NIST Internal Report
NIST IR 8485

Face Analysis Technology Evaluation (FATE) Part 11: Face Image Quality Vector Assessment

Specific Image Defect Detection

Joyce Yang
Patrick Grother
Mei Ngan
Kayee Hanaoka
Austin Hom

This publication is available free of charge from:
https://doi.org/10.6028/NIST.IR.8485

NATIONAL INSTITUTE OF
STANDARDS AND TECHNOLOGY
U.S. DEPARTMENT OF COMMERCE

NIST Internal Report
NIST IR 8485

Face Analysis Technology Evaluation (FATE) Part 11: Face Image Quality Vector Assessment

Specific Image Defect Detection

Joyce Yang
Patrick Grother
Mei Ngan
Kayee Hanaoka
Austin Hom
Image Group
Information Access Division
Information Technology Laboratory

This publication is available free of charge from:
https://doi.org/10.6028/NIST.IR.8485

September 2023

U.S. Department of Commerce
Gina M. Raimondo, Secretary

National Institute of Standards and Technology
Laurie E. Locascio, NIST Director and Under Secretary of Commerce for Standards and Technology

Disclaimer

Certain equipment, instruments, software, or materials are identified in this paper in order to specify the experimental procedure adequately. Such identification is not intended to imply recommendation or endorsement of any product or service by NIST, nor is it intended to imply that the materials or equipment identified are necessarily the best available for the purpose.

Institutional Review Board

The National Institute of Standards and Technology's Research Protections Office reviewed the protocol for this project and determined it is not human subjects research as defined in Department of Commerce Regulations, 15 CFR 27, also known as the Common Rule for the Protection of Human Subjects (45 CFR 46, Subpart A).

Publication History
Approved by the NIST Editorial Review Board on 2023-09-05

How to cite this NIST Technical Series Publication:
Joyce Yang, Patrick Grother, Mei Ngan, Kayee Hanaoka, Austin Hom (2023) Face Analysis Technology Evaluation (FATE) Part 11: Face Image Quality Vector Assessment. (National Institute of Standards and Technology, Gaithersburg, MD), NIST IR 8485. https://doi.org/10.6028/NIST.IR.8485

Contact Information
frvt@nist.gov

September 2023

Abstract

This report summarizes the results of the FATE Quality Vector assessment track, which tests face image quality algorithms' ability to detect specific defects such as non-frontal pose and background non-uniformity in the context of facial images. All algorithms submitted have some success at measuring various quality-related parameters.

Keywords

Face; defect; detection; image; quality; quality component; quality measure; specific.

Executive Summary

This report summarizes results from the Face Analysis Technology Evaluation (FATE) Quality Specific Image Defect Detection (SIDD) activity. It contains results for seven submissions from five participants: Digidata, FRP, Secunet, Neurotechnology, and Rank One. All algorithms submitted have some success at measuring various quality-related parameters. The measures that were implemented by all developers include total faces present, pitch, yaw, roll, eyes open, and inter-eye distance. As ISO/IEC 29794-5:2024 is finalized, we will continue to add, replace, and extend test cases and test datasets, to support development and to identify core capability. As this report is scrutinized by developers and end-users, comments and participation from other developers are welcomed.

RELEASE NOTES

2023-09-19: The FATE Quality SIDD track remains open.

- ▷ This document is the first release of the Quality SIDD report. It contains results for seven submissions from five participants: Digidata, FRP, Secunet, Neurotechnology, and Rank One.

  The procedure and format of submissions to the evaluation can be found in the API document [PDF].

Table of Contents

List of Tables

List of Figures

September 2023

Acknowledgments

The authors would like to thank the U.S. Department of Homeland Security Office of Biometric Identity Management (DHS OBIM) and the U.S. Department of Homeland Security's Science and Technology Directorate (S&T) for their collaboration and contributions to this activity. The authors are also grateful to staff in the NIST Biometrics Resesarch Laboratory for infrastructure supporting rapid evaluation of algorithms.

Other Relevant Reports

Results from the Face Recognition Technology Evaluation (FRTE) and Face Analysis Technology Evaluation (FATE) activities appear in the series of NIST Interagency Reports tabulated below. From 1999 to July 2023, FRTE and FATE were collectively known as FRVT.

Date	Link	Title	NISTIR
2014-03-20	PDF	FATE Performance of Automated Age Estimation Algorithms	7995
2015-04-20	PDF	FATE Performance of Automated Gender Classification Algorithms	8052
2014-05-21	PDF	FRTE Performance of Face Identification Algorithms	8009
2017-03-07	PDF	Face In Video Evaluation (FIVE) Face Recognition of Non-Cooperative Subjects	8173
2017-11-23	PDF	The 2017 IARPA Face Recognition Prize Challenge (FRPC)	8197
2020-01-03	Draft	FRTE - Part 1: Verification	Draft
2019-09-11	PDF	FRTE - Part 2: Identification	8271
2019-12-11	PDF	FRTE - Part 3: Demographic Effects	8280
2020-03-04	PDF	FATE - Part 4: MORPH - Performance of Automated Face Morph Detection	8292
2020-03-06	Draft	FATE - Part 5: Face Image Quality Assessment	Draft
2020-07-24	PDF	FRTE - Part 6A: Face Recognition Accuracy with Face Masks using Pre-COVID-19 Algorithms	8311
2022-01-20	PDF	FRTE - Part 6B: Face Recognition Accuracy with Face Masks using Post-COVID-19 Algorithms	8331
2022-07-13	PDF	FRTE - Part 7: Identification for Paperless Travel and Immigration	8381
2022-09-30	PDF	FRTE - Part 8: Summarizing Demographic Differentials	8429
2022-09-30	PDF	FRTE - Part 9A: Face Recognition Verification Accuracy on Distinguishing Twins	8439
2023-09-20	PDF	FATE - Part 10: Performance of Passive, Software-based Presentation Attack Detection (PAD) Algorithms	8491
2023-09-20	PDF	FATE - Part 11: Face Image Quality Vector Assessment: Specific Image Defect Detection	8485

Details appear on pages linked from https://www.nist.gov/programs-projects/face-projects.

1. Introduction

Consider the procedure of taking a passport photo: whether for renewal or obtaining a visa, there are formal standards that the capture subject and photographer must follow in order to take an acceptable photo. These standards vary in the required photo size, but commonly require a frontal viewpoint, open eyes, a neutral expression, a uniform background, and other criteria to be fulfilled; for a detailed discussion of these criteria, see Annex D1 of ISO/IEC 39794-5:2019 . These standards and the practices needed to conform to them are intended to support highly accurate face recognition by ensuring that the captured photo can serve as a high quality reference photo in a machine readable travel document (e.g. passport) or in a reference database (e.g. the IDENT system in the US, or the EU-VIS BMS system in Europe).

This Face Analysis Technology Evaluation (FATE) track, Specific Image Defect Detection (SIDD), is being conducted to support quality assessment in general, and to support assessment of quality component algorithms that implement the quality checks of ISO/IEC 29794-5:2024 (under development). That standard enumerates checks on face photos that derive from ISO/IEC 19794-5:2011, which established photographic and subject appearance requirements for enrollment images in the European Entry-Exit-System (according to EU-EES implementing decision 2019/329), and ISO/IEC 39794-5:2019, which refined and extended photograph specification and will be used for e-Passports from 2030 onwards.

The existing FATE Quality Summarization Track is an ongoing track that examines the relation between quality score and false non-match rates in order to gauge how well a quality component algorithm can predict false negative errors. However, it does not differentiate between different factors (i.e. quality components) that affect quality. In the SIDD track, we delve deeper into a nuanced discussion of quality measures.

The procedure and format of submissions to our evaluation are described in the Quality SIDD Assessment API document.

2. Test Sets

2.1. Development

The Quality SIDD assessment proceeds by passing photographs to algorithms using a NIST-defined C++ API. The test sets consist of images sequestered at NIST, i.e. developers do not have access to the images. NIST has curated sets specifically to evaluate the performance of algorithms measuring the quantities given in Table 3. For example, there are various sets of frontal and non-frontal images to evaluate pose estimation accuracy; the images in these sets have known pitch and yaw angles.

We formulate ground truth for test sets using three approaches:

- Camera placement: At the time of capture, a camera is placed at a specific angle to the subject.
- Manual labeling: We determine ground truth by human inspection, by measuring the desired quantity using software such as GIMP.
- Synthetic degradation: We generate images with different degrees of a defect (blur, overexposure and underexposure) by applying varying amounts of a defect to a natural image.

This section contains results from all algorithms submitted from the inception of the Quality SIDD evaluation in July 2022.

2.2. Limitations

For several measures, such as inter-eye distance and mouth aperture, we use ground truth that is determined by human inspection. This style of testing, in which ground truth is a continuous variable with some measurement error, means that the software can never be perfect. This is in contrast to a recognition test, for example, where labels are discrete and, ideally, error-free.

2.3. Test Set Sizes

Table 1 lists the number of images in each test set for the quality measures that were implemented so far.

Table 1. Set sizes. This table presents the quality measures in the SIDD track and the number of images in each test set.

Dataset	Number of Images
TotalFacesPresent	92
SubjectPosePitch-1	6291
SubjectPosePitch-2	7145
SubjectPoseYaw-1	6267
SubjectPoseYaw-2	34014
SubjectPoseRoll	12000
EyesOpen	107
InterEyeDistance-1	40
InterEyeDistance-2	39
Resolution	8000
MouthOpen	145
BackgroundUniformity	229
Underexposure	250
Overexposure	250
EyeGlassesPresent	279
SunGlassesPresent	40
CompressionArtifacts	500
FaceOcclusion	30
MotionBlur	6000
PixelsFromEyeToLeftEdge	40
PixelsFromEyeToRightEdge	40
PixelsFromEyesToTop	40
PixelsFromEyesToBottom	40

3. Algorithms and Results

3.1. Algorithms

Table 2 lists the participants who submitted algorithms to the Quality SIDD Assessment.

Table 2. Quality SIDD Assessment Participants

Participant Name	Short Name	Sequence Number	Submission Date
Digidata	digidata	001	2022.09.29
FRP LLC	frpkauai	000	2022.10.28
Secunet Security Networks AG (part of OFIQ)	secunet	001	2023.02.16
Secunet Security Networks AG (part of OFIQ)	secunet	002	2023.04.21
Neurotechnology	neurotechnology	002	2023.07.10
Rank One Computing	rankone	005	2023.07.14
Neurotechnology	neurotechnology	003	2023.08.10

3.2. Quality Measures Supported

Table 3 lists the quality measures defined in our API and the algorithms that implement them. The first row indicates whether the quality measure must be checked for an image to be used as a reference image in a machine-readable travel document (MRTD) such as a passport, which is Use Case 1 (UC1) as listed in ISO/IEC 29794-5:2024.

Table 3. Quality Measures Supported. This table presents the participating algorithm name and which SIDD quality measures were implemented. A 'Y' (for 'Yes') indicates that the quality measure is implemented; a blank space indicates that it is not implemented. The first row indicates which quality measures are required to be checked for use as a reference photo in a machine-readable travel document (MRTD) according to ISO/IEC 29794-5:2024.

Algorithm	TotalFacesPresent	SubjectPosePitch	SubjectPoseYaw	SubjectPoseRoll	EyesOpen	InterEyeDistance	Resolution	MouthOpen	BackgroundUniformity	Underexposure	Overexposure	PixelsFromEyeToLeftEdge	PixelsFromEyeToRightEdge	PixelsFromEyesToTop	PixelsFromEyesToBottom	EyeGlassesPresent	SunGlassesPresent	CompressionArtifacts	FaceOcclusion	MotionBlur	UnifiedQualityScore
Required for MRTD	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y					Y	Y
digidata-001	Y	Y	Y	Y	Y	Y	Y	Y		Y	Y										Y
frpkauai-000	Y	Y	Y	Y	Y	Y	Y		Y												
neurotechnology-002	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
neurotechnology-003	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
rankone-005	Y	Y	Y	Y	Y	Y	Y	Y				Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
secunet-001	Y	Y	Y	Y	Y	Y		Y	Y	Y	Y										
secunet-002	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y										Y

3.3. Timing

The duration of execution of quality algorithm (QA) software is important in those applications where fast quantification is needed to support usability by providing usable feedback to a capture subject. It may be important also, for example, in running QA software over large legacy collections. This section gives duration of the various implementations running on a common hardware platform.

Figure 1 shows the timing performance for the participants who submitted algorithms to the Quality SIDD Assessment.

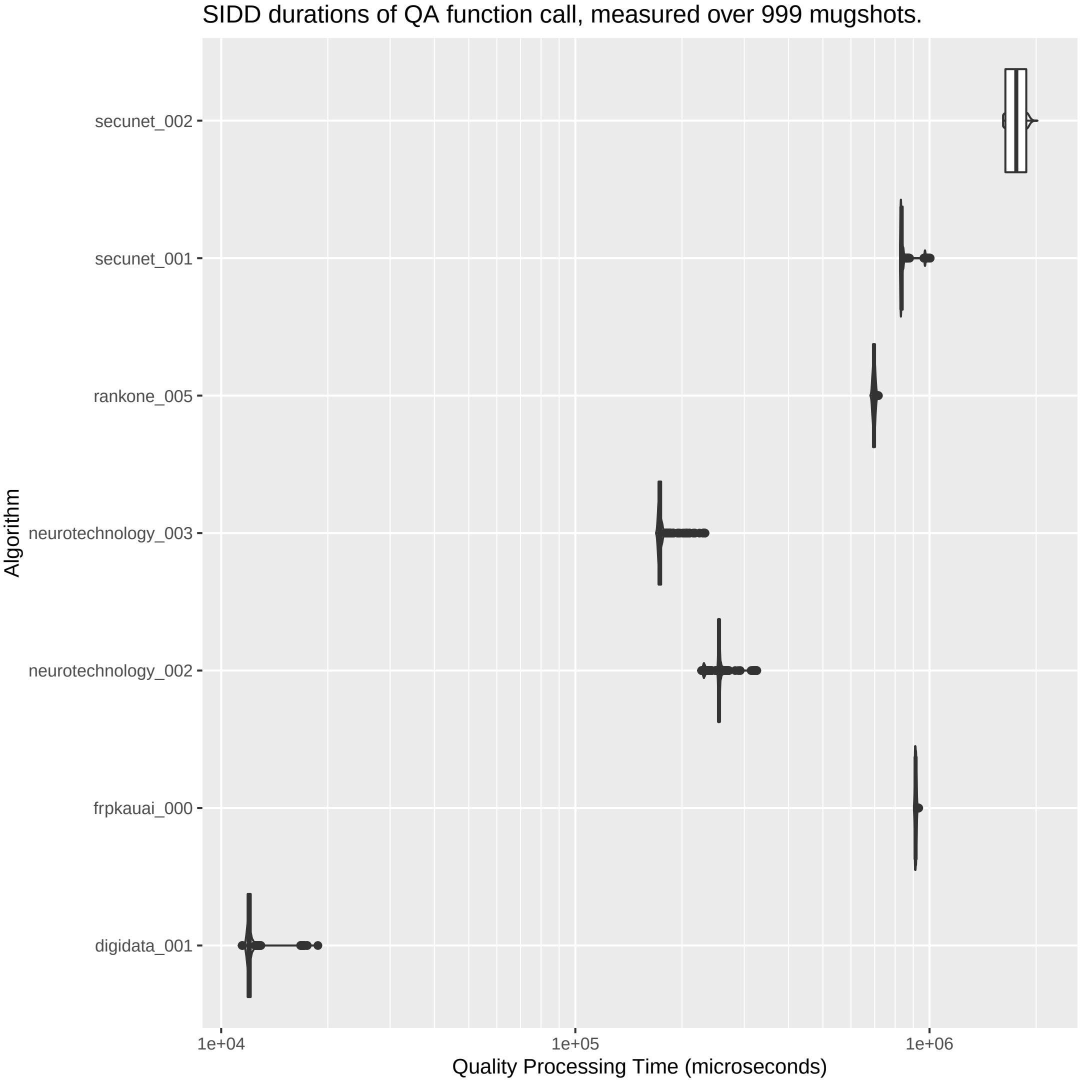


Fig. 1. Distribution of time required for the quality algorithm (QA) function call, measured over 999 mugshots. The implementations vary, in part, because they are computing different quality components – see Table 3. Durations are measured on a fixed Intel Xeon Gold 6140 CPU running at 2.30 GHz. Durations are measured by wrapping the function call in a high resolution timer.

