NISTIR 8197

The 2017 IARPA Face Recognition Prize Challenge (FRPC)

Patrick Grother
Mei Ngan
Kayee Hanaoka
Information Technology Laboratory, NIST

Chris Boehnen
Intelligence Advanced Research Projects Activity (IARPA)

Lars Ericson
Science Applications International Corporation (SAIC)

This publication is available free of charge from:
https://www.nist.gov/programs-projects/face-recognition-prize-challenge
https://doi.org/10.6028/NIST.IR.8197

November 2017

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

NISTIR 8197

The 2017 IARPA Face Recognition Prize Challenge (FRPC)

Patrick Grother
Mei Ngan
Kayee Hanaoka
Information Technology Laboratory, NIST

Chris Boehnen
Intelligence Advanced Research Projects Activity (IARPA)

Lars Ericson
Science Applications International Corporation (SAIC)

This publication is available free of charge from:
https://www.nist.gov/programs-projects/face-recognition-prize-challenge
https://doi.org/10.6028/NIST.IR.8197

November 2017

U.S. Department of Commerce
Wilbur Ross, Secretary

National Institute of Standards and Technology
Walter Copan, Director

ACKNOWLEDGEMENTS

The authors would like to thank the Intelligence Advanced Research Projects Activity for supporting this work, and for administering the $50 000 prize fund.

We are grateful to staff at Noblis for their development and curation of imagery used in this study.

Similarly we thank the staff of SAIC for collection of imagery used in this study. We thank the DHS S&T Homeland Security Advanced Research Agency Air Entry/Exit Re-engineering (AEER) Directorate for their support of that work.

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.

IDENTIFICATION IN TOP R RANKS FROM DATABASE OF SIZE N		VERIFICATION	
FNIR(N, R, T)	"False negative identification rate"	FNMR(T)	"False non-match rate"
FPIR(N, T)	"False positive identification rate"	FMR(T)	"False match rate"

EXECUTIVE SUMMARY

▷ *Overview:* This report documents NIST's execution of the Intelligence Advanced Research Projects Activity (IARPA) Face Recognition Prize Challenge (FRPC) 2017. The (FRPC) was conducted to assess the capability of contemporary face recognition algorithms to recognize faces in photographs collected without tight quality constraints, e.g. non-ideal images collected from individuals who are unaware of, and not cooperating with, the collection. Such images are characterized by variations in head orientation, facial expression, illumination, and also occlusion and reduced resolution.

▷ *Background:* Face recognition has recently been revolutionized by the availability of advanced machine learning algorithms, free software implementations thereof, fast processors, vast web-scraped ground-truthed face image databases, open performance benchmarks, and a vibrant academic literature for both machine learning and face recognition.

The new convolutional neural network technologies have largely been developed to exhibit invariance to the pose, illumination and expression variations characteristic in photojournalism and social media images. The initial research [7, 9] employed large numbers of images of relatively few ($O(10^4)$) individuals to learn invariance. Inevitably much larger populations ($O(10^7)$) were employed for training [8] but the benchmark remained verification at very high false match rates - LFW with an EER metric [3]. A large scale identification benchmark duly followed [6] yet its primary metric, rank one hit rate, contrasts with the high threshold discrimination task required in large-population governmental applications of face recognition, namely credential de-duplication, law enforcement and intelligence searches. There, only one or two images of at least $O(10^7)$ individuals must be recognized with very low false positive identification rates. The FRPC was conducted with both a large population ($O(10^6)$) and low false positive rate metrics.

From a field of 16 commercial and academic entries, the FRPC awarded prizes in three categories.

▷▷ *Verification accuracy*: The $20 000 prize is awarded to the algorithm that can most accurately verify the identity of faces appearing in photojournalism images. The verification task is the fundamental biometric operation - to determine whether two images are of the same face or not. The award criterion is to produce the lowest false non-match rate FNMR at a false match rate FMR of 0.001. The winner of this prize is **NTechLab**, which achieved FNMR = 0.22 well ahead of second place developer Yitu Technology.

▷▷ *Identification accuracy*: The $25 000 prize is awarded to the algorithm that can most accurately retrieve a face cropped from a video frame when searching a gallery composed of N = 691282 faces from cooperative portrait photos, while simultaneously producing a false positive outcome in only 1 in 1000 searches, i.e. to produce the lowest false negative identification rate (FNIR) at a false positive identification rate (FPIR) of 0.001. The winner of this prize is **Yitu Technology** whose algorithm produces superior FNIR values at the lower false positives rates required in one-to-many applications for which many searches do not have a corresponding enrolled entry.

▷▷ *Identification speed*: A $5 000 prize is awarded to the algorithm that executes a one-to-many search in the shortest possible time, and still has high accuracy. The formal criterion is to produce the lowest median duration when executing searches while also producing FNIR less than two times that of the most accurate algorithm. The winner is **NTechLab**, one of whose algorithms returns candidates from a gallery of N = 691282 identities, in just 590 ± 50 microseconds. This is achieved using one process running on a single core of a conventional c. 2016 server-class CPU. This is accompanied by sub-linear search time, such that a 30-fold increase in the gallery size N only incurs a 3-fold increase in search duration. This is achieved, however, using a proprietary fast-search data structure that takes almost 11 hours to build from N = 691282 input templates.

Readers might also consider reports from NIST's Face Recognition Vendor Test (FRVT) which remains open to new algorithm developers. Comments and questions on FRPC and FRVT should be directed to frpc@nist.gov and frvt@nist.gov, respectively.

IDENTIFICATION IN TOP R RANKS FROM DATABASE OF SIZE N		VERIFICATION	
FNIR(N, R, T)	"False negative identification rate"	FNMR(T)	"False non-match rate"
FPIR(N, T)	"False positive identification rate"	FMR(T)	"False match rate"

Contents

List of Tables

List of Figures

IDENTIFICATION IN TOP R RANKS FROM DATABASE OF SIZE N		VERIFICATION	
FNIR(N, R, T)	"False negative identification rate"	FNMR(T)	"False non-match rate"
FPIR(N, T)	"False positive identification rate"	FMR(T)	"False match rate"

1 The verification prize challenge

Background: Verification is perhaps the most common application of biometrics, being widely deployed in applications such as access control and authentication. While such uses usually involve cooperative subjects, the FRPC includes a verification task using non-cooperative and unconstrained photojournalism imagery because one-to-one comparison of single images present the simplest way to assess core algorithmic efficacy.

	Developer	Config[1]	Template		GPU	Comparison Time (ns)[3]	
	Name	Data (KB)	Size (B)	Time (ms)[2]		Genuine	Impostor
1	3DiVi	190867	[13]4096 ± 0	[14]1236 ± 51	No	[1]499 ± 14	[2]501 ± 23
2	Ayonix	58505	[9]1036 ± 0	[1]18 ± 3	No	[2]613 ± 26	[3]621 ± 34
3	CyberExtruder	121469	[3]256 ± 0	[11]889 ± 22	No	[4]993 ± 17	[5]988 ± 16
4	Deep Sense	354779	[8]1028 ± 0	[9]541 ± 7	No	[3]647 ± 16	[4]646 ± 28
5	Digital Barriers	209340	[12]2056 ± 0	[6]200 ± 1	No	[12]12423 ± 204	[13]12426 ± 163
6	HB Innovation	273006	[5]520 ± 0	[8]298 ± 19	No	[11]5283 ± 484	[12]4888 ± 71
7	Imperial College London	274821	[11]2048 ± 0	[15]1367 ± 10	Yes	[6]1911 ± 51	[7]1888 ± 42
8	Innovatrics	0	[4]276 ± 0	[4]152 ± 12	Yes	[10]4002 ± 77	[11]3665 ± 126
9	Morpho	794266	[6]788 ± 0	[7]254 ± 5	Yes	[7]3112 ± 63	[9]3171 ± 126
10	Neurotechnology	413202	[15]4780 ± 0	[16]1560 ± 44	No	[16]73520 ± 1921	[16]72674 ± 1429
11	NTechLab	657997	[16]4825 ± 1	[13]943 ± 16	No	[15]55004 ± 80	[15]55042 ± 93
12	Rank One	0	[1]144 ± 0	[2]37 ± 1	No	[13]30366 ± 177	[1]307 ± 41
13	Smilart UG	107947	[7]1024 ± 0	[3]58 ± 0	Yes	[9]3443 ± 66	[10]3442 ± 69
14	VisionLabs	343661	[2]204 ± 0	[12]943 ± 8	No	[5]1013 ± 40	[6]1030 ± 34
15	Vocord	918293	[10]1280 ± 0	[5]195 ± 0	Yes	[8]3271 ± 94	[8]2413 ± 99
16	Yitu	2226850	[14]4136 ± 0	[10]703 ± 1	No	[14]33991 ± 62	[14]34048 ± 134

Notes	
1	The size of configuration data does not capture static data included in the libraries. We do not include the size of the libraries because some algorithms include common ancilliary libraries for image processing (e.g. openCV) or numerical computation (e.g. blas).
2	The median template creation times are measured on Intel®Xeon®CPU E5-2630 v4 @ 2.20GHz processors or, in the case of GPU-enabled implementations, NVidia Tesla K40m equipped with 12GB of memory.
3	The median comparison durations, in nanoseconds, are estimated using std::chrono::high_resolution_clock which on the machine in (2) counts clock ticks of duration 1 nanosecond. Precision is somewhat worse than that however. The ± value is the median absolute deviation times 1.48 to give consistency with 1σ of a Normal distribution.

Table 1: Summary of 1:1 verifications algorithms evaluated in this report. The red superscripts give ranking for the quantity in that column.

Participation: The participants electing to submit algorithms to the FRPC verification track are listed in Table 1.

Images: The photojournalism set uses 141331 faces images of 3548 adults. The images are closely cropped from the parent images as shown in Figure 1. The images are primarily collected by professional photographers and as such are captured, and selected, to not exhibit exposure and focus problems. All of the images are live capture, none are scanned. Resolution varies widely as these images were posted to the internet with varying resampling and compression practices. The primary difficulties for face recognition is unconstrained yaw and pitch pose variation, with some images extending to profile view. Additionally faces can be occluded, including by hair and hands.

The images are cropped prior to passing them to the algorithm. The cropping is done per human-annotated rectangular bounding boxes. The algorithm must further localize the face and extract features, returning a recognition template. The templates from the images are used in $N_G = 7\,846\,208$ genuine and $N_I = 39\,942\,674$ impostor comparisons. The impostor trials are zero-effort, meaning any template is compared with any other template - no effort is made to pair on such variables as sex, age, race or appearance. While zero-effort impostors are easier to correctly reject, the technique is ubiquitous when assessing core recognition accuracy.

Accuracy metrics: Scores from the genuine comparisons are used in the false non-match rate (FNMR) computation, which states the proportion of genuine scores below threshold, T:

$$\mathrm{FNMR}(T) = 1 - \frac{1}{N_G}\sum_{i=1}^{N_G} H(s_i - T) \tag{1}$$

IDENTIFICATION IN TOP R RANKS FROM DATABASE OF SIZE N		VERIFICATION	
FNIR(N, R, T)	"False negative identification rate"	FNMR(T)	"False non-match rate"
FPIR(N, T)	"False positive identification rate"	FMR(T)	"False match rate"

Figure 1: Examples of "in the wild" photojournalism stills. The top row gives the full original images; the second row gives the manually specified face region that is cropped and passed to the algorithms. The source images in this figure are published on the internet under Creative Commons licenses.

where the step function H(x) is 1 if $x \geq 0$ and 0 otherwise. In cases where an algorithm fails to produce a template from an input image - the so-called failure to enroll outcome - the FNMR computation proceeds by assigning a low score, $-\infty$, to any comparison involving that template. This simulates false rejection of a user.

