Demographic Disparities in 1-to-Many Facial Identification

Aman Bhatta[1], Gabriella Pangelinan[2], Michael C. King[2], Kevin W. Bowyer[1]
[1]University of Notre Dame, [2]Florida Institute of Technology

Abstract

Demographic disparities in face recognition accuracy have received negative coverage in the media. Concern about the impact of accuracy disparities often focuses on instances of false arrest after 1-to-many facial identification. This work investigates accuracy disparities in 1-to-many matching across male / female and African-American / Caucasian demographics. We compare the accuracy of rank-one 1-to-many facial identification across demographics using the d' separation between mated and non-mated score distributions, using a non-parametric value for the separation in the tails of the mated and non-mated distributions, and the distributions of (mated - non-mated) difference values. Our results indicate that rank-one 1-to-many identification accuracy is best for African-American male, followed by Caucasian male, Caucasian female and African-American female. Females have lower accuracy than males, and the accuracy difference between African-American male and Caucasian male is the smallest of the differences.

1. Introduction

Face recognition technology has received considerable negative media attention regarding demographic bias. Media coverage often points to females and African-Americans as having lower accuracy from face recognition [27, 24, 39, 35]. Even publications like *Nature* have indulged in sensational headlines of the type - "Is facial recognition too biased to be let loose?" [11].

Unequal face recognition accuracy across demographic groups is a current hot topic in the research community. Researchers are analyzing possible causes of unequal accuracy [6, 3, 8, 41, 10] and proposing approaches to mitigate differences [32, 12, 34, 38, 37, 40, 31]. Most of this research has focused on analyzing accuracy differences in 1-to-1 matching; that is, the use of face recognition for identity verification. Face recognition is also used for 1-to-many matching. In 1-to-many matching, an image of an unknown person (a "probe") is matched against enrolled images of many known persons (the "gallery") in order to find a candidate identity(ies) for the unknown person. One style

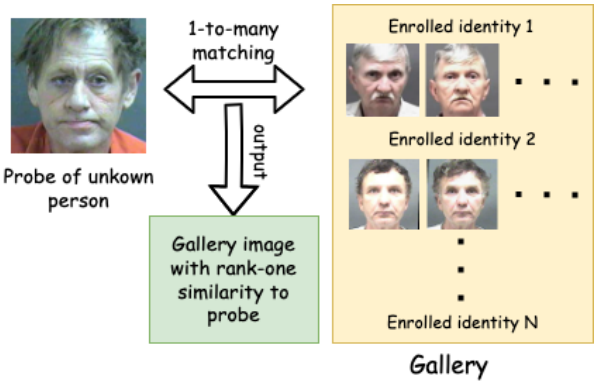


Figure 1: Does 1-to-many rank-one match error rate vary across demographics? In 1-to-many identification, an image of a person with unknown identity (the probe) is matched against a list of persons with known identity (the gallery) to find a candidate identity.

of 1-to-many matching returns the gallery image with the greatest similarity to the probe, the "rank-one match". 1-to-many matching can also be configured to return the top N matches, and / or to return only matches above some threshold similarity.

Media coverage about bias in face recognition is often concerned with instances of false arrest in which face recognition was (mis)used [18, 33, 36]. In these instances, 1-to-many facial identification search is used to find a candidate identity for the person in a probe image. Thus it is important to understand how accuracy varies across demographics in 1-to-many search.

Contributions of this work include the following:

- Using a state-of-the-art instance of ArcFace and the MORPH dataset, we compare 1-to-1 verification accuracy and 1-to-many identification accuracy across female / male and African-American / Caucasian demographics, and find that the ranking of demographics for 1-to-many accuracy is different than the ranking for 1-to-1 accuracy.

- We provide an interpretation of the effects that (1) increasing the number of enrolled identities decreases one-to-many accuracy and (2) increasing the number

of images per identity increases one-to-many accuracy, in terms of the changes in the mean and standard deviation of the non-mated and mated distributions.

- We evaluate 1-to-many accuracy across demographics based on (1) d' between non-mated and mated rank-one score distributions, (2) a non-parametric measure of separation of the distributions in the tails where false positive identifications occur, and (3) the distribution of (mated - non-mated) difference in similarity scores. The ranking of demographic groups is the same across the three approaches, but the analysis suggests that using d' alone could miss some subtle differences.

- Our results indicate that (1) the larger demographic differences are between females and males, for both African-American and Caucasian, (2) that African-American males have accuracy at least as high as any of the other three groups, and (3) African-American females have the lowest accuracy.

2. Literature Review

For a broad overview of issues related to demographic bias in biometrics, see the survey by Drozdowski et. al. [17]. We briefly touch on related works in two areas: (1) demographic differences in face recognition accuracy and (2) 1-to-many matching for face recognition.

The earliest observation that face recognition accuracy varies across demographic groups may be the 2002 Face Recognition Vendor Test [29], which found lower accuracy for females than for males. The well-known study by Klare et al [25] analyzed results from multiple algorithms and found that, "The female, Black, and younger cohorts are more difficult to recognize for all matchers used in this study (commercial, nontrainable, and trainable)". It is particularly interesting that [25] found that *training on demographic-balanced training data did not balance accuracy across demographics*. These papers [29, 25] were before the advent of deep-CNN-based face recognition. One recent study involving deep CNN matchers opposes the result that younger cohorts are more difficult to recognize [4]. Another recent study agrees with the result that demographic balancing of the training data does not result in balanced accuracy across demographics [7]. The consensus across recent studies is that females, compared to males, have a worse impostor distribution and a worse genuine distribution and that African-American males have a worse impostor distribution compared to Caucasian males, but also a better genuine distribution [6, 8, 26, 37]. All of the above studies are in the context of 1-to-1 matching.

