficult for persons to be recognized as themselves. Rightward displacement indicates that individuals are more often confused a somebody else (i.e. false positives).

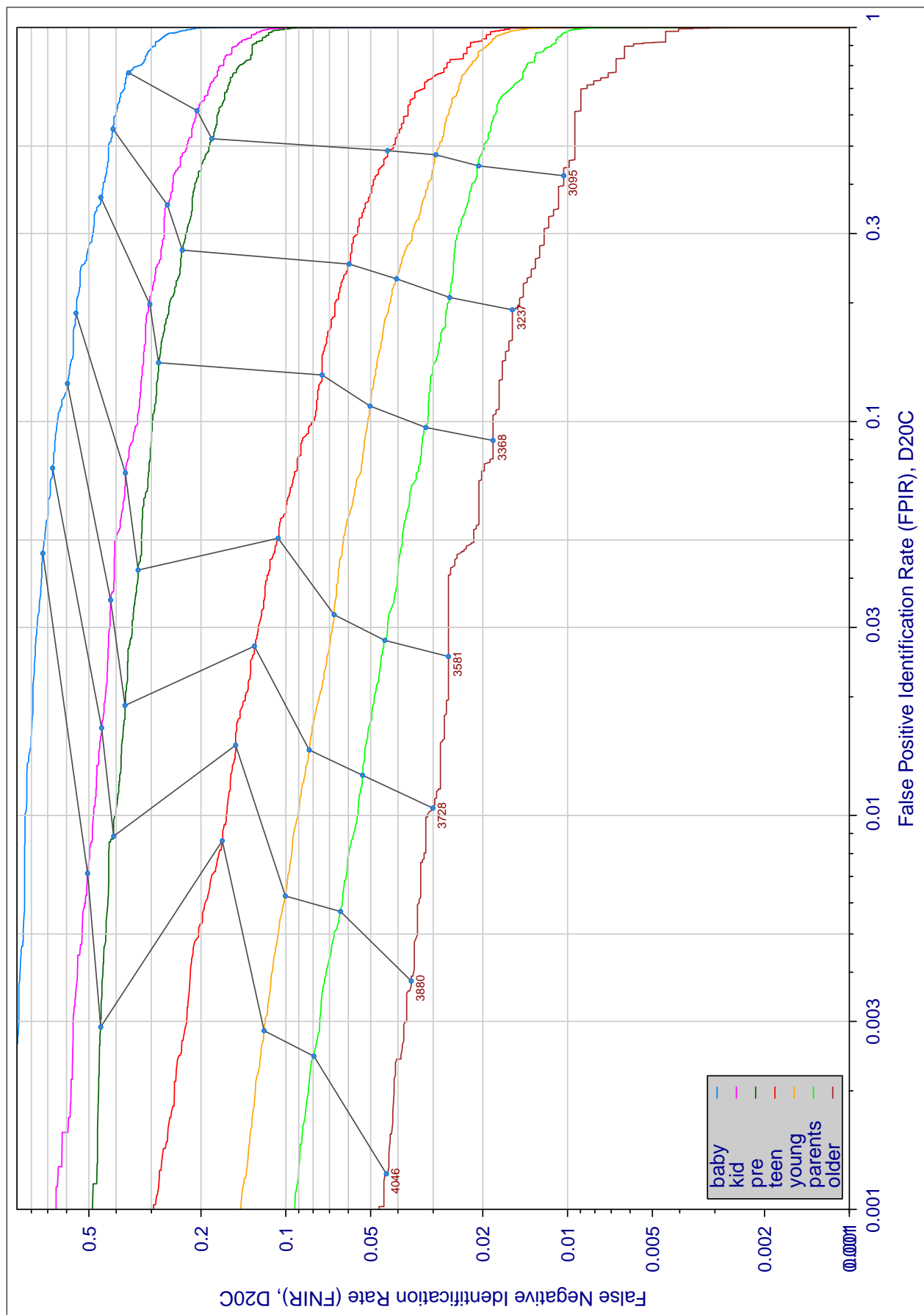


Figure 69: The effect of age on accuracy D20C. Each trace is a detection error tradeoff characteristic for individuals within one of seven age groups at the time their search image was collected. Upward vertical displacement indicates increased difficult for persons to be recognized as themselves. Rightward displacement indicates that individuals are more often confused a somebody else (i.e. false positives).

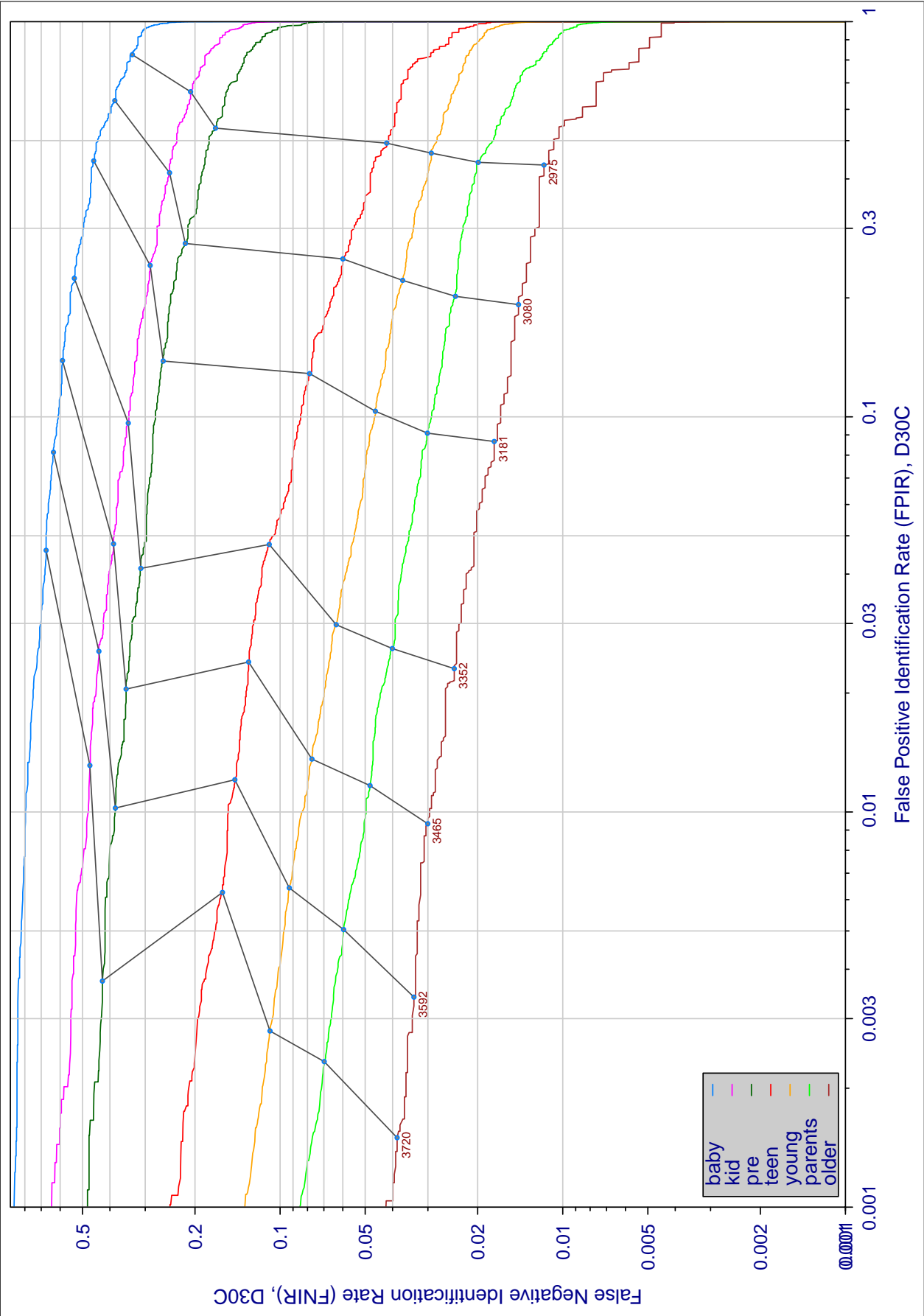


Figure 70: The effect of age on accuracy D30C. Each trace is a detection error tradeoff characteristic for individuals within one of seven age groups at the time their search image was collected. Upward vertical displacement indicates increased difficult for persons to be recognized as themselves. Rightward displacement indicates that individuals are more often confused a somebody else (i.e. false positives).

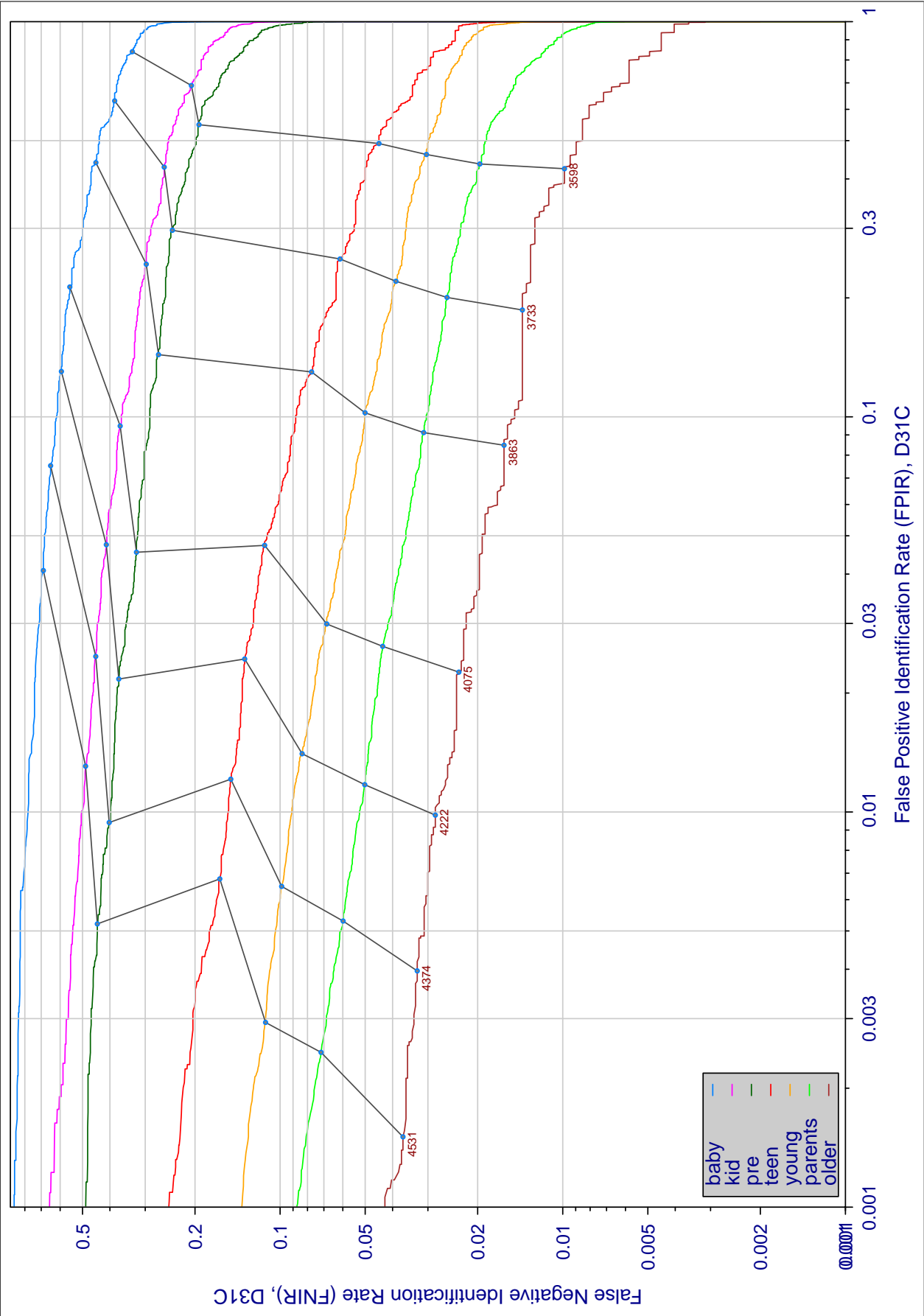


Figure 71: The effect of age on accuracy D31C. Each trace is a detection error tradeoff characteristic for individuals within one of seven age groups at the time their search image was collected. Upward vertical displacement indicates increased difficult for persons to be recognized as themselves. Rightward displacement indicates that individuals are more often confused a somebody else (i.e. false positives).