3.4. Total Faces Present

3.4.1. Images Used

The images in the Total Faces Present dataset are captured in a border-crossing setting with a variety of poses and some background non-uniformity. The input images generally have one primary face that is larger than the others. We count faces manually, where a face is counted if its inter-eye distance is estimated to be larger than 0.02 times the width of the image.

3.4.2. Results for Total Faces Present

Figure 2 summarizes the performance of all algorithms that implemented the Total Faces Present measure. Note that there are more missed detections (below the diagonal) than false detections (above the diagonal). Missed detection rate is the number of missed faces divided by the total number of faces; false detection rate is the average number of wrong detections per image. For both false detection rate and missed detection rate, lower values are better.

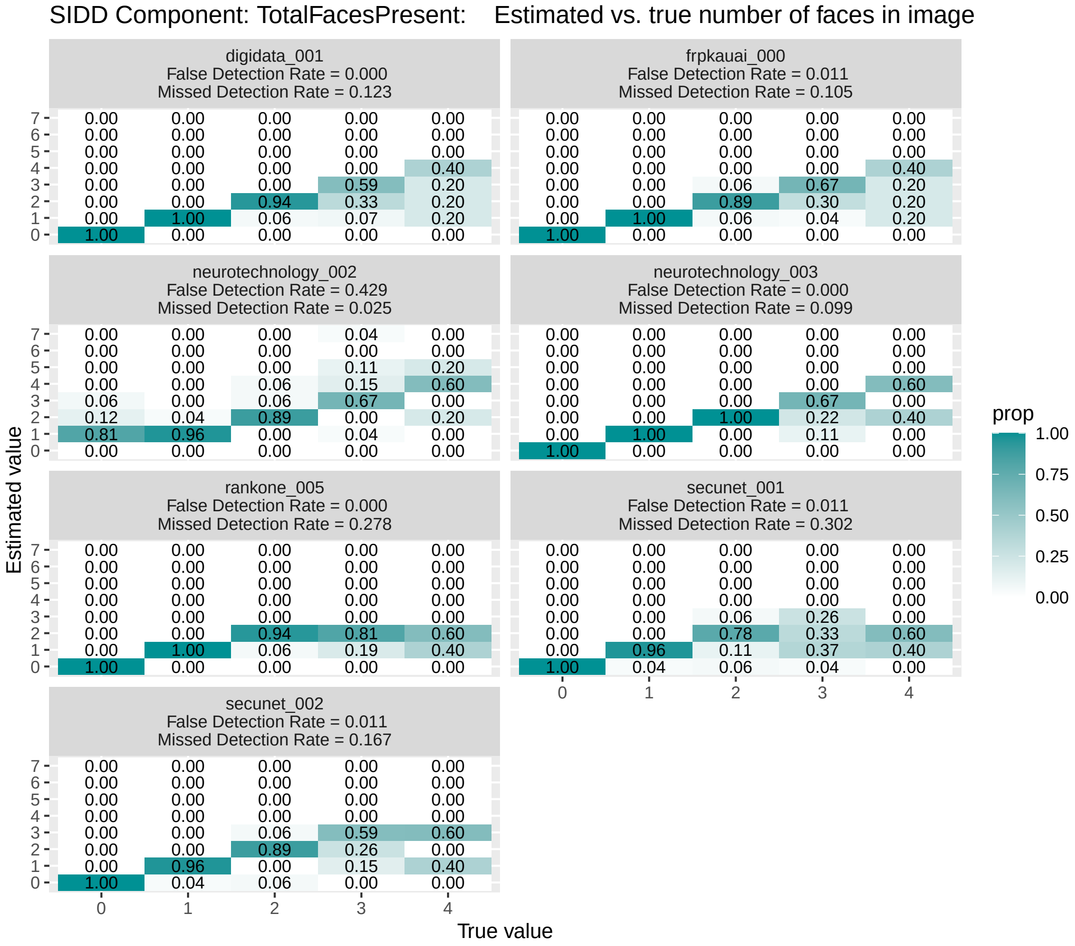


Fig. 2. Estimated number of faces vs. known number of faces. Perfect performance corresponds to zero on off-diagonal entries and 1 on each of the diagonal entries. Each column is normalized by the sum along that column, so that the numbers give estimates of the rate at which the software gives a false detection (above the diagonal), missed detection (below the diagonal), and correct detection (on the diagonal). Missed detections can occur because algorithms are generally configured to only detect faces larger than a certain size. The value at (x, y) is the proportion of images with known number of faces x and detected number of faces y. Missed detection rate is the number of missed faces divided by the total number of faces; false detection rate is the average number of wrong detections per image. For both false detection rate and missed detection rate, lower values are better.

3.5. Yaw Angle

3.5.1. Images Used

The images for the Yaw quality measure are from two sets of sequestered photos. The images in these sets have a well-illuminated setting with a uniform background.

1. For the first set, at the time of capture, a camera is placed to the right or left of the subject at varying angles, with the subject remaining frontal and stationary. Yaw is recorded at the time of collection.
2. For the second set, at the time of capture, the subject turns the head to look at a target to the left or right. Yaw is recorded at the time of collection.

Camera placement to the subject's right corresponds to yaw being positive. Camera placement to the subject's left corresponds to yaw being negative. This sign convention is consistent with the ISO/IEC 39794-5:2019 standard.

3.5.2. Results for Yaw Angle

Table 4, Figure 3, and Figure 4 summarize algorithm performance. The Median Absolute Error (MAE), where error is computed as the difference between ground truth and reported value, is shown for each algorithm. Lower MAE is better.

Table 4. SIDD PoseYaw Median Absolute Error.

Algorithm	Dataset	MAE (in degrees)
digidata_001	Yaw Set 1	8.6
frpkauai_000	Yaw Set 1	4.7
neurotechnology_002	Yaw Set 1	8.0
neurotechnology_003	Yaw Set 1	8.0
rankone_005	Yaw Set 1	10.0
secunet_001	Yaw Set 1	2.9
secunet_002	Yaw Set 1	4.5
digidata_001	Yaw Set 2	15.8
frpkauai_000	Yaw Set 2	7.0
neurotechnology_002	Yaw Set 2	8.0
neurotechnology_003	Yaw Set 2	8.0
rankone_005	Yaw Set 2	10.0
secunet_001	Yaw Set 2	8.3
secunet_002	Yaw Set 2	8.4

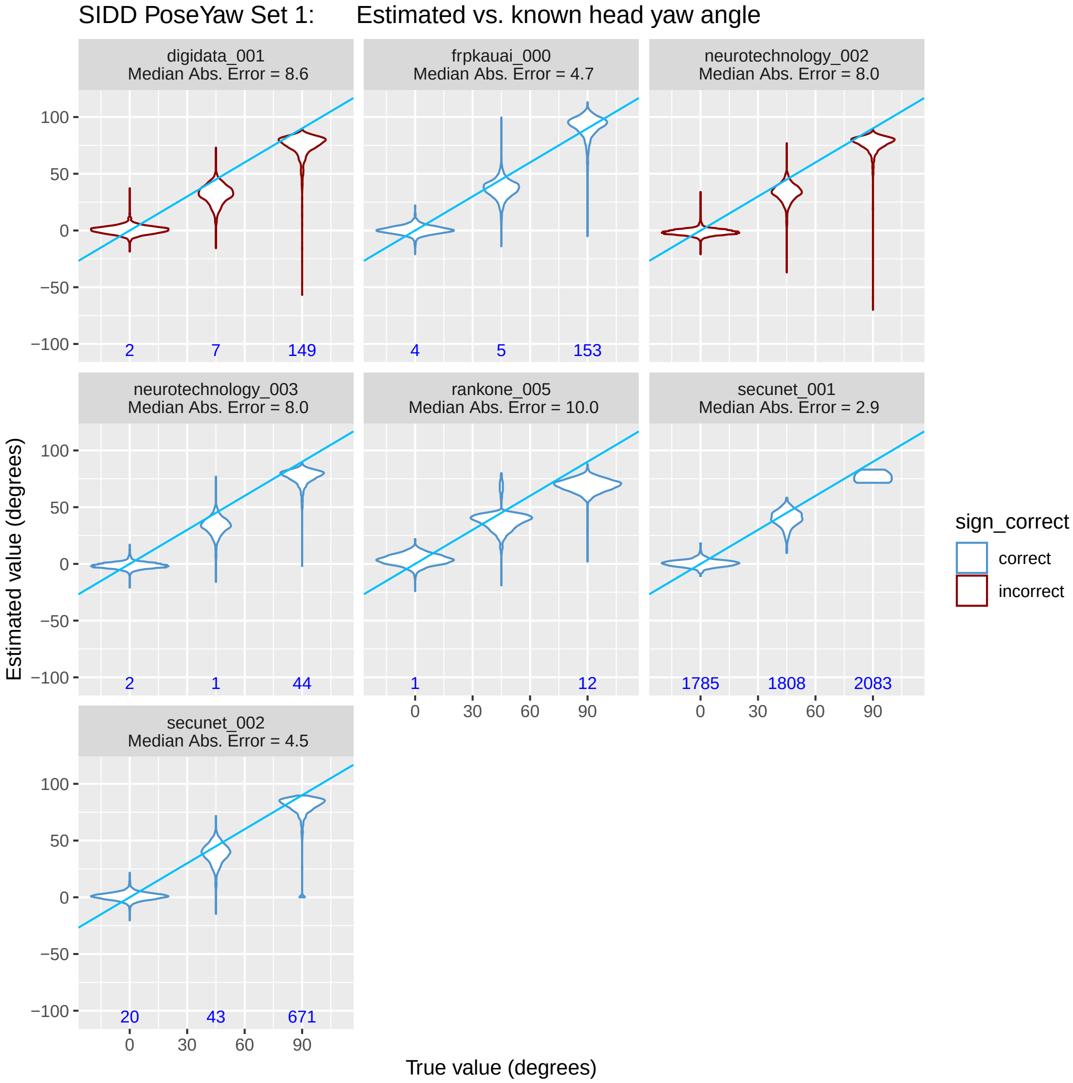


Fig. 3. Estimated vs. known values of yaw angle. For Yaw Set 1, ground truth yaw values are determined by the placement of the camera at the time of capture; the subject remains stationary and frontal. The blue line $(y = x)$ represents perfect performance. The plot shows violins at true yaw values with the tails extending to the minimum and maximum estimated values. The small numbers along the horizontal line at -100 represent the count of faces when the software did not return an estimate; for example, when it did not detect a face. The dark red color-coding indicates that the developer uses the opposite sign convention, and should negate the scores in the next submission.

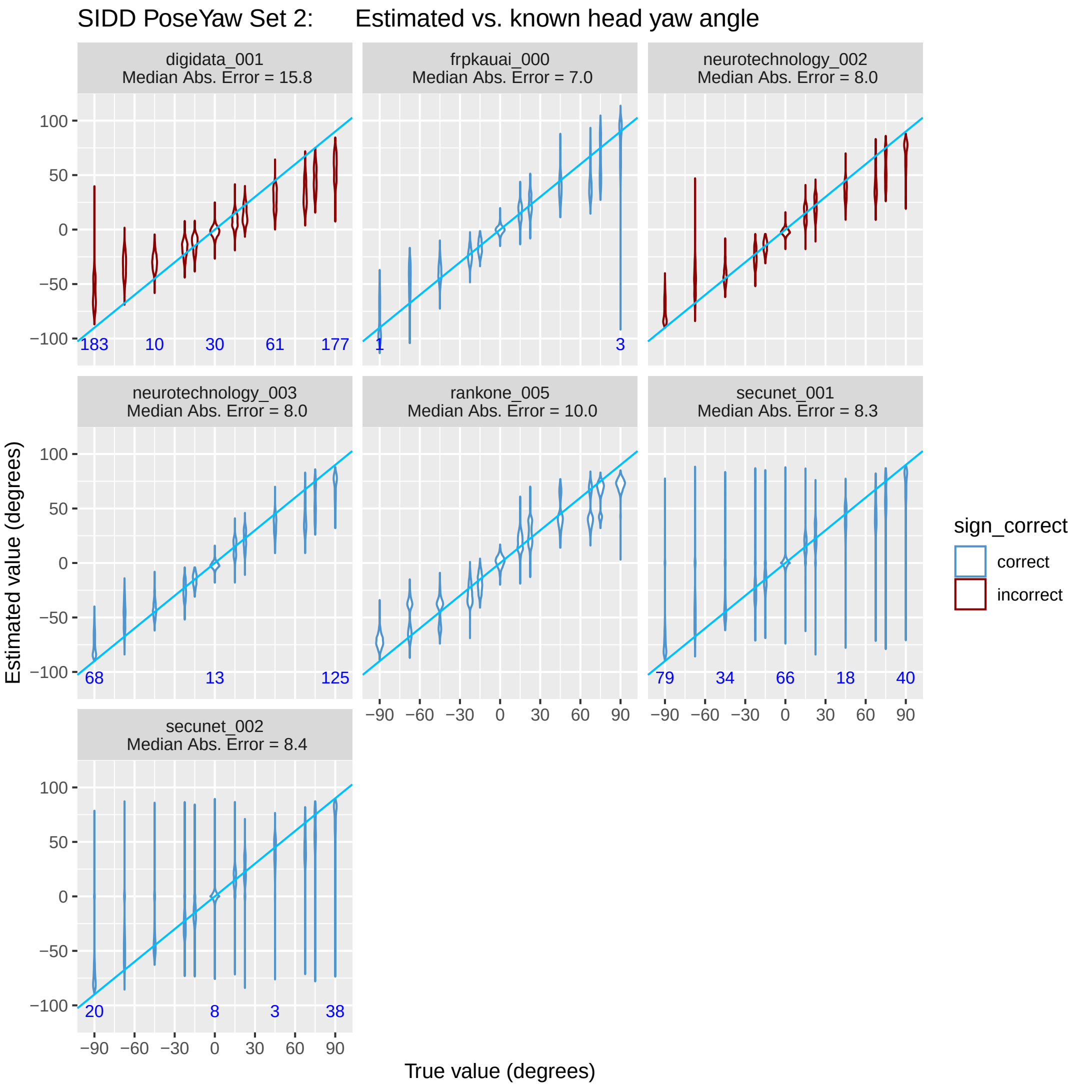


Fig. 4. Estimated vs. known values of yaw angle. For Yaw Set 2, the subject turns the head to look at targets placed to the right or left. The blue line $(y = x)$ represents perfect performance. The plot shows violins at true yaw values with the tails extending to the minimum and maximum estimated values. The small numbers along the horizontal line at -100 represent the count of faces when the software did not return an estimate; for example, when it did not detect a face. The dark red color-coding indicates that the developer uses the opposite sign convention, and should negate the scores in the next submission.