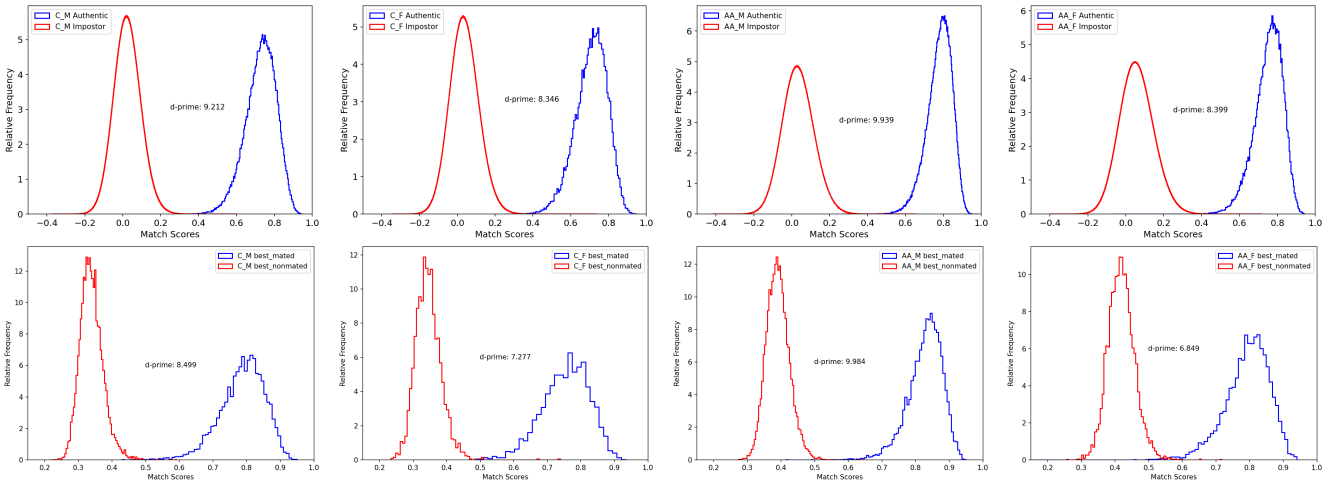
Grother and co-workers at NIST have released important reports on 1-to-many facial identification [22] and on demographic effects [23, 21]. Their report on 1-to-many identification [22] analyzes results from large numbers of matchers and images. They note the high one-to-many rank-one accuracy of current matchers, the fact that accuracy decreases linearly with the number of enrolled identities, and that enrolling multiple images for an identity increases accuracy. Their report on demographic effects [23] notes results similar to those shown later in this paper for African-American / Caucasian, and for female / male. They also discuss the specialized forms of one-to-many matching that are designed to report multiple results for a human analyst to review. The most recent NIST report [21] summarizes demographic comparisons across groups from various locations around the globe, and notes that demographic differences in false positive identification rate (FPIR) are larger than in false negative identification rate (FNIR), and that universal generalizations are difficult. They also note that some commerical matchers do not implement one-to-many matching in the straightforward manner [21], and so comparisons across commercial matchers must be done with care.

Krishnapriya et al [26] also look at 1-to-many matching accuracy, using the MORPH dataset, and also report that when the person in the probe image does have an enrolled image, the 1-to-many search is highly likely to return the correct identity. Both [23] and [26] emphasize that a 1-to-many search when the person in the probe image does not have an enrolled image can by definition only return a false-positive identification (FPI). Recent work by Drozdowski et. al. [16] examines the "watchlist imbalance effect" that occurs in 1-to-many matching when different demographics represent different fractions of the watchlist. They compare theoretical and empirical estimates of FPIR across demographics for balanced and unbalanced watchlists. They also demonstrate that equitable 1-to-1 matching doesn't necessarily ensure equitable 1-to-many matching.

3. Dataset and Matcher

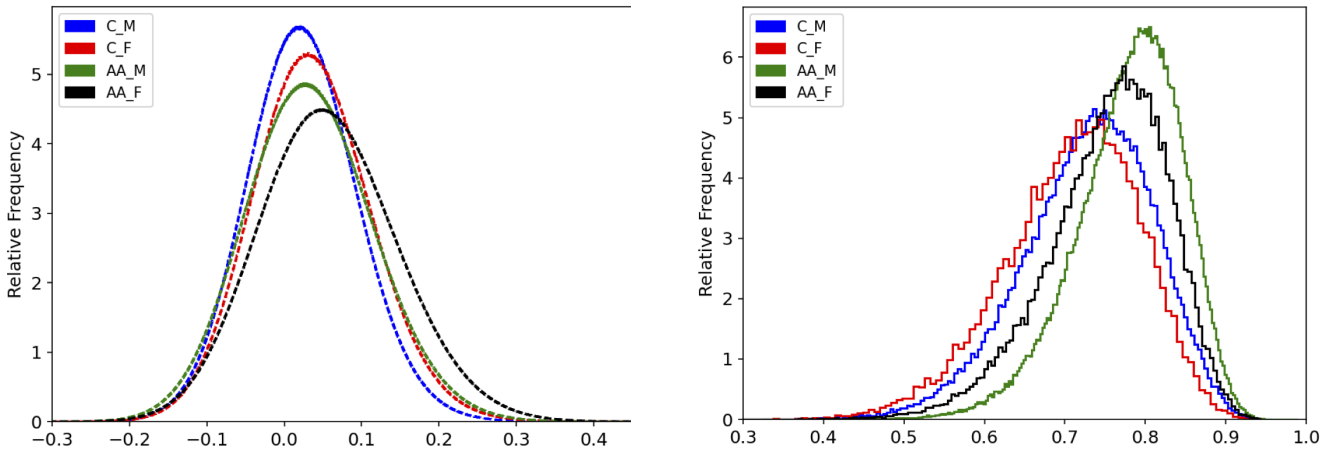
MORPH [2, 30] may be the best known dataset in face aging research and is also widely used in studying demographic accuracy variation [3, 6, 4, 16, 19, 26, 37]. As Drozdowski et al. [16] note, MORPH is particularly appropriate for demographic accuracy studies "... due to its large size, relatively constrained image acquisition conditions, and the presence of ground-truth labels (from public records) for sex, race, and age of the subjects". We use the same version of MORPH used by [8], which has 35,276 images of 8,835 Caucasian males, 10,941 images of 2,798 Caucasian females, 56,245 images of 8,839 African-American males, and 24,857 images of 5,929 African-American females. The average number of images per identity varies across demographics, averaging about 4 per identity for Caucasian male and Caucasian female, 4.2 for African-American female and 6.4 for African-American male.