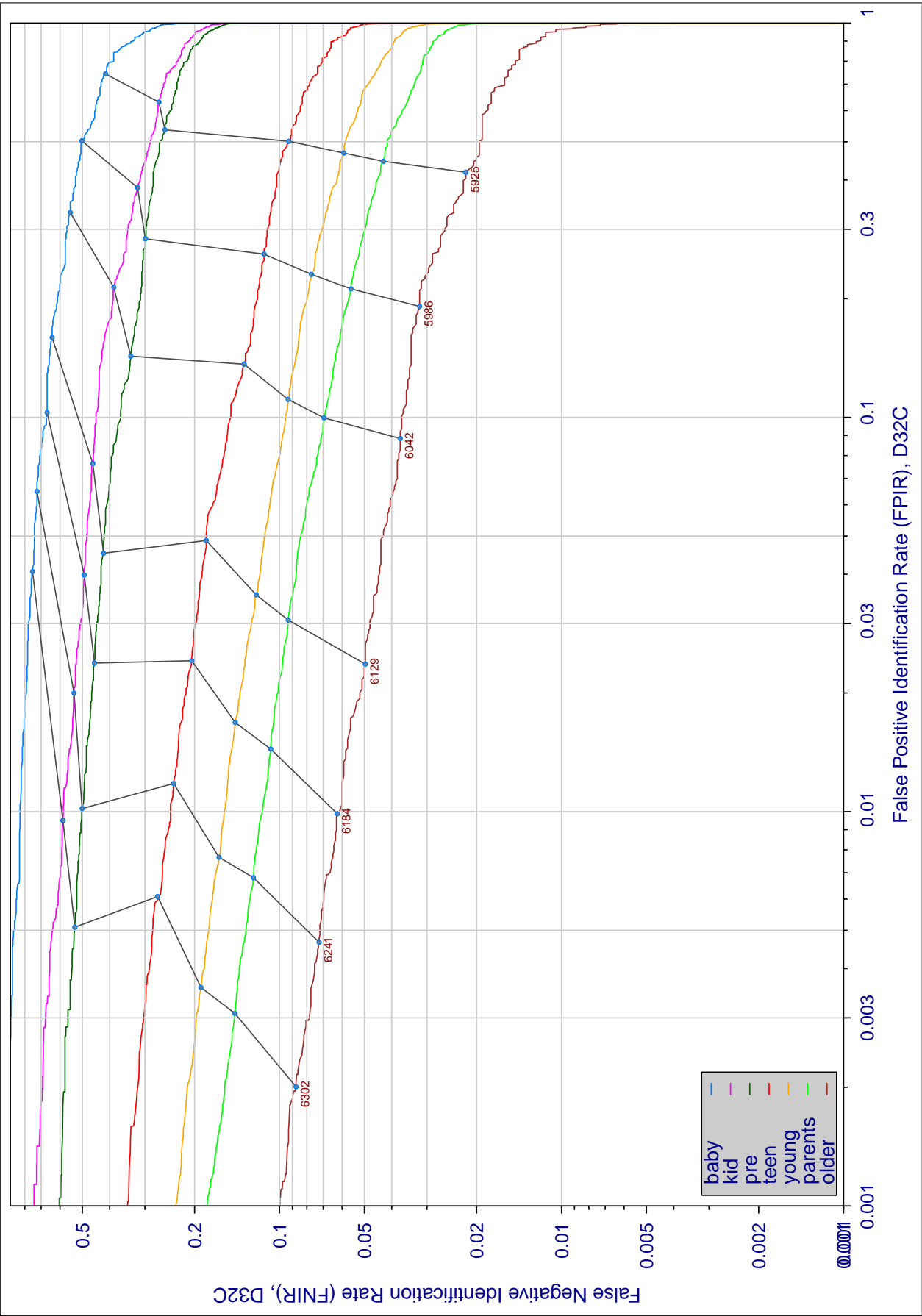


Figure 72: The effect of age on accuracy D32C. Each trace is a detection error tradeoff characteristic for individuals within one of seven age groups at the time their search image was collected. Upward vertical displacement indicates increased difficult for persons to be recognized as themselves. Rightward displacement indicates that individuals are more often confused a somebody else (i.e. false positives).

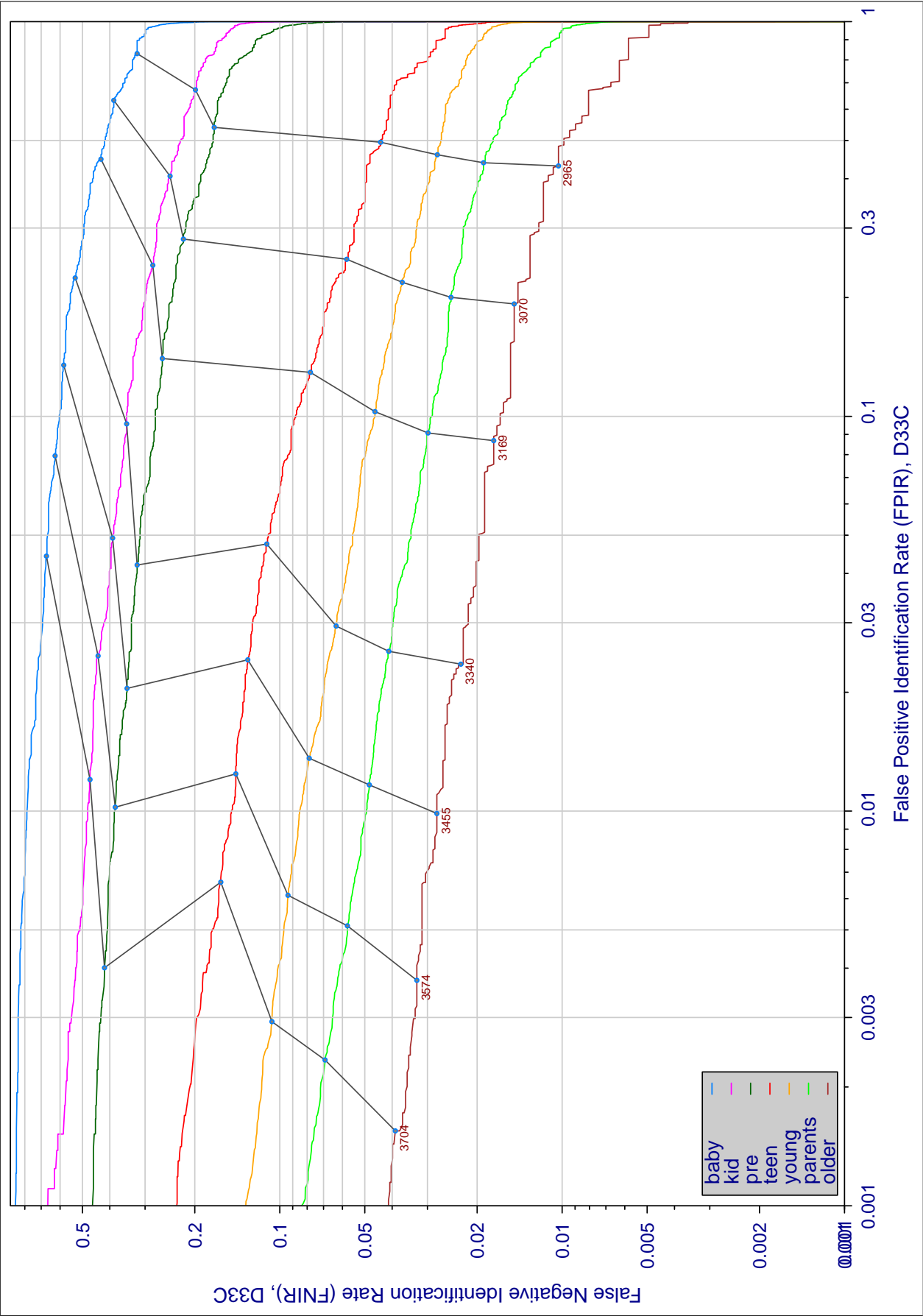


Figure 73: The effect of age on accuracy D33C. Each trace is a detection error tradeoff characteristic for individuals within one of seven age groups at the time their search image was collected. Upward vertical displacement indicates increased difficult for persons to be recognized as themselves. Rightward displacement indicates that individuals are more often confused a somebody else (i.e. false positives).

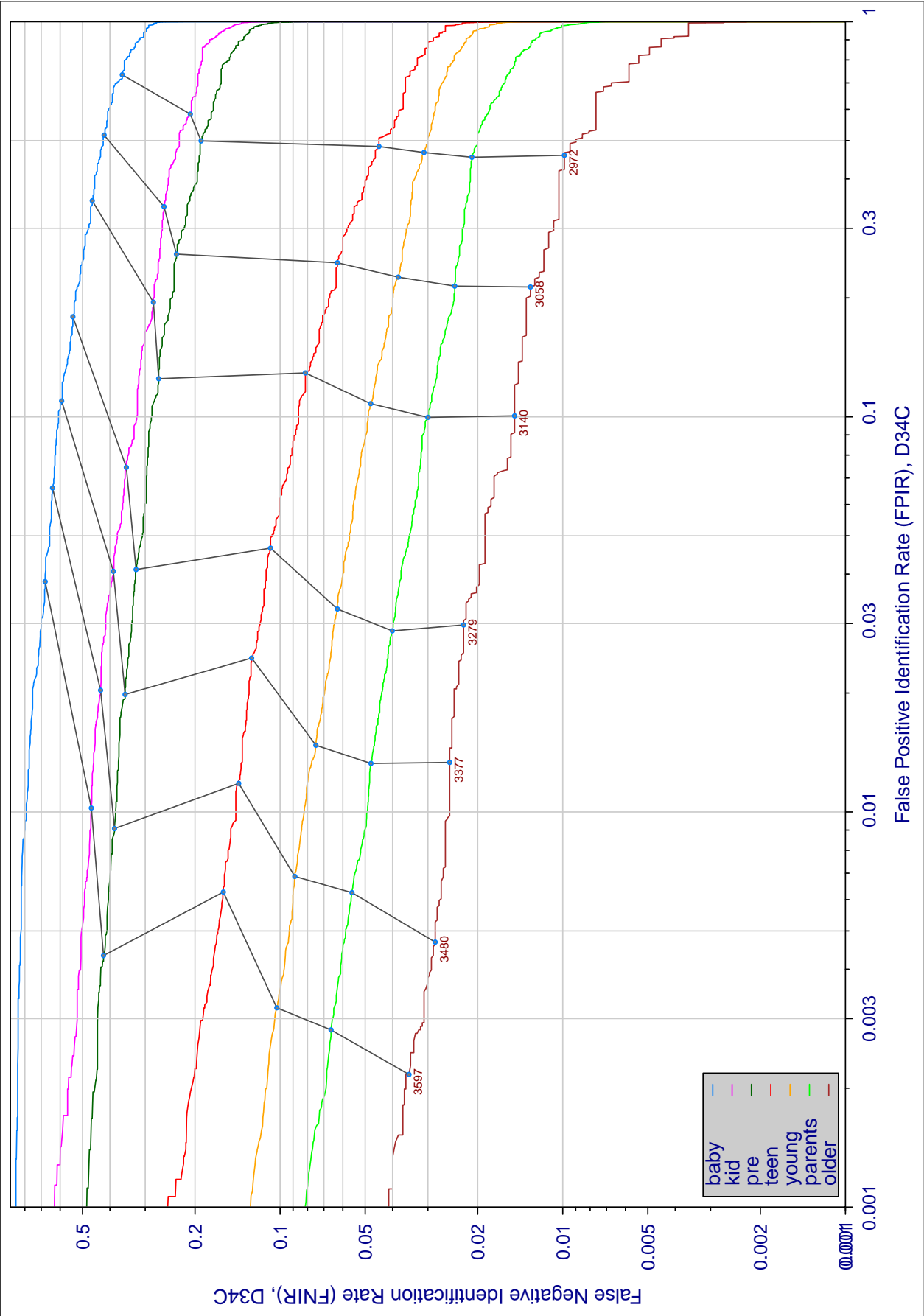


Figure 74: The effect of age on accuracy D34C. Each trace is a detection error tradeoff characteristic for individuals within one of seven age groups at the time their search image was collected. Upward vertical displacement indicates increased difficult for persons to be recognized as themselves. Rightward displacement indicates that individuals are more often confused a somebody else (i.e. false positives).

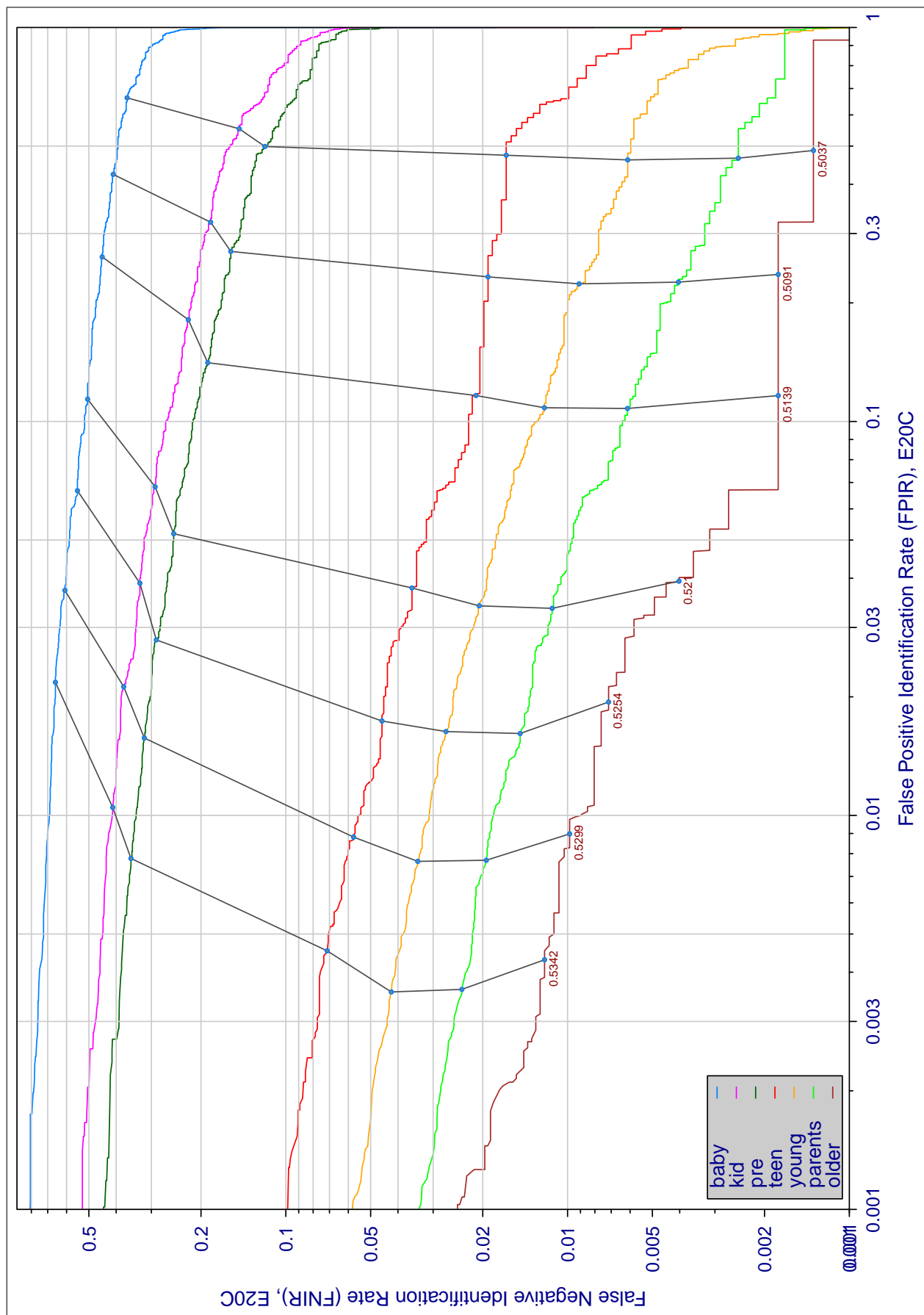


Figure 75: The effect of age on accuracy E20C. Each trace is a detection error tradeoff characteristic for individuals within one of seven age groups at the time their search image was collected. Upward vertical displacement indicates increased difficulty for persons to be recognized as themselves. Rightward displacement indicates that individuals are more often confused as somebody else (i.e. false positives).