3.6. Pitch Angle

3.6.1. Images Used

The images for the Pitch quality measure are from two sets of sequestered photos.

1. In Set 1, the subject generally has a neutral position and is standing against a mostly uniform background, with some shadows behind the subject. At the time of capture, a camera is placed at varying heights; the subject is asked to be frontal. Pitch is recorded at the time of collection.

2. In Set 2, the subject is seated and is against a uniform background. At the time of capture, a camera is placed at varying heights; the subject is asked to be frontal. Pitch is recorded at the time of collection.

Camera placement above the subject, with the top of the head being more exposed, corresponds to pitch being positive, and placement below the subject, with the chin being more exposed, corresponds to pitch being negative. This sign convention is consistent with the ISO/IEC 39794-5:2019 standard.

3.6.2. Results for Pitch Angle

Table 5, Figure 5, and Figure 6 summarize the performance of the algorithms in our evaluation when estimating pitch angle. The Median Absolute Error (MAE), where error is computed as the difference between ground truth and reported value, is shown for each algorithm. Lower MAE is better.

Note that the definition of zero-pitch is not as well-defined as zero-roll and zero-yaw.

Table 5. SIDD PosePitch Median Absolute Error.

Algorithm	Dataset	MAE (degrees)
digidata_001	Pitch Set 1	6.4
frpkauai_000	Pitch Set 1	12.3
neurotechnology_002	Pitch Set 1	6.0
neurotechnology_003	Pitch Set 1	6.0
rankone_005	Pitch Set 1	12.0
secunet_001	Pitch Set 1	4.2
secunet_002	Pitch Set 1	5.1
digidata_001	Pitch Set 2	11.1
frpkauai_000	Pitch Set 2	7.7
neurotechnology_002	Pitch Set 2	10.0
neurotechnology_003	Pitch Set 2	10.0
rankone_005	Pitch Set 2	10.0
secunet_001	Pitch Set 2	13.9
secunet_002	Pitch Set 2	13.9

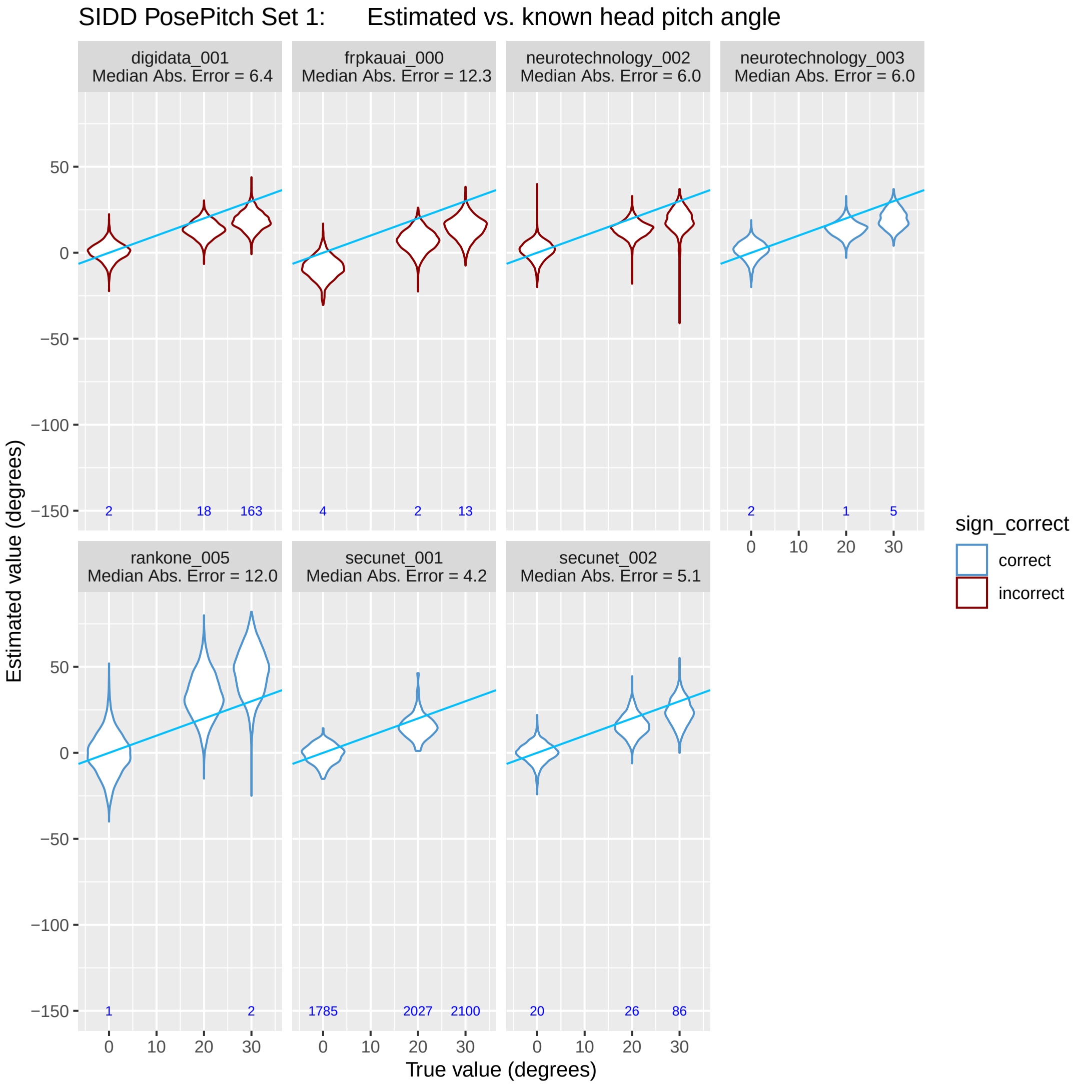


Fig. 5. Estimated vs. known values of pitch angle. For Pitch Set 1, ground truth pitch values are determined by the placement of the camera at the time of capture; the subject remains stationary and frontal. The blue line $(y = x)$ represents perfect performance. The plot shows violins at true pitch values with the tails extending to the minimum and maximum estimated values. The small numbers at $y = -150$ represent the count of faces when the software did not return an estimate; for example, when it did not detect a face. The dark red color-coding indicates that the developer uses the opposite sign convention, and should negate the scores in the next submission.

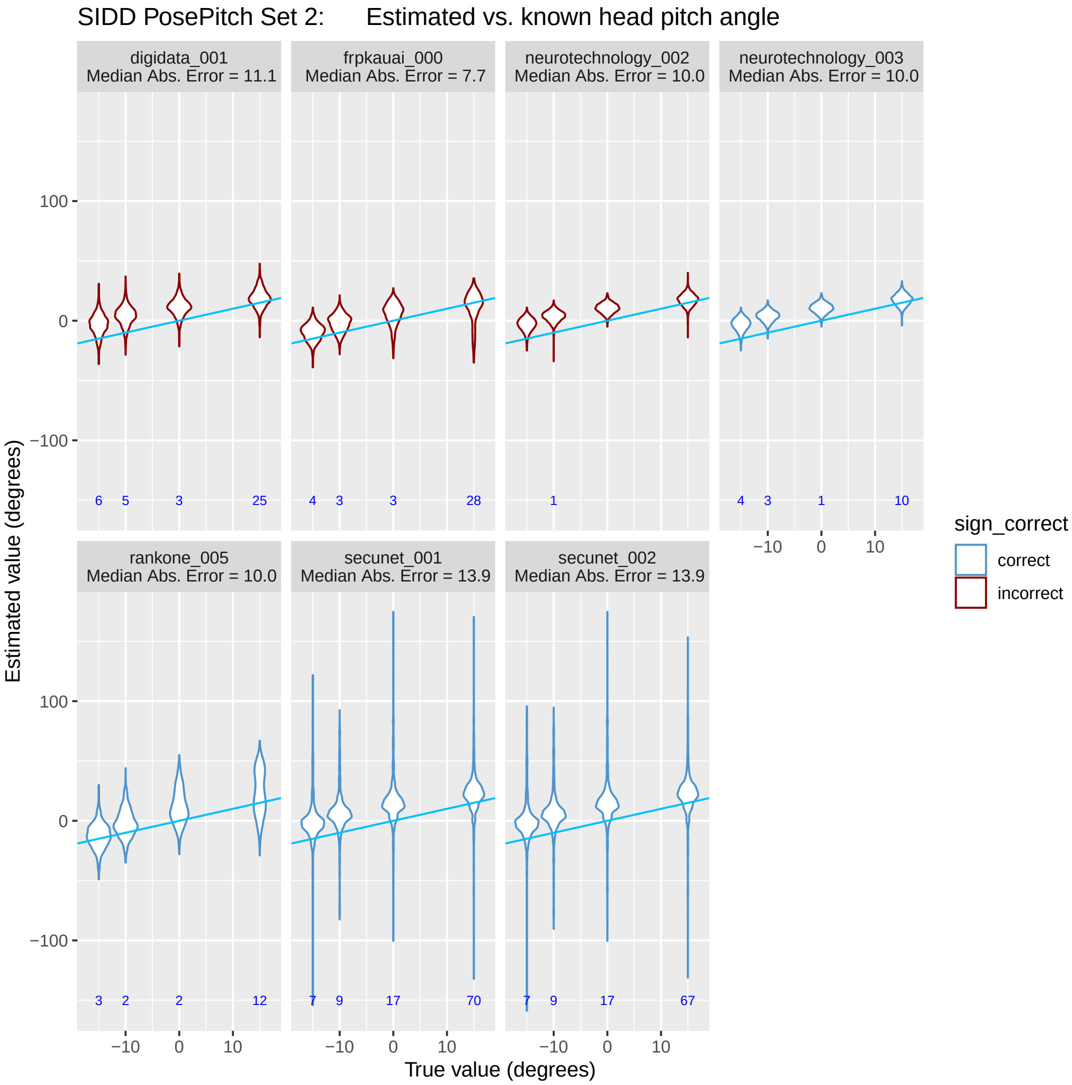


Fig. 6. Estimated vs. known values of pitch angle. For Pitch Set 2, ground truth pitch values are determined by the placement of the camera at the time of capture; the subject remains stationary and frontal. The blue line $(y = x)$ represents perfect performance. The plot shows violins at true pitch values with the tails extending to the minimum and maximum estimated values. The small numbers at $y = -150$ represent the count of faces when the software did not return an estimate; for example, when it did not detect a face. The dark red color-coding indicates that the developer uses the opposite sign convention, and should negate the scores in the next submission.

3.7. Roll Angle

3.7.1. Images Used

The images in the Roll dataset are mugshots that are rotated by a roll angle ranging from -30 to 30 degrees. In particular, we do not include images with a roll angle of 90 degrees. Rotation towards the subject's right shoulder corresponds to a positive roll angle. Rotation towards the subject's left shoulder corresponds to a negative roll angle. This sign convention is consistent with the ISO/IEC 39794-5:2019 standard.

3.7.2. Results for Roll Angle

Table 6 and Figure 7 summarize the performance of the algorithms in our evaluation when estimating roll angle. The Median Absolute Error (MAE), where error is computed as the difference between ground truth and reported value, is shown for each algorithm. Lower MAE is better.

Table 6. SIDD PoseRoll Median Absolute Error

Algorithm	MAE (in degrees)
digidata_001	1.8
frpkauai_000	1.3
neurotechnology_002	2.0
neurotechnology_003	2.0
rankone_005	1.2
secunet_001	1.5
secunet_002	1.5

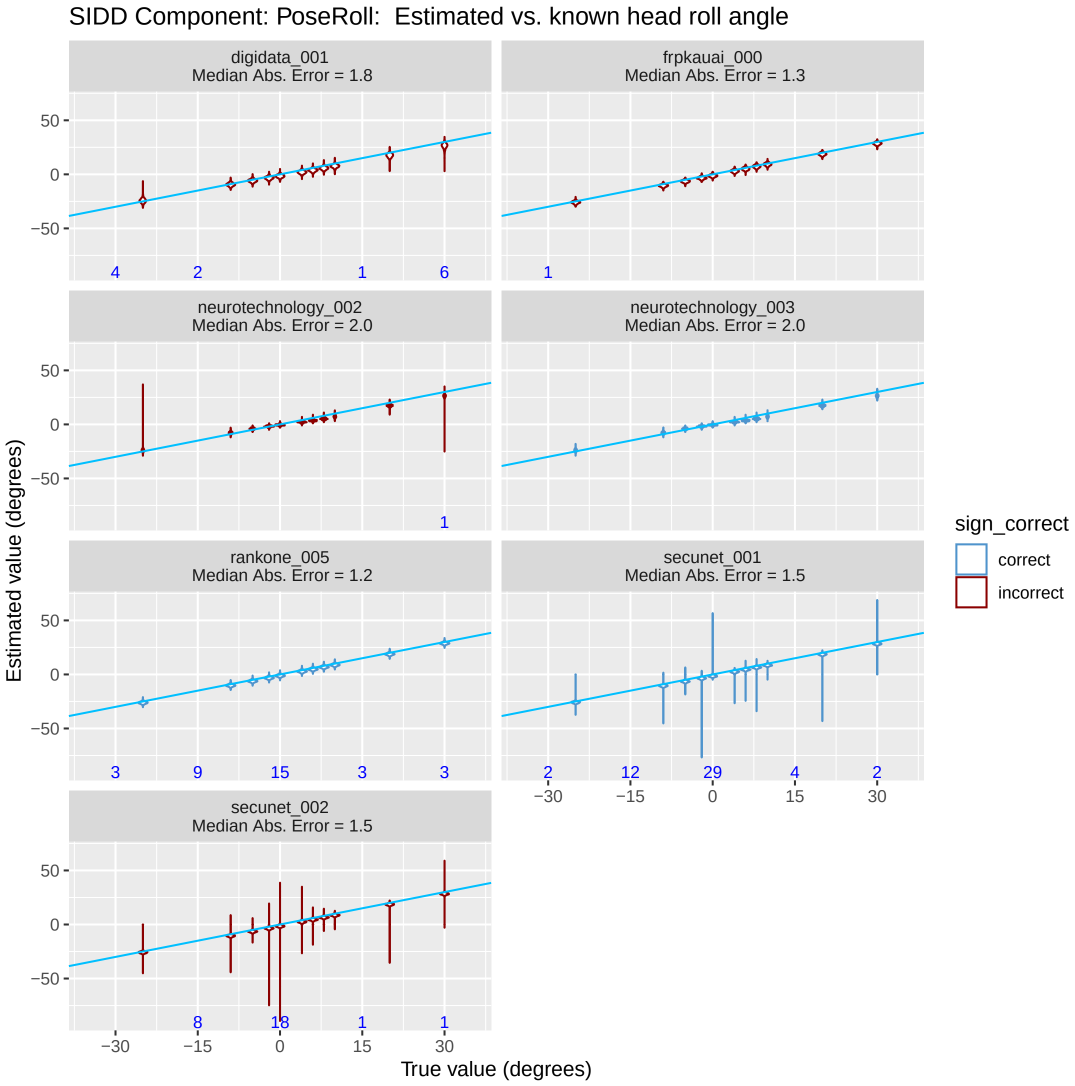


Fig. 7. Estimated vs. known values of roll angle. Ground truth roll values were determined by rotating mugshot images by a known angle. The blue line $(y = x)$ represents perfect performance. The plot shows violins at true roll values with the tails extending to the minimum and maximum estimated values. The small numbers along the horizontal line at -90 represent the count of faces when the software did not return an estimate; for example, when it did not detect a face. The dark red color-coding indicates that the developer uses the opposite sign convention, and should negate their scores in the next submission.

