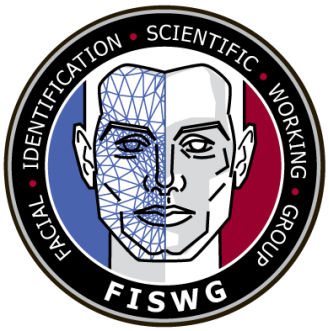


Disclaimer:

As a condition to the use of this document and the information contained herein, the Facial Identification Scientific Working Group (FISWG) requests notification by e-mail before or contemporaneously to the introduction of this document, or any portion thereof, as a marked exhibit offered for or moved into evidence in any judicial, administrative, legislative, or adjudicatory hearing or other proceeding (including discovery proceedings) in the United States or any foreign country. Such notification shall include: 1) the formal name of the proceeding, including docket number or similar identifier; 2) the name and location of the body conducting the hearing or proceeding; and 3) the name, mailing address (if available) and contact information of the party offering or moving the document into evidence. Subsequent to the use of this document in a formal proceeding, it is requested that FISWG be notified as to its use and the outcome of the proceeding. Notifications should be sent to: chair@fiswg.org

Redistribution Policy:

FISWG grants permission for redistribution and use of all publicly posted documents created by FISWG, provided that the following conditions are met:

Redistributions of documents, or parts of documents, must retain the FISWG cover page containing the disclaimer.

Neither the name of FISWG, nor the names of its contributors, may be used to endorse or promote products derived from its documents.

Any reference or quote from a FISWG document must include the version number (or creation date) of the document and mention if the document is in a draft status.

Face Recognition Systems Operation Assurance: Deployment Testing

1. Scope

1.1 This is the next document in the FISWG “Operational Assurance” document series. The reader is encouraged to review these public documents because they detail a sequential process of testing a facial recognition algorithm for an operational deployment and provides an explanation of the various methodologies and performance curves used.

1.2 This document uses the processes defined in the Operational Assurance series for sequential and iterative testing but expands the testing to a wider range of image cohorts, all of which can be present in operational deployments that need to be verified to ensure balanced performance. This testing follows NIST practices but uses a vendor specific facial algorithm and a suite of applications built for this specific testing. This type of testing can be conducted with other facial algorithms as needed.

1.3 The intended audience of this document is agencies that need to execute pre and post deployment verification testing of an FRS. This document serves as a reference document that can be given to an integrator, vendor, or contractor that shows how facial algorithm testing can be performed resulting in output metrics that meet agency performance requirements and legal mandates. It is also possible that the agency could have internal resources to perform the testing.

1.4 What is unique about this document from other FISWG Operational Assurance documents is that the intended audience, how this document should be used, and who is capable of executing it are targeted to very specific agency use cases and to established and experienced integrators, vendors, or contractors who can properly execute it with their existing knowledge base. This is not a learning document as the earlier FISWG Operational Assurance documents have been. This document is targeted to personnel who have proven experience producing results in this technology arena.

1.5 This document does not address applying test results to unique Mission specific operational workflows.

2. Referenced Documents

2.1 NIST

NISTIR 8280 Face Recognition Vendor Test (FRVT) Part 3: Demographic Effects[1]

NISTIR 8271 DRAFT SUPPLEMENT - Face Recognition Vendor Test (FRVT) - Part 2: Identification[2]

2.2 FISWG

Understanding and Testing for Face Recognition Systems Operation Assurance

Facial Recognition Systems Operation Assurance: Part 2, Identity Ground Truth

Facial Recognition Systems Operation Assurance: Part 3, Image Quality Assessment

[1] https://nvlpubs.nist.gov/nistpubs/ir/2019/nist.ir.8280.pdf
[2] https://nvlpubs.nist.gov/nistpubs/ir/2019/nist.ir.8271.pdf

Facial Recognition Systems Operation Assurance: Part 4, Manual Facial Localization

Facial Recognition Systems Operation Assurance: Part 5, Scoring Thresholds

3. Terminology

3.1 Acronyms

3.1.1 *CMC, n*—Cumulative Match Characteristic

3.1.2 *DET, n*—Detection error tradeoff

3.1.3 *FAR, n*—False acceptance rate

3.1.4 *FRR, n*—False reject rate

3.1.5 *FRS, n*—Facial recognition system

3.1.6 *IOD, n*—Interocular Distance (pixels)

3.1.7 *ISO, n*—International Organization for Standardization

3.1.8 *OCD, n*—Ocular chin distance (pixels)

4. Summary of Guide

4.1 It is now more important than ever to do proper accuracy testing before and after an FRS is deployed to ensure you are meeting your expected performance. In addition to having good accuracy, it is also important that the FRS performs well in

terms of minimizing differential algorithmic variations, for example any performance differentials between different demographic groups that would disadvantage one demographic group in comparison to another. NIST has been documenting this need for many years as evidenced in FRVT reports:

> Operational implementations usually employ a single face recognition algorithm. Given algorithm specific variation, it is incumbent upon the system owner to know their algorithm. While publicly available test data from NIST and elsewhere can inform owners, it will usually be informative to specifically measure accuracy of the operational algorithm on the operational image data, perhaps employing a biometrics testing laboratory to assist.

Figure 1: From NISTIR 8280

4.2 Once an agency has determined that an FRS will be deployed or updated, the agency needs to gather Mission and legal requirements that the solution must address. Utilize the FISWG document “Principles for Responsible Use of Facial Recognition Technology” to assist in this process.

4.3 The agency needs to carefully consider and define what types of image cohorts need to be tested based on FRS Mission requirements and legal mandates present. These cohorts must reflect operational data for agency specific use cases. These requirements need to be delivered to the integrator, vendor, or a contractor for acceptance. This document can then be used to define expectations of the agency

specific test results desired. The agency will have to supply agency specific facial data to test with.

4.4 The integrator, vendor, or a contractor shall:

4.4.1 Have developmental skills to integrate vendor specific facial algorithms into an application that can be used for testing

4.4.2 Be fluent with current and legacy NIST and ISO test reports that address facial biometric testing and performance

4.4.3 Be fluent with all terms, definitions, and acronyms regarding facial biometric deployments

4.4.4 Have the ability to produce, understand, and explain the basic biometric accuracy charts for 1:N deployments: FAR, FRR, DET, ROC, CMC

4.4.5 Have the ability to use the test results to assist the agency in an FRS deployment and provide operational support

4.5 Desired outcomes from utilizing this document for FRS testing will result in a large collection of image analysis and accuracy results across a wide range of areas that include:

4.5.1 Facial image file properties

4.5.1.1 File sizes, file format, file compression, date of capture, pose

4.5.2 Facial algorithm metrics

4.5.2.1 Image quality, pose estimations, size of the face, soft biometrics

4.5.3 Identity based metrics

4.5.3.1 Recidivism, sex, race, age

4.5.4 Identity ground truth verification

4.5.5 Accuracy profiles with specific cohorts

4.5.5.1 Sex, race, age, pose, image quality

4.5.6 Summary of performance across all cohorts

4.5.7 Areas where the performance is reduced and steps that can be done to address these gaps

4.5.8 The image analysis described in Sections 4.5.1 through 4.5.4 are the first phases in this testing process because their outputs define what is known about the data before actual testing can be started. Section 7 in this document presents these details.

4.5.9 The biometric performance steps described in Sections 4.5.5 through 4.5.7 are the second phases in this testing process because their outputs define the accuracy profiles that are the overall goals of this testing. Sections 8 - 11 in this document present these results.

5. Significance and Use

5.1 Given the importance of regularly testing the performance of facial recognition technology, the remainder of this document will present a facial algorithm testing process that can be referenced or followed by an agency. The process was defined and followed for specific test scenarios for the test data set used. Agencies can modify these processes as needed.

5.2 This section summarizes the results of specific testing done with this image set and the specific differential testing desired:

5.2.1 Section 9 covers Demographic differentials: sex, race, age

5.2.2 Section 10 covers Facial pose variations: mixed poses, frontal, profile

5.2.3 Section 10 also covers Image quality variations: small frontal faces (IOD), small profile faces (OCD), low vendor quality, manually localized faces

5.2.4 Section 11 covers reducing IOD sizes: 50, 40, 30, 20, 10 pixels

5.3 A summary of the results include the following findings.

5.3.1 All image cohorts have similar accuracy results except for those that were considered low quality imagery, small IOD/OCD, or reducing IOD to low ranges (10-30 IOD). The low-quality cohort testing showed that some adjustments to search parameters are needed to maintain an equivalent identification rate as shown in the CMC charts.

5.4 Other potential results could include the following findings.

5.4.1 If specific accuracy variations occur based on the image quality of specific cohorts, the agency could explore improvements to the algorithm, the algorithm vendor, different capture methods, image file format type, or image size.

5.4.2 If specific accuracy variations occur based on facial pose, the agency could explore improvements at the point of capture.

5.5 Any enhancements to agency specific standard operating procedures that could improve facial accuracy of improve forensic examination procedures.

6. Procedure

6.1 An overview of key aspects in this testing process includes these areas

6.1.1 Proper authorizations and personnel:

6.1.1.1 Authorization to use data for analysis

6.1.1.2 Access to a facial algorithm be used for image quality and accuracy metrics

6.1.1.3 Computational infrastructure

6.1.1.4 Access to appropriate software (Excel, MATLAB, Power BI)

6.1.1.5 Personnel resources (developers, data analysts, scientists)

6.2 Workflow summary

6.2.1 An operationally relevant image sample size must be determined and will vary depending on agency requirements and use cases. Agencies should perform an analysis selecting a sample size and sample content to address Mission requirements. This can be done in several ways:

6.2.1.1 Extract a fixed percentage of the deployed gallery for testing. If the deployed gallery has 10 million enrollments, extract one million faces for testing

6.2.1.2 Extract known cohorts known to be in the deployed gallery. This technique would need to cover a range of encounters and would involve careful selection of sex, race, age, pose, and image quality to produce a gallery for testing that has operational relevance and will assure various demographic differential testing can be accomplished.

6.2.1.3 A combination of these two where gallery size and operational relevance are properly addressed.

6.2.2 The dataset to be tested is gathered and presented to the facial algorithm extracting key image metrics

6.2.3 The desired testing scenarios based on agency focus areas are performed

6.2.4 Analysis and visualization of the test results

6.2.5 Determination of operational impacts

6.3 Basic Sequential Workflow

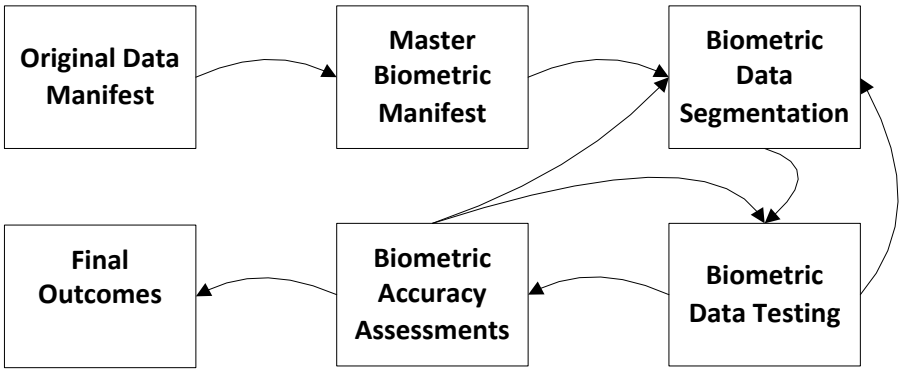


Figure 2: Key Workflow Steps

6.3.1 This workflow is a step-by-step process focusing on a linear sequence of data awareness, data preparation, data segmentation, testing, analysis, and outcomes.

6.3.1.1 Original Data Manifest: Extract basic image file properties and correlate any metadata supplied with the images

6.3.1.2 Master Biometric Manifest: Extract and integrate biometric metrics from the facial algorithm with the Original Data Manifest

6.3.1.3 Biometric Data Segmentation: Segment the facial data into separate cohorts based on the testing to be done

6.3.1.4 Biometric Data Testing: Perform testing on each cohort

6.3.1.5 Biometric Accuracy Assessments: Create accuracy assessment charts on each cohort

6.3.1.6 Final Outcomes: Review the results and determine outcomes on each cohort

6.3.2 Each step has a finite list of inputs, processes, and outputs. As the steps are sequenced and results reviewed, a return to a previous step can be done to address refinements, gaps, or anomalies observed.

6.3.3 Key steps are focused on data awareness while other steps must adapt to support various facial algorithms

6.3.4 Dependencies between steps are minimized

6.3.5 Output artifacts from key steps are reusable for future test scenarios

6.4 Processing steps omitted

6.4.1 Verification testing (1:1) was not performed since this specific test was focused on 1:N performance.

6.4.2 Image encoding and search speeds were not measured since it was done on a workstation that did not have recommended capabilities for a full solution deployment.

6.4.3 The computational resources needed were not measured since the facial algorithm accuracy and data evaluations were the primary focus.

6.5 Key results.

6.5.1 This workflow will produce a wide range of information about the facial imagery used, how the facial algorithm analyzed the data, and if the 1:N accuracy varied:

6.5.1.1 Demographic variations

6.5.1.2 Facial pose variations

6.5.1.3 Facial image quality variations

6.5.1.4 Facial localization variations

6.5.1.5 Facial size:

- IOD: distance between eye centers (frontal poses)
- OCD: distance from the chin to the eye center line (frontal and profile poses)

6.5.1.6 Facial pose:

- Yaw: left to right rotation

6.5.1.7 Soft biometrics available in this specific vendor facial algorithm include:

- Sex
- Race
- Age

7. Original and Biometric Data Manifest Outputs

7.1 A data set of 107,207 facial images was used for this test. The imagery was unclassified mugshots captured with two different camera systems and included information on identity, sex, race, date of birth, date of capture, and facial pose (frontal or profile).

