Demographic Variability in Face Image Quality Measures

Wassim Kabbani [1] Kiran Raja [2] Raghavendra Ramachandra [3] Christoph Busch [4]

Abstract:

Face image quality assessment (FIQA) algorithms are being integrated into online identity management applications. These applications allow users to upload a face image as part of their document issuance process, where the image is then run through a quality assessment process to make sure it meets the quality and compliance requirements. Concerns about demographic bias have always been raised about biometric systems, given the societal issues this may cause. Therefore, it is important that any demographic variability in FIQA algorithms is assessed so it can be mitigated. In this work, we study demographic variability in all face image quality measures included in the ISO/IEC 29794-5 international standard across three demographic variables: age, gender, and skin tone. The results are rather promising and show no clear bias toward any specific demographic group for most measures. Only two quality measures are found to have rather considerable variations in their outcomes for different groups on the skin tone variable.

Keywords: Facial Biometrics, FIQA, Bias, Demographic Variability

1 Introduction

Facial biometric systems have become an integral part of our daily lives. We interact with them when we unlock our phones, cross an automatic border control (ABC) gate, gain access to a physical or virtual space, or apply for an identity document. These systems provide a seamless and convenient experience and have been shown to achieve high performance on several benchmark tests [DK20]. However, concerns about bias and demographic fairness have recently accompanied the deployment of these systems [Dr20], and algorithmic bias has been considered one of the important open challenges in biometrics [Ro19].

Studies have shown that facial image datasets as well as facial biometric algorithms such as face detection, face recognition, and gender classification are biased toward certain demographic groups [BG18, Te20, Ha21]. Thus, questions of how we assess and mitigate any demographic biases in any proposed AI system, and particularly in this context, facial biometric systems, are very often raised [Mi22]. To help address these concerns and to standardize how we measure and talk about demographic fairness, an international standard ISO/IEC 19795-10 on quantifying biometric system performance variation across demographic groups is currently under development [IS23].

[1] IIK, Info. Sec. and Comm. Technology, Gjovik, Norway, wassim.h.kabbani@ntnu.no
[2] IIK, Info. Sec. and Comm. Technology, Gjovik, Norway, kiran.raja@ntnu.no
[3] IIK, Info. Sec. and Comm. Technology, Gjovik, Norway, raghavendra.ramachandra@ntnu.no
[4] IIK, Info. Sec. and Comm. Technology, Gjovik, Norway, christoph.busch@ntnu.no

The international standard ISO/IEC 2382-37 on biometrics vocabulary [IS22] defines the term *demographic differential* as the *extent of difference in outcome of a biometric system across socially recognized sectors of the population*. Hence, we use the term *demographic differential* instead of *demographic bias* when referring to any variations in the outcome of the algorithms between different demographic groups.

In this work, we study demographic differentials in the area of face image quality assessment (FIQA), particularly assessing the quality measures included in the recent version of the international standard ISO/IEC 29794-5 on Biometric Sample Quality—Part 5: Face Image Data [IS24]. In its most recent version (2024-07-02), the Face Analysis Technology Evaluation (FATE) report on Specific Image Defect Detection (SIDD) from NIST, which summarizes the evaluation results of face image quality algorithms submitted to its platform, includes a section on demographic differentials by "subject region of birth" and for only five quality measures [Ya24]. Instead, we focus on demographic differentials across three demographic variables: age, gender, and skin tone, which is a recently proposed alternative to traditional race categories when exploring performance differentials [Ho21].

Face image quality assessment (FIQA) refers to the process of assessing the utility of a face image to face recognition systems [Sc21] as well as assessing the image against a set of regulatory specifications [IS18, IS19]. The ISO/IEC 29794-5 standard defines a set of quality *measures* that apply to a face image. It categorizes these *measures* into two main groups: (1) *unified quality score* which gives an overall assessment that is not necessarily correlated with a particular aspect of the image. (2) *quality component* which assesses a specific aspect of the face image such as illumination uniformity, dynamic range, luminance mean, and background uniformity [IS24].

