The IJB-A Face Identification Challenge

Performance Report

Patrick Grother and Mei Ngan

Caution: This report quantifies face recognition performance using data supplied by external research and development organizations. Its results are derived from self-administered experiments on the fully public IJB-A dataset. As such the results can be manipulated by various means that may not be operationally realistic. Therefore, end users of face recognition technology should prefer results from NIST's sequestered testing campaigns, FRVT or FIVE, or on similar independent evaluations of face recognition. Developers whose algorithms exhibit good performance here are encouraged to submit their algorithms to those sequestered test programs.

This report is generated automatically. It will be updated as new algorithms are evaluated, and as new analyses are added. Automated notifications can be obtained via the mailing list. Correspondence should be directed to the authors via FaceChallenges@nist.gov.

This report was last updated on April 26, 2017.

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

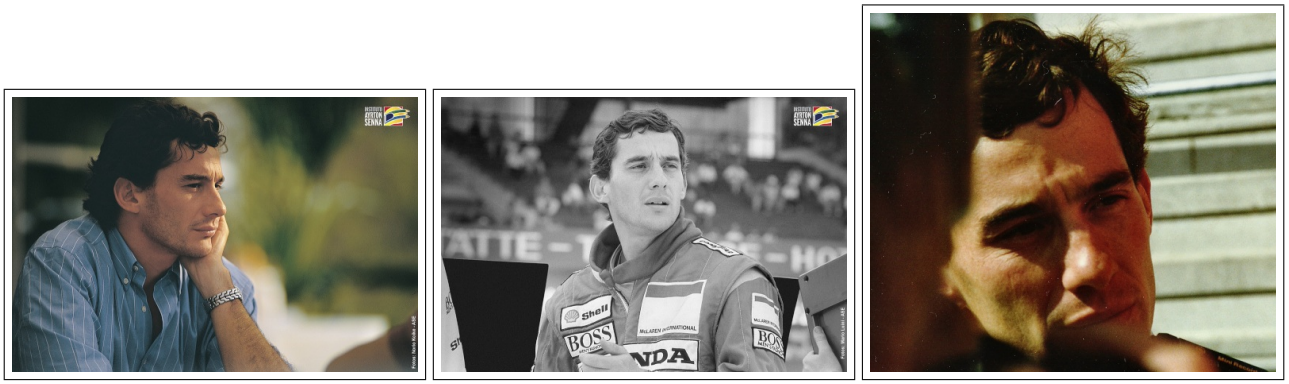

Figure 1: Three images of one subject in the IJB-A dataset. The entire dataset is available online. Many photos were taken by photo journalists and, as such, are well exposed, well focused, and specifically selected as suitable for public display. For face recognition, they nevertheless remain challenging due to wide variations in pose, illumination, expression and occlusion.

1 Introduction

Three IARPA Janus Benchmark A challenges are described by Klare et al. in the paper *Pushing the Frontiers of Unconstrained Face Detection and Recognition*[3]. The second of these, the IJB-A 1:N challenge, quantifies performance of face identification algorithms ("same person or not?") on challenging photo-journalism images of the kind shown in Figure 1. They are considerably more difficult to recognize than the portraits mandated by facial recognition standards[1].

IJB-A 1:N is a "take-home" test in that it is based on fully public data. It follows the design of the LFW protocol in requiring many pairs of samples to be compared in isolation[2]. This corresponds to recognition tasks like passport identification or forensic comparison where there is just a pair of samples and no central database or gallery.

The IJB-A 1:N challenge departs from LFW as follows:

- **Face selection**: LFW contains faces that could be detected with the Viola-Jones face detection algorithm. This limits difficulty. IJB-A on the other hand, uses manually located and annotated faces.
- **Landmarks**: The IJB-A tests include landmark coordinates (eyes and nose) whereas LFW provides just raw images, and aligned (funneled) images.
- **Multi-image samples**: LFW compared single images. IJB-A uses richer samples containing $1 \leq K \leq 202$ images, including frames from video sequences.
- **More impostor pairs**: IJB-A 1:N uses many more impostor comparisons that genuines. In LFW, the ratio was 1 which precluded computation of false match rates at usefully low values.

2 Metrics

This section describes the open-set one-to-many identification accuracy metrics used in this report.

[1] NIST maintains a challenge for such images based on the mugshots of NIST Special Database 32 ("MEDS")[1]. This is intended as a stepping stone prior for developers prior to entering NIST's ongoing fully sequestered FRVT identification test.