7.2 Basic Image File Properties

Image Size (Height and Width in Pixels)	Count	Percentage
384x480	99,793	93%
960x1280	7,406	7%

Figure 3: Image Size

Image Resolution	Count	Percentage
72	2,321	2%
96	7,406	7%
150	22,718	21%
300	74,758	70%

Figure 4: Image Resolution

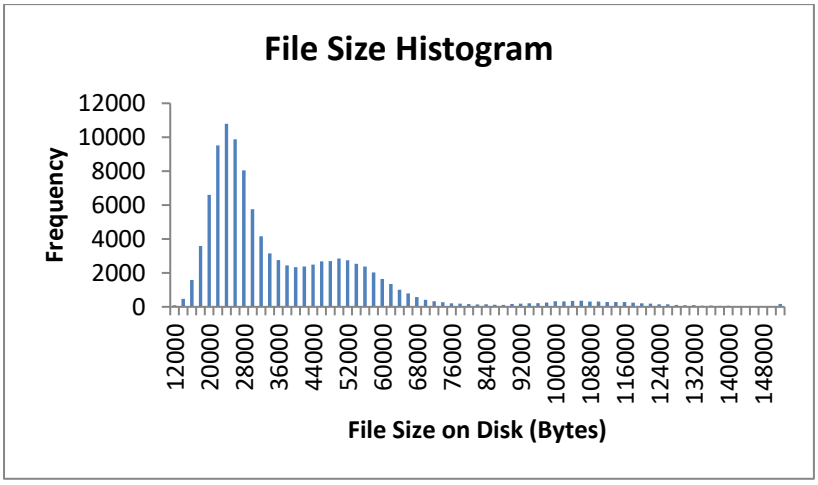


Figure 5: File Size (Bytes on Disk)

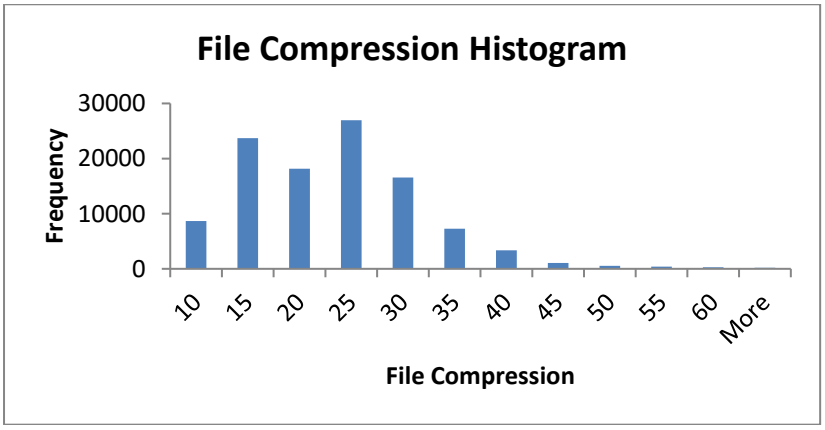


Figure 6: File Compression

7.3 Defined Sex and Pose

Image	Count	Percentage

Male	92,073	85%
Female	15,134	14%
Frontal Pose	54,247	50%
Profile Pose	52,960	49%

Figure 7: Sex/Pose Content

7.4 Defined Identity Recidivism

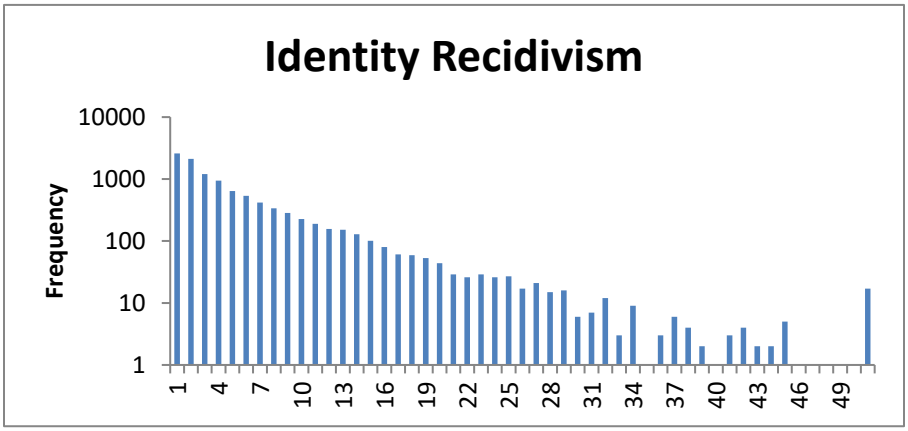


Figure 8: Identity Recidivism

7.5 Defined Race

7.5.1 This specific data set had 19 race codes provided with the imagery. Figure 8 shows the races selected for testing.

Race	Count	Percentage
Black	31173	29%
Hispanic	29361	27%
White	30489	28%

Figure 9: Race Content

7.6 Facial Metrics

7.6.1 Facial algorithm derived metrics were extracted and then combined with the basic image file properties. This step involved software development to allow processing the images through the facial algorithm.

7.6.2 IOC (interocular distance) and OCD (ocular chin distance)

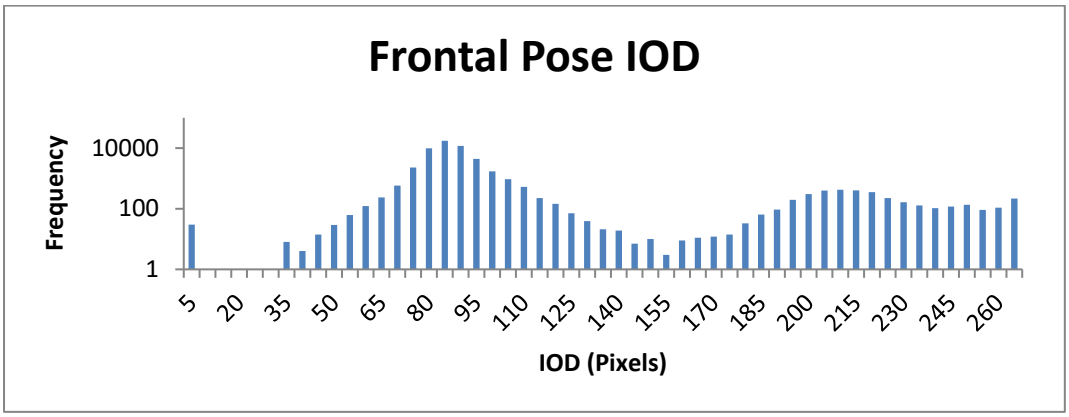


Figure 10: Frontal Pose IOD (pixels)

7.6.2.1 Figure 9 shows the frontal pose IOD having two clusters of ~90 and ~215 pixels. This is caused by the two different image capture solutions used producing image of different sizes.

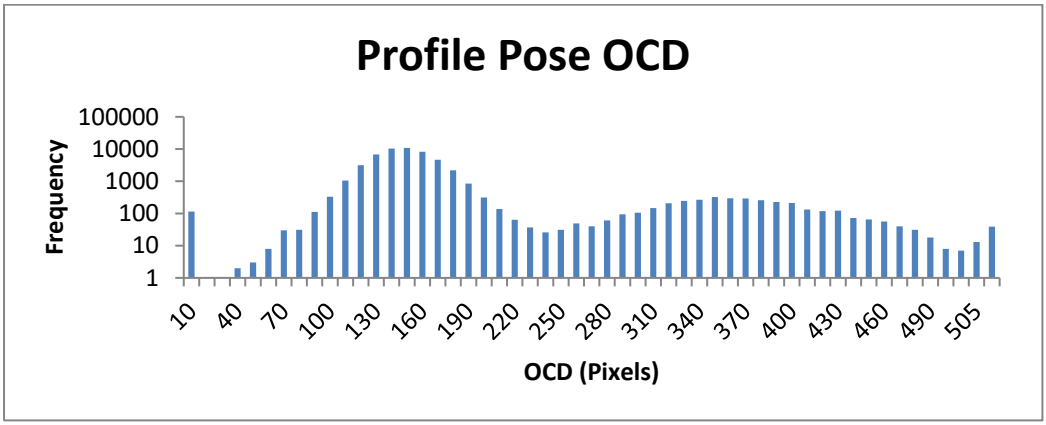


Figure 11: Profile Pose OCD (pixels)

7.6.2.2 Figure 10 shows the profile poses OCD has two clusters of ~140 and ~370 pixels. This is caused by the two different image capture solutions used producing image of different sizes.

7.6.3 Image Quality

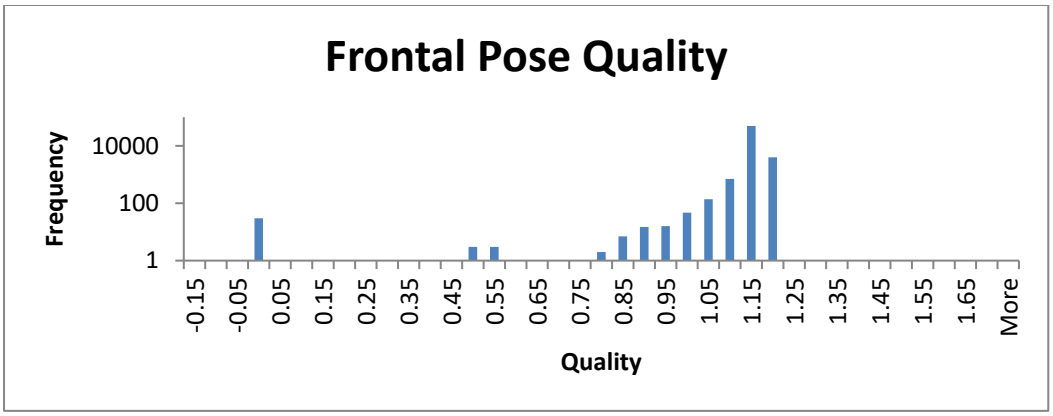


Figure 12: Frontal Pose Quality

7.6.3.1 Figure 11 shows that the frontal pose quality. The implied value in this metric is vendor specific.

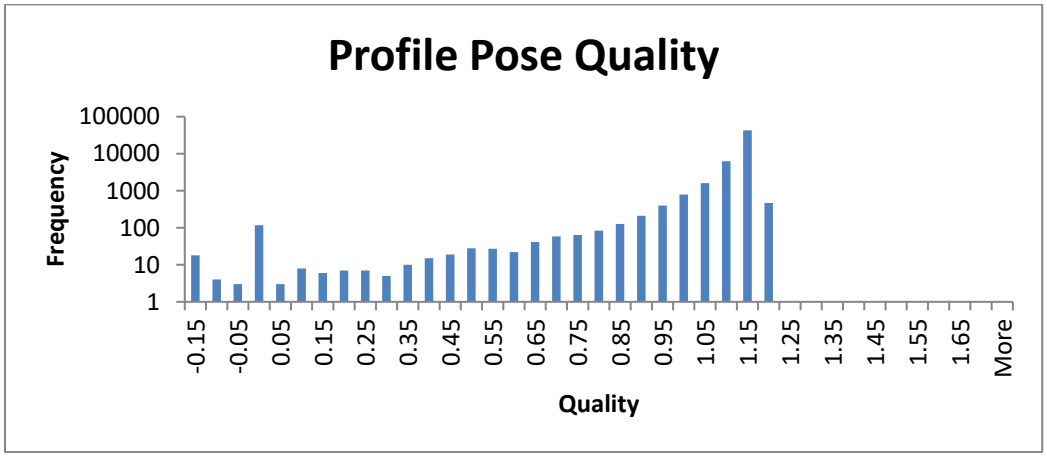


Figure 13: Profile Pose Quality

7.6.3.2 Figure 12 shows the profile pose quality has larger variations than the frontal poses. The implied value in this metric is vendor specific. Assumptions on the cause of this difference includes:

- The yaw angle in the profile pose varies more than the frontal pose most likely due to inconsistencies in the pose of the person at the time of image capture.

- It is also likely that more obstructions are present in profile poses (hair) that could impact the yaw pose metric

7.6.4 Defined Age

7.6.4.1 The image data was delivered with a date of capture and a date of birth. From these two items the age of the person in the image was derived and presented in Figures 16 and 17.

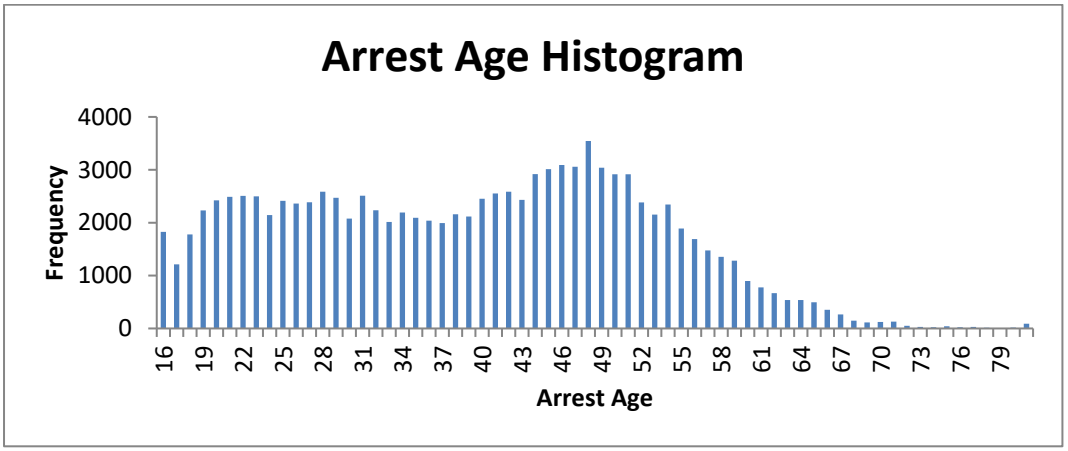


Figure 14: Arrest Age provided with the Images

Age Range	Count	Percentage
Between 16 and 20	6,510	6%
Between 20 and 30	24,276	22.6%
Between 30 and 40	21,323	19.9%
Between 40 and 50	28,687	26.7%
Between 50 and 60	20,401	19%
Greater than 60	5,370	5%

Figure 15: Arrest Age provided with the Images

7.6.4.2 The facial algorithm returned an estimated age from the person in the image. The comparison of the defined arrest age and the age returned by the facial algorithm is shown in Figure 15. The standard deviation in the age difference was 5.64 years.

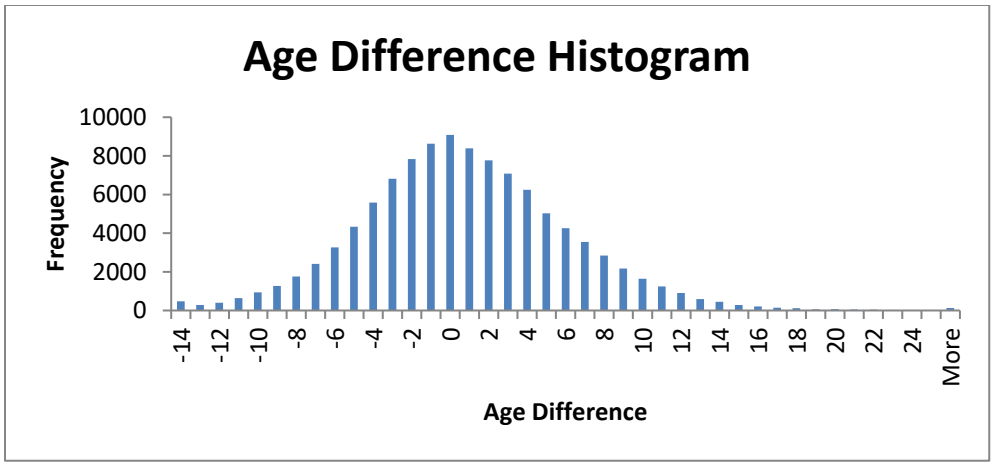


Figure 16: Arrest Age difference as returned from the facial algorithm

7.6.4.3 The facial algorithm returned a sex and race estimation from the person in the image. The comparison of the defined sex and race and the sex and race estimated by the facial algorithm is shown in Figures 16.

Metric	Difference Between Data Supplied and Algorithm Estimation
Male	0.75%
Female	6.63%
White Race	18.66%
Black Race	4.29%
Hispanic Race	62.89%

Figure 17: Sex and Race Variations

7.6.5 The facial algorithm calculated the facial pose yaw and is shown in Figure 17. This shows that the mixture of poses varied from yaws of -80 to + 90 degrees.