Scores from the impostor comparisons are used in the false match rate (FMR) computation, which states the proportion of impostor scores at or above T:

$$\text{FMR}(T) = \frac{1}{N_I} \sum_{i=1}^{N_I} H(s_i - T) \tag{2}$$

In cases where an algorithm fails to produce a template from an input image, a low score is again assigned as the result of any comparison involving that template. This practice actually benefits (reduces) FMR.

Figure of Merit: The prize is awarded to the algorithm that achieves the lowest false non-match rate at a threshold set to achieve a false match rate of 0.001. This is the most common way to state recognition accuracy, and it serves as a simple way to compare core algorithm recognition capability.

Prize winner: By consulting Figure 2, the most accurate verification algorithm on this dataset is developed by NTechLab http://ntechlab.com/.

Discussion: The NTechLab algorithm gives FNMR = 0.22 with FMR = 0.001. This FNMR would be intolerably high for an access control application, but is achieved with images of non-cooperating subjects that have very few of the image quality constraints that are engineered into, for example, border crossing gates. In particular, as discussed later in section 4, the winning algorithm here has superior capability at recognizing individuals whose head orientations vary widely.

IDENTIFICATION IN TOP R RANKS FROM DATABASE OF SIZE N		VERIFICATION	
FNIR(N, R, T)	"False negative identification rate"	FNMR(T)	"False non-match rate"
FPIR(N, T)	"False positive identification rate"	FMR(T)	"False match rate"

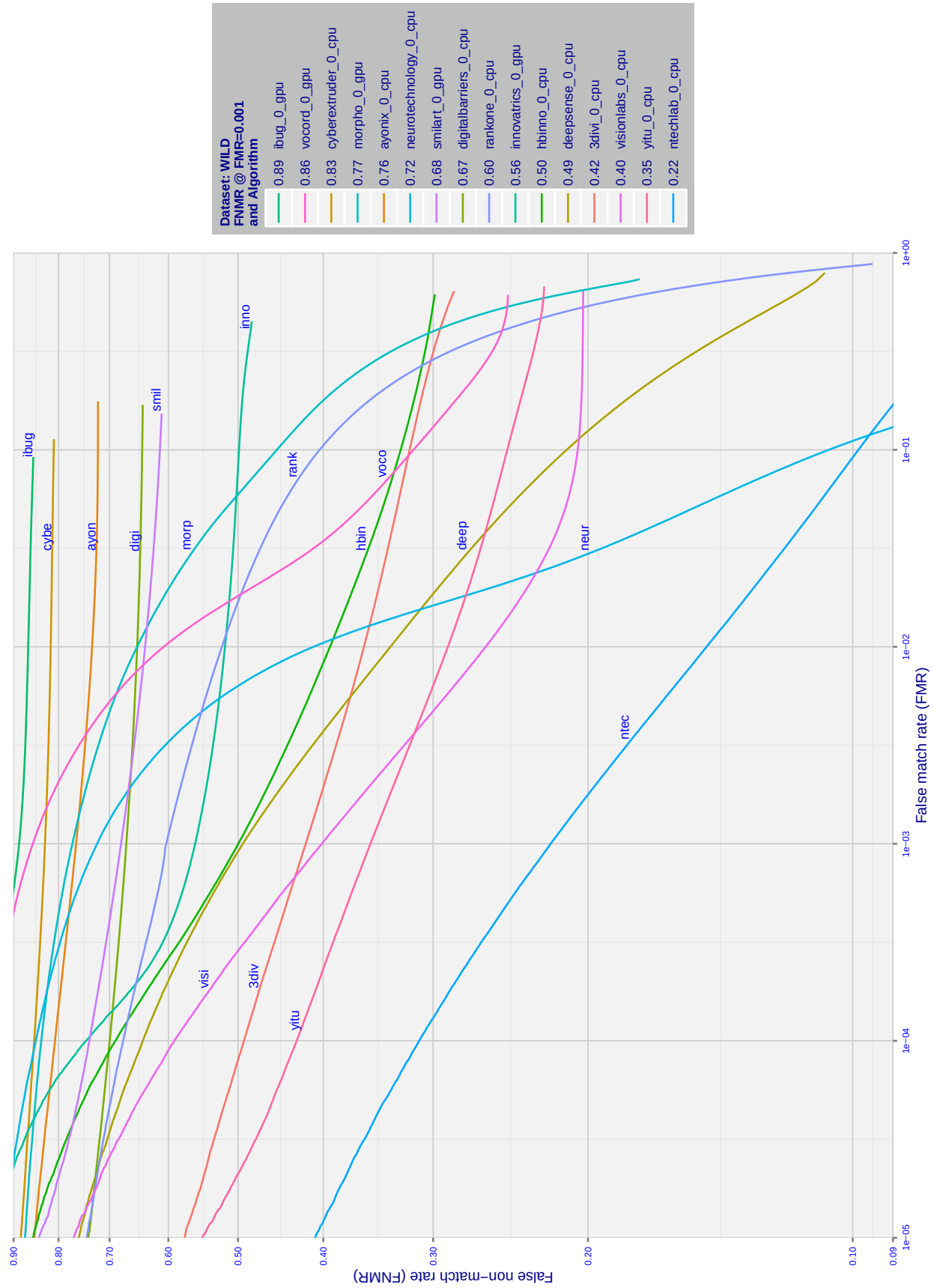


Figure 2: For the FRPC *verification algorithms, the graph shows show error tradeoff characteristics for comparisons of wild images.*

IDENTIFICATION IN TOP R RANKS FROM DATABASE OF SIZE N		VERIFICATION	
FNIR(N, R, T)	"False negative identification rate"	FNMR(T)	"False non-match rate"
FPIR(N, T)	"False positive identification rate"	FMR(T)	"False match rate"

	Developer	Seq.	GPU	Config[1]	Template		Impostor Search Time (microseconds)[3]				
	Name	No.		Data (KB)	Size (B)	Time (ms)[2]	N = 16000	N = 48000	N = 160000	N = 320000	N = 691282
1	3DiVi[4]	0	No	190867	[18]4096 ± 0	[20]1121 ± 54	[12]5223 ± 10	[12]15709 ± 10	[12]52411 ± 34	[12]106109 ± 64	[12]229400 ± 851
2	3DiVi[4]	0	Yes	190867	[17]4096 ± 0	[7]153 ± 52	[15]6853 ± 10	[15]20788 ± 17	[13]69584 ± 57	[13]138443 ± 78	[13]302072 ± 254
3	3DiVi[4]	1	No	191636	[20]4224 ± 0	[21]1121 ± 55	[3]285 ± 40	[3]818 ± 109	[4]2852 ± 332	[3]6792 ± 593	[3]14452 ± 1768
4	3DiVi[4]	1	Yes	191636	[21]4224 ± 0	[8]153 ± 52	[4]395 ± 52	[4]921 ± 251	[3]2798 ± 409	[4]7249 ± 601	[4]16001 ± 1456
5	CyberExtruder	0	No	121469	[4]256 ± 0	[17]793 ± 107	[11]4722 ± 5	[11]14157 ± 10	[11]48199 ± 112	[11]96578 ± 95	[11]208519 ± 132
6	CyberExtruder	1	No	89921	[1]128 ± 0	[15]515 ± 111	[10]4711 ± 4	[10]14039 ± 18	[10]47197 ± 82	[10]96130 ± 71	[10]205348 ± 127
7	Deep Sense	0	No	354779	[11]1028 ± 0	[13]498 ± 10	[13]6240 ± 59	[13]19998 ± 227	[15]69789 ± 925	[15]143146 ± 2179	[14]317479 ± 4647
8	Deep Sense	1	No	354779	[10]1028 ± 0	[14]499 ± 9	[14]6243 ± 59	[14]20038 ± 255	[14]69744 ± 989	[14]142888 ± 2002	[15]317877 ± 4708
9	Digital Barriers	0	No	209216	[14]2056 ± 0	[6]90 ± 28	[23]53198 ± 293	[22]164269 ± 617	[22]552162 ± 2051	[21]1089861 ± 2954	[21]2377165 ± 26073
10	Digital Barriers	1	No	209216	[15]2056 ± 0	[5]89 ± 27	[24]54102 ± 291	[23]168442 ± 1197	[23]574761 ± 2051	[22]1122960 ± 6583	[22]2474337 ± 29008
11	Imperial College London	0	Yes	274821	[13]2048 ± 0	[22]1365 ± 10	[17]11670 ± 404	[17]35771 ± 285	[17]128969 ± 300	[17]261700 ± 462	[17]606752 ± 1295
12	Innovatrics	0	No	0	[2]138 ± 0	[19]944 ± 33	[6]1213 ± 1	[5]3629 ± 2	[6]12172 ± 13	[6]24502 ± 22	[6]52994 ± 32
13	Innovatrics	1	Yes	0	[5]276 ± 0	[4]84 ± 21	[9]2259 ± 91	[9]6254 ± 182	[7]20328 ± 346	[7]40674 ± 533	[7]86716 ± 812
14	Morpho	0	Yes	794266	[7]788 ± 0	[10]244 ± 5	[8]1590 ± 7	[7]5556 ± 106	[9]28156 ± 94	[9]56673 ± 192	[9]126932 ± 303
15	Morpho	1	Yes	198517	[6]404 ± 0	[2]75 ± 4	[7]1302 ± 2	[6]4140 ± 16	[8]20963 ± 70	[8]42731 ± 229	[8]92624 ± 257
16	Neurotechnology	0	No	413202	[23]4780 ± 0	[23]1387 ± 77	[22]52081 ± 65	[24]173035 ± 232	[24]772830 ± 2145	[24]2092898 ± 10433	[24]7340539 ± 31367
17	Neurotechnology	1	No	413202	[22]4780 ± 0	[24]1391 ± 76	[21]25648 ± 75	[21]99376 ± 281	[21]539105 ± 853	[23]1651293 ± 2866	[23]6471346 ± 19941
18	NTechLab	0	No	875851	[24]5784 ± 1	[16]626 ± 21	[16]7912 ± 38	[16]24932 ± 98	[16]91830 ± 418	[16]192315 ± 868	[16]433640 ± 1840
19	NTechLab	1	No	288973	[9]987 ± 0	[11]361 ± 21	[1]208 ± 13	[1]344 ± 47	[1]508 ± 56	[1]558 ± 50	[1]592 ± 51
20	Rank One	0	No	0	[3]144 ± 0	[1]70 ± 26	[5]804 ± 298	[8]5729 ± 1214	[5]12165 ± 4110	[5]17908 ± 5222	[5]32512 ± 9976
21	Vocord	0	Yes	918293	[12]1280 ± 0	[9]191 ± 4	[19]21738 ± 20	[19]66094 ± 61	[19]219715 ± 120	[19]438762 ± 226	[19]947782 ± 467
22	Vocord	1	Yes	1089798	[8]896 ± 0	[3]78 ± 5	[18]15224 ± 13	[18]46034 ± 42	[18]153935 ± 128	[18]307279 ± 217	[18]664020 ± 388
23	Yitu	0	No	2226850	[19]4136 ± 0	[18]844 ± 18	[20]25641 ± 40	[20]77823 ± 257	[20]286455 ± 9072	[20]1071714 ± 3395	[20]1320849 ± 17367
24	Yitu	1	No	2262178	[16]2260 ± 0	[12]436 ± 11	[2]270 ± 57	[2]506 ± 128	[2]2144 ± 622	[2]5006 ± 1558	[2]11885 ± 3663

Notes	
1	The size of configuration data does not capture static data included in the libraries. We do not include the size of the libraries because some algorithms include common ancilliary libraries for image processing (e.g. openCV) or numerical computation (e.g. blas).
2	The median template creation times are measured on Intel®Xeon®CPU E5-2630 v4 @ 2.20GHz processors or, in the case of GPU-enabled implementations, NVidia Tesla K40m equipped with 12GB of memory.
3	The median impostor search durations, in milliseconds, are estimated using std::chrono::high_resolution_clock which on the machine in (2) counts clock ticks of duration 1 nanosecond. Precision is somewhat worse than that however. The ± value is the median absolute deviation times 1.48 to give consistency with 1σ of a Normal distribution.
4	Four entries appear for 3DiVi who NIST asked to submit a CPU variant of their main GPU submission. This allowed NIST to expedite testing. The report includes timing results for both CPU and GPU variants. Accuracy numbers are included only once, as accuracy is identical for CPU and GPU implementations.

