Appeared in the Proceedings of the 10th IEEE International Conference on Biometrics: Theory, Applications and Systems (IEEE BTAS), Tampa Bay, USA, September 2019

The Effect of Broad and Specific Demographic Homogeneity on the Imposter Distributions and False Match Rates in Face Recognition Algorithm Performance

John J. Howard and Yevgeniy B. Sirotin
The Maryland Test Facility
{john, yevgeniy}@mdtf.org

Arun R. Vemury
Department of Homeland Security,
Science and Technology Directorate
arun.vemury@hq.dhs.gov

Abstract

The growing adoption of biometric identity systems, notably face recognition, has raised questions regarding whether performance is equitable across demographic groups. Prior work on this issue showed that performance of face recognition systems varies with demographic variables. However, biometric systems make two distinct types of matching errors, which lead to different outcomes for users depending on the technology use case. In this research, we develop a framework for classifying biometric performance differentials that separately considers the effect of false positive and false negative outcomes, and show that oft-cited evidence regarding biometric equitability has focused on primarily on false-negatives. We then correlate demographic variables with false-positive outcomes in a diverse population using a commercial face recognition algorithm, and show that false match rate (FMR) at a fixed threshold increases >400-fold for broadly homogeneous groups (individuals of the same age, same gender, and same race) relative to heterogeneous groups. This was driven by systematic shifts in the tails of the imposter distribution impelled primarily by homogeneity in race and gender. For specific demographic groups, we observed the highest false match rate for older males that self-identified as White and the lowest for older males that self-identified as Black or African American. The magnitude of FMR differentials between specific homogeneous groups (<3-fold) was modest in comparison with the FMR increase associated with broad demographic homogeneity. These results demonstrate the false positive outcomes of face recognition systems are not simply linked to single demographic factors, and that a careful consideration of interactions between multiple factors is needed when considering the equitability of these systems.

1. Introduction

Machine learning algorithms are increasingly being used in ways that affects people's lives. Consequently, it is important that these systems are not only accurate when executing their given task but *equitable*, i.e. have fair outcomes for all people. Face recognition technology leverages machine learning algorithms to compute the similarity score between photos of people's faces. This similarity score can be used to find the identity associated with a photo based on its similarity to a gallery of photos with known identities. This process is called identification. Face similarity scores can also be used to verify identity claims by measuring the similarity of a new photo with a photo associated with a claimed identity. This process is called verification. Algorithm performance in these tasks depends critically on the difference between distributions of scores generated when comparing photos that belong to the same person (*mated* scores) versus scores generated when comparing photos of different people (*non-mated* scores). Match and no-match decisions made by comparing this score to a set threshold can result in impactful identity decisions.

Prior to the advent of machine vision, face recognition tasks were solely performed by people using dedicated neural networks for face processing [8]. However, since the 1960s [1] computers have steadily increased their face recognition performance such that artificial neural networks have recently surpassed human performance on some difficult recognition tasks [18]. Automation of face recognition has spurred its burgeoning use in both commercial and government applications. With increasing use of the technology it is important to consider whether it performs equitably. Indeed, there is significant public interest in this issue with the popularity of the search term "facial recognition bias" increasing steadily in the United States since 2013, and clearly skewing higher than historical averages in 2017, 2018, and in projections for 2019 (Figure 1). There has also been an associated increase in the number of popular news articles reporting on the topic of bias in face recognition

978-1-7281-1522-1/19/$31.00 ©2019 IEEE

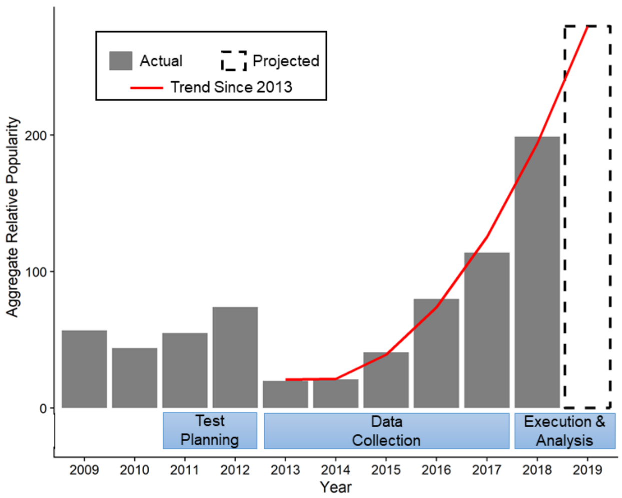


Figure 1. Aggregate relative popularity of the search term "facial recognition bias" in the United States according to Google Trends (as of March 2019) overlaid on planning, collection, execution, and analysis time frames for our study.

systems since 2016 [24, 16, 6, 25, 2, 23, 12].