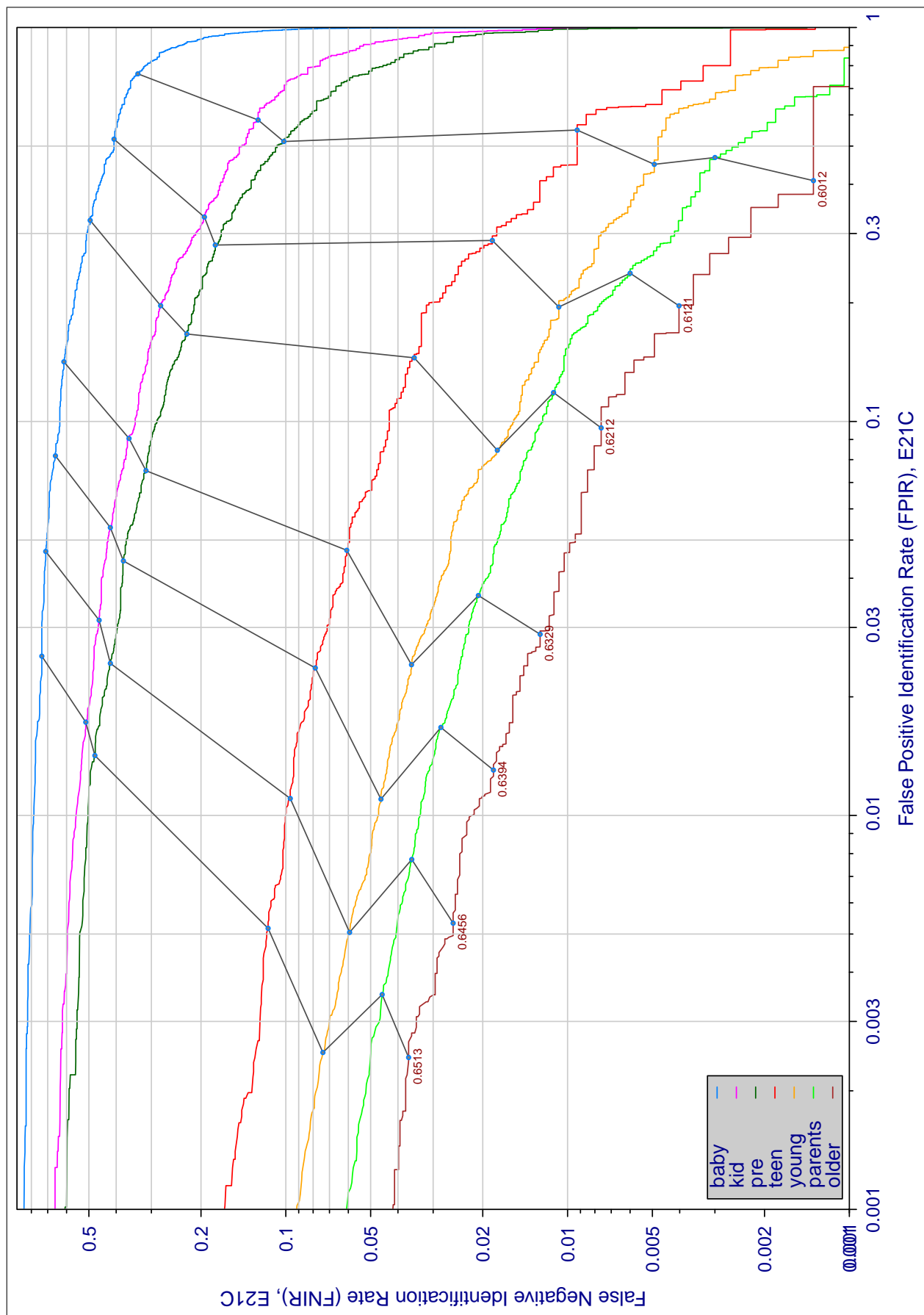


Figure 76: The effect of age on accuracy E21C. Each trace is a detection error tradeoff characteristic for individuals within one of seven age groups at the time their search image was collected. Upward vertical displacement indicates increased difficult for persons to be recognized as themselves. Rightward displacement indicates that individuals are more often confused a somebody else (i.e. false positives).

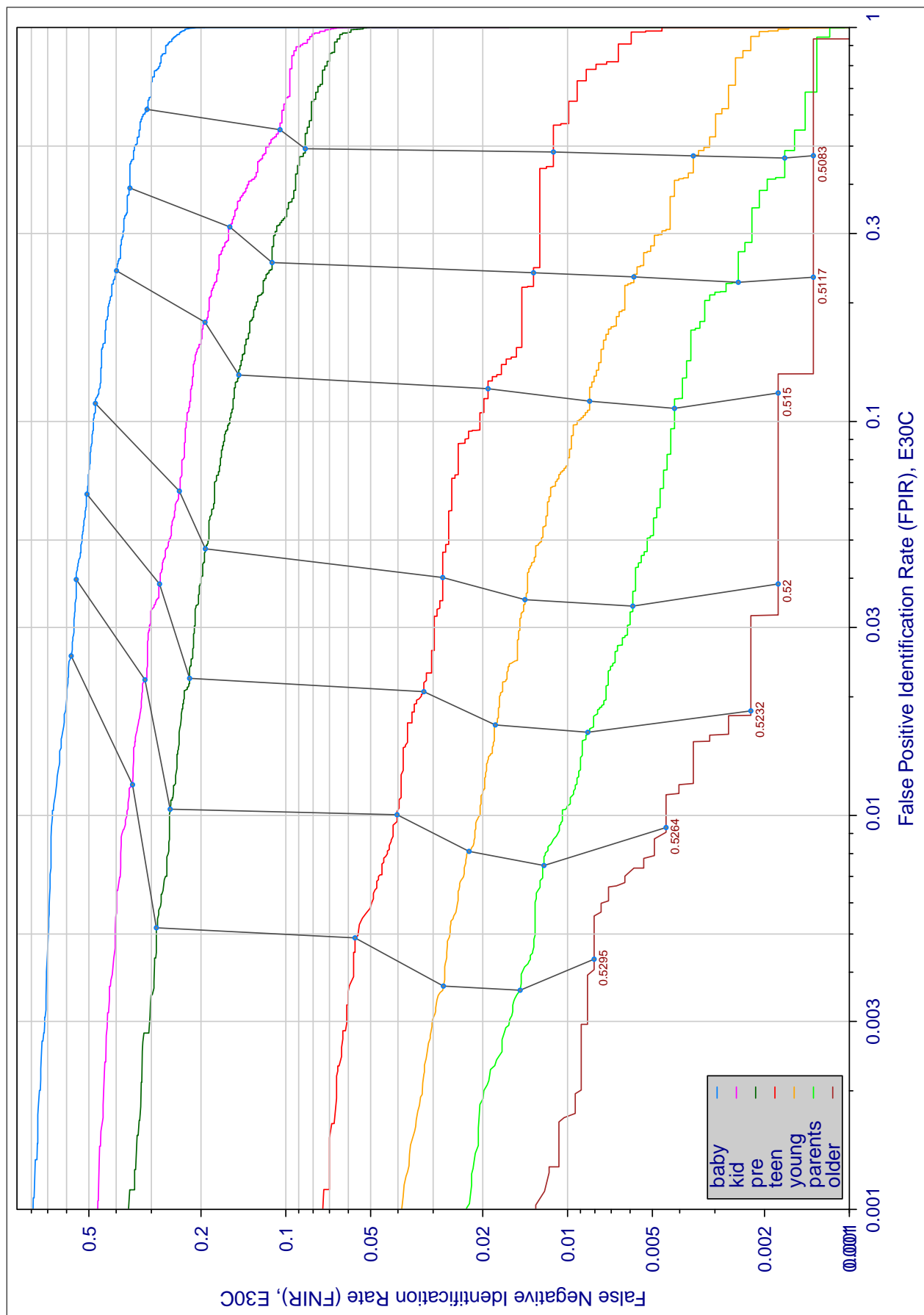


Figure 77: The effect of age on accuracy E30C. Each trace is a detection error tradeoff characteristic for individuals within one of seven age groups at the time their search image was collected. Upward vertical displacement indicates increased difficult for persons to be recognized as themselves. Rightward displacement indicates that individuals are more often confused a somebody else (i.e. false positives).

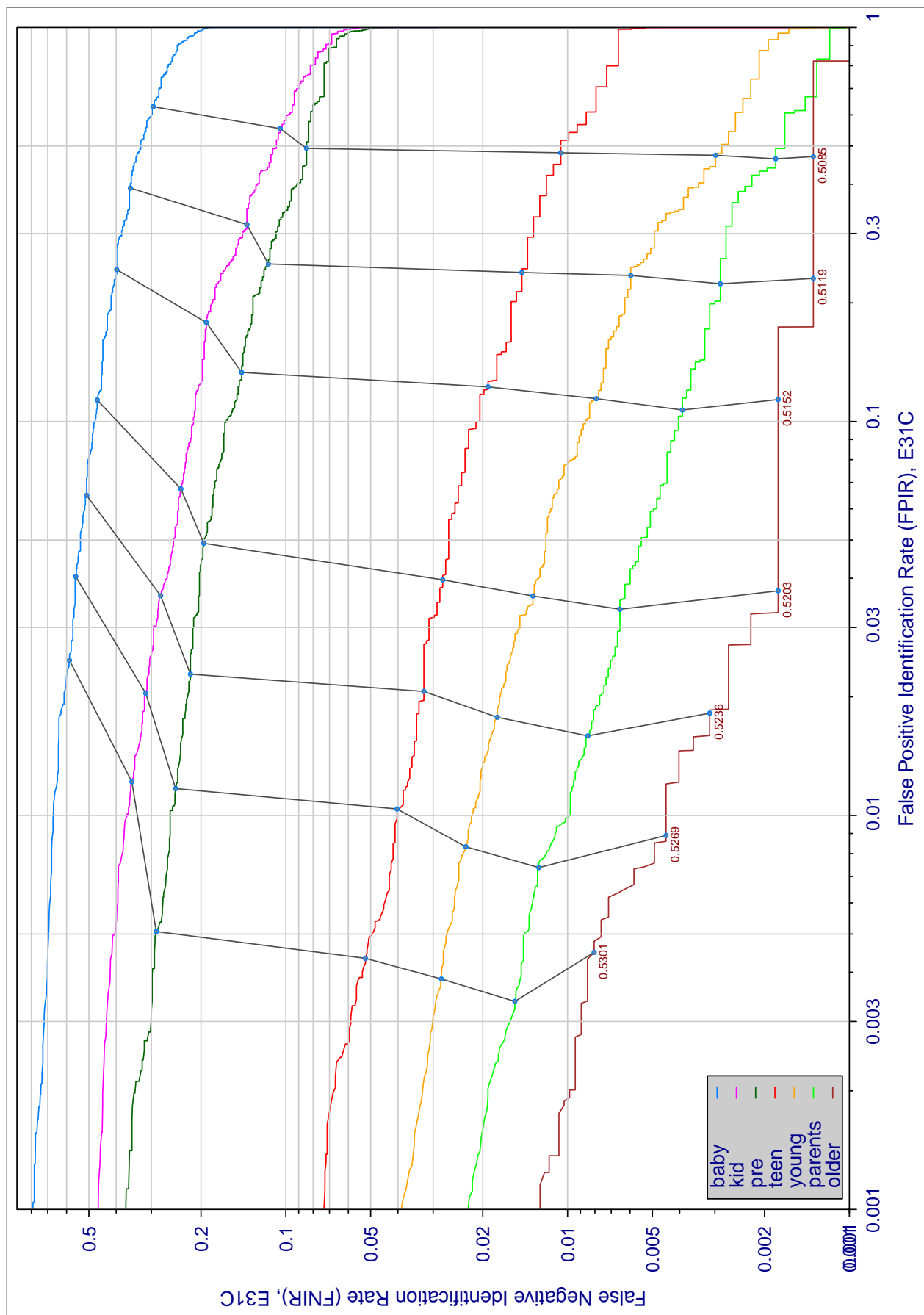


Figure 78: The effect of age on accuracy E31C. Each trace is a detection error tradeoff characteristic for individuals within one of seven age groups at the time their search image was collected. Upward vertical displacement indicates increased difficulty for persons to be recognized as themselves. Rightward displacement indicates that individuals are more often confused as somebody else (i.e. false positives).

