Literature Review of Sampling and Evaluation Bias in Facial Analysis and Recognition Algorithms

A Technical Report
presented to the faculty of the
School of Engineering and Applied Science
University of Virginia

by

Hana Nur

April 26, 2021

On my honor as a University student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments.

Hana Nur

Technical advisor: Daniel Graham, Department of Computer Science

Literature Review of Sampling and Evaluation Bias in Facial Analysis and Recognition Algorithms

Hana Nur
University of Virginia
hrn4ch@virginia.edu

ABSTRACT

This project aims to analyze the sources of sampling bias and evaluation bias responsible for direct discrimination in facial analysis and recognition algorithms. In the field of facial analysis, bias can occur throughout data collection as well as the development process. This can lead to sampling bias, which occurs when data used for a model is not randomly sampled, or evaluation bias, which occurs when an algorithm includes ill-fitting criterion. Previous work has investigated examples of facial recognition applications that have been found to discriminate against those with non-Caucasian features, ranging from differences in skin tone to eye shape. It is imperative bias is mitigated as facial recognition applications become more widespread, particularly when utilized in fields with a history of discrimination such as surveillance. In analyzing existing literature in the field to determine collective findings and disagreements, this project will aid future research into solutions for reducing bias.

1 Introduction

Artificial intelligence (AI) has allowed for the innovation of facial analysis and recognition algorithms, leading to extensive applications of facial recognition software in society. Sample bias occurs as a result of non-randomly sampled data, whereas evaluation bias occurs when an algorithm makes decisions using ill-fitting benchmarks [5]. The presence of these biases in algorithms skew decision making towards inaccurate and unfair predictions [5].

Bias in facial recognition algorithms may result in false matches, false non-matches, or longer transaction time, whereas bias in facial analysis software may result in incorrect classification of images [7,8]. Outcomes of algorithmic bias may have critical implications due to high stakes applications of facial analysis and recognition algorithms [6]. Facial analysis has been employed by law enforcement; one example being Amazon Rekognition, a cloud-based facial analysis platform with facial recognition capabilities. Rekognition uses deep learning to conduct analysis against an extensive database of faces, objects, and scenes [3]. In an investigation conducted by the American Civil Liberties Union (ACLU), Rekognition was found to match 28 members of congress with mugshots of other individuals arrested for criminal offenses. The ACLU reported that despite 20% of the congress members being people of color, 39% of the members falsely matched were people of color [3]. In response, Amazon implemented a one-year moratorium on law enforcement use of Rekognition and advocated for stronger regulations on use of facial analysis [4]. That same month, the ACM U.S. Technology Policy Committee called for the suspension of current and future use of facial recognition technology by the private and public sector “in all circumstances known or reasonably foreseeable to be prejudicial to established human and legal rights” [18].

Through surveying the sources of sample and evaluation bias present in facial recognition algorithms, this literature review will aid future researchers in determining solutions.

2 Background

The other-race effect, wherein individuals recognize faces of their own race more than other races, affects facial recognition in humans and has even been found to decrease confidence in eyewitness identification [9,11]. The contact hypothesis proposes the increased contact individuals have with members of their own race is correlated with the other-race effect, however previous studies have disproven this claim [10]. An other-race effect has been identified in facial recognition algorithms. A 2009 study demonstrated the effect by making a fusion of Western facial recognition algorithms and a fusion of East Asian algorithms. The Western algorithm more accurately recognized Caucasian faces, while the East Asian algorithm more accurately recognized East Asian faces, and the performance of these algorithms was found to be less stable than human performance [12]. Because of the other-race effect, performance and accuracy of facial recognition algorithms vary with demographic origin of both the algorithm and the test subjects.

Biometrics are the physical, behavioral, or adhered attributes of humans that can be used to identify an individual. Physical attributes, such as height, weight, eye color, or skin color, can be combined into more complex attributes, such as face or iris [16]. Behavioral attributes are identifiers such as gait or keystroke, and adhered attributes are pieces adorned by the individual such as clothes, tattoos, and accessories [16]. Facial analysis and recognition algorithms utilize biometrics to analyze and identify the image of an individual.

