Report of the Ad Hoc Group on Demographic Variability of Face Image Quality Measures

Christoph Busch[1*], Andre Doersch[1], Pierre Gacon[1], Marcel Ginzler[1], Patrick Grother[1], Rudolf Haraksim[1], Daniel Hartung[1], Olaf Henniger[1], John Howard[1], Wassim Kabbani[1], C.J. Lee[1], Johannes Merkle[1], Lisa Mugnano[1], Torsten Schlett[1], Kerry Shannon[1], Yevgeniy Sirotin[1], Anna Stratman[1], Benjamin Tams[1], Joyce Yang[1]

[1*]ISO/IEC JTC1 SC37 WG3 .

*Corresponding author(s). E-mail(s): christoph.busch@h-da.de;

Abstract

This report addresses the challenge of demographic variability of biometric recognition systems, which are based on face image analysis and which are incorporating biometric sample quality assessment algorithms. When dealing with operational systems, the quality of captured face images is relevant as it will impact the recognition accuracy. Thus, it is required to measure the utility of a face sample with a quality score but also with complementary measures that can provide actionable feedback. Acceptability of biometric systems requires fairness of biometric algorithms and artificial neural networks that are used. It is important to determine if face recognition systems are/are not biased towards a specific demographic group. In order to investigate this challenge SC37 WG3 has started in July 2024 an Ad Hoc group on demographic variability of face image quality measures. This is the first report of the groups' work from July to December 2024.
Disclaimer-01: It is desirable to investigate the demographic variability for sample quality assessement algorithms for fingerprint images and other. However this report is limited to face images.
Disclaimer-02: For the sake of providing a self-contained document, we included textual components from ISO/IEC standards [1–3] that we have developed and papers or reports [4–6] which we have published recently.

Keywords: Biometric face recognition systems, Sample quality, Biometric fairness

1 Introduction

Face recognition today is widely adopted and has reached high significance in a variety of applications, ranging from authentication with smart personal devices (e.g. mobile phones), over access control (e.g. border crossing) to forensic applications (e.g. video surveillance), which all constitute relevant operational systems.

For the face capture process it is relevant to fulfill the quality requirements for enrolment samples and recognition samples: the capture subject should frontally face the capture device to ensure a frontal pose, neutral facial expression and appropriate lighting conditions. ISO/IEC 29794-5 [2] describes uses cases, which require quality assessment:

- **UC1**: Collection of reference samples for ID documents. The face image will be stored on a document, used for example for a maximum of 10 years and should support human examination.
- **UC2**: System enrolment, current or later creation of a reference, delayed recognition. Acquisition of face images where quality should be high enough to ensure later usage and interoperability.
- **UC3**: Collection of probe samples for instantaneous recognition. Single use face image with instantaneous response.

Requirements on face image quality, which are relevant to these use cases are formulated in ISO/IEC 39794-5:2019 [7] for UC1 and in ISO/IEC 19794-5:2011 [8].

Biometric performance is addressing the recognition accuracy in terms of low error rates for false positive of false negative errors [9]. In order to reach a good recognition accuracy the quality of biometric samples plays an important role. Only when both reference and probe samples are of good quality a reliable comparison score can be achieved.

However a remaining challenge is that biometric algorithms shall treat different demographic groups in a fair manner, meaning with the same recognition accuracy and equal chances to fulfill low quality score thresholds. Addressing this challenge is fundamental to reach wide acceptability of biometrics in society.

To address this challenge, we recommend re-examining each individual quality measure algorithm for potential bias. Specially crafted test data are required for this purpose. More precisely, test data has to cover relevant demographic variables including gender, age, ethnicity, among others. An innovative approach would be the use of synthetically generated data, which offer the advantage of analyzing quality-related defects in an isolated way with homogeneous quality across all demographic groups. In contrast, the conventional approach of creating composite databases from various sources is less suitable here - but all that can be done in this version of the report. Differences in compression, cameras used, or other factors can lead to unintended distortions. Synthetic data, on the other hand, would enable controlled homogeneity and increase the reliability of the results.

The WG 3 Ad Hoc group was tasked to investigate the variation of the unified quality score and of other quality measures of interest for the demographic variables (e.g. age, gender and skin color). This WG 3 Ad Hoc group aims to collect from

operators data regarding the distribution of quality measures with respect to the three use cases defined in ISO/IEC 29794-5.

The output of the Ad Hoc group is this report containing also recommendations for further actions.

2 Face image quality

For two dimensional (2D) face recognition, capture requirements have been formulated in the international standard ISO/IEC 39794-5 [7], including for example a decent resolution [10], a full frontal perspective, good contrast, and good lighting. Furthermore certain acquisition criteria such as a neutral facial expression or the precondition that the face region and specifically the landmarks shall not be covered by hair, and the absence of (reflective) glasses or headgear should be met. If compliance of a face image with these requirements is not fulfilled, then the biometric system may recognize the capture subject only with low probability. Not very often the pose (i.e. perspective) and the expression of the face is fully identical in the reference and in the presented probe sample. In essence the drawback of the 2D approach is: the biometric performance is sensitive to pose variations, illumination changes, sensor conditions, and other disturbance factors that degrade the image quality.

It can be assumed that a biometric comparison algorithm delivers good and reliable results when high-quality images are presented and, conversely, delivers worse results when low-quality images are presented. Recently strong innovation is observable for face image quality assessment. One of the driving factors is the launch of the European Entry Exit System (EES)[11] which requires that the EU member states will conduct the biometric enrolment at border control points in accordance with Implementing Decision 2019/329 [12].

Capturing high-quality biometric samples still remains a difficult task. Examples of factors that have a negative impact on the quality of a face image can be seen in Figure 1. To this end, great efforts have been placed into developing quality assessment algorithms for various biometric characteristics to estimate the quality of a captured biometric sample and ensure that its quality is sufficient.

An overview of methods to assess face image quality was recently given in [13]. These methods focus on unified quality scoring approaches that describe the utility of an image for face recognition. The algorithms should have predictive power, meaning that a low quality score indicates a low comparison score to be expected when that image is used in a biometric comparison. Such low score should prevent the face image to be inserted into the EES enrolment database. But also complementary measures are needed that allow actionable feedback to the capture subject such as the correctness of the pose or information to the biometric attendant such as the sharpness of the face image (among many others). Requirements for a face image to be compliant to a canonical face image definition are expressed in the Biometric Data Interchange Standard ISO/IEC 19794-5:2011 as *Frontal image type* [8] and in the more recent

Fig. 1: Various examples of face image defects (i.e. factors) of a captured sample that negatively impact the recognition performance. As a result, the images shown are not compliant with requirements formulated in ISO/IEC 39794-5 [7]. Facial images taken from [7].

Extensible Biometric Data Interchange Standard ISO/IEC 39794-5 in Annex D.1 [7][1] following the ICAO requirements for reference facial images for MRTDs [14].

The prediction capability of a unified quality score is determined with error versus discard characteristic curve (EDC)[1] based on the false non match rate (FNMR)[9] as an expression of recognition performance (i.e. false negative outcomes). To examine the full impact of discarding low quality samples on performance, false non-match EDC and false-match EDC should be explored together [1]. The EDC can illustrate how quickly the FNMR will decrease, when poor quality samples are discarded from the dataset in a step-wise manner. This is illustrated in the example in Figure 2, where for a chosen discard fraction the FNMR decreases faster for the MagFace algorithm [15] (the green line) as compared to alternativealgorithms indicating for MagFace a better prediction of the biometric recognition performance[2]. For this analysis it is important to demonstrate that a unified quality scoring method can generalise over many recognition algorithms [16, 17].

Soon a standardisation process for a unified quality algorithm and complementary quality measures (i.e. actionable feedback) is about to be completed with ISO/IEC 29794-5 [2]. The quality score is a holistic measure for the entire sample, which is predictive of recognition performance and is an integer number in the range 0 to 100 (with higher being better). Along with the standard ISO/IEC 29794-5 the Open Source Face Image Quality (OFIQ) project [6] does provide an open-source reference implementation of standardised algorithms, which was released also in 2024. This open source software can be deployed in commercial and governmental applications[3]. The MagFace algorithm [15] was selected for the unified quality scoring, as it showed the

[1] According to ICAO TAG/TRIP/4 decision from October 2023, passport inspection system must be able to handle ISO/IEC 39794-5 face image data by 2026-01-01

[2] The FNMR in Figure 2 is computed in the NIST FATE SIDD evaluation

[3] For more information on OFIQ visit the BSI website: https://bsi.bund.de/dok/OFIQ-e

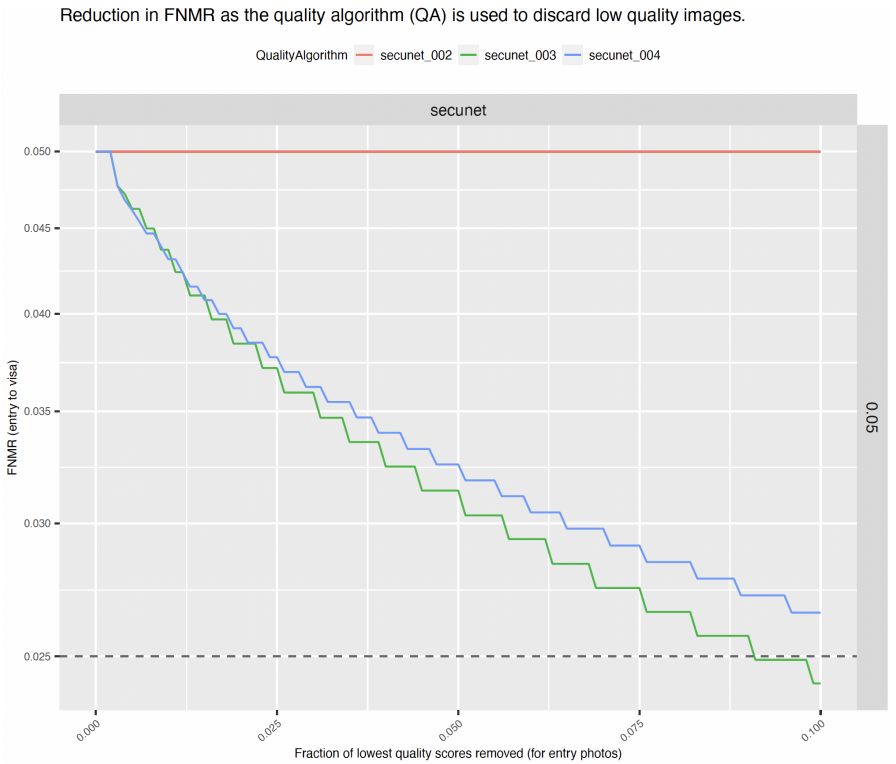

Fig. 2: Reduction in False Non-Match Rate (FNMR) as a function of the fraction of lowest Unified Quality Score images discarded, for an initial FNMR value of 5 percent. Mated comparison scores are from comparison of high quality visa-like application photos with medium quality airport arrival webcam photos. Quality is computed only on the webcam photos. A steeply declining curve connotes a better QA. FNMR decreases faster for the OFIQ-UQS (green) algorithm as compared to the alternative algorithms..

best EDC curve in the generalisation over the 15 best performing face recognition systems in the Face Analysis Technology Evaluation (FATE) Part 11: Face Image Quality Vector Assessment - Specific Image Defect Detection [17].

