Literature Review for the GSA Identity Proofing Equity Study

I. OVERVIEW

Biometric recognition is automated recognition of individuals based on their biological and behavioral characteristics [16]. Biometric recognition has broad applications for border security, e-commerce, financial transactions, health care, and benefit distribution. Applications can be classified into two broad ways of operating: 1:1 or “verification” and 1:many or “identification”. Biometric verification, or one-to-one matching, is defined as the ”process of confirming a biometric claim through comparison” [16]. Biometric identification, or one-to-many matching, is defined as the “process of searching against a biometric enrolment database to find and return the biometric reference identifier(s) attributable to a single individual” [16].

With its explosion in use, there are concerns about the fairness of solutions across the broad spectrum of individuals, based on factors such as age, race, ethnicity, gender, education, socioeconomic status, etc. In particular, since biometric verification has a possibility of error, both false negatives (false rejection) and false positives (false acceptance), the expectation is that solutions have a performance that is “fair” across demographic groups. Buolamwini, et al. found that gender classification based on a single face image had a higher error rate for darker-skinned females with a high 34.7% error rate, compared to other groups (intersections of skin types and genders) [1], [2]. While focused on gender classification rather than face recognition, these papers brought considerable attention to this issue. Others found demographic differences in face recognition for some algorithms and systems [3], [4]. This has led to extensive research on assessing variance across demographic groups for face recognition [5], [6], [12], [22-63].

While this research has shone a light on the area, most research studies have focused on the matching algorithm, i.e, the “set of instructions and rules for processing biometric samples" [16] that performs the biometric comparison, i.e., “estimation, calculation or measurement of similarity or dissimilarity between a biometric probe(s) and a biometric reference(s) [16]. According to the International Organization for Standardization (ISO) standard 19795-1 [17], this would be considered a “technology evaluation”, defined as “offline valuation of one or more algorithms for the same biometric modality using a pre-existing or especially-collected corpus of samples" [17]. ISO defines a scenario evaluation as “evaluation that measures end-to-end system performance in a prototype or simulated application with a test crew”. A scenario evaluation adds the full context of the use case and includes the end-to-end system which incorporates image-capture hardware and software, user experience, quality control, etc. In other words, a technology test of the algorithms is a good first step to assess a biometric recognition system, but should be followed by scenario testing of the complete end-to-end system such that the full impact of remaining components can be assessed for the system’s intended use in a real world environment.

The study being undertaken by GSA would be considered a scenario evaluation, as it is an end-to-end evaluation of the software. The application is remote identity proofing software that utilizes face recognition as one of its steps. The process of face verification includes the user taking a photo of their own drivers license and taking a selfie which includes a liveness check. Liveness (also called presentation attack detection) of the individual are hardware and software components that prevent spoofing using a fake biometric, such as a printed photo of a face. A presentation attack is defined as “presentation to the biometric data capture subsystem with the goal of interfering with the operation of the biometric system” [20]. Facial verification in the GSA study is not simply matching two images (as in a technology test), but includes the full end-to-end experience of the individual. By implementing an end-to-end scenario evaluation of the remote identity proofing solutions, GSA will be able to learn whether or not the legitimate users' experiences measured by performance metrics varies by demographic group.

Outside of academia, the most closely related datasets and studies are the National Institute of Standards and Technology’s (NIST) Face Recognition Vendor Test [3, 63] and Department of Homeland Security’s (DHS) yearly Biometric Technology Rally [39, 40]. NIST’s Face Recognition Vendor Test (FRVT) is a technology test that seeks to “inform discussion and decisions about the accuracy, utility, and limitations of face recognition technologies. Its intended audience includes policy makers, face recognition algorithm developers, systems integrators, and managers of face recognition systems concerned with mitigation of risks implied by demographic differentials.” As opposed to

the new GSA study, FRVT is a technology evaluation based on commercial algorithms that are submitted to NIST. FRVT does not try to relate recognition errors to skin tone or any other phenotypes evident in faces in their image sets. FRVT is only focused on the algorithm, not the full end-to-end system. FRVT does not consider the role of the camera or the subject-camera interaction and how these possibly impacted the results. FRVT does not consider the errors that might be associated with the presentation attack detection (PAD) component.

The recent Department of Homeland Security Biometric Technology Rallies [39, 40] explored scenario evaluations for one-to-many matching. In these studies, participants' pictures were captured with a high resolution camera under ideal lighting to create a reference "gallery". This gallery was then incorporated into multiple face recognition systems. Participants would then enter a "check-in/security gate" environment either by themselves [39] or in groups [40] while each of the systems under test would capture one best photo of the participant and try to recognize and match these images to the people from the reference gallery.

DHS is planning a Remote Identity Validation Technology Demonstration in 2023. The information we've gathered so far indicates that it is a technology test focused on fraud rejection and thus it does not overlap with GSA's proposed study.

The proposed GSA study leverages the design and architecture of the DHS's Rally with important distinctions:

1. The GSA study not conducted in a laboratory setting and thus the data will feature a wide range of imaging conditions (ie. camera model, environmental lighting and setting, participant's proficiency taking pictures, etc.);
2. The GSA study is recruiting a larger participant pool (4,000 versus DHS's ~600); and
3. The GSA study expands on the identity proofing methods under test by incorporating consumer history checks that look for identity markers in physical address and financial records as well as phone account validation and device risk assessment products that are not usually tested.