Table 2: Summary of 1:N identification algorithms evaluated in this report. The red superscripts give ranking for the quantity in that column.

IDENTIFICATION IN TOP R RANKS FROM DATABASE OF SIZE N	VERIFICATION
FNIR(N, R, T) "False negative identification rate"	FNMR(T) "False non-match rate"
FPIR(N, T) "False positive identification rate"	FMR(T) "False match rate"

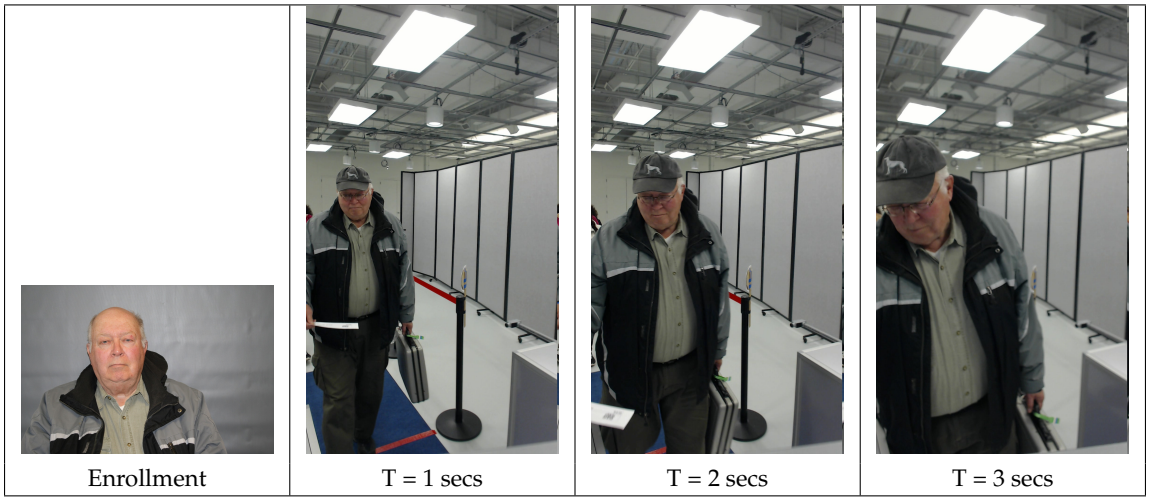


Figure 3: Enrollment (left) and non-cooperative video-frame search examples from a boarding gate process. The algorithm received the enrollment image as is, and faces cropped from the video search frames. The images are from subject 79195746 in the DHS/ S&T AEER dataset. He consented to release of his images in public reports. For those individuals who did not consent to publication, their faces were masked (yellow circles).

2 The identification prize challenge

Background: This section documents the one-to-many identification experiments performed under FRPC. Generically, one-to-many biometric identification is more difficult than one-to-one verification because a search of an N person database must either correctly reject either $N - 1$ or N identities depending, respectively, on whether the search has a mated enrollment or not. Given its difficulty, and implied computational expense, identification algorithms are by far the largest revenue segment of the face recognition marketplace.

Participation: The participants electing to submit algorithms to the FRPC identification track are listed in Table 2.

Experimental design: The identification experiments proceeds by searching non-cooperative face images against enrollment galleries built from cooperative portrait images.

- ▷ **Enrolled portraits:** The portrait images are either visa images, mugshot images, or dedicated portraits collected from test subjects. These were collected typically using an SLR camera, ample two point light, and a standard uniform grey background. We defined five galleries containing, respectively, N = $\{16000, 48000, 160000, 320000, 691282\}$ images and people, i.e. exactly one image per person. These galleries include 825 portraits of the people who ap-

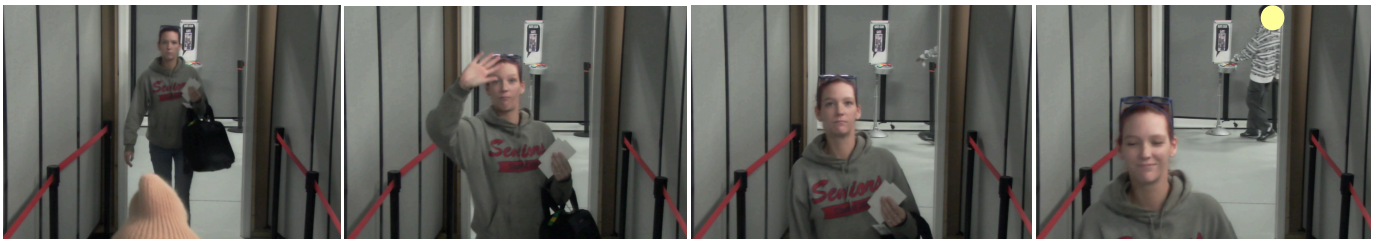
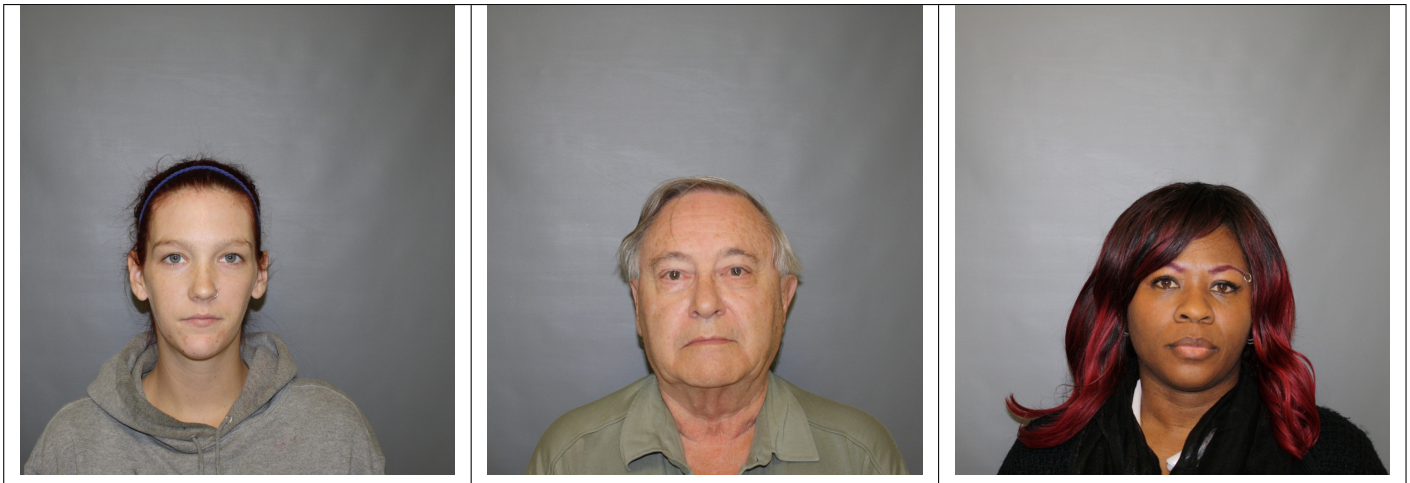


Figure 4: Example images from the ceiling mounted camera for the free movement scenarios from videos collected on an aircraft boarding ramp. The images in this table are from the subject S1115 in the DHS / S&T provided AEER dataset. The subject gave written opt-in permission to allow public release of all imagery. Where consent from individuals in the background was not obtained, their faces were masked (yellow circle).

IDENTIFICATION IN TOP R RANKS FROM DATABASE OF SIZE N		VERIFICATION	
FNIR(N, R, T)	"False negative identification rate"	FNMR(T)	"False non-match rate"
FPIR(N, T)	"False positive identification rate"	FMR(T)	"False match rate"



Subject S1155 (Perm Granted)	Subject S2880 (Perm Granted)	Subject S1848 (Perm Granted)

Figure 5: Examples of enrollment images collected with an SLR camera. The face images in this figure are from the DHS / S&T provided AEER dataset. The included subjects consented to release their images in public reports.

pear in the mated search sets described next. Examples of the portraits appear in section 5.

- ▷ **Mated search images:** The non-cooperative face images are faces cropped from video clips collected in surveillance settings. Examples of the cropped faces and the parent video frames are shown in Figures 4 and 3
- ▷ **Non-mated search images:** A separate set of $N_I = 79403$ faces cropped from video that are known not to contain any of the enrolled identities are used to estimate false positive accuracy.

Accuracy metrics: Scores from the mated searches are used in the false negative identification rate (FNIR) computation. FNIR is defined as the number of mated searches which fail to produce the enrolled mate in the top R ranks with score above threshold, T. FNIR is therefore known as a miss rate. It's value will generally increase with the size of the enrolled database, N, because the recognition algorithm is tasked with assigning a low score to *all* $N-1$ non-mated enrollments. Thus for each of M mated searches the algorithm returns $1 \leq r \leq L$ candidates with hypothesized identities and similarity scores. If the identity of the search face is ID_i and that of the r-th candidate is ID_r then

$$\text{FNIR}(N, R, T) = 1 - \frac{1}{M} \sum_{i=1}^{M} \sum_{r=1}^{R} H(s_{ir} - T)\ \delta(\text{ID}_i, \text{ID}_r) \tag{3}$$

where s_{ir} is the r-th highest score from the i-th search, the step function H(x) is 1 if $x \geq 0$ and 0 otherwise, and the function $\delta(x, y)$ is 1 if $x = y$, and 0 otherwise.

In cases where an algorithm fails to produce a template from an input image - the so-called failure to enroll outcome - the FNIR computation proceeds by assigning a low score, $-\infty$, and high rank, $L+1$, This simulates a miss.

Scores from the non-mated searches are used in the false positive identification rate (FPIR) computation, which states the proportion of non-mate searches yielding *any* candidates at or above a threshold T:

$$\text{FPIR}(T) = \frac{1}{N_I} \sum_{i=1}^{N} H(s_i - T) \tag{4}$$

In cases where an algorithm fails to produce a template from an input image, a low score is again assigned as the result of any comparison involving that template. This practice actually benefits (reduces) FPIR.

IDENTIFICATION IN TOP R RANKS FROM DATABASE OF SIZE N		VERIFICATION	
FNIR(N, R, T)	"False negative identification rate"	FNMR(T)	"False non-match rate"
FPIR(N, T)	"False positive identification rate"	FMR(T)	"False match rate"

Figure of Merit: The prize is awarded to the algorithm that achieves the lowest FNIR when the threshold is set to produce FPIR at or below 0.001. This was determined using N = 691,282, and probes from the travel concourse dataset. This criterion differs substantially from many benchmarks and academic studies which try to maximize "rank one hit rate", i.e. to minimize $\text{FNIR}(N, 10)$. The criterion here, instead, seeks to minimize $\text{FNIR}(N, L, T)$ by demanding that mated candidates exceed a score threshold that is adopted to minimize false positives. Use of a high threshold is an imperative in the many operations which feature high search volumes and a low prior probability that the search is mated. An example, is a casino "watch list" surveillance application in which card sharps are a small minority of the customer base.

Prize winner: By consulting Figure 7, the most accurate identification algorithm on this dataset is developed by Yitu. Using probes from the travel concourse dataset to search in a dataset of N = 691282 portraits, the first Yitu algorithm gives FNIR = 0.204 with FPIR = 0.001.