Despite this rising public awareness and media attention, the term "bias" as it applies to biometric systems is not well defined. The term bias itself is vastly overloaded with, often different connotations in statistics, computer science, psychology, and historical public discourse. In machine learning specifically, the sources of bias can vary, from the historical processes used to generate data, to the selection criteria used to collect data, to the evaluation criteria used to optimize a particular model [22]. These distinctions are critically important when discussing how to address bias in biometric systems.

Our group has been studying demographic factors in biometric performance since 2010 and, since 2011, has been planning and collecting data in order to report on how these factors affect the performance of facial recognition systems (Figure 1). The initial results of this effort were published in [4] . Here we extend this work and propose a language and quantitative framework for measuring the equitability of biometric identity systems. We base this framework on the concept of differential performance across demographics, and show how these may lead to separate differential outcomes for mated and non-mated face comparisons (Section 2.2). Using this framework, we show that prior studies of biometric equitability, including our own, have focused largely on *mated* or *genuine* score distributions, and how these lead to false rejections. We then proceed to use our framework to measure how differences in similarity scores for *non-mated* or *imposter* comparisons across demographics lead to false matches using a commercial face recognition algorithm.

2. Background and Methodology

2.1. Biometric Error Rates and Their Causes

Biometric samples are inherently noisy. They come from biological systems that are subject to physiological, behavioral, and environmental changes. Biometric recognition systems are tasked with evaluating the correlations between pairs of biometric samples, and establishing if they are "similar enough" to be declared from the same biometric source. This operation necessarily involves a decision or discrimination threshold that defines the degree of similarity required to declare a pair of samples as matching or non-matching. A false non-match (FNM) error occurs when a mated pair of biometric samples is found to exhibit similarity that is less than a given system's decision threshold.

Biometric samples also inevitably share some common patterns. They come from biological systems that share a certain degree of genetic and random similarity. A false match (FM) error occurs when a non-mated pair of biometric samples is found to exhibit similarity that is greater than a given system's decision threshold. The rates at which FNM and FM errors occur are dictated by the degree to which the genuine and imposter distributions overlap and the location of the decision threshold (Figure 2). Importantly, the relative position of both genuine and imposter distributions can be different when considering comparisons between individuals belonging to different demographic groups.

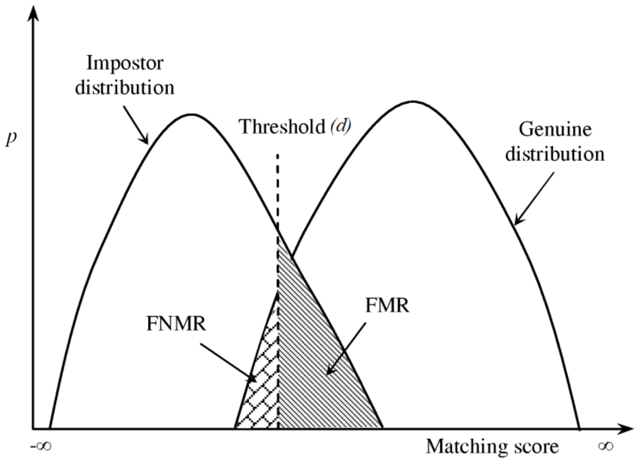


Figure 2. The relationship between genuine distributions, imposter distributions, decision threshold, false non-match rate (FNMR) and false match rate (FMR) in biometric systems. Figure modified from [14].

.

2.2. Biometric Equitability Framework

The equitability of a biometric system must be considered in the context of specific biometric task being performed. To facilitate discussion, we introduce the following set of terms:

- **Differential Performance.** We define differential performance as a difference in the genuine or imposter distributions between specific demographic groups independent of any decision threshold. This is closely related to the concept of "biometric menagerie", a phenomena in which subject-specific genuine and imposter distributions are statistically different [5, 26, 19, 21]. Differential performance is this same effect, not for specific subjects, but for specific demographic groups.

- **Differential Outcome.** We define differential outcome as a difference in FM or FNM rates between different demographic groups relative to a decision threshold. Similarity scores in and of themselves are not the outcome of an identity decision. They must be be re-cast to match/no-match decisions using a decision threshold. These match decisions can then can be used to calculate FM and FNM error rates.

- **False Negative Differential.** We define the term False Negative Differential to mean a greater or lesser tendency for one demographic group to experience a false negative error relative to another group. That is, a tendency of the group member to fail to be identified as themselves.

- **False Positive Differential.** We define the term False Positive Differential to mean a greater or lesser tendency for one demographic group to experience a false positive error relative to another group. That is, a tendency to mistake the group member for somebody else.

The concepts of False Negative and False Positive Differentials helps frame the effect a biometric system can have on individuals when performing a specific task. For example, in a law enforcement scenario, the False Negative Differential Outcome is that a bad actor is *not* identified, and therefore incorrectly *not* investigated. The False Positive Differential Outcome is that an innocent is mistaken for a bad actor and, therefore incorrectly stopped or investigated. In both examples, these differentials are undesirable but the outcomes themselves are very different.