GSA's study is the first of its kind in the U.S.; the study will attempt to gather sufficient data to determine the severity of bias in remote identity-proofing scenarios under "real world" conditions. GSA will recruit two to five times more respondents than similar studies; obtain representation from all parts of the United States, which will result in more robust statistical results; and ensure that participants are evenly distributed across demographics. To our knowledge, this is the largest public scenario evaluation of its kind.

The lack of similar studies and the need for this work is underscored by NIST's request for seeking related to establishing "metrics and testing methodologies to allow for assessment of performance and understanding of impacts across user populations (e.g., bias in artificial intelligence)" and recommending "operational testing to determine if the image capture technologies have introduced unintentional biases" in the Enrollment and Identity Proofing (SP-800-63A-4) sub-document [64]. GSA's proposed study will be a valuable reference for agencies and institutions seeking to fairly implement NIST's standards.

Statistical Overview

Another important aspect of this study is the statistical methods that will be developed to determine if there is a meaningful difference between groups or if that difference might be a difference seen by chance. The following paragraphs describe the equitability metrics and statistical approach.

To quantify the equitability of the various face recognition algorithms, multiple metrics have been proposed to evaluate fairness. The Fairness Discrepancy Rate (FDR) weights the two types of errors seen in biometric recognition (false accept and false reject rates), either equally or otherwise, and balances FDR across groups [5]. The U.S. National Institute of Standards and Technology (NIST) introduced the Inequity Rate (IR) metrics for face recognition algorithm performance testing and [6] proposed two interpretability criteria for biometric systems, i.e. Functional Fairness Measure Criteria (FFMC) and Gini Aggregation Rate for Biometric Equitability (GARBE). In other artificial intelligence (AI) research, evaluation metrics include demographic parity, equalized odds, and equal opportunity [7] [8] [9] [10]. Currently there is ongoing development of a new ISO standard on quantifying biometric system

performance variation across demographic groups [18].

However, with all of these active research and analyses, there has been limited contribution towards recommending appropriate statistical methods for determining when two or more groups are "equal" or not. This is essential, as any metric when measured in a sample, will have uncertainty which is a function of variability, correlation, number of groups, and other factors. This uncertainty can be measured through statistical methods, e.g. confidence intervals, to determine the likelihood that differences are found by chance or are a true difference. Given that exact "equality" is unlikely, if not impossible, for a set of groups, these methods allow for appropriate conclusions to be drawn from results.

This study will incorporate statistical methods for fairness solely for false negatives. Biometric solutions used widely by the public are typically based on "verification" or one-to-one matching. A false negative error is when the correct individual is falsely rejected, e.g., does not match their enrollment on a mobile device, passport, bank, or government benefits provider. This "error" may block an individual from accessing benefits which they are entitled to. TThe number of demographic groups being compared impacts the variation as an increased number of groups increases the chances that a difference between groups may be found "by chance", and thus adjustments need to be made in the test due to this effect, often called multiplicity [11].

Prior work includes development of two approaches to detecting differences in FNMR between demographic groups [21]. Additionally, we explore the trade-off among variation parameters based on simulations of a hypothetical equity study. In addition to giving guidance on expected outcomes of such a study, this paper provides suggestions for "practical" thresholds that could be used for when to say that a group is different that would minimize the possibility that that difference was based upon chance alone.

II. DETAILED LITERATURE REVIEW

There has been significant attention to artificial intelligence as a whole and as it relates to equity. Association for Computing Machinery, the world's largest body of computer scientists, urges an immediate suspension to private and governmental use of face recognition technologies, citing ".. clear bias based on ethnic, racial, gender, and other human characteristics.." [14] Federal Trade Commission the U.S. Federal Trade Commission released new guidance on AI fairness, highlighting that "It's essential to test your algorithm [for discrimination] based on race, gender, or other protected classes" [15]. Extensive evaluation such as testing proposed by GSA are one of the ways to ensure the technology is being used appropriately and fairly.

Summary of NIST Evaluation for Demographic Differentials

NIST has performed the most extensive evaluation of biometric verification as part of a technology evaluation [3]. Commercial software biometric algorithms are submitted to NIST for testing. Evaluation is performed across a variety of datasets including border, visa application, and mugshot images and for both identification (1:N) and verification (1:1). Performance is reported in terms of FNMR and FMR for verification and FNIR and FPIR for identification. Example results are shown in Figure 1. This figure from NIST shows results for 1:1 testing comparing photos submitted with a visa application with photos captured at the border. The figure shows false non-match rate, or probability the correct individual does not match, for demographic groups including country of origin, age (>45 and <45 years of age) and gender (male, female). Confidence intervals are provided based on bootstrapping. Results are continually updated at https://pages.nist.gov/frvt/html/frvt11.html