[2] IJB-A 1:N does not cross-compare galleries and probesets; it has no concept of such. It does not attempt to measure both identification and identification accuracy from the same similarity score matrix; it does not pin the prior probabilities of impostor vs. genuine pairs i.e. O(n^2) vs. O(n).

2.1 Quantifying false alarms

False alarm incidence is computed over K searches, each involving imagery from a person who is known to not to be present in the enrolled gallery. Each search yields a list of L = 20 candidate identities sorted in order of non-negative scalar similarity scores. The false postitive identifcation rate (FPIR) is defined as the proportion of searches with any candidates at or above threshold threshold T.

$$\text{FPIR}(N, T) = 1 - \frac{1}{K} \sum_{i=1}^{K} H(s_{i1} - T) \tag{1}$$

where H is a unit step, and s_{i1} is the first (highest) score on the i-th nonmated candidate list. The enrolled population (gallery) size is N.

This metric does not account for searches that produce several above-threshold candidates. The appropriate metric there is selectivity, which counts the number of false positives expected from a nonmate search (at some threshold).

2.2 Quantifying false rejection

False rejection is computed over M searches each involing imagery from a person who is known to be present in the enrolled gallery. Each search yields a list of L = 20 candidate identities sorted in order of non-negative scalar similarity scores. Zero or one of the candidates will be from the search individual. The false negative identification rate is is defined as the proportion of scores for which the known individual is outside the top R ranks, or has similarity below threshold T.

$$\text{FNIR}(N, R, T) = 1 - \frac{1}{M} \sum_{a=1}^{M} H(s_{ic} - T) H(R - r_{ic}) \tag{2}$$

The FNIR defnition supports two use cases:

- ▷ *Forensic*: In a high profile case, or in an application where only a few searches are ever conducted, a human analyst might examine all L candidates or perhaps just the top $R \leq L$ identities. The appropriate metric then is the cumulative match charateristic(CMC) which gives the fraction of hits at rank R or better, CMC(N, R, L) = 1 - FNIR$(N, R, 0)$. By ignoring scores, this metric allows "weak" hits to count as strongly as high-scoring "strong" hits. The CMC metric is relevant to operations in which (trained) human reviewers will traverse candidate lists in pursuit of hits. This is possible when the volume of searches is low enough, and when the CMC is favorable enough, to utilize all available labor.
- ▷ *Surveillance*: On the other hand, in applications such as public-space surveillance, where prior probabilities of mates are low, or where search volumes are very high and where human labor has limited availability, it becomes impossible to review all candidate lists. To limit workload a threshold T is applied so that only candidates with score at or above threshold are flagged for examination. The appropriate metric then is FNIR(N, N, T) because rank becomes irrelevant. High thresholds suppress false positives, but elevate false negatives. For example, the German trial of a surveillance system in the Mainz train station[2] configured thresholds on the algorithms to target FPIR$= 0.001$.

FNIR is colloquially referred to as "miss rate". Its complement, true positive identification rate, TPIR, is the "hit rate".

2.3 Non-equivalence of 1:1 and 1:N performance

A 1:N search can most simply be implemented by executing N 1:1 comparison and sorting the results. However, a number of algorithmic techniques exist to improve 1:N accuracy and to expedite search.

From a metrics point-of-view, 1:1 accuracy is stated via a plot of false non-match vs. false match rates, FNMR(T) vs. FMR(T). 1:N accuracy is plotted as FNIR(T) vs. FPIR(T). The difference is not just terminology because FPIR is estimated using the highest highest nonmated score (i.e. a sample drawn from an extreme value distribution) while FMR is estimated from scores drawn from the entire impostor distribution.

Practitioners sometimes regard a 1:N error tradeoff characteristic as being identical to a 1:1 DET with the horizontal axis scaled by N. This is a first order model obtained from the binomial approximation when a 1:N search is indeed the result of N 1:1s.

IJBA includes separate identification and verification tasks to encourage improved search algorithms, both with respect to accuracy and speed. Indeed the IJBA 1:N task makes no assumptions of how search is implemented. It regards a search as an atomic operation.