Direct discrimination occurs when an individual's attribute results in negative outcomes [5]. This is exemplified in the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) software which measures risk of recidivism. COMPAS disproportionately predicted higher recidivism rates in Black defendants [24]. As Black defendants faced worse outcomes due to a sensitive attribute, their race, this constitutes direct discrimination. In facial recognition or analysis algorithms, an example would be lower false match rates for those within a particular demographic.

3 Related Work

Previous surveys of algorithmic bias have determined the definitions for types of bias used in this project. Both sample bias and evaluation bias have been defined in the context of machine learning algorithms [5]. The types of errors were viewed from the dimensions of data and algorithms, however, were not categorized in this way due to the feedback loop phenomenon. This phenomenon dictates that the decisions made by a trained algorithm impacts the outcomes that produce future data for later trained algorithms. Therefore, the types of data are not independent of one another and categorizations of bias in data versus bias in algorithms are insufficient [5]. Definitions of fairness and types of discrimination by cause of algorithmic bias have also been determined [5]. Studies have detailed sources of historical bias in facial recognition algorithms, citing examples in photography and the chemistry of film [14]. Auditing bias in AI algorithms has been found to reduce classification bias in facial recognition algorithms [13]. A survey on facial analysis systems determined that most facial analysis software is biased and categorized methods to reduce bias [15]. A literature review concerning demographic bias in biometrics described sources of bias based on gender, race, and several other identities. The other-race effect was found to impact biometric data acquisition and comparison [1]. The negative impacts of bias in applications of facial recognition algorithms were studied in the context of automated border control. Bias resulting in increased false matches among specific demographics may be exploited, allowing falsified identification developed with open-source deep models to match subjects' images through a morphing attack [17]. A morphing attack uses a fake portrait, developed by combining the features of two existing portraits, to match with two individuals' portraits. This exploitation also results in increased verification error rates for individuals in affected demographics [17].

4 System Design

A literature search through IEEE Xplore, ACM digital library, and Google Scholar resulted in 14 papers being selected. These articles were obtained by a combination of keyword searching "facial recognition", "bias", and "biometrics" as well as backtracking through references of relevant research. Of these 14 papers, 6 discussed background or outcomes of bias in facial analysis and recognition software, 4 papers discussed sources of sample bias, and 4 papers discussed sources of evaluation bias. The 8 articles on sample and evaluation bias [6-7], [9], [19-23], comprise the literature review. Table 1 summarizes the reviewed articles. reviewed

4.1 Sample Bias

The articles reviewed unanimously agreed that sampling bias impacted the performance or accuracy of their respective facial analysis and recognition algorithms. However, the ways in which sampling bias affects the algorithm's outcome is dependent on its implementation and the subject demographics examined. In algorithms trained on datasets unrepresentative of minority races, verification and identification was found to be less reliable [23]. The impacts of sample bias on algorithms using a convolutional neural network (CNN) were found to be different than impacts on algorithms using principal component analysis (PCA).

Imbalanced training set data was determined to be the main source of bias in the CNN examined by Robinson et. al. [22]. When looking at the intersections of gender and race in facial recognition bias, the study found worse algorithmic performance for minorities to be correlated with their lack of representation in the training data. White males were found

Table 1. Summary of Research on Sampling and Evaluation Bias

Reference	Type of Algorithm	Subgroups Examined	Key Finding
Buolamwini et. al. [6]	Gender classification	Gender, skin color	All gender classifiers performed worst on darker skinned females and performed best on lighter skinned males
Klare et. al. [19]	Verification and identification	Gender, race, age	Commercial facial recognition software tested performed worse on female, Black, and younger subjects; all three subgroups were underrepresented in the dataset.
Grother et. al. [20]	Verification and identification	Gender, race, age, nationality	Majority of biometric systems employed fixed global threshold. High false match rates in West and East African, Asian, African American, and Native American populations across varying image qualities.
Quinn et. al. [21]	Verification and identification	Gender, age, nationality	Effectiveness of biometric traits varies across demographics.
Robinson et. al. [22]	Verification and identification	Gender, race	For facial recognition using CNN and imbalanced dataset, Asian females performed the worst and White males performed the best.
Cook et. al. [7]	Verification and identification	Gender, age, skin reflectance	Efficiency of biometric acquisition is affected by demographics, and those with glasses, lower skin reflectance, are younger, or are female have worse overall performance.
Yucer et. al. [23]	Verification and identification	Race	Augmenting imbalanced data pre-processing improved performance of DCNN facial recognition algorithm.
Furl et.al. [9]	Verification and identification	Race	Developmental contact hypothesis found to result in other-race effect in facial recognition algorithm.