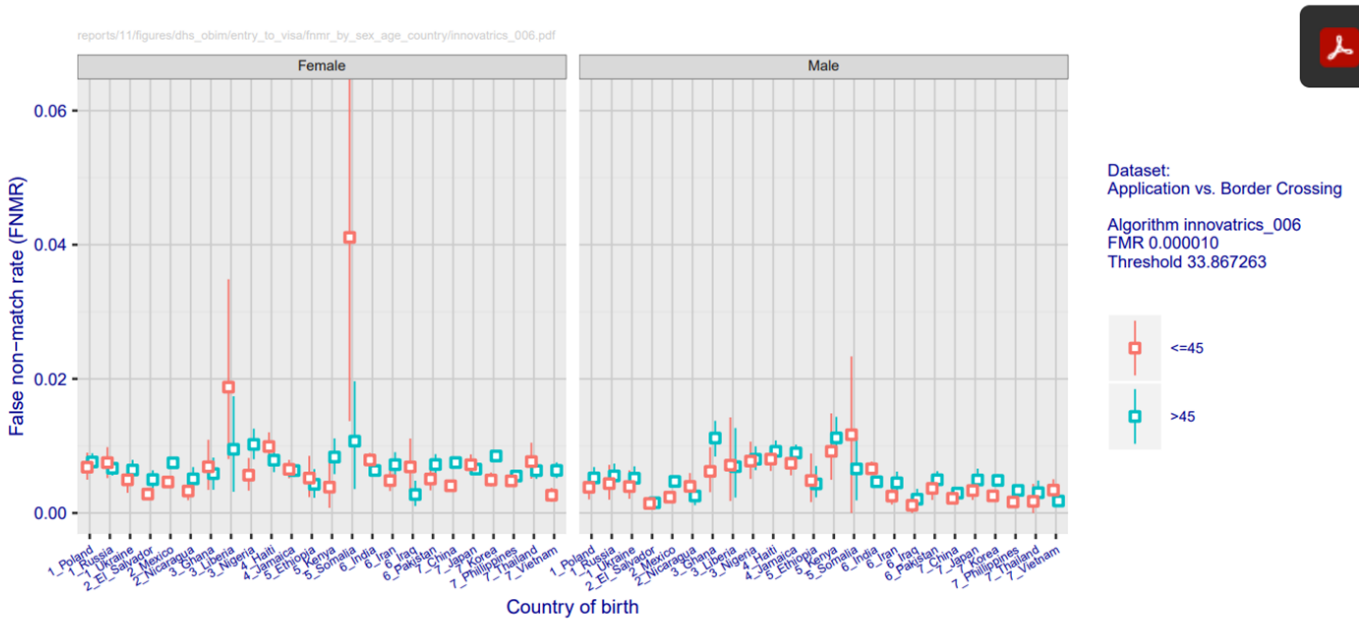

Figure 57: FNMR by sex, age and country of birth, innovatrics-006

Figure 1. Example results from NIST FRVT (technology evaluation) for demographic impact. False non-match rate, or probability the correct individual does not match, is plotted across demographic groups including country of origin, age (>45 and <45 years of age) and gender (male, female) for a specific commercial algorithm (innovatrics-006) [3].

The GSA study differs from NIST FRVT in the following ways:

- Scenario evaluation testing full end-to-end systems
- Comparison is based on photo of an identity document compared to a selfie
- Incorporates presentation attack detection (PAD)
- Incorporate user capture interface that will be used in the field

Academic Research on Equity in Face Recognition

There have been multiple academic papers that have looked at equity of face recognition [22 to 62]. The majority of work can be characterized as Technology Tests. Most are also focused on face matching and do not consider presentation attack detection and other components of an end-to-end system.

Previous studies have demonstrated biometric technologies are built around maleness, whiteness, and their ability to categorize within default categories, which measures the lack of inclusion of gender, class, race, and ability in biometric data collection [23, 35, 41, 49, 50, 53, 54, 55, 56, 58]. Earlier efforts [22] to reduce performance bias concentrated on matching methods i.e. failure to completely match all identities and failure to obtain exact matches. To that aspect, Howard et. al. [32] proposed two terms, “differential performance” and “outcome” to classify biometric performance differentials. In the last decade, facial recognition gained prominence, which exacerbated the issues of inclusion as the facial recognition algorithms showed systemic bias as an automated decision-making system [31]. The study further reveals that demographic factors have a significant influence on the performance of facial biometric algorithms i.e. a lower biometric performance for females, dark-skinned females, and youngest subjects [31]. The biometric performance biases are not limited to hand-crafted algorithms but can expand to deep neural network architectures in terms of biases in their learning process, which can translate to detection problems i.e. gender detection of face images [38]. Continued research showed variation across demographics and that results differ depending on specific algorithms, capture conditions, use cases and a host of additional factors [32, 46, 47, 48, 51, 52]. Different and inclusive training strategies can help overcome the learning bias of deep algorithms [25, 34, 36, 37, 57, 59, 60, 61]. For example, Zhang et. al. [61] proposed a robust face representation method for large training datasets, called ranged loss. Ranged loss learns the face representation in the dataset and to deal with learning biases of large training datasets. However, despite the best efforts of the academia to conclusively define the terms, “different

training strategies" or "inclusive training set ", they are not well defined universally. Recently, national and international standard institutes and researchers are making an effort to define the definitions of bias but they still lack a universal cohesiveness [3], [13], [17], [18], [63], [64].

The following paragraphs provide more detail on literature in this area.

Cavazos et al. compares the accuracy of different face recognition algorithms and their performance in recognizing faces of different races. The study found that some algorithms exhibit higher error rates in recognizing faces of certain racial groups, particularly those with darker skin tones. The authors call for further research and development of more accurate and fair face recognition algorithms [19].

Li and Abd-Almageed propose an information-theoretic approach to assess the bias in learned representations of pretrained face recognition models. The approach is based on quantifying the mutual information between different facial features and the identity labels, and the authors demonstrate its effectiveness on various datasets. The results show that the proposed method can accurately detect and quantify bias in pretrained face recognition models [24].

Wang et al. present an approach to reduce racial bias in face recognition systems called the Information Maximization Adaptation Network (IMAN). The proposed method learns to transform the feature representation of faces to minimize the influence of racial bias while preserving identity information. The authors demonstrate the effectiveness of IMAN on several benchmark datasets and show that it outperforms other state-of-the-art methods in reducing racial bias in face recognition [25].