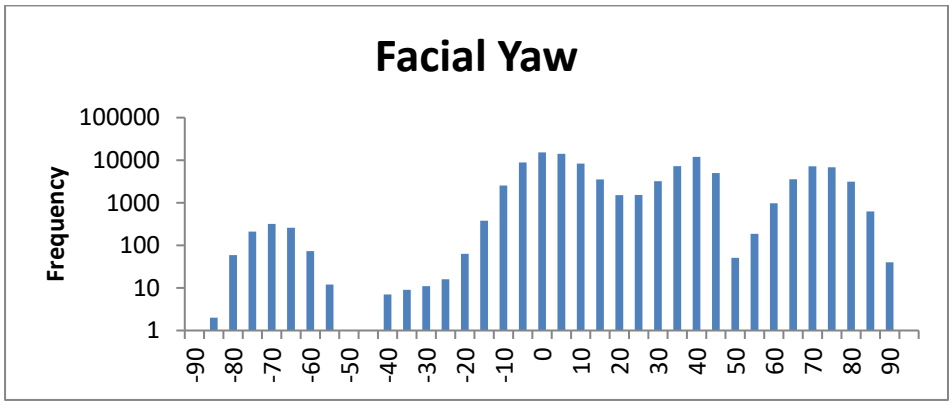


Figure 18: Facial Yaw Histogram (Degrees)

7.7 When creating templates, 84 images failed to template due to various reasons as shown in Figures 18-21:

Figure 19: Non faces

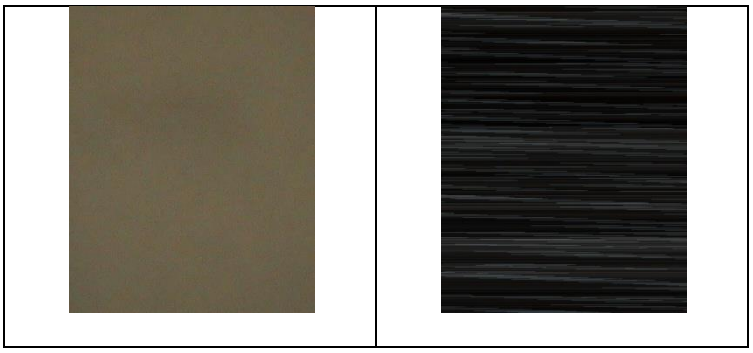


Figure 20: Blank Images

Figure 21: Extreme Obstructions

Figure 22: Extreme Yaw and Tilt

7.7.1 Manual facial localization was used to recover 19 images by manually assigning coordinates to both eyes and the chin and feeding these into the facial algorithm for template creation.

7.8 This completes the biometric data manifest step as defined in Figure 4 “Key Workflow Steps”.

8. Ground Truth Test

8.1 This starts the biometric data segmentation and testing steps as defined in Figure 4 “Key Workflow Steps”.

8.2 Ground truth verification

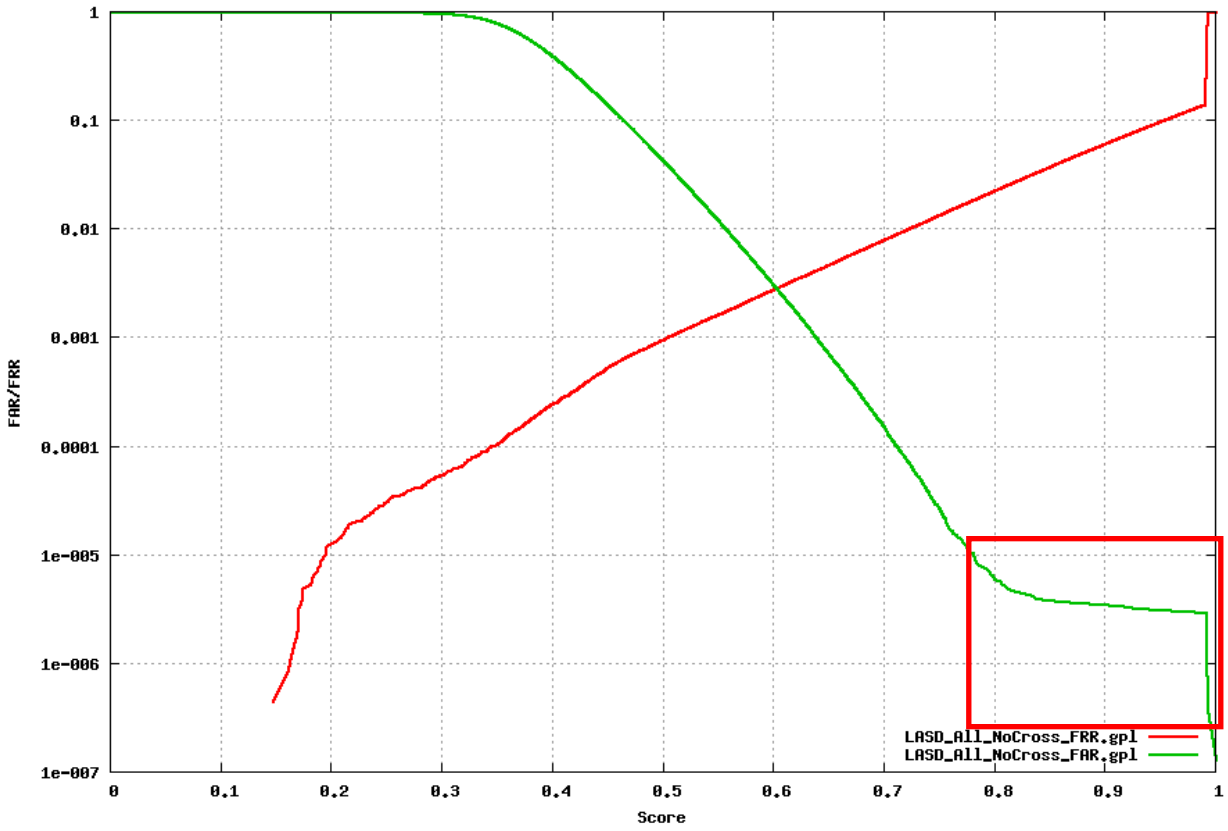
8.2.1 Ground truth must be present for all images in the data set. If identities are not known for images they should not be used.

8.2.2 Templates should be created from all images and loaded into a searchable gallery.

8.2.3 All the images should be searched against the gallery.

8.2.4 Output charts should be produced and analyzed for identity ground truth.

306



307

308 **Figure 23: All Imagery FAR/FRR before Ground Truth**

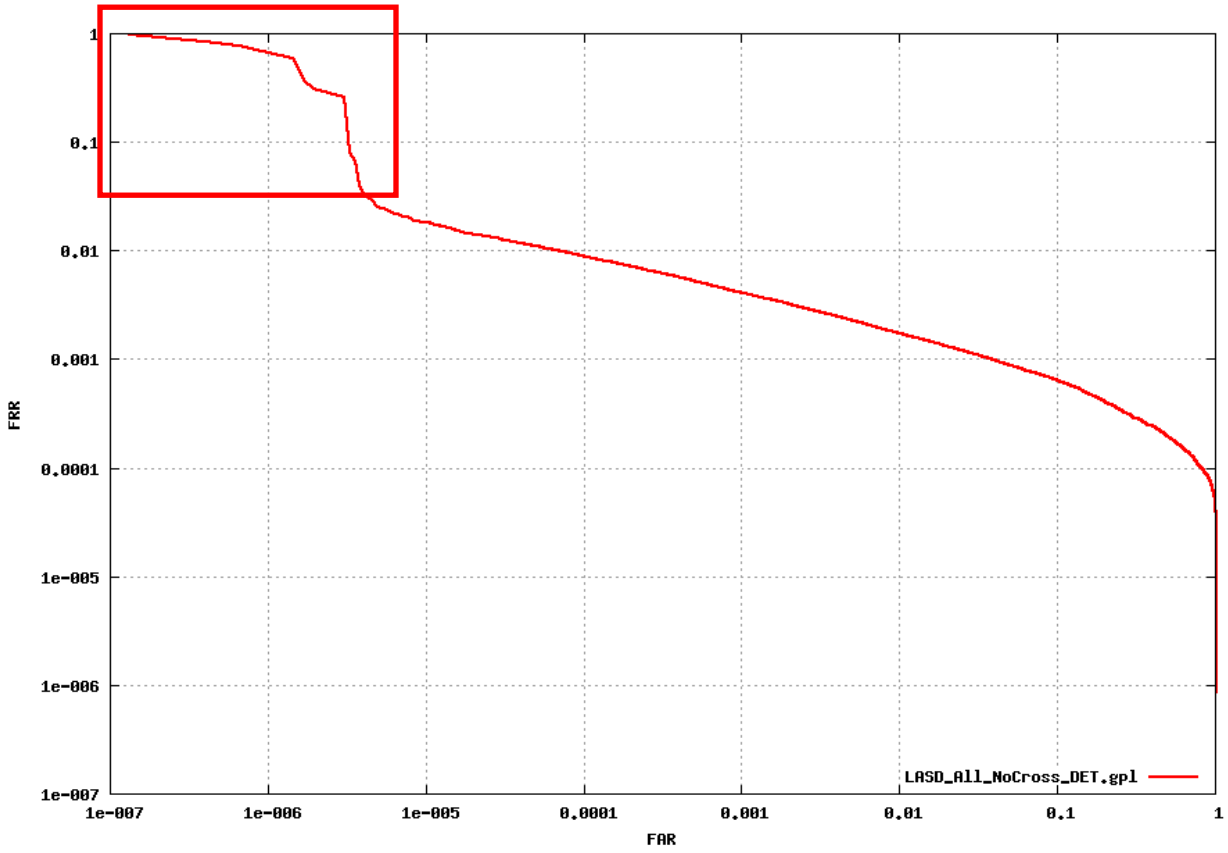


Figure 24: All Imagery DET before Ground Truth

8.2.5 Results

8.2.5.1 The results indicate potential identity ground truth errors:

- Figure 21 shows high FAR scores in the lower right red box. The FAR score increase from 0.8 to 1.0 suggests that imposters are scoring very high.
- Figure 24 shows a DET curve anomaly in the upper left red box. This is caused by the high FAR scores.

8.2.6 High score imposters were analyzed, and ground truth errors were located in 37 identities. These could be manually corrected if needed but for this test these identities were removed.

8.2.6.1 Verify ground truth Corrections

8.2.6.2 Templates should be created from all images and loaded into a searchable gallery.

8.2.6.3 All the images should be searched against the gallery.

8.2.7 Output charts should be produced and analyzed for identity ground truth verification.

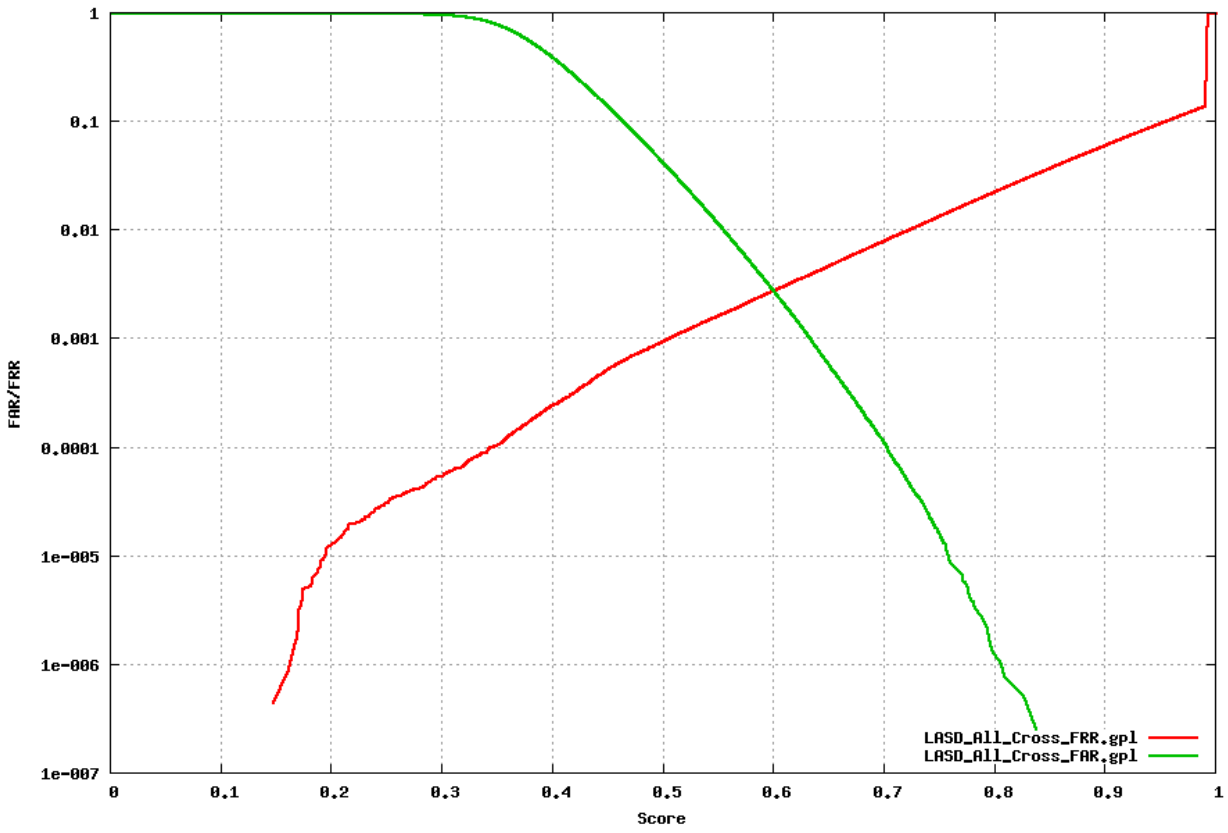


Figure 25: All Imagery FAR/FRR after Ground Truth

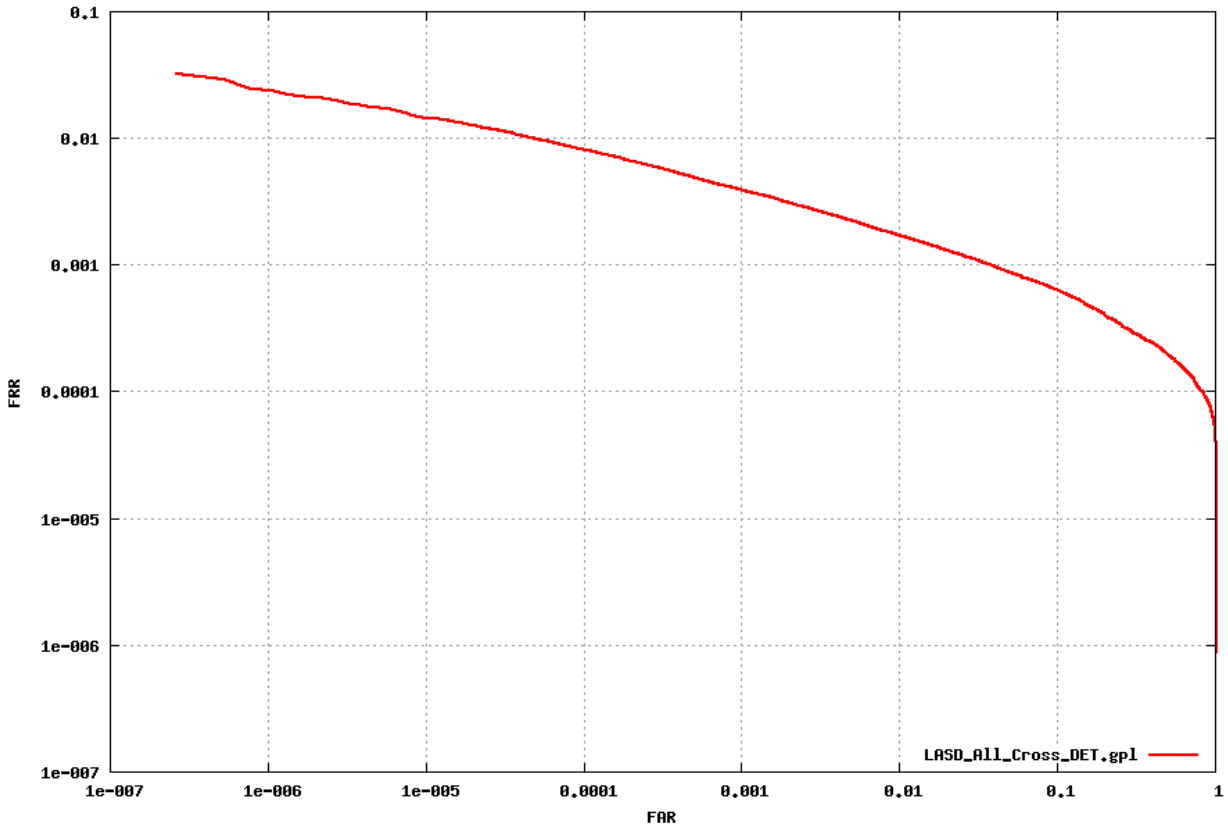


Figure 26: All Imagery DET after Ground Truth

8.2.8 Results

8.2.8.1 Figures 24 and 25 show the FAR, FRR and DET curves which now have the ground truth errors corrected.

9. Demographic Differential Test

9.1 This starts the biometric data segmentation, testing steps, and accuracy measurements as defined in Figure 4 “Key Workflow Steps”.