3 Results

3.1 Comparing accuracy

The graphs that follow include results for several classes of algorithms that are differentiated by their development date, and use of landmarks and training data - see Table 1. This latter issue is nuanced and yet critical to understanding how and whether algorithms can be compared. Historically commercial algorithms have been provided and used in an entirely off-the-shelf manner - the representation is fixed and the user in no way adapts (trains) the algorithm to his native data. The academic community, meanwhile, almost always isolates some portion of the data for the express purpose of adapting the algorithm. The result is a refined set of parameters, or explicit data "models" (most prosaically, a PCA basis set). The academic community, ignoring marketplace practice, has noted that recognition accuracy is improved by training, and training is improved through detailed exploitation of large training sets. Why then do commercial implementations not roll-out training facilities within their commercial off the shelf products. The answer partly rests on the observation that succesful training and adaptation is a fine art that, empirically, cannot be canned in a simple function call.

That said, one particular kind of training is possible operationally: Gallery training occurs after templates have been enrolled into a gallery. The range of techniques is varied from simple O(N) aggregation of statisics to O(N^2) feature space comparison, separation or clustering techniques. In commercial cases, it is usually a trade secret. Gallery training is effective particularly when it can be assumed that the N enrolled items come from N distinct individuals. The efficacy of the techniques can depend on the integrity of the ground truth identity labels, and it is potentially retrograde to conduct this form of training on an un-consolidated set in which the same individual is present in the gallery under several unknown identifiers. Some commercial implementations do compute data across gallery entries. This data is used or retained to improve recognition accuracy.

Algorithm	Development thru	Training data	Gallery training	Role of provided IJB-A landmarks
BENCH-MK1	2010	External training data only		
ANON-2013	2013	External training data only	Unknown, likely	Algorithm was provided only with image cropped from bounding box
MSU-071715	2015	External training data and IJB-A training splits	Yes	Algorithm was provided only with full IJBA-specified landmarks
JANUS*	2015-09	External training data and IJB-A training splits	Yes	Algorithm was provided only with full IJBA-specified landmarks
RankOne-011816	2016-01	External training data	No	Developer asserts IJBA bounding boxes and landmark points were not used.

Table 1: **Context of use:** Comparison of the algorithms should be conducted in the context of variations in when they were developed, with what training data and on whether the ran in fully automated mode or were assisted by the provision of geometric information. Note that the ANON-2013 algorithm was developed before the IJB-A challenge was assembled and was provided to NIST without the expectation that it would be run on images of this type.

Table 2: **Leaderboard:** The table shows various statements of false negative identification rate (FNIR) and algorithm rankings based on those. Different applications of face identification algorithms require different metrics or operating points. Some algorithms can be tuned for such. From left, threshold-based "hit" rates at FPIR of 0.01 and 0.1, then rank-based hit rates at ranks 1 and 10. Lower FPIR values cannot be sustained given the limited number of subjects. The full error tradeoff characteristics appear in Figure 4. **The COTS algorithms were developed before the IJB-A set was developed, and were provided to NIST without any training nor expectation that they would be run on images of this type.**

Result	Algorithm	TPIR@FPIR=0.01	R1	TPIR@FPIR=0.1	R2	CMC(R=1)	R3	CMC(R=10)	R4	Enrol Tsize(bytes)	Search Tsize(bytes)
1	NUS-032917	0.86	1	0.94	1	0.96	1	0.99	1	3779	3661
2	VCOG-021317	0.50	4	0.77	2	0.89	2	0.97	2	8208	4112
3	JanusD-071715	0.36	8	0.66	5	0.88	3	0.97	3	16384	16384
4	JanusB-092015	0.52	3	0.75	3	0.87	4	0.95	7	93762	93756
5	JanusA-090815	0.49	5	0.71	4	0.85	5	0.96	6	1280	1280
6	JanusC-090815	0.38	6	0.60	9	0.85	6	0.96	5	4144	4136
7	MSU-072115	0.37	7	0.60	8	0.82	7	0.96	4	400	400
8	JanusB-071315	0.12	14	0.64	6	0.80	8	0.93	8	128000	128000
9	JanusA-071715	0.35	9	0.56	10	0.73	9	0.86	10	1280	1280
10	JanusC-071515	0.17	12	0.35	13	0.71	10	0.91	9	4122	4122
11	ANON-2013	0.53	2	0.62	7	0.65	11	0.72	12	2529	2529
12	RankOne-011816	0.28	10	0.45	11	0.64	12	0.69	13	85	81
13	BENCH-MK1	0.17	13	0.39	12	0.58	13	0.76	11	1	74422
14	RankOne-091015	0.18	11	0.31	14	0.50	14	0.59	14	71	71