Results in this paper are computed using the instance of



(a) Caucasian Male (b) Caucasian Female (c) African-American Male (d) African-American Female

Figure 2: Baseline 1-to-1 (verification) and 1-to-many (identification) distributions. Top row shows 1-to-1 impostor and genuine distributions for ArcFace; d-prime is greatest for African-American Male and lowest for Caucasian Female. Bottom row shows 1-to-many mated and non-mated distributions; d-prime is greatest (lowest false positive identification rate) for African-American Male and lowest (highest FPIR) for African-American Female.



(a) Overlaid Impostor Distributions (b) Overlaid Genuine Distributions

Figure 3: Demographic Comparison of Impostor and Genuine. For fixed decision threshold, high-similarity impostor tails show highest FMR is African-American female; low-similarity genuine tails show highest FNMR is Caucasian female.

ArcFace [15] trained on Glint-360K(R100) [9] with weights available at [1]. We have found that this instance of ArcFace, trained on a larger training set, achieves higher accuracy than one trained on MS1MV2 [15]. The input to ArcFace is an aligned face resized to 112x112, and the output is a 512-d feature vector that is matched using cosine similarity. Faces are detected and aligned using img2pose [5], which in our experience, slightly outperforms RetinaFace [14].

4. 1-to-1 Accuracy Across Demographics

The top row of Figure 2 shows the 1-to-1 impostor and genuine distributions for the demographics in MORPH. The impostor distribution contains similarity scores for all pairs of images representing different identities. The genuine distribution contains similarity scores for all pairs of images representing the same identity. We also give the d' for separation of the impostor and genuine distributions [13]:

$$d' = \frac{|\mu_G - \mu_I|}{\sqrt{\frac{\sigma_G^2 + \sigma_I^2}{2}}} \quad (1)$$

where μ_G and σ_G are the mean and standard deviation for the genuine distribution, and μ_I and σ_I for the impostor distribution. Larger d' means greater separation between impostor and genuine distributions, implying greater matching accuracy. (Using d' assumes that the distributions are reasonably approximated as Gaussian.)

It can be seen from Figure 2 that the African-American male has the largest d' of the four demographics, followed by Caucasian male, African-American female, and Caucasian female. However, the difference in d' for African-American female and Caucasian female is small. This pattern of d' implies that if identity verification was performed separately for each demographic, the highest accuracy – lowest false non-match rate (FNMR) at a fixed false match rate (FMR) – could be achieved for African-American male, followed by Caucasian male, African-American female, and Caucasian female. However, the normal operational scenario for 1-to-1 matching is that a fixed threshold is used for classifying all image pairs, regardless of demographic. Based on the fixed threshold, each demographic experiences its own combination of false match rate (FMR) and false non-match rate (FNMR). The relative FMR can be determined from the high-similarity tails of the impostor distributions and the relative FNMR from the low-similarity tails of the genuine distributions. From the high-similarity tails of the impostor distributions in Figure 3(a), the highest FMR would be for African-American female, followed by Caucasian female, African-American male and Caucasian male. From the low-similarity tails of the genuine distributions in Figure 3(b), the highest FNMR would be for Caucasian female, followed by Caucasian male, African-American female and African-American male.

5. 1-to-Many Accuracy Across Demographics

For our analyses of one-to-many matching, the most recent image of each identity is designated as its probe image and the older images as its enrolled images for the gallery. A handful of identities have just one image in MORPH, and so have a probe image but no enrolled images. Given the high rank-one matching accuracy of current matchers [22, 23, 26], a dataset the size of MORPH will not yield statistically meaningful numbers of false positive identifications. However, we can still make useful comparisons across demographics in at least three ways: (1) based on d' for separation of rank-one non-mated and mated score distributions, (2) separation in the FPIR-centric tails of the non-mated and mated distributions, and (3) the distribution of (mated – non-mated) score differences. Because it has a relatively straightforward analogy to impostor and genuine distributions in 1-to-1 matching, we begin with rank-one non-mated and mated score distributions.

The 1-to-many mated distribution contains one score for each identity that has one or more enrolled images. That score is the maximum similarity score of the probe to any enrolled image of the same identity. The 1-to-many non-mated distribution also contains one score for each probe, and that score is the maximum similarity of the probe to any enrolled image of a different identity.

1-to-many matching is also done in scenarios other than reporting just the rank-one match. For example, 1-to-many matching might report the N highest-similarity matches, or report only matches above a threshold similarity [22, 23]. In such scenarios, the candidate matches are then evaluated by a human analyst. In this paper, we present results for standard rank-one identification.