9.2 Disclaimer: FISWG does not endorse the use of demographic filter categories, but these may be included with the subject metadata.

9.3 All charts that follow compare the results from all imagery to a specific subset of imagery:

9.3.1 Sex: male and female

9.3.2 Race: black, Hispanic, white, and other mixed races not in these categories

9.3.3 Arrest age variations using 10-year ranges

9.4 Sex Differential Testing

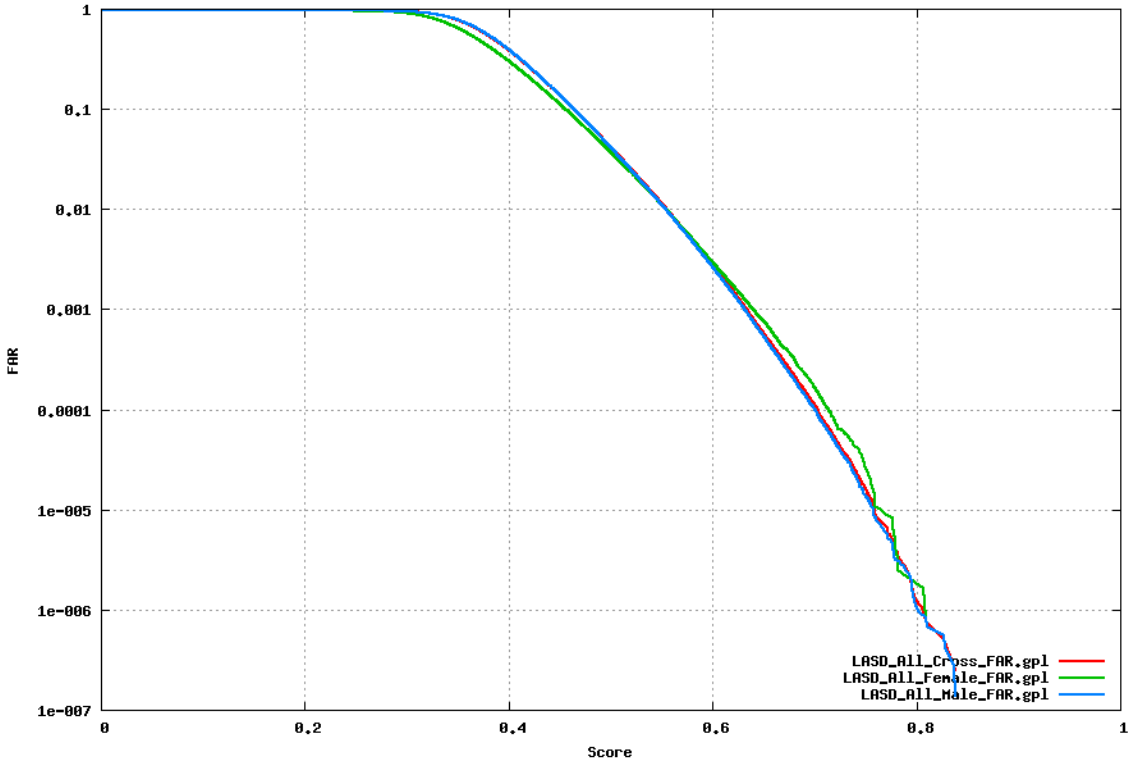


Figure 27: Sex Variations: FAR

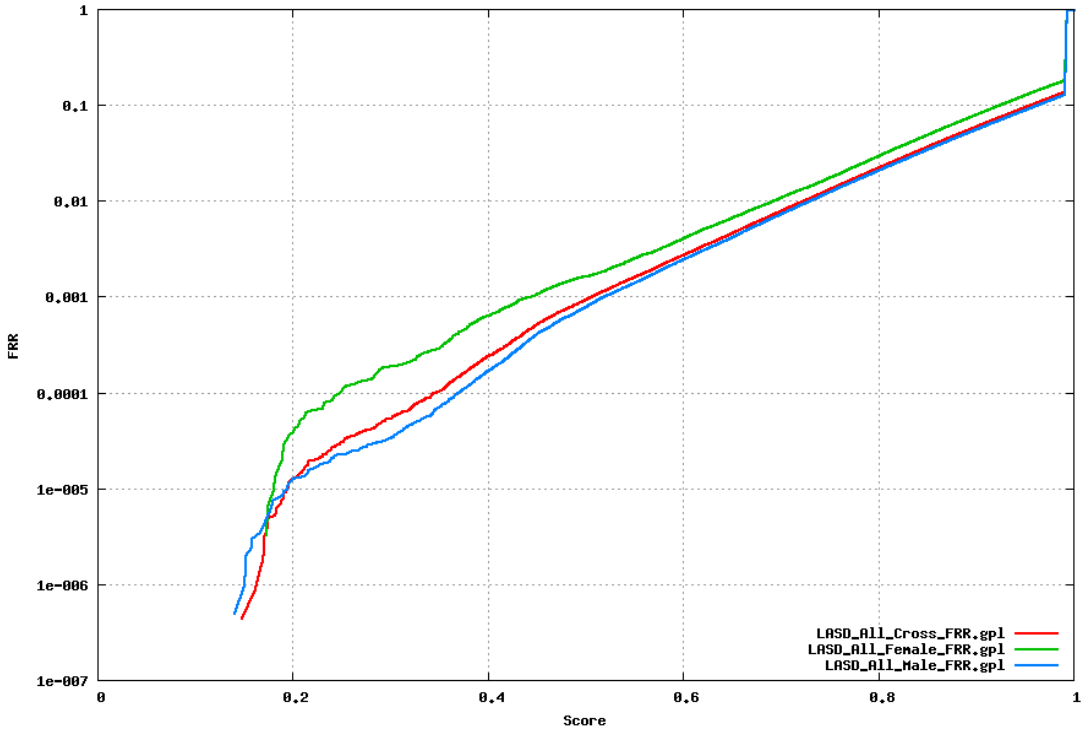


Figure 28: Sex Variations: FRR

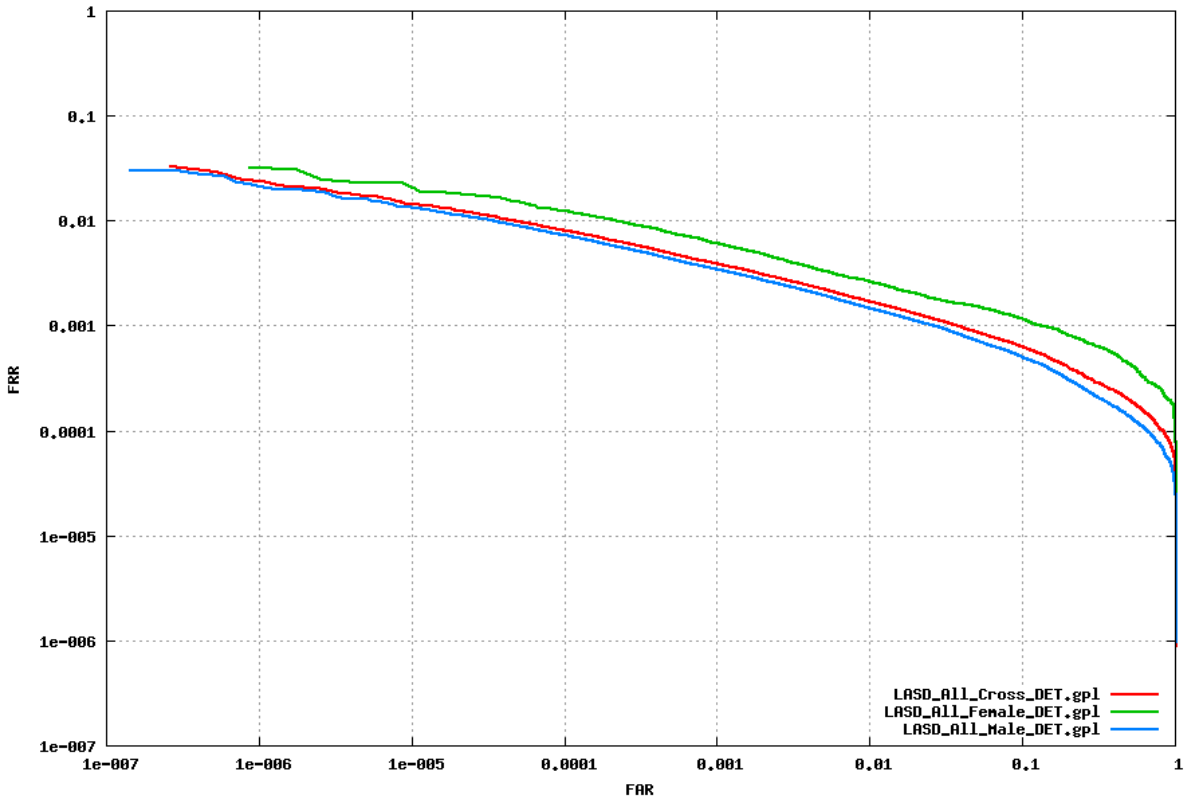


Figure 29: Sex Variations: DET

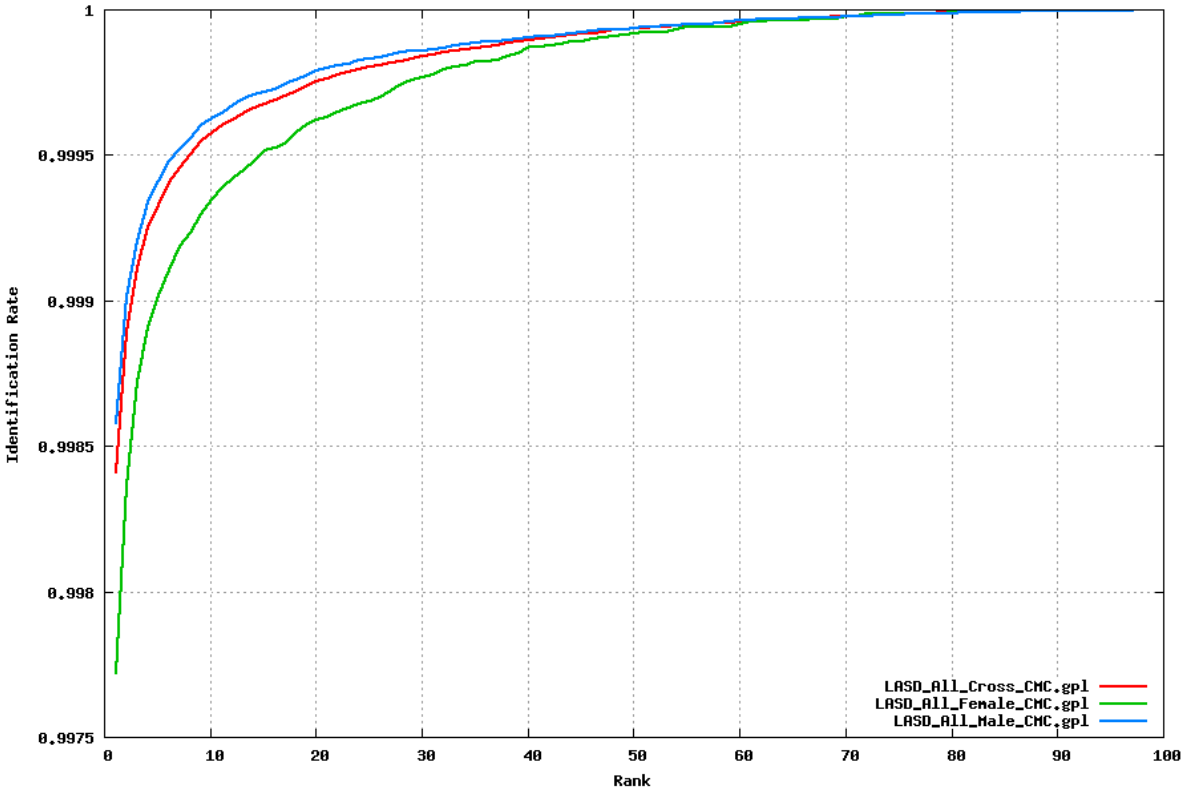


Figure 30: Sex Variations: CMC

9.4.1 Results

9.4.1.1 There are little variations when testing accuracy for male and female sex. Figures 26 and 27 show similar FAR and FRR scoring, while Figure 31 shows consistent DET performance.