2.3. Equitability Framework and Prior Work

Uniformly, the media articles referenced in Section 1 cite a 2016 Georgetown study, "The Perpetual Lineup" which states as one of its main findings that "Face Recognition Algorithms Exhibit Racial Bias" [9]. This report, in turn, cites the academic study [15], which states that "Several leading algorithms performed worse on African Americans, women, and young adults than on Caucasians". However, this study looked at false non-match rate (FNMR) at a specific false match rate (FMR) and decision threshold to find "lower matching accuracies on the same cohorts (females, Blacks and age group 18 to 30)". This is an example of False Negative Differential Outcome. As pointed out in the Perpetual Lineup report, in a law-enforcement context, this finding means that African Americans present in a law-enforcement gallery are actually *less likely* to be identified compared to Caucasian individuals present in the same gallery. This study [15] does not show that African Americans are more likely to be mistaken for others.

Our group as well has previously investigated the effect of demographic factors on biometric performance of eleven commercial biometric systems [4] . Among other effects, we found that mated match scores for people with lower skin reflectance tended to be lower than mated match scores for people with higher skin reflectance. This is another clear example of False Negative Differential Performance, which may or may not manifest as False Negative Differential Outcome depending on the camera technology utilized. Our study also suggests that individuals in a law-enforcement gallery that have lower skin reflectance are *less likely* to be identified using face recognition than individuals with higher skin reflectance present in the same gallery. Our study does not show that individuals in a with lower skin reflectance are more likely to be mistaken for others.

However, False Positive Differential Performance and Outcome are legitimate issues, and must be examined separately. Facial structure, skin tone, and anatomical features are all genetic traits, meaning they are more likely to be shared by those that share genomic background, such as those with a similar ancestry. Recently, large scale tests of face recognition algorithms have supported this notion showing higher FMR for individuals from specific countries [11]. If some demographic groups are notably more likely than others to be erroneously matched during a facial recognition gallery searches (False Positive Differential) it would raise reasonable concerns regarding the use of that technology.

Finally, we note there is a large existing body of work regarding disparate impact, discrimination analysis, and fairness [20, 7]. While most of these methods were not developed specifically for facial recognition applications, they can still be applied in some cases. A full review of the various methods and the issues associated with their use is outside the scope of this research. However, one specific definition of fairness, known as the "four-fifths" rule has been advocated by the United States Equal Employment Opportunity Commission (EEOC) [3]. This rule, typically applied to certify selection rates in hiring are non-discriminatory, states that rates across all groupings should be within 80%, or four-fifths, of the highest group rate. For example, if female applicants are hired at a rate of 20%, male applicants should be hired at a rate no less than 16% and vice-versa. We will reference this specific definition of fairness in our investigation of False Positive Differential Outcomes.

2.4. Test Design

We examined False Positive Differential Performance and Outcome using face biometric samples collected during the 2018 U.S. Department of Homeland Security, Science and Technology Directorate (DHS S&T) Biometric Technology Rally . The methods for data collection have been published elsewhere [13] [4] . Briefly, face images were collected from 363 test subjects, diverse in age, gender, and race within a controlled environment. For the purposes of this study, we limited our sample to subjects who self-reported as White (W), Black or African American (B), male (M), and female (F) (Figure 3). Face images were identified against a historic gallery of face samples using a leading commercial face recognition algorithm tested in [11]. Here we examine the non-mated biometric comparisons performed as part of these identifications, together with the self-reported subject demographics to measure imposter distributions and false positive rates.

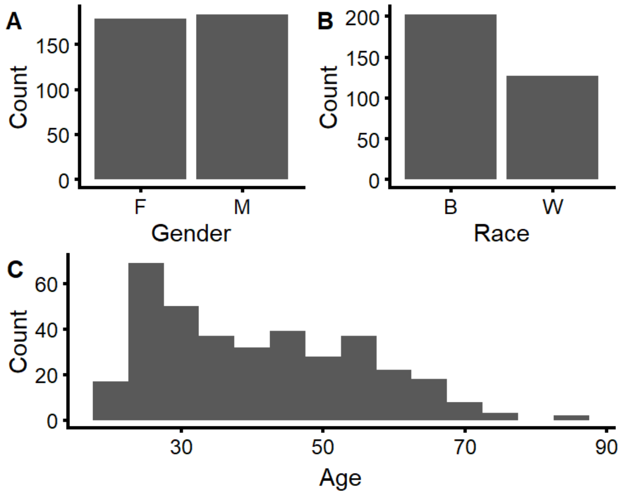


Figure 3. Distributions of the demographic variables self-reported by test subjects. **A.** Distribution of subject gender: (F) female; (M) male. **B.** Counts of subjects identifying with each racial category: (B) Black or African American; (W) White; **C.** Distribution of test subject ages.