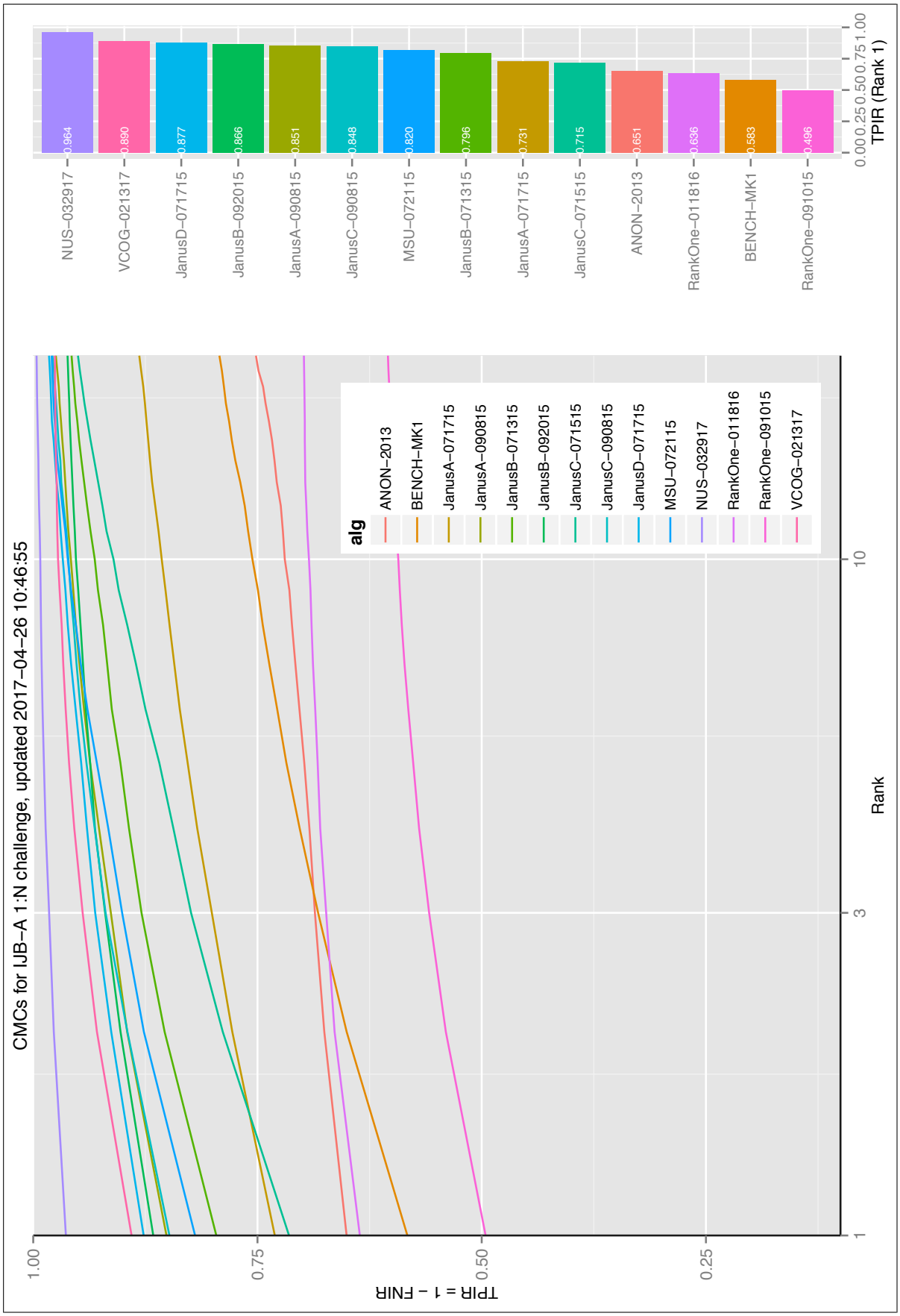

Figure 2: **CMCs:** At left, are full cumulative match characteristics (CMCs) for the IJB-A algorithms. At right is a summary statistic corresponding to a particular rank. This is included to ease lookup.