Deuschel, Finzel, and Rieger present a study on the potential bias in facial expressions used to train and evaluate emotion recognition systems. The authors analyze a dataset commonly used for this purpose and find that the distribution of facial expressions is biased towards certain demographic groups, leading to inaccurate and unfair results. The paper highlights the need for more diverse and representative datasets in emotion recognition research [26].

Alshareef et al. investigate the gender bias in face presentation attack detection systems, where attackers attempt to bypass the face recognition system by presenting fake faces. The authors analyze the performance of several state-of-the-art face presentation attack detection systems and find that they exhibit gender bias, with higher error rates for female faces. The paper proposes a mitigation approach based on adversarial training, which improves the overall performance and reduces gender bias in these systems [27].

Singh et al. investigates the robustness of face recognition algorithms against both adversarial attacks and bias. The authors evaluate the performance of state-of-the-art face recognition models on several benchmark datasets and demonstrate their vulnerability to adversarial attacks and the presence of bias. The paper highlights the needs for developing a more robust and unbiased face recognition algorithm [28].

Bharati et al. investigate the performance bias in facial retouching detection algorithms towards in-group and out-group faces, i.e., faces from the same and different racial/ethnic groups as the algorithm's training data. The paper highlights the need for developing more diverse and representative datasets and improving the generalizability of facial retouching detection algorithms. The authors evaluate the performance on several state-of-art facial retouching models on a diverse datasets and show that they exhibit in group bias, with higher accuracy for in group faces [29].

Krishnan, Neas, and Rattani investigate the bias in facial recognition algorithms using near-infrared (NIR) spectrum imaging, which is commonly used in low-light conditions. The authors evaluate the performance of state-of-the-art face recognition models on a dataset captured using NIR spectrum imaging and compare it with a dataset captured using visible light. In addition, it shows that the face recognition models exhibit similar bias towards certain demographic groups in both visible light and NIR spectrum imaging, also highlighting the need for developing more robust and unbiased face recognition algorithms for low-light conditions [30].

Howard, Sirotin, and Vemury investigate the impact of demographic homogeneity on the performance of face recognition algorithms, specifically the imposter distributions and false match rates. The authors analyze several demographic factors, such as age, gender, and race, and evaluate their effect on the performance of face recognition algorithms on a diverse dataset. The study also demonstrates the performance of face recognition algorithms is significantly affected by demographic homogeneity and highlights the need for developing more robust and unbiased algorithms [32].

Dooley et al. compare human and machine bias in face recognition and conclude that both humans and machines

exhibit bias, with humans being slightly more biased than machines. The study also highlights the importance of diversifying datasets used for training machine learning algorithms to reduce bias [33].

Gong, Liu, and Jain propose a joint framework for debiasing face recognition and demographic attribute estimation by simultaneously learning the shared features between them. The proposed method outperforms existing methods in terms of both face recognition accuracy and demographic attribute estimation accuracy, while reducing bias against certain demographic groups. The study emphasizes the importance of addressing bias in face recognition systems and suggests a practical approach for doing so [34].

Kärkkäinen and Joo present a new dataset called FairFace, which is designed for measuring and mitigating bias in face recognition systems. FairFace consists of face images labeled with attributes such as race, gender, and age, with an emphasis on balancing these attributes to reduce bias. The authors demonstrate the utility of FairFace by evaluating the bias of several state-of-the-art face recognition models on the dataset [35].

Das, Dantcheva, and Bremond propose a multi-task convolutional neural network approach for mitigating bias in gender, age, and ethnicity classification. The proposed method leverages shared representations across multiple tasks to reduce bias in each individual task. Experimental results demonstrate the effectiveness of the proposed method in reducing bias in classification [36].

Gong, Liu, and Jain propose a group adaptive classifier approach for mitigating bias in face recognition. The proposed method uses demographic group information to adaptively adjust the decision boundary for each group, reducing the effect of bias on recognition accuracy. Experimental results on benchmark datasets demonstrate the effectiveness of the proposed method in mitigating bias while maintaining high recognition accuracy [37].

Robinson et al. examine the issue of bias in face recognition technology, specifically in terms of accuracy and fairness across different demographic groups. The authors conducted experiments using various datasets and found evidence of significant biases that need to be addressed in order to improve the overall performance of face recognition systems. The paper concludes by proposing several recommendations for mitigating these biases [41].

Sukthanker et al. explore the impact of different architectures and hyperparameters on fairness in face recognition technology. The authors use multiple datasets and conduct experiments to evaluate the performance of various models, finding that certain architectures and hyperparameters can significantly affect fairness. The paper concludes by suggesting future research directions for improving fairness in face recognition systems [42].

Burns et al. address the issue of gender bias in image captioning models and propose methods to overcome it. The authors specifically focus on the example of snowboarding, a sport where women are often underrepresented. They introduce a new dataset and perform experiments to demonstrate the effectiveness of their approach in reducing gender bias in image captioning [43].

Krishnan, Almadan, and Rattani investigate the fairness of ocular biometrics among different age groups: young, middle-aged, and older adults. The study uses eye images collected from a diverse group of participants and evaluates the performance of an ocular recognition system. The results show that the system exhibits a similar level of accuracy across all age groups, indicating fairness in its performance [44].

Bharati et al. propose a supervised deep learning approach for detecting facial retouching. The method uses a convolutional neural network (CNN) to identify inconsistencies in image regions that have been retouched. The proposed approach outperforms existing methods and can be used for forensic analysis of retouched images [45].