3.8. Eyes Open

3.8.1. Images Used

The images for the Eyes Open test are mugshot images. We calculate the EyesOpen measure by comparing the left and right maximum apertures of the eyes as shown in Fig. 8, taking the minimum of the two values, and dividing the result by the inter-eye distance. This procedure correctly assigns a ground truth value of zero for eyes that are closed.

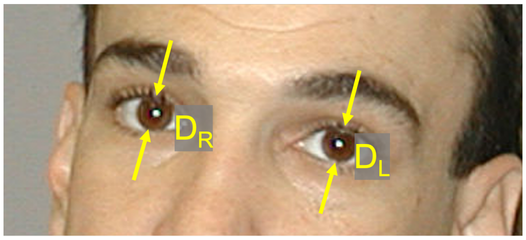


Fig. 8. The EyesOpen measure is computed by comparing the left and right maximum apertures of the eyes, taking the minimum of the two values, and dividing the result by inter-eye distance. Image from NIST Special Database 32, MEDS.

3.8.2. Results for Eyes Open

Table 7 and Figure 9 summarize algorithm performance. The Median Absolute Error (MAE), where error is computed as the difference between ground truth and reported value, is shown for each algorithm. Lower MAE is better.

Table 7. SIDD EyesOpen Median Absolute Error.

Algorithm	MAE (dimensionless)
digidata_001	0.03
frpkauai_000	0.85
neurotechnology_002	0.02
neurotechnology_003	0.02
rankone_005	0.17
secunet_001	0.03
secunet_002	0.03

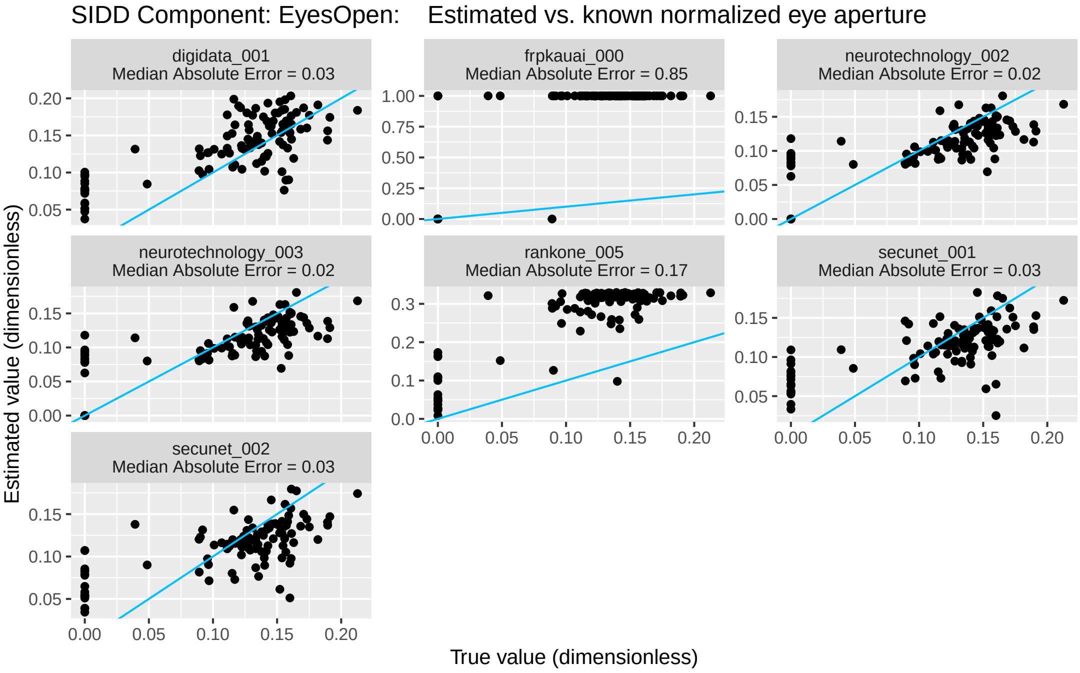


Fig. 9. Estimated vs. known values of the ratio of eye aperture to inter-eye distance. Ground truth preparation is discussed in Section 3.8.1. The blue line $(y = x)$ represents perfect performance. The vertical line of dots at true value zero corresponds to closed eyes. Note that the plots have different y-axis ranges.

3.9. Inter-Eye Distance

3.9.1. Images Used

The images for the Inter-Eye Distance test are from two sets.

1. In the first set, image sizes range from 310 to 1000 pixels in width, and 240 to 1330 pixels in height.
2. In the second set, image sizes range from 720 to 5200 pixels in width, and 1080 to 3500 pixels in height.

In order to determine ground truth, we manually find the eye-centers by determining the two points where eyelids meet for each eye and averaging the two points. The distance between the two eye-centers is used as the ground truth inter-eye distance, as shown in Fig. 10. This procedure is identical to that mandated in ISO/IEC 39794-5:2019 and is effective even when the eyes are closed. Note that subjects may have properties that make detection of eyelid corners challenging. Examples include drooping eyelids, makeup on the eyes, and long eyelashes. As a result, there may be times when ground truth is imperfect.

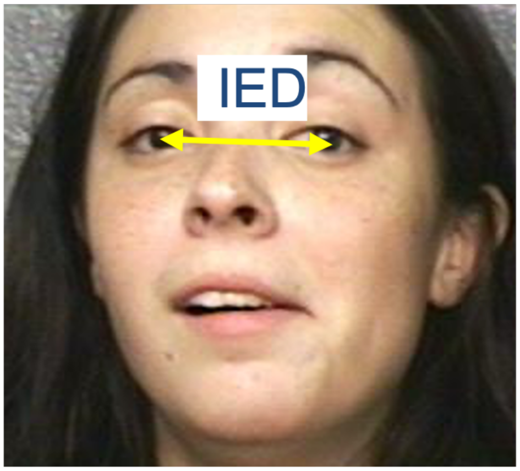


Fig. 10. Inter-eye distance is calculated by averaging the canthi for each eye and taking the distance of the two resulting points. Image from NIST Special Database 32, MEDS.

3.9.2. Results for Inter-Eye Distance

Table 8 and Figure 11 summarize algorithm performance. The Median Absolute Error (MAE), where error is computed as the difference between ground truth and reported value, is shown for each algorithm. Lower MAE is better.

Table 8. SIDD InterEyeDistance Median Absolute Error

Algorithm	Dataset	MAE (px)
digidata_001	IED Set 1	9.5
frpkauai_000	IED Set 1	3.8
neurotechnology_002	IED Set 1	5.2
neurotechnology_003	IED Set 1	5.2
rankone_005	IED Set 1	2.7
secunet_001	IED Set 1	2.0
secunet_002	IED Set 1	2.7
digidata_001	IED Set 2	20.4
frpkauai_000	IED Set 2	20.2
neurotechnology_002	IED Set 2	13.2
neurotechnology_003	IED Set 2	13.2
rankone_005	IED Set 2	11.9
secunet_001	IED Set 2	12.0
secunet_002	IED Set 2	12.5

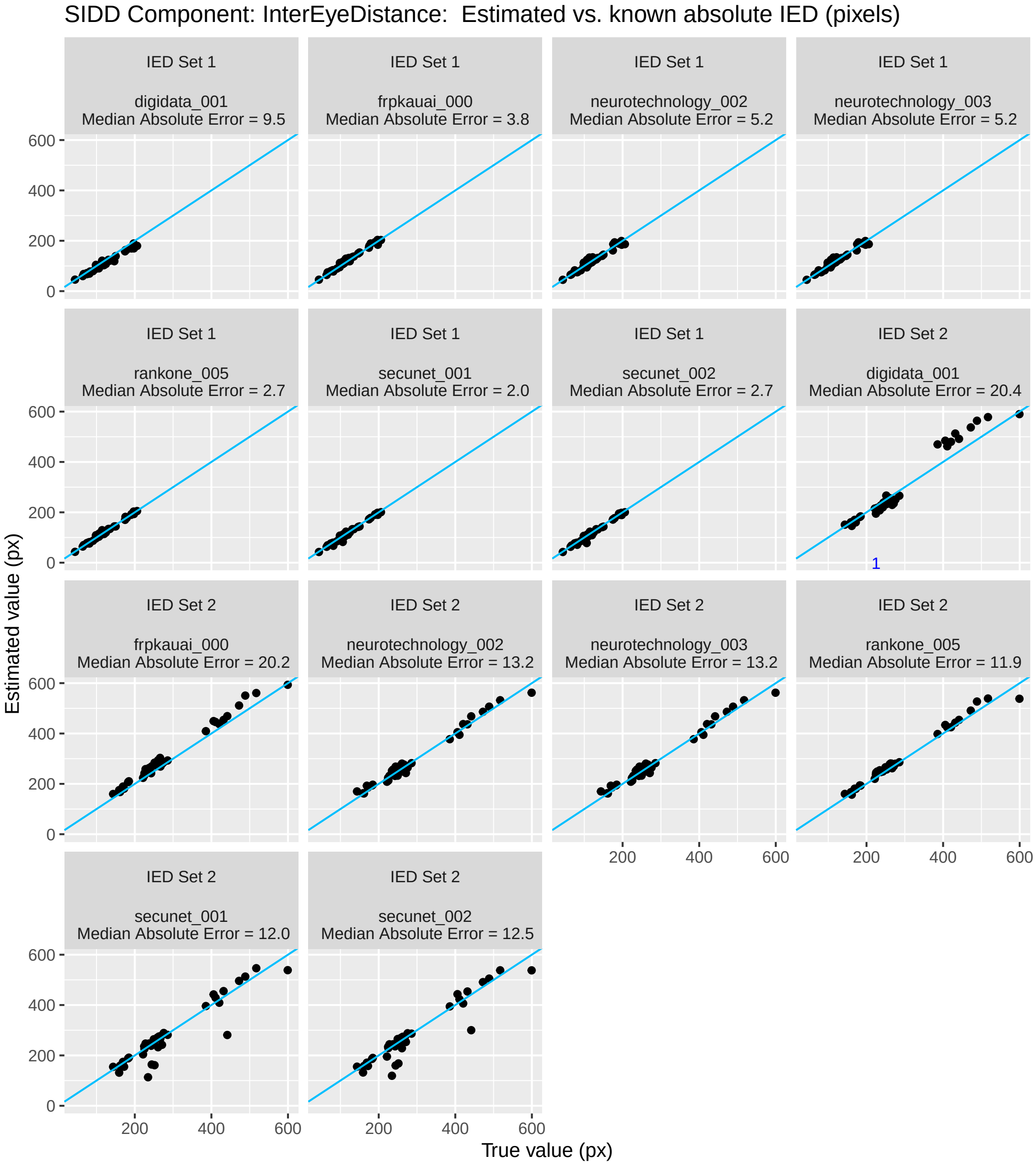


Fig. 11. Estimated vs. known values of inter-eye distance. Ground truth preparation is discussed in Section 3.9.1. The blue line ($y = x$) represents perfect performance. The small numbers at $y = 0$ represent the count of faces when the software did not return an estimate; for example, when it did not detect a face.

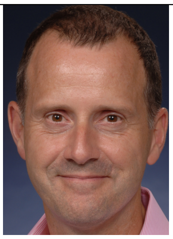
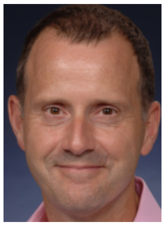
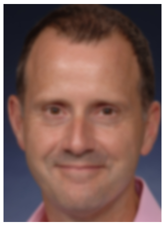
3.10. Resolution

3.10.1. Images Used

The images for the Resolution measure are produced by blurring mugshots. The mugshots are selected to have no significant blur, motion blur, or compression artifacts to begin with. We use the *convert* command from the ImageMagick package with the argument *gaussian-blur*, as illustrated in Table 9. This command convolves each pixel in the input image with a Gaussian kernel. Higher values of the σ parameter, the standard deviation of the Gaussian, correspond to lower resolution.

For the Resolution test, the images range in size from 128 to 3456 pixels in width and 120 to 2719 pixels in height. The inter-eye distance (IED) in the images ranges from approximately 15 to 600 pixels. We use eight values of sigma, ranging from 0 to 7. The highest value of sigma, 7, corresponds to not being able to discern the canthi accurately, but still being able to detect that the eyes are open. Note that the resolution perceived by a reader of this report depends on the handling of the image by LaTeX, the device used to display the image, and other optical factors.

Table 9. Resolution Illustration. Images are used with the permission of the subject.

Standard deviation σ	0	2	5
Result of convert -gaussian-blur 0xσ			

3.10.2. Effect of Gaussian Blur on IED Error

Figure 12 shows a violin plot of the IED error at 8 different values of σ, ranging from 0 to 7. Error is measured with respect to developer-reported IED, so that the error at σ=0 corresponds to the developer measurement of IED for an original, unblurred image.