For the optimisation of the capture process and the involved individuals, namely the capture subject and the biometric attendant, actionable feedback should be provided. Quality components (e.g. the pose angle) are assessing properties of the biometric sample and the compliance with the requirements for a canonical face image (e.g. frontal perspective to the capture device with zero pose angle). Beyond subject related measures, capture device related measures are also of interest, primarily for the capture system set-up and calibration. Here the standard provides algorithms to assess the sharpness / focus of the camera. Component measures have also been included in ISO/IEC 29794-5 [2] and its reference implementation Open Source Face Image Quality (OFIQ) [6]

3 Biometric fairness overview

The successful deployment of biometric systems requires good acceptability in the target population, respectively in the society, when it comes to public biometric system operations. Acceptability in turn requires on the one hand that the interaction of individuals with capture devices is considered as convenient, meaning a good usability of the interaction scheme. Acceptability on the other hand also requires that data subjects have confidence that they are treated in a fair manner by the biometric algorithm. Fairness is specifically expected for biometric recognition algorithms. The NISTIR 8280 [18] investigated to which extent the biometric performance[4] for face recognition systems shows a differential performance[5], meaning a difference in the mated and non-mated comparison score distributions. Such investigation is specifically of interest for categorical demographic variables[6], such as the gender categories *male*, *female* or *neutral*. Another interest is the differential performance related to continuous demographic variables[7], such as the skin color of an individual. While skin color has in the past been considered as categorical classes (e.g. the Fitzpatrick Skin Tpyes (FST) [19]) the recent literature considers as alternative the Monk Skin Tone Scale (MST) constituting a categorical variable that is more appropriate [20, 21] and for which test data exists [22].

It is important to validate prior to deployment that a face recognition system is not biased towards a specific demographic group. An overview of the effect of algorithmic bias in biometric systems and a survey on the recent literature is given in [23]. The reasons for bias are manifold and range from unbalanced training datasets to systematic effects in the training procedures [24–26].

On the path to reach fair biometric systems, a testing methodology is needed. Recent proposals for fairness measures [25, 27] have been included as testing methodology in the International Standard ISO/IEC 19795-10 [3]. However the challenge remains open, as testing methodology for quality measures that can ensure a fair sample quality assessment process are still in their infancy [5]. This report is elaborating in Section 7 on concepts that have been proposed so far.

4 Demographic variables of interest

The International Standard ISO/IEC 19795-10 [9] has introduced the following demographic variables that are of interest for this report:

- **Gender**: is defined as the classification of individuals as male, female or additional categories based on social, cultural or behavioural qualities. An individual's gender identity can consist of multiple, distinct categories. An individual's gender can also

[4] Following the International Standard ISO/IEC 19795-1 [9] biometric performance is reported in terms of false match rate (FMR) and false non-match rate (FNMR) for verification systems and in terms of false positive identification rate (FPIR) and false negative identification rate (FNIR) for identification systems

[5] The International Standard ISO/IEC 19795-10 [3] defines differential performances as "difference in biometric system metrics across different demographic groups"

[6] The International Standard ISO/IEC 19795-10 [3] defines categorical demographic variable as "demographic variable of an individual that is nominally or ordinally described"

[7] The International Standard ISO/IEC 19795-10 [3] defines continuous demographic variable as "demographic variable of an individual that is observable, measurable, and that is not necessarily constrained to discrete categories"

change over time. When gender is included in the evaluation, gender should be determined through self-reporting. Gender self-reporting options presented to the capture subject shall be documented. In some evaluations that include gender, it is not always possible to obtain self-reported information.

- **Sex**: is defined as the state of being male or female as it relates to biological factors such as DNA, anatomy and physiology. Sex typically consists of two categories, "male" and "female". Female individuals generally possess two copies of the X chromosome. Male individuals generally possess one copy each of an X and a Y chromosome. Important exceptions do occur and complicate binary classification. The tester should establish appropriate categories for sex. If necessary, the tester can extend the general binary classification model of male/female.
- **Ethnicity**: in the context of biometric evaluations, ethnicities are classifications of individuals within a society based on shared qualities that are generally considered distinct within that society. Categories can reflect common physical characteristics, ancestry, language, community, religious affiliation, cultural heritage or other common qualities.
- **Skin tone**[8]: is the perceptual lightness or darkness value of an individual's skin. Skin tone is primarily determined by the amount of melanin in an individual's skin cells. Skin tone or the amount of melanin in skins cells, can be impacted by ethnicity as well as external factors, such as exposure to ultraviolet radiation or levels of vitamin A in the body.
- **Birthplace**: refers to the geographic location (e.g. a region or country) where an individual was born. When birthplace is included in the evaluation, birthplace shall be established through voluntary self-reporting or from available ID data or documents. In evaluations that include birthplace, the tester shall prepare a statement that documents the method for determining birthplace. If utilizing self-reporting to establish birthplace, the tester shall prepare a statement that documents birthplace self-reporting options presented to the data subject. If birthplace is recorded more finely than nation state (e.g. by a region within a country), the tester shall prepare a statement that documents how this granularity was established (see 7.3). Birthplace is a distinct demographic variable from ethnicity and shall not be used as a proxy for ethnicity.
- **Age**: the age of an individual is the quantity of time that has elapsed since the moment of the individual's birth. Age is commonly expressed in months or years. When age is included in the evaluation, age shall be established through self-reporting. Age can be subsequently verified via identity documents (e.g. a driver's license, passport, birth certificate, etc.).

Additional categories that are contained in ISO/IEC 19795-10 are height, weight, place of residences and native language. These categories are not considered in this first phase of the Ad Hoc group work.

The following demographic variable that can influence face image quality component values is also of interest for this report. **Wearing of eyeglasses**: Face images show individuals who either wear eyeglasses or no eyeglasses on their nose.

[8]Skin tone is used intentionally as both skin lightness (ISO/IEC 19795-10) and hue may contribute to face image quality

5 Quality measures of interest

The discussion of the Ad Hoc group concluded that all quality measures are of interest, including the Unified Quality Score (UQS) and the 27 Component Quality Measures (CQM), which are defined in ISO/IEC 29794-5 [2]. Note that not all reports contained in the following sub-sections address all quality measures of interest and/or all demographic variables of interest due to the lack of data respectively ground truth data.

6 Demographic variability reports

6.1 DV report based on MST

6.1.1 Experimental Setup

Disclaimer: The report that is presented in Section 6.1 has been published recently at BIOSIG 2024 [28].

Datasets

Kabbani et al. [28] use four datasets in total for the evaluation: FRLL [29], FRGCv2 [30], LFW [31], and MST-E [22]. All four datasets include images of subjects with different skin tones, genders, ages, and ethnicities. The FRLL dataset features images taken in a controlled studio environment, while LFW has images taken in the wild. FRGC and MST-E have images taken indoors, outdoors, and with different lighting, poses, and expression conditions. The MST-E Dataset is meant to be a reference dataset for the Monk Skin Tone Scale (MST) [21] such that human observers can use it to train on how to label subjects on this scale, thus it includes ground truth labels about skin tone for each of the subjects [20, 22]. The other datasets do not have ground-truth skin tone labels. The FRLL dataset has ground-truth labels for age, gender, and ethnicity. LFW has manually verified gender labels [32].

To overcome the lack of ground truth age, gender, and ethnicity labels for some datasets, Kabbani et al. [28] use Face Attribute Classification (FAC) [33] to extract these labels when they are missing. The predicted labels are averaged on all images of the same subject to obtain more accurate results. However, there is no reliable automated method for predicting the real skin tone labels, so to make sure Kabbani et al. [28] obtain credible results for the skin tone analysis, they manually label 902 subjects from two datasets according to the MST's guidance. As per the Monk Scale Tone guidance, they use the MST-E dataset as a reference, and label all subjects in the FRLL dataset and 800 subjects in the LFW dataset[9][10]. The subjects in the LFW dataset are selected based on those that have the largest number of images to make sure that as many images of the same subject as possible are investigated, before giving them an MST scale value.

[9]The MST labels are available at: https://github.com/wkabbani/dv-fiqa

[10]Extracting labels from unlabelled images to replace missing ground-truth labels for age, gender and ethnicity appears error-prone. Classification errors and age-estimation errors may occur. Future version of this report should focus on data with given ground-truth labels

Dataset	#Images	#Subjects	#Skin Tone Labelled Subjects
MST-E	887	19	19
LFW	12684	5556	800
FRLL	597	102	102
FRGC	18154	227	-

Table 1: Overview of the evaluation datasets. The numbers are for the actual number of subjects and images used in the evaluation after discarding images where no face is detected.

FIQA Algorithms

To evaluate the FIQA measures defined in ISO/IEC FDIS 29794-5, Kabbani et al. [28] use the reference implementation in the Open Source Face Image Quality (OFIQ) framework[11]. OFIQ provides implemented algorithms for all quality measures.