9.4.1.2 Figure 29 shows a CMC rank one identification rate for both male and female above 99.75%.

9.5 Race Differential Testing

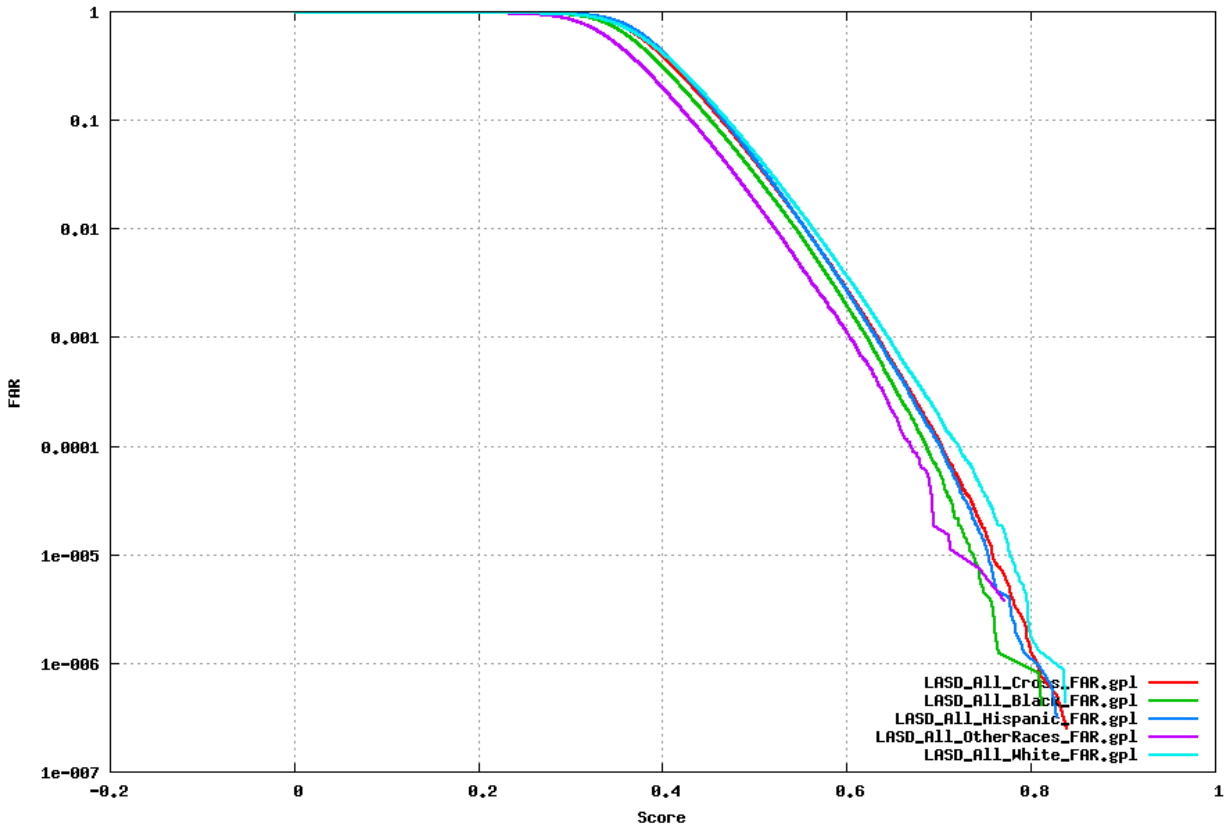
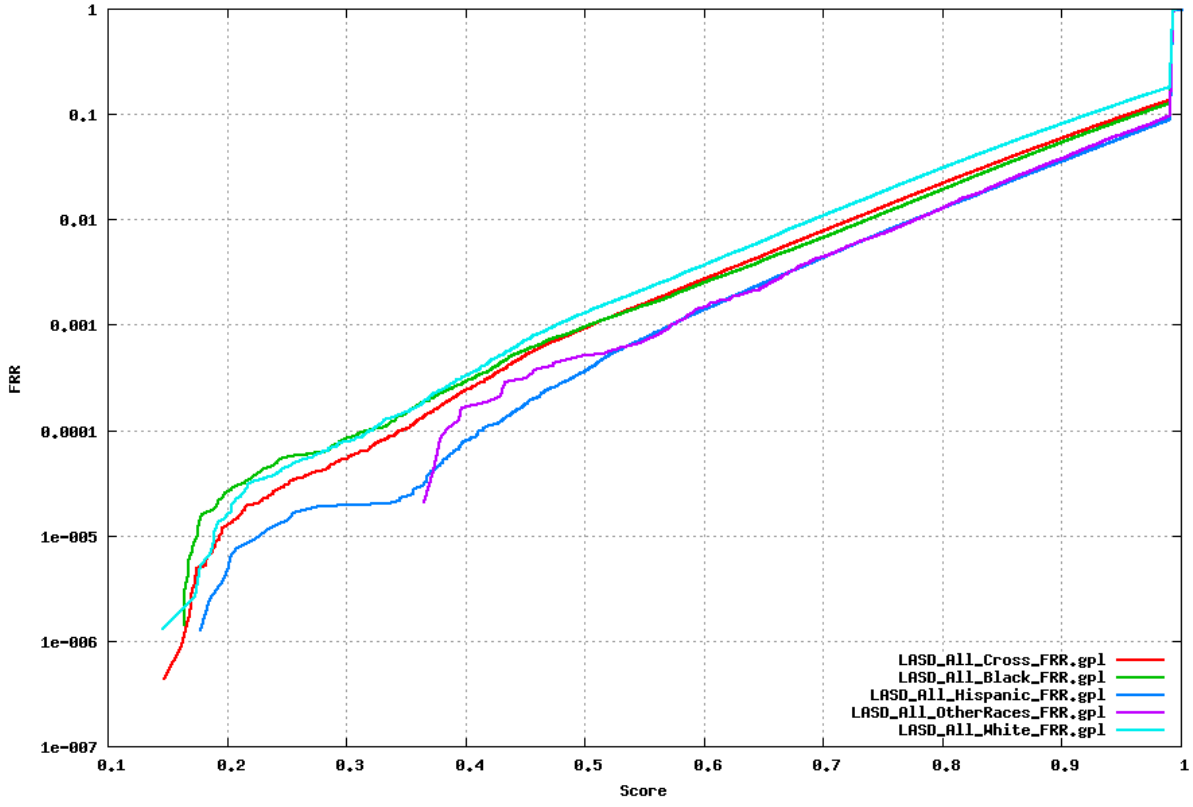
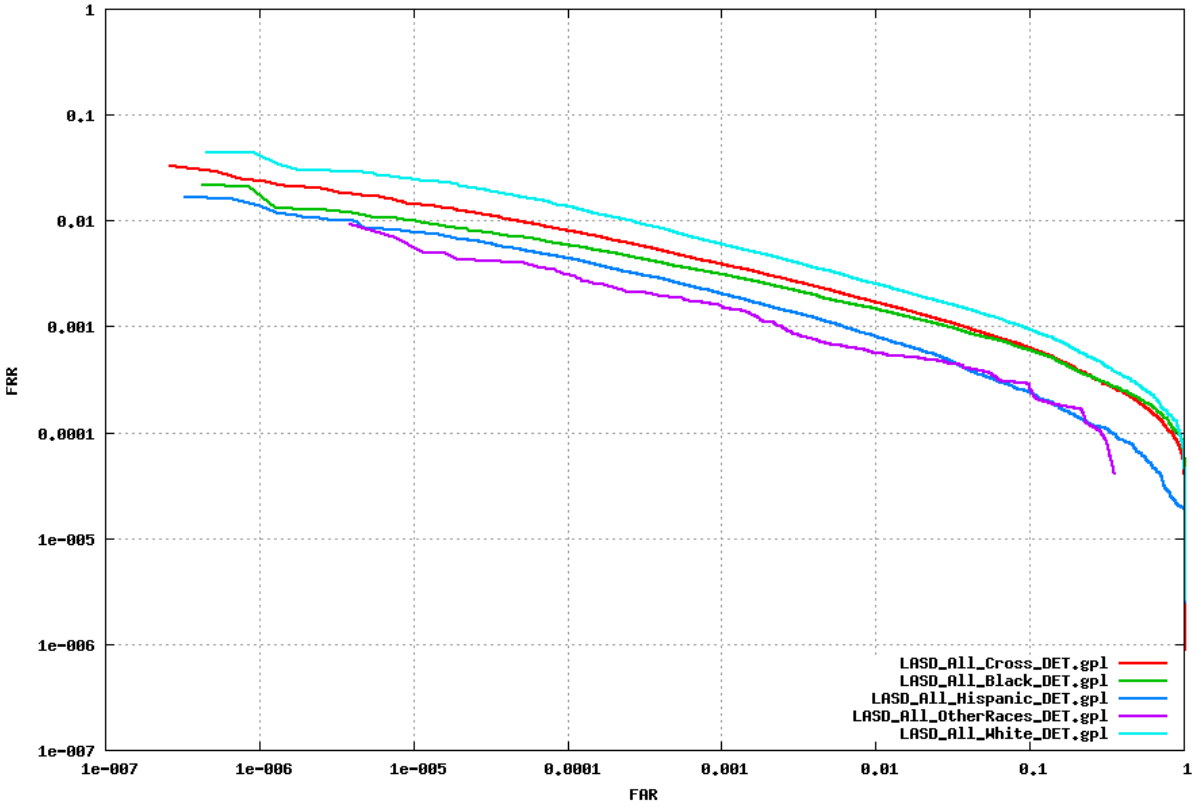


Figure 31: Race Variations: FAR



363 **Figure 32: Race Variations: FRR**



364
365 **Figure 33: Race Variations: DET**

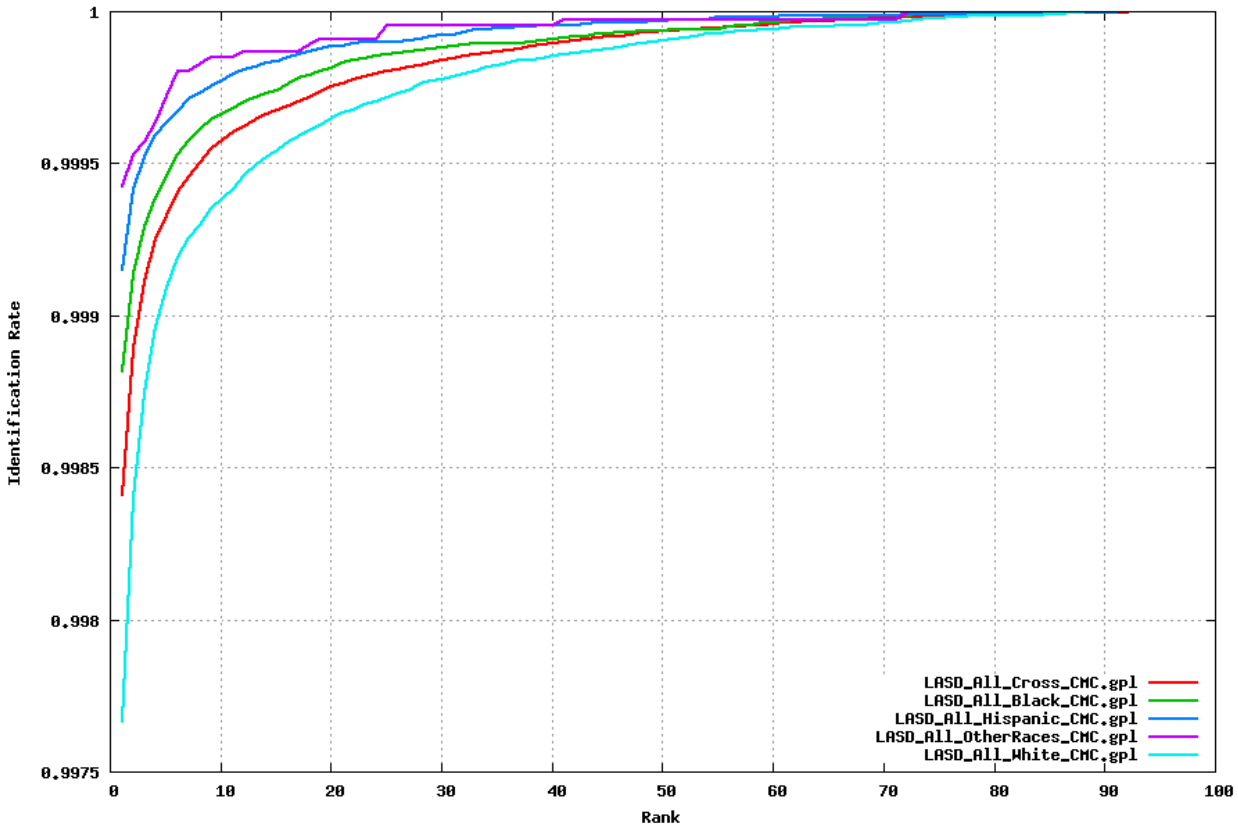


Figure 34: Race Variations: CMC

9.5.1 Results

9.5.1.1 There are little variations when testing accuracy for race. Figures 30 and 31 show similar FAR and FRR scoring, while Figure 32 shows consistent DET performance.