to perform best while Asian females were the worst performing demographic, resulting in 5% more errors in face matching for Asian females [22]. White subjects made up 70% and 85% of the two datasets tested [22]. Most bias was found when comparing images of subjects in the same subgroup, resulting in intra-subgroup error being the highest followed by inter-subgroup error with subjects of the same gender. Imbalanced training data sets may introduce sensitive attributes such as race into a neural network. In facial recognition algorithms, removing dependency from attributes such as race disturbs information shared between important attributes, like facial features [23]. In place of removing dependency, algorithms may be augmented pre-processing to improve both accuracy and fairness [23]. Yucer et. al. improved performance of a model employing a deep convolutional neural network (DCNN) through adversarial augmentation pre-processing, increasing matching accuracy [23]. By augmenting data attributes before model training, sensitive attributes such as race become irrelevant for verification and identification. An augmented dataset was found to provide more accurate verification for Asian and African subjects than imbalanced datasets, decreasing the standard deviation from 2.91 to 2.45 [23]. Klare et. al. found that all commercial-of-the-shelf algorithms examined disproportionately performed worse on female, Black, and younger subgroups [19]. Both trainable and non-trainable models performed worse on these subgroups. In changing the training dataset to be all Black, the Spectrally Sampled Structural Subspace Features (4SF) facial recognition algorithm was nearly 2% more accurate for Black subjects. When changed to all White, accuracy increased by 1.5% for White subjects. Accuracy for younger subjects increased by nearly 2% when 4SF was trained on 18 to 30 year olds. This suggests facial recognition algorithms trained on a particular demographic perform better than an algorithm trained on a balanced or imbalanced dataset to use for all subjects.

However, there are difficulties when considering race as a valid label in facial recognition algorithms [6]. All studies concerning sample bias used demographics of subject as variables as opposed to phenotype. This has been found to result in discrepancies when considering the diversity of phenotypes in particular demographics. While categorizing subjects by race is useful for finding bias in algorithm performance post-processing, diversity in phenotype creates variation in bias within a subgroup. For example, representation within the IJB-A dataset was found to vary among Black subjects of different shades [6]. This is particularly evident with Hispanic subgroups due to the mixture of races (White European, Black, Native American) evident within this demographic. Klare et. al. attributed the diversity in the subgroup to the lower accuracy when 4SF is trained on exclusively Hispanic subjects and hypothesized that diversity in the subgroup would hinder improvements on accuracy for the demographic [23]. Additionally, all papers read used female and male labels for subgroups when analyzing for gender bias. This eliminates all representation of transgender identities and does not address any potential biases for subjects outside of cisgender female and male subgroups [6]. Due to the absence of representation in the dataset, sampling bias may affect performance for transgender subjects. The same is true for other gender identities outside of the traditional gender binary.

Accuracy for minority subjects of PCA-based facial recognition algorithms had different outcomes with imbalanced and balanced datasets. Furl et. al. examined the effect of contact hypothesis, the development hypothesis, and the non-contact hypothesis on other-race effect on facial recognition algorithms. Imbalanced datasets displayed a bias for Asian faces instead of Caucasian faces, the majority subgroup. This occurred due to the distinctiveness of Asian features compared to majority features in the set of learned faces making it easier to match subjects [9]. The lack of representation in the dataset also allowed for Asian faces to be distinct from their neighboring face space when matching [9]. Developmental contact theory was represented through the use of a PCA algorithm and a Fisher discriminant analysis (FDA) to separate faces in space, creating dense spaces where there are faces with similar features. The three models displayed the other-race effect with higher matching accuracy for Caucasian faces than Asian faces [9]. The non-contact hypothesis was investigated by ensuring representations of faces in space are not dependent on the learning history of the algorithm but resulted in mixed results with no consistent other-race effect. The impact of the other-race effect was largely impacted by the learning process for the model, as opposed to simply the representation of minorities in the dataset, as was seen in the CNN algorithms.