Discussion: The Yitu algorithms would win this prize at all tested gallery sizes. The algorithms, however, would not win had the figure-of-merit been a zero-threshold, rank-based metric. As can be seen in Figure 8 the first NTechLab algorithm gives lowest FNIR(N, R, 0) for R = 1 and all N values, i.e. the NTechLab algorithm places more correct mates at rank 1 but does not do so with a score high enough to survive a threshold. Figure 10 shows error tradeoff characteristic of NTechLab is superior to Yitu at very low thresholds (high FPIR), but Yitu has a flatter response and quickly dominates all other algorithms for FPIR below about 0.88. Whether this would be sustained at very low values of FPIR, for example below 0.0001, is unknown given the limited test size $N_I = 79403$.

IDENTIFICATION IN TOP R RANKS FROM DATABASE OF SIZE N		VERIFICATION	
FNIR(N, R, T)	"False negative identification rate"	FNMR(T)	"False non-match rate"
FPIR(N, T)	"False positive identification rate"	FMR(T)	"False match rate"

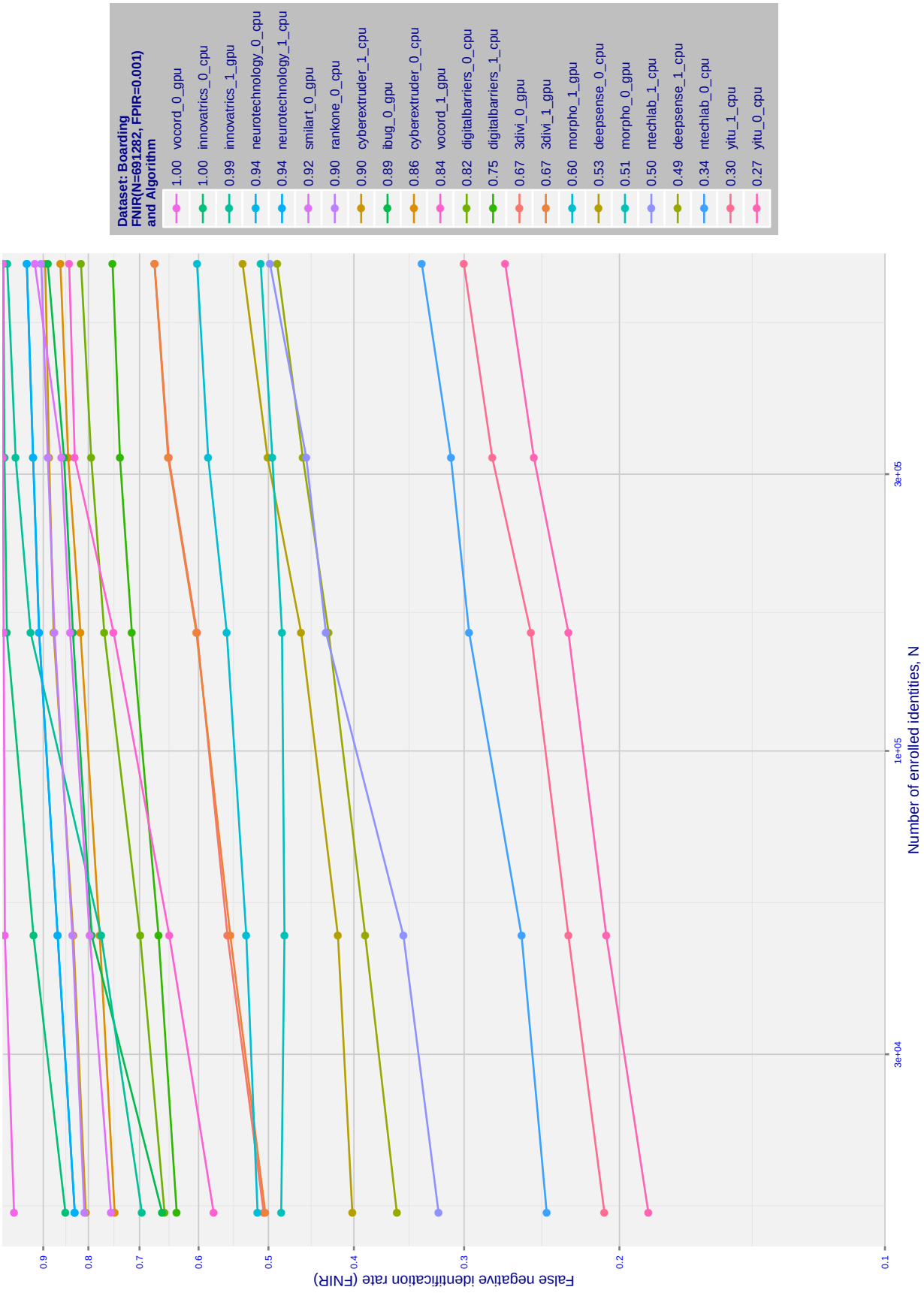


Figure 6: For the boarding gate dataset, the curves show false negative identification rates (FNIR) versus enrolled population when the threshold is set to a high value sufficient to limit false positive outcomes, FPIR = 0.001. This metric is relevant to automated watchlist applications, where most searches are from individuals who are not enrolled.

IDENTIFICATION IN TOP R RANKS FROM DATABASE OF SIZE N		VERIFICATION	
FNIR(N, R, T)	"False negative identification rate"	FNMR(T)	"False non-match rate"
FPIR(N, T)	"False positive identification rate"	FMR(T)	"False match rate"

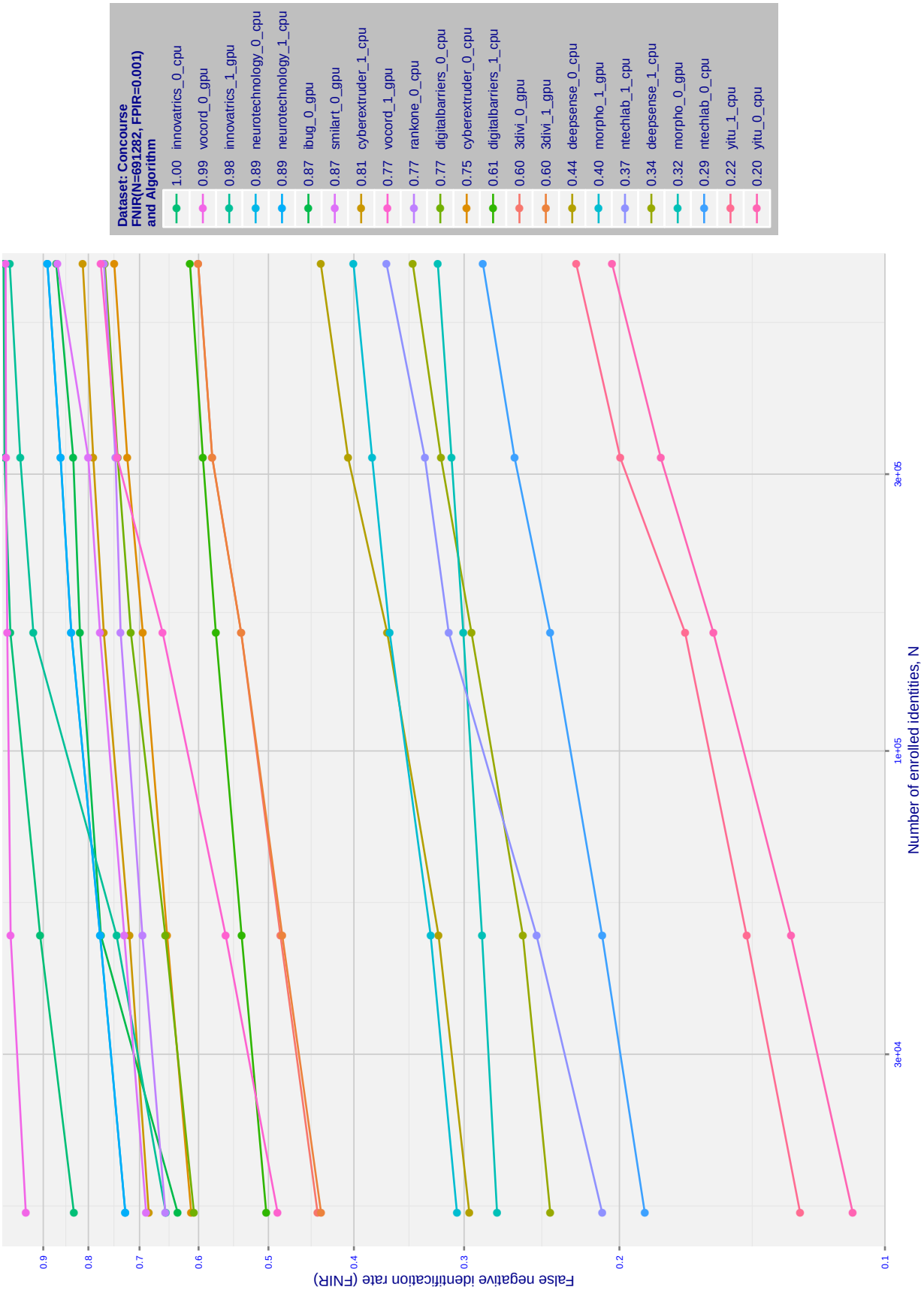


Figure 7: For the travel concourse dataset, the curves show false negative identification rates (FNIR) versus enrolled population when the threshold is set to a high value sufficient to limit false positive outcomes, FPIR = 0.001. This metric is relevant to automated watchlist applications, where most searches are from individuals who are not enrolled.

IDENTIFICATION IN TOP R RANKS FROM DATABASE OF SIZE N		VERIFICATION	
FNIR(N, R, T)	"False negative identification rate"	FNMR(T)	"False non-match rate"
FPIR(N, T)	"False positive identification rate"	FMR(T)	"False match rate"