6.1.2 Experiments and Results

Skin Tone

Kabbani et al. [28] evaluate the FIQA measures on the MST-E, FRLL, and LFW datasets where the ground truth skin tone labels are available. As shown in Figure 3, the score distributions of the unified quality score do not show any noticeable differences between the various skin tone groups. The scores are rather distributed along the same value ranges on each dataset, with a higher concentration around the median values. The score distributions for most of the other quality measures show rather the same behavior where there are not clear differences between the groups. Kabbani et al. [28] show one such example for the illumination uniformity measure in Figure 4. However, the distributions of two quality measures show clear differences in the results for different skin tone groups. Figure 5 shows the distributions of the dynamic range quality values. It is clear from the results that lighter skin tones are getting relatively higher quality values. This effect is mildly noticeable in LFW but clearly visible in FRLL. In MST-E, the distributions are almost split into two groups, with the lighter skin tone group having relatively higher quality values. Another quality measure where skin tone is having a very noticeable effect on the quality values is the luminance mean. Figure 6 shows the distributions of the luminance mean quality values. The results of the MST-E dataset are clearly split into the same two groups as for the dynamic range, but with a much larger gap. The two peeks in the distribution are understandable given that MST-E explicitly features images under good and bad illumination settings. The effect of skin tone on this measure is also much more noticeable on FRLL and LFW. On both datasets, the distributions are clearly becoming more concentrated toward lower quality values as the skin tone gets darker.

[11] https://github.com/BSI-OFIQ/OFIQ-Project

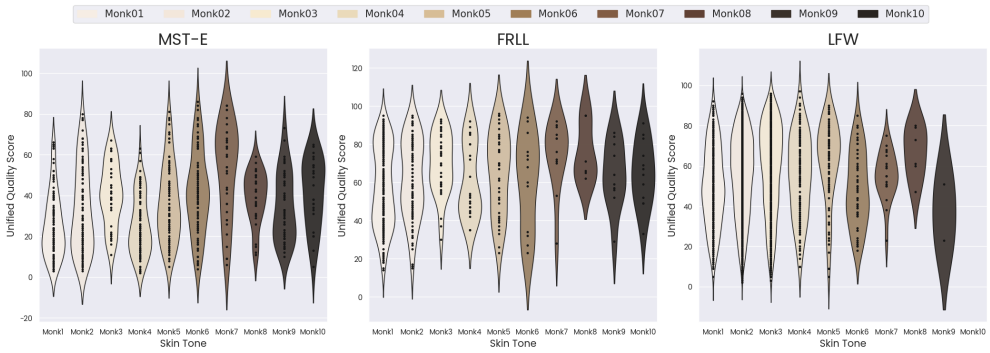

Fig. 3: Unified quality score distributions across the MST 10 skin tone scale.

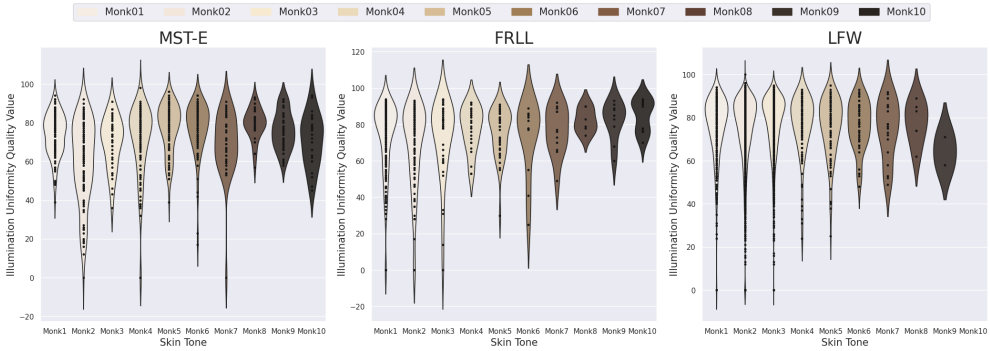

Fig. 4: Illumination Uniformity quality value distributions across the MST 10 skin tone scale.

Age

Kabbani et al. [28] evaluate the FIQA measures on all four datasets. They divide the age label into age groups and retain only the groups that have sufficient representation in one or more datasets. These are age groups: 20–40, 40–60, and 60–80. The evaluation results for all measures show no clear differences between the three age groups. Hence, Kabbani et al. [28] choose to show only the distributions of the unified quality scores in Figure 7.

Gender

The term *gender* refers to a classification based on social, cultural, or behavioral factors as per the international standard ISO/IEC 2382-37 on biometrics vocabulary [34]. In our study, Kabbani et al. [28] confine gender to two genders only, given that the ground truth labels and the face attribute classification models report gender in terms of male and female only. Similar to the results for age, there are no clear differences between the two genders in any of the evaluated measures. Figure 8 shows the distributions for the unified quality scores, and as evident from the results, the distributions are

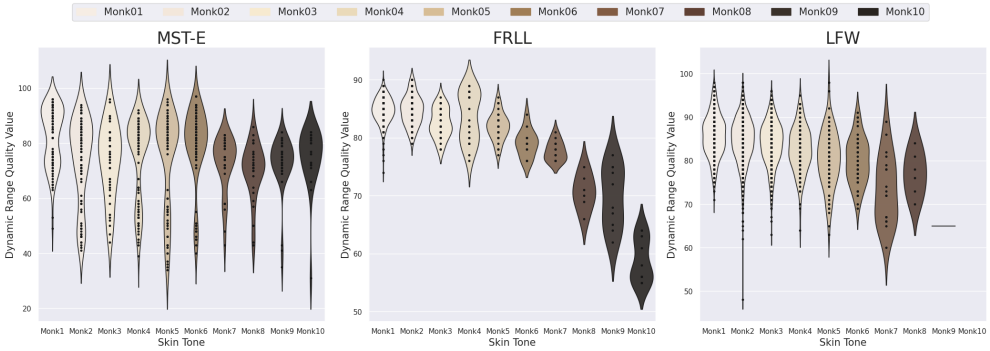

Fig. 5: Dynamic Range quality value distributions across the MST 10 skin tone scale.

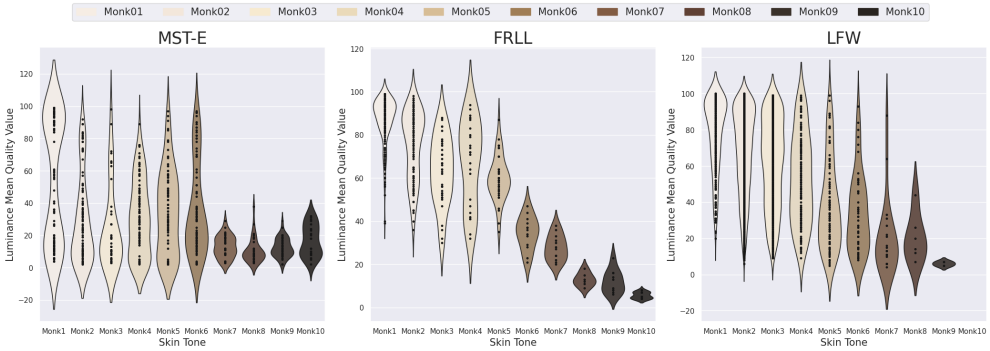

Fig. 6: Luminance Mean quality value distributions across the MST 10 skin tone scale.

very similar, with slightly more concentration of higher quality values for the male gender on FRLL and LFW, but on the other hand, slightly more concentration of lower quality values on MST-E. The results on FRGC are rather identical. Figure 9 shows the quality value distributions for the expression neutrality measure. It is also evident that the distributions are rather identical across the four datasets. The two peeks are also understandable, given that the datasets explicitly feature images of neutral and non-neutral expressions.

6.1.3 Discussion

The findings of the study are rather promising, because unlike what one might expect given the documented demographic bias in facial biometric systems, most FIQA algorithms, studied over four different datasets, have not demonstrated any substantial differences in their results across the demographic variables. The only two measures that have shown variations in their results on the skin tone variable are the luminance mean and the dynamic range. While it might be expected that these two aspects are different for individuals with different skin tones, it is worth noting that the FIQA algorithms are supposed to produce quality values that reflect how good a given image

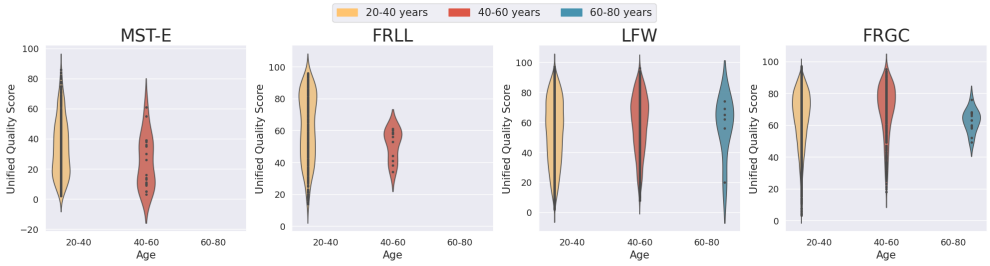

Fig. 7: Unified quality score distributions across the 3 age groups. MST-E and FRLL have no subjects in the age group 60-80.

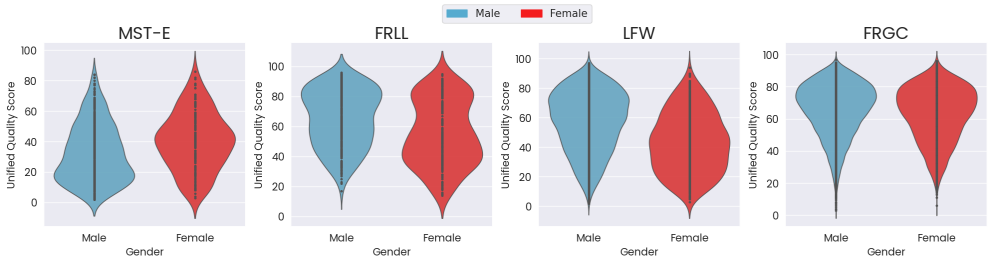

Fig. 8: Unified quality score distributions across the 2 gender groups.

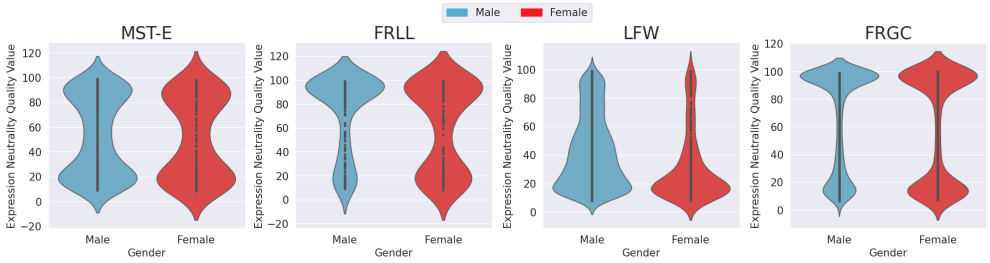

Fig. 9: Expression Neutrality quality value distributions across the 2 gender groups.

is with regard to the aspect assessed by the algorithm. Hence, it is not acceptable that these algorithms have variations in their outcomes only due to differences in skin tone.