2.5. Subject-Specific 99th Percentile Non-Mated Score as a Measure of Differential Performance

Figure 2 shows that biometric error rates are driven by the behavior of the tails of the genuine and imposter distributions. Consequently, in this research, we quantify shifts in the 99th percentile score of the imposter distribution. We measure the 99th percentile non-mated score separately for each individual in our dataset for reference galleries composed of individuals with homogeneous or heterogeneous demographics. For age, subjects who were within ±10 years of each other were considered homogeneous or "same" and vice versa. Gender and race similarity was defined according to the categories in Figure 3. In Equation 1, $S_{99,m}$ is the subject-specific 99th percentile non-mated score, $\mathcal{I}_{(n)}(m)$ is the ordered set of non-mated similarity scores for subject m, and $n = \lceil .99 * |\mathcal{I}| \rceil$.

$$S_{99,m} = \mathcal{I}_{(n)}(m) \tag{1}$$

2.6. Conditional Probability as a Measure of Differential Outcome

Conditional probability is a parsimonious technique to measure differential outcome. It is the measure of the probability of some event A given some condition B, denoted $P(A \mid B)$. Using this method, [15] showed that:

$$P(\text{FNM} \mid \text{G} \in \text{F}) > P(\text{FNM} \mid \text{G} \in \text{M})$$
$$P(\text{FNM} \mid \text{R} \in \text{B}) > P(\text{FNM} \mid \text{R} \in \text{W})$$
$$P(\text{FNM} \mid \text{A} \in [18{,}30] > P(\text{FNM} \mid \text{A} \in [30{,}50])$$
$$P(\text{FNM} \mid \text{A} \in [18{,}30] > P(\text{FNM} \mid \text{A} \in [50{,}70])$$

where G is gender, which has a value of male (M) or female (F), R is race, which has a value of Black or African American (B) or White (W) and A is age, which has a range of 18 to 30, 30 to 50, or 50 to 70.

This study will use the methods discussed in Sections 2.5 and 2.6 to explore False Positive Differentials across both broad homogeneous groups (same race, same gender, same age) and demographically specific homogeneous groups (white, males, older, etc.). When approaching the latter, the possible number of demographic combinations multiplies rapidly and the chosen subsetting order can mask datapoints. For example, we never calculate the FMR of all females compared to all males if we first subset our population by race. Therefore, to intelligently select this ordering, we use the concept of Shannon Entropy to quantify the amount of information gained about FMR by knowing the demographic labels of race, gender, and age. Information gain is the change in entropy from a prior state $E(\text{T})$ to a state where some information is known $E(\text{T} \mid \text{X})$. It is defined in Equations 2 & 3, where p_i is the probability of being in state $i \in \{\textit{false match, true no-match}\}$.

$$IG(\text{T}, \text{X}) = E(\text{T}) - E(\text{T} \mid \text{X}) \tag{2}$$

$$E(\text{T}) = -\sum_{i} = p_i \log_2(p_i) \tag{3}$$

This technique also produces the exact FMR for each specific demographic subgroup, allowing us to quantify the

False Positive Differential Outcome of the face recognition algorithm on different demographic subgroups.

3. Results

3.1. The Effect of Broad Demographic Homogeneity on Face Recognition

Prior work has shown that comparisons of faces belonging to similar demographic groups produce higher similarity scores than faces that belong to different demographic groups [11][10]. We examined how the imposter distribution generated by our algorithm varied when comparing demographically heterogeneous (different race, gender, and age) versus demographically homogeneous individuals (same race, gender, and age; see 2.5). Figure 4 shows that $S_{99,m}$ increases steadily with increasing homogeneity. Indeed, we see a nearly twofold increase in the average $S_{99,m}$ from 0.193 when comparing demographically heterogeneous individuals to 0.373 when comparing demographically homogeneous individuals.

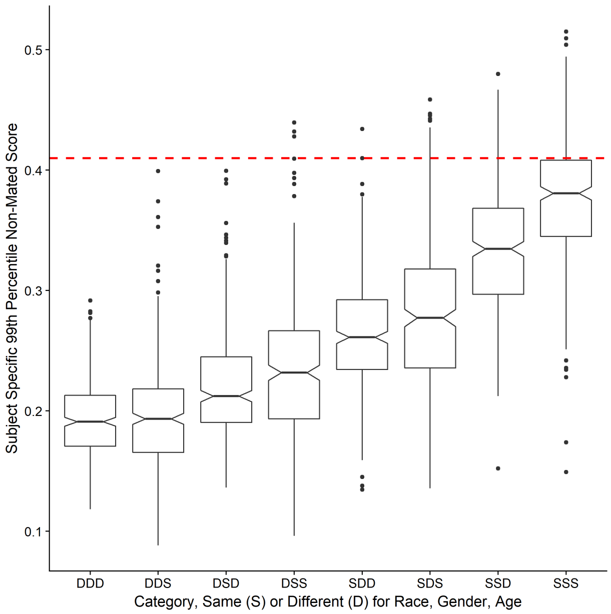


Figure 4. Distributions of the 99th percentile subject-specific non-mated scores across broad homogeneous versus heterogeneous race, gender, and age categories.