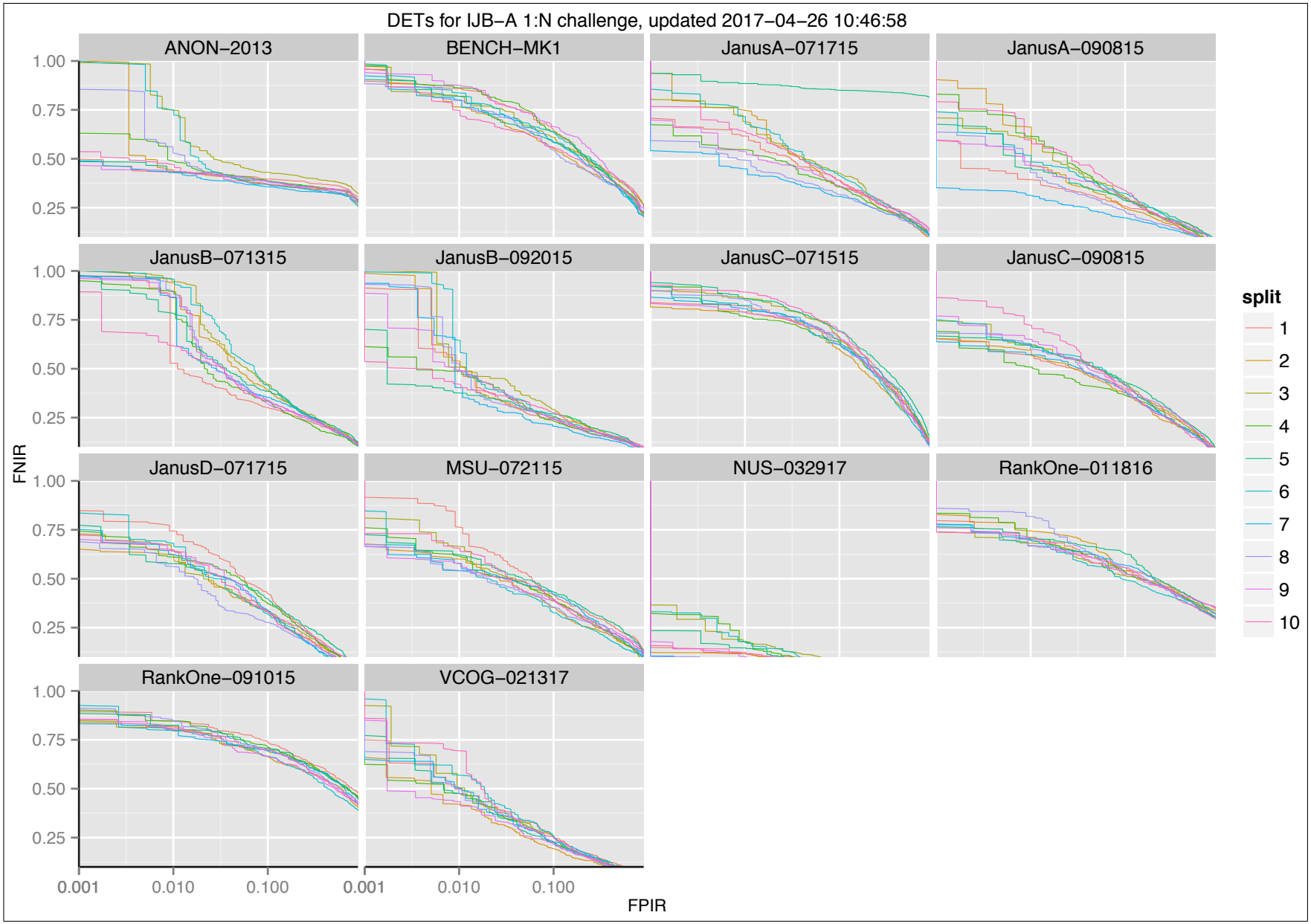

Figure 3: **DETs by split:** DETs, as before, but with individual traces for each of the ten splits present in the IJB-A set.