6.2 DV report from NIST

6.2.1 NIST FATE SIDD Evaluation

Automated face recognition and age estimation are sensitive to quality problems in images. Standards set requirements on images; for machine-readable travel documents, these requirements are found in ISO/IEC 39794-5.

The NIST Specific Image Defect Detection (SIDD) evaluation tests automated tools used to check photo requirements, which are quantified in ISO/IEC 29794-5.

The images used by NIST for the current work are high-quality, front-facing *immigration application*-type images– these images are generally frontal and well-illuminated; subjects generally have mouths closed and eyes open.

The six measures evaluated are EyesOpen2 (eye openness normalized by chin-to-eyes distance), MouthOpen2 (mouth openness normalized by chin-to-eyes distance), Overexposure, Underexposure, Resolution, and Unified Quality Score.

Further information can be found in the NIST FATE Quality SIDD evaluation.

6.2.2 Regions of Birth

For this study, countries of origin were selected to have low levels of transcontinental migration. The countries were grouped into six regions of birth: East Africa, East Asia, East Europe, South Asia, Southeast Asia, and West Africa. For this analysis, the SIDD results are also separated by sex (female and male).

6.2.3 Results

Evaluation of the OFIQ algorithm shows the following (figures 10 and 11):

- slightly low values of **Eye Openness** (normalized by chin-to-eyes distance) for West African female subjects and East Asian male subjects
- high values of **Mouth Openness** (normalized by chin-to-eyes distance) for East African subjects
- high values of **Overexposure** for East European and East Asian subjects
- high values of **Underexposure** for East African and West African subjects
- slightly low values of **Resolution** for Southeast Asian subjects
- slightly low values for **Unified Quality Score** for West African subjects of both sexes and East African female subjects.

The implementation for OFIQ uses secunet_003 for EyesOpen2, MouthOpen2, Resolution, and Unified Quality Score, and secunet_005 for Overexposure and Underexposure.

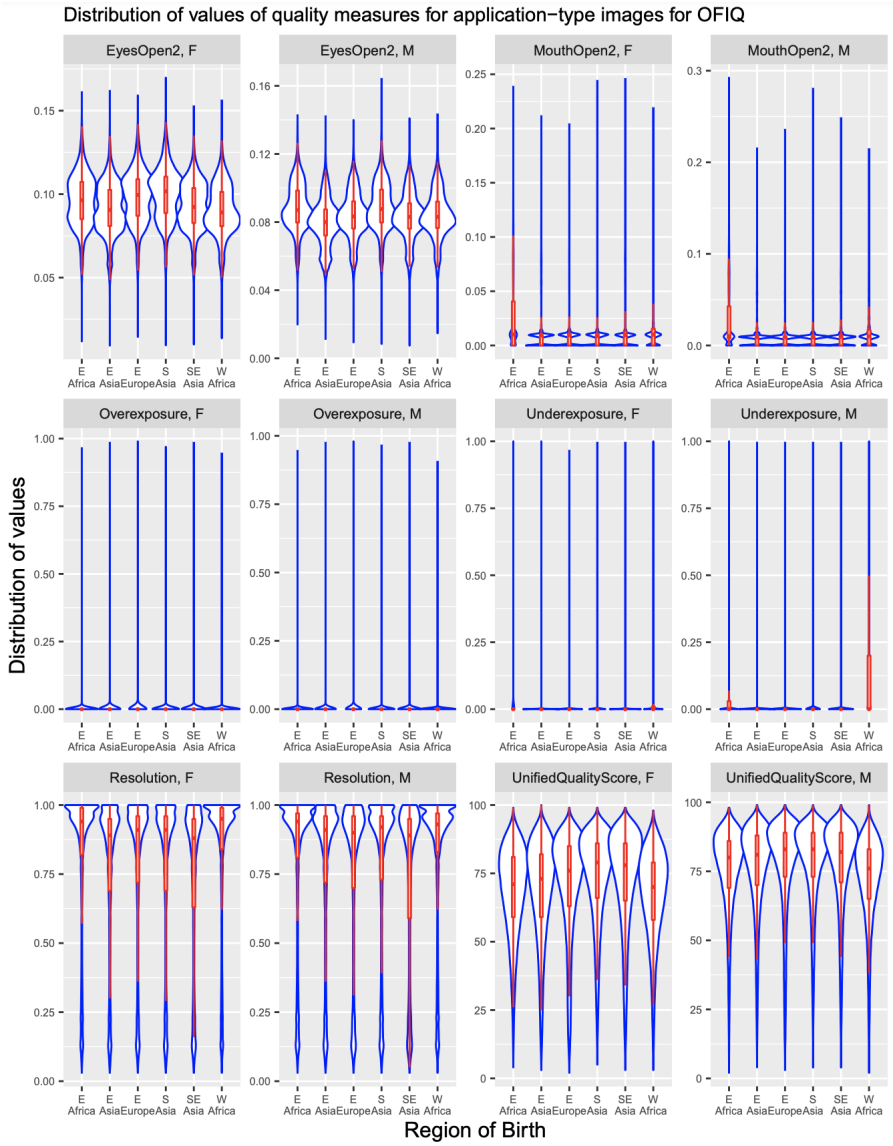

Fig. 10: For application-type photos, violin plots show distribution of values for six regions of birth and two sexes.

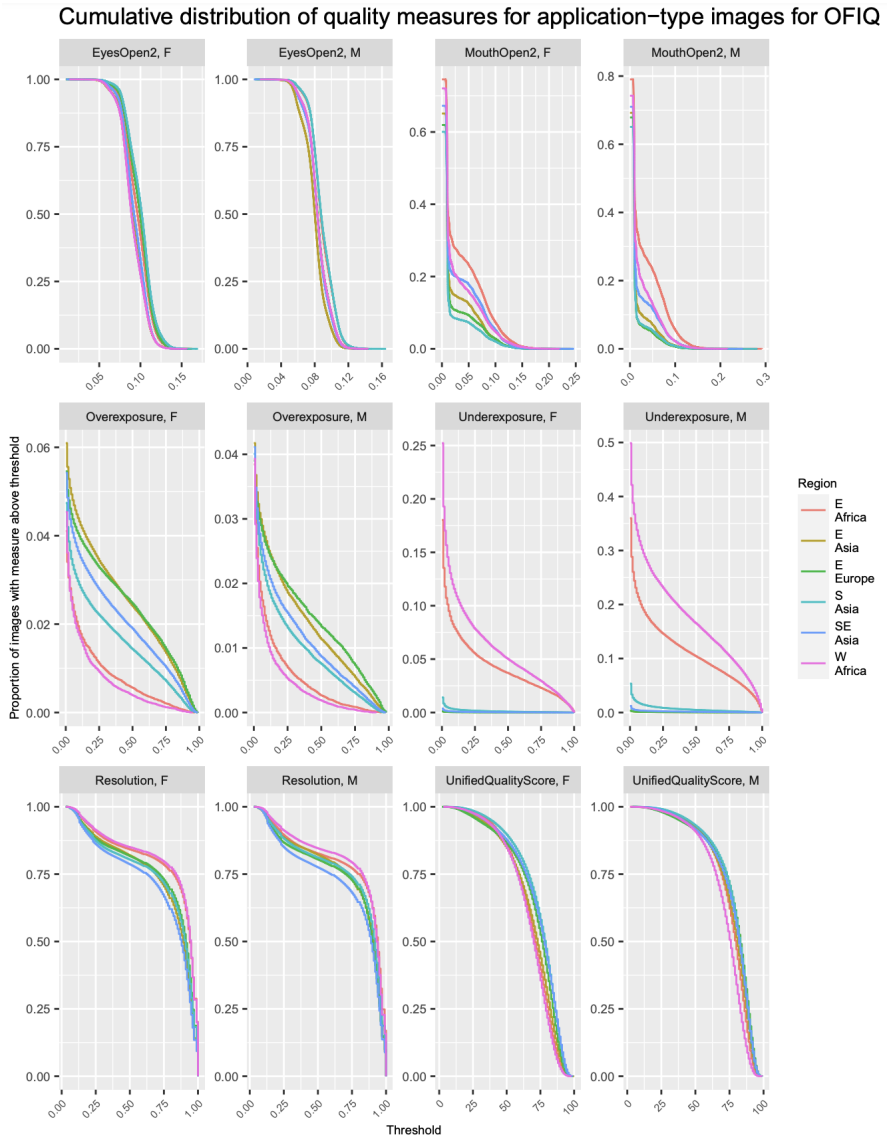

Fig. 11: For application-type photos, cumulative distribution function plots show distribution of values for six regions of birth and two sexes.

6.3 DV report regarding glasses

6.3.1 Experimental setup

Algorithm for measuring the percentage of face occlusion

For measuring the degree of face occlusion, we use the face occlusion prevention quality component of OFIQ [6], OFIQ uses the face segmentation approach face3d0725 [35], which is recommended in the emerging standard [2].

Data set

For evaluating the performance of the occlusion measurement method, a subset of well-lit frontal face images with neutral facial expression randomly selected from the Multi-PIE face image data set [36] was used, which is publicly available for research purposes.

This data set was chosen, because it contains face images known to be without occlusion as required for UC1 and UC2. Transparent eyeglasses are being worn in 25 probe images. In 60 probe images no eyeglasses are worn.

6.3.2 Experimental results

Figure 12 shows box and whisker plots for the distributions of the native OFIQ quality measure (percentage of face occlusion) and for the distributions of the face occlusion prevention quality component of OFIQ (mapped to the range from 0 to 100) in the subset of the Multi-PIE data set for face images with transparent eyeglasses and without any eyeglasses. A box is drawn between the first and third quartiles with a line in between marking the second quartile (median value); crosses represent mean values. Whiskers are drawn at the greatest/smallest percentage of occlusion smaller/-greater than 1,5 times the inter-quartile range (between the first and third quartiles) above/below the third/first quartile. Scores beyond the whiskers are outliers. Figure 12 shows that the measured percentage of occlusion is considerably higher (in the data set under consideration up to 15 %) when eye glasses are worn. At a 95 % confidence level, two-sample t-tests show that the differences between the mean values in Fig. 12 are statistically significant.

