Approaches to Facial Image Quality

Contributors:
DeWayne Halfen, Srinivasan Rajaraman
and James Wayman
Office of Biometric Identity Management

obim@hq.dhs.gov

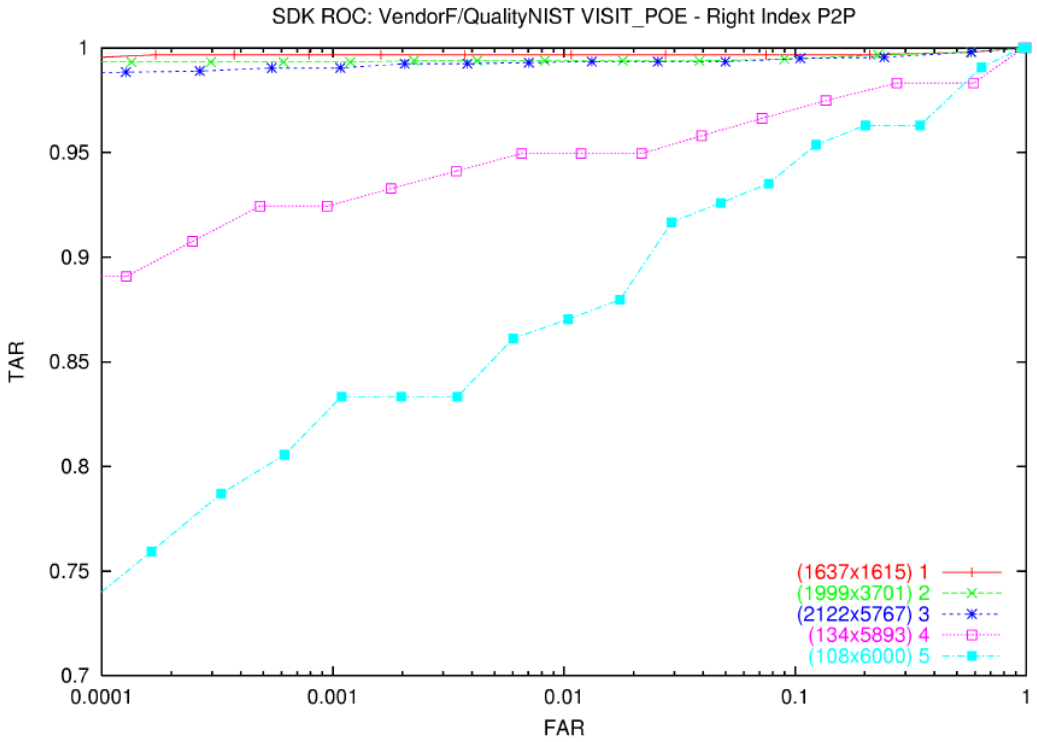
Usual Caveats: This presentation represents our scientific findings and opinions only, and not necessarily the position of the US government

Defining Fingerprint Image Quality

- "We define fingerprint image quality as a predictor of matcher performance before a matcher algorithm is applied. This means presenting the matcher with good quality fingerprint images will result in ***high matcher performance***, and vice versa, the matcher will perform poorly for poor quality fingerprints…Predicting matcher performance is also valuable for biometric fusion of multiple fingerprints because the fingerprints with ***the best image quality can be assigned higher weight in the fusion***" – E. Tabassi, C.L. Wilson, C.I. Watson, "Fingerprint Image Quality", NISTIR 7151, August 2004 ***(bolded italics added)***

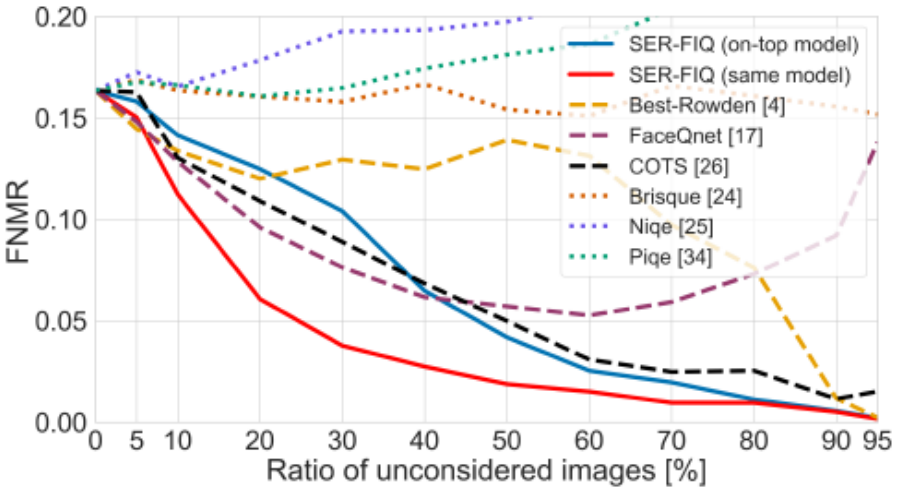
Defining "Matcher/Comparator Performance"

- "The DET, and the equivalent ROC, are the commonest statement of performance of a verification system….Therefore, quality measure q should be indicative of the degree to which the match distribution M (s_m) is separated from the non-match distribution N(s_n)." -- NISTIR 5171



Other Definitions of Comparator Performance

- "Quality values are most useful as **predictors of false negative outcomes**, arising from low genuine scores. The alternative, as predictors of false positives, is considered less feasible because these arise from high impostor scores which should result only from facial (e.g. anatomical) similarity of the input image pair…**This standard requires quality algorithms to predict false negative recognition outcomes**…. " P. Grother, M. Ngan, K. Hanaoka "Face Recognition Quality Assessment: Concept and Goals v1.0" NIST 04/23/2019
- (**bold** added)



Terhorst, Philipp, et al. "SER-FIQ: Unsupervised estimation of face image quality based on stochastic embedding robustness." Proc. IEEE/CVPR, 2020.

Other Definitions of Comparator Performance

False Match Rate Only

DEA/DOJ Electronic Prescriptions for Controlled Substances
Interim Final Rule with Request for Comments

"…after extensive consultation with NIST, and based on NIST recommendations… . The biometric subsystem must operate at a false match rate of 0.001 or lower. …DEA is not establishing a false non-match (rejection) rate; while users may be interested in this criterion, DEA does not have an interest in setting a requirement for a tolerance level for false rejections for electronic prescription applications….This testing must be performed by the National Institute of Standards and Technology (NIST) or another DEA approved (government or nongovernment) laboratory."

16236 Federal Register / Vol. 75, No. 61 / Wed, Mar.31, 2010 / Rules and Regulations/ Dept of Justice Drug Enforcement Administration
21 CFR Parts 1300, 1304, 1306, and1311

Resistance to Presentation Attack Instruments exploiting data noise

Quality Comes in Pairs

The current received theory for fingerprints (NISTIR 7151):

Therefore, pair wise quality Q as defined below, should be predictive of recognition performance of pair ($x_{gallery}$, x_{probe}) .

$$Q = H(q_{gallery}, q_{probe}) \qquad \text{(eq. 5)}$$

Extensive testing at NIST…has shown that recognition errors are triggered by low quality samples. That is, H(.) is simply a min function of the individual numbers q_{probe} and $q_{gallery}$, and so pair wise quality is defined in equation 6.

$$Q=\min(q_{gallery}, q_{probe}) \qquad \text{(eq. 6)}$$

(Notation of 7151 altered slightly)

Question: Does this apply if $q_{gallery}$ and q_{probe} are measures such as pose angle or inter-ocular distance?

Other Reduction Methods Useful for Facial Image Quality

- Max: $Q=\max(q_{gallery}, q_{probe})$
- Mean or sum: $Q=(q_{gallery}+q_{probe})/2$
- Harmonic mean: $Q=\left[\frac{1}{2qg}+\frac{1}{2qp}\right]^{-1}$
- Difference: $Q=\|q_{gallery}-q_{probe}\|$
- Probe only: $Q=q_{probe}$

Can Poor Quality Images Be Discarded?

- At the point of collection, if more images available
- OBIM provides facial comparison services to other groups who supply images and act on outcomes.
- OBIM does not collect facial images.
- Generally, OBIM cannot discard any images.
- OBIM facial recognition services use monomodal fusion.
- Quality metrics must support the NISTIR 7151 vision that “the best image quality can be assigned higher weight in the fusion”

Our Experimental Data

- 20k mated face image comparison scores
- 200k non-mated face image comparison scores
- 50% of both data sets sequestered for results validation
- All scores accompanied with array of 12 “quality” metrics for each image in each comparison.
- Some values are the same for all images encountered.
- Names given these metrics are not assumed to have ”reified” meaning
- Metrics are uncorrelated but not independent

“GenQual”; “Face”; "Frontal"; “Yaw”; “Roll”; “Pitch”; rEye x; rEye y; lEye x; lEye y; Eye Dist; “Qual”

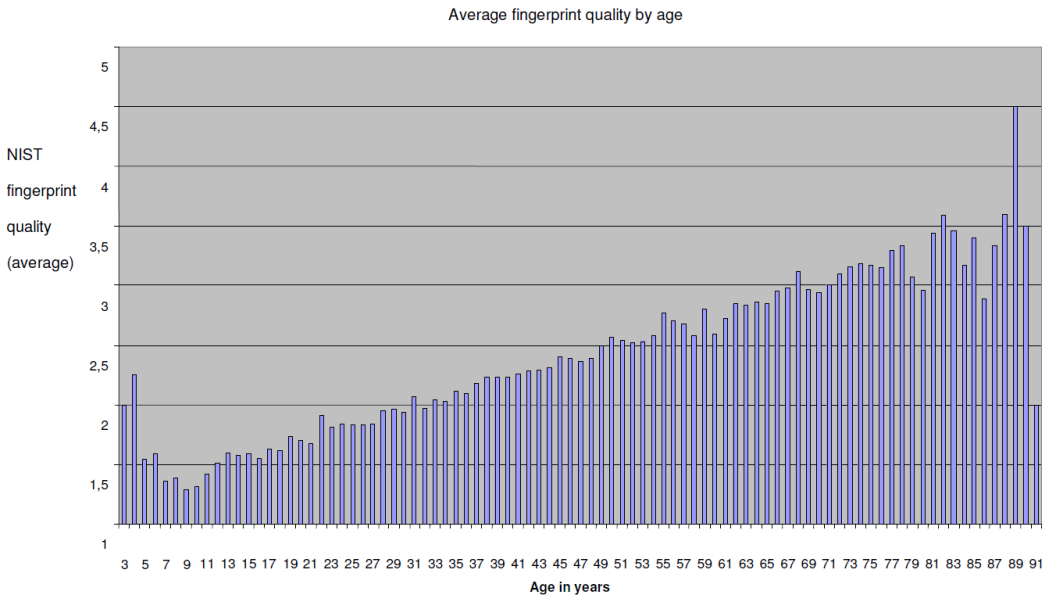
If It's Reified, It's Potentially Actionable

- Reified quality metrics:
  - Pose angle (yaw, pitch, roll)
  - Motion blur
  - Focus blur
  - Illumination variables
  - Interocular distance

- Non-reified quality metrics
  - Faceness
  - Overall quality
  - Frontal

Hypothesis: Reified metrics have fewer hidden demographic biases

Quality and Demographic Biases of the Comparison Algorithm



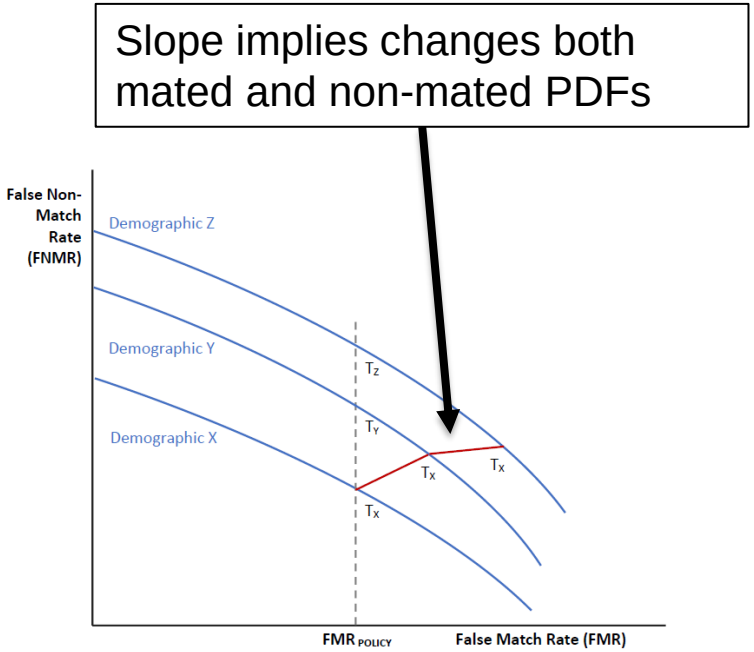


Figure 29: The figure shows the effect of setting thresholds to achieve the target FMR on demographics X and Y.