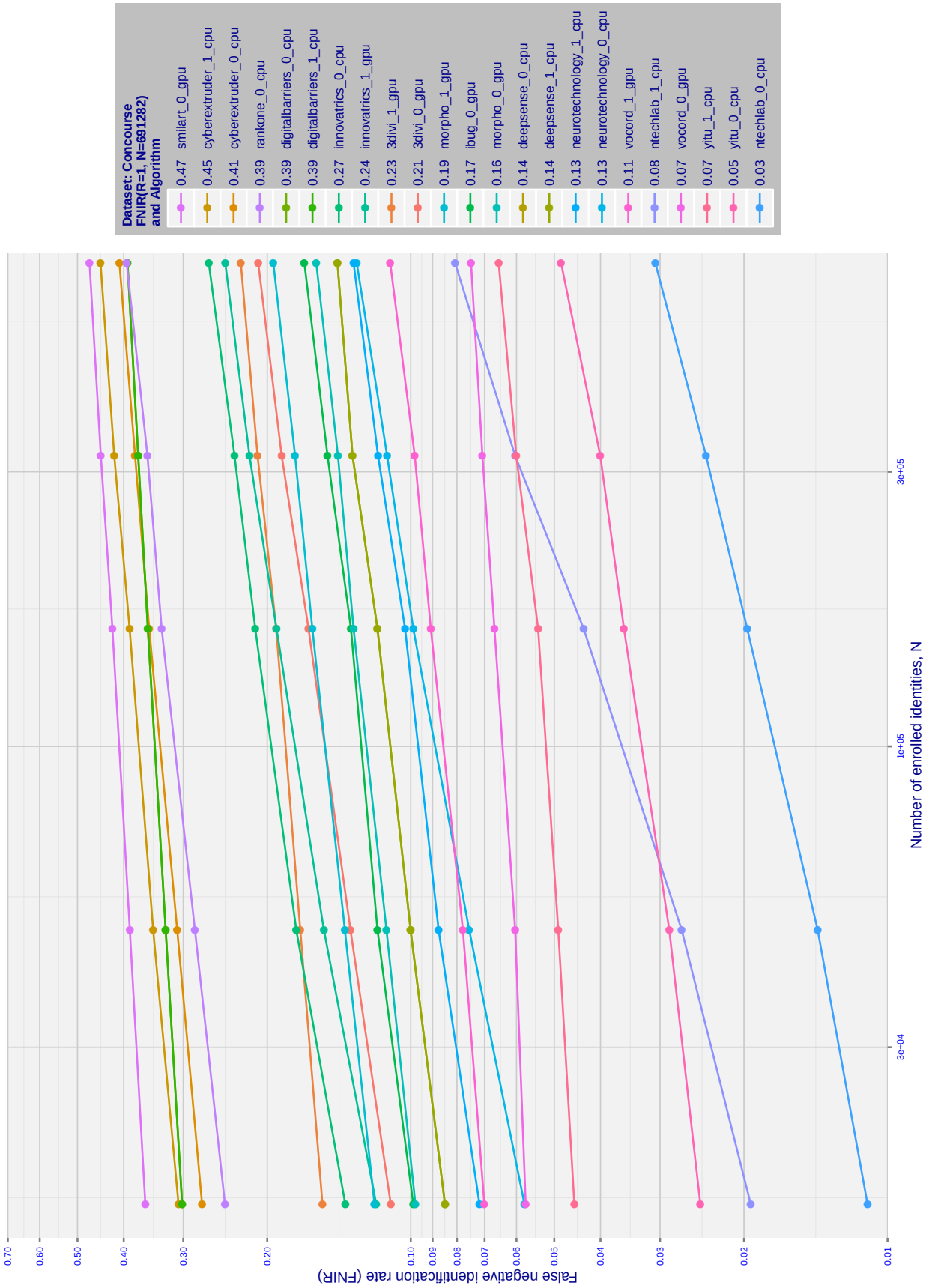


Figure 8: For the travel concourse dataset, the curves show false negative identification rates (FNIR) at rank 1 versus population size, N. The threshold is set to zero. This metric is relevant to human reviewers who will traverse candidate lists in pursuit of investigations.

IDENTIFICATION IN TOP R RANKS FROM DATABASE OF SIZE N		VERIFICATION	
FNIR(N, R, T)	"False negative identification rate"	FNMR(T)	"False non-match rate"
FPIR(N, T)	"False positive identification rate"	FMR(T)	"False match rate"

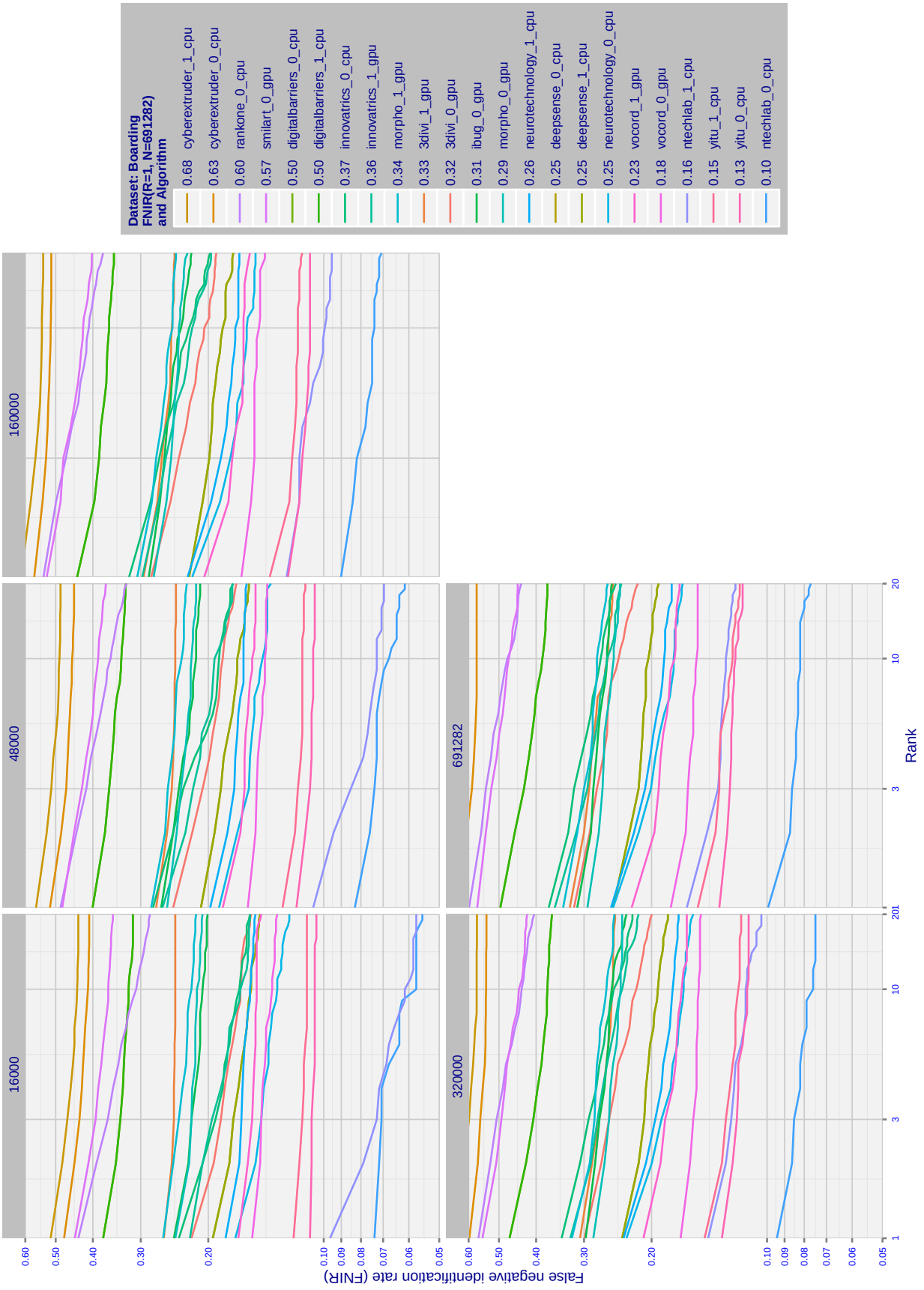


Figure 9: For the boarding gate dataset, the curves show false negative identification rates (FNIR) versus rank when the threshold is set to zero. This metric is relevant to human reviewers who will traverse candidate lists checking whether any of the returned identities match to the search imagery.

IDENTIFICATION IN TOP R RANKS FROM DATABASE OF SIZE N		VERIFICATION	
FNIR(N, R, T)	"False negative identification rate"	FNMR(T)	"False non-match rate"
FPIR(N, T)	"False positive identification rate"	FMR(T)	"False match rate"

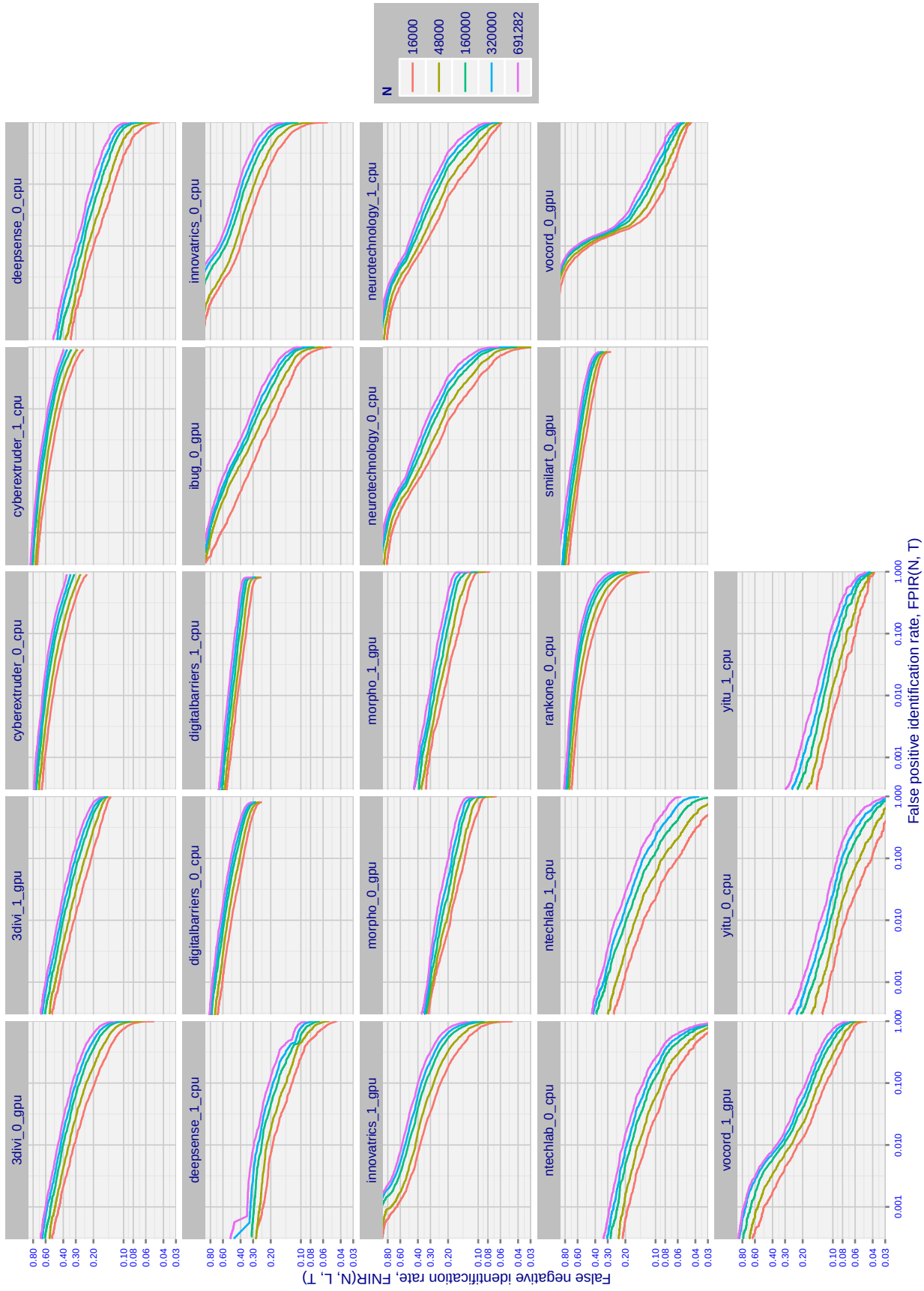


Figure 10: For the travel concourse dataset, the curves show FNIR *vs.* FPIR*, parametrically on threshold T. High thresholds are at the left side of the graphs.*

IDENTIFICATION IN TOP R RANKS FROM DATABASE OF SIZE N		VERIFICATION	
FNIR(N, R, T)	"False negative identification rate"	FNMR(T)	"False non-match rate"
FPIR(N, T)	"False positive identification rate"	FMR(T)	"False match rate"

3 The identification speed challenge

Background: Prior tests have documented search speeds spanning up to three orders of magnitude. Given the implications for hardware procurement, it becomes essential to measure speed and to only invest in slow algorithms if they offer tangible accuracy advantages. Further, given very large operational databases, the scalability of algorithms is important. It has been reported previously [2] that search duration can scale sublinearly with enrolled population size N. Further there has been considerable recent research on indexing, exact [4] and approximate nearest neighbor search [1, 4] and fast-search [5].