Figure 13 is a box and whisker plot showing the distributions of the OFIQ unified quality score in the subset of the Multi-PIE data set for face images with transparent eye glasses and without any eyeglasses. At a 95 % confidence level, a two-sample t-test shows that the difference between the mean score values for face images with and without eyeglasses are not statistically significant in the data set of high-quality images under consideration. Note that in data sets of lower-quality images showing, e.g., reflection artefacts on eyeglasses, there can be a significant difference between the mean score values for face images with and without eyeglasses.

6.3.3 Discussion

For UC1 concerning passport photographs, ICAO prohibits face occlusions except in specific exceptional cases [14]. For UC2 concerning system enrolment, occlusions are not permitted either [8]. When lenses are transparent and eyes are not occluded,

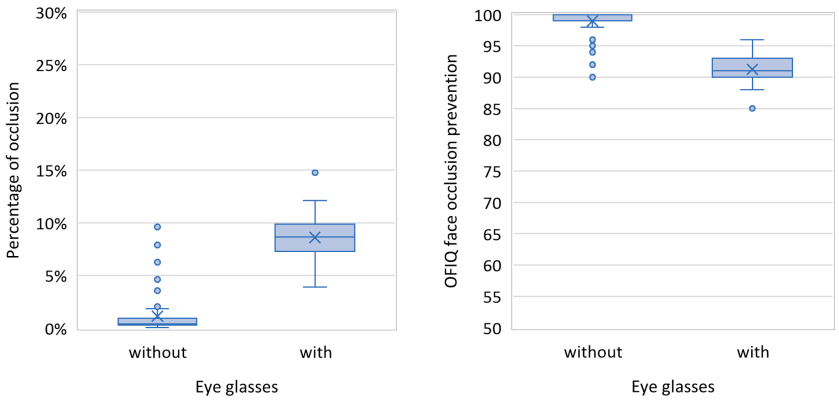

Fig. 12: Measured percentage of face occlusion and corresponding quality component of OFIQ for unoccluded face images with and without transparent eyeglasses

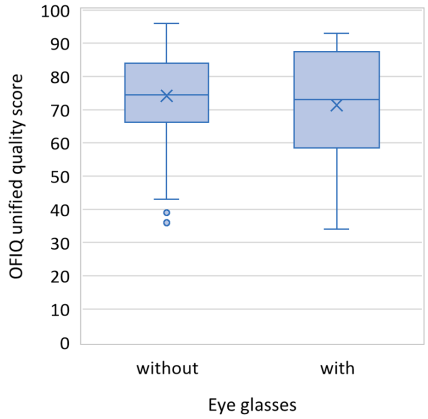

Fig. 13: Unified quality score of OFIQ for unoccluded face images with and without transparent eyeglasses

eye glasses affect neither mated nor non-mated comparison scores of face images (see Fig. 14 . Hence, both documents [8, 14] allow subjects to wear eye glasses with transparent lenses not occluding the eyes. The face segmentation approach face3d0725 [35], however, counts transparent eye glasses as occlusion even if the frame is neither extremely thick nor occluding the eyes. This deviates from the requirements [8, 14].

To avoid bias against the demographic group of wearers of glasses, for UC1 and UC2 the discard threshold for the measured face occlusion should be chosen to be at least 14,6 % (maximum value in above experiment). As this allows other, unwanted face occlusions to be ignored, it may be even better to retrain the occlusion segmentation model so that it does not consider transparent eye glasses as occlusions.

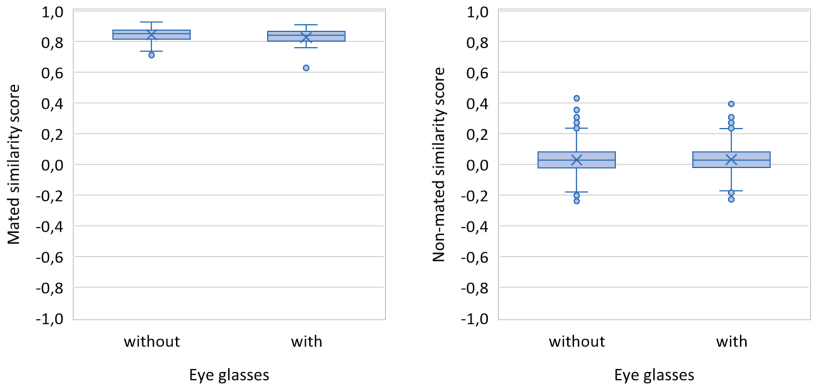

Fig. 14: Mated and non-mated MagFace similarity scores when comparing against high-quality reference face images from the Multi-PIE data set, for unoccluded face images with and without transparent eyeglasses. At a 95 % confidence level, a two-sample t-test shows that there is no statistically significant difference between the mean score values for face images with and without eyeglasses

The presence of glasses can be correlated with age. Therefore, bias against wearers of glasses can contribute to bias against elderly.

For UC3 concerning probe images for instantaneous recognition, some degree of occlusion is permissible as long as a given FNMR target is not exceeded. Determining how much face occlusion is permissible at a chosen FNMR target[12] is beyond the scope of this report.

[12]For automated border control, best practice is an FRR of at most 5 % at an FAR of 0.1 % [37]

7 Methodology

7.1 Objectives

The questions that a methodology for assessing the demographic variability of face image quality measures should help to answer include:

- How to decide whether slight differences in metrics for multiple demographic groups are statistically significant?
- How many biometric samples per demographic group are necessary and how many are sufficient for detecting demographic variability?

Disclaimer: The methodology that is presented in Section 7.2 has been published recently at ICPR 2024 [5].

7.2 Fairness metrics for biometric quality assessment

In this section we propose and compare multiple fairness measures for evaluating quality components across demographic groups.

In most cases, the result of a biometric quality assessment is a unified quality score [2], [1] (UQS), which is a single scalar value, representing the captured biometric sample quality. Alternatively, the output of a biometric quality assessment can also be a vector of quality values (i.e. quality components measures (QCM)), measuring various quality-related properties [13]. While there already exists numerous FIQA methods (see e.g. [38], [39], [15], [40]), the OFIQ algorithm is expected to be most influential, since it is used as a reference implementation for the International Standard ISO/IEC 29794-5 [2].

To measure demographic performance differences between various demographic groups, ISO/IEC 19795-10 [3] introduced the term *differential performance measure* (DPM), which is equivalent to the term *demographic differential* listed in the standard ISO/IEC 2382-37 [34]. In our context, a DPM is defined by a formula or algorithm that receives as input quality scores of different demographic groups and reflects how fair the underlying quality assessment algorithm is. Although ISO/IEC 19795-10 [3] specifies methods and statistical techniques for calculating DPMs, there is no dedicated standardised approach for assessing fairness of quality components across demographic groups.

The recent paper by Doersch et al. [5] introduces and compares new statistical approaches based on quality score distributions, for assessing fairness of quality components across demographic groups.[13] Proposed measures could be used as potential candidates for defining a fairness measure in an upcoming standard.

7.2.1 Background and Related Work

To ensure that quality algorithms provide equivalent results across demographic groups and investigate potential biases, various reports have been proposed in the scientific literature.

[13] The source code and data of this work is made available at: https://github.com/dasec/QA-Fairness-Measures

In the current NIST FATE SIDD report [17], FIQA algorithms for five quality measures are evaluated to quantify demographic performance differentials. These performance differentials were investigated across six demographic groups. It was found that only certain algorithms for the quality measures *Eyes Open 2* and *Resolution* exhibit demographic bias, while several algorithms for the quality measures *Mouth Open 2*, *Underexposure* and *Overexposure* exhibit demographic bias. As there is no standardised DPM for FIQA quality components yet, results shown were only visualized in the form of violin plots. In [41] Babnik et al. investigated demographic biases in FIQA methods. Although no specific quality components were analysed, it was found that FIQA methods generally exhibit significant bias and tend to favour white individuals. Terhörst et al. [42] evaluated FIQA algorithms with respect to potential bias in ethnicity and age. It was found, that for all evaluated FIQA algorithms, demographic performance differentials were observed.

These reports and studies demonstrate the importance of developing a standardised method for measuring demographic performance differentials in quality assessments and underlying algorithms in order to reveal potential biases in quality components.

7.2.2 Differential Performance Measures

Gini Coefficient

The Gini coefficient (GC) is a statistical measure of dispersion of a set of numbers [43]. This index, originally used to calculate income inequality, can be applied to various scenarios, including biometric measures. One biometric DPM based on the GC can be found, for example, in the ISO/IEC 19795-10 [3] standard for calculating performance differences for multiple groups. Since there is not yet a standardised DPM for assessing the fairness of quality assessment across demographic groups, we decided to use the GC as the backbone for this approach. Therefore, either the mean or median quality scores for any quality component Q across each demographic group d_i are utilized as inputs to the GC as follows:

$$\mathrm{GC} = \left(\frac{n}{n-1}\right)\left(\frac{\sum_i^n \sum_j^n \left|Q_{d_i} - Q_{d_j}\right|}{2n^2\overline{Q}}\right) \quad \forall d_i, d_j \in D \tag{1}$$

where n represents the number of demographic groups, Q_{d_i} is either the mean or median quality score of the demographic group d_i and D is the set of all demographic groups to be evaluated. However, a disadvantage of using median quality scores over mean quality scores is that, given slightly different demographic distributions of quality scores with relatively few outliers, all groups may receive exactly the same median score, even if there exists a slight bias. In addition, as previously for other DPMs listed in ISO/IEC 19795-10 [3], we adopt the approach of Howard et al. [25] by multiplying our result by a factor of $n/(n-1)$ to account for group self-comparisons ($i = j$ in equation 1), which can be especially relevant for smaller group numbers (n).