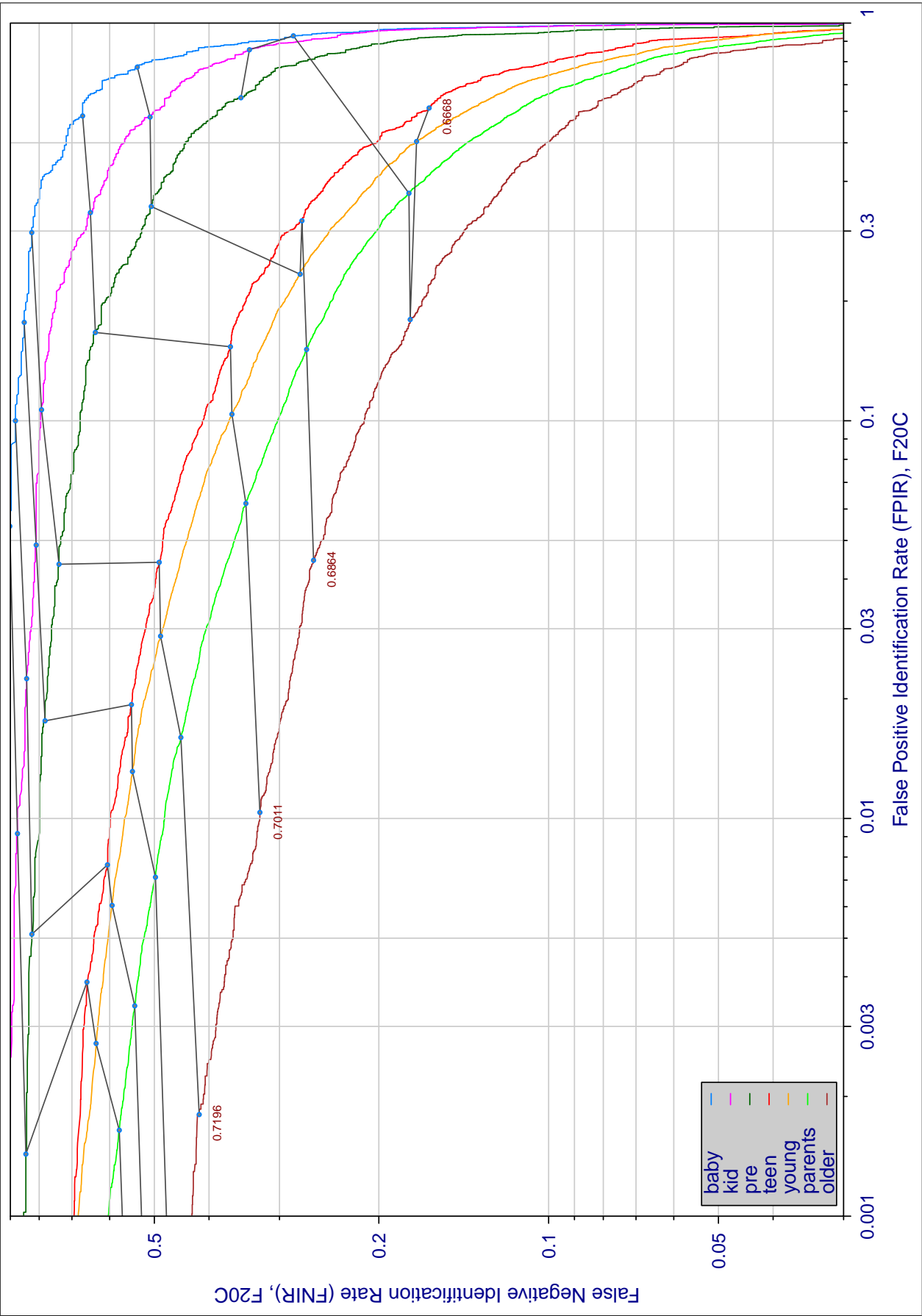


Figure 79: The effect of age on accuracy F20C. Each trace is a detection error tradeoff characteristic for individuals within one of seven age groups at the time their search image was collected. Upward vertical displacement indicates increased difficult for persons to be recognized as themselves. Rightward displacement indicates that individuals are more often confused a somebody else (i.e. false positives).

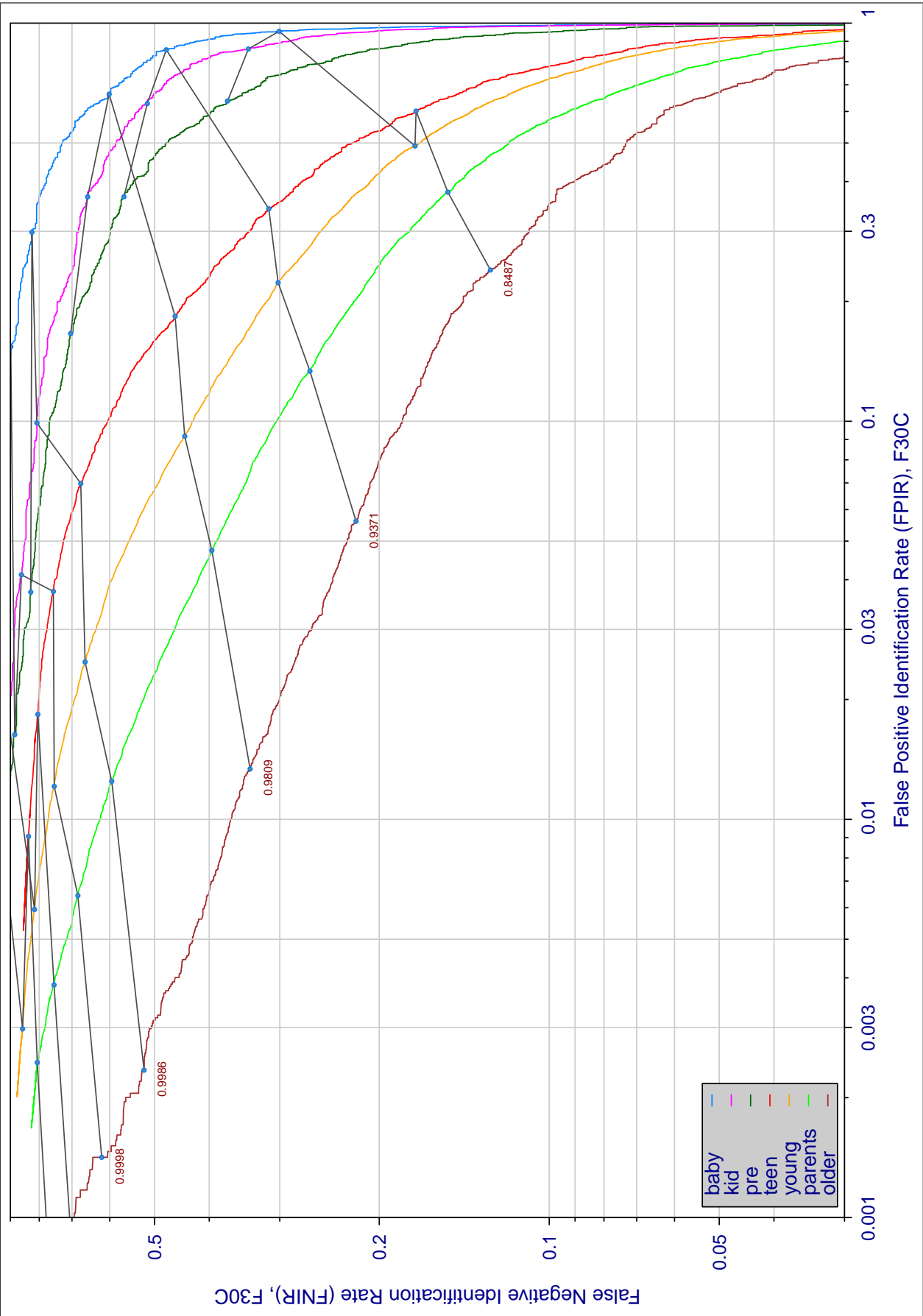


Figure 80: The effect of age on accuracy F30C. Each trace is a detection error tradeoff characteristic for individuals within one of seven age groups at the time their search image was collected. Upward vertical displacement indicates increased difficult for persons to be recognized as themselves. Rightward displacement indicates that individuals are more often confused a somebody else (i.e. false positives).

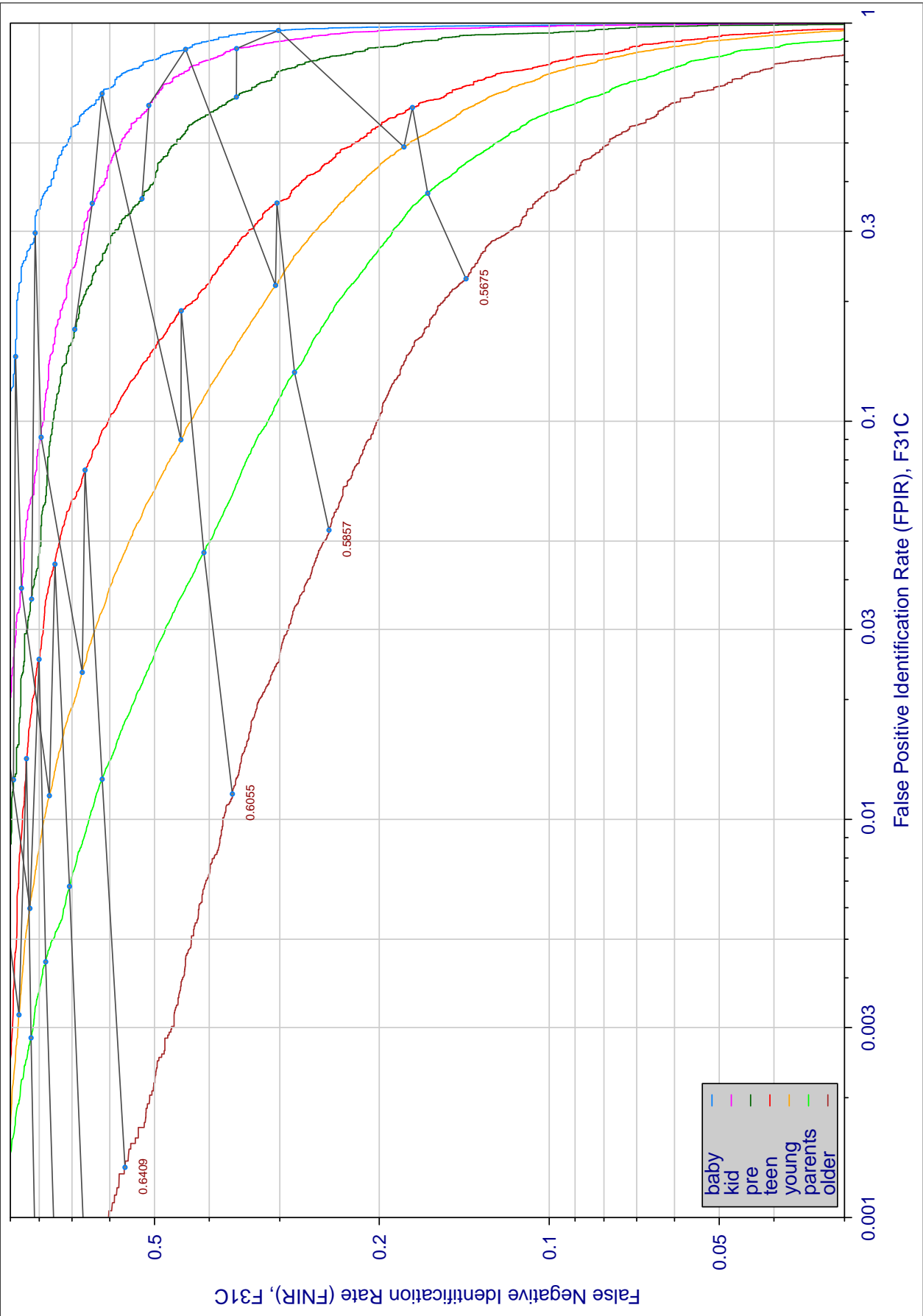


Figure 8I: The effect of age on accuracy F31C. Each trace is a detection error tradeoff characteristic for individuals within one of seven age groups at the time their search image was collected. Upward vertical displacement indicates increased difficult for persons to be recognized as themselves. Rightward displacement indicates that individuals are more often confused a somebody else (i.e. false positives).