The movement in the tails of the imposter distributions observed in Figure 4 for the more homogeneous groups will result in greater FMR at some decision thresholds. This is reflected in Table 1, which shows the FMR at a hypothetical threshold of 0.41 as a function of broad demographic homogeneity. From Table 1 we ascertain that demographically homogeneous individuals in our study are 428 times more likely to match each other than demographically heterogeneous individuals. Table 1 shows that some demographic variables have a larger effect on FMR. Taking the FMR for demographically heterogeneous individuals as baseline, we see that limiting comparisons to those of the same age results in a 1.3-fold increase in FMR, limited to those of the same gender results in a 9-fold increase in FMR, and limited to those of the same self-reported race results in a 25-fold greater FMR. These homogeneity effects are not strictly linear. For example, taking the age, gender, and race multipliers of 1.3, 9 and 25, we would naively expect FMR for individuals sharing these traits to be $1.3 * 9 * 25 = 292.5$-fold greater for heterogeneous individuals, under-estimating the true FMR by nearly a third.

Table 1. False match rate at a threshold of 0.41 across homogeneous versus heterogeneous race, gender and age categories.

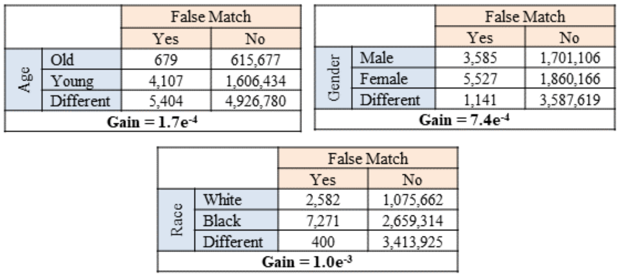
Race	Gender	Age	FMR	Multiplier
Different	Different	Different	1.7e-5	1
Different	Different	Same	2.3e-5	1.3
Different	Same	Different	1.6e-4	9
Different	Same	Same	3.3e-4	19
Same	Different	Different	4.3e-4	25
Same	Different	Same	8.3e-4	49
Same	Same	Different	2.8e-3	162
Same	Same	Same	7.3e-3	428

3.2. The Effect of Specific Demographic Homogeneity on Face Recognition

Figure 4 and Table 1 show that broad demographic homogeneity alters the imposter distribution so as to increase FMR, with the largest FMR increases for individuals of similar race, followed by gender, and finally age. We next examine whether these effects are different for specific demographic groups within our sample (i.e. False Positive Differentials) using the information gain metric, discussed in Section 2.6. The specific demographic groups we consider are individuals who self-identified their race (R) as either White (W), or Black or African American (B), individuals who self-identified their gender (G) as male (M) or female (F), and individuals who have an age difference of less than 10 years and were older than 40 (Old or O) or who have an age difference of less than 10 years and were younger than 40 (Young or Y).

Comparing all probe images to all non-mated historic gallery images (see Section 2.4) produced a set of ~ 7.2 million non-mated comparisons between 52,020 non-mated subject pairs. 10,253 of these comparisons, from 1690 non-mated subject pairs produce a similarity score of greater than 0.41, yielding a FMR at this threshold for the entire population of $1.4e^{-3}$. This FMR (between 1 in a 1000 and 1 in 500) is reasonable for an operational system with a low imposter incidence rate (such as office access control). Per Equation 3, the entropy of this full set is $1.6e^{-2}$.

To determine which covariate (race, gender, or age) provided the most information about false match rate, we calculated the information gain by splitting the full set of comparisons by each covariate in turn (see 2.6). Figure 5 shows that information gain for race is greater than the information gain for any of the other covariates.



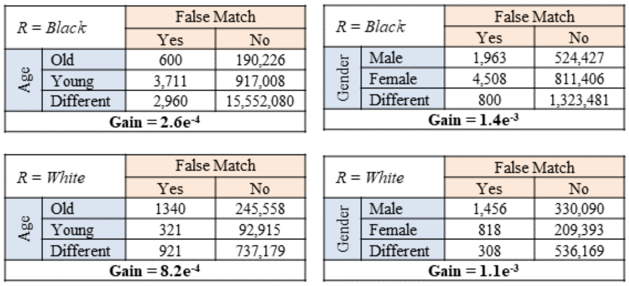
		False Match	
		Yes	No
Age	Old	679	615,677
	Young	4,107	1,606,434
	Different	5,404	4,926,780
Gain = 1.7e^{-4}			

		False Match	
		Yes	No
Gender	Male	3,585	1,701,106
	Female	5,527	1,860,166
	Different	1,141	3,587,619
Gain = 7.4e^{-4}			

		False Match	
		Yes	No
Race	White	2,582	1,075,662
	Black	7,271	2,659,314
	Different	400	3,413,925
Gain = 1.0e^{-3}			

Figure 5. Information gain starting from the full set of non-mated comparisons for the age, gender and race covariates

After subsetting the full set by race, the information gain of next subsetting by either gender or age was individually calculated. Figure 6 shows that splitting the racially homogeneous subsets next by gender yields a greater information gain than splitting by age.