3.10.3. Results for Resolution

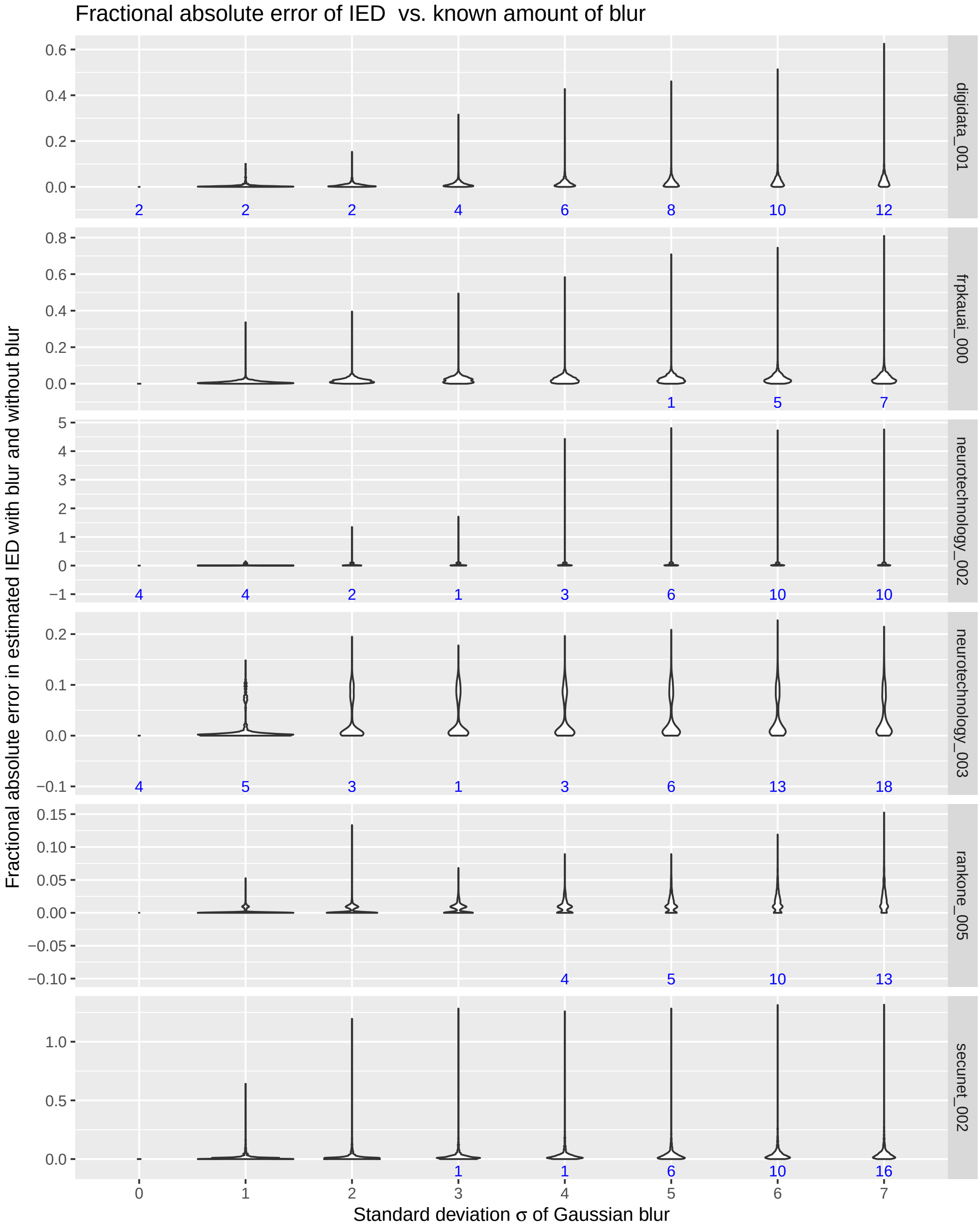


Fig. 12. Fractional absolute value of error in inter-eye distance estimates vs. σ parameter of Gaussian blur. The higher the σ, the more extreme the fractional error in IED estimates. The small blue numbers below the x-axis represent the count of faces when the software did not return an estimate, for instance, when it did not detect a face.

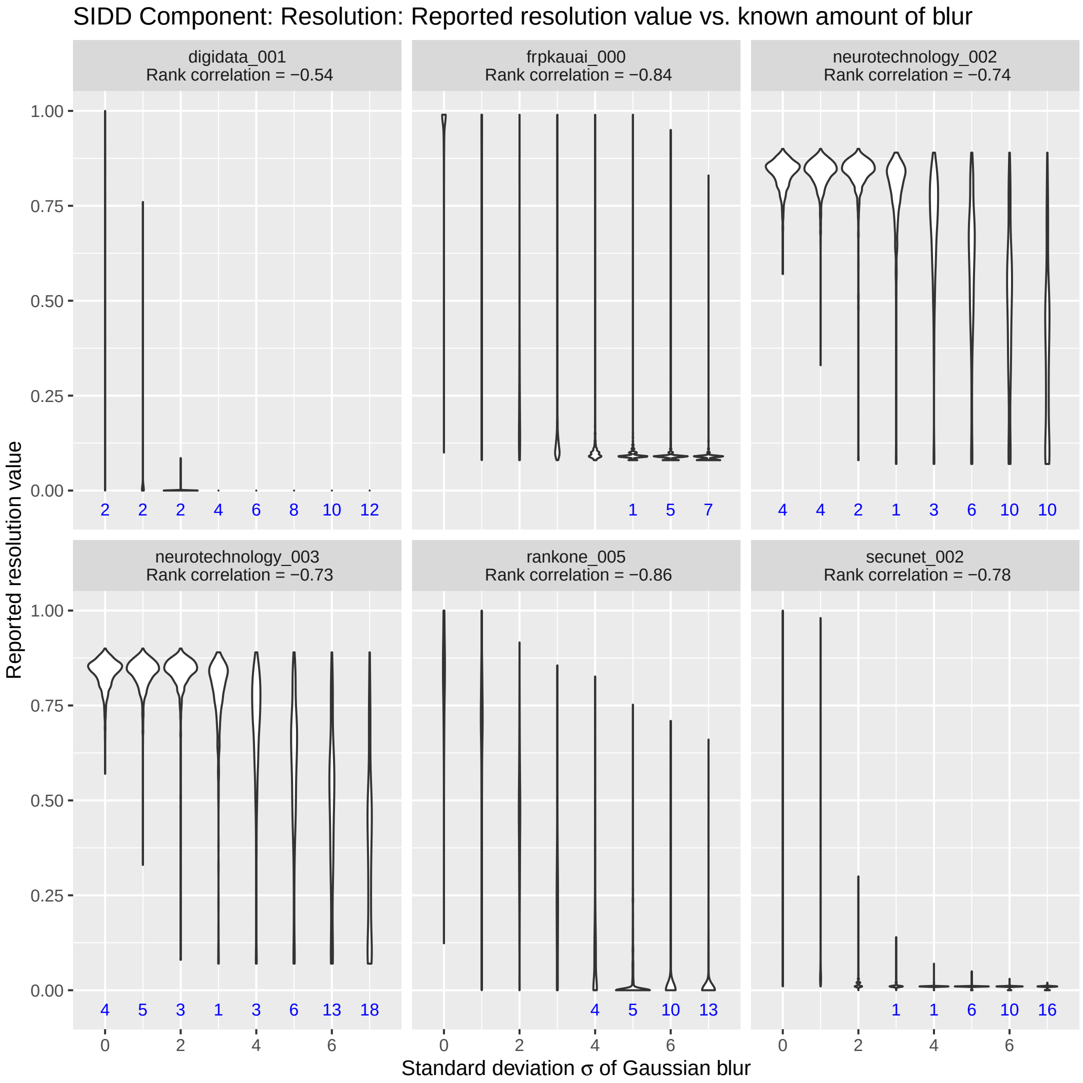


Fig. 13. Distribution of estimated resolution vs. σ parameter of Gaussian blur applied to a set of mugshot images. The higher the σ, the more extreme the blur. The small blue numbers at $y = 0$ represent the count of faces when the software did not return an estimate, for instance, when it did not detect a face. Perfect performance would correspond to monotonically decreasing resolution estimates as σ increases, and a rank correlation value of -1.

3.11. Mouth Open

3.11.1. Images Used

We use mugshot images for the Mouth Open measure. The maximum distance from the bottom of the upper lip to the top of the lower lip is measured, then divided by the inter-eye distance to determine ground truth, as shown in Figure 14. This procedure assigns a ground truth value of zero for mouths that are closed.

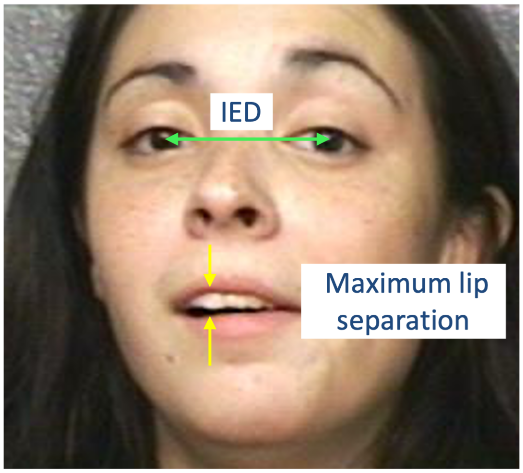


Fig. 14. The MouthOpen measure is computed by taking the maximum distance from the bottom of the upper lip to the top of the lower lip, and dividing the result by inter-eye distance. Image from NIST Special Database 32, MEDS.

3.11.2. Results for Mouth Open

Table 10 and Figure 15 summarize algorithm performance. The Median Absolute Error (MAE), where error is computed as the difference between ground truth and reported value, is shown for each algorithm. Lower MAE is better.

Table 10. SIDD MouthOpen Median Absolute Error.

Algorithm	MAE (dimensionless)
digidata_001	0.03
neurotechnology_002	0.00
neurotechnology_003	0.00
rankone_005	0.03
secunet_001	0.02
secunet_002	0.02

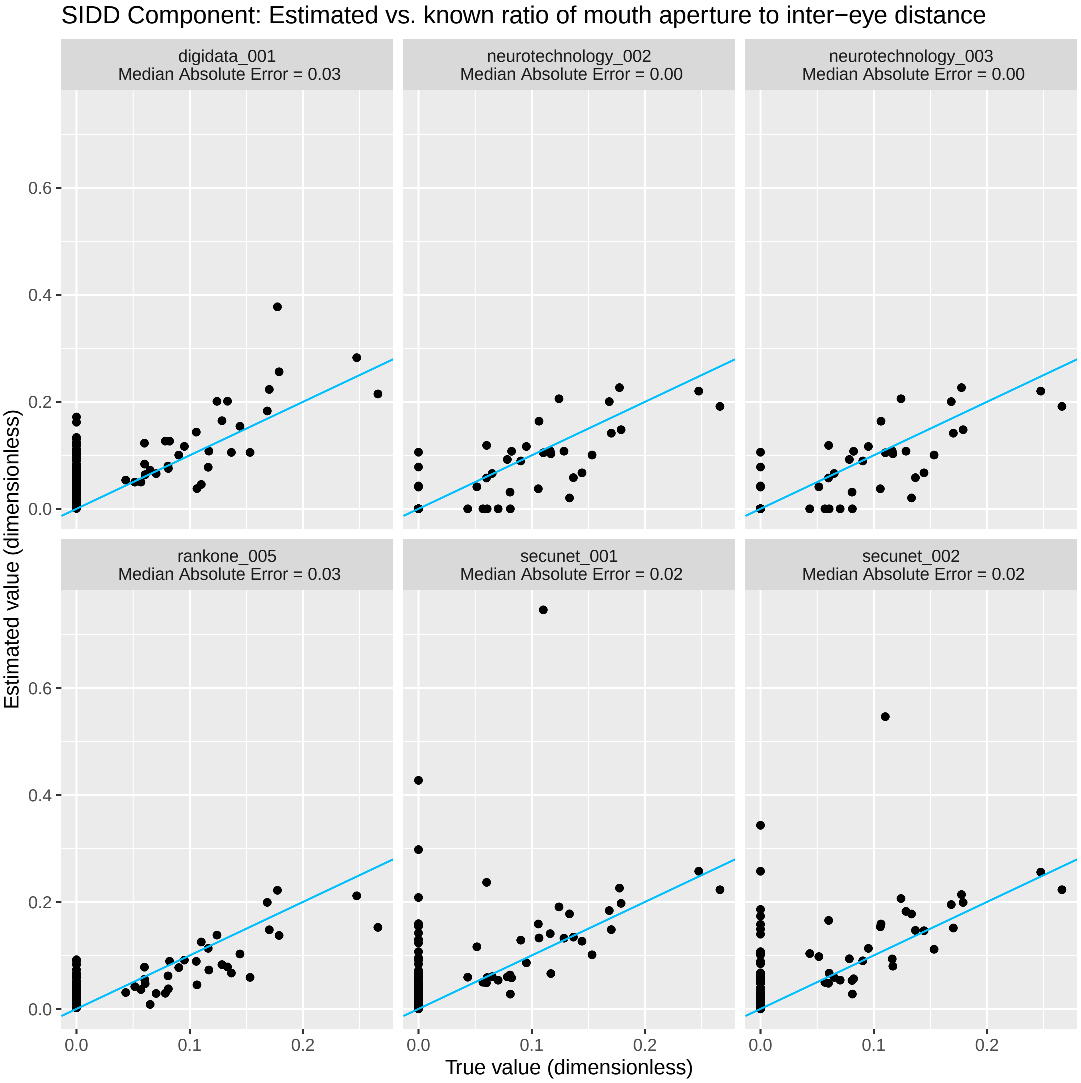


Fig. 15. Estimated vs. known values of the ratio of mouth aperture to inter-eye distance. Ground truth preparation is discussed in Section 3.11.1. The blue line $(y = x)$ represents perfect performance. The vertical line of dots at true value zero corresponds to closed mouths.

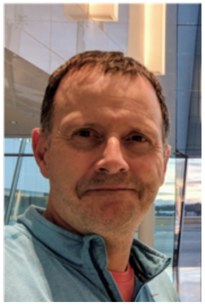
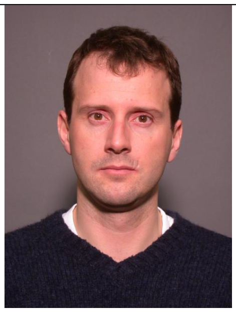
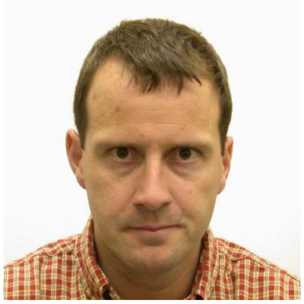
3.12. Background Uniformity

3.12.1. Images Used

We use mugshot images for the Background Uniformity measure. We categorize images into three categories: Uniform, Attempt at Uniform, and Cluttered.

Uniform images have a plain background, with no brick or shadows behind the subject. Images in the Attempt at Uniform category might have a background with concrete or brick texture. Alternatively, they may have shadows behind the subject, but no other significant non-uniformity. We categorize all other images as Cluttered. The images in the Cluttered category include backgrounds containing furniture, walls with writing behind the subject, and significant variation in background color. Examples are in Table 11.

Table 11. Images in order of increasing background uniformity; images are used with the permission of the subject.

Category	Cluttered	Attempt at Uniformity	Uniform
Example			

3.12.2. Results for Background Uniformity

Figure 16 summarizes algorithm performance for background uniformity.