To study demographic differentials across the skin tone variable, a method for categorizing human skin tones is needed. The commonly used method in the literature is the Fitzpatrick Skin Types (FST) Scale. This scale describes a person's skin type in terms of its response to ultraviolet radiation (UVR) exposure and groups skin tones into six categories [Wa18]. However, while it may be useful for dermatological use cases, it has been shown to have a tendency toward lighter skin tones because they have more UV sensitivity [Ho21, TJX23]. In their study, Howard et al. report that FST is *poorly predictive of skin tone and should not be used as such in evaluations of computer vision applications* [Ho21]. To address issues in FST and to offer a more inclusive scale that can be used to assess and mitigate bias in computer vision systems, Google has recently released the Monk Skin Tone Scale (MST) [Mo19]. Introduced by Dr. Ellis Monk and made open source by Google, the MST introduces a more inclusive 10-tone scale explicitly designed to represent a broader range of communities. A study conducted by Google to understand how well participants across diverse communities felt their own skin tone was represented within the scale reported that participants found the MST Scale to be more inclusive than the Fitzpatrick Scale and better at representing their skin tone [Mo19]. Together with the scale, Google also released the Monk Skin Tone Examples (MST-E) dataset, meant as a reference dataset to train human annotators on producing consistent skin tone annotations according to the 10-point MST scale [Sc23].

The approach we follow in this work is to study the score distributions of all the FIQA measures included in the ISO/IEC 29794-5 standard on four different datasets and across the three chosen demographic variables, namely: skin tone, age, and gender. To the best of our knowledge, this is the first study on demographic differentials in FIQA that considers three demographic variables and all the quality measures included in the newest version of the ISO/IEC 29794-5 standard [3]. Besides, this is the first work in FIQA, and facial biometrics in general, to use the newly released and more inclusive Monk Skin Tone Scale (MST) to study and report skin tone bias.

2 Experimental Setup

2.1 Datasets

We use four datasets in total for the evaluation: FRLL [DJ21], FRGCv2 [Ph05], LFW [Hu07], and MST-E [Sc23]. All four datasets include images of subjects with different skin tones, genders, ages, and races. 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) 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 [Mo19, Sc23]. The other datasets do not have ground-truth skin tone labels. The FRLL dataset has ground-truth labels for age, gender, and race. LFW has manually verified gender labels [AA19].

To overcome the lack of ground truth age, gender, and race labels for some datasets, we use Face Attribute Classification (FAC) [Pa22] 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 we obtain credible results for the skin tone analysis, we manually label 902 subjects from two datasets according to the MST's guidance. As per the Monk Scale Tone guidance, we use the MST-E dataset as a reference, and we label all subjects in the FRLL dataset and 800 subjects in the LFW dataset. The subjects in the LFW dataset are selected based on those that have the largest number of images to make sure that we investigate as many images of the same subject as possible before giving them an MST scale value.

---

[3] Due to the limited space, we include only the plots for the most important measures and interesting results for each demographic variable. But results for all quality measures on all three demographic variables, along with the produced skin tone ground truth labels, will be made publicly available.

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

Tab. 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.

2.2 FIQA Algorithms

To evaluate the FIQA measures defined in ISO/IEC DIS 29794-5, we use the reference implementation in the Open Source Face Image Quality (OFIQ) framework [4]. OFIQ provides implemented algorithms for all quality measures.

3 Experiments and Results

3.1 Skin Tone

We 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 1, 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. We show one such example for the illumination uniformity measure in Figure 2. However, the distributions of two quality measures show clear differences in the results for different skin tone groups. Figure 3 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 4 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 peaks 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.

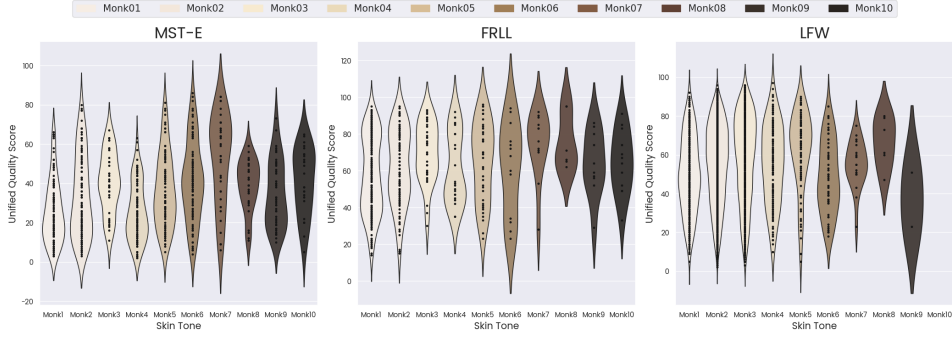[4] https://github.com/BSI-OFIQ/OFIQ-Project