Figure of merit: The FRPC therefore included a prize for the fastest search algorithm but with the requirement that it also has competitive accuracy. Formally the prize went to the algorithm with the lowest template search duration and which gave FNIR no larger than twice the best FNIR. The false negative identification rate in question here is FNIR(N, N T) with N = 691282, and T set to give FPIR = 0.001. The figure of merit did not include the time taken to prepare the search template which is independent of N and which dominates search time up to some crossover population size whereupon search duration is larger.

Participation: The challenge was open to all participants in the identification accuracy challenge, as listed in Table 2.

Prize winner: Figure 11 charts the speed measurements presented earlier in Table 2. By consulting the figure, the fastest identification algorithm is the second algorithm developed by NTechLab `http://ntechlab.com/`. The algorithm gives search duration that grows sub-linearly fitting neither a logarithmic nor Power-law model exactly. It is faster but somewhat less accurate than its linear sibling.

Note that we did not differentiate between CPU and GPU based implementations - developers were free to submit algorithms using either kind of hardware. For those algorithms listed in Table 2 as CPU, the search duration is measured on an Intel(R) Xeon(R) CPU E5-2630 v4 running at 2.20GHz. For GPU algorithms, the hardware is an NVidia Tesla K40m equipped with 12GB of memory. However the FRPC test infrastructure did not record whether search was actually conducted on the CPU or GPU - it could have been either.

4 Effect of head orientation

Invariably the most influential parameter on recognition outcomes has been the orientation of the head in one photograph relative to that in a prior image.

- **Verification dependence of yaw in pairs of images:** Using wild photographs and yaw estimates obtained from an automated, government owned, pose-estimation tool we quantify the dependence of face recognition accuracy on yaw. The ability of algorithms to compensate for viewing angle is summarized in Figure 13 which shows false non-match rate as a function of yaw angle, θ, of the face in enrollment and verification images. These vary over ± 90 degrees. Each panel encodes false non-match rate FNMR for an algorithm at a particular threshold. This is set to give a false match rate of 0.001 for images of frontal pose i.e. those with $|\theta| \leq 15$. The FNMR values are generally lowest for frontal pairs, then for pairs with the same yaw angle, and they increase with difference in yaw.

  At this fixed threshold, Figure 14 shows how FMR itself varies with the pair of yaw angles. This figure is relevant in applications where a global threshold is set and pose varies widely. It would not be relevant in cases where a specific pair of poses is designed-in, and a dedicated threshold could be set. In all panels the center cell has FMR = 0.001, by design. The results for other yaw angles show different behaviors. First, the more accurate algorithms often have weak dependence of FMR on yaw angles (prevalence of grey). Others give consistently low FMR when angles differ (prevalence of blue) consistent with an inability to match. A final class of algorithms give *higher* FMR when yaw angles differ (prevalence of red in the periphery). This is typically unexpected and undesirable.

IDENTIFICATION IN TOP R RANKS FROM DATABASE OF SIZE N		VERIFICATION	
FNIR(N, R, T)	"False negative identification rate"	FNMR(T)	"False non-match rate"
FPIR(N, T)	"False positive identification rate"	FMR(T)	"False match rate"

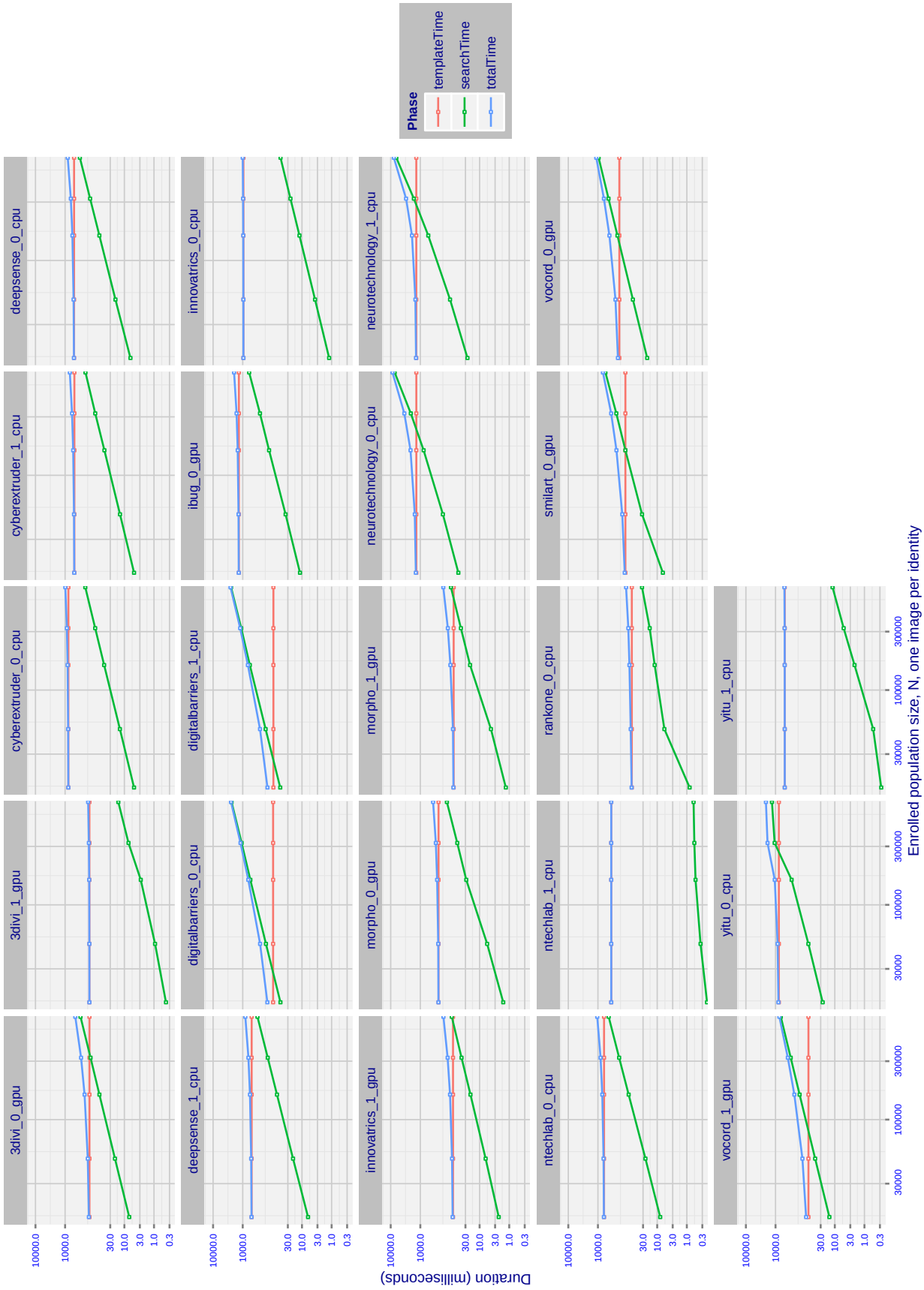


Figure 11: The figure shows the template creation time, search time, and the both combined, as a function of enrolled population size.

IDENTIFICATION IN TOP R RANKS FROM DATABASE OF SIZE N		VERIFICATION	
FNIR(N, R, T)	"False negative identification rate"	FNMR(T)	"False non-match rate"
FPIR(N, T)	"False positive identification rate"	FMR(T)	"False match rate"

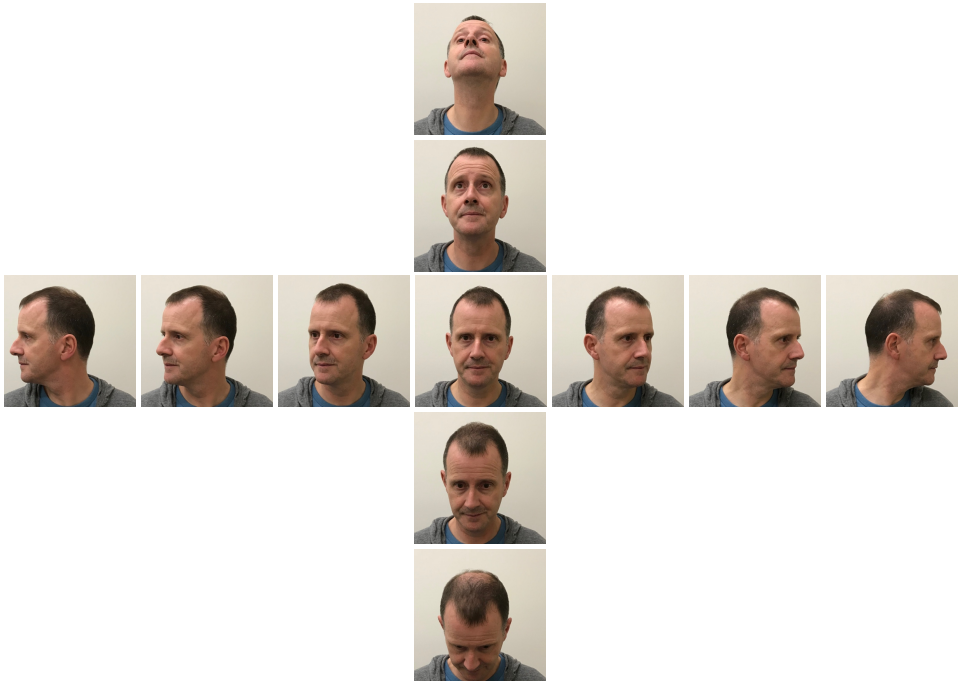


Figure 12: Approximate examples of the images passed to the FRPC *identification algorithms for the off-angle experiments summarized by figures*

▷ **Identification with frontal enrollment**: It is often the case that a cooperative enrollment photograph that is collected to be an authoritative reference image placed in a credential (passport, driving license) or database (e.g. mugshot database) will conform to a standard prescription of frontal pose, with roll, pitch and yaw all being zero degrees. Accuracy then is determined by the pose relative to that. In the general case the three head angles - roll, pitch and yaw - can vary independently taking on values up to (and beyond) 90 degrees from a frontal (0, 0, 0) view. The relative yaw angle can then ascend to ± 90 degrees, while pitch is usually constrained by the range of motion of the neck to say ± 60 degrees. Roll alone is not usually considered to be serious impediment to face recognition since an implementation that detects eyes can perform an in-plane rotation to remove roll. However, compound rotation of the head, as might be seen if a non-cooperative subject was lying down, has presented severe challenges to face recognition.

Using dedicated controlled non-frontal search images, of the kind shown in Figure 12, for enrolled mates present in galleries of size N $= \{16000, 48000, 160000, 320000, 691282\}$, we plot both rank 1 accuracy, $1-\text{FNIR}(N, 1, 0)$ and high-threshold accuracy $1-\text{FNIR}(N, L, T)$ against yaw angle relative to a zero degree frontal. The results are shown in three figures as follows. The first two, Figures 15 and 16, show the sensitivity of rank one hit rate to pitch and yaw, respectively. Many algorithms give excellent accuracy with same-day frontal images, but degrade markedly with pitch of ± 40 degrees. Similarly with yaw, most, but not all, algorithms fail to identify profile-view probes.

Figure 17 shows yaw dependence again but for FNIR at high threshold, as would be set in a surveillance application. This exposes earlier declines in accuracy, as yaw depresses similarity scores below the threshold.