Wehrli et al. explore the issue of bias in deep-learning-based face recognition, highlighting the potential consequences of biased algorithms. It argues that a lack of awareness and ignorance of the potential biases present in such algorithms can lead to harmful outcomes, particularly for marginalized communities. The authors call for greater awareness and accountability in the development and deployment of these technologies [46].

Guo and Zhang provide an overview of the state-of-the-art in deep learning-based face recognition. It covers a wide range of topics, including face detection, face alignment, face feature extraction, and face classification. The article also discusses the challenges and future directions in this field [47].

Gong's PhD dissertation explores several aspects of face recognition, including the representation of facial features, intrinsic dimensionality, capacity, and demographic bias. The author investigates how these factors impact the accuracy and fairness of face recognition algorithms. The findings suggest that addressing issues of bias and increasing the diversity of training data can lead to more accurate and equitable face recognition systems [48].

Terhörst et al. reports on a study that explores biases present in face recognition technology beyond demographic

characteristics. The study analyzed five algorithms on a diverse dataset and found biases present beyond race and gender. The article concludes by emphasizing the need to address biases in the development and deployment of face recognition technology [49].

Yucer et al proposes a method for measuring hidden bias in face recognition technology using racial phenotypes. The study found that the proposed method can effectively detect hidden bias in face recognition algorithms. The article suggests that the proposed method can be used to improve the fairness and accuracy of face recognition technology [50].

Cheong, Kalkan, and Gunes discuss the problem of bias in affect recognition systems and propose a causal structure learning approach to identify and mitigate bias. The authors use a dataset of facial expressions to demonstrate the effectiveness of their method which is able to reduce bias. This approach has potential applications in creating more fair and inclusive affect recognition systems [51].

Yucer et al. explores the impact of lossy image compression on racial bias within face recognition technology. The authors conduct experiments using different levels of compression on images of people from different racial backgrounds, and test the accuracy of face recognition algorithms on these compressed images. The results suggest that lossy compression can have a negative impact on the accuracy of face recognition for all racial groups, but does not necessarily exacerbate existing racial biases [52].

Xu et al. investigates bias and fairness in facial expression recognition technology. The authors use a dataset of images with labeled expressions and evaluate the performance of a deep learning model on this dataset for different demographic groups. They find evidence of bias in the model's performance, particularly for certain demographic groups, and propose methods for mitigating this bias [53].

Pahl et al. evaluates bias in data and algorithms used for affect recognition in faces. The authors specifically focus on biases related to age, gender, and race. They find evidence of bias in both the training data and the deep learning models which is used for affect recognition, and proposed strategies for mitigating this bias [54].

Chouldechova et al. present a benchmarking framework for evaluating bias in unsupervised and semi-supervised face recognition systems. The authors use a variety of metrics to measure bias in demographic subgroups, such as race and gender. They also demonstrate the effectiveness of their framework on several datasets and provide recommendations for improving the fairness of face recognition systems [55].

Taati et al. discusses algorithmic bias in facial analysis technology when applied to older adults with dementia. The study evaluates the accuracy of a facial analysis algorithm in detecting emotion and gender in this group and identifies significant bias. The authors propose a set of recommendations to improve the performance of facial analysis technology in clinical populations [56].

Chen and Joo discuss the issue of annotation bias in facial expression recognition, which can lead to inaccurate models and unfair results. The authors propose a method to identify and mitigate this bias by incorporating diverse and representative data during training. Experimental results show that their approach improves the accuracy and fairness of facial expression recognition models [57].

Hussein et al. highlights the issue of racial bias in facial expression analysis and discusses its ethical implications. The authors provide a review of current research on this topic and suggest several approaches to mitigate the bias, such as using diverse datasets and developing algorithms that are less reliant on facial features. They also emphasize the importance of promoting transparency and accountability in the development and deployment of facial expression analysis systems [58].

Lin, Kim, and Joo present a fairness-aware gradient pruning method called Fairgrape for face attribute classification. Fairgrape is designed to address the issue of unfairness in facial recognition systems by identifying and removing bias from the gradient updates during training. The authors show experimentally that Fairgrape achieves improved fairness and accuracy in face attribute classification compared to other state-of-the-art methods [59].

Salvador et al. proposes a method called Faircal for calibrating the fairness of face verification systems. Faircal adjusts the decision threshold of the system to achieve a desired level of fairness based on demographic parity. The authors demonstrate through experiments that Faircal can effectively increase the fairness of face verification systems while maintaining high accuracy [60].

Equity Metrics and Statistical Methods

In this section, we discuss other work on statistical methods for comparison of bio-authentication across demographic groups. The NIST Information Technology Laboratory (ITL) quantifies the accuracy of face recognition algorithms for the demographic groups of sex, age, and race [3]. A component of the evaluation focuses on FNMR for one-to-one verification algorithms on four large datasets of photographs collected in U.S. governmental applications (domestic mugshots, immigration application photos, border crossing, and visa applications). For high-quality photos, FNMR was found to be low and it is fairly difficult to measure false negative differentials across demographics. Compared to high-quality application photos, the FNMR is higher for lower-quality border crossing images. Similar observations regarding image quality have been made by others, e.g. [4]. A measure of uncertainty is calculated for each demographic group based on a bootstrapping approach. In bootstrapping, the genuine scores are sampled 2,000 times and the 95% interval is plotted providing bounds for each group. No method was presented to suggest when an algorithm might be ”fair” under uncertainty. A notional approach might be to declare an algorithm fair if the intervals plotted overlap across all combinations of groups. This, however, does not fully address the possibility of Type I errors.