The 1-to-many non-mated and mated distributions for the demographics in MORPH are shown in the bottom row of Figure 2. The 1-to-many non-mated and mated distributions are centered at higher similarity than the 1-to-1 impostor and genuine distributions. This is because the 1-to-many score for a probe is the maximum similarity score across a set of images, whereas in 1-to-1 impostor and genuine distributions include similarity scores for all image pairs.

The d' values for the 1-to-1 matching distributions and the 1-to-many matching distributions shown in Figure 2 are summarized in Table 1, columns "1-to-1" and "1-to-many original". As with the 1-to-1 distributions, African-American male has the largest d' for the 1-to-many distributions, followed by Caucasian male. However, the order of the d' values flips for Caucasian female and African-American female, so that African-American female has the lowest d' in the 1-to-many matching results. This is an example of how demographic differences in 1-to-many matching do not necessarily repeat differences observed in 1-to-1 matching.

The results in Table 1 suggest that, other factors equal, African-American males should have the lowest false positive identification rate (FPIR). However, there are some caveats to this. One is that the number of enrolled identities and the number of images per identity vary across demographics in the results in the bottom row of Figure 2. Another is that d' may not adequately reflect variations in skew in the tails of the different distributions across the demographics. We address these two concerns in later sections. Still another concern is that operational scenarios are typically "open-set" in the sense that there is no knowledge of whether the identity in the probe image has any image enrolled in the gallery. And, every 1-to-many search using a probe that has no image in the gallery can only result in a FPI, as pointed out in [22, 26]. For this reason, a general conclusion about the FPIR across demographics in an operational scenario requires knowledge of the rate of open-set searches across demographics.

Across the more than 25,000 1-to-many searches represented in Figure 2(b) (8,800+ probes each for African-American male and Caucasian male, 5,900+ for African-

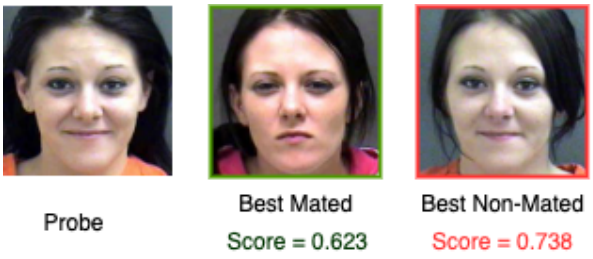


Figure 4: Instance of false-positive identification. The probe image (left) has a higher similarity score to the non-mated image (right) than it does to tbe highest similarity image of the same identity (middle).

American female and 2,700+ for Caucasian female), there is **one** instance of a rank-one FPI. The one probe image whose non-mated similarity score was higher than its mated score is for a Caucasian female identity, and the images are shown in Figure 4. Examining the images, this instance of FPI may be due to the facial expression being more similar across the non-mated image pair than across the mated pair. One false positive identification in 25,000+ 1-to-many identification searches indicates the very high accuracy of modern face recognition algorithms.

5.1. Effect of Number of Enrolled Identities

The number of identities in the gallery affects one-to-many accuracy [20, 22]. To investigate this, we split each demographic into four parts that are approximately balanced on number of subjects and average number of images per subject. This allows us to compare the non-mated and mated distributions for 25%, 50%, 75% and 100% of the identities, keeping the average number of images per identity approximately the same. These non-mated and mated distributions for different size galleries are shown in Figure 5. For each of the demographics, increasing gallery size decreases the separation between non-mated and mated distributions, and so increases the chances of FPI. This can be understood in terms of changes in the mean and standard deviation of the distributions with increasing numbers of images per identity, as shown in Figure 7a. The mean of the non-mated distribution increases, while the mean of the mated distribution and the standard deviations of the distributions undergo no significant change.

5.2. Effect of Number of Images Per Identity

To investigate the effect of the number of images per enrolled identity, we first select the subset of identities in MORPH that have 16 or more total images. This results in 495 identities for African-American male, 151 for Caucasian male, 124 for African-American female and 58 for Caucasian female. The most recent image of each identity is again kept as the probe. Then, 1, 5, 10 and 15 images are randomly selected from the remaining images as the various size galleries. The 10-image-per-identity gallery is a subset of the 15-image-per-identity gallery, and so on. This allows us to compare non-mated and mated distributions with the only difference being number of enrolled images per identity. These distributions are shown in Figure 6.

Demographic	1-to-1	1-to-many original	1-to-many balanced
A-A M	9.94	9.98	9.45
C M	9.21	8.50	8.46
C F	8.35	7.28	6.84
A-A F	8.40	6.85	6.74

Table 1: Summary of d' Comparisons Across Demographics. The d' values are summarized for 1-to-1 impostor and genuine, 1-to-many non-mated and mated with original unbalanced dataset, and 1-to-many non-mated and mated with demographics balanced on number of identities and images.

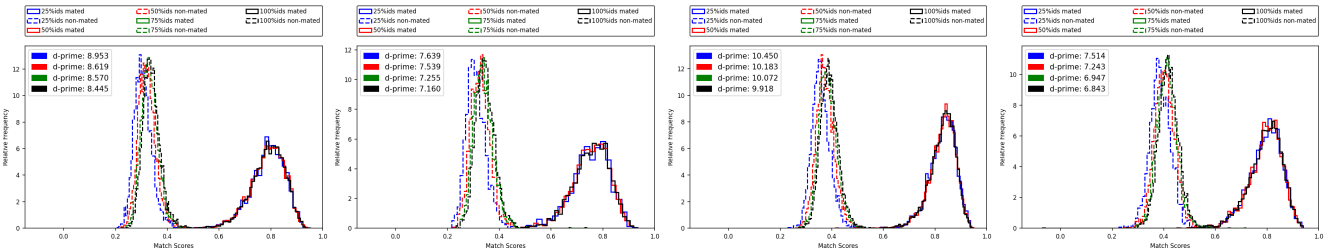
Across the demographics, increasing the number of images per identity *increases the d' for distributions*. This result may seem surprising. The trends in the mean and standard deviation of the distributions with increasing numbers of images per identity are shown in Figure 7b. Increasing the number of images per identity increases the means of the non-mated and the mated distributions, and also decreases the standard deviation of the mated distribution. The increase in the mean of the non-mated distribution is outweighed by the combination of the increase in the mean of the mated distribution and the decrease in the standard deviation of the mated distribution.