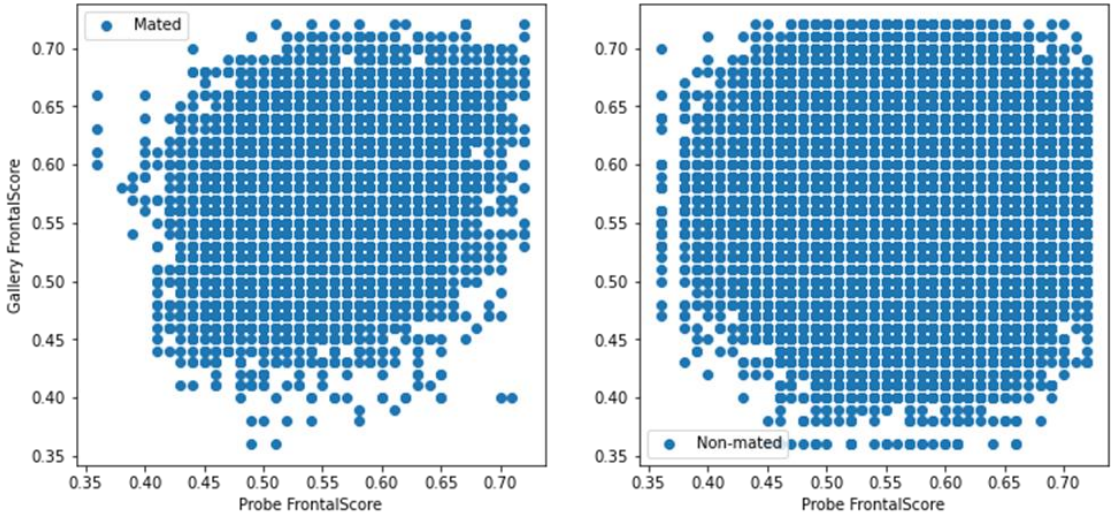
"Evaluation Report Biometrics Trial 2b or not 2b", Dutch Ministry of the Interior and Kingdom Relations, Tech. Rep., 2004 http://www.dematerialisedid.com/PDFs/88_630_file.pdf

P. Grother, M.Ngan, and K.Hanaoka. "Face Recognition Vendor Test: Part 3, Demographic Effects", NISTIR 8280, 2019. https://nvlpubs.nist.gov/nistpubs/ir/2019/NIST.IR.8280.pdf

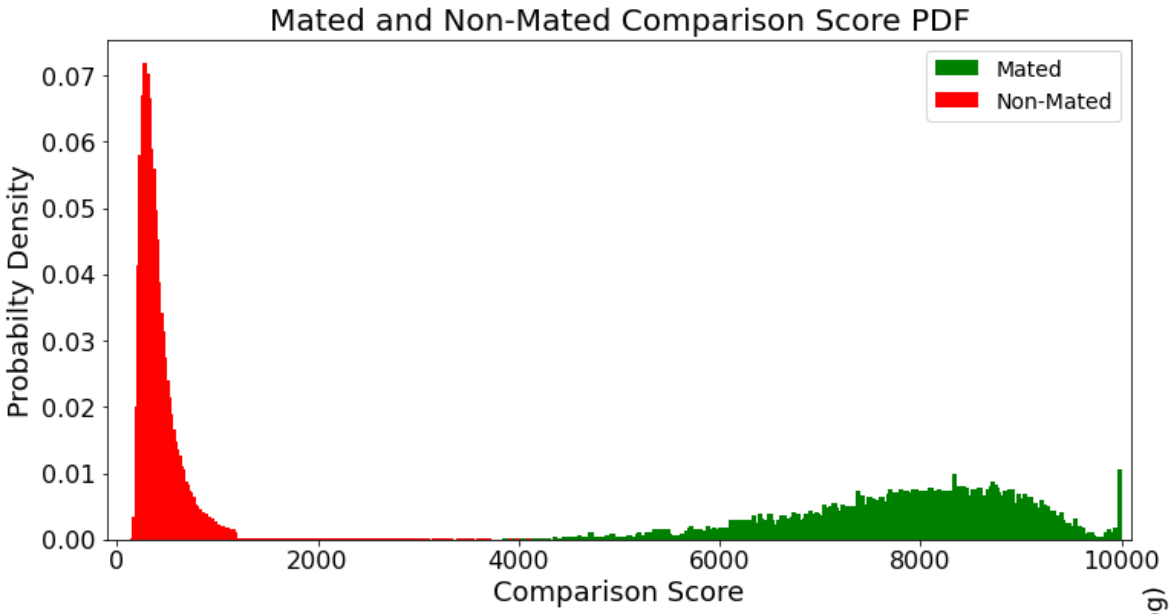
Persistent Quality Metrics

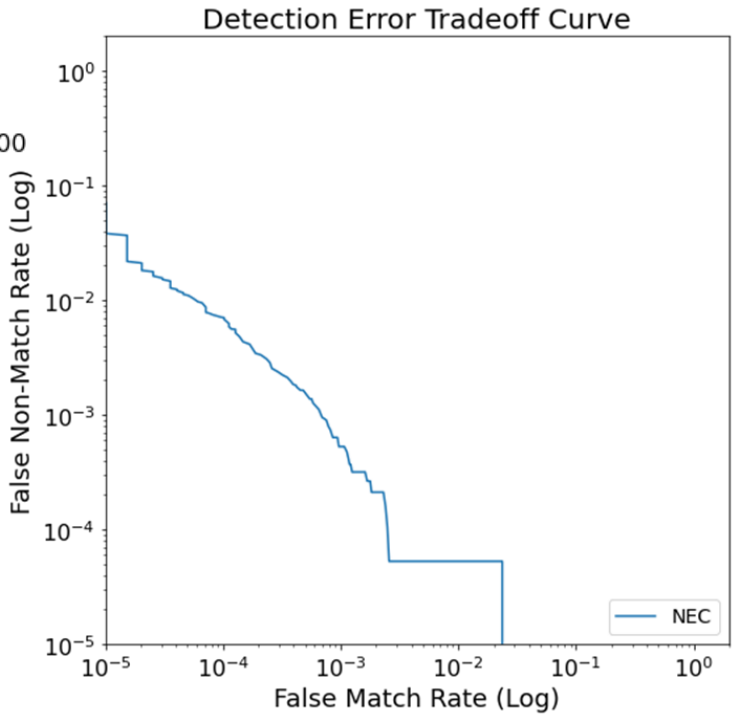
	Mated	Non-mated
OverallQuality	0.003900	-0.001022
FaceScore	0.115536	0.001704
FrontalScore	0.307005	0.000693
Pan	0.077105	0.004799
Roll	0.082337	0.002072
Tilt	0.253753	0.004872

	Mated	Non-mated
RightEyeX	0.122837	-0.003684
RightEyeY	0.127075	-0.004903
LeftEyeX	0.116512	-0.003616
LeftEyeY	0.126789	-0.004618
EyeDistance	0.115344	-0.001773
QualityScore	0.185734	0.001354



Quality Agnostic Data



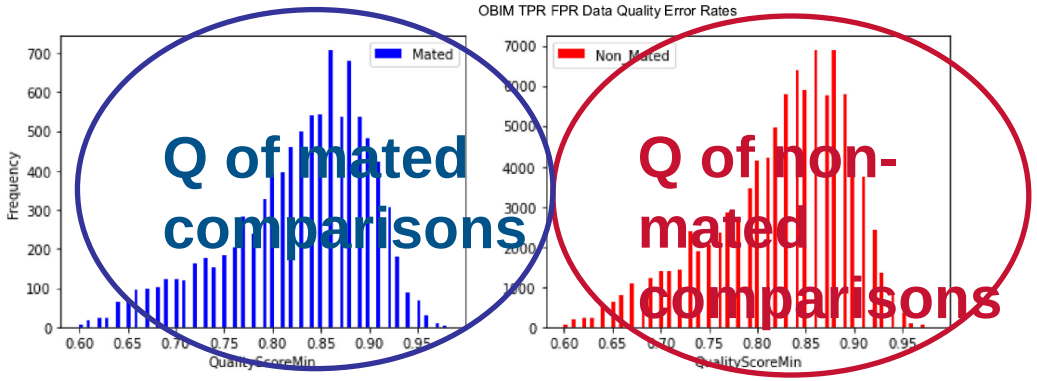


Evaluation of Individual Quality Metrics

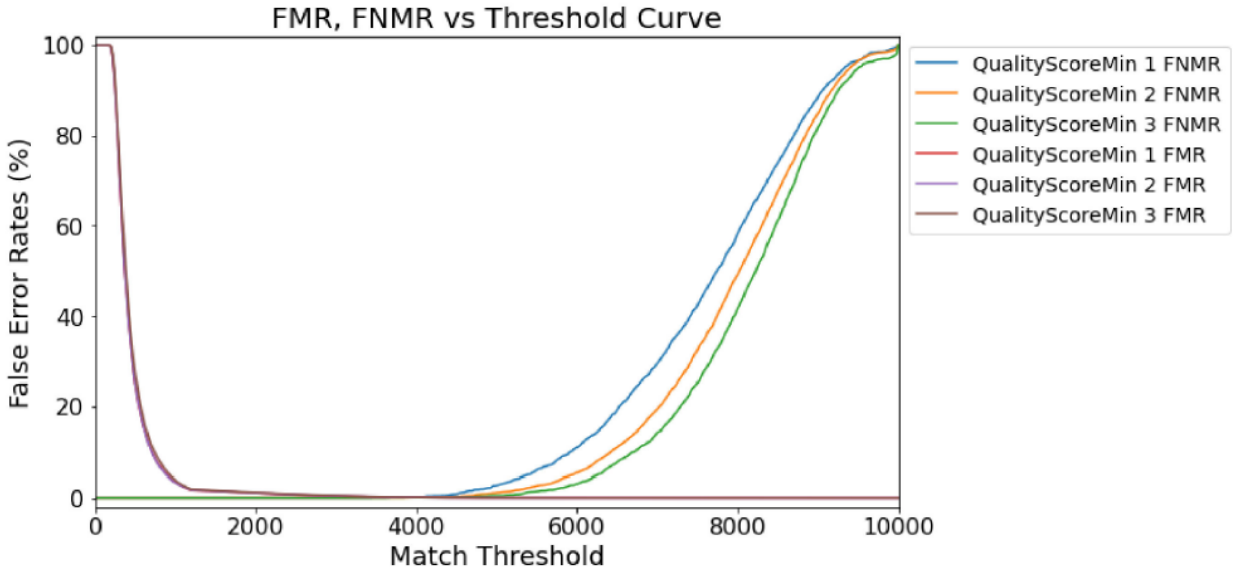
GenQual; Face; "Frontal"; Yaw; Roll; Pitch; rEye x; rEye y; lEye x; lEye y; Eye Dist; Qual