4.2 Evaluation Bias

The majority of facial recognition or analysis systems utilizing biometrics employ a fixed threshold for classification, verification, and identification [20]. This results in evaluation bias, as these benchmarks are not suited to all demographics as a consequence of other forms of bias or the implementation of the algorithm itself. Fixed operating thresholds eliminate the ability to consider camera model, image conditions, or demographics when matching or classifying [20]. Buolamwini et. al. determined that fixed operating benchmarks were largely responsible for the bias observed in gender classification algorithms produced by Microsoft, IBM, and Face++ [6]. The gender classifiers all had the highest error rates for darker skinned females and the lowest error rate for lighter skinned males. The *lowest* error rate for dark females was 20.8% while the *highest* error rate for lighter males was 0.8% [6]. The algorithms examined did not allow for the confidence values to be modified to account for variations dependent on demographic or phenotype. A score of 0 denotes low confidence and 1 denotes high confidence. While lighter males had near perfect confidence scores, darker females had confidence scores ranging from approximately 0.75 to 1 [6]. The lower confidence values for darker females display the evaluation bias in using a fixed threshold for confidence. This same bias affected all female and darker subjects, not only their intersection.

Fixed operating thresholds also exacerbate bias stemming from selection of biometric features. Specific biometric features may provide more information about particular demographics than others, leading to more confidence in matching and resulting in evaluation bias. For example, including eye color as a biometric would be advantageous to demographics with varying eye colors [21]. Some biometric traits have been determined to be harder to identify than others. In both trainable and non-trainable models, female subjects and Black subjects were harder to identify than their counterparts [19]. Because the non-trainable models were not influenced by bias in data, this indicates the variation in effectiveness among biometric traits. Evaluation bias occurs as the varied effectiveness of biometric traits impacts their confidence values, resulting in varied outcomes dependent on these biometric traits. Quinn et. al. determined that fixed thresholds contributed to male subjects having greater false match rates than women as well as Indian and Chinese subjects having greater false match rates than Cubans or Mexicans [21]. False match rates decreased as age increased except for those over 58, where the false match rate increased [21]. Filtering the potential matches to those only with shared biometric traits was found to increase accuracy and decrease false match rates, however more false match rates occur when comparing subjects that share the same biometric traits.

Impacts on verification or identification confidence due to image acquisition may also lead to evaluation bias. Some biometric traits have less efficient biometric acquisition, resulting in lower mated similarity scores for those traits [7]. Performance of facial recognition software was found to be strongly affected by attributes related to demographics, such as skin reflectance [7]. Cook et. al. determined that subjects with lower skin reflectance, or darker skin, had lower average mated similarity scores than those with higher reflectance [7]. Similarly, female subjects, younger subjects, and subjects with eyewear had lower average mated similarity scores [7]. This discrepancy in accuracy decreased with the use of more advanced biometric acquisition systems, reversing the impacts of biometric traits like skin reflectance and underscoring the effect of demographics on efficiency of biometric acquisition [7]. As biometric acquisition improves for particular traits, thus improving confidence scores, global thresholds will have a lesser effect on demographics with these traits.

4.3 Intersection of Sampling Bias and Evaluation Bias

Outcomes of sampling bias and evaluation bias are not isolated from one another. Training a CNN on an unrepresentative dataset leads to performance results biased by subgroup, and the distances between faces in the mappings created by a CNN vary depending on distribution of similarity scores. As a consequence of bias, these distances are different for each demographic, making it difficult to utilize a global threshold [22]. The use of fixed thresholds on mappings that vary by demographic will result in biased outcomes, as these benchmarks are not suited for minority subgroups [6]. This was seen in gender classifiers, where unrepresentative datasets were hypothesized to be the cause of confidence score variation in darker females [6].

4.4 Direct Discrimination

Bias in the reviewed articles overwhelmingly led to worse outcomes for demographics of sensitive attributes, such as race, ethnicity, nationality, gender, and age. This is consistent with definitions of direct discrimination and indicates specific demographics will be more likely to face negative outcomes from applications of facial recognition and analysis software. Of the demographics reviewed, race, nationality, and ethnicity were found to result in the greatest discrepancy in outcome. Demographics affected were mostly non-Caucasian, ranging from East and West African populations to Asian populations (South, East, and Southeast Asian). Regarding gender, 5 out of the 6 articles examining gender determined that female subjects faced worse outcomes than male subjects, with the disagreeing article displaying similar outcomes for younger females. Age had less consistent results, but across all studies reviewed middle aged subjects performed the best, indicating bias against children and elderly populations. Overall, there was sufficient evidence for direct discrimination as a consequence of sampling bias, evaluation bias, or both in all algorithms reviewed.