Analysis of 1-to-many matching is often limited to 1 enrolled image per identity. However, operational scenarios may allow the number of images per identity to grow over time. If the number of images per identity varies across demographics, as it does in MORPH, the number of enrolled images per identity is effectively another possible source of watchlist imbalance.

5.3. Comparison With Equal Identities and Images

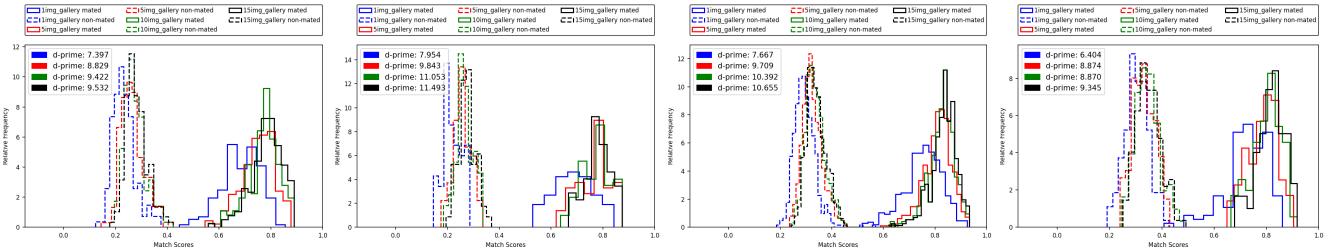
Results in the bottom row of Figure 2 shows that 1-to-many accuracy varies across demographics. However, the number of identities and the average number of gallery images per identity vary across demographics in these results. Results in Figures 5 and 6 show that 1-to-many accuracy changes based on the number of identities in the gallery, and the number of enrolled images per identity, respectively. Therefore, a more controlled comparison across demographics should control the number of enrolled identities and the number of enrolled images per identity.

The Caucasian female demographic has the least number of identities, 2,797. So we randomly select 2,797 identities



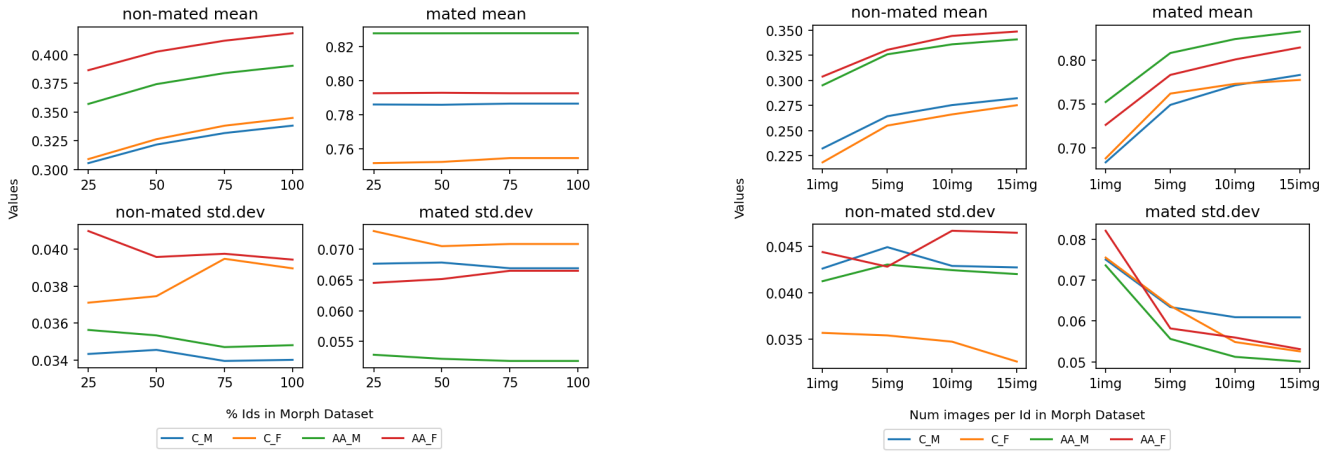
(a) Caucasian Male (b) Caucasian Female (c) African-American Male (d) African-American Female

Figure 5: Effect of Number of Enrolled Identities on 1-to-many Matching. For each of the demographics, the primary effect is to shift the non-mated distribution toward the mated, thereby on average increasing the FPIR.



(a) Caucasian Male (b) Caucasian Female (c) African-American Male (d) African-American Female

Figure 6: Number of images effect. Increasing the number of images per identity enrolled in the gallery has no clear effect on the false match identification rate, at least up until 15 images per identity enrollment.



(a) Number of person effect (b) Number of images effect

Figure 7: Mean and Standard deviation for mated and non-mated distribution. Left two columns(7a) are for increasing number of identities enrolled in the gallery. Right two columns(7b) are for increasing number of images per identity.

from each of the other demographics. Again, the most recent image of each identity is used as the probe. To equalize the number of enrolled images per identity, the next most recent image is selected as the one enrolled image per identity. This results in each of the four demographic groups having 2,797 probe images and 2,797 gallery images. Further, to check that the age difference between probe and enrolled image for an identity is approximately balanced across demographics, the distribution of time between mated images is shown in Figure 8. The differences across demographics in the distribution of probe-gallery image age difference are too small to cause noticeable accuracy difference.