While the GC is used in many scenarios, one notable drawback is that it is rather insensitive to outliers. Table 2 shows synthetically generated mean and median quality scores for three demographic groups. These quality scores can be interpreted as

descriptive results of an arbitrary Quality Assessment Algorithm referring to any quality component Q. The quality scores for the fictitious quality component Q_1 were generated in such a way that one of the three groups exhibits a slight bias (a deviation of approximately 4 to 5 quality score points on average) compared to the other two groups, as shown visually in Figure 15. Table 3 shows another set of synthetically generated mean and median quality scores for the same three demographic groups. In this second fictitious quality component Q_2, a more prominent bias is simulated (a deviation of approximately 13 to 14 quality score points on average) compared to the others, which is shown in Figure 16. The *Sample Quality Fairness Rate* (SQFR), which follows a "higher is better" semantic outputs a fairness score in the range 0-1 and serves as our DPM.

In this report, the term SQFR for the general concept is annotated with a prefix depending on the method used. When using mean quality scores as input to the GC as fairness metric, the Mean-GC-SQFR is calculated as follows:

$$\text{Mean-GC-SQFR} = 1 - \text{GC}(Q_{d_n}) \tag{2}$$

where Q_{d_n} represents the mean quality scores of a quality component Q for a set of demographic groups D.

When instead using median quality scores as input to the GC as fairness metric, the Median-GC-SQFR is calculated as follows:

$$\text{Median-GC-SQFR} = 1 - \text{GC}(Q_{d_n}) \tag{3}$$

where Q_{d_n} represents the median quality scores of a quality component Q for a set of demographic groups D.

The SQFR scores for the setups discussed are shown in Table 4. For the two fictitious scenarios presented, the resulting SQFR scores are surprisingly high. This should not be the case, especially for the strongly biased quality component Q_2 (see e.g. Table 3 or Figure 16), as a significant deviation should result in a general lower SQFR score. Even for quality component Q_1, where one group slightly deviates from the others, one would not expect Mean-GC-SQFR or Median-GC-SQFR scores of *0.98* and *0.99*, which describe a near maximum fair system. To this end, to obtain a more reliable SQFR, the GC must be adjusted or alternative solutions developed as it does not sufficiently capture the underlying bias.

Table 4: SQFR results for quality component Q_1 and Q_2

SQFR	Quality component Q_1	Quality component Q_2
Mean-GC-SQFR	0.98	0.95
Median-GC-SQFR	0.99	0.95

Cubed Sample Quality Fairness Rate

To address the identified drawbacks of the traditional GC we provide an adapted approach, called Cubed Sample Quality Fairness Rate (CSQFR), which is designed

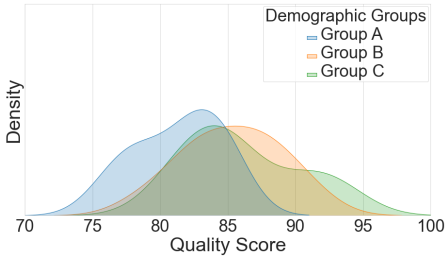

Fig. 15: Fictitious quality component Q_1 (Slightly biased): KDE Plot of the demographic score distribution

Table 2: Fictitious quality component Q_1 (Slightly biased): Synthetic Mean and Median Quality Scores of different demographic groups

	Group		
	A	B	C
Mean	81.3	85.3	86.1
Median	82	85.5	85

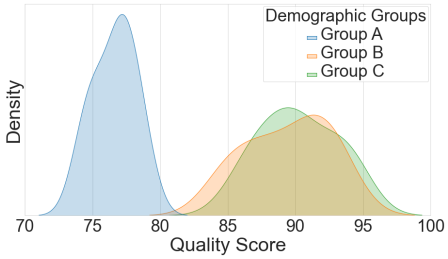

Fig. 16: Fictitious quality component Q_2 (Strongly biased): KDE Plot of the demographic score distribution

Table 3: Fictitious quality component Q_2 (Strongly biased): Synthetic Mean and Median Quality Scores of different demographic groups

	Group		
	A	B	C
Mean	76.6	89.4	90.2
Median	77	90	90

to achieve lower fairness scores in scenarios with demographic bias. This approach places greater emphasis on biased scenarios by cubing the result, making cases of lower fairness more visible. The adapted CSQFR when using mean quality scores as input to the GC, resulting in the Mean-GC-CSQFR is calculated as follows:

$$\text{Mean-GC-CSQFR} = (1 - \text{GC}(Q_{d_n}))^3 \tag{4}$$

where Q_{d_n} represents the mean quality scores[14] of a quality component Q for a set of demographic groups D. Different Quality Score scenarios for three demographic groups are provided in Table 5.

For the scenario *One group exhibits strong bias* Group A received significantly lower quality scores on average than the other two groups. The Mean-GC-SQFR score assesses this scenario a score of 0.73, indicating moderate fairness. However, the Mean-GC-CSQFR approach more reliably reflects the underlying bias by assigning a value of 0.38 to this scenario. For the scenario *One group exhibits slight bias* Group A received slightly lower quality scores on average than the other two groups. While the GC approach assesses this scenario with a high Mean-GC-SQFR score of 0.91,

[14]Due to the potential disadvantages of median scores compared to mean quality scores described, median values should not be used.

Table 5: Comparison of Mean-GC-SQFR and Mean-GC-CSQFR scores for different scenarios

Quality Score Scenarios	Mean QS Group A	Mean QS Group B	Mean QS Group C	Mean-GC -SQFR	Mean-GC -CSQFR
One group exhibits strong bias	35	95	89	0.73	0.38
One group exhibits slight bias	67	82	89	0.91	0.75
All groups receive different QS	30	50	95	0.63	0.25
All groups receive similar QS	84	89	87	0.98	0.94

the CSQFR captures the underlying slight bias much better, resulting in a moderate Mean-GC-CSQFR fairness score of 0.74. Furthermore, for the last two scenarios *All groups received different* and *All groups received similar QS*, the Mean-GC-CSQFR reflects the average quality scores more precise than the traditional Mean-GC-SQFR.

Low-Weighted-Mean Scores

Another DPM as an alternative to inputting the mean or medium quality scores into the GC, is our proposed approach of Low-Weighted-Mean (LWM) Scores. This method performs a linear weighting (from lowest to highest) of quality scores in a given demographic distribution, resulting in lower quality scores being weighted higher. Since captured biometric samples associated with lower scores would be rejected by more potential thresholds and this could disadvantage a group with in general lower quality scores more easily, this approach attempts to place a greater focus on fairness. This LWM weighting approach is calculated as follows: For each quality score q, a weight w is calculated as follows:

$$w(q) = 1 - \left(\frac{q - min(Q)}{max(Q) - min(Q)} \right) \tag{5}$$

where Q represents the union set of all quality scores across the demographic groups to be evaluated. This inverted min-max normalization ensures that our proposed method generalizes to quality scores at arbitrary scale while assigning higher weight to lower quality scores. For each quality score q the calculated weights (multiple occurrences of the same quality score) are accumulated and used to calculate a weighted arithmetic mean of the corresponding quality scores. In the unlikely special case where $min(Q) = max(Q)$ (i.e. there exists only a single quality score across demographic groups), this single quality score could then alternatively be used as output.

The adapted SQFR when using the LWM, resulting in the LWM-GC-SQFR is calculated as follows:

$$\text{LWM-GC-SQFR} = (1 - \text{GC}(\text{LWM}(Q_{d_n}))) \tag{6}$$

where Q_{d_n} represents the quality scores of a quality component Q for a set of demographic groups D.

A fictitious scenario for demonstration the approach of LWM is visualised in Figure 17. Even though the quality scores of Group A and Group B have similar mean and median values (see Table 6), the underlying quality score distributions are very

different. In this setup, more biometric samples of Group A would be rejected using lower operational thresholds than for Group B. However, as the LWM approach takes this behaviour into account, this results in a lower SQFR score[15], as demonstrated in Table 7.

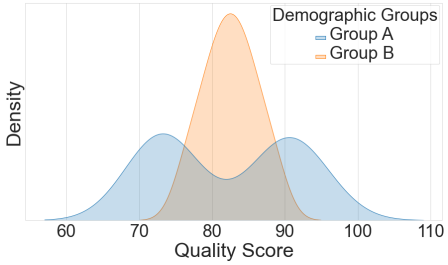

Fig. 17: Fictitious quality component Q_3: KDE Plot of the demographic score distribution

Table 6: Fictitious quality component Q_3: Synthetic Mean and Median Quality Scores of different demographic groups

	Group A	Group B
Mean	81.95	82.5
Median	81.5	82.5
LWM	75.4	81.4

The LWM approach achieves the lowest SQFR score in this scenario. When using the CSQFR with the LWM approach, the resulting LWM-GC-CSQFR becomes even lower, considering the different underlying distributions and placing greater emphasis on biased scenarios. On the other hand, one should be aware that the quality scores of Group A shown in Figure 17 represents a rather unrealistic distribution in an operational environment and thus the advantage of the LWM approach could potentially be better in theory than in reality.

Table 7: SQFR results for quality component Q_3

SQFR	Quality component Q_3
Mean-GC-SQFR	0.997
Median-GC-SQFR	0.994
LWM-GC-SQFR	0.962
LWM-GC-CSQFR	0.889

Mean-Discard-Gap

The last DPM that we propose in this research paper is the Mean-Discard-Gap (MDG). This approach first calculates the proportion of the biometric samples of a demographic group that are below a certain number of relevant thresholds. Relevant thresholds are selected as follows:

$$\text{Thresholds} = \{\min(QS) + 1, \min(QS) + 2, \ldots, \max(QS)\} \tag{7}$$

[15]SQFR Scores in Table 7 have been rounded to 3 decimal places as potential misinterpretations could occur with this setup

where QS represents the set of quality scores from all demographic groups to be evaluated. Thus, thresholds are limited to quality scores that exist in the demographic data set. To avoid a zero-distance discard, the first relevant threshold starts at $\min(QS)+1$. For all defined relevant threshold, a discard-percentage-gap is computed, which is the result of the distance between the minimum and maximum of the discard-percentage values across the groups. The final fairness measure is then derived by taking the mean discard-percentage-gap value of all min-max distances for all considered thresholds.

The adapted SQFR when using the MDG, resulting in the MDG-SQFR is calculated as follows:

$$\text{MDG-SQFR} = 1 - \text{MDG} \tag{8}$$

A fictitious scenario for illustrating the behaviour of the MDG-SQFR can be seen in Figure 18.