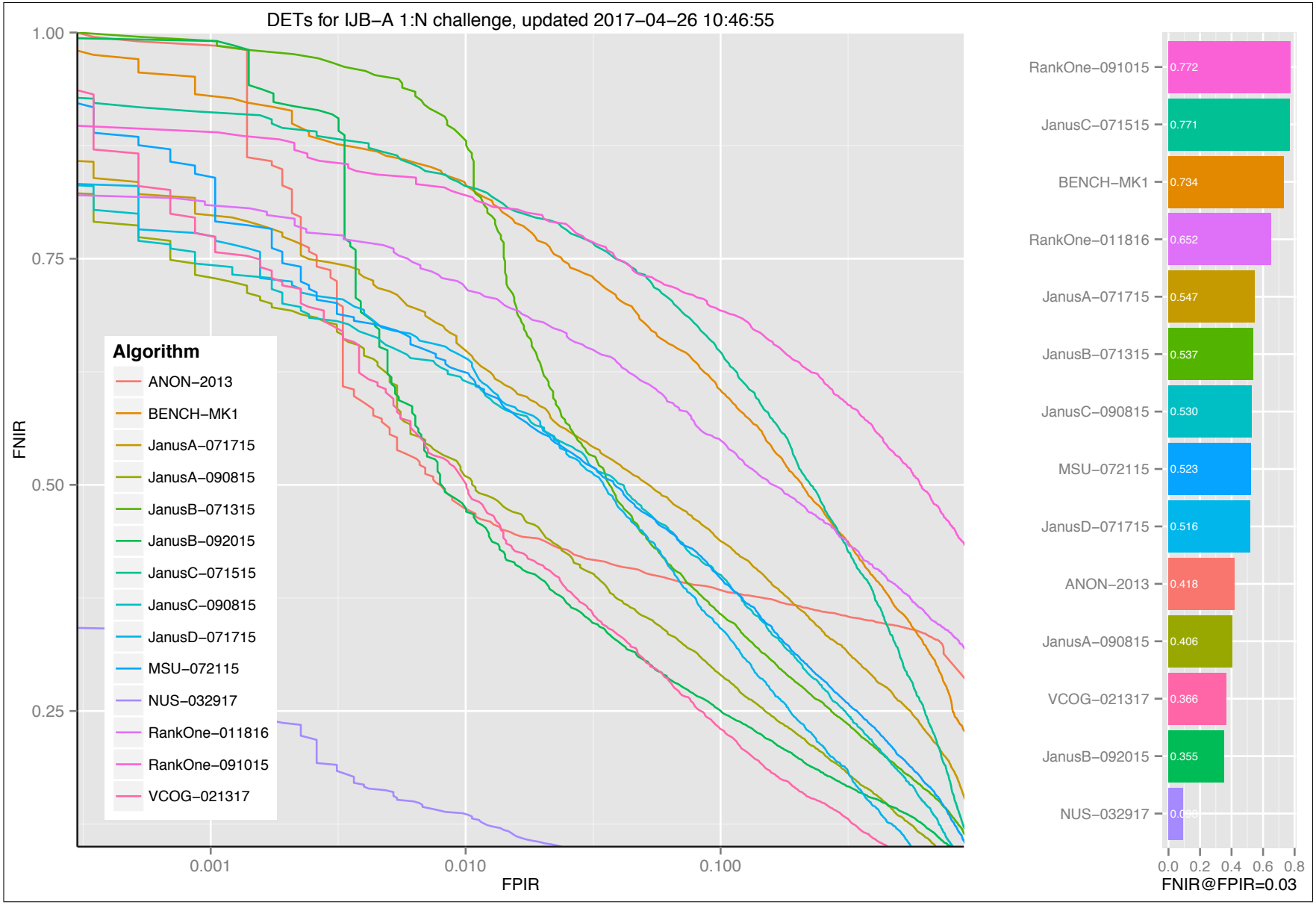

Figure 4: **DETs:** At left, are full error tradeoff characteristics (DETs) for the IJB-A algorithms. At right is a summary statistic corresponding to a vertical slice from the DETs. This is included to ease lookup. While it implies one ranking of the algorithms, it is notable that some DETs cross such that false rejection accuracy depends heavily on the FPIR operating point. Algorithm developers can shape the DET characteristics according to a known use-case. Plotting FNIR vs. FPIR, this 1:N error tradeoff characteristic differs from a traditional 1:1 DET plots of FNMR vs. FMR in potentially complicated ways - see discussion in 2.3. To first order the x-axis can be scaled by a factor of the enrolled population size N (here 500) as FMR $\sim$ FPIR / N.