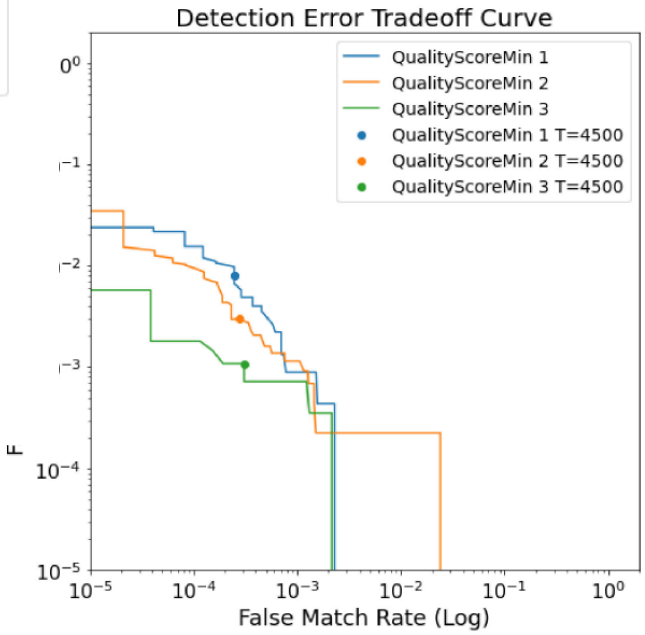
- Each image pair (except “Face” and “Gen Qual”, which were nearly static) evaluated under min, mean, max, harmonic mean, difference rules
- Values partitioned into N quality levels: Low,… Medium…., High
  - Generally, 3 levels displays best: (0-25)%,(25-75)%,(75-100)%
- Cumulative mated and non-mated distributions and DET evaluated using percentile approach
- Best metric, function for Q and number of partitions N all depend upon figure of merit (false match rate, false non-match rate, DET)