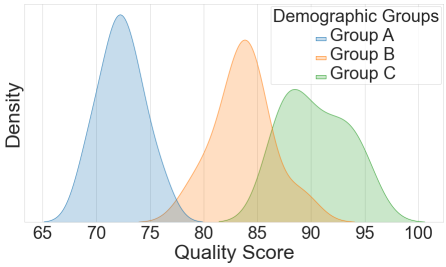

Fig. 18: Fictitious quality component Q_5: KDE Plot of the demographic score distribution

Table 8: Fictitious quality component Q_5: Synthetic Mean and Median Quality Scores of different demographic groups

	Group		
	A	B	C
Mean	72.3	83.7	90.4
Median	72	83.5	90

On average, the quality scores of the three demographic groups differ from each other by 7-8 quality score points, which can be seen in Table 8. The resulting SQFR Scores with all proposed measures for the fictitious quality component Q_5 can be seen in Table 9.

Table 9: SQFR results for quality component Q_5

SQFR	Quality component Q_5
Mean-GC-SQFR	0.93
Median-GC-SQFR	0.93
LWM-GC-SQFR	0.93
LWM-GC-CSQFR	0.81
MDG-SQFR	0.3

Looking at the SQFR scores from Table 9, a clear trend can be seen: The resulting fairness score for the MDG-SQFR measure (0.3) is significantly lower than the previously presented measures (fairness scores of 0.93 and 0.81). This is due to the

property of MDG that it considers the largest possible fairness difference per threshold and ignores groups in between. For this scenario, therefore, only groups A and C are considered for the fairness evaluation, when using MDG.

7.2.3 Discussion

Table 10: Comparison of proposed SQFR scores for different scenarios of 5 groups

QS scenarios of groups	Mean QS Group A	Mean QS Group B	Mean QS Group C	Mean QS Group D	Mean QS Group E	Mean-GC-SQFR	Mean-GC-CSQFR	LWM-GC-SQFR	LWM-GC-CSQFR	MDG-SQFR
One has strong bias	31.4	84.4	84.9	85.2	86.8	0.85	0.61	0.85	0.62	0.13
Two have strong bias	31.1	26.7	85	85.1	87.1	0.72	0.38	0.73	0.38	0.12
One has slight bias	79.1	85.6	85	85.1	86.9	0.98	0.94	0.98	0.94	0.52
Two have slight bias	76	77.5	85.6	86.9	85.8	0.96	0.89	0.97	0.9	0.47
All have similar QSs	85.7	87.5	85.6	86.6	86.5	0.99	0.98	0.99	0.98	0.71
All have equal QSs	87.5	87.5	87.5	87.5	87.5	1	1	1	1	1
All have different QSs	87.5	72.2	25	14.3	47.3	0.61	0.22	0.61	0.22	0.06

In this report, several measures for the fairness assessment of biometric quality have been presented. A distinction can be made between DPMs that use the Gini coefficient or variations of it, and measures that treat fairness differently, such as the MDG-SQFR measure.

In contrast to the GC-based measures, MDG-SQFR only considers the worst possible fairness difference per threshold ($max(discard\%) - min(discard\%)$) across demographic groups, not considering groups in between. Consequently, the number of demographic groups to be evaluated does not affect the resulting fairness score. A significant deviation of a single group (a disadvantaged group) is sufficient to receive a relatively low fairness score. This behavior can have an advantage over the GC-based measures for biased scenarios, as the resulting fairness score is generally lower. At the same time, however, this can also lead to a fairness score that is too low for scenarios where the quality score histograms are relatively similar across demographic groups, which is demonstrated in the scenario *All groups receive similar QS* in Table 10.

The GC-based measures, on the other hand, are group size sensitive, consequently returning different fairness scores for different group sizes. However, it should be noted that the number of groups is usually rather limited (e.g., there are typically 2 groups for gender comparisons and ethnic and age-specific characteristics are often binned). Furthermore, the GC-based measures and variations may be less robust, as they use scalar approximation values such as the mean quality score of a demographic group, which may not accurately reflect the underlying quality score distribution. For a comparison and overview of different scenarios of all the measures presented in this paper, see Table 10.

In general, if there is a preference of achieving even lower fairness scores for biased scenarios while slightly reducing the score of fairer scenarios, we recommend using a variation of the CSQFR over a variation of the SQFR. A promising CSQFR variant could be the LWM-GC-CSQFR, as it behaves similarly to the Mean-GC-CSQFR and additionally has the property of giving higher weight to lower quality scores. On the other hand, this weighting of the LWM-GC-CSQFR may not be necessary for quality score distributions in the field, as these are unlikely to include edge cases as demonstrated in Figure 17 and therefore the simpler Mean-GC-CSQFR may already be sufficient for quality score distributions in the field.

7.3 Addendum regarding the Mean-Discard-Gap

Section 7.2.2 introduced the "Mean-Discard-Gap (MDG)" computation in the context of integer quality scores, which should be applicable e.g. to OFIQ (introduced in section 2). The MDG concept does however generalise to arbitrary floating-point quality score distributions, if that is required: Taking figure 19 for example, the MDG value for the range of relevant quality score thresholds can be generally computed as the area between the minimum and maximum discard percentage curves, divided by the relevant threshold span.

The "range of relevant quality score thresholds" can be the range between the lowest and highest functionally distinct thresholds, or the range of possible algorithm output values (such as $[0, 100]$ for OFIQ), or some other range (e.g. by explicitly excluding higher thresholds above a limit that is no longer deemed to be operationally relevant). In the example's case the MDG value is 41.31% for the $[0, 100]$ range.

Additionally, note that this concept could easily be modified to e.g. use the mean of the distances of each group's discard percentage to the best group discard percentage per threshold, instead of intentionally focusing only on the distance between the worst group value (maximum curve) and the best group value (minimum curve) per threshold.

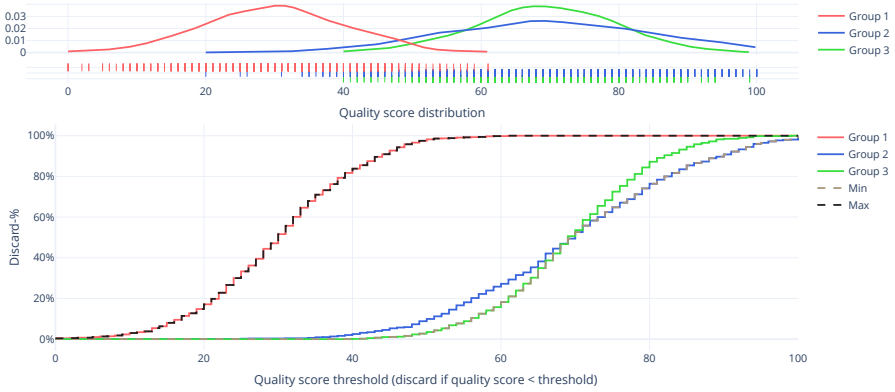

Fig. 19: A synthetic Mean-Discard-Gap (MDG) example with three groups. The first plot shows the quality score distribution density, while the second plot shows the corresponding discard percentages across the functionally relevant quality score threshold range. The "Min"/"Max" curves are the curves for the minimum/maximum discard percentage values across all group curves. Note that the discard percentage curves should typically use stepwise interpolation to reflect that a set of samples is discarded per threshold (as no fractions of samples can be discarded).

7.4 Incorrect Sample Discard Rate

ISO/IEC 29794-1 [1] defines the incorrect sample discard rate as the proportion of biometric samples incorrectly discarded when they would reach correct match decisions by a comparison subsystem. The incorrect sample accept rate is defined as the proportion of biometric samples incorrectly retained when they ultimately result in false non-matches by the comparison subsystem [1]. Both error rates should be reported together because both depend on the same quality score threshold used for discarding/retaining. To assess the impact of demographic variability when face image quality scores are used to make decisions on whether to discard or to retain images for further processing, for each demographic group the incorrect sample discard rate and the incorrect sample accept rate can be reported as functions of the quality score threshold

8 Conclusion and recommendations

This report article has looked at demographic variability of face image quality assessment methods in face recognition systems. The validation of algorithm fairness is fundamental. It remains important to demonstrate that face quality measures are not biased towards a specific demographic group.

The Ad Hoc group proposes to WG3 the following steps, which should be initiated in the January 2025 meeting:

- Transform Sections 1 and 2 to 6 of this Ad Hoc group report into an ISO/IEC TR. Via plenary resolution the DE NB is invited to submit as soon as possible a new

work item proposal "Technical Report on Demographic Variability of Face Image Quality Measures".
- Transform Sections 1, 2, 4 and 7 of this Ad Hoc group report into an ISO/IEC IS. Via plenary resolution the DE NB is invited to submit as soon as possible a new work item proposal "Fairness Metrics for Biometric Quality Assessment". Alternative titles to be considered are: "Fairness Evaluation for Biometric Quality Assessment" or "Evaluation of Demographic Differential of Biometric Quality Assessment Algorithms".

9 Acknowledgements

Authors include members of the WG3 Ad Hoc Group on Demographic Variability of Face Image Quality Measures. Thanks to all members of the Ad Hoc group for the intensive discussion we had in the virtual meetings. The views expressed in this paper reflect the work of the group and integrated in parts work of individual members, that was produced prior to the groups existence. They do not necessarily reflect the opinions or endorsements of all contributing authors.