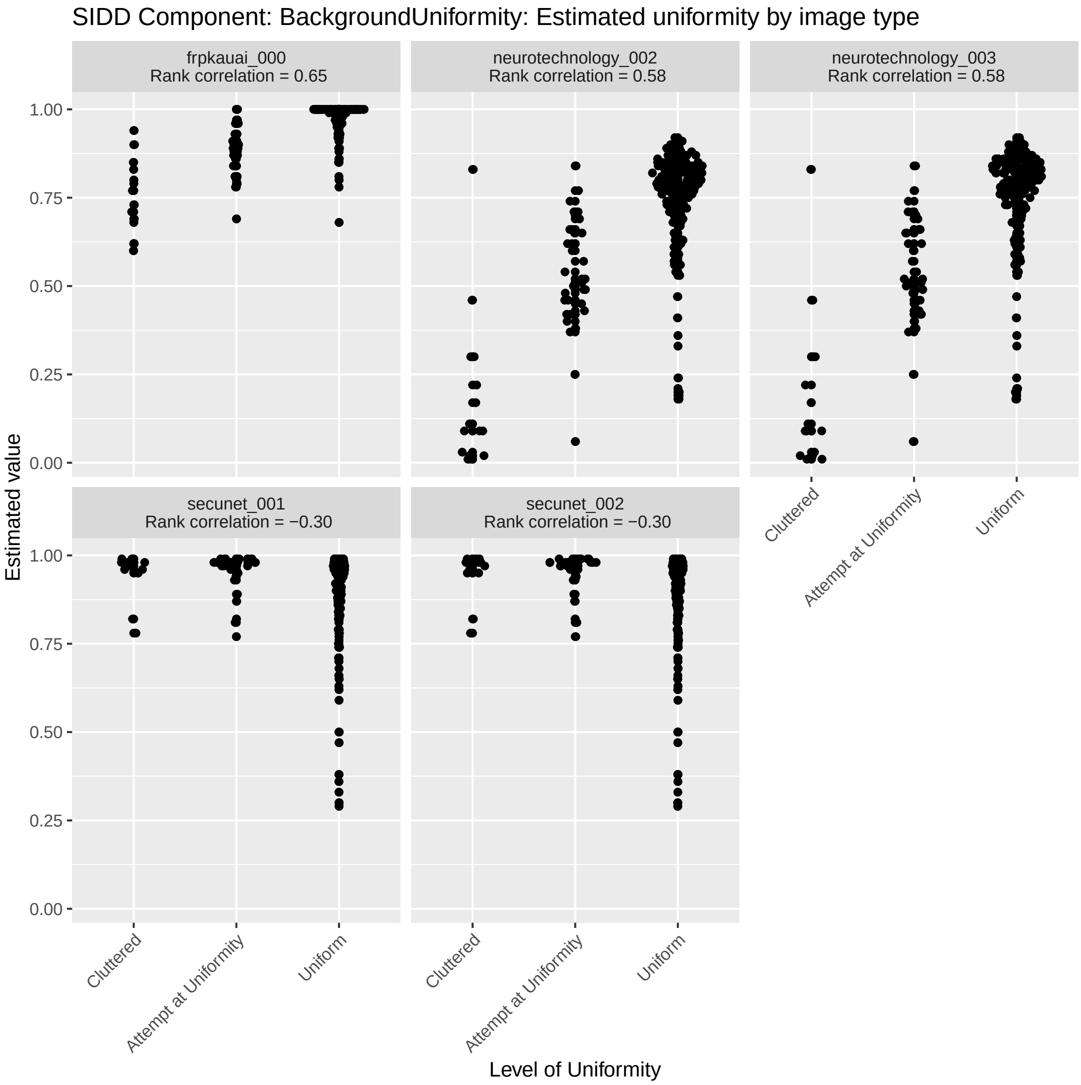


Fig. 16. Estimated degree of background uniformity by category (Cluttered, Attempt at Uniformity, Uniform). Perfect performance corresponds to clusters that shift upward as uniformity increases, and a rank correlation value of 1.

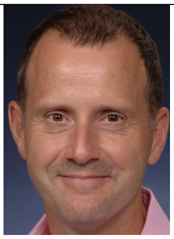
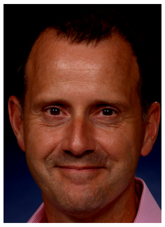
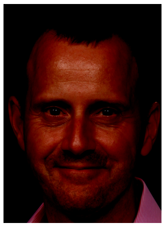
3.13. Underexposure

3.13.1. Images Used

To generate ground truth for the underexposure measure, we start by using mugshot images. We use the *convert* command from the ImageMagick package with argument *brightness-contrast*, as illustrated in Table 17. We use five values of d ranging from 0 to 32. For this measure, more negative values correspond to more underexposure.

Note that the two parameters for brightness and contrast d_1 and d_2 are both inputs, separated by the symbol x. We use $d_1 = -d_2$ to ensure that the two values are inversely proportional and have equal ranges of values.

Fig. 17. Underexposure Illustration. Images are used with the permission of the subject.

Brightness and contrast (d_1, d_2)	(0,0)	(-16,16)	(-32,32)
Result of convert -brightness-contrast d_1xd_2			

3.13.2. Results for Underexposure

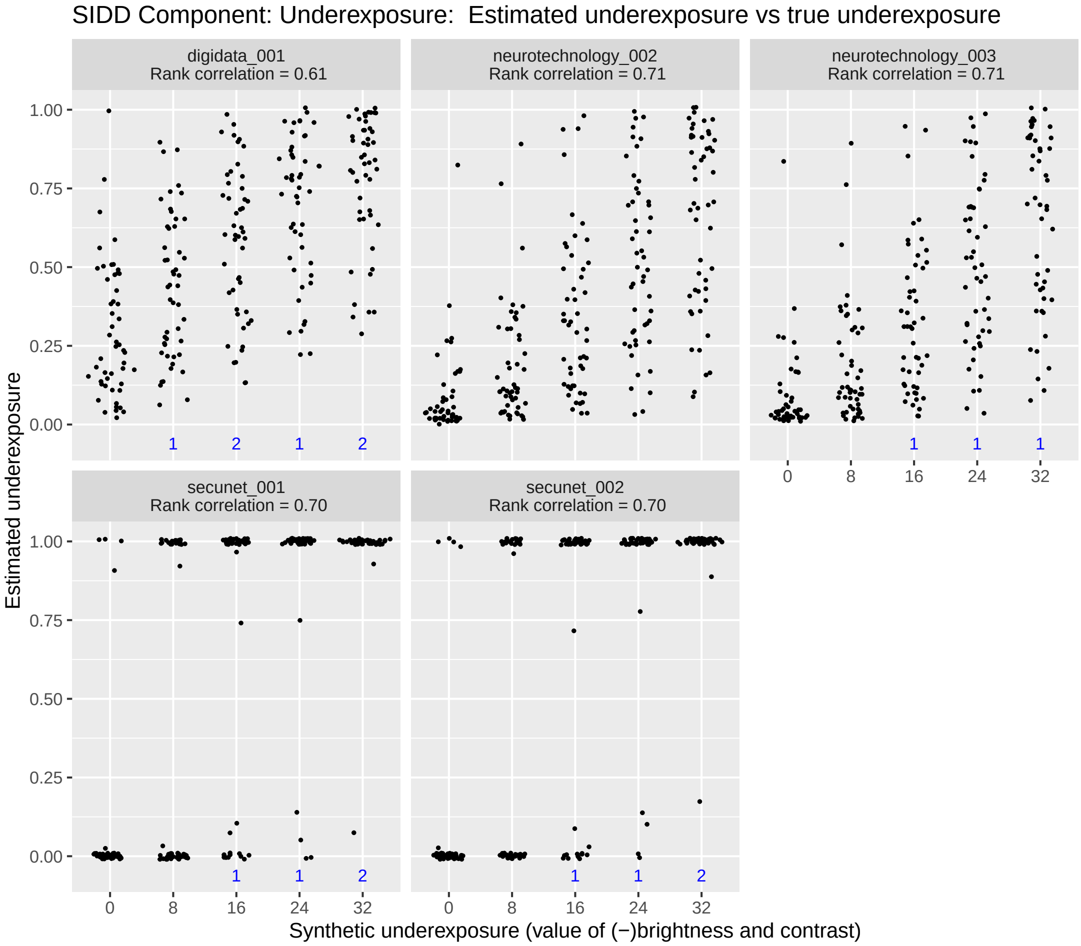


Fig. 18. Distribution of estimated underexposure vs. known underexposure. The x-values represent the contrast and magnitude of decreased brightness of an image. The higher the x value, the more extreme the underexposure. Perfect performance corresponds to clusters shifting upward as x increases, and a rank correlation value of 1. The small numbers along the horizontal line at $y = 0$ represent failures to detect a face.

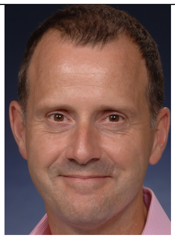
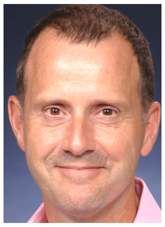
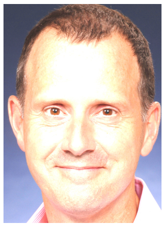
3.14. Overexposure

3.14.1. Images Used

We start with images from mugshot sets for the overexposure measure. We then use the *convert* command from the ImageMagick package with the argument *brightness-contrast*, as illustrated in Table 12. We use five values of d ranging from 0 to 40. For this measure, higher brightness corresponds to more overexposure.

Note that the two parameters for brightness and contrast d_1 and d_2 are both inputs, separated by the symbol x. We use $d_1 = d_2$ to ensure that the two values increase linearly with each other and lie on the same range.

Table 12. Overexposure Illustration. Images are used with the permission of the subject.

Brightness and contrast (d_1, d_2)	(0,0)	(20,20)	(40,40)
Result of convert -brightness-contrast d_1xd_2			

3.14.2. Results for Overexposure

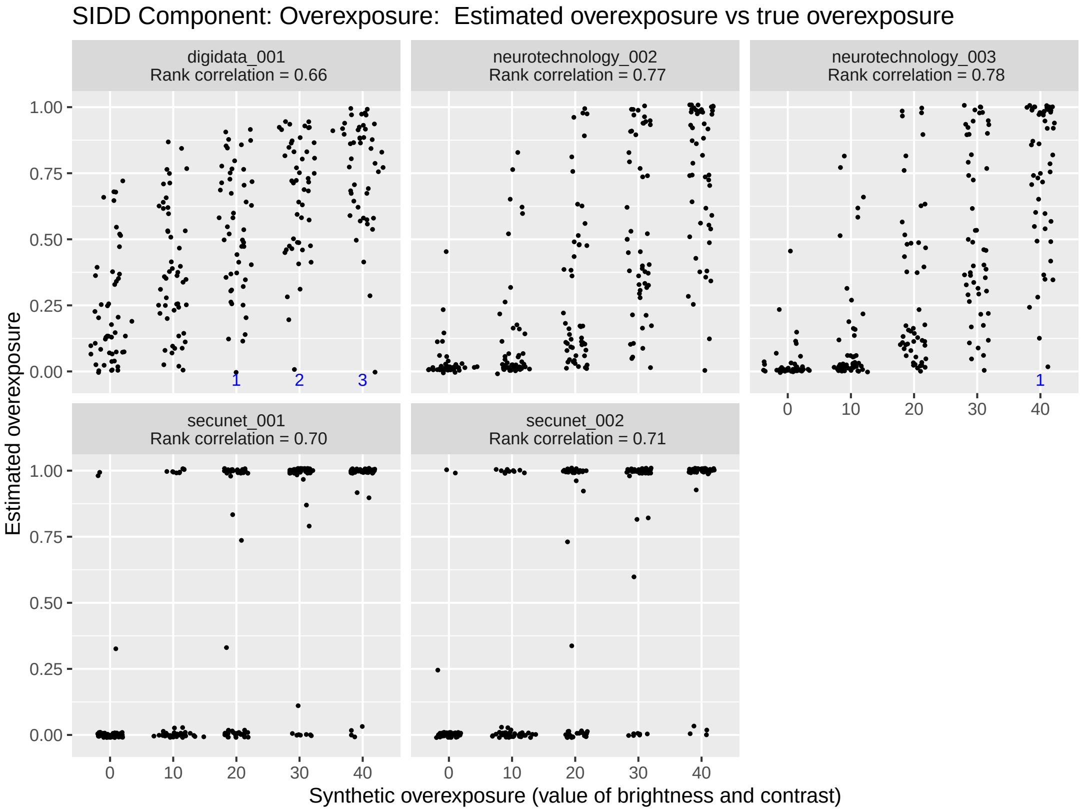


Fig. 19. Distribution of estimated overexposure vs. known overexposure. The x-values represent the contrast and magnitude of brightness applied to an image. The higher the x value, the more extreme the overexposure. Perfect performance corresponds to clusters shifting upward as x increases, and a rank correlation value of 1. The small numbers along the horizontal line at $y = 0$ represent failures to detect a face.

3.15. Eyeglasses Present

3.15.1. Images Used

The images in the Eyeglasses set are mugshot images, in which pose is generally frontal and background is generally uniform. We assign ground truth value of 1 for images in which the subject is wearing eyeglasses (transparent or sunglasses), and 0 otherwise.

3.15.2. Results for Eyeglasses Present

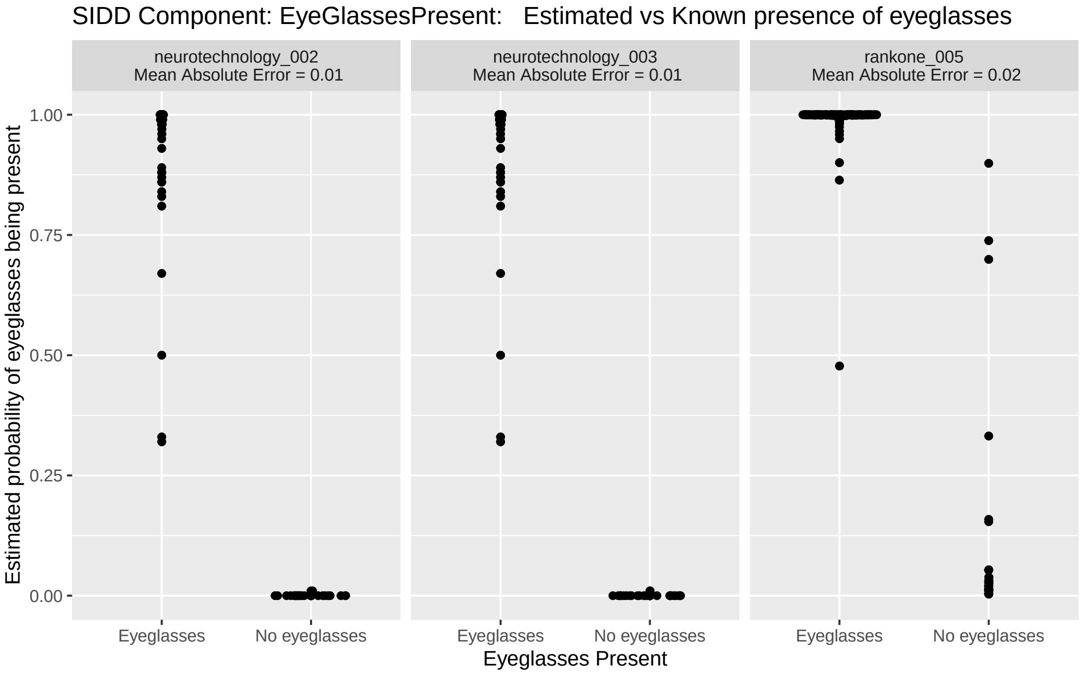


Fig. 20. Estimated vs. known presence of eyeglasses. Perfect performance corresponds to one cluster at 1 (for eyeglasses) and one cluster at 0 (no eyeglasses).

3.16. Sunglasses Present

3.16.1. Images Used

The images in the Sunglasses set are images in a natural setting, including non-frontal poses and non-uniform background. We evaluate submissions on images from three categories: opaque, semi-opaque, and transparent.