Key to the Next Slides

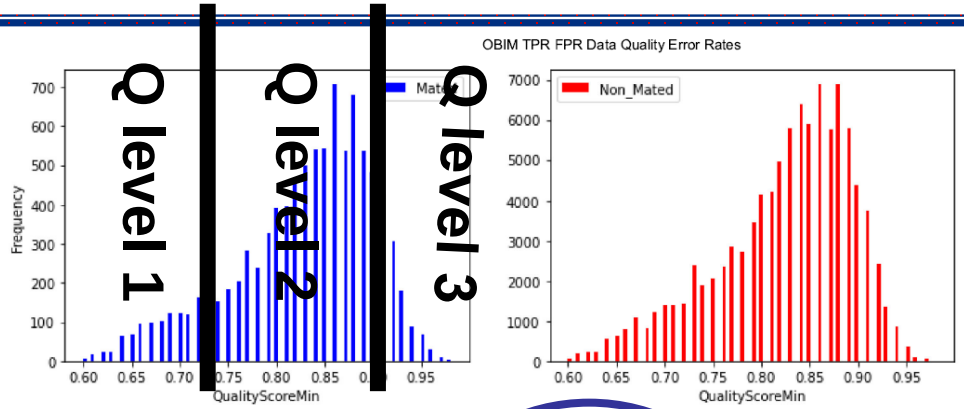


Comparison score PDFs not shown

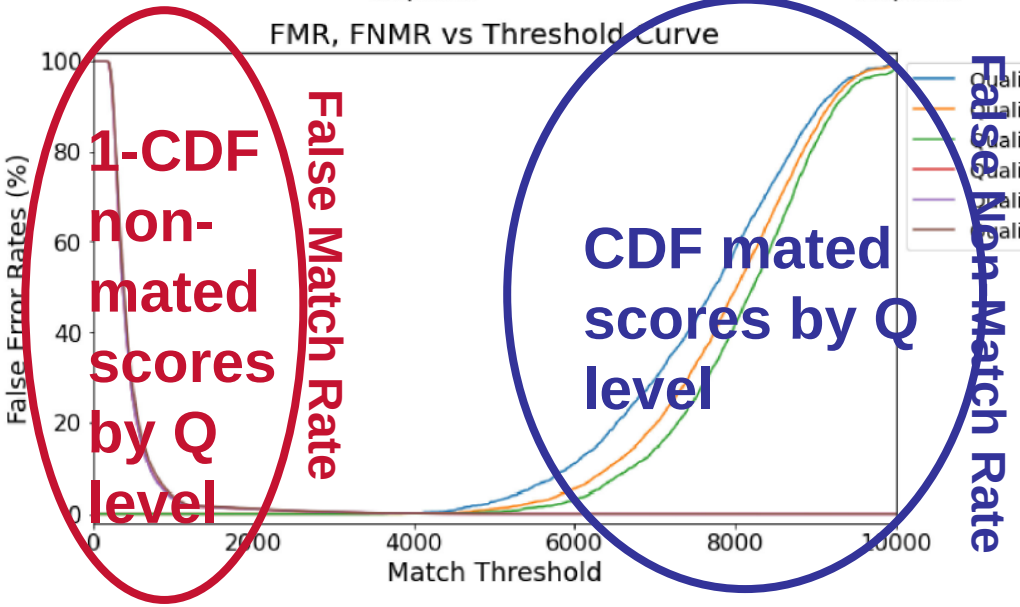


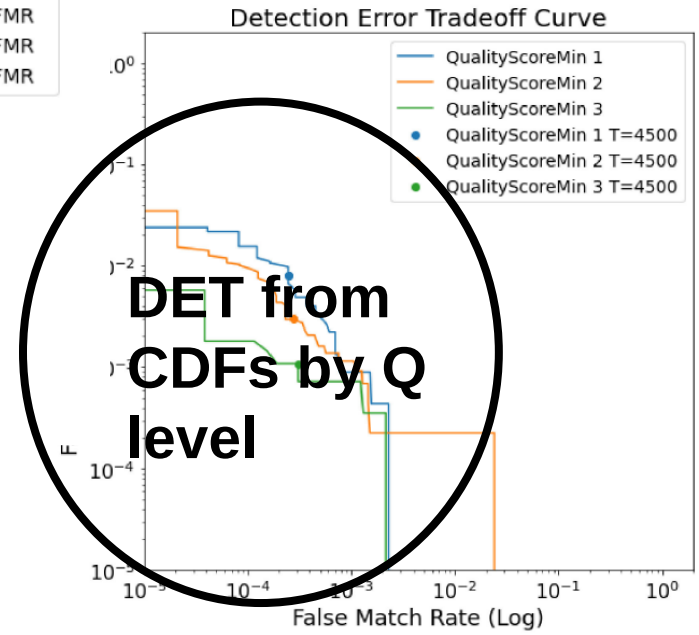


Key to the Next Slides

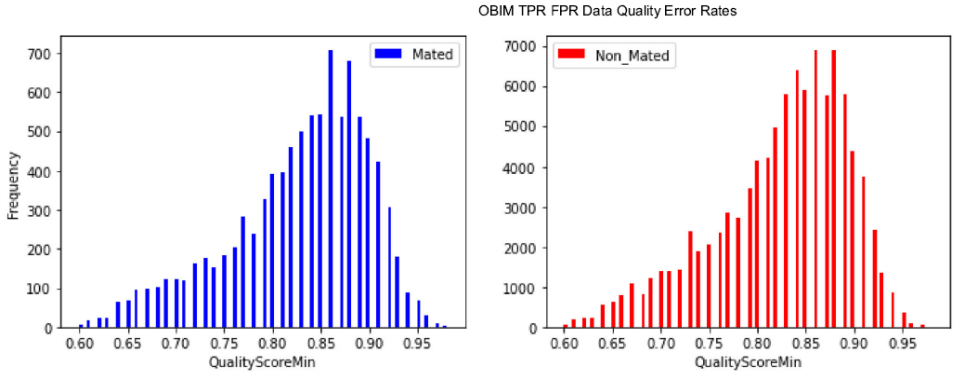


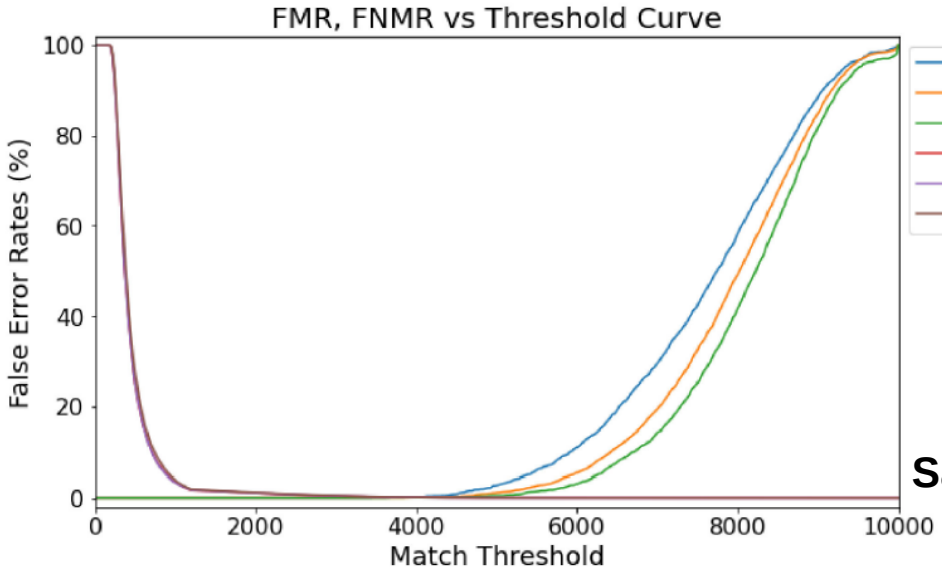
Comparison score PDFs not shown



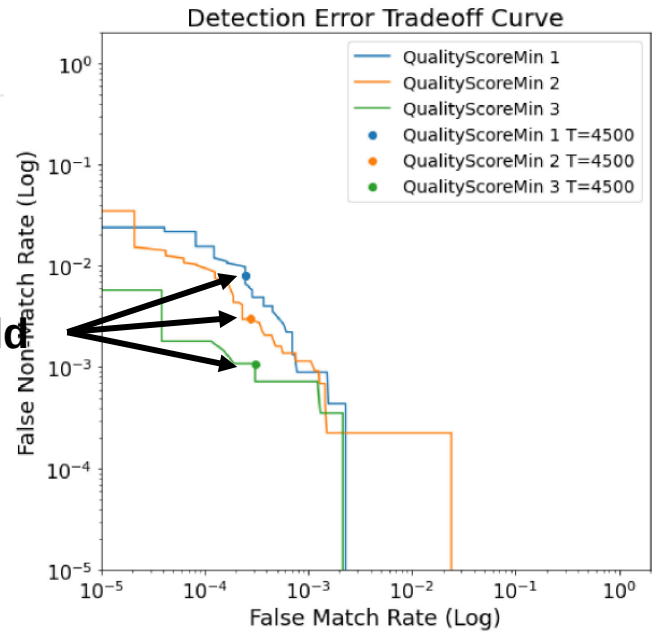


“Quality” Score: Q=Min

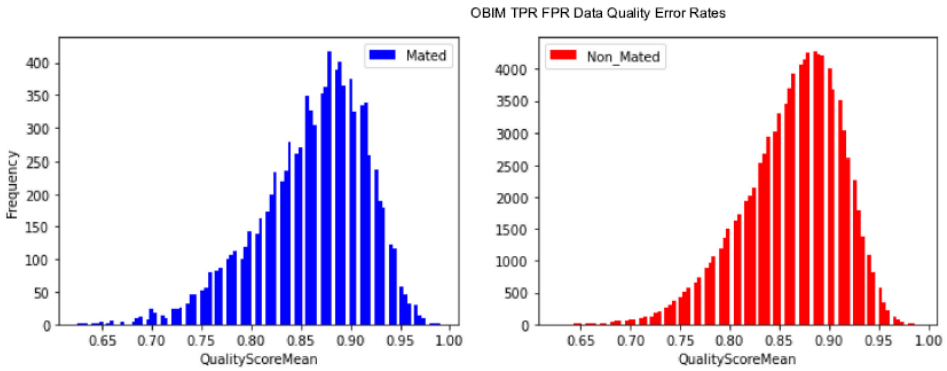


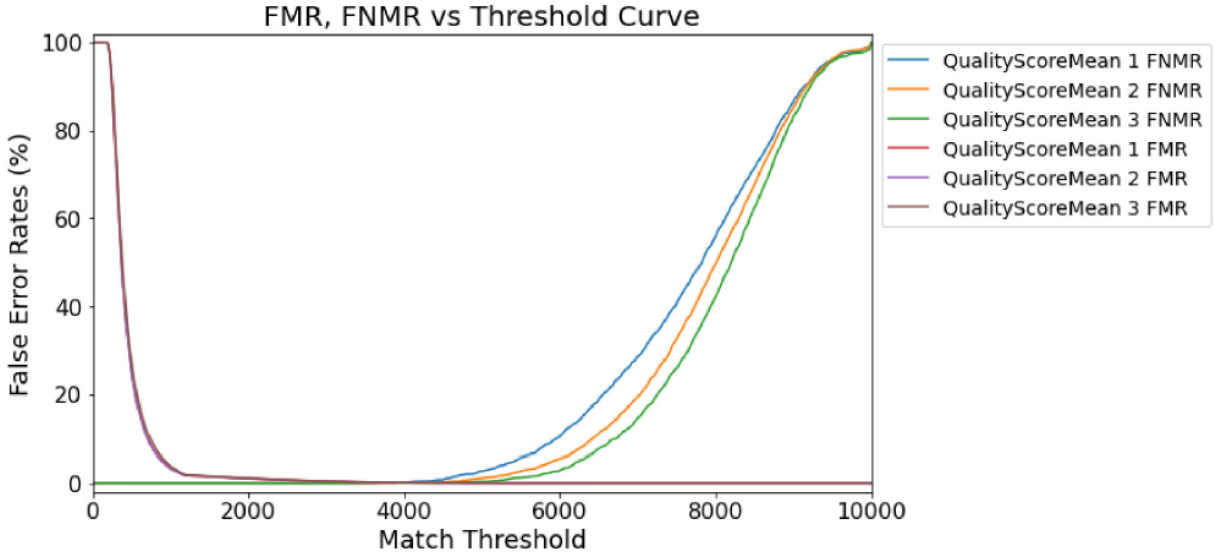


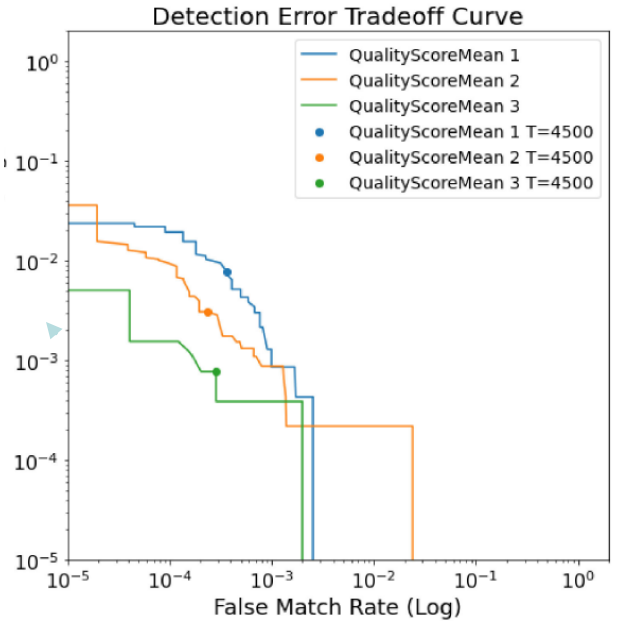
Same threshold



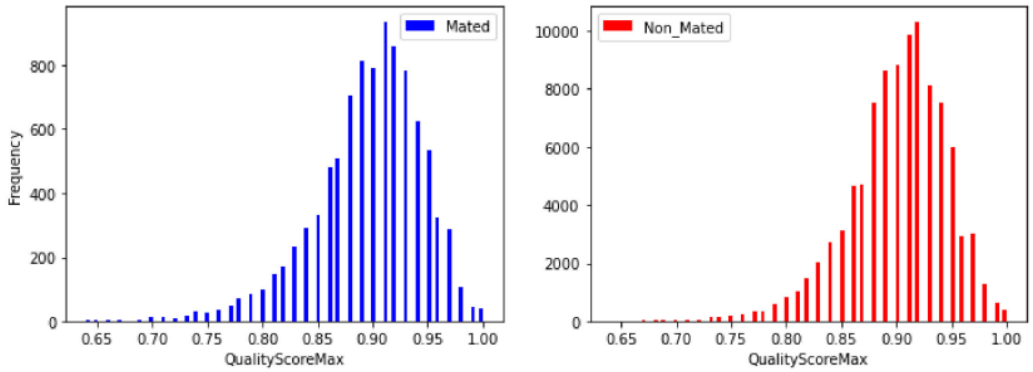
“Quality” Score: Q=Mean

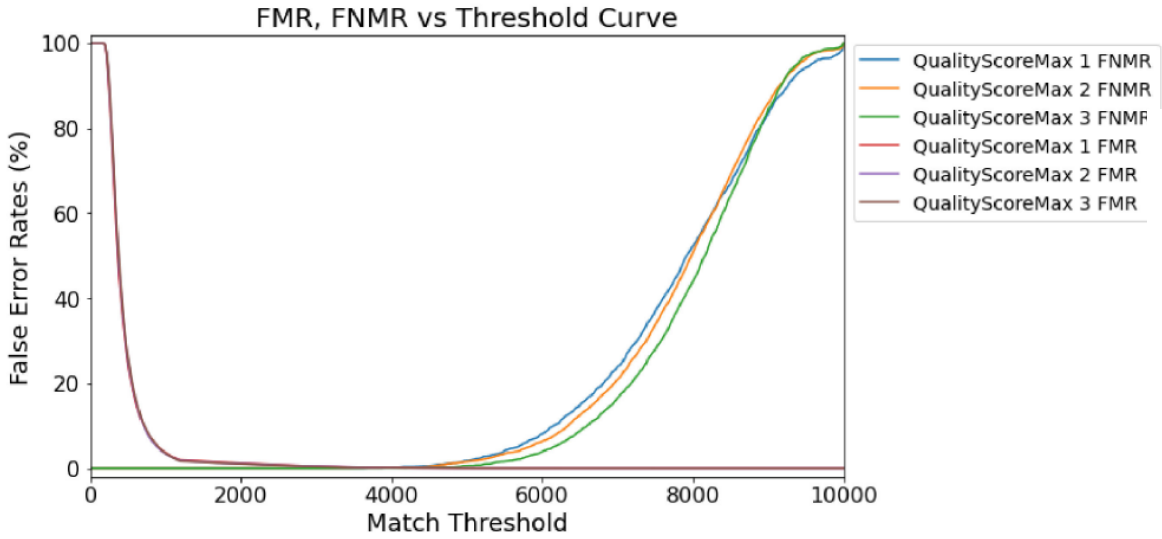


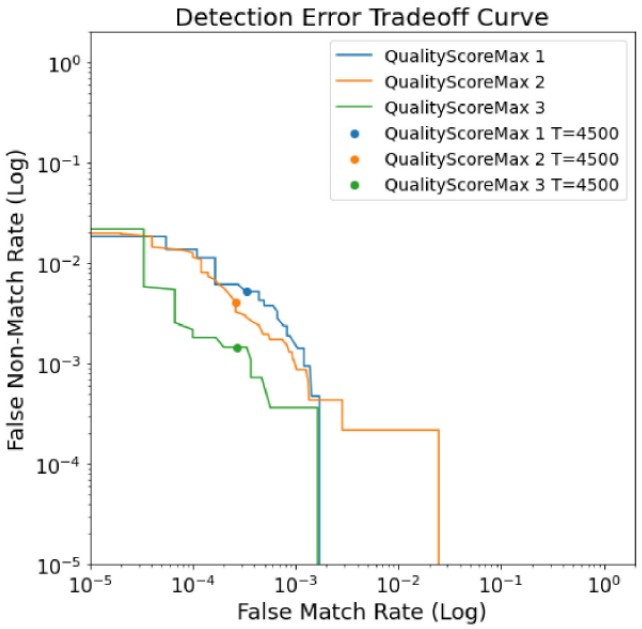




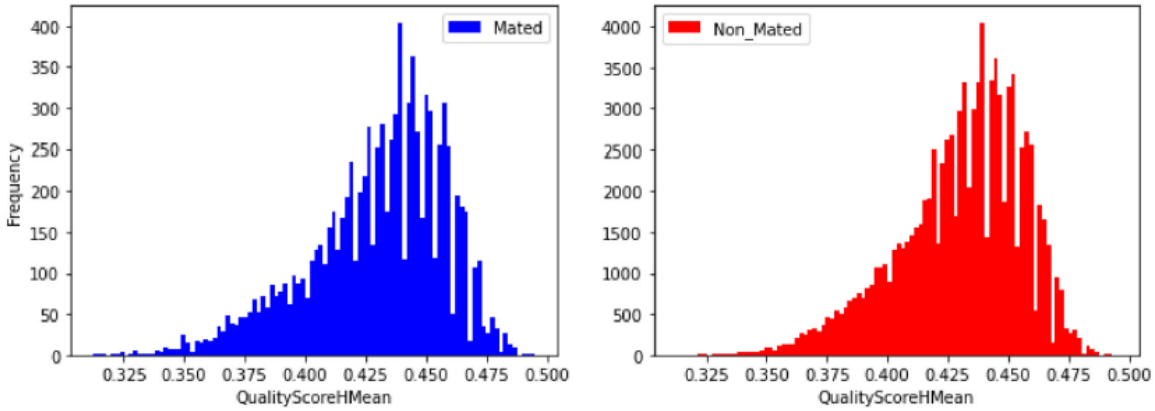
“Quality” Score: Q=Max