Cook et al. [4] examined the effect of demographic factors on the performance of the eleven commercial face biometric systems tested as part of the 2018 United States Department of Homeland Security, Science and Technology Directorate (DHS S&T) Biometric Technology Rally. Each participating system was tasked with acquiring face images from a diverse population of 363 subjects in a controlled environment. Biometric performance was assessed by measuring both efficiency (transaction times) and accuracy (mated similarity scores using a leading commercial algorithm). The authors quantified the effect of relative facial skin reflectance and other demographic covariates on performance using linear modeling. Both the efficiency and accuracy of the tested acquisition systems were significantly affected by multiple demographic covariates including skin reflectance, gender, age, eyewear, and height, with skin reflectance having the strongest net linear effect on performance. Linear modeling showed that lower (darker) skin reflectance was associated with lower efficiency (higher transaction times) and accuracy (lower mated similarity scores) [4]. While statistical significance of demographic factors was considered based on a linear model of match scores, this approach may not be applicable for assessing commercial systems which operate at a fixed threshold.

de Freitas Pereira and Marcel [5] introduce the Fairness Discrepancy Rate (FDR) which is a summary of system performance accounting for both FNMR and FMR. Their approach uses a “relaxation constant” rather than trying to assess the sampling variation or statistical variation between FNMR’s from different demographic groups. Howard et al. [6] present an evaluation of FDR noting its scaling problem. To address this scaling problem, the authors propose a new fairness measure called Gini Aggregation Rate for Biometric Equitability (GARBE).

Other research have also performed extensive evaluations of face recognition across demographic groups, e.g. [12], but have not presented statistical methods as part of their work.

III. REFERENCES

[1] J. Buolamwini and T. Gebru, "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification," in *Proceedings of Machine Learning Research, Conference on Fairness, Accountability, and Transparency*, 2018, pp. 1–15.

[2] J. A. Buolamwini, "Gender shades : intersectional phenotypic and demographic evaluation of face datasets and gender classifiers," 2017, MSc Thesis; http://hdl.handle.net/1721.1/114068; Last accessed: July 10, 2022.

[3] P. Grother, M. Ngan, and K. Hanaoka, "Face Recognition Vendor Test (FRVT) Part 3: Demographic Effects," United States National Institute of Standards and Technology, Tech. Rep., 2019, NIST.IR 8280, https://doi.org/10.6028/NIST.IR.8280.

[4] C. M. Cook, J. J. Howard, Y. B. Sirotin, J. L. Tipton, and A. R. Vemury, "Demographic effects in facial recognition and their dependence on image acquisition: An evaluation of eleven commercial systems," *IEEE Transactions on Biometrics, Behavior, and Identity Science*, vol. 1, no. 1, pp. 32–41, 2019.

[5] T. de Freitas Pereira and S. Marcel, "Fairness in Biometrics: A Figure of Merit to Assess Biometric Verification Systems," *IEEE Transactions on Biometrics, Behavior, and Identity Science*, vol. 4, no. 1, pp. 19–29, 2022.

[6] J. J. Howard, E. J. Laird, Y. B. Sirotin, R. E. Rubin, J. L. Tipton, and A. R. Vemury, "Evaluating proposed fairness models for face recognition algorithms," *arXiv preprint arXiv:2203.05051*, 2022.

[7] M. Hardt, E. Price, and N. Srebro, "Equality of opportunity in supervised learning," Tech. Rep., 2016, https://doi.org/10.48550/arXiv.1610.02413.

[8] R. K. E. Bellamy, K. Dey, M. Hind, S. C. Hoffman, S. Houde, K. Kannan, P. Lohia, J. Martino, S. Mehta, A. Mojsilovic´, S. Nagar, K. N. Ramamurthy, J. Richards, D. Saha, P. Sattigeri, M. Singh, K. R. Varshney, and Y. Zhang, "AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias," *IBM Journal of Research and Development*, vol. 63, no. 4/5, pp. 4:1–4:15, 2019.

[9] IBM, "AI Fairness 360," IBM toolkit; https://aif360.mybluemix.net/, Last accessed: July 11, 2022.

[10] O. A. Osoba, B. Boudreaux, J. Saunders, J. L. Irwin, P. A. Mueller, and S. Cherney, "Algorithmic equity: A framework for social applications," RAND Corporation, Tech. Rep., 2019.

[11] J. Hsu, *Multiple Comparisons: Theory and Methods*. Chapman & Hall/CRC, 1996.

[12] K. S. Krishnapriya, V. Albiero, K. Vangara, M. C. King, and K. W. Bowyer, "Issues related to face recognition accuracy varying based on race and skin tone," *IEEE Transactions on Technology and Society*, vol. 1, no. 1, pp. 8–20, 2020.

[13] S. Verma and J. Rubin, "Fairness Definitions Explained," in *Fair-Ware'18*. ACM, 2018, https://doi.org/10.1145/3194770.3194776.

[14] A. Eisgrau, "ACM US Technology Policy Committee urges suspen- sion of private and govern mental use of facial recognition technologies," 2020, https://www.acm.org/media-center/2020/june/ustpc-issues-statement-on-facial-recognition-technologies ; Last accessed: July 7, 2022.

[15] E. Jillson, "Aiming for truth, fairness, and equity in your company's use of AI," 2021, https://www.ftc.gov/business-guidance/blog/2021/04/aiming-truth-fairness-equity-your-companys-use-ai; Last accessed: July 7, 2022.