3.16.2. Results for Sunglasses Present

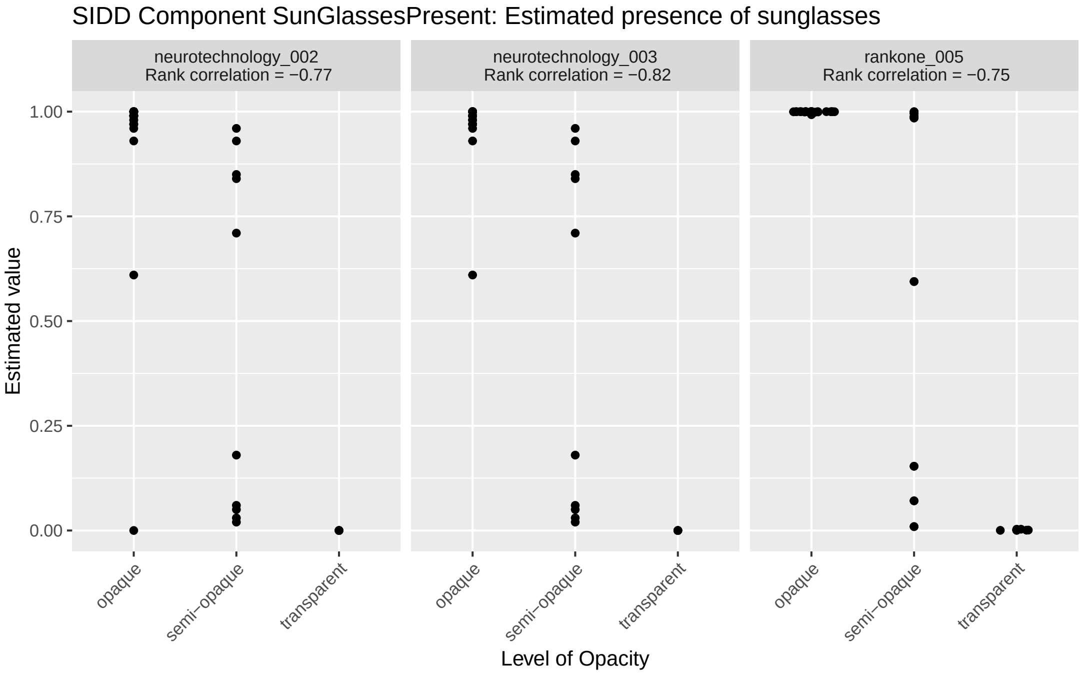


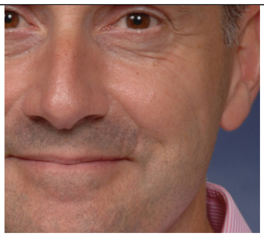
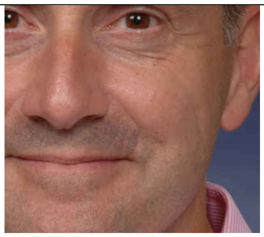
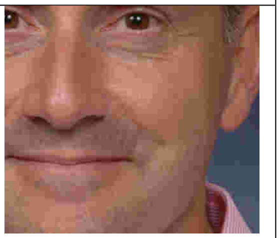
Fig. 21. Estimated vs. known presence of sunglasses. Perfect performance would correspond to monotonically decreasing clusters across the three categories, and a rank correlation value of -1.

3.17. Compression Artifacts

3.17.1. Images Used

We start by using mugshots for the Compression Artifacts set. We then use the imageMagick *convert* function with the argument *-quality* to apply JPEG compression to the original images. Table 13 shows the effect of blur at three values of compression d.

Table 13. Compression Artifacts Illustration. Images are used with the permission of the subject.

Compression parameter d	90	40	10
Result of convert -quality d			

3.17.2. Results for Compression Artifacts

Figure 22 summarizes the performance of the algorithms who have implemented detection of compression artifacts.

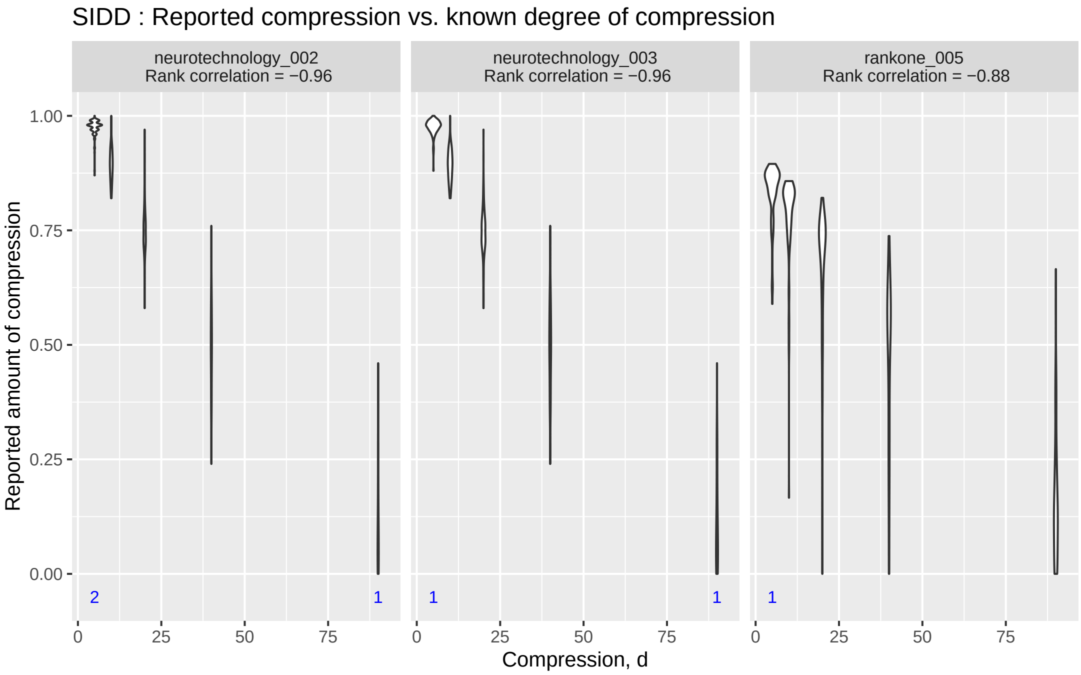


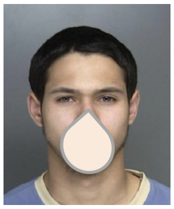
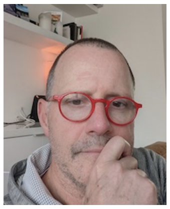
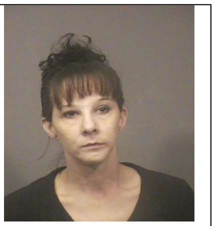
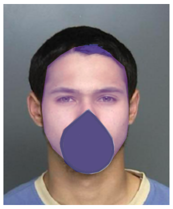
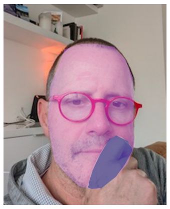
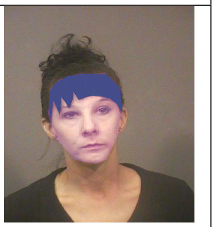
Fig. 22. Distribution of estimated amount of compression vs. d parameter of compression applied to a set of mugshot images. The higher the value of d, the lower the compression value. Perfect performance would correspond to monotonically decreasing estimates as d increases, and a rank correlation value of -1.

3.18. Face Occlusion

3.18.1. Images Used

For the Face Occlusion set, we use images that are generally frontal and well-illuminated. We then compute the occluded area and take the ratio of the occluded area to the total area of the facial region, as described in our API document. Table 14 illustrates values for three example images.

Table 14. Face Occlusion Illustration. The first and third images are from NIST Special Database 32, MEDS; the second image is used with permission of the subject.

Original image			
Image with occluded area shown in blue			
Ratio of occluded area to total area	0.27	0.11	0.36

3.18.2. Results for Face Occlusion

Figure 23 summarizes the performance of the algorithms who have implemented face occlusion.

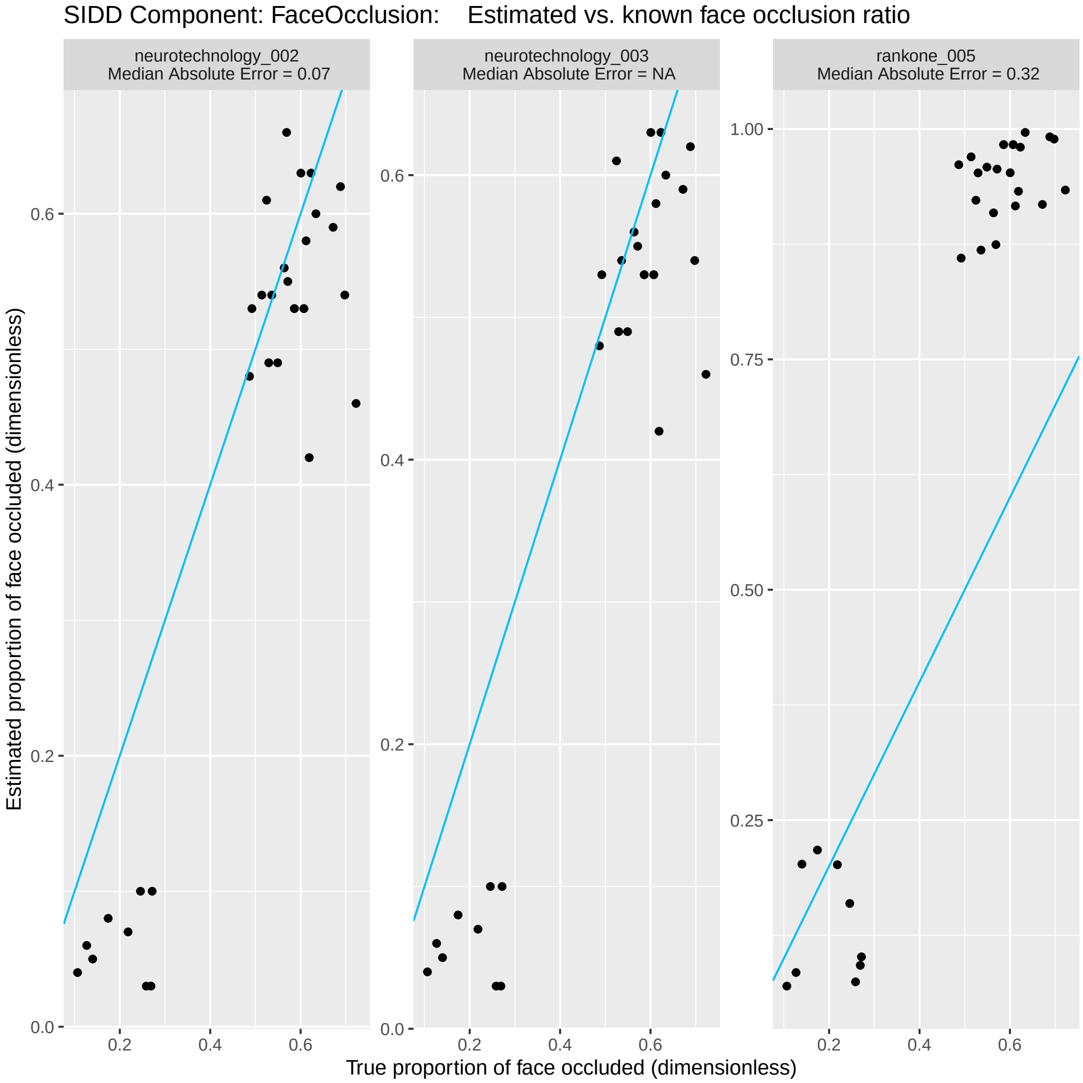


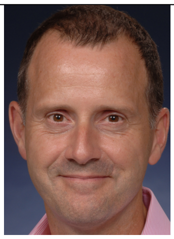
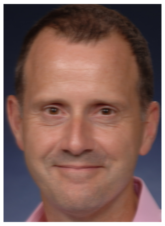
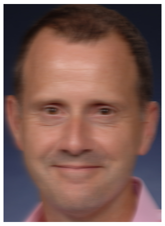
Fig. 23. Estimated vs. known ratio of occluded area to total area of the face. The blue line represents perfect performance. Note that the plots have different y-axis ranges. When points fall significantly above or below the blue line, the developer is likely implementing a different definition of the occluded area; for example, including beards or frames of eyeglasses when they should not be considered occlusion.

3.19. Motion Blur

3.19.1. Images Used

We start by using mugshots for the Motion Blur set. We then use the imageMagick *convert* function with the argument -*motion-blur* to apply motion blur to the original images, which are selected to have no visible blur, motion blur, or compression artifacts to begin with. Table 15 shows the effect of blur at three values of displacement d. For our test we use six values of displacement ranging from 0 to 20.

Table 15. Motion Blur Illustration. Images are used with the permission of the subject.

Displacement d	0	8	16
Result of convert -motion-blur 0xd			

3.19.2. Results for Motion Blur

Figure 24 summarizes the performance of the algorithms who have implemented motion blur.

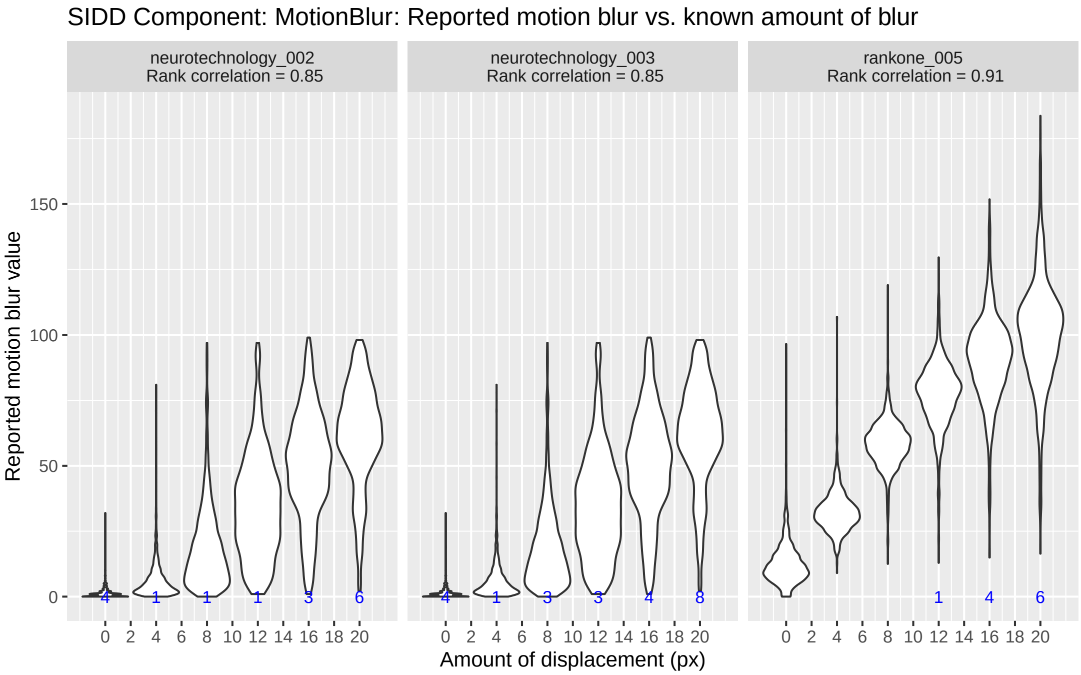


Fig. 24. Distribution of estimated motion blur vs. d parameter of motion blur applied to a set of mugshot images. The higher the value of d, the more extreme the blur. Perfect performance would correspond to monotonically increasing estimates as d increases, and a rank correlation value of 1.

3.20. Distance from Eyes to Edges

3.20.1. Images Used

We use mugshots for the four distance-from-eye-to-edge quality measures. Pose is generally frontal and backgrounds are generally uniform. In order to determine ground truth, we manually find the eye-centers by determining the two points where eyelids meet for each eye and averaging the two points. We then calculate the following:

1. The distance from the left edge to the closest eye-center
2. The distance from the right edge to the closest eye-center
3. The distance from the top edge to the average of the eye-centers
4. The distance from the bottom edge to the average of the eye-centers

These quantities are shown in figure 25.