Fig. 1: Unified quality score distributions across the MST 10 skin tone scale.
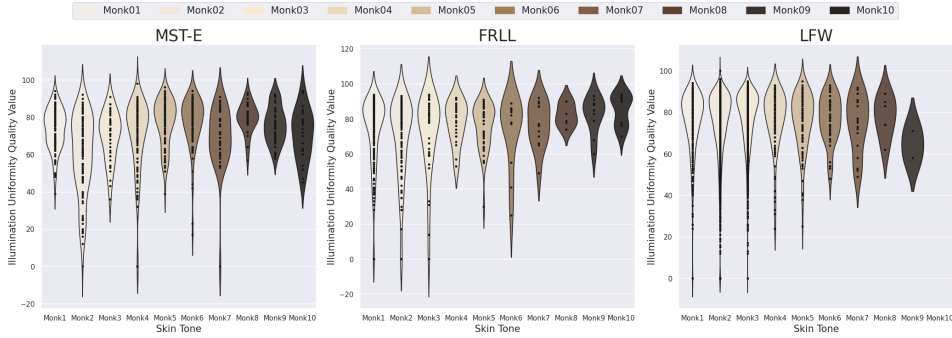

Fig. 2: Illumination Uniformity quality value distributions across the MST 10 skin tone scale.

3.2 Age

We evaluate the FIQA measures on all four datasets. We 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, we choose to show only the distributions of the unified quality scores in Figure 5.

3.3 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 [IS22]. In our study, we 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 6 shows the distributions for the unified quality scores, and as evident from the results, the distributions are very similar, with slightly more concentration of higher quality values for the male gender on FRLL and LFW, but on the other hand,
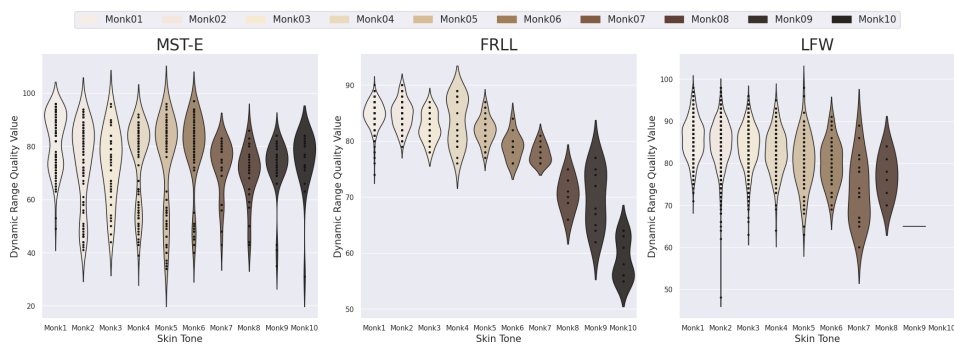
Fig. 3: Dynamic Range quality value distributions across the MST 10 skin tone scale.
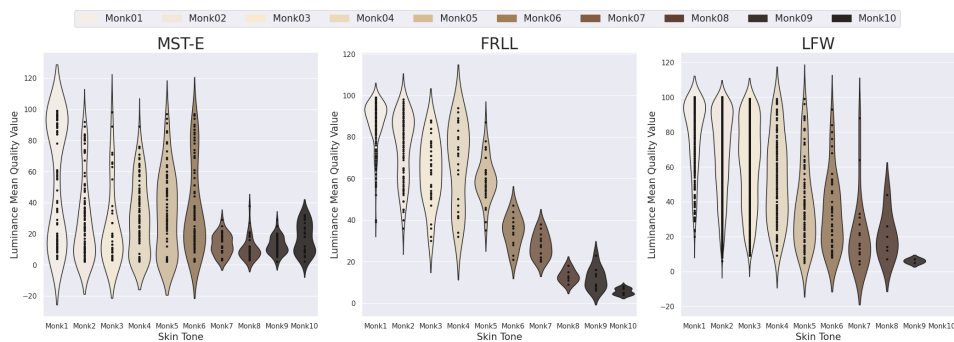

Fig. 4: Luminance Mean quality value distributions across the MST 10 skin tone scale.

slightly more concentration of lower quality values on MST-E. The results on FRGC are rather identical. Figure 7 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.