[16] ISO/IEC 2382-37:2022(en) Information technology — Vocabulary — Part 37: Biometrics. https://www.iso.org/obp/ui/#iso:std:iso-iec:2382:-37:ed-3:v1:en:term:37.08.03

[17] ISO/IEC 19795-1:2021(en) Information technology — Biometric performance testing and reporting — Part 1: Principles and framework.https://www.iso.org/standard/73515.html

[18] ISO/IEC CD 19795-10 Information technology — Biometric performance testing and reporting — Part 10: Quantifying biometric system performance variation across demographic groups (draft). https://www.iso.org/standard/81223.html

[19] J. G. Cavazos, P. J. Phillips, C. D. Castillo and A. J. O'Toole, "Accuracy Comparison Across Face Recognition Algorithms: Where Are We on Measuring Race Bias?," in *IEEE Transactions on Biometrics, Behavior, and Identity Science*, vol. 3, no. 1, pp. 101-111, Jan. 2021, doi: 10.1109/TBIOM.2020.3027269.

[20] ISO/IEC 30107-1:2016(en) Information technology — Biometric presentation attack detection — Part 1: Framework

[21] Schuckers, M., Purnapatra, S., Fatima, K., Hou, D. and Schuckers, S., 2022. Statistical Methods for Assessing Differences in False Non-Match Rates Across Demographic Groups. arXiv preprint arXiv:2208.10948.

[22] Rosenbaum, P. R., & Rubin, D. B. (1985). The bias due to incomplete matching. *Biometrics*, 103-116.

[23] Wevers, R. (2018). Unmasking biometrics' biases: facing gender, race, class and ability in biometric data collection. *TMG Journal for Media History*, *21*(2).

[24] Li, Jiazhi, and Wael Abd-Almageed. "Information-Theoretic Bias Assessment Of Learned Representations Of Pretrained Face Recognition." *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*. IEEE, 2021.

[25] Wang, Mei, et al. "Racial faces in the wild: Reducing racial bias by information maximization adaptation network." *Proceedings of the ieee/cvf international conference on computer vision*. 2019.

[26] Deuschel, Jessica, Bettina Finzel, and Ines Rieger. "Uncovering the bias in facial expressions." *arXiv preprint arXiv:2011.11311* (2020).

[27] Alshareef, N.; Yuan, X.; Roy, K.; Atay, M. A Study of Gender Bias in Face Presentation Attack and Its Mitigation. Future Internet 2021, 13, 234. https://doi.org/10.3390/fi13090234

[28] Singh, Richa, et al. "On the robustness of face recognition algorithms against attacks and bias." *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 34. No. 09. 2020.

[29] A. Bharati, E. Connors, M. Vatsa, R. Singh and K. Bowyer, "In-group and Out-group Performance Bias in Facial Retouching Detection," *2022 IEEE International Joint Conference on Biometrics (IJCB)*, Abu Dhabi, United Arab Emirates, 2022, pp. 1-10, doi: 10.1109/IJCB54206.2022.10007942.

[30] A. Krishnan, B. Neas and A. Rattani, "Is Facial Recognition Biased at Near-Infrared Spectrum as Well?," *2022 IEEE International Symposium on Technologies for Homeland Security (HST)*, Boston, MA, USA, 2022, pp. 1-7, doi: 10.1109/HST56032.2022.10025433.

[31] Drozdowski, P., Rathgeb, C., Dantcheva, A., Damer, N., & Busch, C. (2020). Demographic bias in biometrics: A survey on an emerging challenge. *IEEE Transactions on Technology and Society*, *1*(2), 89-103.

[32] J. J. Howard, Y. B. Sirotin and A. R. Vemury, "The Effect of Broad and Specific Demographic Homogeneity on the Imposter Distributions and False Match Rates in Face Recognition Algorithm Performance," *2019 IEEE 10th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, Tampa, FL, USA, 2019, pp. 1-8, doi: 10.1109/BTAS46853.2019.9186002.

[33] Dooley, Samuel, et al. "Comparing human and machine bias in face recognition." *arXiv preprint arXiv:2110.08396* (2021).

[34] Gong, Sixue, Xiaoming Liu, and Anil K. Jain. "Jointly de-biasing face recognition and demographic attribute estimation." *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIX 16*. Springer International Publishing, 2020.

[35] K. Kärkkäinen and J. Joo, "FairFace: Face Attribute Dataset for Balanced Race, Gender, and Age for Bias Measurement and Mitigation," *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, Waikoloa, HI, USA, 2021, pp. 1547-1557, doi: 10.1109/WACV48630.2021.00159.

[36] Das, Abhijit, Antitza Dantcheva, and Francois Bremond. "Mitigating bias in gender, age and ethnicity classification: a multi-task convolution neural network approach." *Proceedings of the european conference on computer vision (eccv) workshops*. 2018.

[37] Gong, Sixue, Xiaoming Liu, and Anil K. Jain. "Mitigating face recognition bias via group adaptive classifier." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2021.

[38] Serna, I., Pena, A., Morales, A., & Fierrez, J. (2021, January). InsideBias: Measuring bias in deep networks and application to face gender biometrics. In *2020 25th International Conference on Pattern Recognition (ICPR)* (pp. 3720-3727). IEEE.

[39] The 2021 Biometric Technology Rally (2021 Rally) at MdTF. https://mdtf.org/Rally2021/Results2021?Length=0. Last Accessed: April 6 2023..

[40] The 2022 Biometric Technology Rally (2022 Rally) at MdTF. https://mdtf.org/Rally2022/Results?Length=0. Last Accessed:

April 6 2023.