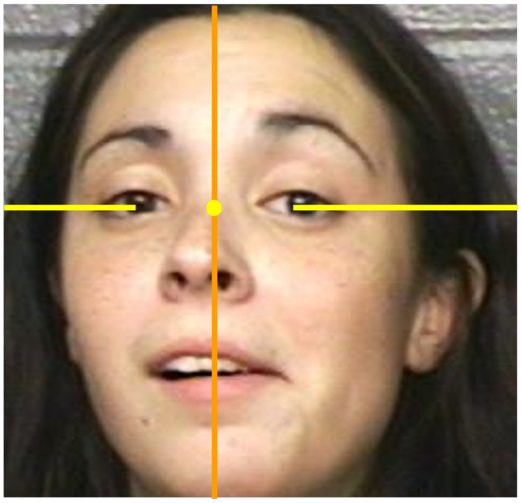


Fig. 25. Image from NIST Special Database 32, MEDS.

This procedure is consistent with that described in the ISO/IEC 29794-5:2024 standard.

3.20.2. Results for Distance from Eyes to Edges

Figure 26, 27, 28, and 29 summarize algorithm performance. The Median Absolute Error (MAE), where error is computed as the difference between ground truth and reported value, is shown for each algorithm. Lower MAE is better.

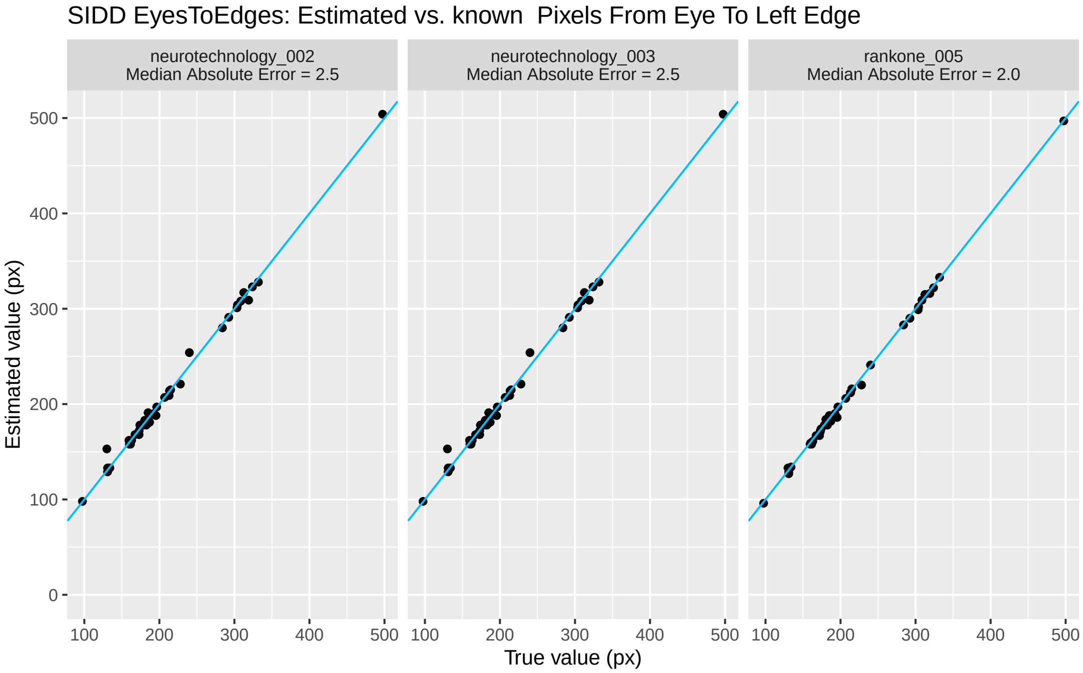


Fig. 26. Estimated vs. known pixels from left edge to the closest eye center. The blue line represents perfect performance.

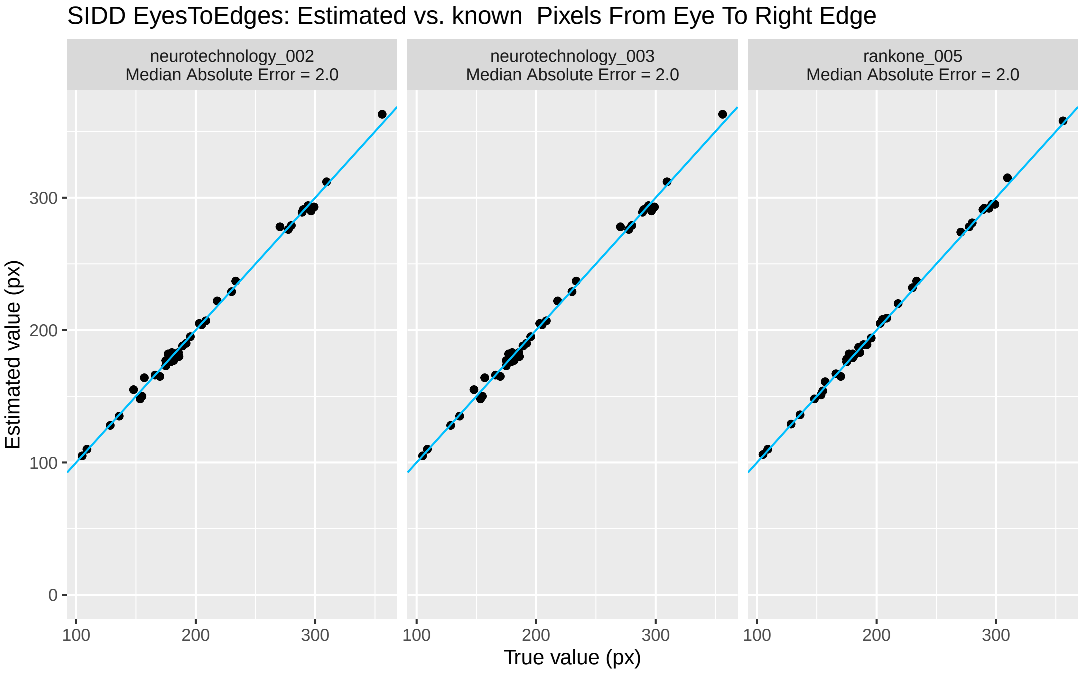


Fig. 27. Estimated vs. known pixels from right edge to the closest eye center. The blue line represents perfect performance.

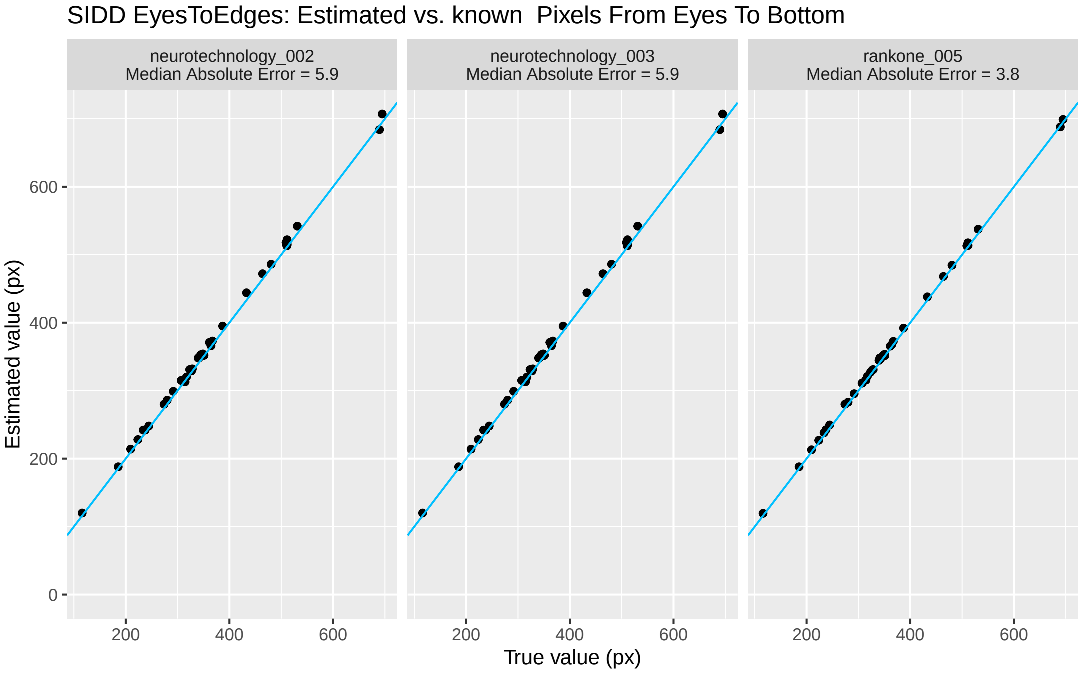


Fig. 28. Estimated vs. known pixels from center of eyes to the bottom of the image. The blue line represents perfect performance.

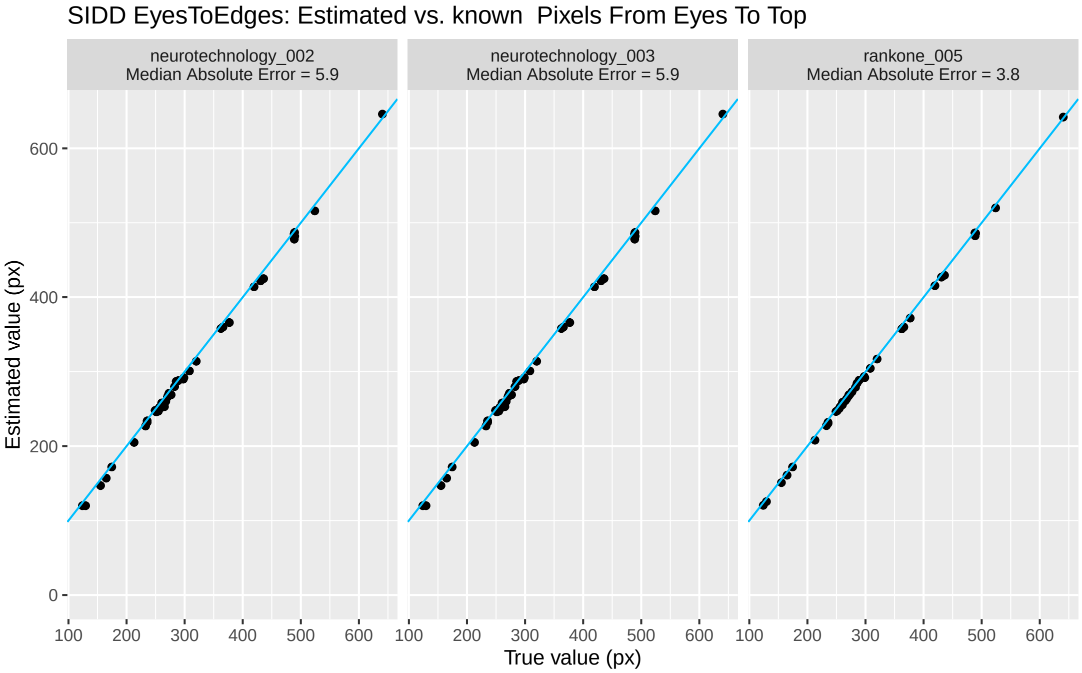


Fig. 29. Estimated vs. known pixels from center of eyes to the top of the image. The blue line represents perfect performance.

3.21. Unified Quality Score

3.21.1. Images Used

Similarity scores are generated from mated comparison of high quality visa-like application photos with medium quality airport arrival webcam photos. Quality is computed only on the webcam photos.

3.21.2. Results for Unified Quality Score

Figure 30 shows false non-match rate (FNMR) gains as a function of the fraction of lowest quality images discarded, for four initial FNMR values: 0.5%, 1%, 2% and 5%.

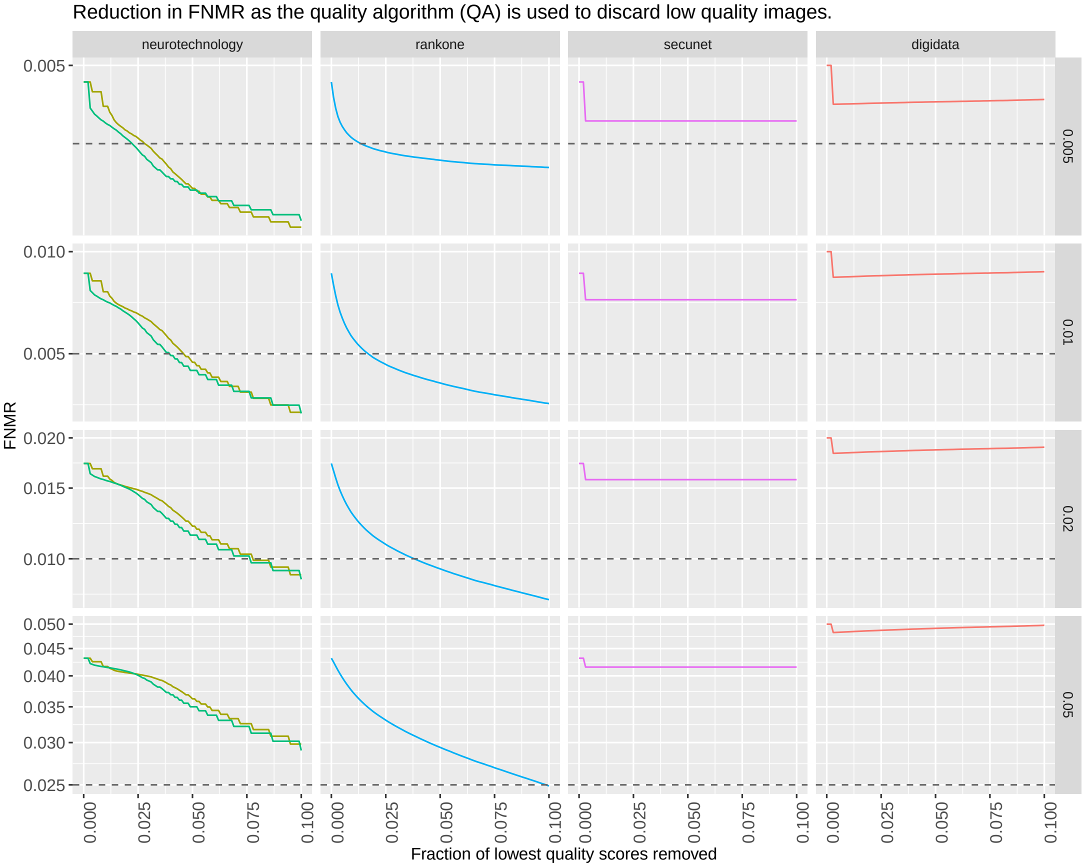


Fig. 30. Reduction in FNMR as the quality algorithm (QA) is used to discard low quality images. Ground truth for quality is set as the false negatives from 22 of the more accurate recognition algorithms, one per developer. The comparison algorithm's thresholds are set to one of four values corresponding to FNMR = 0.005, 0.01, 0.02, or 0.05 given in the grey row strips. Similarity scores are from mated comparison of high quality visa-like application photos with medium quality airport arrival webcam photos. Quality is computed only on the webcam photos. The dotted line gives either half the initial FNMR, or the lowest observed value. A steeply declining curve connotes a better QA.