4 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 four 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 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.
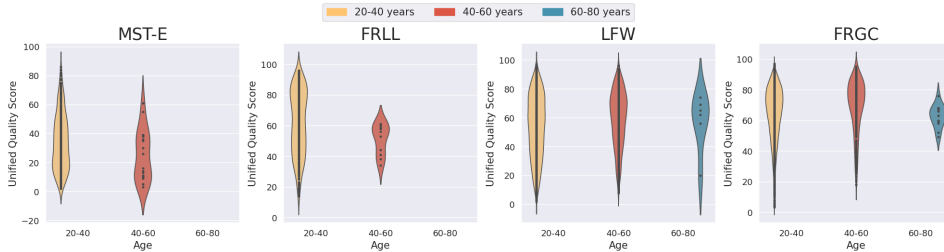
Fig. 5: Unified quality score distributions across the 3 age groups. MST-E and FRLL have no subjects in the age group 60-80.
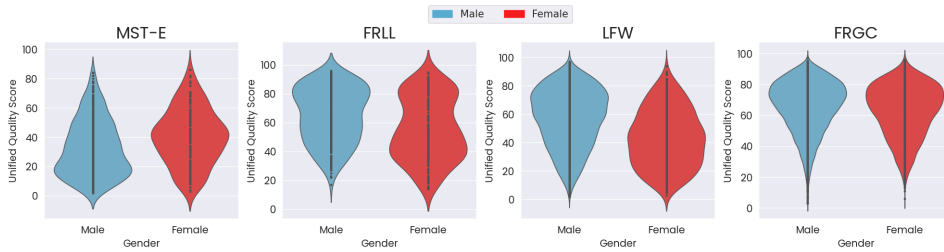

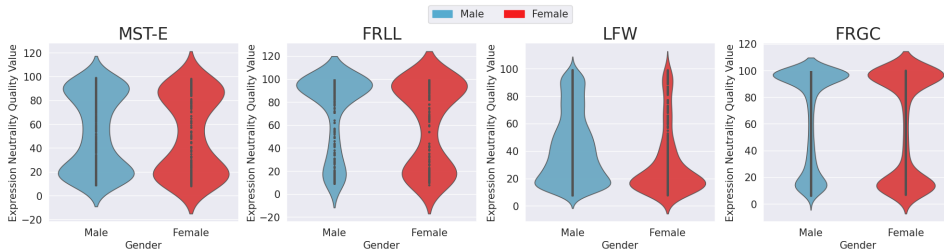Fig. 6: Unified quality score distributions across the 2 gender groups.

Fig. 7: Expression Neutrality quality value distributions across the 2 gender groups.

5 Conclusion

In this work, we presented a study on the less studied field of bias in face image quality assessment (FIQA). We conducted a comprehensive evaluation of demographic differentials in all quality measures included in the most recent version of the ISO/IEC 29794-5 international standard. We studied variations in these measures across three demographic variables: age, gender, and skin tone. We studied skin tone differentials on the newly introduced Monk Skin Tone Scale, used its reference labeled dataset, and manually labeled subjects in two other datasets to make sure the results are more reliable. The results of the study give us more confidence when deploying these algorithms as part of online identity management and document issuance applications and point to the specific algorithms that need some attention to alleviate the detected bias.

Acknowledgment

References

[AA19]   Afifi, Mahmoud; Abdelhamed, Abdelrahman: 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.

[BG18]   Buolamwini, Joy; Gebru, Timnit: Gender shades: Intersectional accuracy disparities in commercial gender classification. In: Conference on fairness, accountability and transparency. PMLR, S. 77–91, 2018.

[DJ21]   DeBruine, Lisa; Jones, Benedict: Face Research Lab London Set. 4 2021.

[DK20]   Dargan, Shaveta; Kumar, Munish: A comprehensive survey on the biometric recognition systems based on physiological and behavioral modalities. Expert Systems with Applications, 143:113114, 2020.

[Dr20]   Drozdowski, Pawel; Rathgeb, Christian; Dantcheva, Antitza; Damer, Naser; Busch, Christoph: Demographic Bias in Biometrics: A Survey on an Emerging Challenge. IEEE Transactions on Technology and Society, 1(2):89–103, 2020.