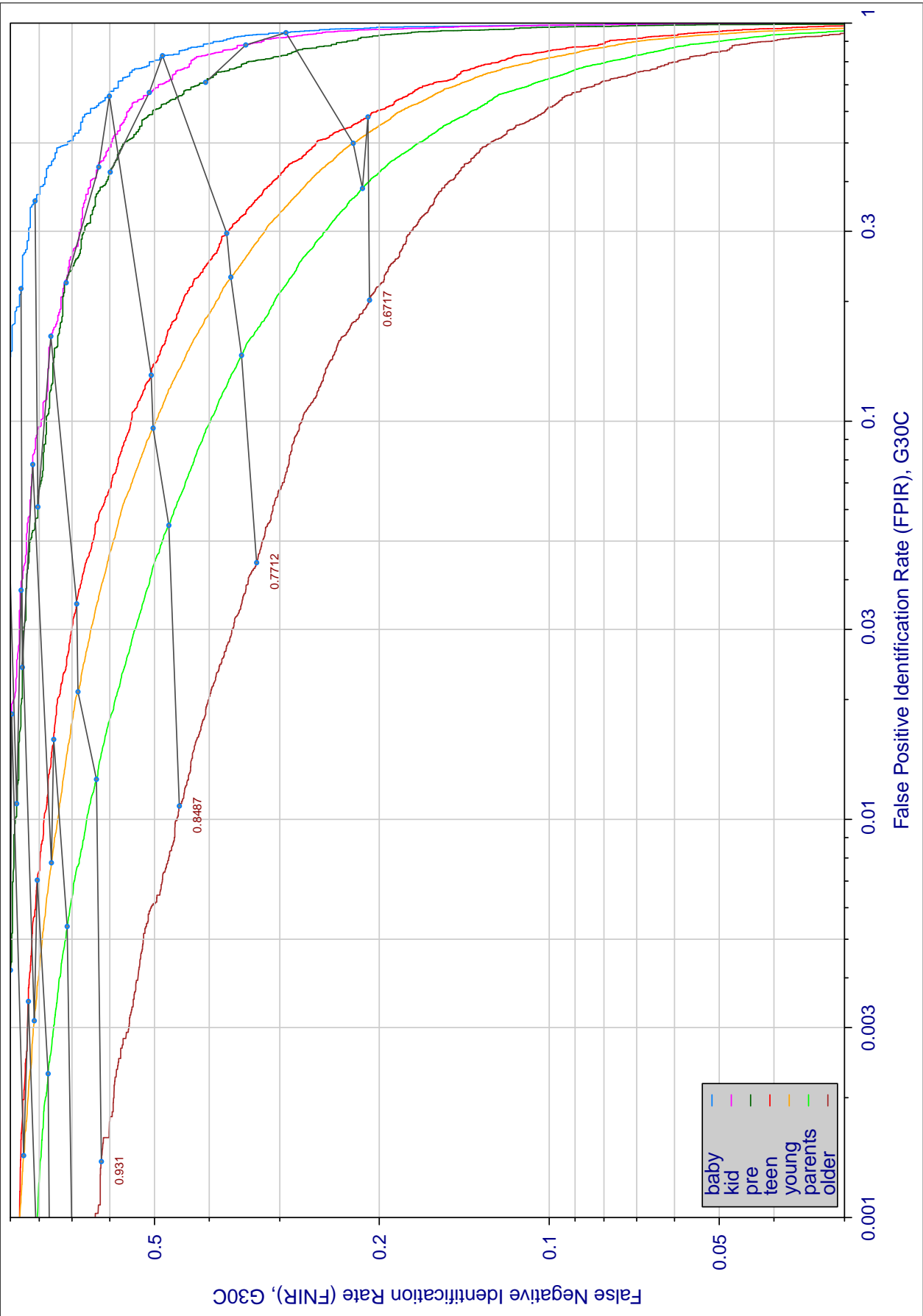


Figure 82: The effect of age on accuracy G30C. Each trace is a detection error tradeoff characteristic for individuals within one of seven age groups at the time their search image was collected. Upward vertical displacement indicates increased difficult for persons to be recognized as themselves. Rightward displacement indicates that individuals are more often confused a somebody else (i.e. false positives).

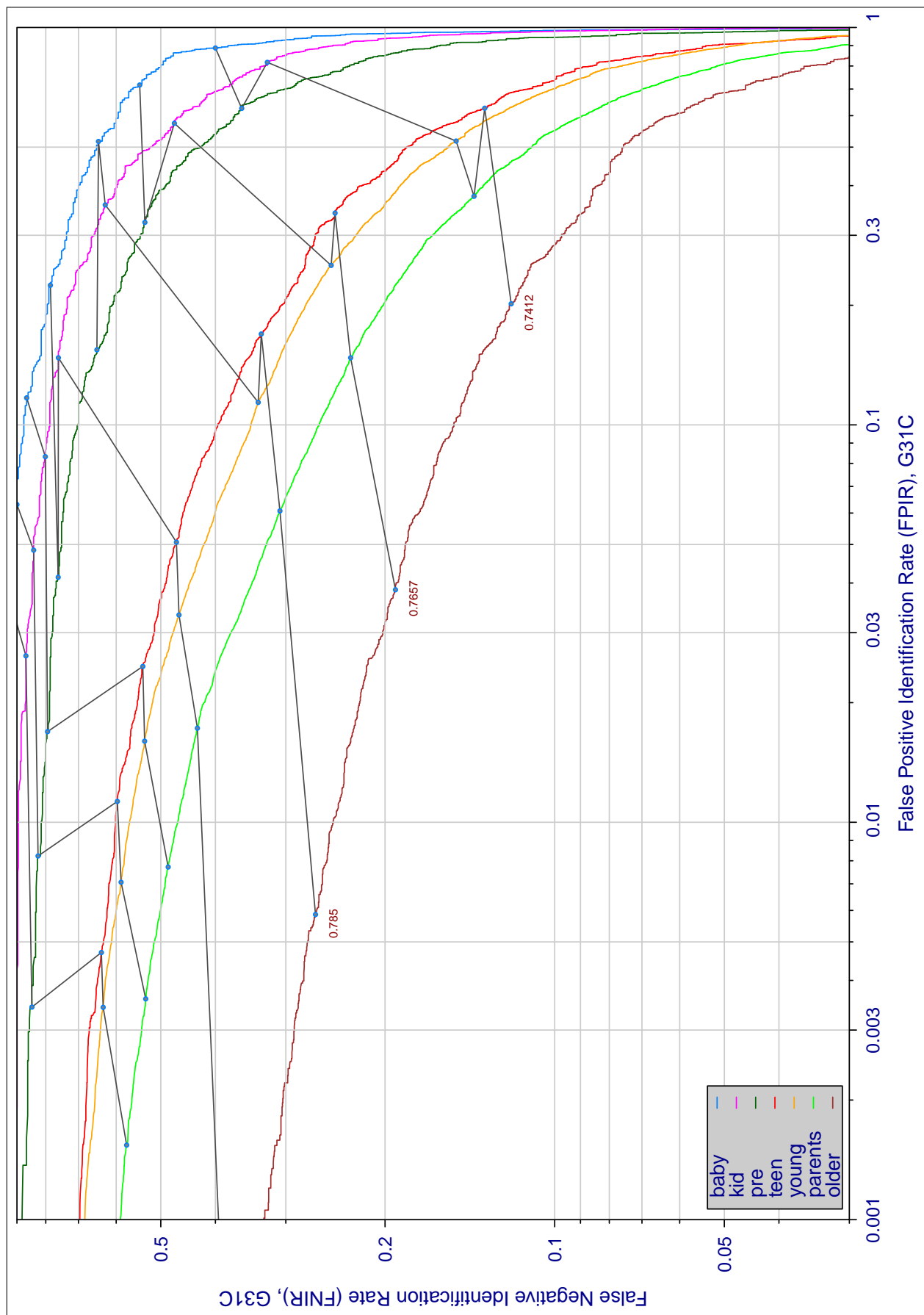


Figure 83: The effect of age on accuracy G31C. Each trace is a detection error tradeoff characteristic for individuals within one of seven age groups at the time their search image was collected. Upward vertical displacement indicates increased difficulty for persons to be recognized as themselves. Rightward displacement indicates that individuals are more often confused a somebody else (i.e. false positives).

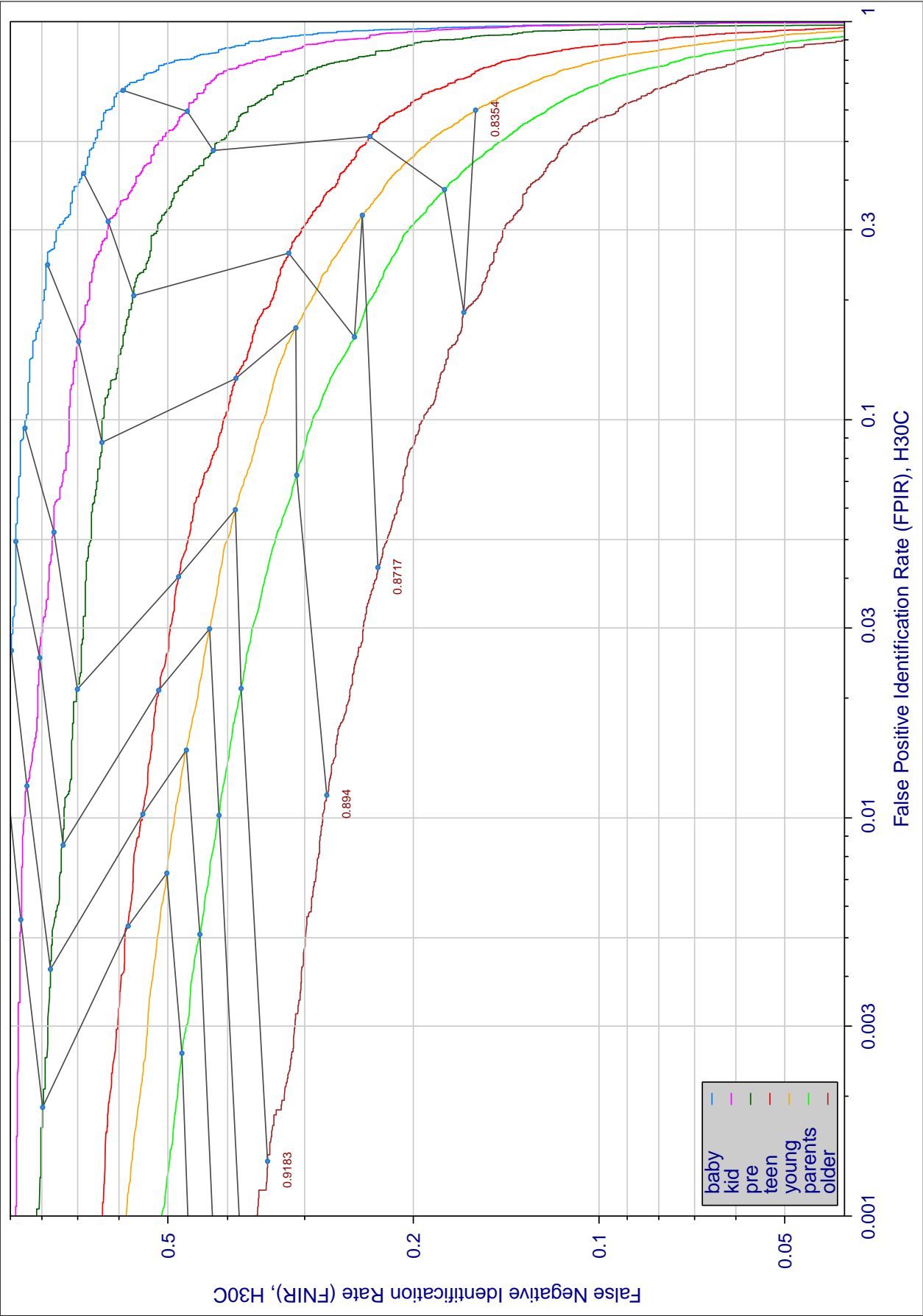


Figure 84: The effect of age on accuracy H30C. Each trace is a detection error tradeoff characteristic for individuals within one of seven age groups at the time their search image was collected. Upward vertical displacement indicates increased difficult for persons to be recognized as themselves. Rightward displacement indicates that individuals are more often confused a somebody else (i.e. false positives).

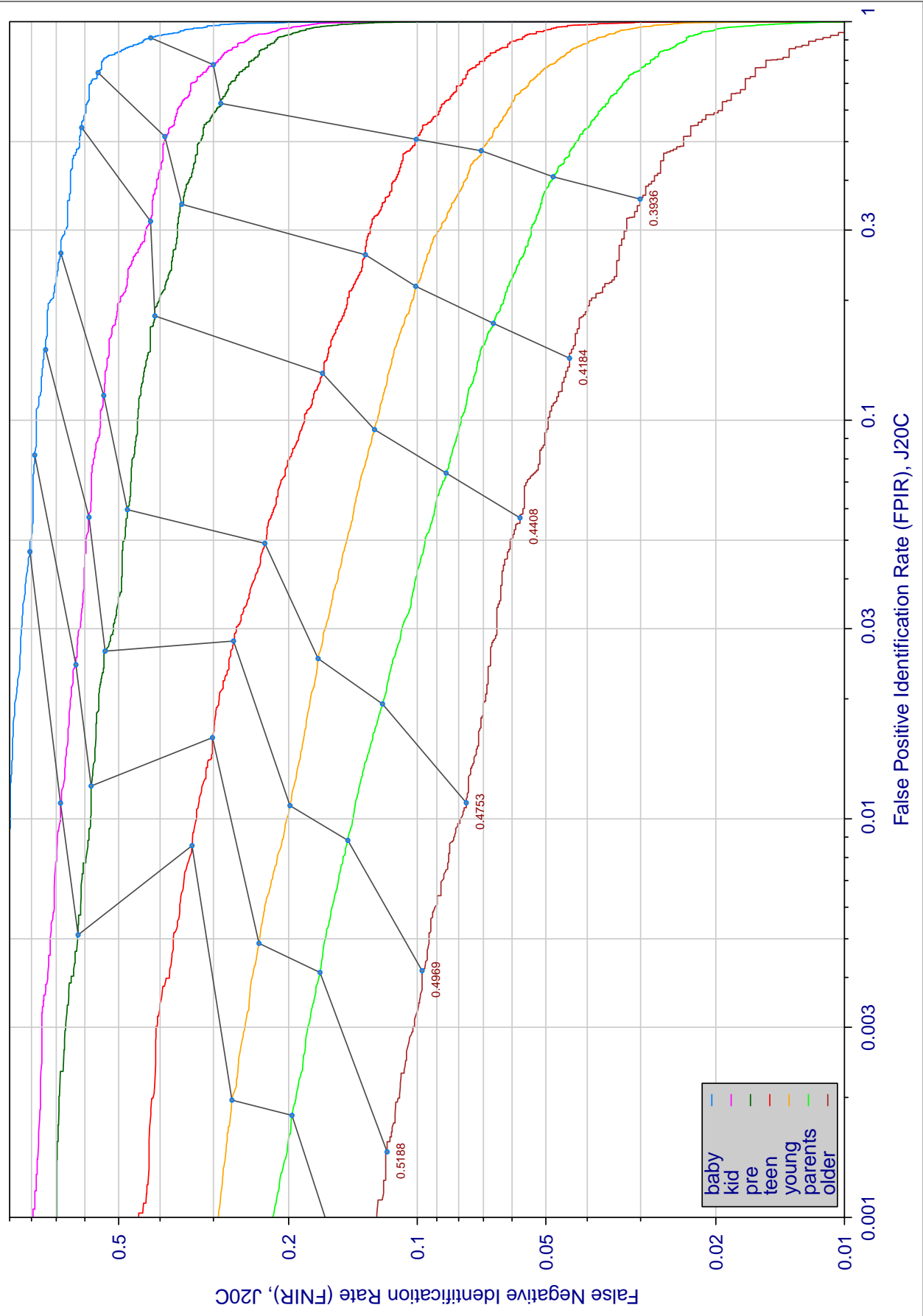


Figure 85: The effect of age on accuracy J20C. Each trace is a detection error tradeoff characteristic for individuals within one of seven age groups at the time their search image was collected. Upward vertical displacement indicates increased difficult for persons to be recognized as themselves. Rightward displacement indicates that individuals are more often confused a somebody else (i.e. false positives).