IDENTIFICATION IN TOP R RANKS FROM DATABASE OF SIZE N		VERIFICATION	
FNIR(N, R, T)	"False negative identification rate"	FNMR(T)	"False non-match rate"
FPIR(N, T)	"False positive identification rate"	FMR(T)	"False match rate"

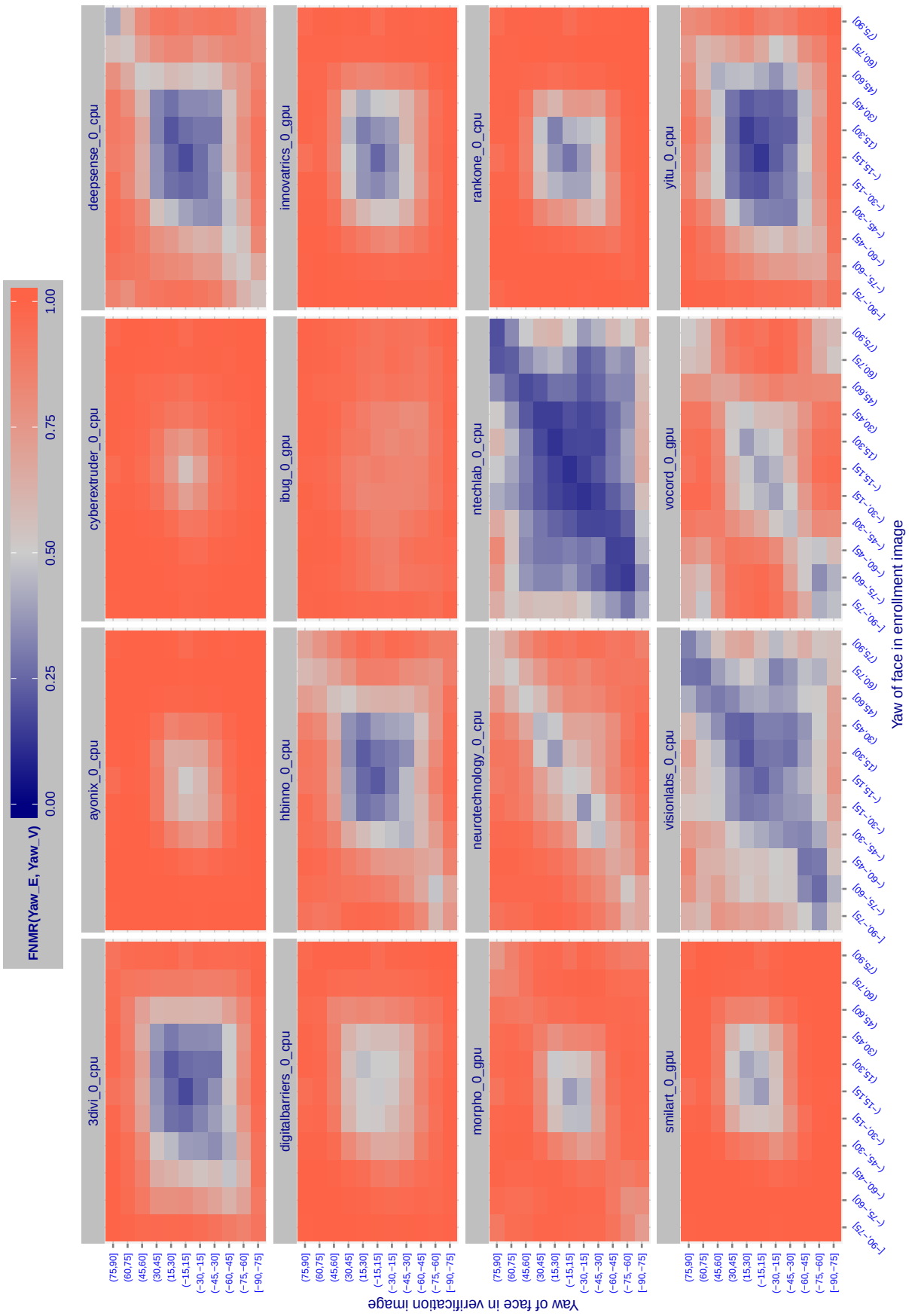


Figure 13: The heatmaps shows FNMR *as a function of the yaw of the enrollment and verification images. The threshold is the same in all cells, and is set to the value that yeilds* FMR *= 0.001 on near frontal pairs i.e. where yaw is in the interval* $(-15, 15]$*. Poor algorithms give generally red figures. The better algorithms show a) diagonal dominance, indicating ability to authenticate when pairs have the same yaw angle, and b) off-diagonal cross-pose capability also. The yaw estimates are from an automated pose estimator, and are themselves noisy. The figure assumes that the pose estimates are not systematically incorrect.*

IDENTIFICATION IN TOP R RANKS FROM DATABASE OF SIZE N		VERIFICATION	
FNIR(N, R, T)	"False negative identification rate"	FNMR(T)	"False non-match rate"
FPIR(N, T)	"False positive identification rate"	FMR(T)	"False match rate"

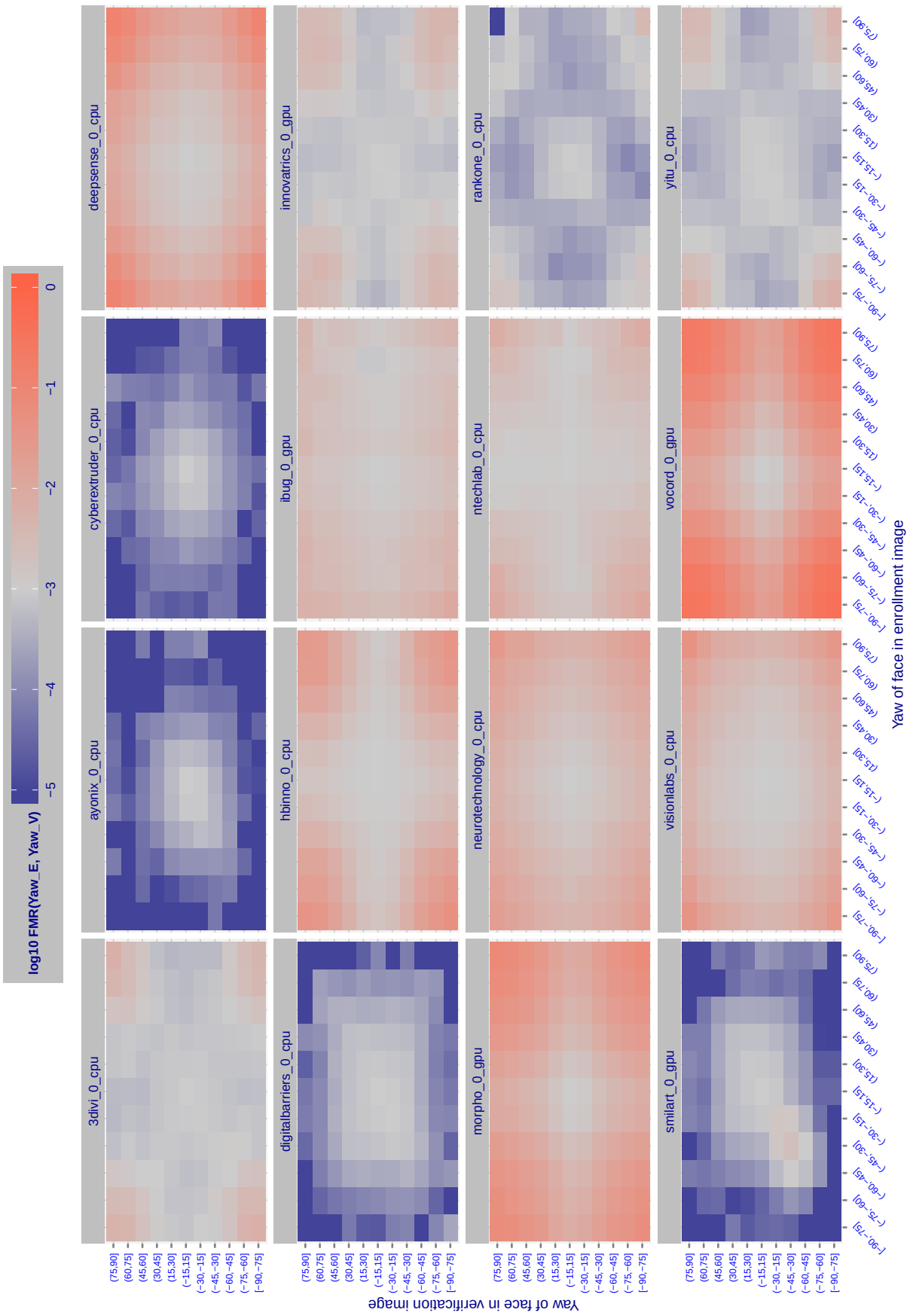


Figure 14: The heatmap shows FMR *dependence on the yaw of the enrollment and verification images. The threshold is the same in all cells, and is set to the value that yields* FMR *= 0.001 on near frontal pairs, i.e. where yaw is in the interval* $(-15, 15]$*. Thus the center of each panel is grey. The desired behavior is that* FMR *does not vary with relative yaw. However, some algorithms gives elevated* FMR *when yaw differs. This would present a security vulnerability in, say, mobile-phone authentication attempts where an off-angle presentation against a (nominally) frontal enrollment would give a much higher chance of impostor access. The yaw estimates are from an automated pose estimator, and are themselves noisy. The figure assumes that the pose estimates are not systematically incorrect.*

IDENTIFICATION IN TOP R RANKS FROM DATABASE OF SIZE N		VERIFICATION	
FNIR(N, R, T)	"False negative identification rate"	FNMR(T)	"False non-match rate"
FPIR(N, T)	"False positive identification rate"	FMR(T)	"False match rate"

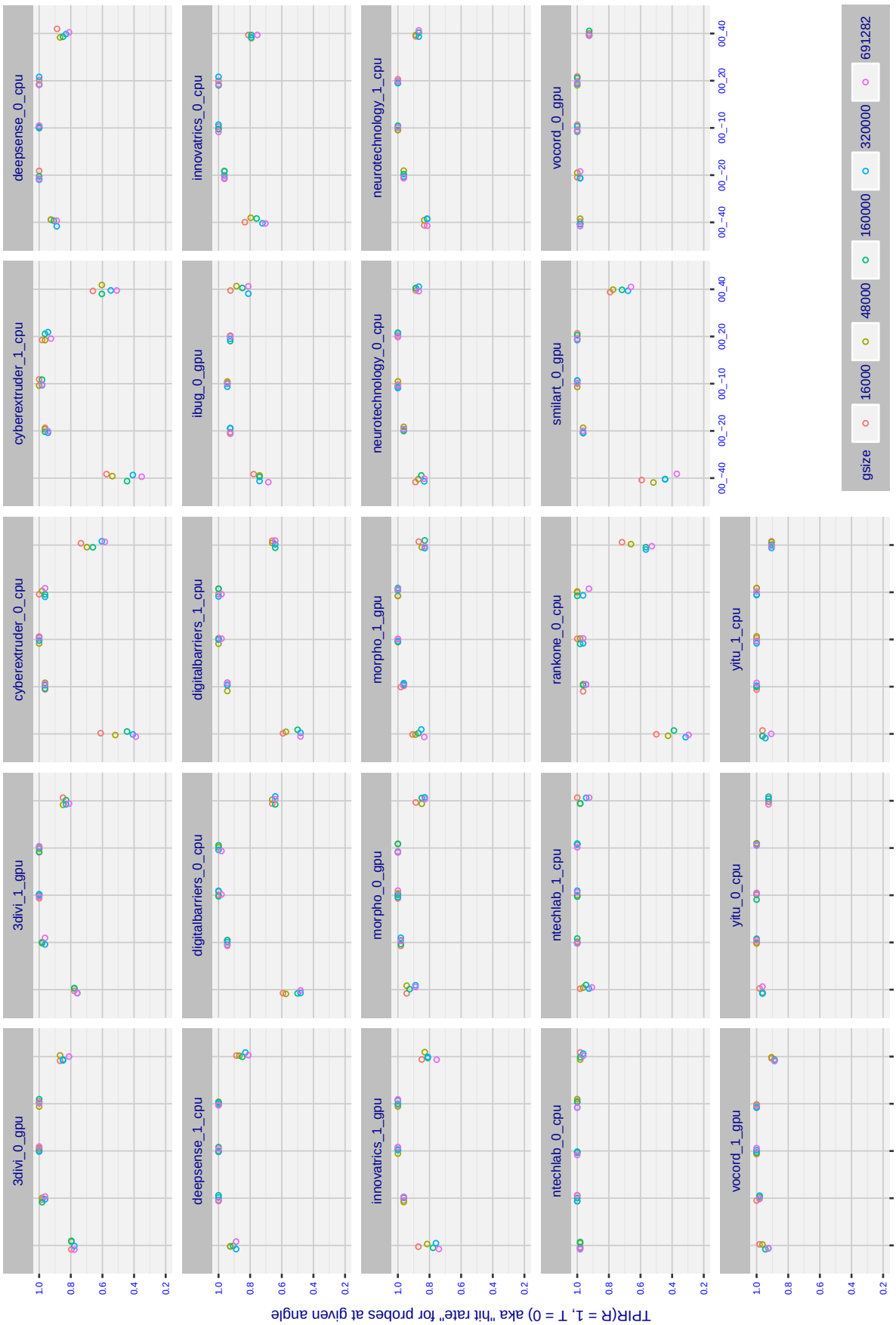


Figure 15: The points show zero threshold, rank one, true positive identification rates (TPIR$= 1-$FNIR)*, aka hit rates, versus pitch difference between frontal enrollment and non-frontal search images. Five gallery sizes are shown. The x-coordinate is jittered to separate the points slightly. The effects induced by pitch are much larger than those due to enrolled population size.*