R = Black		False Match	
		Yes	No
Age	Old	600	190,226
	Young	3,711	917,008
	Different	2,960	15,552,080
Gain = 2.6e^{-4}			

R = Black		False Match	
		Yes	No
Gender	Male	1,963	524,427
	Female	4,508	811,406
	Different	800	1,323,481
Gain = 1.4e^{-3}			

R = White		False Match	
		Yes	No
Age	Old	1340	245,558
	Young	321	92,915
	Different	921	737,179
Gain = 8.2e^{-4}			

R = White		False Match	
		Yes	No
Gender	Male	1,456	330,090
	Female	818	209,393
	Different	308	536,169
Gain = 1.1e^{-3}			

Figure 6. Information gain starting from racially homogeneous subsets of non-mated comparisons for age and gender

Figure 7 shows the full entropy based classification tree. Figure 7 also shows the FMR at each stage of the tree. In our sample, the FMR for subjects who identified as Black or African American is similar to FMR for subjects who identified as White ($2.7e^{-3}$ vs. $2.3e^{-3}$, respectively). At the next level of the tree, the FMR is highest for females who identified as Black or African American ($5.5e^{-3}$), while the lowest FMR is observed for males who identified as Black or African American ($3.7e^{-3}$). Finally, for demographically homogeneous comparisons, FMR was highest for males who identified as White and were similarly aged, over 40 ($9.0e^{-3}$), and the lowest FMR was for males who identified as Black or African American and were similarly aged, over 40 ($3.5e^{-3}$).

Some of these error rate spreads, such as those for Black or African American and White subjects, meet the EEOC definition of fairness discussed in Section 2.3. However, the error rates for other groups, such as the FMR rate for all Females ($3.0e^{-3}$) and all Males ($2.1e^{-3}$) do not meet this particular definition of fairness (rates calculated from Figure 5).

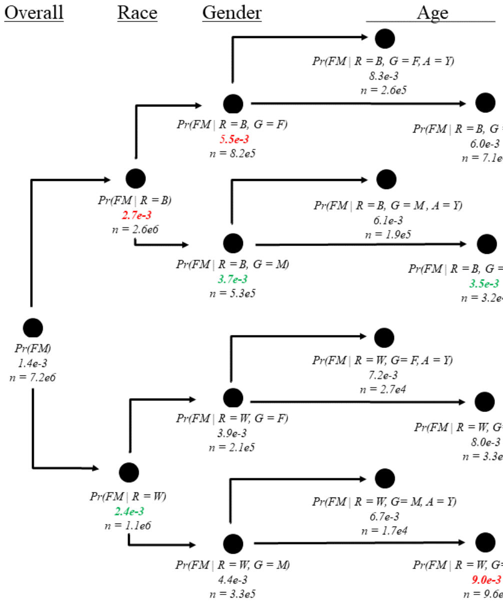


Figure 7. Entropy based classification tree showing the different false match rates across demographic groups. The worst, i.e. highest, false match rates per demographic subset are highlighted red. The best false match rates are highlighted green. The number of comparisons, i.e. the denominator in the false match rate calculation, is also shown for each node in the tree. Race (R), gender (G), and age (A) demographic covariates are reported by Black or African American (B) vs. White (W), male (M) vs. female (F), and similarly aged old (O) vs. similarly aged young (Y), respectively.

Our observation of highest FMR for older White males and lowest FMR for older Black or African American males was surprising in that it was not expected from prior work [10] which showed positive shifts in the imposter distributions for Black males relative to White males. However, this prior work examined only younger individuals (aged <40) whereas our population was more diverse in age (Figure 3). Previous work showed that the absolute value of continuous variables, such as age, can have a strong effect on biometric performance [17]. We therefore examined the effect of age, and consistency with prior work, by examining separately the shifts in the imposter distribution for different homogeneous groups. Figure 8 resolved the discrepancy showing that the imposter distributions shifted higher for homogeneous groups of young (aged <40) individuals identifying as male and Black or African American relative to groups of old (aged >40) individuals identifying as male and Black or African American. Surprisingly, this effect was reversed in individuals identifying as male and White, with the imposter distribution shifted higher in older individuals compared to the younger homogeneous group, consistent with the observed FMR.