9.5.1.2 Figure 33 shows a CMC rank one identification rate for all races above 99.75%.

9.6 Arrest Age Differential Testing

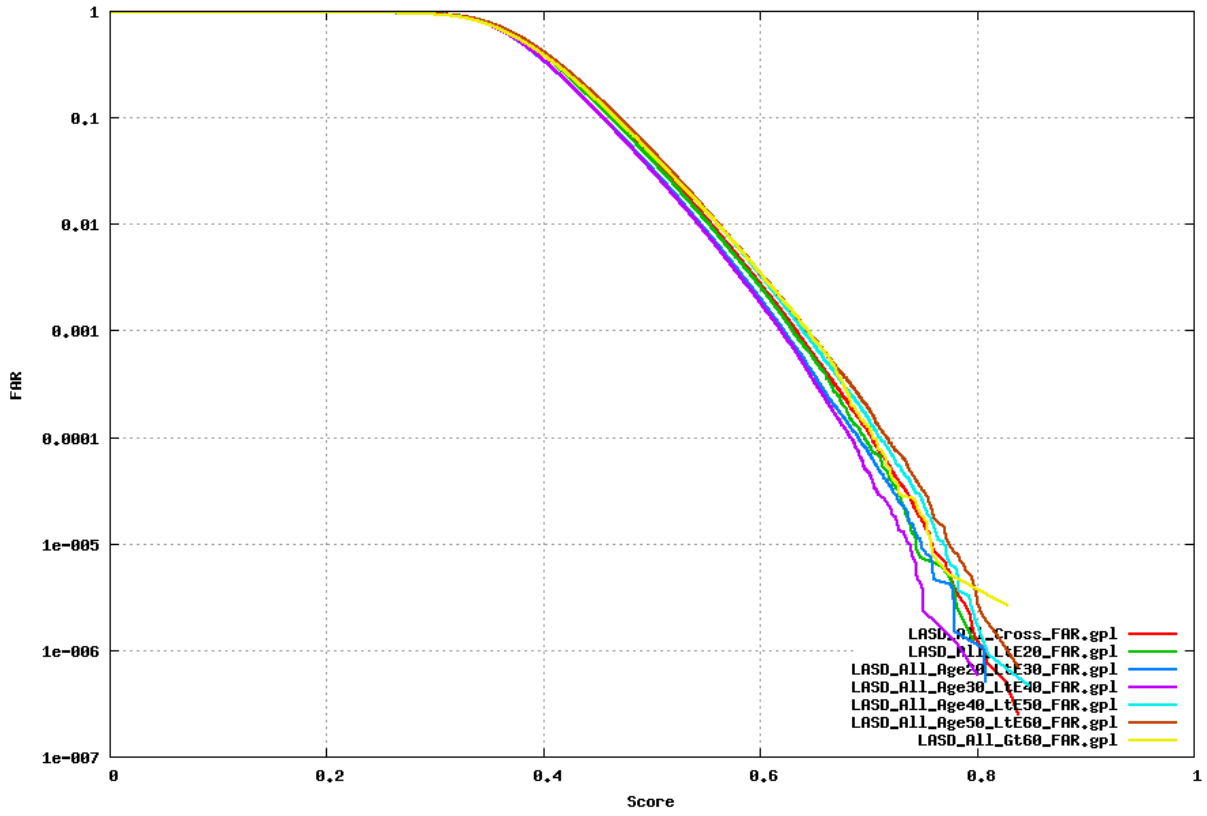


Figure 35: Arrest Age: FAR

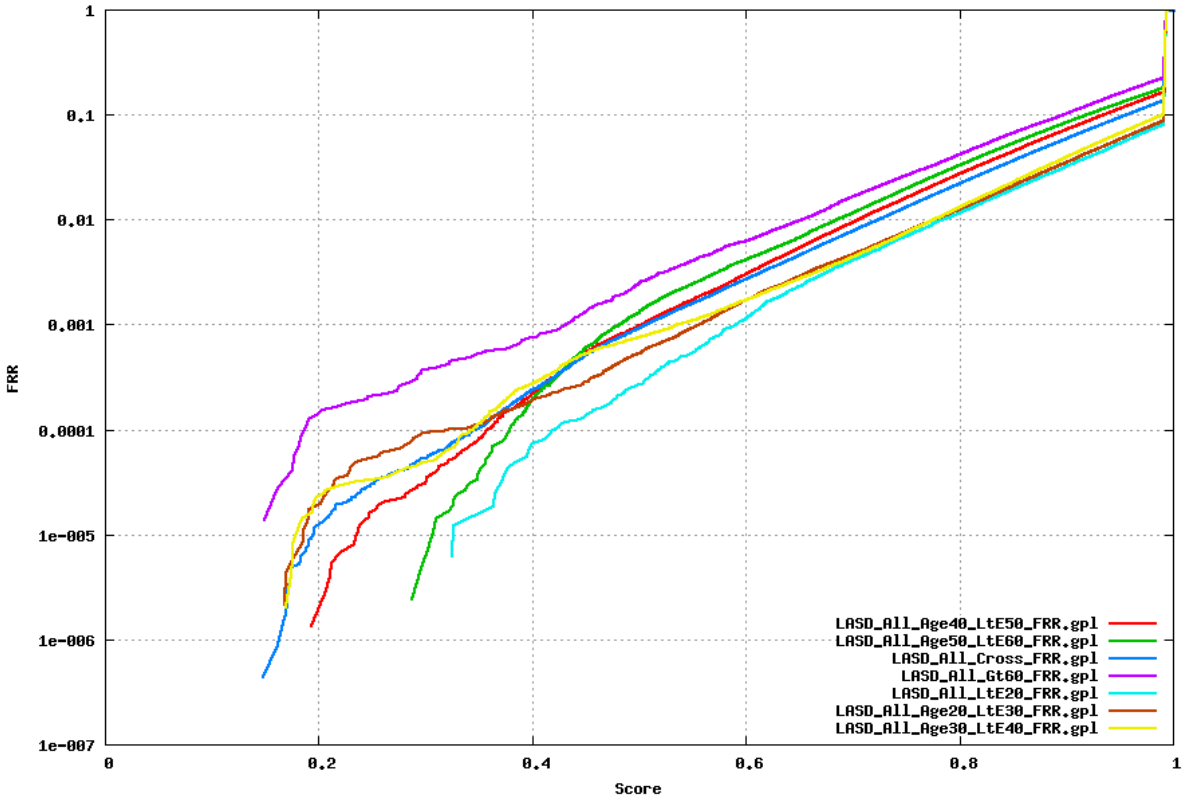


Figure 36: Arrest Age: FRR

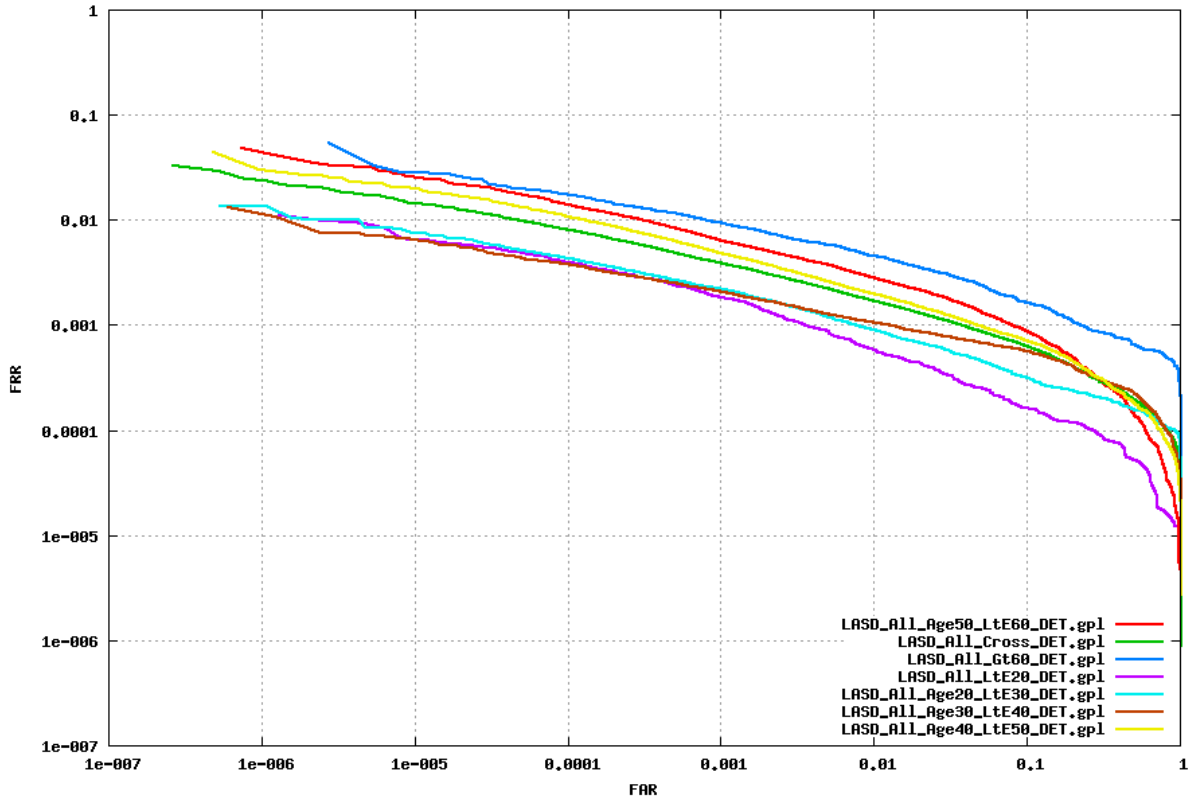


Figure 37: Arrest Age: DET

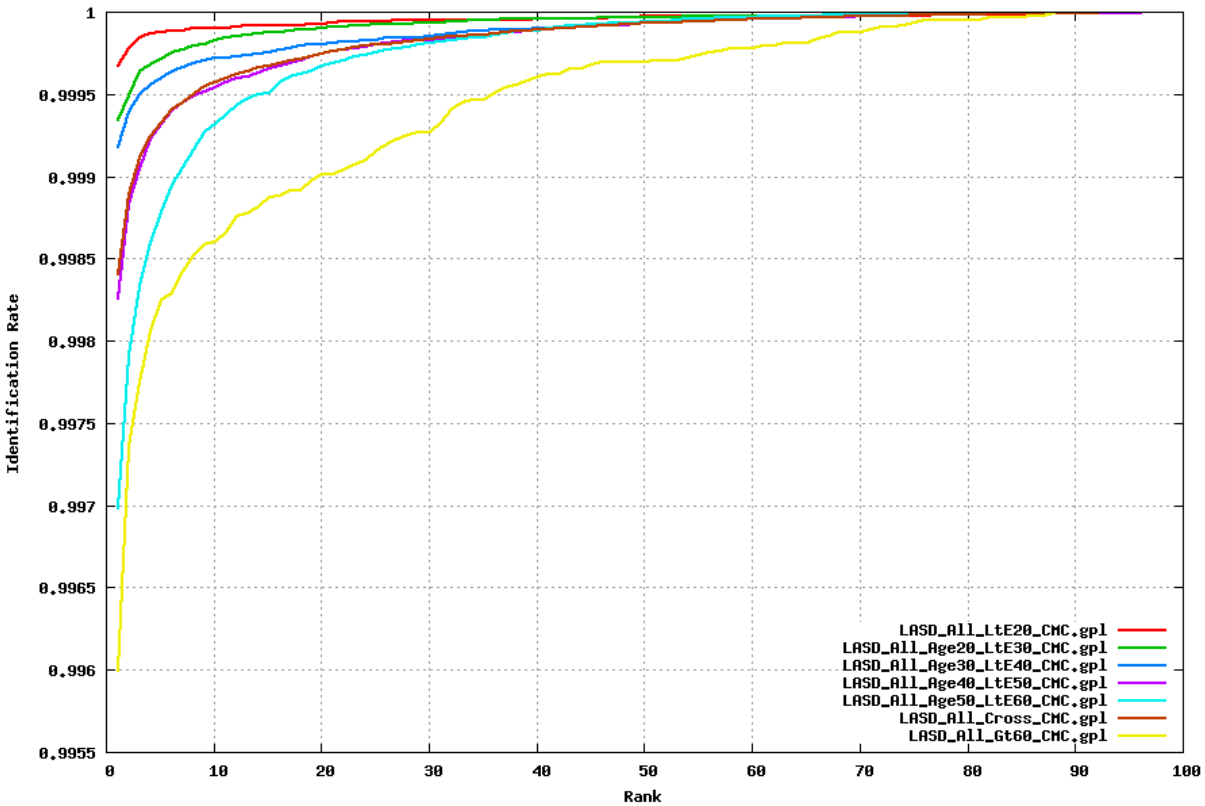


Figure 38: Arrest Age: CMC

9.6.1 Results

9.6.1.1 There are little variations when testing accuracy for age. Figures 34 and 35 show similar FAR and FRR scoring, while Figure 36 shows consistent DET performance.

9.6.1.2 Figure 37 shows a CMC rank one identification rate for all age ranges above 99.6%:

- Age greater than 60
- Ages between 50 and 60
- Ages between 40 and 50
- Ages between 30 and 40
- Ages between 20 and 30
- Age less than 20

10. Extended Testing

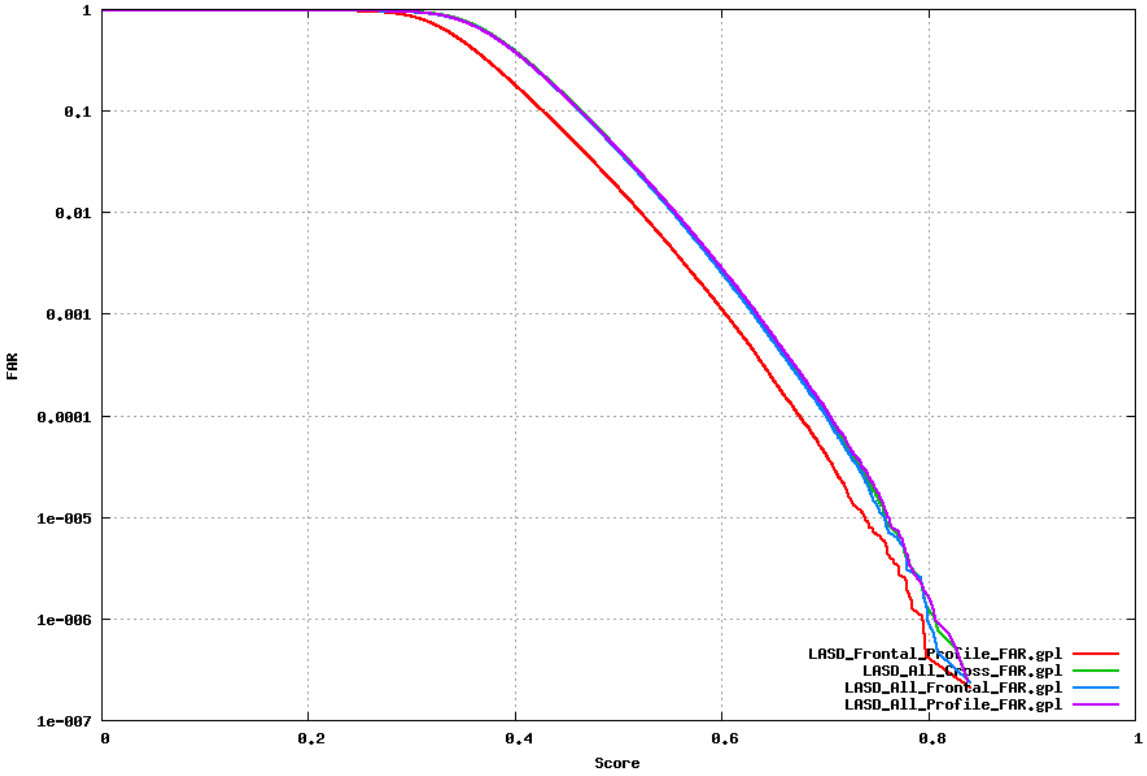
10.1 All charts that follow (Figures 38-45) compare the results from all imagery to a specific subset of imagery:

10.1.1 Pose: Entire gallery searched with frontal and profile poses, frontal gallery searched with profile poses

400 10.1.2 Low quality: low IOD frontal, low OCD profile, low vendor image quality