IDENTIFICATION IN TOP R RANKS FROM DATABASE OF SIZE N		VERIFICATION	
FNIR(N, R, T)	"False negative identification rate"	FNMR(T)	"False non-match rate"
FPIR(N, T)	"False positive identification rate"	FMR(T)	"False match rate"

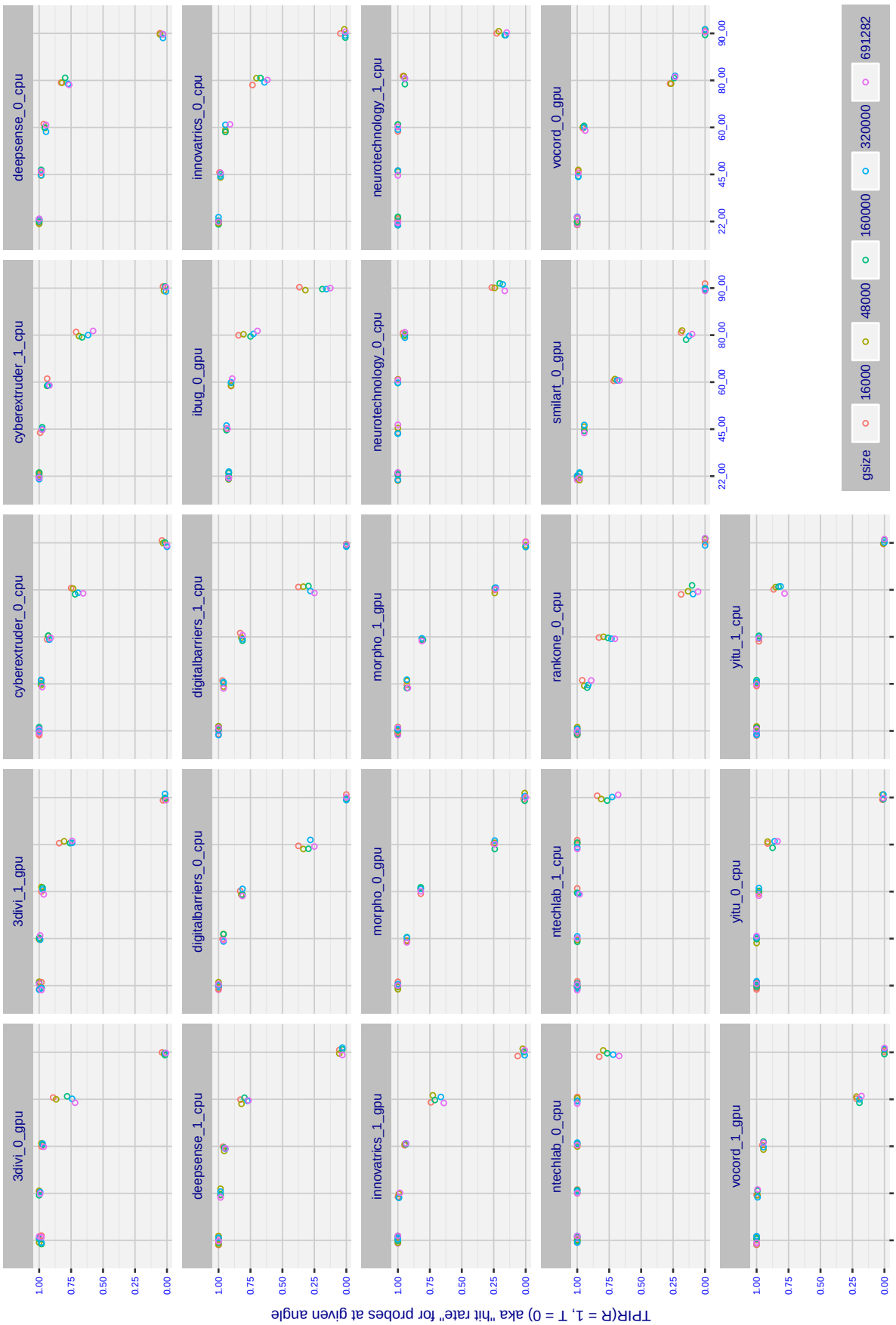


*Figure 16: The points show zero threshold, rank one, true positive identification rates (*TPIR$= 1-$FNIR*), aka hit rates, versus yaw difference between frontal enrollment and non-frontal search images. Five gallery sizes are shown. The x-coordinate is jittered to separate the points slightly. The effects induced by yaw are much larger than those due to enrolled population size.*

IDENTIFICATION IN TOP R RANKS FROM DATABASE OF SIZE N		VERIFICATION	
FNIR(N, R, T)	"False negative identification rate"	FNMR(T)	"False non-match rate"
FPIR(N, T)	"False positive identification rate"	FMR(T)	"False match rate"

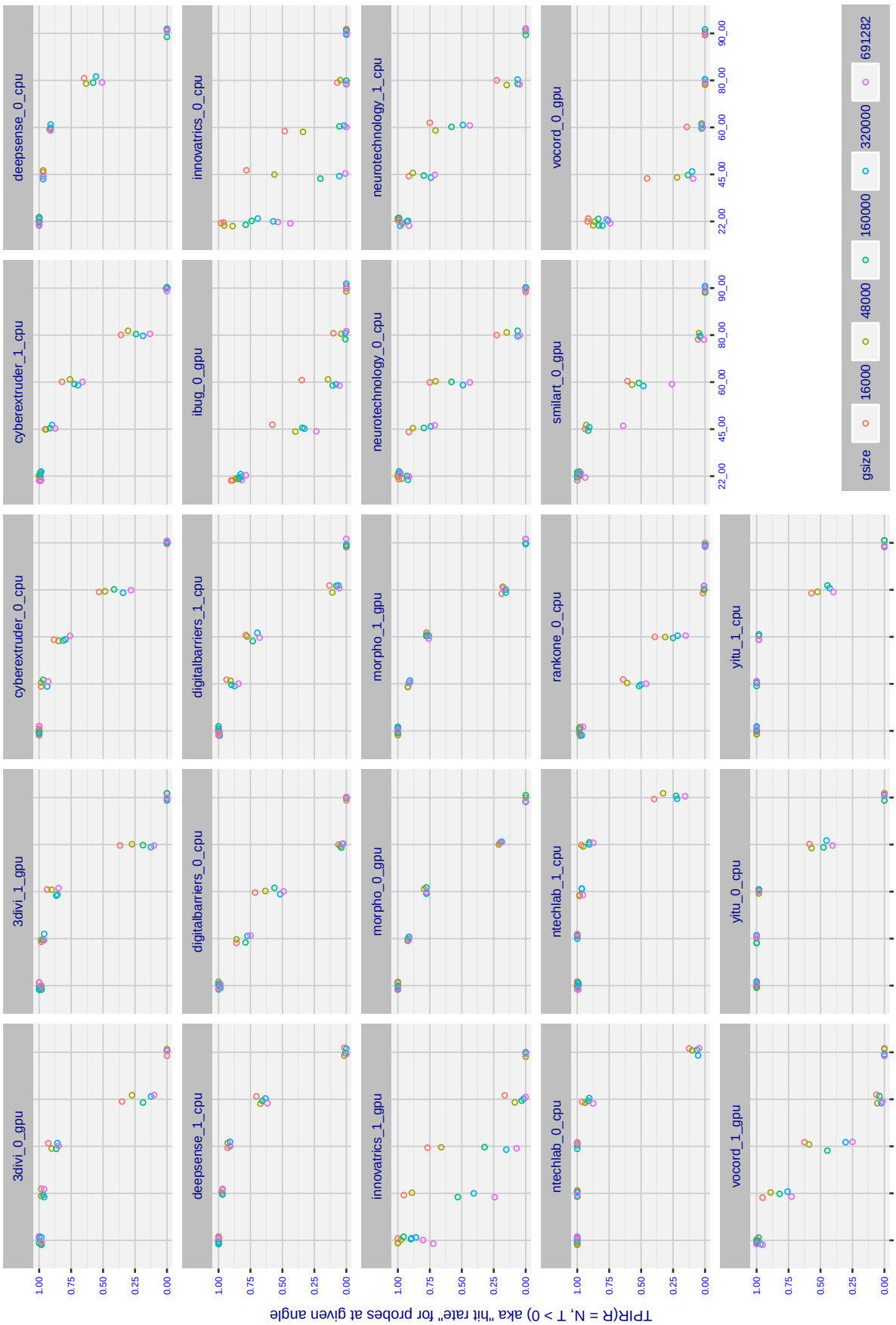


Figure 17: The points show high threshold true positive identification rates ($\text{TPIR} = 1 - \text{FNIR}$*), aka hit rates, versus the yaw angle of the probe face. The threshold is set to a globally high value, set to achieve* $\text{FPIR}(N) = 0.001$*. The enrolled gallery mate has frontal pose. Five gallery sizes are shown. The x-coordinate is jittered to separate the points slightly.*

IDENTIFICATION IN TOP R RANKS FROM DATABASE OF SIZE N		VERIFICATION	
FNIR(N, R, T)	"False negative identification rate"	FNMR(T)	"False non-match rate"
FPIR(N, T)	"False positive identification rate"	FMR(T)	"False match rate"

References

[1] Artem Babenko and Victor Lempitsky. Efficient indexing of billion-scale datasets of deep descriptors. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[2] Patrick Grother and Mei Ngan. Interagency report 8009, performance of face identification algorithms. *Face Recognition Vendor Test (FRVT)*, May 2014.

[3] Gary B. Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst, October 2007.

[4] Masato Ishii, Hitoshi Imaoka, and Atsushi Sato. Fast k-nearest neighbor search for face identification using bounds of residual score. In *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, pages 194–199, Los Alamitos, CA, USA, May 2017. IEEE Computer Society.

[5] Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with gpus. *CoRR*, abs/1702.08734, 2017.

[6] Ira Kemelmacher-Shlizerman, Steven M. Seitz, Daniel Miller, and Evan Brossard. The megaface benchmark: 1 million faces for recognition at scale. *CoRR*, abs/1512.00596, 2015.

[7] O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. In *British Machine Vision Conference*, 2015.

[8] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. *CoRR*, abs/1503.03832, 2015.

[9] Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. Deepface: Closing the gap to human-level performance in face verification. In *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '14, pages 1701–1708, Washington, DC, USA, 2014. IEEE Computer Society.