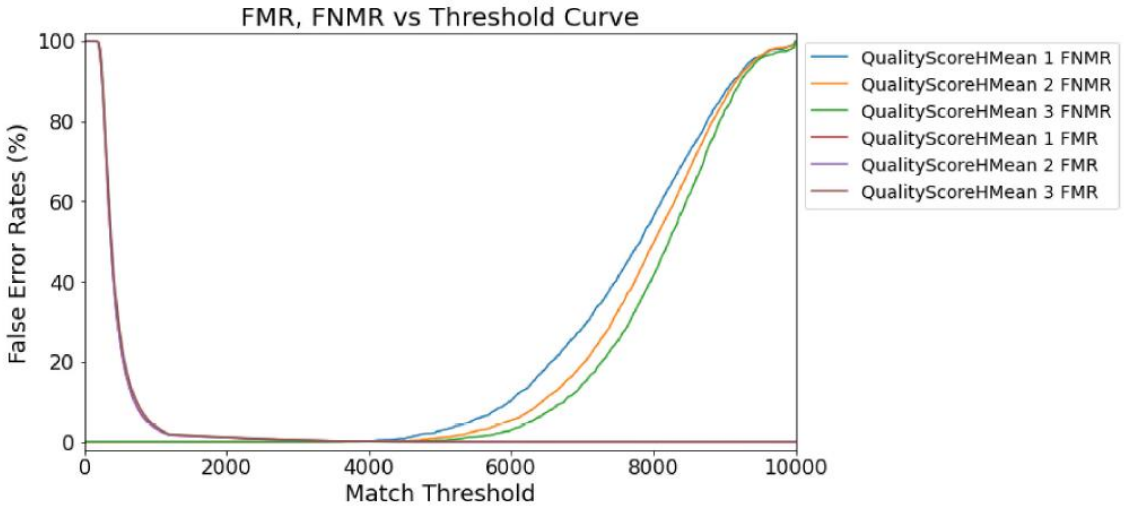


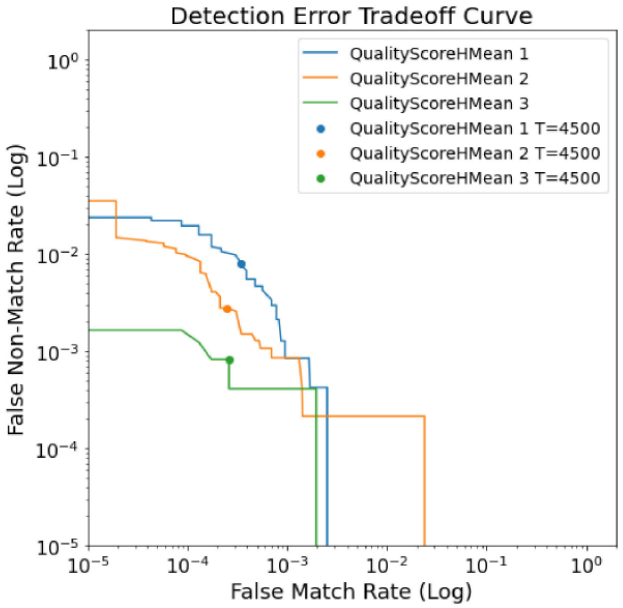




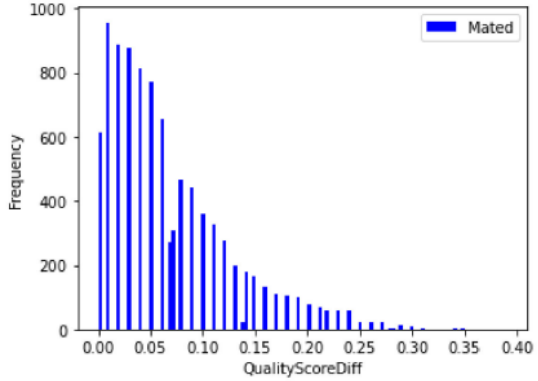
“Quality” Score: Q=Harmonic Mean

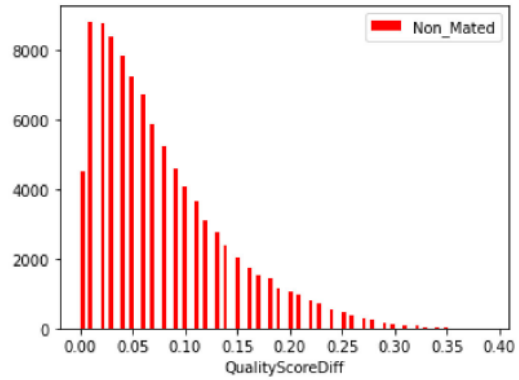


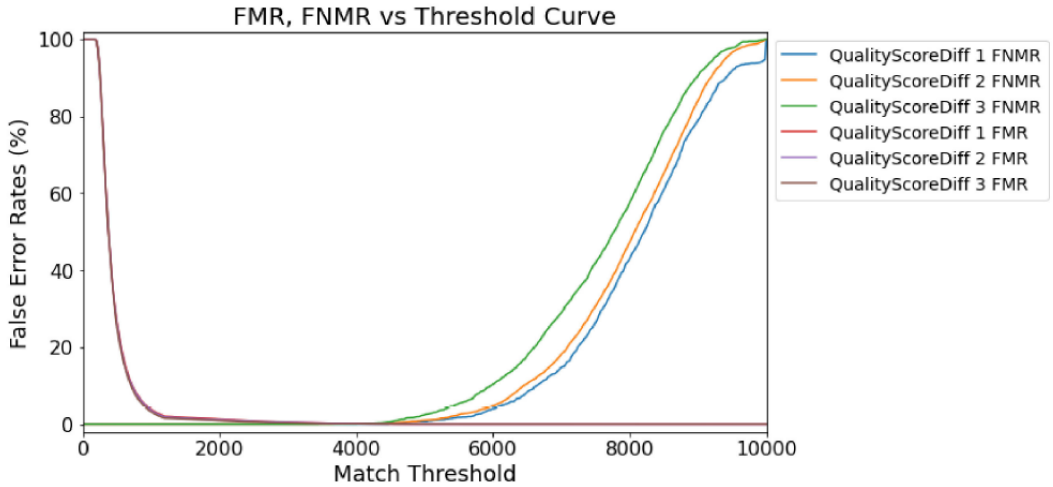


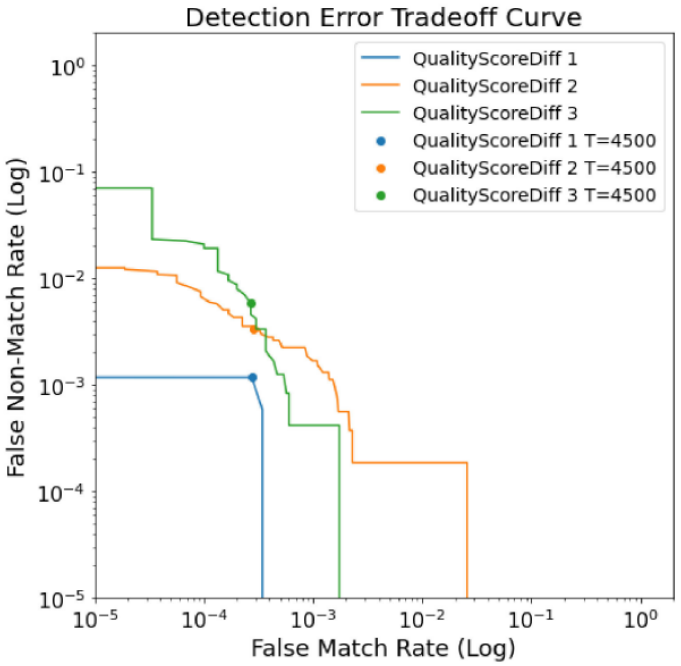


“Quality” Score: Q=Diff

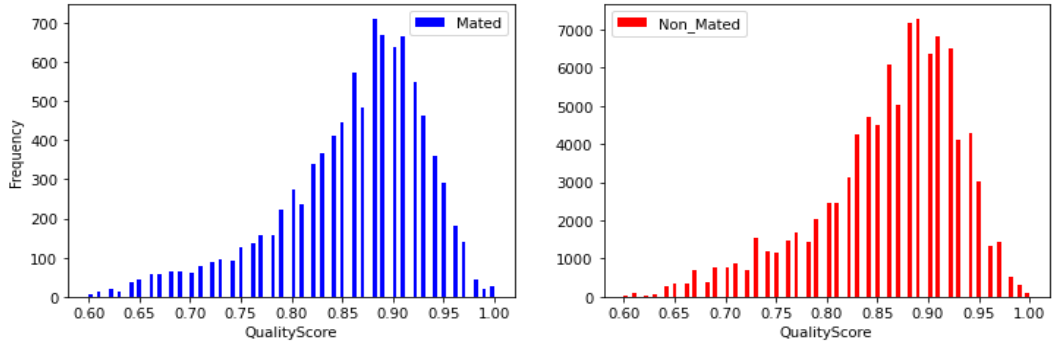


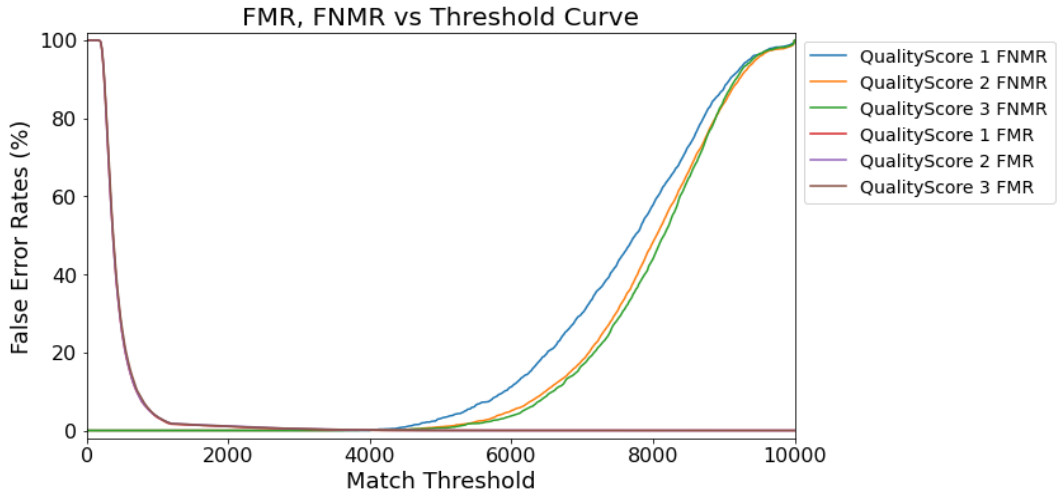


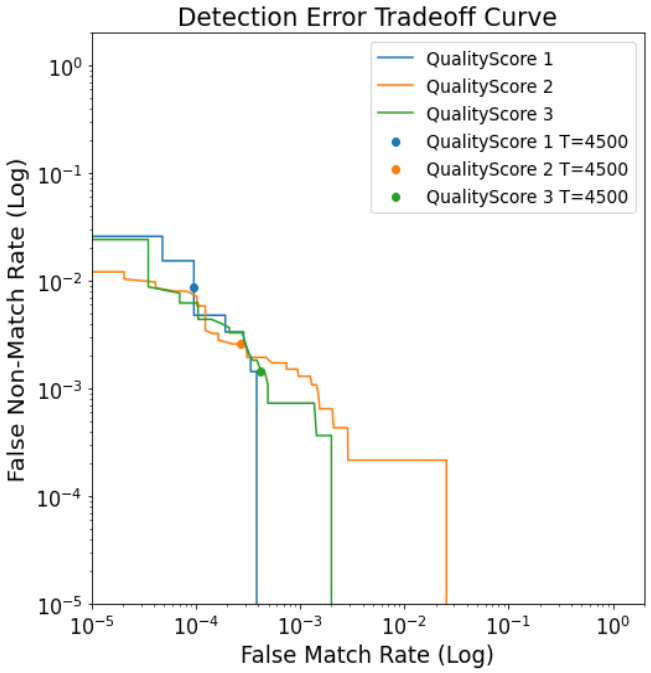




“Quality” Score: Q=Probe only

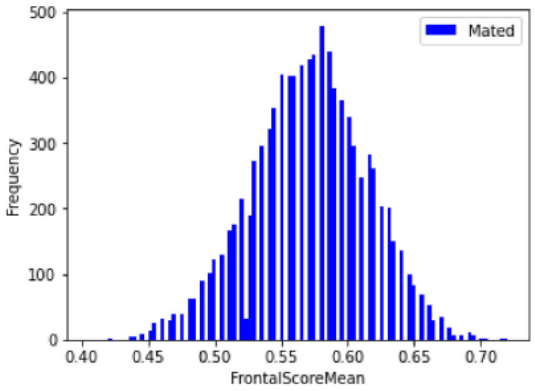


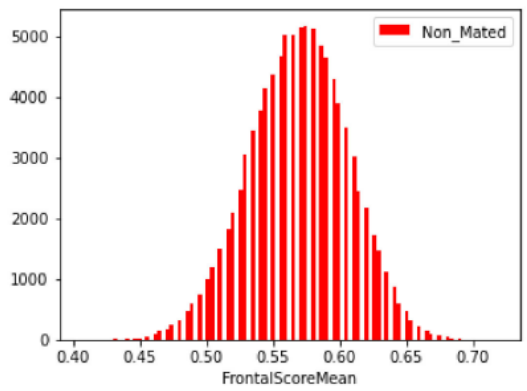


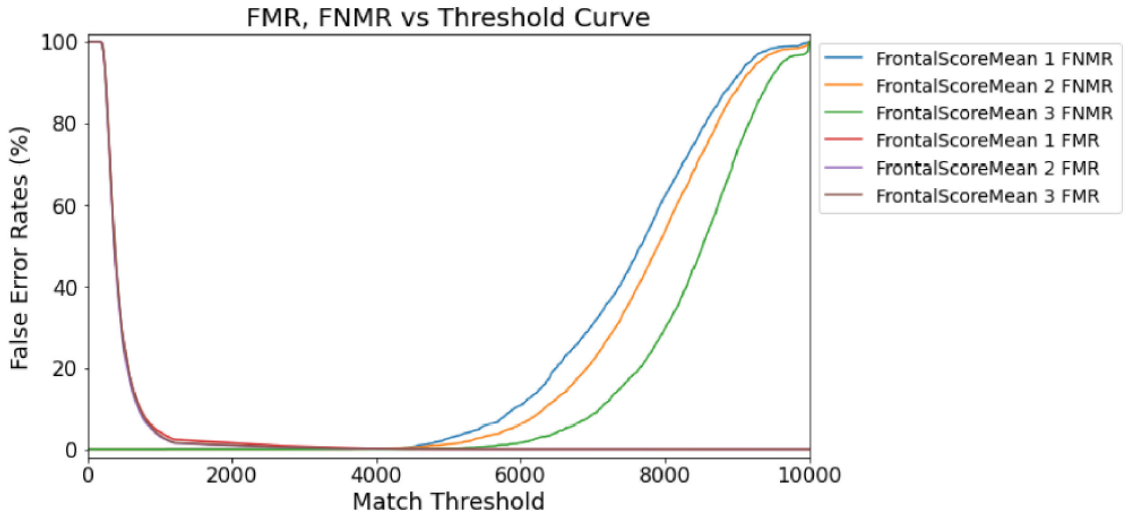


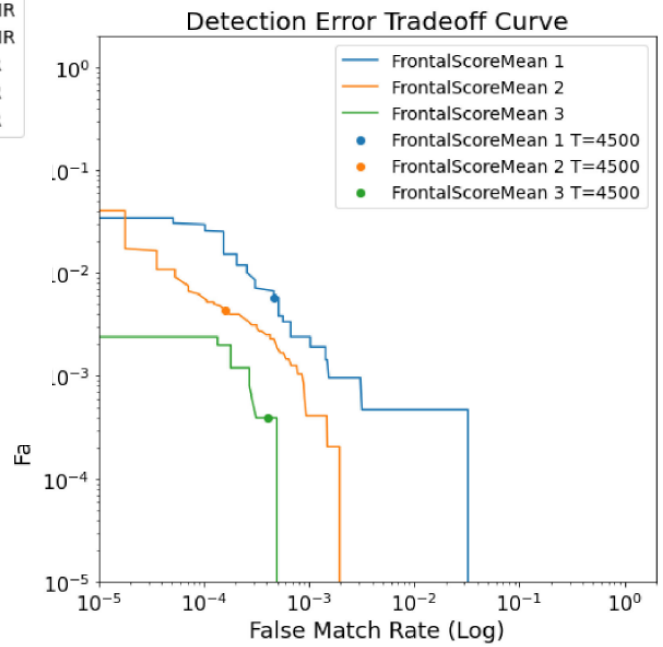
A Few of Our Favorites

"Frontal" Score: Q=Mean

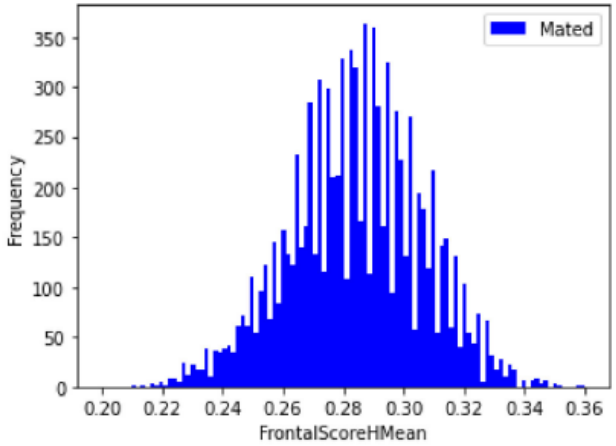


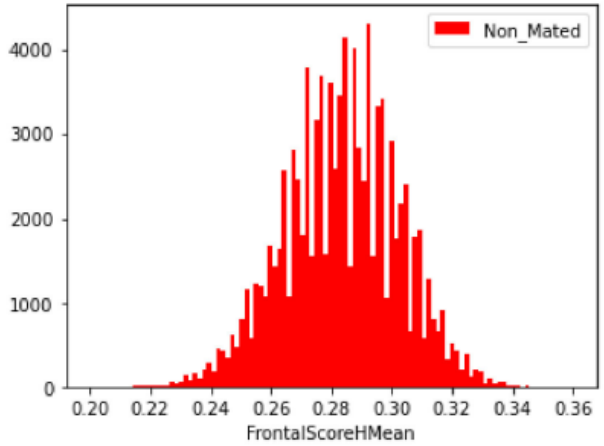