[41] J. P. Robinson, G. Livitz, Y. Henon, C. Qin, Y. Fu and S. Timoner, "Face Recognition: Too Bias, or Not Too Bias?," 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Seattle, WA, USA, 2020, pp. 1-10, doi: 10.1109/CVPRW50498.2020.00008.

[42] Sukthanker, Rhea, et al. "On the Importance of Architectures and Hyperparameters for Fairness in Face Recognition." *arXiv preprint arXiv:2210.09943* (2022).

[43] Burns, K., Hendricks, L. A., Saenko, K., Darrell, T., & Rohrbach, A. (2018). Women also snowboard: Overcoming bias in captioning models. *arXiv preprint arXiv:1803.09797*.

[44] A. Krishnan, A. Almadan and A. Rattani, "Investigating Fairness of Ocular Biometrics Among Young, Middle-Aged, and Older Adults," *2021 International Carnahan Conference on Security Technology (ICCST)*, Hatfield, United Kingdom, 2021, pp. 1-7, doi: 10.1109/ICCST49569.2021.9717383.

[45] A. Bharati, R. Singh, M. Vatsa and K. W. Bowyer, "Detecting Facial Retouching Using Supervised Deep Learning," in *IEEE Transactions on Information Forensics and Security*, vol. 11, no. 9, pp. 1903-1913, Sept. 2016, doi: 10.1109/TIFS.2016.2561898.

[46] Wehrli, S., Hertweck, C., Amirian, M., Glüge, S., & Stadelmann, T. (2022). Bias, awareness, and ignorance in deep-learning-based face recognition. *AI and Ethics*, *2*(3), 509-522.

[47] Guo, Guodong, and Na Zhang. "A survey on deep learning based face recognition." *Computer vision and image understanding* 189 (2019): 102805.

[48] Gong, Sixue. *Face Recognition: Representation, Intrinsic Dimensionality, Capacity, and Demographic Bias*. Michigan State University, 2021.

[49] P. Terhörst *et al*., "A Comprehensive Study on Face Recognition Biases Beyond Demographics," in *IEEE Transactions on Technology and Society*, vol. 3, no. 1, pp. 16-30, March 2022, doi: 10.1109/TTS.2021.3111823.

[50] Yucer, Seyma, et al. "Measuring hidden bias within face recognition via racial phenotypes." *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2022.

[51] J. Cheong, S. Kalkan and H. Gunes, "Causal Structure Learning of Bias for Fair Affect Recognition," *2023 IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW)*, Waikoloa, HI, USA, 2023, pp. 340-349, doi: 10.1109/WACVW58289.2023.00038.

[52] S. Yucer, M. Poyser, N. Al Moubayed and T. P. Breckon, "Does lossy image compression affect racial bias within face recognition?," *2022 IEEE International Joint Conference on Biometrics (IJCB)*, Abu Dhabi, United Arab Emirates, 2022, pp. 1-10, doi: 10.1109/IJCB54206.2022.10007956.

[53] Xu, Tian, et al. "Investigating bias and fairness in facial expression recognition." *Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16*. Springer International Publishing, 2020.

[54] Pahl, Jaspar, et al. "Female, white, 27? bias evaluation on data and algorithms for affect recognition in faces." *2022 ACM Conference on Fairness, Accountability, and Transparency*. 2022.

[55] Chouldechova, Alexandra, et al. "Unsupervised and semi-supervised bias benchmarking in face recognition." *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XIII*. Cham: Springer Nature Switzerland, 2022.

[56] B. Taati *et al*., "Algorithmic Bias in Clinical Populations—Evaluating and Improving Facial Analysis Technology in Older Adults With Dementia," in *IEEE Access*, vol. 7, pp. 25527-25534, 2019, doi: 10.1109/ACCESS.2019.2900022.

[57] Chen, Yunliang, and Jungseock Joo. "Understanding and mitigating annotation bias in facial expression recognition." *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021.

[58] Sham, Abdallah Hussein, et al. "Ethical AI in facial expression analysis: Racial bias." *Signal, Image and Video Processing* 17.2 (2023): 399-406.

[59] Lin, Xiaofeng, Seungbae Kim, and Jungseock Joo. "Fairgrape: Fairness-aware gradient pruning method for face attribute classification." *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022,*

Proceedings, Part XIII. Cham: Springer Nature Switzerland, 2022.

[60] Salvador, T., Cairns, S., Voleti, V., Marshall, N., & Oberman, A. (2021). Faircal: Fairness calibration for face verification. *arXiv preprint arXiv:2106.03761*.

[61] X. Zhang, Z. Fang, Y. Wen, Z. Li, and Y. Qiao, "Range loss for deep face recognition with long-tailed training data," in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Venice, Italy, Oct. 2017, pp. 5419–5428.

[62] "Beyond identity: What information is stored in biometric face templates?" in Proc. IEEE Int. Joint Conf. Biometrics (IJCB), Houston, TX, USA, Sep. 2020, pp. 1–10.

[63] NIST FRVT Demographics webpage, accessed 4/7/2023, https://pages.nist.gov/frvt/html/frvt_demographics.html

[64] SP 800-63-4 (Draft), Digital Identity Guidelines, Date Published: December 16, 2022, Authors: David Temoshok (NIST), Diana Proud-Madruga (Electrosoft), Yee-Yin Choong (NIST), Ryan Galluzzo (NIST), Sarbari Gupta (Electrosoft), Connie LaSalle (NIST), Naomi Lefkovitz (NIST), Andrew Regenscheid (NIST)