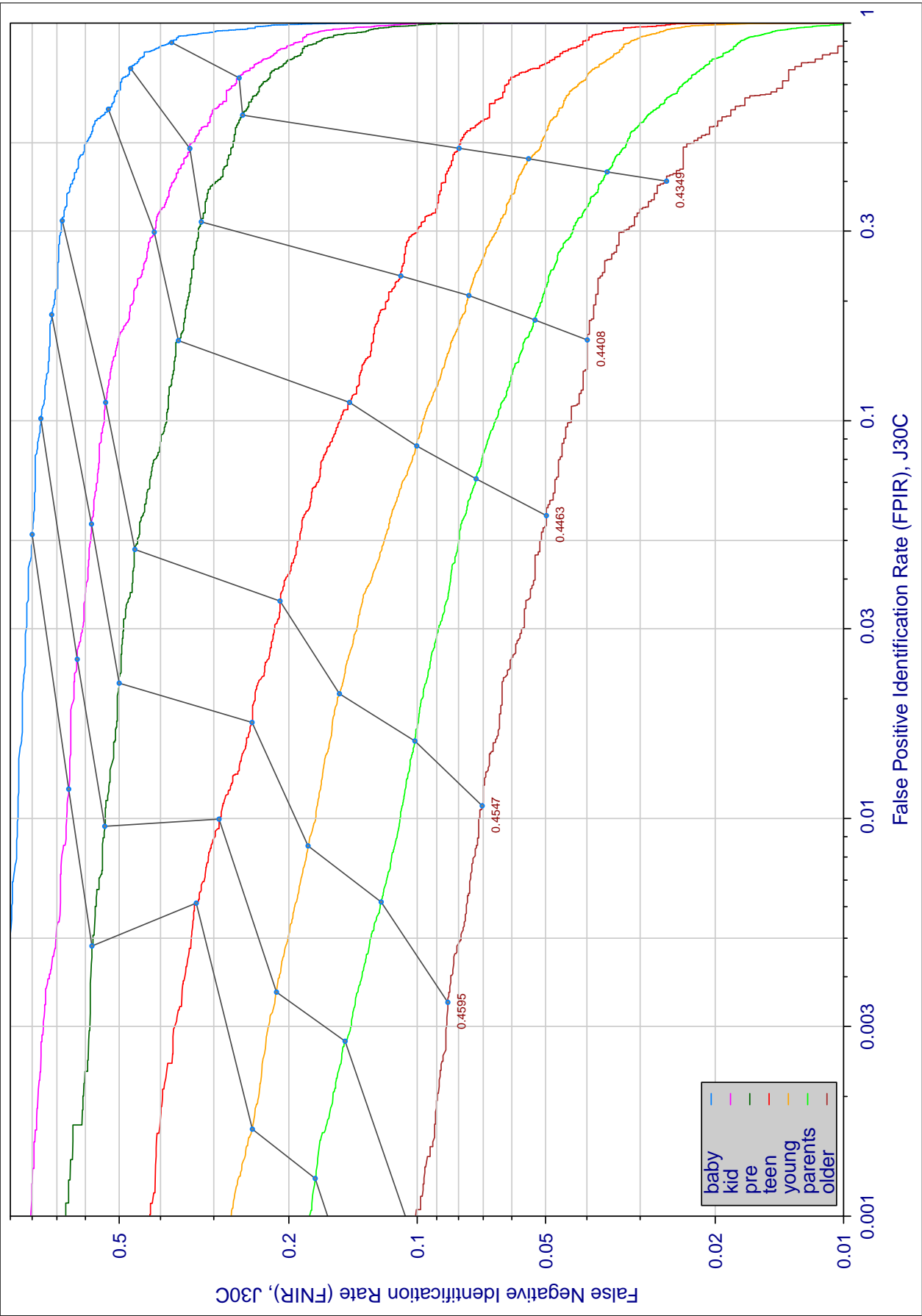


Figure 86: The effect of age on accuracy J30C. Each trace is a detection error tradeoff characteristic for individuals within one of seven age groups at the time their search image was collected. Upward vertical displacement indicates increased difficult for persons to be recognized as themselves. Rightward displacement indicates that individuals are more often confused a somebody else (i.e. false positives).

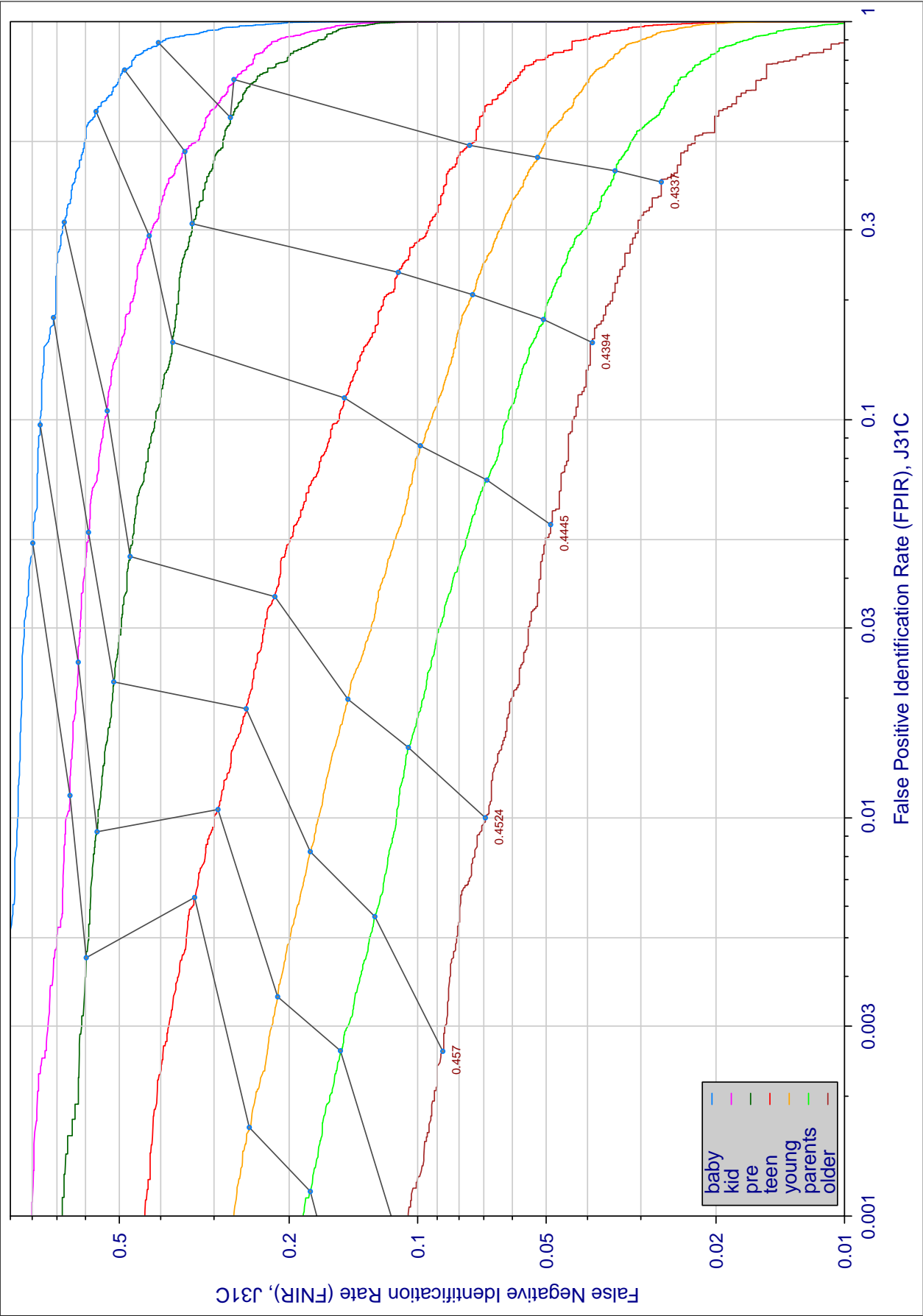


Figure 87: The effect of age on accuracy J31C. Each trace is a detection error tradeoff characteristic for individuals within one of seven age groups at the time their search image was collected. Upward vertical displacement indicates increased difficult for persons to be recognized as themselves. Rightward displacement indicates that individuals are more often confused a somebody else (i.e. false positives).

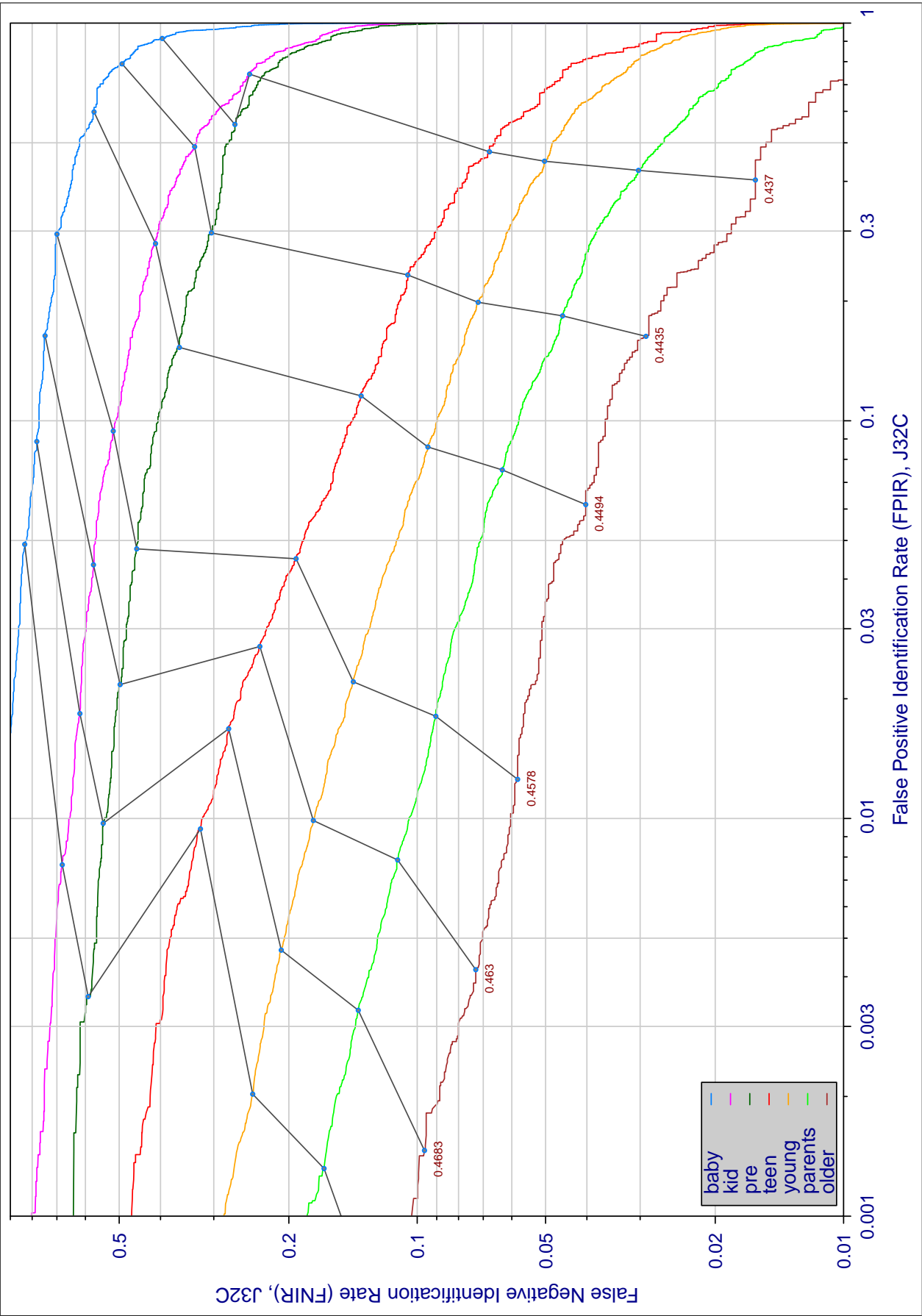


Figure 88: The effect of age on accuracy J32C. Each trace is a detection error tradeoff characteristic for individuals within one of seven age groups at the time their search image was collected. Upward vertical displacement indicates increased difficult for persons to be recognized as themselves. Rightward displacement indicates that individuals are more often confused a somebody else (i.e. false positives).