The d' values for the non-mated and mated distributions based on the balanced image sets are given in Table 1 ("1-to-many balanced" column). The d' order across demographics is the same for the balanced datasets as for the original unbalanced datasets. However, the differences in d-prime between the demographics are smaller for the bal-

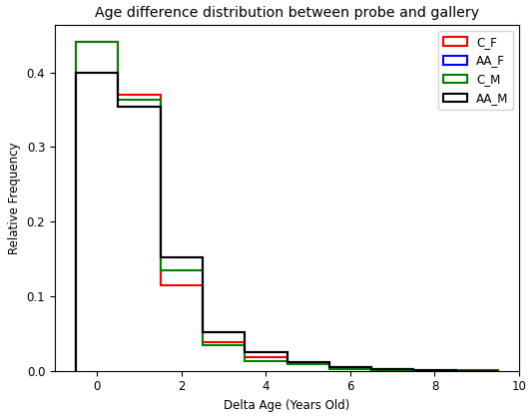


Figure 8: Distribution of Probe-Gallery Age Difference Across Demographics.

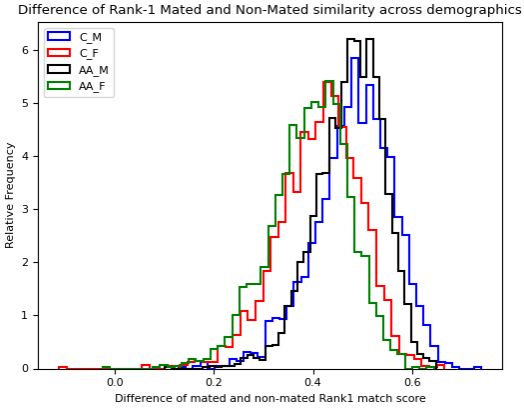


Figure 9: Distributions of (mated - non-mated) score difference across demographics.

anced comparison.

5.4. Metric for Separation In FPI Region

As mentioned earlier, using d' assumes that the distributions are reasonably approximated as Gaussian. If the high-similarity tail of the non-mated distribution and/or the low-similarity tail of the mated distribution are skewed differently enough across demographics, then comparison based d' could give a misleading impression of relative FPIR.

To check whether comparison based on d' is adequate, we also compute a non-parametric value aimed at the separation between the tails of the non-mated and mated distributions. We compute the distance between the threshold for the 1-in-1,000-th highest similarity non-mated similarity score and the lowest 1-in-1,000-th mated score. This metric makes no assumption about the form of the non-mated or mated distribution, and focuses on the tails of the distributions relevant to instances of false positive identification. A higher value of this metric indicates greater separation of the FPIR-focused tails of the distributions.

Demographic	Δ original	Δ balanced
A-A M	0.073	0.060
C M	0.010	0.059
C F	-0.005	0.018
A-A F	-0.066	-0.044

Table 2: Non-parametric"Recognition Power" Across Demographics. Δ is the distance between the 1-in-1000 threshold in the high-similarity tail of the non-mated distribution and the low-similarity tail of the mated.

	$\Delta(mated - non\_mated\ value)$					
	0.1	0.15	0.2	0.25	0.3	0.35
A-A M	0	0.03	0.10	0.39	1.43	4.36
C M	0	0.03	0.28	0.71	2.07	6.43
C F	0.14	0.32	0.965	3.00	8.29	20.80
A-A F	0.10	0.42	1.39	4.47	12.52	26.84

Table 3: Cumulative Fraction of (mated - non-mated) Difference Distribution at Low Difference Values.

Values of this metric are shown in Table 2, computed both for the original dataset with varying numbers of identities and images across demographics, and for the balanced dataset. Note that the order of the demographics is the same for the unbalanced and the balanced datasets. However, the difference between African-American male and Caucasian male is now minimal for the balanced dataset.

5.5. Distribution of (Mated - Non-mated) Difference

Evaluation of the potential for FPI based on the overall non-mated and mated distributions, or from a metric focused on the tails of the distributions, is in a sense indirect. For a given probe image, a negative (mated - non-mated) score difference represents an FPI, and larger positive differences represent being further away from possible FPI. In this sense, the distribution of (mated - non-mated) score differences is a more direct indication of potential for FPI. The distributions of (mated - non-mated) differences, for the balanced datasets, are shown in Figure 9. The two male distributions are relatively similar to each other. The two female distributions are also relatively similar to each other. The female distributions are centered at lower values than the male distributions, indicating greater potential for FPI. These results suggest that the main FPIR difference is between females and males, and that any difference between Caucasian and African-American of the same gender is much smaller. Table 3 summarizes the cumula-

tive fraction of these difference distributions across lower difference values. Except for the 0.1 difference value, the African-American female demographic has the largest fraction of the distribution at each value. This indicates that African-American female has the highest potential for FPI. The African-American male demographic has the smallest fraction of the distribution at each value, indicating low potential for FPI. Figure 9 indicates that the fractions of the African-American and Caucasian male difference distributions flip at larger difference values, but the lower tails of the distributions are most relevant to potential for FPI.