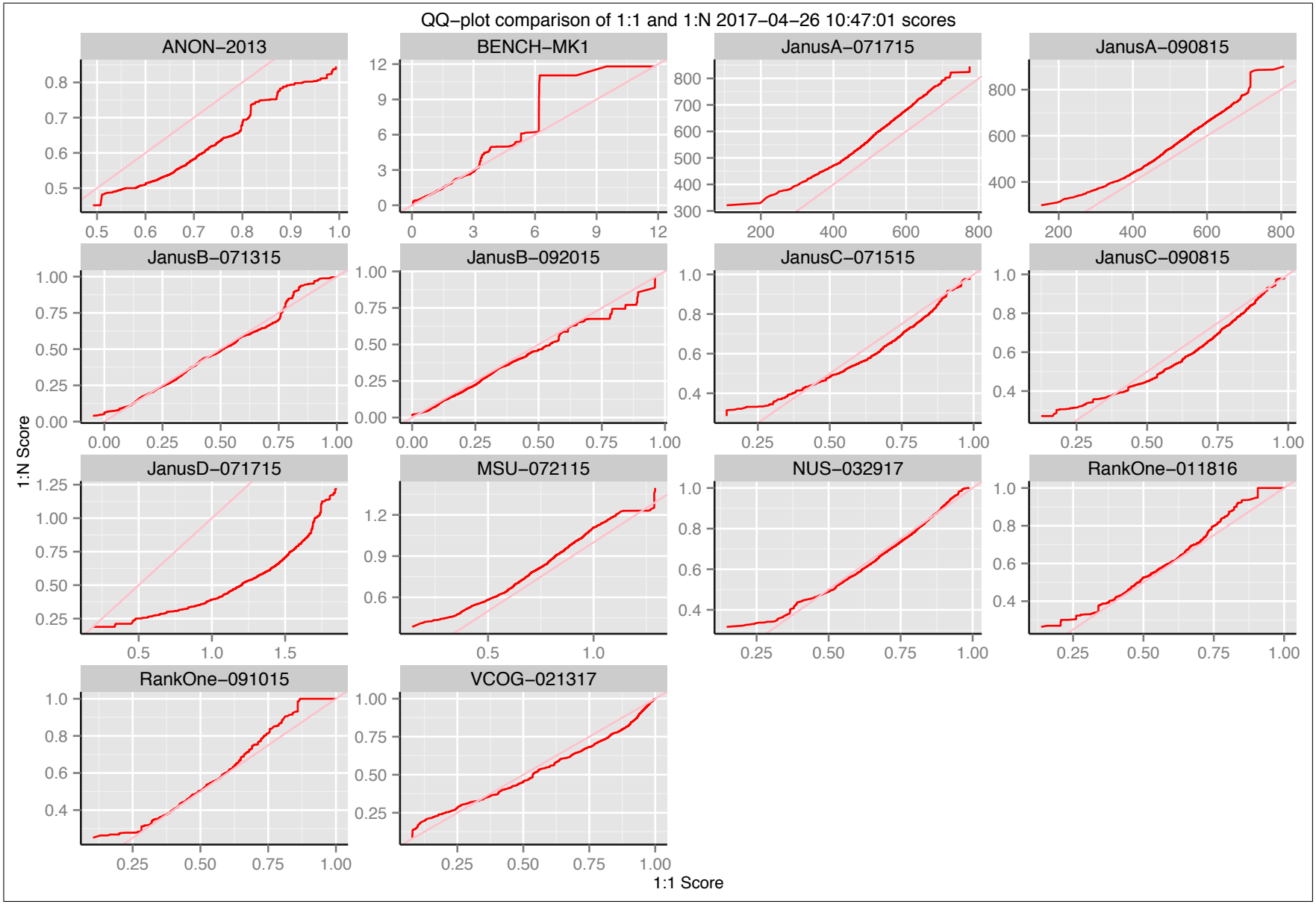

Figure 5: **Are 1:N and 1:1 equivalent**: Each panel gives a QQ plot of genuine "hit" scores from 1:N searches against genuine scores from 1:1 comparisons. If these two scores have the same distribution the red line is straight. A curved line indicates that a 1:N scores are simply the result of comparing two samples using the 1:1 algorithm.

4 References

[1] 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. 1

[2] Foto-Fahndung. Face recognition as a search tool. Technical report, Bundeskriminamt (BKA), Thaerstrasse 11, 65193, Wiesbaden, Germany, February 2007. 2

[3] Brendan F. Klare, Ben Klein, Emma Taborsky, Austin Blanton, Jordan Cheney, Kristen Allen, Patrick Grother, Alan Mah, and Anil K. Jain. Pushing the frontiers of unconstrained face detection and recognition: Iarpa janus benchmark a. In *Proc. IEEE CVPR*, June 2015. 1