5 Conclusion

Sampling bias affects performance and accuracy of facial recognition and analysis algorithms. Imbalanced datasets that do not represent subjects within a demographic were found to lead to worse performance and accuracy for that demographic. In a CNN, this imbalance introduces sensitive attributes into the algorithm, at which point removing this dependency becomes difficult without altering important data [23]. In algorithms employing PCA, the distinctiveness of minority faces when utilizing an imbalanced dataset resulted in advantages for minority subjects. However, when modeling developmental contact theory in PCA training, an other-race effect was evident [9]. Use of demographics in place of phenotype generalizes the effect of imbalanced datasets on subjects within a demographic. Variation in phenotype within a demographic may indicate variation in performance within the demographic. Evaluation bias was found to be most evident in algorithms using fixed global thresholds for gender classification or facial recognition. When an algorithm employs a global threshold for confidence in classification or matching, demographics with lower confidence scores face worse outcomes. Lower confidence may stem from the use of biometric traits that do not effectively distinguish subjects of all demographics or worse performance in biometric acquisition for particular demographics. Additionally, lower confidence for specific demographics due to sampling bias may also lead to evaluation bias when utilizing fixed thresholds, resulting in worse performance and accuracy due to both forms of bias. Direct discrimination occurs as a result of these biases due to dependencies on sensitive attributes.

6 Future Work

Future studies should confirm effects of phenotypic traits closely associated with minority demographics, in addition to skin reflectance, on performance and accuracy. Additionally, future work is needed to confirm the effectiveness of phenotypic biometric traits as opposed to demographic biometric traits when examining phenotypically diverse populations. Finally, future work should address the effect of largely ignored demographics, namely more inclusive gender identities, on facial recognition and analysis software.

REFERENCES

[1] Pawel Drozdowski, Christian Rathgeb, Antitza Dantcheva, Naser Damer and Christopher Busch. 2020. Demographic Bias in Biometrics: A Survey on an Emerging Challenge. In *IEEE Transactions on Technology and Society*, vol. 1, no. 2, pp. 89-103. doi: 10.1109/TTS.2020.2992344.

[2] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine Bias: there's software used across the country to predict future criminals. And it's biased against blacks, *ProPublica*. propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

[3] Jacob Snow. 2018. Amazon's Face Recognition Falsely Matched 28 Members of Congress With Mugshots, *American Civil Liberties Union*. https://www.aclu.org/blog/privacy-technology/surveillance-technologies/amazons-face-recognition-falsely-matched-28

[4] Paul Grother, George Quinn and P J. Phillips. 2010. Report on the Evaluation of 2D Still-Image Face Recognition Algorithms. In *NIST Interagency/Internal Report (NISTIR), National Institute of Standards and Technology, Gaithersburg, MD*, https://doi.org/10.6028/NIST.IR.7709

[5] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman and Aram Galstyan. 2019. A Survey on Bias and Fairness in Machine Learning, *Cornell University*, https://arxiv.org/abs/1908.09635

[6] Joy Buolamwini and Timnit Gebru. 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency, PMLR 81:77-91.*

[7] Cynthia M. Cook, John J. Howard, Yevgeniy B. Sirotin, Jerry L. Tipton and Arun R. Vemury. 2019. Demographic Effects in Facial Recognition and Their Dependence on Image Acquisition: An Evaluation of Eleven Commercial Systems. In *IEEE Transactions on Biometrics, Behavior, and Identity Science, vol. 1, no. 1, pp. 32-41*, doi: 10.1109/TBIOM.2019.2897801.

[8] Amazon. 2020. *We are implementing a one-year moratorium on police use of Rekognition*. https://www.aboutamazon.com/news/policy-news-views/we-are-implementing-a-one-year-moratorium-on-police-use-of-rekognition

[9] Nicholas Furl, P. Jonathon Phillips and Alice J O'Toole. 2002. Face recognition algorithms and the other-race effect: computational mechanisms for a developmental contact hypothesis, In *Cognitive Science, Volume 26, Issue 6, 797-815*, https://doi.org/10.1016/S0364-0213(02)00084-8.