6. Conclusions and Discussion

One-to-many facial identification is an algorithmic module that is often embedded in a human decision-making system. As one example, the New York City police commissioner described their use of 1-to-many identification [28]. There are three important steps where human decision-making is involved. In the earliest step, human judgement determines which images are run through 1-to-many facial identification and which are not. In the second step, when an image is run through 1-to-many facial identification, the results include more than only the rank-one match, and human decision-making determines the single best lead from among the candidates, or that there is no plausible lead among the identities returned from 1-to-many identification. In the instances where a single best lead is returned to the detective who requested the search, human decision-making is involved in how the lead is followed up. Any of the stages that involve human decision-making could introduce an element of different results for different demographics that is beyond what we can analyze. Our analysis is focused on the degree to which a 1-to-many matching algorithm produces different accuracy across image datasets representing different demographics.

Comparison of 1-to-many accuracy across demographics should be balanced on (at least) the number of identities and number of images per identity. Other factors equal, increasing the number of enrolled identities can increase FPIR. Other factors equal, increasing the average number of images per enrolled identity can decrease FPIR. To isolate accuracy difference that is due to demographics alone, image sets should be balanced on number of enrolled identities and number of images per enrolled identity. Other factors that might impact accuracy, such as time lapse between mated images, should also be evaluated. Recent work even suggests consideration of hairstyle balancing [8, 10].

The d' for 1-to-many non-mated and mated distributions may not be an adequate single tool for comparing demographic accuracy. The d' ranking across demographics, with balanced datasets, shows African-American male with a substantially better d' than Caucasian male, then a substantial drop in d' to Caucasian female, and Caucasian female and African-American female as almost equal. (See Table 1.) The ranking based on the difference in 1-in-1,000 thresholds in the tails of the non-mated and mated distributions ranks African-American male and Caucasian male as almost equal, Caucasian female as an increment worse, and African-American female as another increment worse. The ranking based on the distribution of probe (mated similarity – non-mated similarity) differences ranks African-American male as best, with a slight advantage over Caucasian male, followed by a large gap to Caucasian female, with a slight advantage over African-American female. The rankings of the four demographics are the same in all three approaches, but the size of d' difference between African-American male and Caucasian male implies a stronger difference than is borne out in the other approaches. For this reason, comparisons based on datasets not large enough to observe large numbers of FPIs should employ multiple tools to analyze demographic differences.

Demographic differences in 1-to-many accuracy do exist, but our results only partly suport what media coverage may have led people to expect. Our results also indicate that the female demographics are noticeably disadvantaged relative to the male demographics in 1-to-many facial identification. However, our results also indicate that the African-American male demographic is not disadvantaged by 1-to-many facial identification search algorithms. In this context, we note again that one-to-many facial identification is an algorithmic module often embedded in a system with multiple human decision-making elements.

References

[1] Insightface: 2d and 3d face analysis project. `https://github.com/deepinsight/insightface/tree/master/model_zoo`.

[2] Morph dataset. `https://uncw.edu/oic/tech/morph.html`.

[3] Vítor Albiero and Kevin W Bowyer. Is face recognition sexist? no, gendered hairstyles and biology are. In *Proc. British Mach. Vision Conf.*, 2020.

[4] Vítor Albiero, Kevin W. Bowyer, Kushal Vangara, and Michael C. King. Does face recognition accuracy get better with age? Deep face matchers say no. In *Winter Conf. on App. of Comput. Vision*, 2020.

[5] Vítor Albiero, Xingyu Chen, Xi Yin, Guan Pang, and Tal Hassner. img2pose: Face alignment and detection via 6dof, face pose estimation. In *Computer Vision & Pattern Recognition (CVPR)*, 2021.

[6] Vitor Albiero, Krishnapriya KS, Kushal Vangara, Kai Zhang, Michael C King, and Kevin W Bowyer. Analysis of gender inequality in face recognition accuracy. In *Proc. Conf. Comput. Vision Pattern Recognition Workshops*, 2020.

[7] Vítor Albiero, Kai Zhang, and Kevin W Bowyer. How does gender balance in training data affect face recognition accuracy? In *Int. Joint. Conf. on Biometrics*, 2020.

[8] Vítor Albiero, Kai Zhang, Michael C King, and Kevin W Bowyer. Gendered differences in face recognition accuracy explained by hairstyles, makeup, and facial morphology. In *Trans. on Inform. Forensics and Security*, 2021.

[9] Xiang An, Xuhan Zhu, Yuan Gao, Yang Xiao, Yongle Zhao, Ziyong Feng, Lan Wu, Bin Qin, Ming Zhang, Debing Zhang, et al. Partial fc: Training 10 million identities on a single machine. In *Int. Conf. Comput. Vision (ICCV)*, pages 1445–1449, 2021.

[10] Aman Bhatta, Vítor Albiero, Kevin W Bowyer, and Michael C King. The gender gap in face recognition accuracy is a hairy problem. In *Winter Conf. on App. of Comput. Vision Workshop*, 2022.

[11] Davide Castelvecchi. Is facial recognition too biased to be let loose. *Nature*, Nov. 18 2020. https://www.nature.com/articles/d41586-020-03186-4.

[12] Abhijit Das, Antitza Dantcheva, and Francois Bremond. Mitigating bias in gender, age and ethnicity classification: a multi-task convolution neural network approach. In *European Conf. Comput. Vision Workshops*, pages 0–0, 2018.

[13] John Daugman. How iris recognition works. In *The essential guide to image processing*, pages 715–739. Elsevier, 2009.

[14] Jiankang Deng, Jia Guo, Evangelos Ververas, Irene Kotsia, and Stefanos Zafeiriou. Retinaface: Single-shot multi-level face localisation in the wild. In *Computer Vision & Pattern Recognition (CVPR)*, pages 5203–5212, 2020.

[15] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Computer Vision & Pattern Recognition (CVPR)*, 2019.