10 Glossary terms

- biometric characteristic: biological and behavioural characteristic of an individual from which distinguishing, repeatable biometric features can be extracted for the purpose of biometric recognition
- biometric feature: number or label extracted from biometric samples and used for comparison
- biometric capture: obtaining and recording of, in a retrievable form, signal(s) of biometric characteristic(s) directly from individual(s), or from representation(s) of biometric characteristic(s)
- biometric capture device: device that collects a signal from a biometric characteristic and converts it to a captured biometric sample
- biometric capture process: series of actions undertaken to effect a biometric capture
- biometric capture subject: individual who is the subject of a biometric capture process
- biometric attendant: agent of the biometric system operator who directly interacts with the biometric capture subject
- bona fide presentation: biometric presentation without the goal of interfering with the operation of the biometric system
- comparison: estimation, calculation or measurement of similarity or dissimilarity between a biometric probe(s) and a biometric reference(s)
- comparison score: numerical value (or set of values) resulting from a comparison

- biometric recognition: automated recognition of individuals based on their biological and behavioural characteristics
- biometric sample: analogue or digital representation of biometric characteristics prior to biometric feature extraction
- biometric reference: one or more stored biometric samples, biometric templates or biometric models attributed to a biometric data subject and used as the object of biometric comparison
- biometric probe: biometric sample or biometric feature set input to an algorithm for comparison to a biometric reference(s)
- biometric utility: degree to which a biometric sample supports biometric recognition performance
- quality component: measurement on the biometric sample that may contribute to the computation of a unified quality score
- quality measure: quality score or quality component
- quality score: quantitative value of the fitness of a biometric sample to accomplish or fulfil the comparison decision
- EDC - Error-versus-Discard-Characteristic: method to evaluate the efficacy of quality assessment algorithms by quantifying how efficiently discarding samples with low quality scores results in improved (i.e., reduced) error. which can for example be the false non-match rate
- canonical face image: face image conformant to an external standard or specification of a reference face image
- FNMR - false non-match rate: proportion of the completed biometric mated comparison trials that result in a false non-match
- FMR - false match rate: proportion of the completed biometric non-mated comparison trials that result in a false match

References

[1] ISO/IEC JTC1 SC37 Biometrics: ISO/IEC 29794-1 Information Technology - Biometric Sample Quality - Part 1: Framework. International Organization for Standardization, (2024). International Organization for Standardization

[2] ISO/IEC JTC1 SC37 Biometrics: ISO/IEC FDIS 29794-5 Information Technology - Biometric Sample Quality - Part 5: Face Image Data. International Organization for Standardization, (2024). International Organization for Standardization

[3] ISO/IEC JTC1 SC37 Biometrics: ISO/IEC 19795-10. Information Technology – Biometric Performance Testing and Reporting – Part 10: Quantifying

Biometric System Performance Variation Across Demographic Groups. International Organization for Standardization, (2024). International Organization for Standardization

[4] Busch, C.: Challenges for automated face recognition systems. Nature Reviews Electrical Engineering (2024)

[5] Dörsch, A., Schlett, T., Munch, P., Rathgeb, C., Busch, C.: Fairness measures for biometric quality assessment. In: Proc. Intl. ICPR 2024 Workshop on Fairness in Biometrics (FairBio) (2024)

[6] Merkle, J., Rathgeb, C., Herdeanu, B., Tams, B., Lou, D., Dörsch, A., Schaubert, M., Dehen, J., Chen, L., Yin, X., Huang, D., Stratmann, A., Ginzler, M., Grimmer, M., Busch, C.: Open Source Face Image Quality (OFIQ) - Implementation and Evaluation of Algorithms. https://www.bsi.bund.de/SharedDocs/Downloads/EN/BSI/OFIQ/Projektabschlussbericht_OFIQ_1_0.pdf (2024)

[7] ISO/IEC JTC1 SC37 Biometrics: ISO/IEC 39794-5:2019 Information Technology - Extensible Biometric Data Interchange Formats - Part 5: Face Image Data. International Organization for Standardization, (2019). International Organization for Standardization

[8] ISO/IEC JTC1 SC37 Biometrics: ISO/IEC 19794-5:2011. Information Technology - Biometric Data Interchange Formats - Part 5: Face Image Data. International Organization for Standardization, (2011). International Organization for Standardization

[9] ISO/IEC JTC1 SC37 Biometrics: ISO/IEC 19795-1:2021. Information Technology – Biometric Performance Testing and Reporting – Part 1: Principles and Framework. International Organization for Standardization, (2021). International Organization for Standardization

[10] Funk, W., Arnold, M., Busch, C., Munde, A.: Evaluation of image compression algorithms for fingerprint and face recognition. In: Proc. IEEE Information Assurance Workshop (2005)

[11] European Council: Regulation 2017/2226 of the European Parliament and of the Council of 30 November 2017 on establishing an Entry/Exit System (EES) to register entry and exit data and refusal of entry data of third-country nationals (2017)

[12] European Council: Commission Implementing Decision 2019/329 of 25 February 2019 laying down the specifications for the quality, resolution and use of fingerprints and facial image for biometric verification and identification in the Entry/Exit System (EES) (2019)

[13] Schlett, T., Rathgeb, C., Henniger, O., Galbally, J., Fierrez, J., Busch, C.: Face

image quality assessment: A literature survey. ACM Computing Surveys (CSUR) (2021)

[14] Portrait quality (reference facial images for MRTD). ICAO Technical Report. https://www.icao.int/Security/FAL/TRIP/Documents/TR-PortraitQualityv1.0.pdf (2018)

[15] Meng, Q., Zhao, S., Huang, Z., Zhou, F.: MagFace: A universal representation for face recognition and quality assessment. In: 2021 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR) (2021)

[16] Schlett, T., Rathgeb, C., Tapia, J., Busch, C.: Considerations on the evaluation of biometric quality assessment algorithms. Trans. on Biometrics, Behavior, and Identity Science (TBIOM) (2023)

[17] Yang, J., Grother, P., Ngan, M., Hanaoka, K., Hom, A.: Face analysis technology evaluation (fate) part 11: Face image quality vector assessment - specific image defect detection. NIST Interagency Report 8485, National Institute of Standards and Technology (2024)

[18] Grother, P., Ngan, M., Hanaoka, K.: Face recognition vendor test (FRVT) part 3: Demographic effect. NIST Interagency Report 8280, National Institute of Standards and Technology (December 2019)

[19] Fitzpatrick, T.: The validity and practicality of sun-reactive skin types I through VI. Archives of Dermatology **124**(6), 869–871 (1988)

[20] Monk, E.: Monk Skin Tone Scale. Last accessed: 2024-12-15 (2019). https://skintone.google

[21] Monk, E.: The monk skin tone scale (MST). Technical report (2023). Harward University

[22] Schumann, C., Ollanubi, G., Wright, A., Monk, E., Heldreth, C., Ricco, S.: Consensus and subjectivity of skin tone annotation for ML fairness. In: Intl. Conf. on Neural Information Processing Systems (NIPS), pp. 30319–30348 (2023)

[23] Drozdowski, P., Rathgeb, C., Dantcheva, A., Damer, N., Busch, C.: Demographic bias in biometrics: A survey on an emerging challenge. Trans. on Technology and Society (TTS) **1**(2), 89–103 (2020)

[24] Drozdowski, P., Rathgeb, C., Busch, C.: The watchlist imbalance effect in biometric face identification: Comparing theoretical estimates and empiric measurements. In: Intl. Conf. on Computer Vision Workshops (ICCVW), pp. 1–9. IEEE, New York (2021)

[25] Howard, J., Laird, E., Rubin, R., Siroting, Y., Tipton, J., Vemury, A.: Evaluating

proposed fairness models for face recognition algorithms. In: Proc. Intl. Conf. on Pattern Recognition (2022)

[26] Rathgeb, C., Drozdowski, P., Frings, D.C., Damer, N., Busch, C.: Demographic fairness in biometric systems: What do the experts say? IEEE Technology and Society Magazine **41**, 71–82 (2022)

[27] Kotwal, K., Marcel, S.: Fairness index measures to evaluate bias in biometric recognition. In: Proc. Intl. Conf. on Pattern Recognition (2022)

[28] Kabbani, W., Raja, K., Raghavendra, R., Busch, C.: Demographic variability in face image quality measures. In: Proc. Intl. Conf. of the Biometrics Special Interest Group (BIOSIG) (2024)

[29] DeBruine, L., Jones, B.: Face Research Lab London Set (2021) https://doi.org/10.6084/m9.figshare.5047666.v5

[30] Phillips, J., Flynn, P., Scruggs, T., Bowyer, K., Chang, J., *et al.*: Overview of the Face Recognition Grand Challenge. In: Conf. on Computer Vision and Pattern Recognition (CVPR), vol. 1, pp. 947–954 (2005)

[31] Huang, G.B., Ramesh, M., Berg, T., Learned-Miller, E.: Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst (October 2007)

[32] Afifi, M., Abdelhamed, A.: AFIF4: Deep gender classification based on AdaBoost-based fusion of isolated facial features and foggy faces. Journal of Visual Communication and Image Representation **62**, 77–86 (2019)

[33] Park, S., Lee, J., Lee, P., Hwang, S., Kim, D., Byun, H.: Fair contrastive learning for facial attribute classification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10389–10398 (2022)

[34] ISO/IEC JTC1 SC37 Biometrics: ISO/IEC 2382-37:2022 Information Technology - Vocabulary - Part 37: Biometrics. International Organization for Standardization, (2022). International Organization for Standardization

[35] Yin, X., Chen, L.: FaceOcc: A diverse, high-quality face occlusion dataset for human face extraction. In: Treatment and Analysis of the Information Methods and Applications (TAIMA) (2022)

[36] Gross, R., Matthews, I., Cohn, J., Kanade, T., Baker, S.: Multi-pie. In: Proc. Intl. Conference on Automatic Face and Gesture (FG), pp. 1–8 (2008)

[37] Best practice technical guidelines for automated border control (ABC) systems. Frontex Technical Report (2015)

[38] Ou, F.-Z., Chen, X., Zhang, R., Huang, Y., Li, S., Li, J., Li, Y., Cao, L., Wang,

Y.-G.: SDD-FIQA: Unsupervised face image quality assessment with similarity distribution distance. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2021)

[39] Boutros, F., Fang, M., Klemt, M., Fu, B., Damer, N.: CR-FIQA: Face image quality assessment by learning sample relative classifiability. In: Conf. on Computer Vision and Pattern Recognition (CVPR), pp. 5836–5845 (2023). IEEE

[40] Kolf, J., Damer, N., Boutros, F.: GraFIQs: Face Image Quality Assessment Using Gradient Magnitudes (2024). https://arxiv.org/abs/2404.12203

[41] Babnik, Z., Štruc, V.: Assessing bias in face image quality assessment. In: 2022 30th European Signal Processing Conference (EUSIPCO), pp. 1037–1041 (2022). https://doi.org/10.23919/EUSIPCO55093.2022.9909867

[42] Terhörst, P., Kolf, J., Damer, N., Kirchbuchner, F., Kuijper, A.: Face quality estimation and its correlation to demographic and non-demographic bias in face recognition. In: Intl. Joint Conf. on Biometrics (IJCB), 2020 (2020)

[43] Gini, C.: Variabilità e Mutabilità, (1912)