[Ha21]   Hazirbas, Caner; Bitton, Joanna; Dolhansky, Brian; Pan, Jacqueline; Gordo, Albert; Ferrer, Cristian Canton: Casual conversations: A dataset for measuring fairness in ai. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. S. 2289–2293, 2021.

[Ho21]   Howard, J.; Sirotin, Y.; Tipton, J.; Vemury, A.: Reliability and Validity of Image-Based and Self-Reported Skin Phenotype Metrics. Trans. on Biometrics, Behavior, and Identity Science (TBIOM), 3(4):550–560, October 2021.

[Hu07]   Huang, Gary B.; Ramesh, Manu; Berg, Tamara; Learned-Miller, Erik: Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments. Bericht 07-49, University of Massachusetts, Amherst, October 2007.

[IS18]   ISO/IEC JTC1 SC17 WG3: Portrait Quality - Reference Facial Images for MRTD. Bericht, International Civil Aviation Organization, April 2018.

[IS19]   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.

[IS22]   ISO/IEC JTC1 SC37 Biometrics: . ISO/IEC 2382-37:2022 Information Technology - Vocabulary - Part 37: Biometrics. International Organization for Standardization, 2022.

[IS23]   ISO/IEC JTC1 SC37 Biometrics: . ISO/IEC DIS 19795-10. Information Technology – Biometric Performance Testing and Reporting – Part 10: Quantifying biometric system performance variation across demographic groups. International Organization for Standardization, 2023.

[IS24]   ISO/IEC JTC1 SC37 Biometrics: . ISO/IEC DIS 29794-5 Information Technology - Biometric Sample Quality - Part 5: Face Image Data. International Organization for Standardization, 2024.

[Mi22] Michael, Katina; Abbas, Roba; Jayashree, Payyazhi; Bandara, Ruwan J.; Aloudat, Anas: Biometrics and AI Bias. IEEE Transactions on Technology and Society, 3(1):2–8, 2022.

[Mo19] Monk, Ellis: , Monk Skin Tone Scale, 2019.

[Pa22] Park, Sungho; Lee, Jewook; Lee, Pilhyeon; Hwang, Sunhee; Kim, Dohyung; Byun, Hyeran: Fair contrastive learning for facial attribute classification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. S. 10389–10398, 2022.

[Ph05] 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). Jgg. 1. IEEE, S. 947–954, June 2005.

[Ro19] Ross, Arun; Banerjee, Sudipta; Chen, Cunjian; Chowdhury, Anurag; Mirjalili, Vahid; Sharma, Renu; Swearingen, Thomas; Yadav, Shivangi: Some research problems in biometrics: The future beckons. In: 2019 international conference on biometrics (ICB). IEEE, S. 1–8, 2019.

[Sc21] Schlett, T.; Rathgeb, C.; Henniger, O.; Galbally, J.; Fierrez, J.; Busch, C.: Face Image Quality Assessment: A Literature Survey. ACM Computing Surveys (CSUR), December 2021.

[Sc23] Schumann, Candice; Olanubi, Gbolahan O; Wright, Auriel; Monk Jr, Ellis; Heldreth, Courtney; Ricco, Susanna: Consensus and Subjectivity of Skin Tone Annotation for ML Fairness. arXiv preprint arXiv:2305.09073, 2023.

[Te20] Terhörst, Philipp; Tran, Mai Ly; Damer, Naser; Kirchbuchner, Florian; Kuijper, Arjan: Comparison-Level Mitigation of Ethnic Bias in Face Recognition. In: 2020 8th International Workshop on Biometrics and Forensics (IWBF). S. 1–6, 2020.

[TJX23] Thong, William; Joniak, Przemyslaw; Xiang, Alice: Beyond Skin Tone: A Multidimensional Measure of Apparent Skin Color. In: ICCV. 2023.

[Wa18] Ward, William H; Lambreton, Fernando; Goel, Neha; Yu, Jian Q; Farma, Jeffrey M: Clinical presentation and staging of melanoma. 2018.

[Ya24] Yang, Joyce; Grother, Patrick; Ngan, Mei; Hanaoka, Kayee; Hom, Austin: Face analysis technology evaluation (FATE) part 11. Specific Image Defect Detection (SIDD). July 2024.