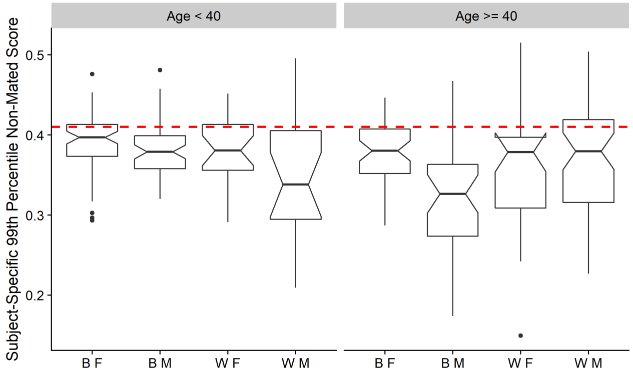


Figure 8. Distributions of the 99th percentile subject-specific non-mated scores across specific homogeneous demographic categories. Black or African American (B), White (W), Male (M), and Female (F) labels are shown along the y axis. Chart is faceted by Young (Age <40) and Old (Age >= 40).

4. Conclusions

This work presents a framework for understanding Differential Performance and Differential Outcome in face recognition across different demographic groups. Critically, it introduces the notion that there are different types of differential measures. We show that oft-cited academic work in this area mostly overlooks False Positive Differential, a tendency for certain demographic groups to be falsely matched. We examined False Positive Differential Performance and Outcome in homogeneous groups broadly (same age, same gender, same race) and within specific homogeneous groups (similarly aged and older, white, male, etc.). Our outcomes are as follows. First, we show that homogeneity in a given population can have an aggregate two order of magnitude (˜400x) effect on FMR, and that this FMR increase is driven primarily by homogeneous race (˜25x) and gender (˜10x), with a more modest contribution of homogenous age (˜1.3x). Second, we show that FMR is generally similar across specific homogeneous groups in our sample, varying at most 3x between the highest FMR, observed for older males self-identifying as White, and the lowest FMR, observed for older males self-identifying as Black or African American. We highlight that some of these variations in error rates actually meet the U.S. Equal Employment Opportunity Commission definition of fairness, while others do not. Finally, we show that False Positive Differentials for the tested biometric algorithm showed a reversal in its association with race. For younger individuals, median 99th percentile subject-specific non-mated scores and FMR were highest for females self-identifying as Black or African American, but for older individuals, scores and FMR were highest for males self-identifying as White. This shows that False Positive Differentials have a complex relationship with specific demographic groups such that effects are not strictly tied to one variable, such as race, but to an interaction between, at least, race, gender, and age.

We hope these contributions, particularly the introduction of False Positive and False Negative Differential terms, leads to a withdrawal of the ambiguous term "bias" from public discourse when discussing demographic effects in facial recognition. As outlined, this term is not descriptive enough to properly describe the problem, and carries with it certain social connotations. We also hope the notion that, in this system, false positive error rates for specific demographic groups were generally similar, can contribute to the ongoing conversations regarding the equitability of facial recognition systems.

Much future work in this area is needed. First, this effect needs to be explored using a variety of algorithms, datasets, and populations. Our test population notably did not include statistically significant sample of subjects that self-identified as Asian, or as Hispanic or Latino. Second, these error rates and the concepts of False Positive and Negative Differential Outcomes need to be explored more carefully under a variety of different fairness models. We believe ongoing work in the broader area of Artificial Intelligence fairness will be helpful in this regard. Finally, it has recently been shown that phenotypic measures offer a superior explanation of demographic effects in face recognition [4]. Future modeling using continuous variables, such as face morphology or skin reflectance, will likely provide a more complete account False Positive Differentials.

Acknowledgements

This research was funded by the Department of Homeland Security, Science and Technology Directorate on contract number W911NF-13-D-0006-0003. The views presented here are those of the authors and do not represent those of the Department of Homeland Security, the U.S. Government, or their employers. The data were acquired using the IRB protocol "Development and Evaluation of Enhanced Screening" number 120180237, approved by New England IRB.

The authors thank the following staff at SAIC's Identity and Data Sciences Lab (IDSL) for their contributions to this report: Andrew Blanchard, Kirsten Huttar, Michael Odio, and Roland Bell for providing software engineering support; Laura Rabbitt and Nelson Jaimes for human factors support; Kevin Slocum, Frederick Clauss, Gary Lake, Jeffrey Chudik, and Brian Glatfeltner for providing network and integration engineering support; Jacob Hasselgren for directing Rally execution; Rebecca Rubin for technical document support and editing; Cynthia Cook for statistics support; as well as Colette Bryant and Rebecca Duncan for support in Rally organization and execution. The authors thank Jerry Tipton and Patty Hsieh for their broad support as well as review and comment on this manuscript.

The paper authors acknowledge the following author contributions: John J. Howard conceived the work, developed and performed analyses, and wrote the paper; Yevgeniy B. Sirotin conceived the work, developed and performed analyses, and edited the paper; Arun Vemury conceived the work and edited the paper.