[10] Hoo Keat Wong, Ian D. Stephen and David R. T. Keeble. 2020. The Own-Race Bias for Face Recognition in a Multiracial Society. In *Frontiers in Psychology, 11(208)*, https://doi.org/10.3389/fpsyg.2020.00208

[11] John C. Brigham, L. Brooke Bennett, Christian A. Meissner and Tara L. Mitchell. 2007. The Influence of Race on Eyewitness Memory. In R. C. L. Lindsay, D. F. Ross, J. D. Read, & M. P. Toglia (Eds.), *The handbook of eyewitness psychology, Vol. 2. Memory for people.* Lawrence Erlbaum Associates Publishers, 257–281.

[12] P. Jonathon Phillips, Fang Jiang, Abhijit Narvekar, Julianne Ayyad, and Alice J. O'Toole. 2011. An other-race effect for face recognition algorithms. *ACM Trans. Appl. Percept*. 8, 2, Article 14 (January 2011), 11 pages. https://doi.org/10.1145/1870076.1870082

[13] Inioluwa Deborah Raji and Joy Buolamwini. 2019. Actionable Auditing: Investigating the Impact of Publicly Naming Biased Performance Results of Commercial AI Products. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society* (*AIES '19*). Association for Computing Machinery, New York, NY, USA, 429–435. https://doi.org/10.1145/3306618.3314244

[14] David Leslie. 2020. Understanding bias in facial recognition technologies: an explainer. *The Alan Turing Institute*, https://doi.org/10.5281/zenodo.4050457

[15] Ashraf Khalil, Soha Glal Ahmed, Asad Masood Khattak and Nabeel Al-Qirim. 2020. Investigating Bias in Facial Analysis Systems: A Systematic Review. In *IEEE Access, vol. 8, pp. 130751-130761*, https://doi.org/10.1109/ACCESS.2020.3006051

[16] Priyamol James, Jeena Thomas and Neena Alex. 2015. A survey on soft Biometrics and their application in person recognition at a distance. In *2015 International Conference*

on Soft-Computing and Networks Security (ICSNS), pp. 1-5, https://doi.org/10.1109/ICSNS.2015.7292416

[17] Raul Vicente Garcia, Lukasz Wandzik, Louisa Grabner and Joerg Krueger. 2019. The Harms of Demographic Bias in Deep Face Recognition Research. In *2019 International Conference on Biometrics (ICB), pp. 1-6*, https://doi.org/10.1109/ICB45273.2019.8987334

[18] ACM U.S. Technology Policy Committee. 2020. *STATEMENT ON PRINCIPLES AND PREREQUISITES FOR THE DEVELOPMENT, EVALUATION AND USE OF UNBIASED FACIAL RECOGNITION TECHNOLOGIES*. https://www.acm.org/binaries/content/assets/public-policy/ustpc-facial-recognition-tech-statement.pdf

[19] Brendan F. Klare, Mark J. Burge, Joshua C. Klontz, Richard W. Vorder Bruegge and Anil K. Jain. 2012. Face Recognition Performance: Role of Demographic Information. In *IEEE Transactions on Information Forensics and Security, vol. 7, no. 6, pp. 1789-1801*, https://doi.org/10.1109/TIFS.2012.2214212

[20] Patrick Grother, Mei Ngan, and Kayee Hanaoka. 2019. Face Recognition Vendor Test (FRVT) Part 3: Demographic Effects. In *NIST Interagency/Internal Report (NISTIR), National Institute of Standards and Technology, Gaithersburg, MD*, https://doi.org/10.6028/NIST.IR.8280

[21] George Quinn and Patrick Grother. 2008. False Matches and Non-independence of Face Recognition Scores. *In 2008 IEEE Second International Conference on Biometrics: Theory, Applications and Systems, pp. 1-5*, https://doi.org/10.1109/BTAS.2008.4699325.

[22] Joseph P Robinson, Gennady Livitz, Yann Henon, Can Qin, Yun Fu and Samson Timoner. 2020. Face Recognition: Too Bias, or Not Too Bias? In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 1-10*, https://doi.org/10.1109/CVPRW50498.2020.00008.

[23] Seyma Yucer, Samet Akçay, Noura Al-Moubayed and Toby P. Breckon. 2020. Exploring Racial Bias within Face Recognition via per-subject Adversarially-Enabled Data Augmentation. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, https://doi.org/10.1109/CVPRW50498.2020.00017