401 10.1.3 Images that needed to be manually localized

402 10.2 Pose Variation Testing



403
404 **Figure 39: Pose Variations: FAR**

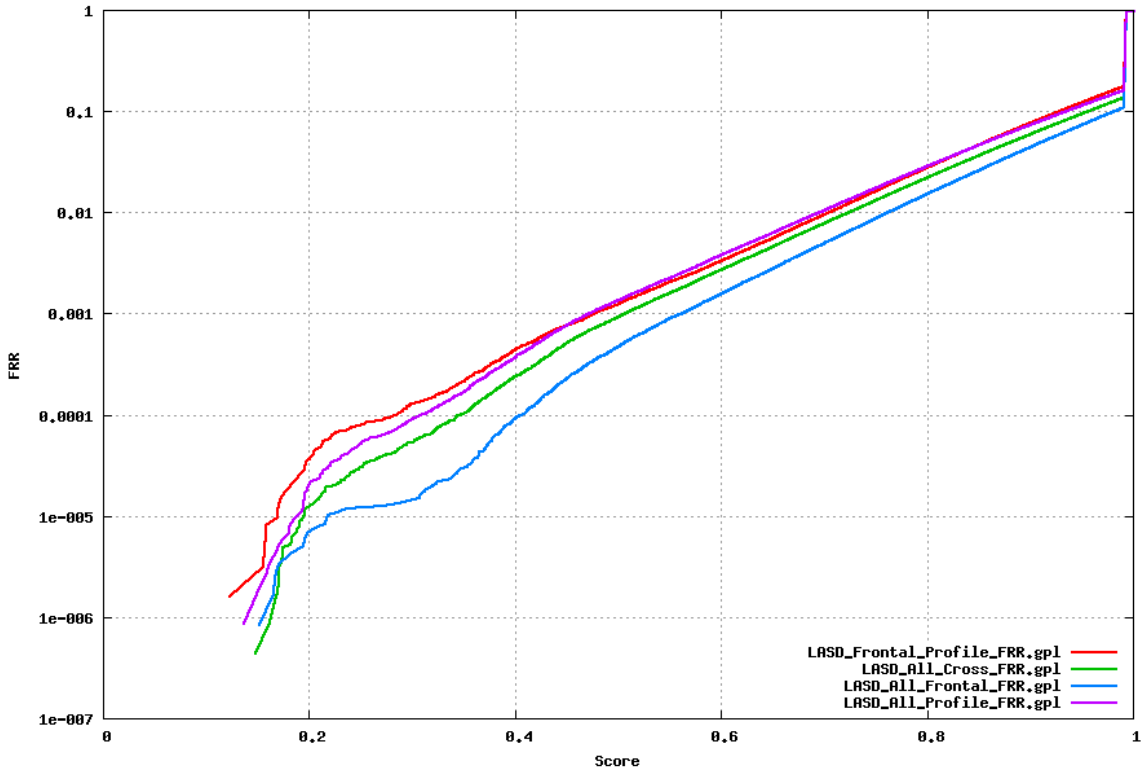


Figure 40: Pose Variations: FRR

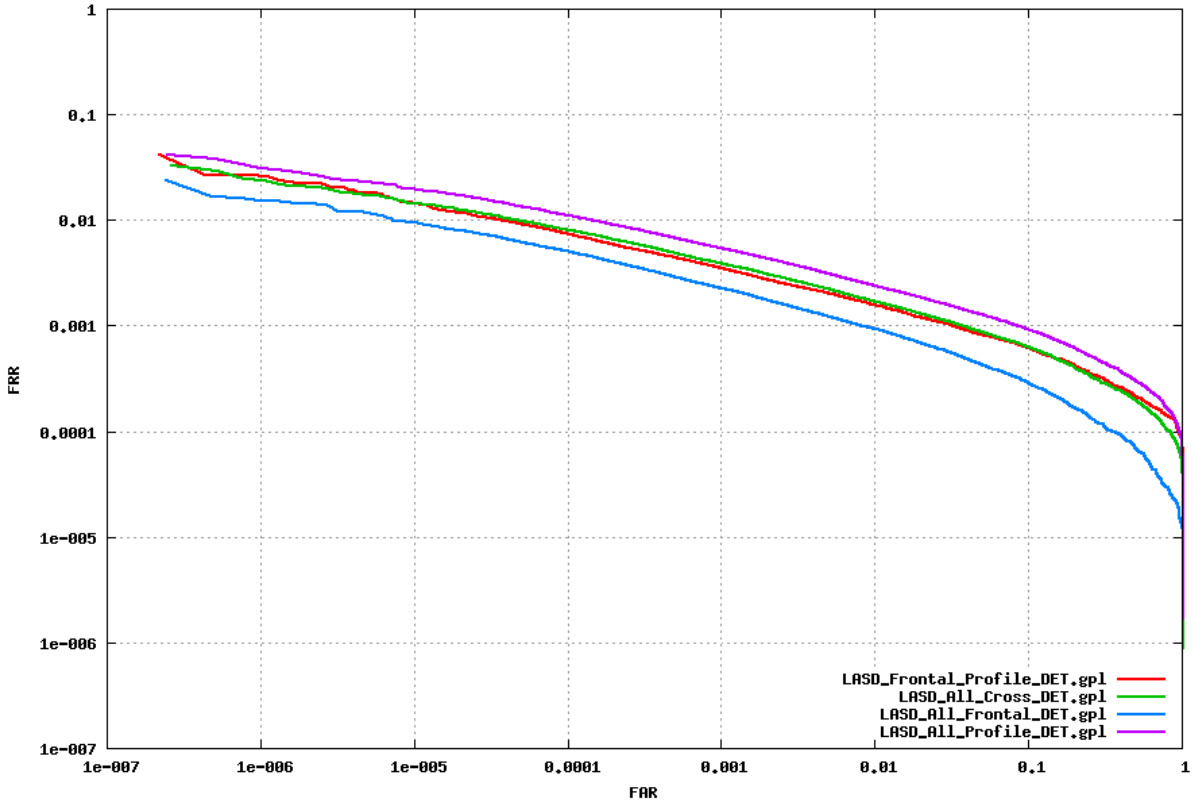


Figure 41: Pose Variations: DET

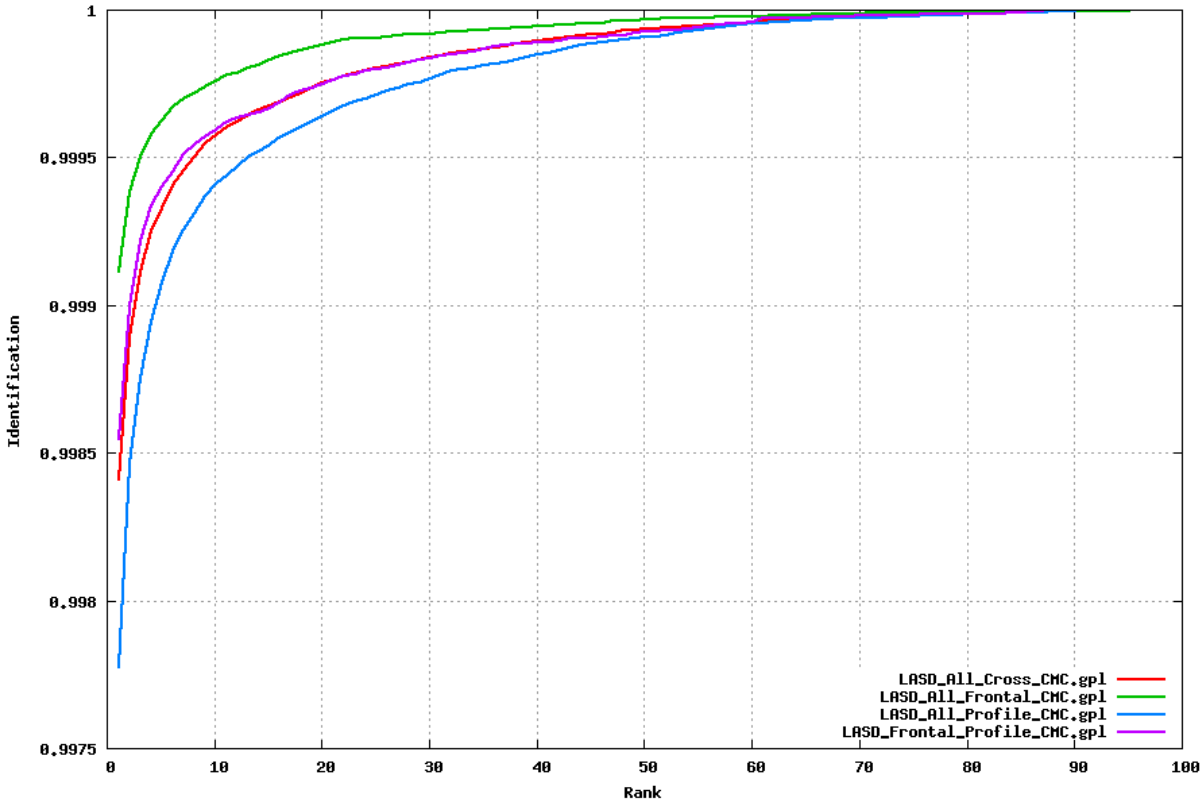


Figure 42: Pose Variations: CMC

10.2.1 Results

10.2.1.1 There are little variations when testing accuracy for facial pose. Figures 38 and 39 show similar FAR and FRR scoring, while Figure 40 shows consistent DET performance.

10.2.1.2 Figure 41 shows a CMC rank one identification rate for all pose variations was above 99.7%:

- Frontal probes searched against the entire gallery
- Profile probes searched against the entire gallery
- Profile poses searched against a frontal only gallery

10.3 Image Quality Variation Testing

10.3.1 Image quality testing has various aspects:

10.3.1.1 Small frontal faces as defined by IOD pixels

10.3.1.2 Small profile faces as defined by OCD pixels

10.3.1.3 Low quality faces as defined by the vendor facial algorithm

10.3.1.4 Manually localized faces which had a failure to encode but could be manually localized. 84 Images failed to template of which 19 had potential for some usability. These 19 were manually localized (both eyes and chin) and then searched.