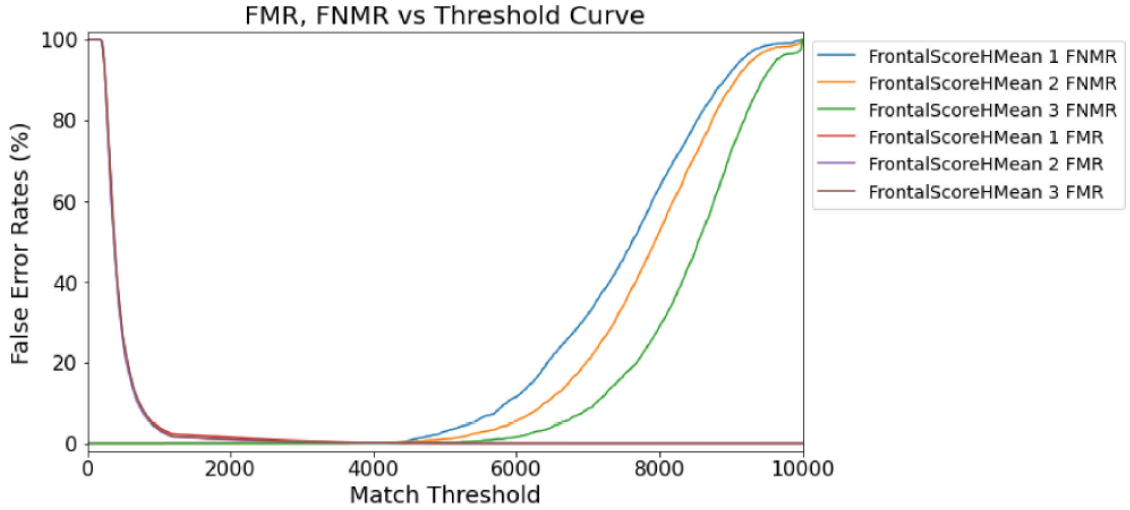


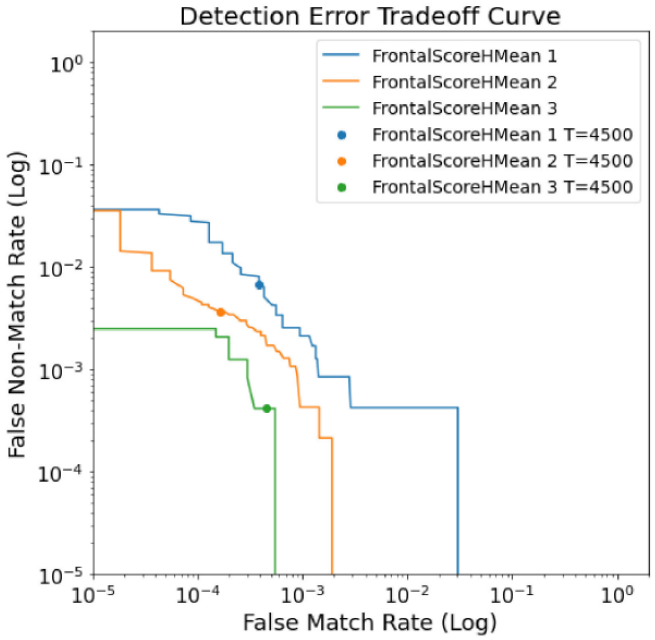


"Frontal" Score: Q=Harmonic Mean

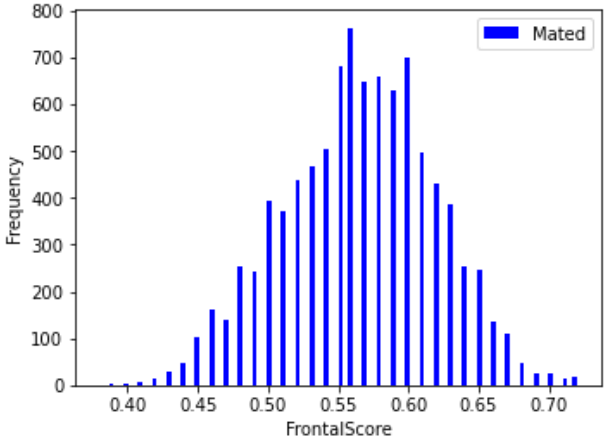


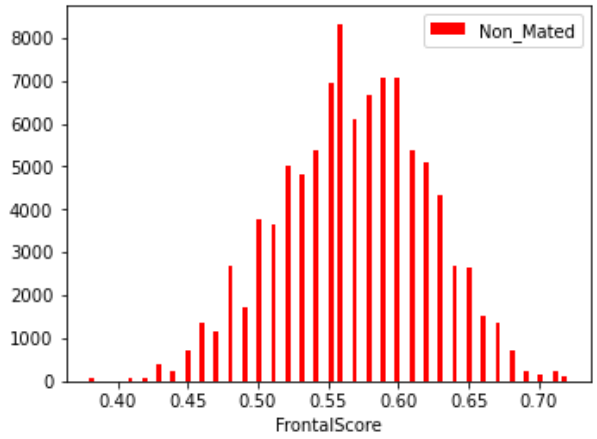


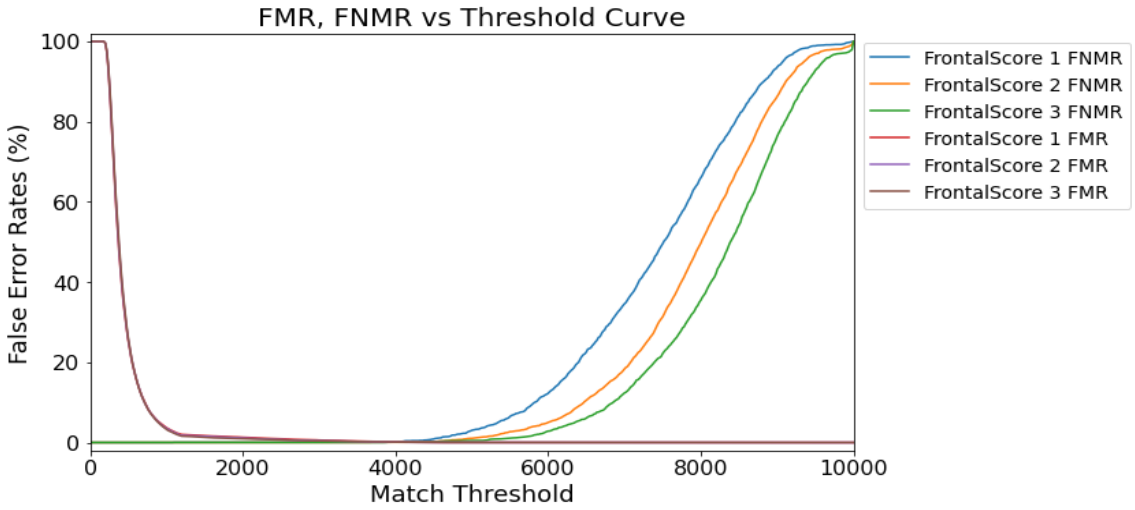


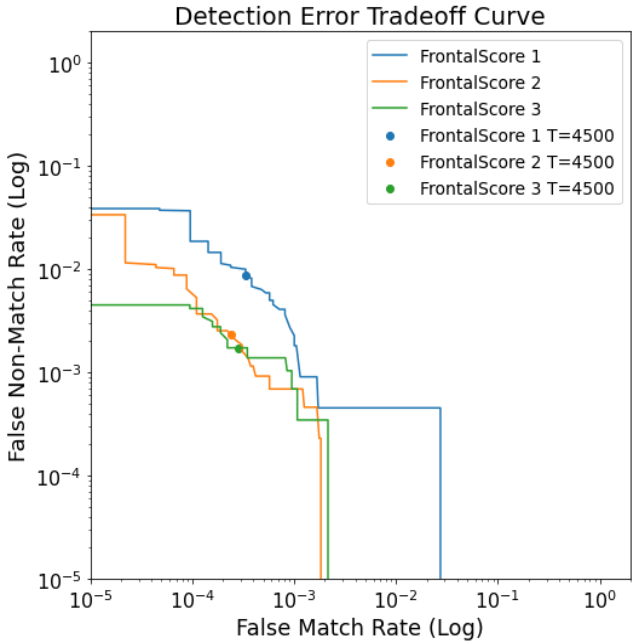


"Frontal" Score: Q=Probe only

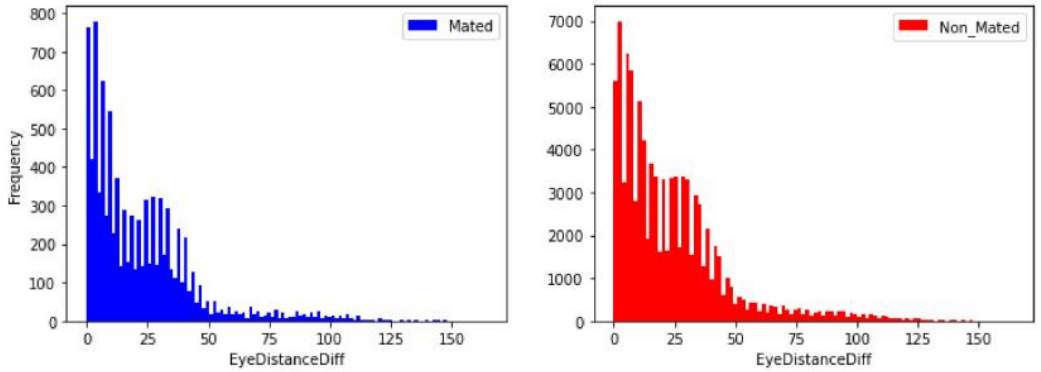


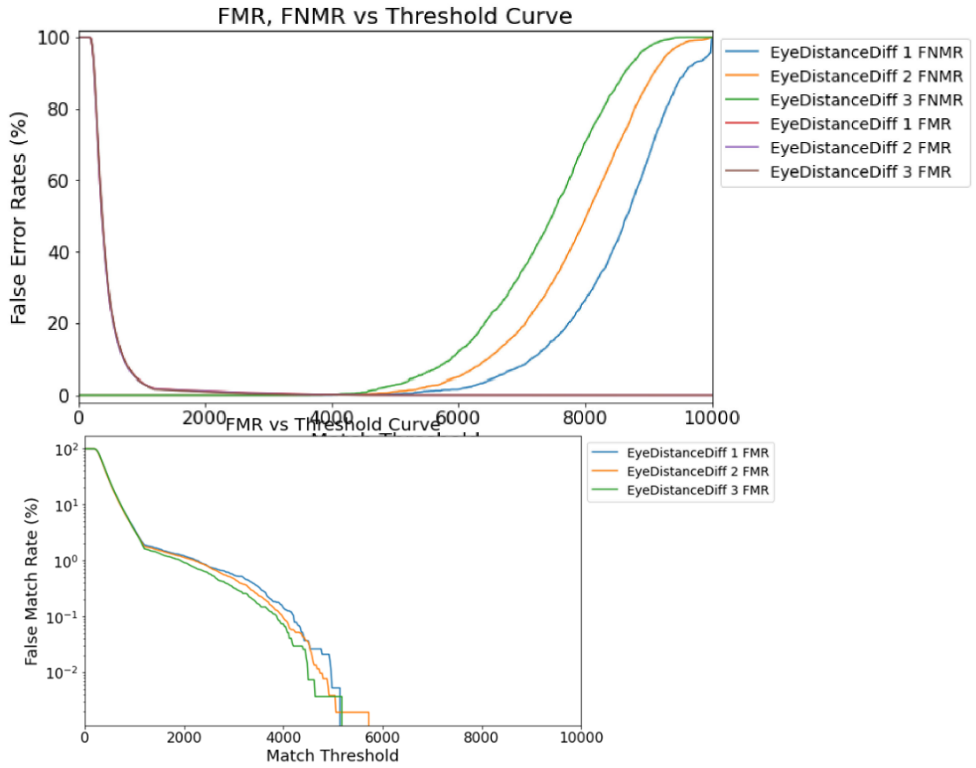


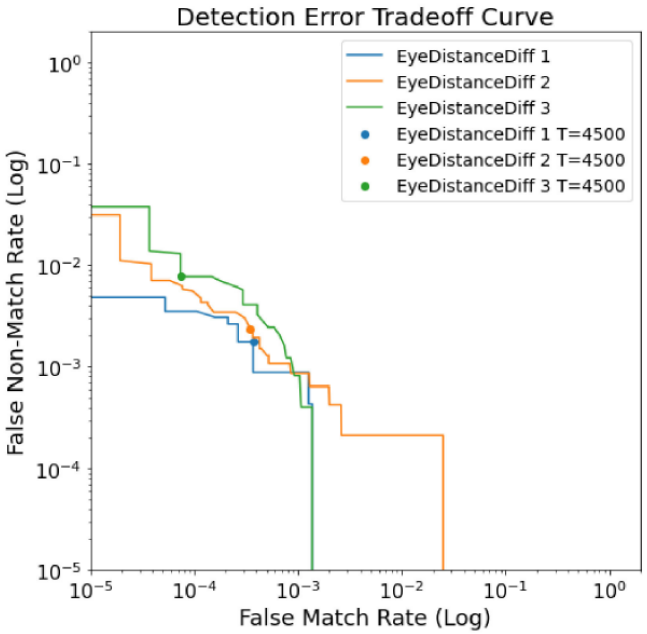




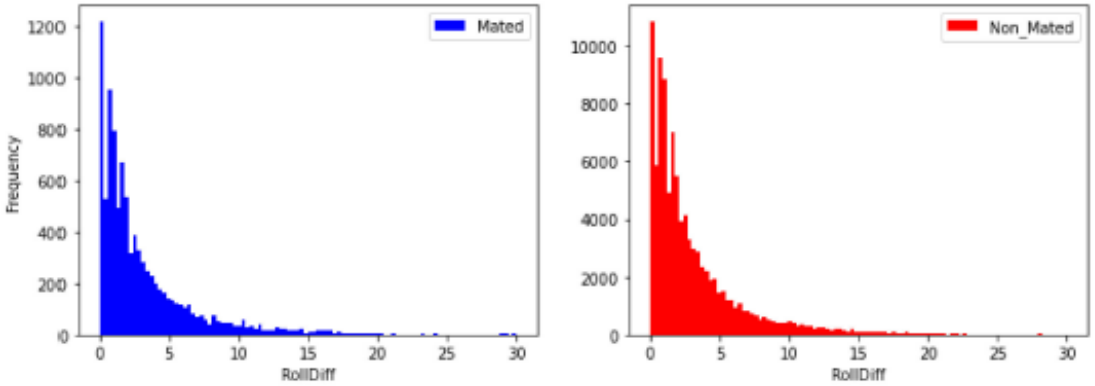
Eye Distance: Q=Diff



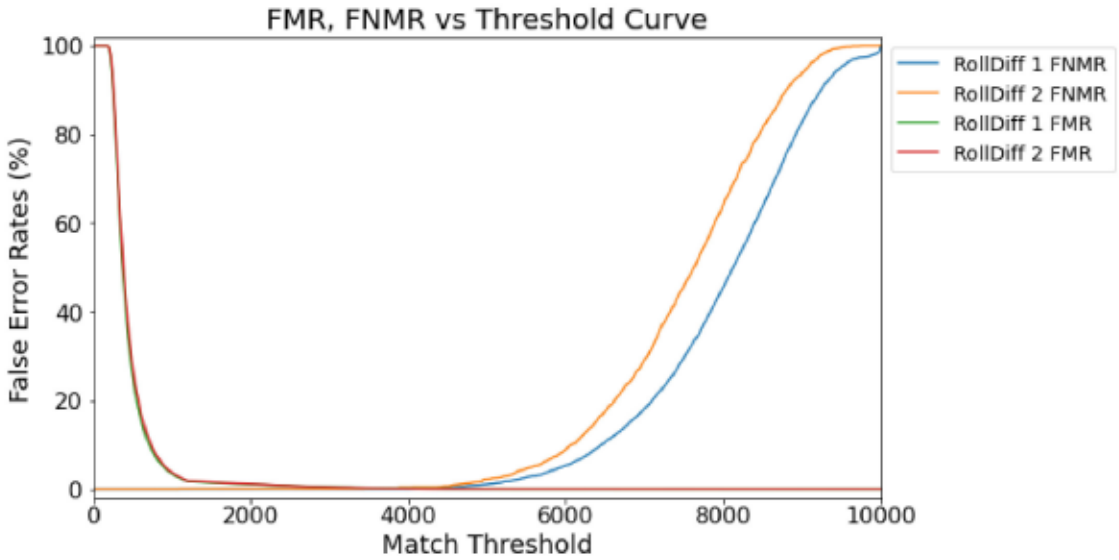


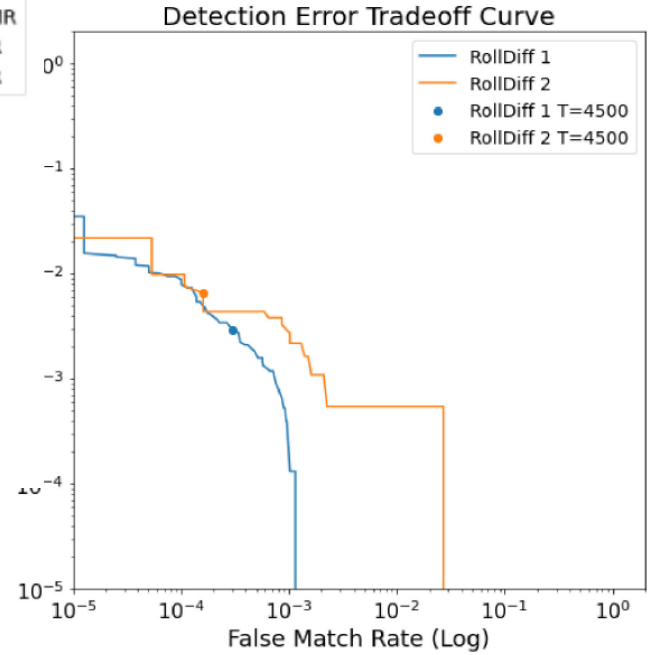


Roll: Q= Diff : 2 quality levels

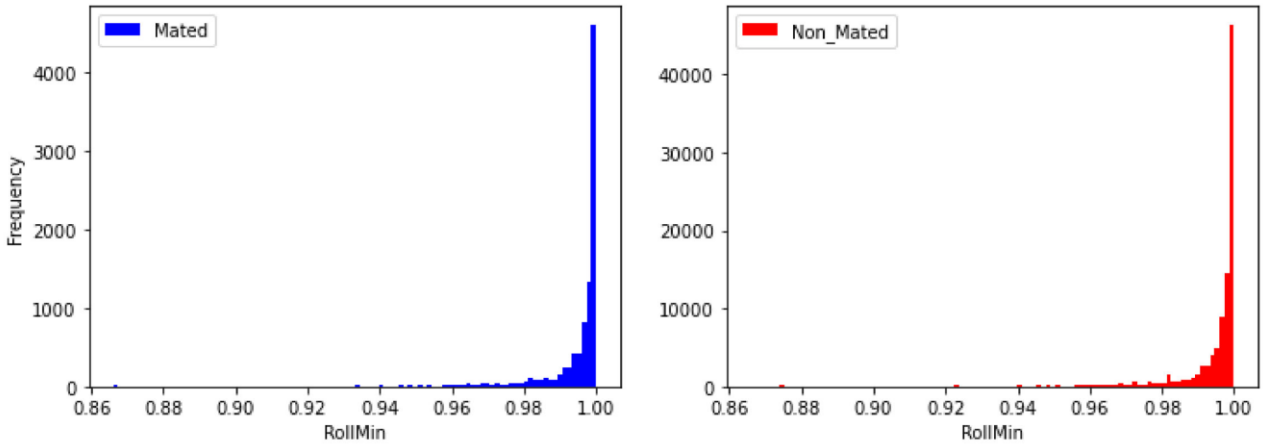


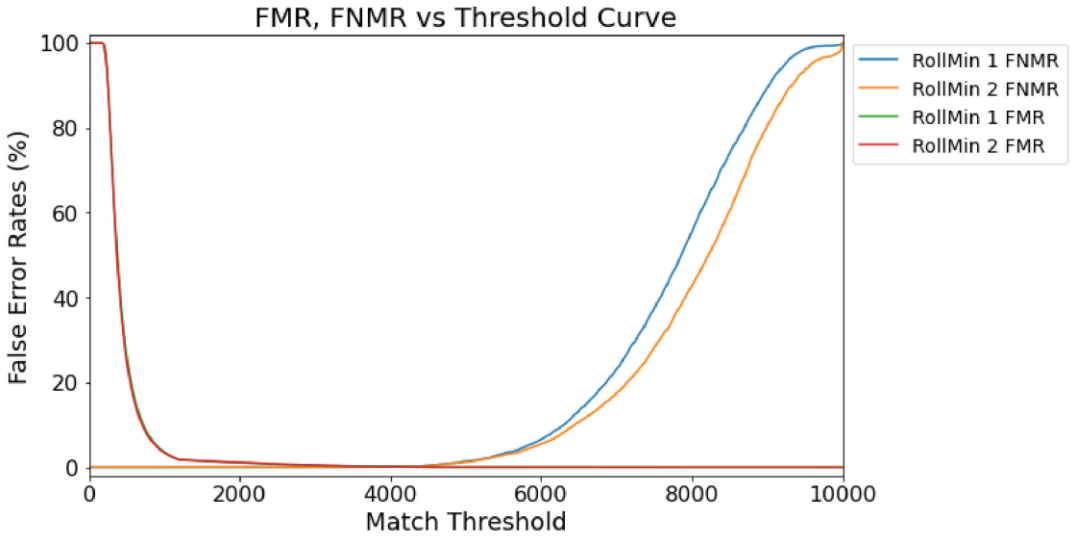