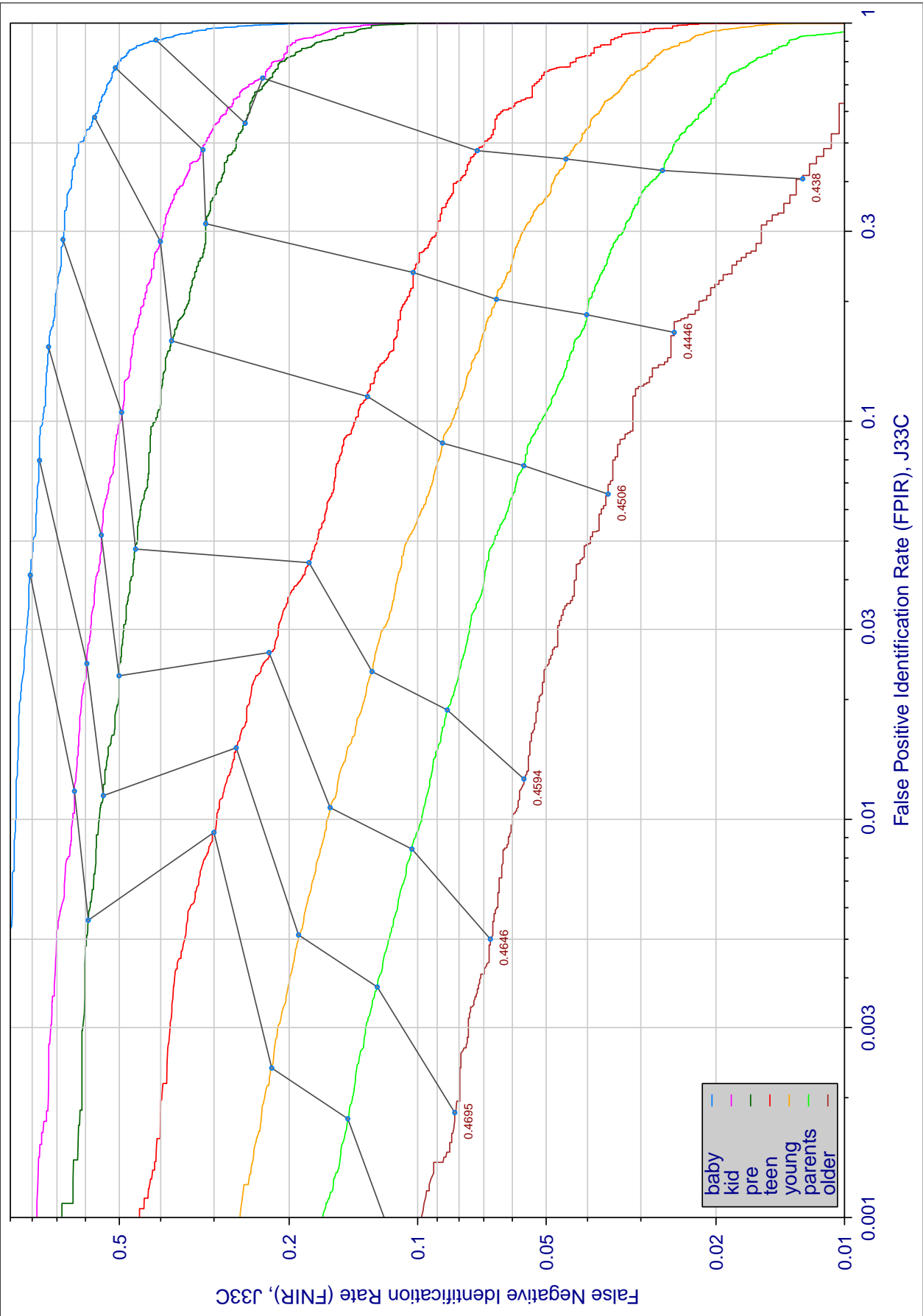


Figure 89: The effect of age on accuracy J33C. Each trace is a detection error tradeoff characteristic for individuals within one of seven age groups at the time their search image was collected. Upward vertical displacement indicates increased difficult for persons to be recognized as themselves. Rightward displacement indicates that individuals are more often confused a somebody else (i.e. false positives).

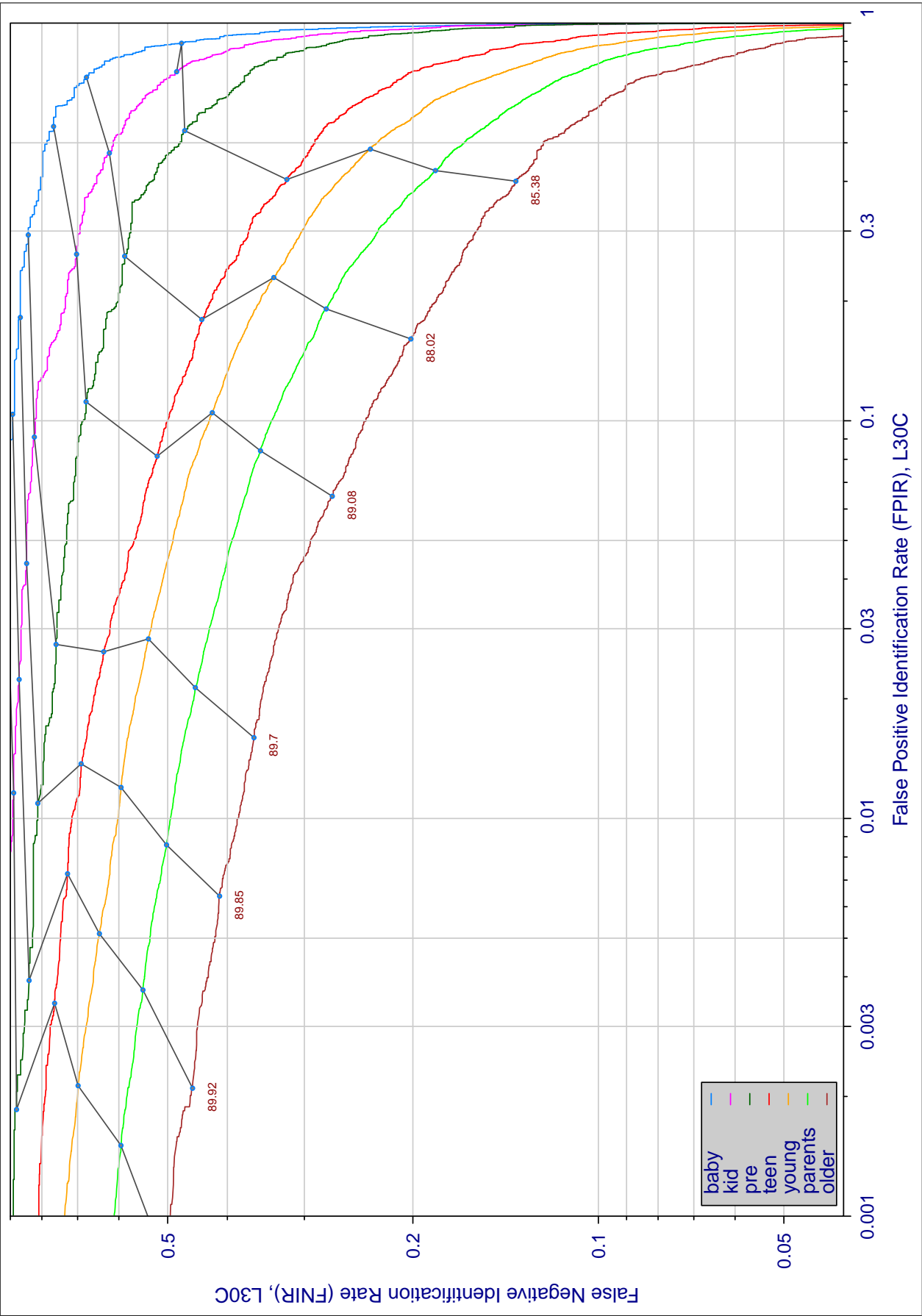


Figure 90: The effect of age on accuracy L30C. Each trace is a detection error tradeoff characteristic for individuals within one of seven age groups at the time their search image was collected. Upward vertical displacement indicates increased difficult for persons to be recognized as themselves. Rightward displacement indicates that individuals are more often confused a somebody else (i.e. false positives).

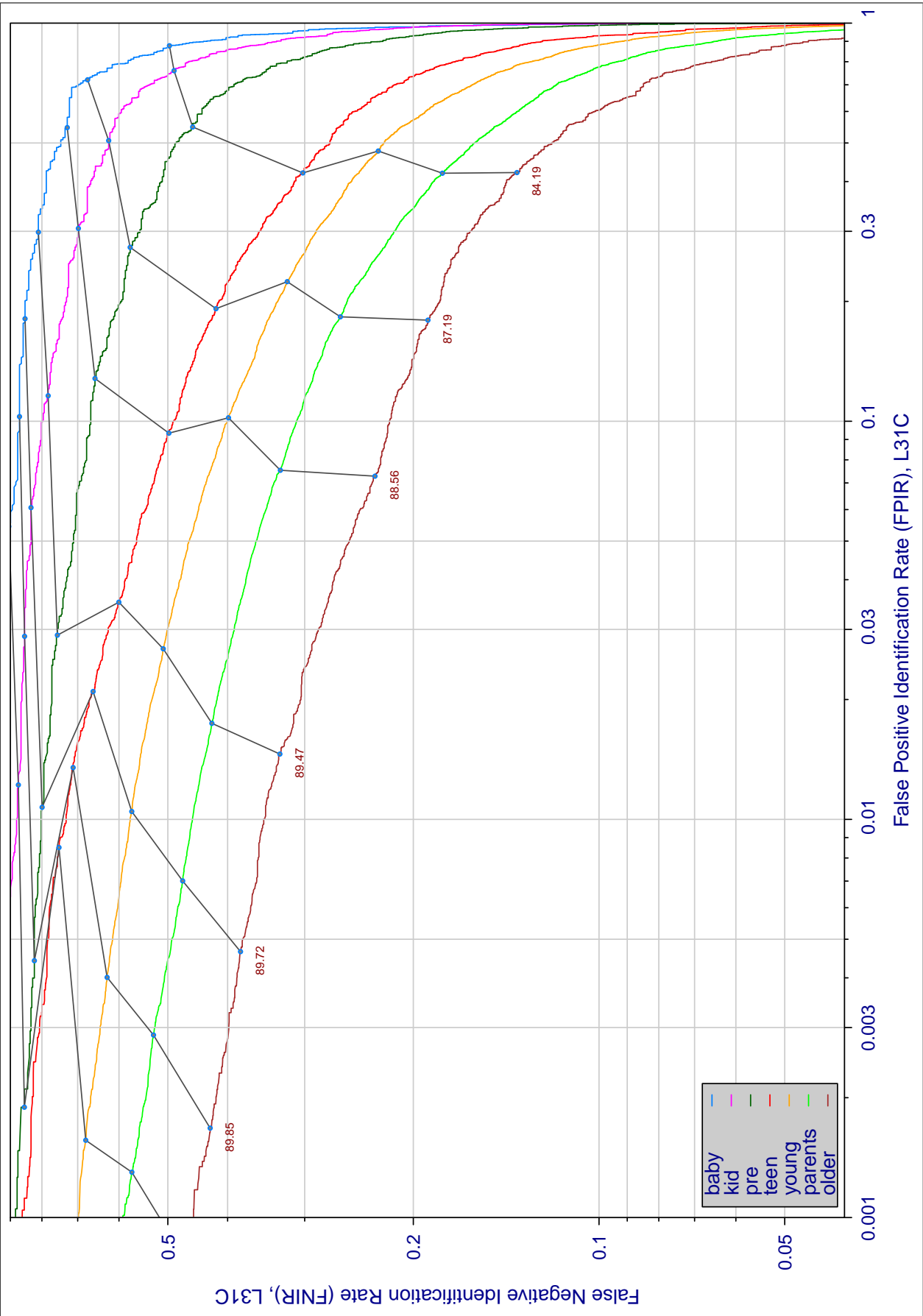


Figure 9I: The effect of age on accuracy L31C. Each trace is a detection error tradeoff characteristic for individuals within one of seven age groups at the time their search image was collected. Upward vertical displacement indicates increased difficult for persons to be recognized as themselves. Rightward displacement indicates that individuals are more often confused a somebody else (i.e. false positives).

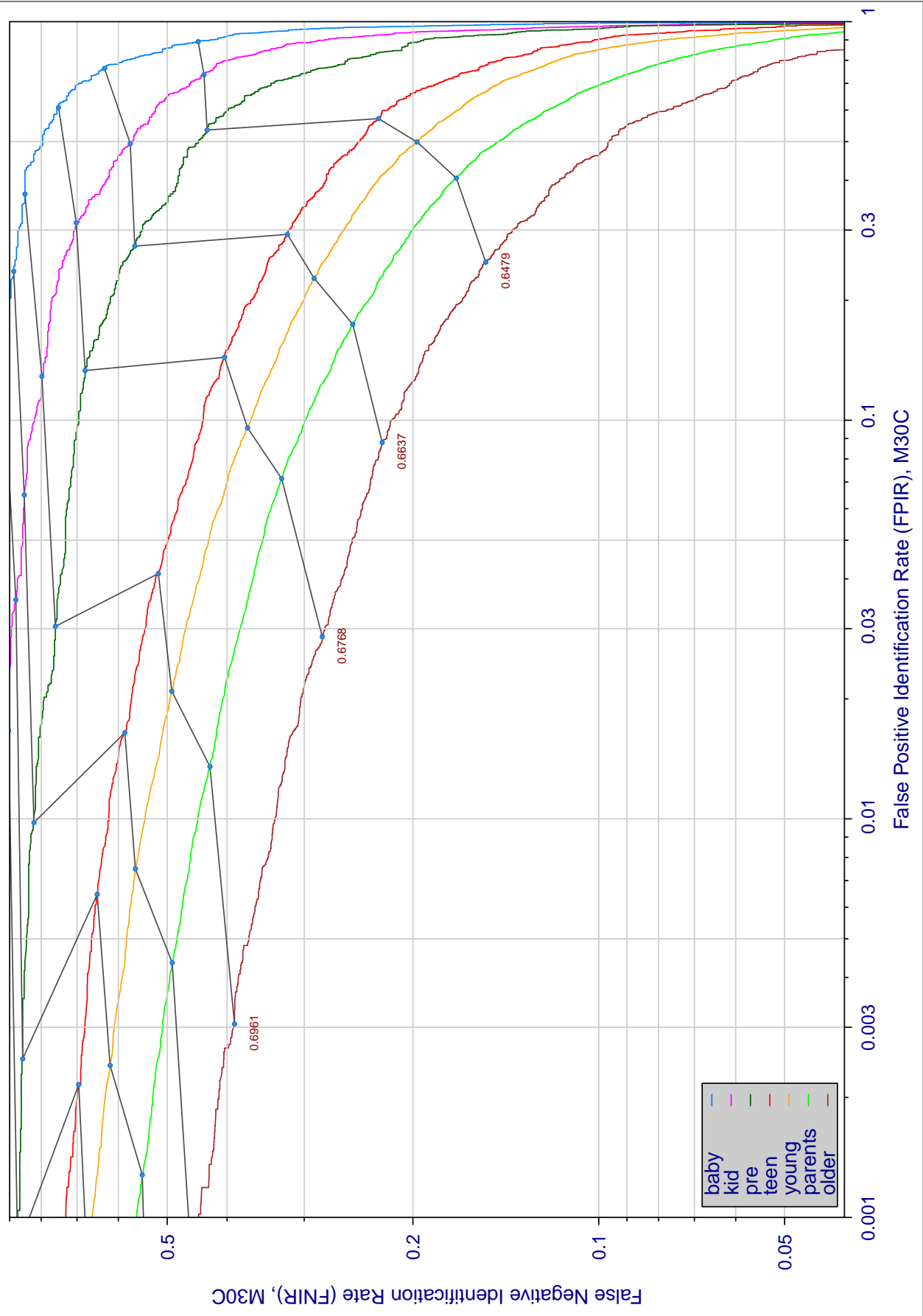


Figure 92: The effect of age on accuracy M30C. Each trace is a detection error tradeoff characteristic for individuals within one of seven age groups at the time their search image was collected. Upward vertical displacement indicates increased difficult for persons to be recognized as themselves. Rightward displacement indicates that individuals are more often confused a somebody else (i.e. false positives).