[16] Pawel Drozdowski, Christian Rathgeb, and Christoph Busch. The watchlist imbalance effect in biometric face identification: Comparing theoretical estimates and empiric measurements. In *Proc. Int. Conf. Comput. Vision Workshops*, pages 3750–3758, 2021.

[17] Pawel Drozdowski, Christian Rathgeb, Antitza Dantcheva, Naser Damer, and Christoph Busch. Demographic bias in biometrics: A survey on an emerging challenge. In *Trans. Technology and Society*, 2020.

[18] John General and Jon Sarlin. A false facial recognition match sent this innocent black man to jail. *CNN Business*, Apr. 29 2021. https://www.cnn.com/2021/04/29/tech/nijeer-parks-facial-recognition-police-arrest/index.html.

[19] Markos Georgopoulos, James Oldfield, Mihalis A. Nicolaou, Yannis Panagakis, and Maja Pantic. itigating demographic bias in facial datasets with style-based multi-attribute transfer. In *Int. J. Comput. Vision*, volume 129, page 2288–2307, 2021.

[20] P. Grother and P.J. Phillips. Models of large population recognition performance. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2, pages II–II, 2004.

[21] Patrick J. Grother. Face recognition vendor test (FRVT) part 8: Summarizing demographic differentials. Technical Report 8429, NIST, July 2022.

[22] Patrick J. Grother, Mei Ngan, and Kayee Hanaoka. Face recognition vendor test (FRVT) part 2: Identification. Technical Report 8271, NIST, Nov. 2018.

[23] Patrick J. Grother, Mei Ngan, and Kayee Hanaoka. Ongoing face recognition vendor test (FRVT) part 3: Demographic effects. Technical Report 8280, NIST, Dec. 2019.

[24] T. Hoggins. 'Racist and sexist' facial recognition cameras could lead to false arrests, Dec. 20 2019. https://www.telegraph.co.uk/technology/2019/12/20/racist-sexist-facial-recognition-cameras-could-lead-false-arrests/.

[25] Brendan F. Klare, Mark J. Burge, Joshua C. Klontz, Richard W. Vorder Bruegge, and Anil K. Jain. Face recognition performance: Role of demographic information. In *Trans. on Inform. Forensics and Security*, volume 7, pages 1789–1801, 2012.

[26] 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. In *Trans. Technology and Society*, 2020.

[27] S. Lohr. Facial recognition is accurate, if you're a white guy. *The New York Times*, Feb. 9 2018. https://www.nytimes.com/2018/02/09/technology/facial-recognition-race-artificial-intelligence.html.

[28] James O'Neill. How facial recognition makes you safer. *The New York Times*, June.

[29] P. Jonathon Phillips, Patrick J. Grother, Ross J. Michaels, Duane M. Blackburn, Elham Tabassi, and J.M. Bone. Face Recognition Vendor Test 2002: Evaluation Report. Technical Report 6965, NIST, 2003.

[30] Karl Ricanek and Tamirat Tesafaye. MORPH: A longitudinal image database of normal adult age-progression. In *Automatic Face and Gesture Recognition*, 2006.

[31] Philip Smith and Karl Ricanek. Mitigating algorithmic bias: Evolving an augmentation policy that is non-biasing. In *Winter Conf. on App. of Comput. Vision Workshop*, pages 90–97, 2020.

[32] Nisha Srinivas, Karl Ricanek, Dana Michalski, David S Bolme, and Michael King. Face recognition algorithm bias: Performance differences on images of children and adults. In *Proc. Conf. Comput. Vision Pattern Recognition Workshops*, pages 0–0, 2019.

[33] Elaisha Stokes. Wrongful arrest exposes racial bias in facial recognition technology. *CBS News*, Nov. 19 2020. https://www.cbsnews.com/news/detroit-facial-recognition-surveillance-camera-racial-bias-crime/.

[34] Philipp Terhörst, Jan Niklas Kolf, Naser Damer, Florian Kirchbuchner, and Arjan Kuijper. Post-comparison mitigation of demographic bias in face recognition using fair score normalization. *Pattern Recognition Letters*, 140:332–338, 2020.

[35] Shreya Tewari. Sexism in facial recognition technology. *Berkman Klein Center*, May 5 2021. https://cyber.harvard.edu/story/2021-05/sexism-facial-recognition-technology.

[36] Ella Torres. Black man wrongfully arrested because of incorrect facial recognition. *abc News*, June 25 2020. https://abcnews.go.com/US/black-man-wrongfully-arrested-incorrect-facial-recognition/story?id=71425751.

[37] Kushal Vangara, Michael C King, Vitor Albiero, Kevin Bowyer, et al. Characterizing the variability in face recogni-

tion accuracy relative to race. In *Proc. Conf. Comput. Vision Pattern Recognition Workshops*, 2019.

[38] Ruben Vera-Rodriguez, Marta Blazquez, Aythami Morales, Ester Gonzalez-Sosa, Joao C Neves, and Hugo Proença. Facegenderid: Exploiting gender information in dcnns face recognition systems. In *Proc. Conf. Comput. Vision Pattern Recognition Workshops*, 2019.

[39] J. Vincent. Gender and racial bias found in amazon's facial recognition technology (again). *The Verge*, Jan. 25 2019. https://www.theverge.com/2019/1/25/18197137/amazon-rekognition-facial-recognition-bias-race-gender.

[40] Mei Wang and Weihong Deng. Mitigate bias in face recognition using skewness-aware reinforcement learning. *arXiv preprint arXiv:1911.10692*, 2019.

[41] Haiyu Wu, Vítor Albiero, KS Krishnapriya, Michael C King, and Kevin W Bowyer. Face recognition accuracy across demographics: Shining a light into the problem. *arXiv preprint arXiv:2206.01881*, 2022.