Cos θ cannot be substituted for θ because $|Cos(\theta_{probe}) - Cos(\theta_{ref})|$ & $|\theta_{probe} - \theta_{ref}|$ are not montonically related

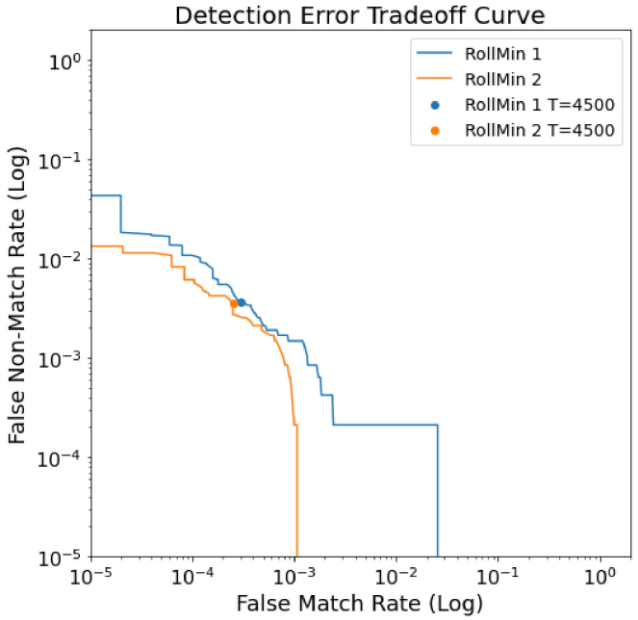




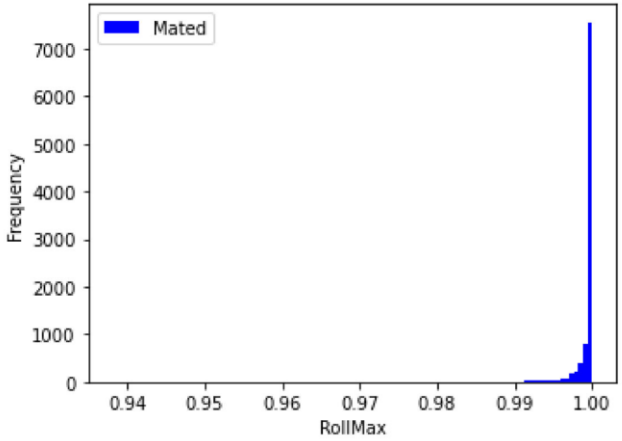
Roll: Q=Min|Θ| or Max cos(Θ): 2 levels

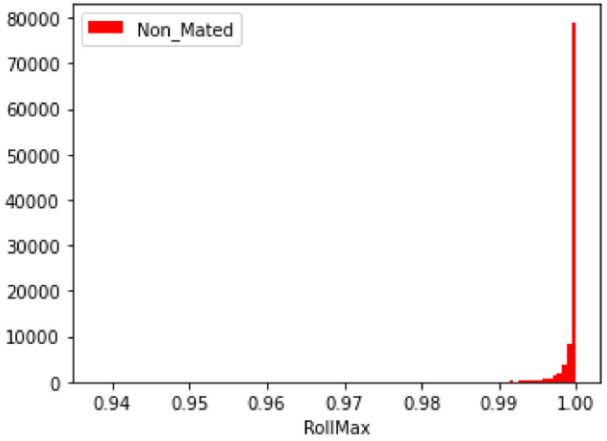


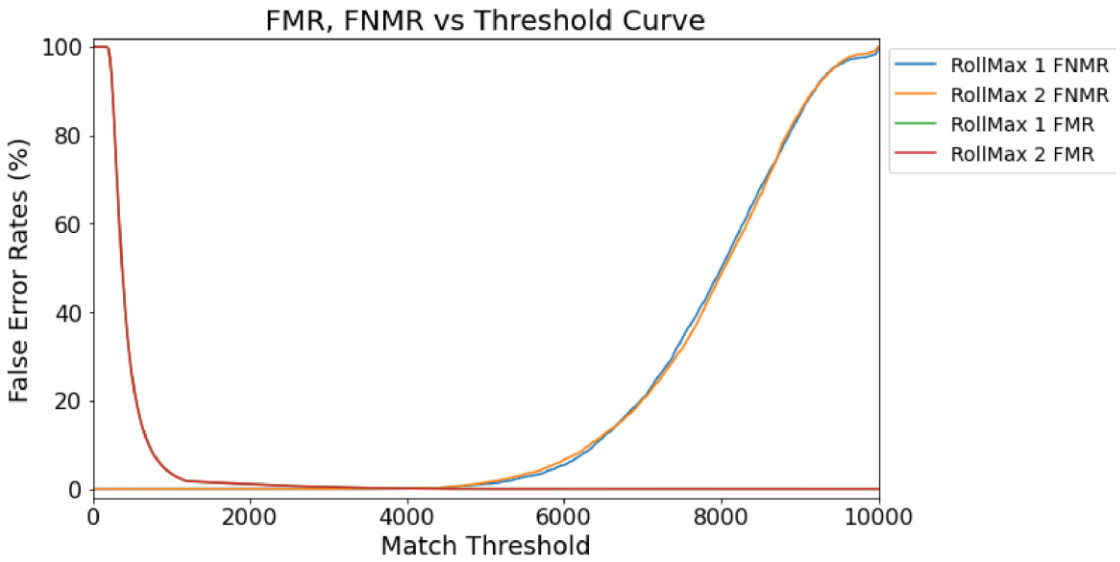


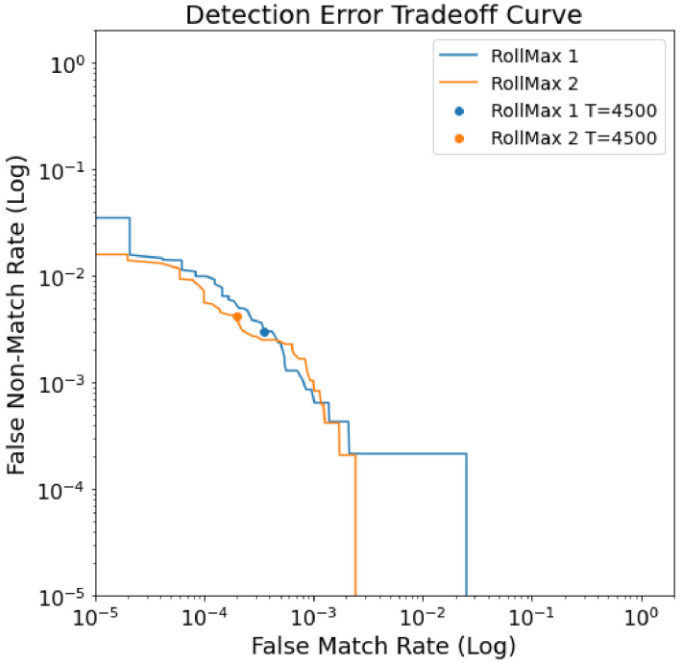


Roll: Q=Max|Θ| or Min cos(Θ): 2 levels









Summarizing the Process

1. Define “matcher/comparator performance”
2. Obtain “quality” metrics for both probe and reference image
3. Find appropriate Q (min, max, mean, harmonic mean, difference, probe only,…)
4. Chose number of levels and percentiles to differentiate performance by level
5. If performance cannot be differentiated by levels for any Q, metric is not a “quality” metric for that definition of “matcher/comparator performance”.

Conclusions

1. Stakeholder requirements differ. "Quality" should include assessment of both mated and non-mated distributions.
2. Quality comes in pairs, but choice of a single value Q (min, max, mean, harmonic mean, difference) depends upon the quality metric and the definition of comparator performance.
3. Basing Q on probe alone is not optimal
4. We speculate that reified metrics are more actionable and have fewer hidden demographic biases than non-reified.
5. OBIM does not discard images.
6. OBIM will use quality metrics in score fusion.

FUTURE WORK

- Validate current quality metrics against ground-truth images
- Add results to the recently published "1:M:N" facial recognition fusion model