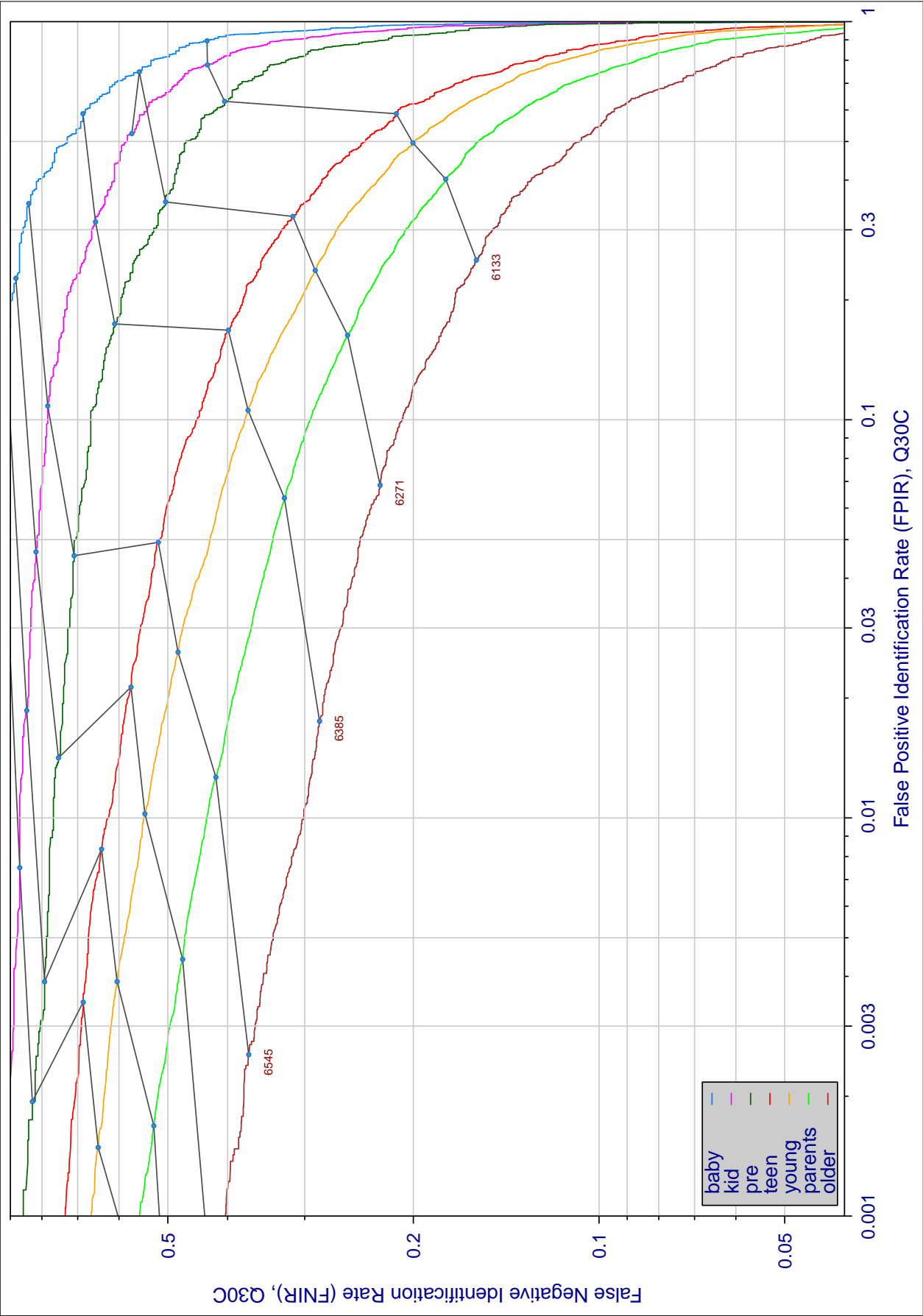


Figure 93: The effect of age on accuracy Q30C. Each trace is a detection error tradeoff characteristic for individuals within one of seven age groups at the time their search image was collected. Upward vertical displacement indicates increased difficult for persons to be recognized as themselves. Rightward displacement indicates that individuals are more often confused a somebody else (i.e. false positives).

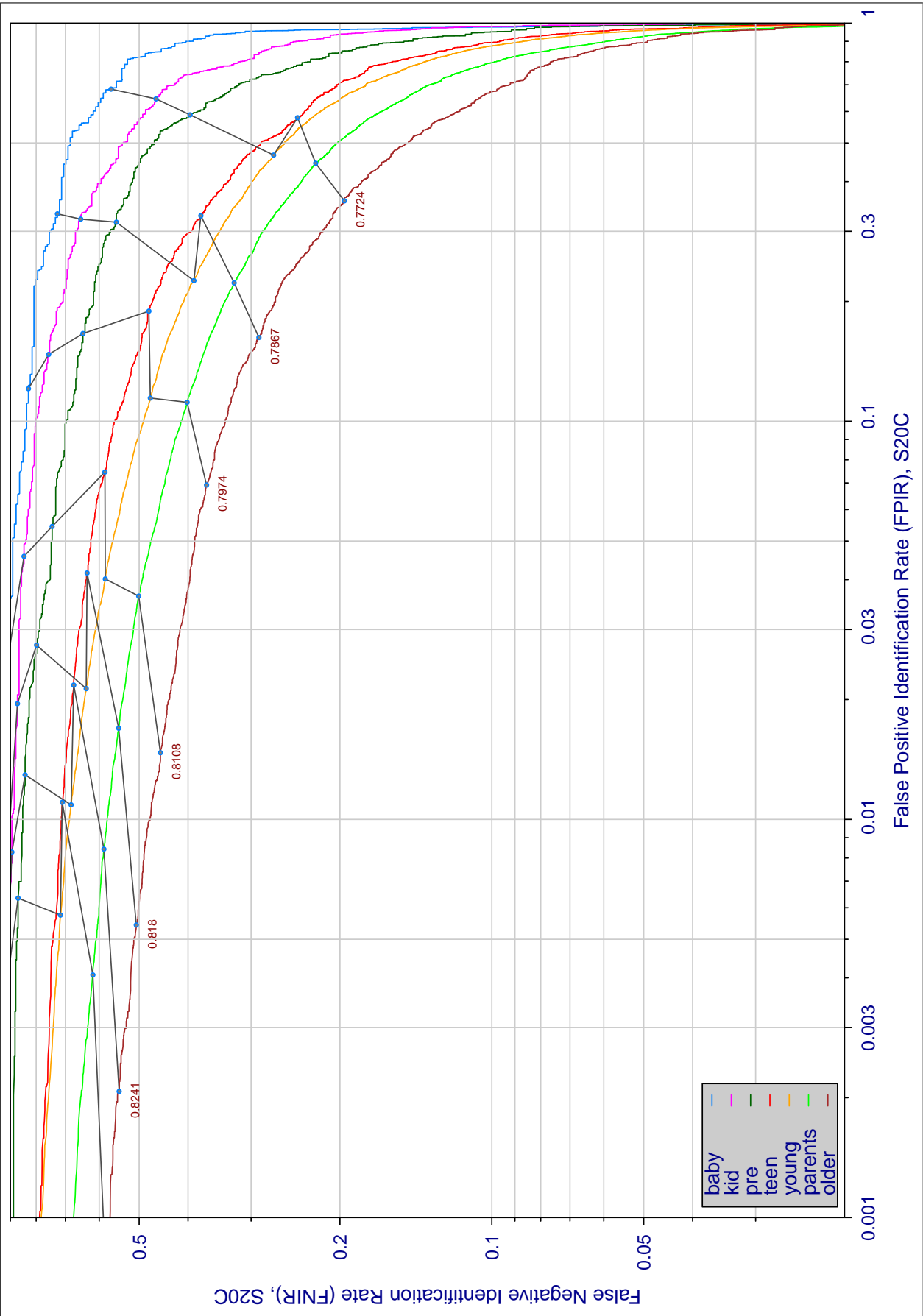


Figure 94: The effect of age on accuracy S20C. Each trace is a detection error tradeoff characteristic for individuals within one of seven age groups at the time their search image was collected. Upward vertical displacement indicates increased difficult for persons to be recognized as themselves. Rightward displacement indicates that individuals are more often confused a somebody else (i.e. false positives).

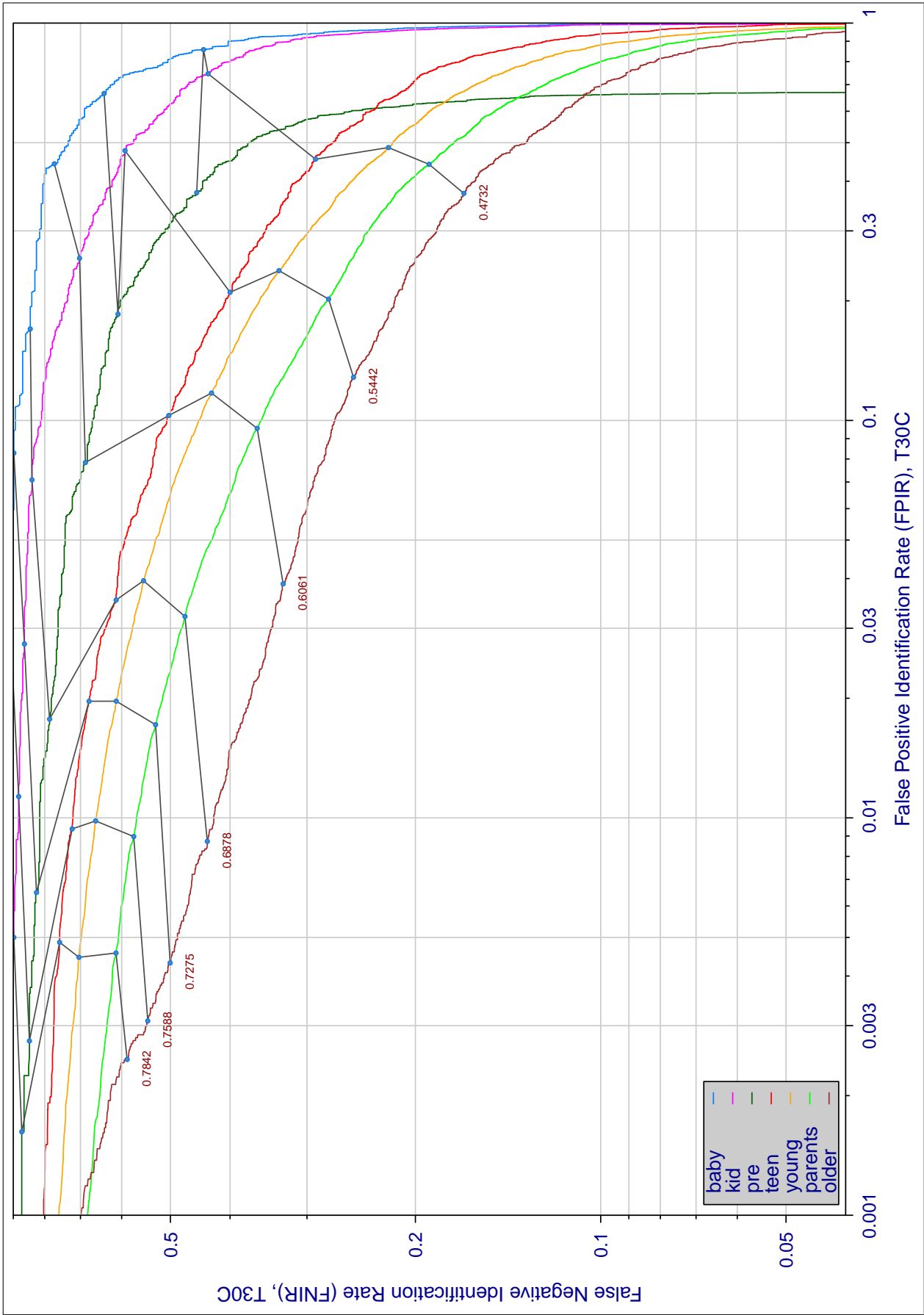


Figure 95: The effect of age on accuracy T30C. Each trace is a detection error tradeoff characteristic for individuals within one of seven age groups at the time their search image was collected. Upward vertical displacement indicates increased difficult for persons to be recognized as themselves. Rightward displacement indicates that individuals are more often confused a somebody else (i.e. false positives).