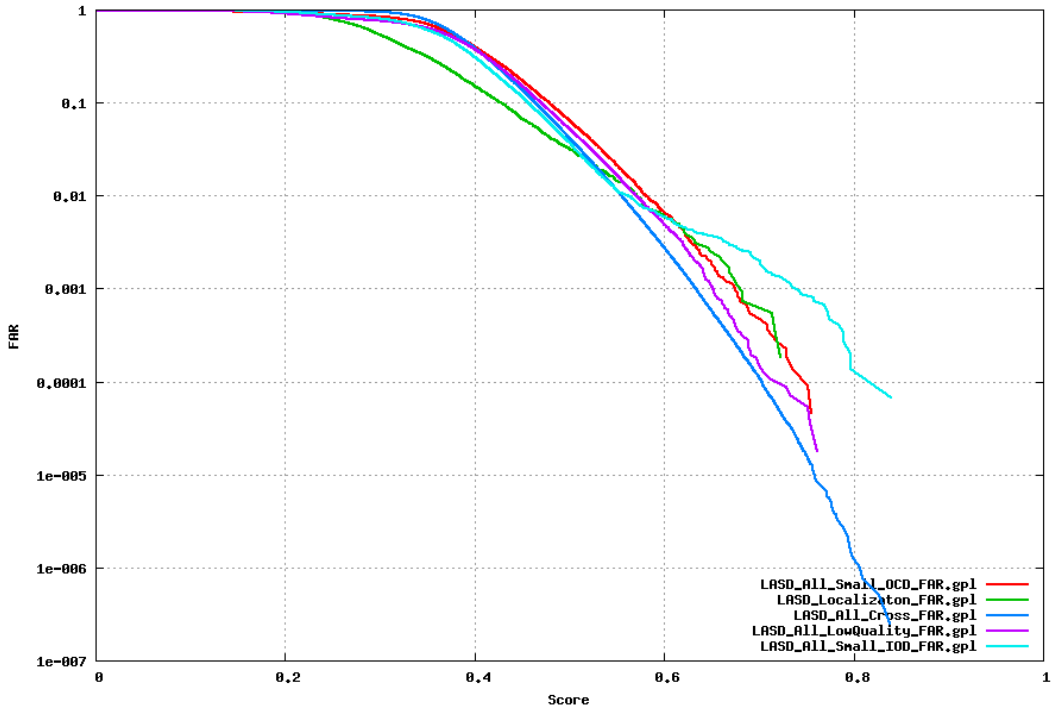


Figure 43: Quality Variations: FAR

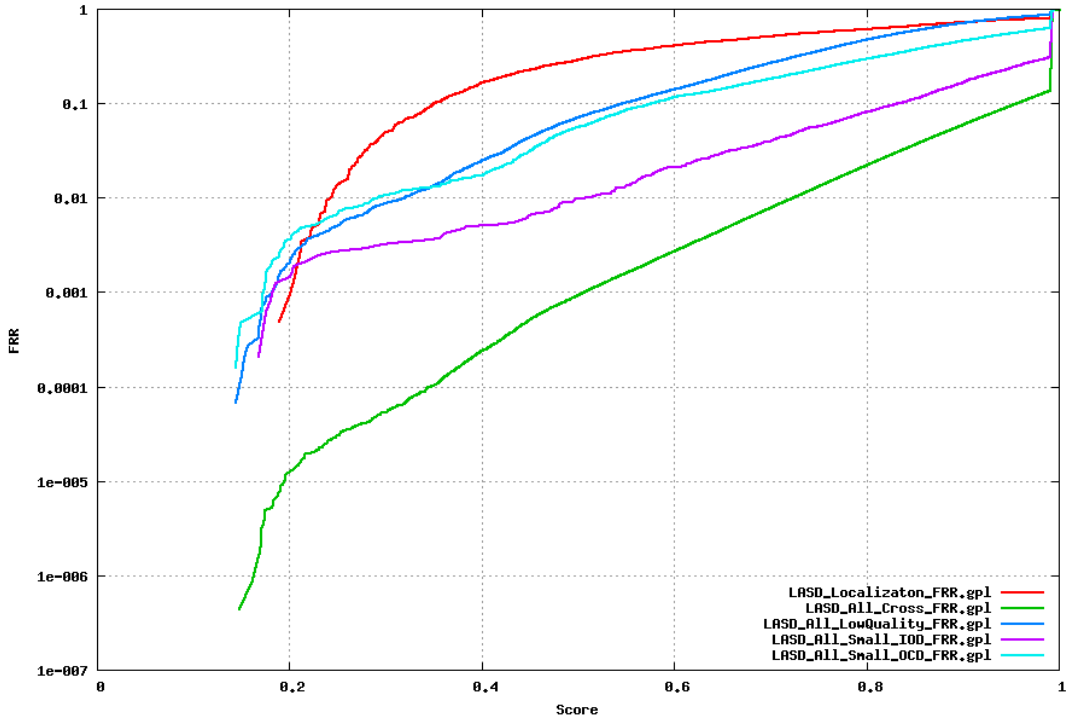


Figure 44: Quality Variations: FRR

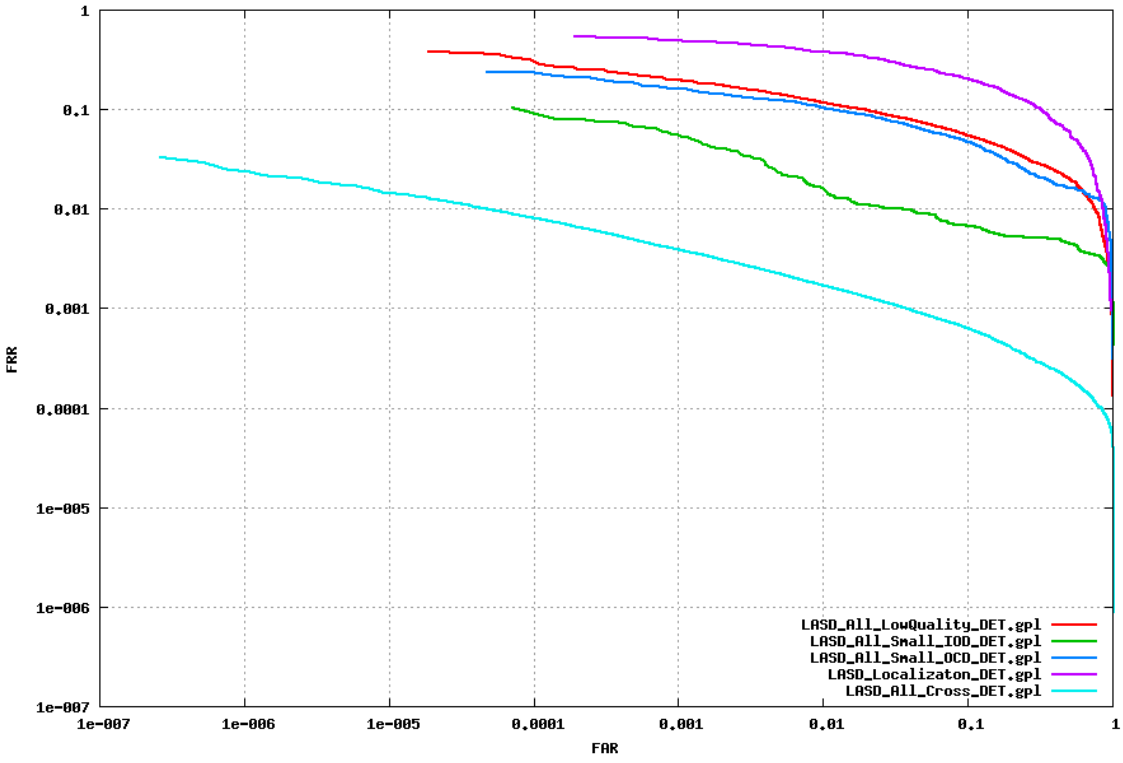


Figure 45: Quality Variations: DET

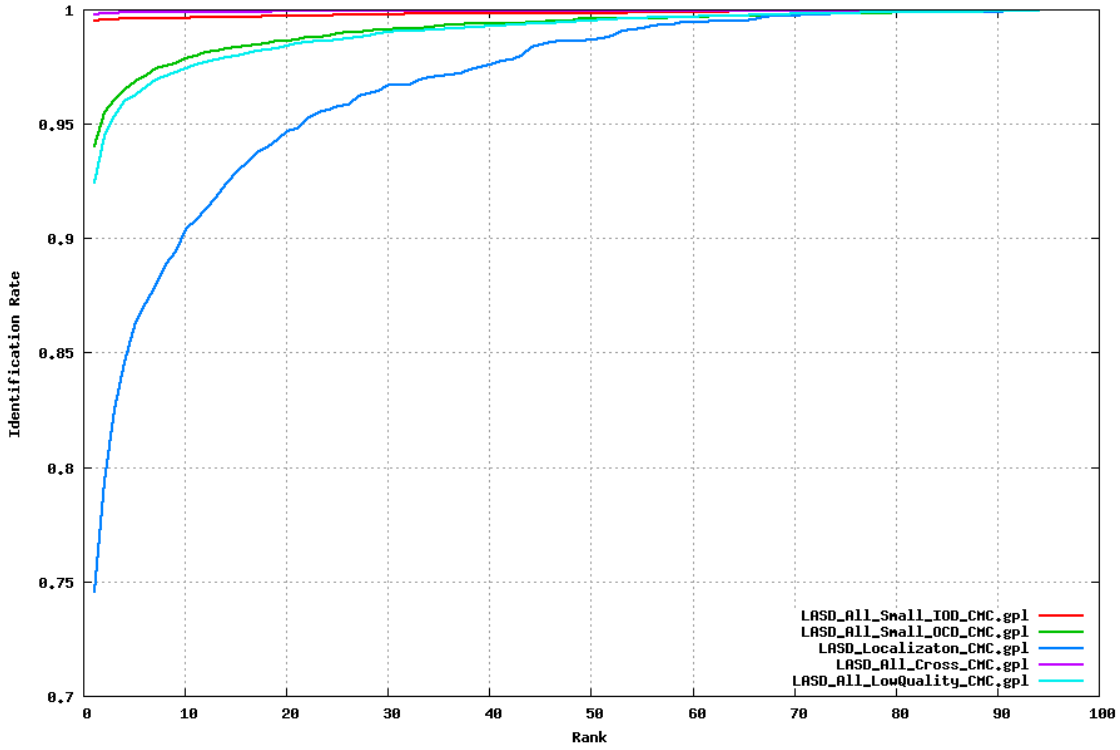


Figure 46: Quality Variations: CMC

10.3.2 Results

10.3.2.1 There are several variations when testing accuracy for image quality. Figure 42 and 43 shows varying imposter and mate scoring, while Figure 44 shows a wide difference in DET performance.

10.3.2.2 Figure 45 shows CMC at rank 1:

- Low IOD rank: 98%
- Low OCD rank: 94%
- Low quality rank: 92%
- Manually localized rank: 75%

11. IOD Variations

11.1 A subset of 20,000 frontal images was extracted and the images reduced in size from their original IOD (>= 90 pixels) to these IOD pixel ranges: 50, 40, 20, 20, and 10 pixels.

11.2 These reduced IOD images were then searched against the original images and accuracy charts produced.

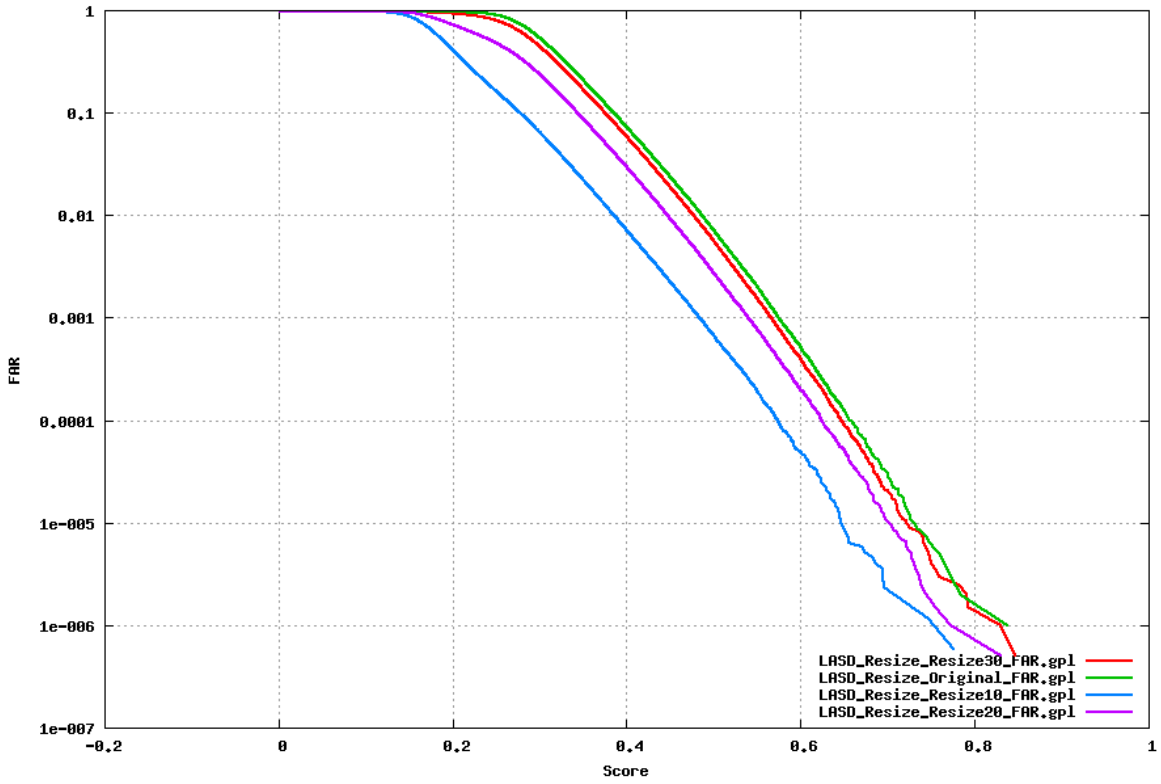


Figure 47: IOD Pixel Reductions: FAR

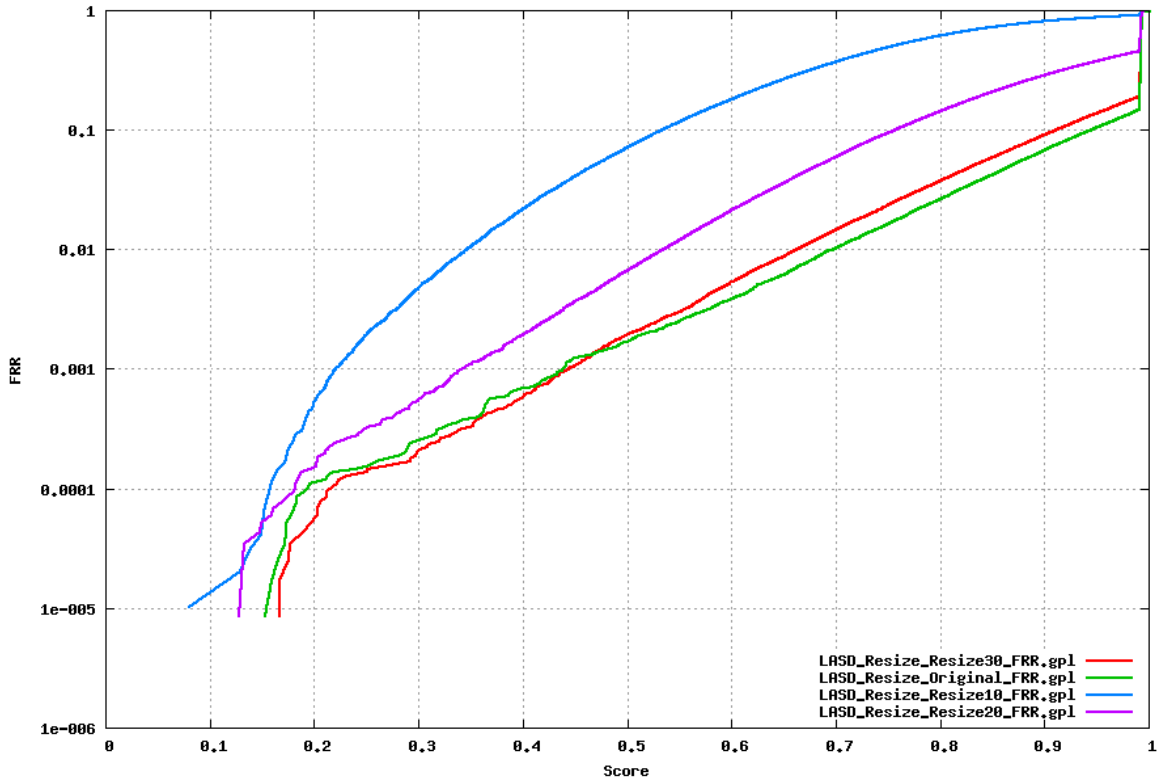


Figure 48: IOD Pixel Reductions: FRR

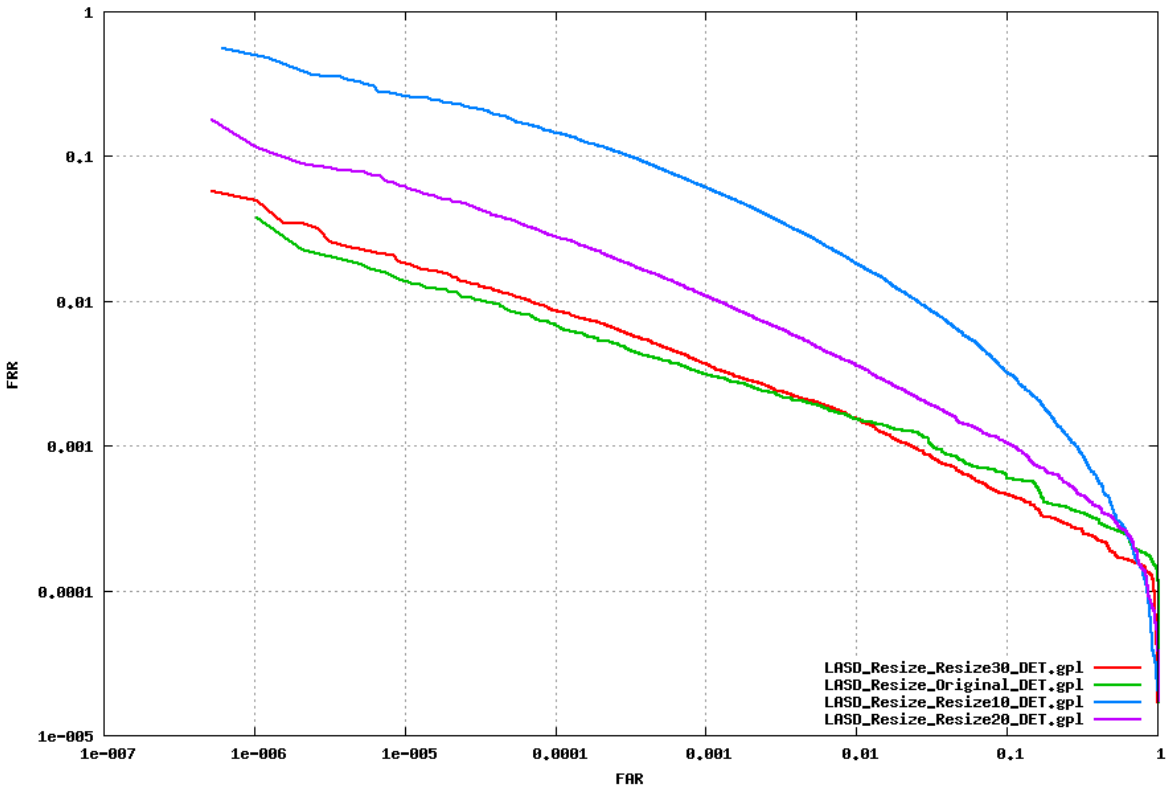


Figure 49: IOD Pixel Reductions: DET

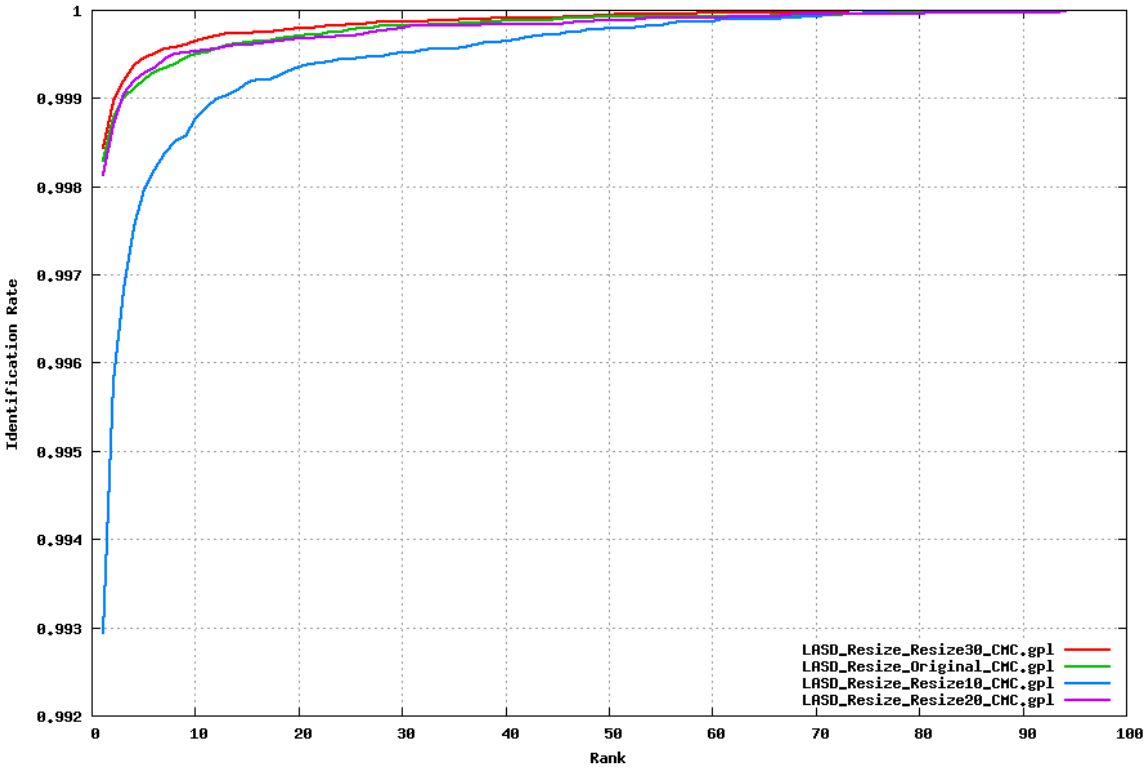


Figure 50: IOD Pixel Reductions: CMC

11.3 Results

11.3.1 FAR, FRR, DET, and CMC performance was consistent down to approximately thirty pixels IOD. Facial imagery with a twenty-pixel IOD showed an accuracy reduction. Facial imagery with a ten-pixel IOD showed a further accuracy reduction.

11.3.2 The results from facial imagery with a ten- or twenty-pixel IOD are large in terms of FAR and FRR score threshold determination for operational deployments as these image cohorts show a large and potentially significant reduction in accuracy translating from an assumed FAR to the resultant FRR.