References

[1] W. W. Bledsoe. The model method in facial recognition. *Panoramic Research Inc., Palo Alto, CA, Rep. PR1*, 15(47):2, 1966. 1

[2] J. Buolamwini. When the robot doesnt see dark skin. *New York Times (Jun 2018). https://www. nytimes. com/2018/06/21/opinion/facial-analysis-technology-bias. html*, 2018. 2

[3] E. E. O. Commission et al. Uniform guidelines on employee selection procedures. *Fed Register*, 1:216–243, 1990. 3

[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*, pages 1–1, 2019. 2, 3, 4, 7

[5] G. Doddington, W. Liggett, A. Martin, M. Przybocki, and D. Reynolds. Sheep, goats, lambs and wolves: A statistical analysis of speaker performance in the nist 1998 speaker recognition evaluation. Technical report, DTIC Document, 1998. 3

[6] K. Draper. ”madison square garden has used face scanning technology on customers”. *New York Times, March, 2018*, page 2018, 2018. 2

[7] M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian. Certifying and removing disparate impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 259–268. ACM, 2015. 3

[8] W. Freiwald, B. Duchaine, and G. Yovel. Face processing systems: from neurons to real-world social perception. *Annual Review of Neuroscience*, 39:325–346, 2016. 1

[9] C. Garvie. *The perpetual line-up: Unregulated police face recognition in America*. Georgetown Law, Center on Privacy & Technology, 2016. 3

[10] P. Grother. Demographic effects in face recognition. Technical report, National Institute of Standards and Technology, Nov 2018. https://nigos.nist.gov/ifpc2018/presentations/17_grother_demographics.pdf, last accessed on 03/24/18. 5, 6

[11] P. Grother, M. Ngan, and K. Hanaoka. Ongoing face recognition vendor test (frvt) part 1: Verification. Technical report, National Institute of Standards and Technology, Apr 2018. https://www.nist.gov/sites/default/files/documents/2018/04/03/frvt_report_2018_04_03.pdf, last accessed on 06/07/18. 3, 4, 5

[12] D. Harwell. ”amazon facial-identification software used by police falls short on tests for accuracy and bias, new research finds”. *Washington Post, January, 2019*, page 2019, 2019. 2

[13] J. J. Howard, A. J. Blanchard, Y. Sirotin, J. Hasselgren, and A. Vemury. An investigation of high-throughput biometric systems: Results of the 2018 department of homeland security biometric technology rally. In *IEEE*, 2018. 4

[14] A. K. Jain, A. Ross, S. Prabhakar, et al. An introduction to biometric recognition. *IEEE Transactions on circuits and systems for video technology*, 14(1), 2004. 2

[15] B. F. Klare, M. J. Burge, J. C. Klontz, R. W. V. Bruegge, and A. K. Jain. Face recognition performance: Role of demographic information. *IEEE Transactions on Information Forensics and Security*, 7(6):1789–1801, 2012. 3, 4

[16] S. Lohr. ”facial recognition is accurate, if youre a white guy”. *New York Times, February, 2018*, page 2018, 2018. 2

[17] D. Michalski, S. Y. Yiu, and C. Malec. The impact of age and threshold variation on facial recognition algorithm performance using images of children. In *2018 International Conference on Biometrics (ICB)*, pages 217–224. IEEE, 2018. 6

[18] P. J. Phillips, A. N. Yates, Y. Hu, C. A. Hahn, E. Noyes, K. Jackson, J. G. Cavazos, G. Jeckeln, R. Ranjan, S. Sankaranarayanan, et al. Face recognition accuracy of forensic examiners, superrecognizers, and face recognition algorithms. *Proceedings of the National Academy of Sciences*, 115(24):6171–6176, 2018. 1

[19] N. Poh and J. Kittler. A methodology for separating sheep from goats for controlled enrollment and multimodal fusion. In *Biometrics Symposium, 2008. BSYM'08*, pages 17–22. IEEE, 2008. 3

[20] A. Romei and S. Ruggieri. A multidisciplinary survey on discrimination analysis. *The Knowledge Engineering Review*, 29(5):582–638, 2014. 3

[21] A. Ross, A. Rattani, and M. Tistarelli. Exploiting the doddington zoo effect in biometric fusion. In *Biometrics: Theory, Applications, and Systems, 2009. BTAS'09. IEEE 3rd International Conference on*, pages 1–7. IEEE, 2009. 3

[22] H. Suresh and J. V. Guttag. A framework for understanding unintended consequences of machine learning. *arXiv preprint arXiv:1901.10002*, 2019. 2

[23] J. Teicher. ”what do facial recognition technologies mean for our privacy?”. *New York Times, July, 2018*, page 2018, 2018. 2

[24] D. Victor. ”study urges tougher oversight for police use of facial recognition”. *New York Times, October, 2016*, page 2016, 2016. 2

[25] N. Wingfield. Amazon pushes facial recognition to police. critics see surveillance risk. *New York Times, May, 2018*, 22:2018, 2018. 2

[26] N. Yager and T. Dunstone. Worms, chameleons, phantoms and doves: New additions to the biometric menagerie. In *Automatic Identification Advanced Technologies, 2007 IEEE Workshop on*, pages 1–6. IEEE, 2007. 3