The Gender Gap in Face Recognition Accuracy Is a Hairy Problem

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

Abstract

It is broadly accepted that there is a "gender gap" in face recognition accuracy, with females having lower accuracy. However, relatively little is known about the cause(s) of this gender gap. We first demonstrate that female and male hairstyles have important differences that impact face recognition accuracy. In particular, variation in male facial hair contributes to a greater average difference in appearance between different male faces. We then demonstrate that when the data used to evaluate recognition accuracy is gender-balanced for how hairstyles occlude the face, the initially observed gender gap in accuracy largely disappears. We show this result for two different matchers, and for a Caucasian image dataset and an African-American dataset. Our results suggest that research on demographic variation in accuracy should include a check for balanced quality of the test data as part of the problem formulation. This new understanding of the causes of the gender gap in recognition accuracy will hopefully promote rational consideration of what might be done about it. To promote reproducible research, the matchers, attribute classifiers, and datasets used in this work are available to other researchers.

1. Introduction

Deep learning algorithms rule the world of face recognition research. Thus a natural reaction to observing different face recognition accuracy across demographic groups is to point to imbalance in the quantity of training data across the demographic groups. But what if the difference in test accuracy is *not* caused by imbalance in the quantity of training data? What if the difference in test accuracy is caused by imbalance in the *quality* of *test* data? We show that the observed gender gap in face recognition accuracy largely disappears when female and male test images are balanced on basic elements of how hairstyle affects appearance.

The observation that face recognition algorithms achieve different accuracy for females and males goes back at least to 2002 [33]. The general observation seen in various research efforts, including the NIST report on demographic effects [32], is that females tend to have a worse impostor distribution (higher false-match rate) and worse genuine distribution (higher false non-match). We replicate this effect for both African-Americans and for Caucasians, for both the ArcFace matcher *and also for a matcher trained with training data gender-balanced on number of identities and number of images*. We then consider elements of hairstyle that differ strongly between females and males, and how those elements impact recognition accuracy. Then we select a subset of the original test data that has hairstyle-balanced female / male face visibility, so that it represents a fair evaluation of accuracy across gender. (See Section 6 for the definition of hairstyle-balanced.) The gender gap in face recognition accuracy largely disappears, for both matchers and datasets, when accuracy is evaluated with hairstyle-balanced test data.

Male Pairs Female Pair

Figure 1: Which of the male impostor pairs in the left four columns is the fairest comparison to the female impostor pair in the fifth column? The (beard, no beard) pairing in the leftmost column, the (bald, not bald) pairing in the second column and the (shorter, longer) hairstyle pairing in the third column all lead to greater dissimilarity due to hairstyle difference. The male impostor pair in the fourth column is most hairstyle balanced with the female impostor pair in the fifth column. No previous paper has looked at hairstyle balance in this level of detail in making an accuracy comparison across demographic groups.

This work suggests that differences in hairstyle-related occlusion of the face is a major causal factor of the gender gap in accuracy. To promote reproducible research, the face matchers, face attribute classifiers and datasets used in this research are all available to other researchers.

This paper is organized as follows. Section 2 presents some of the related recent works. Section 3 introduces the

datasets and matchers used. Section 4 documents the baseline gender gap in recognition accuracy across the datasets and matchers. Section 5 analyses how each of three dimensions of hairstyle – baldness, male facial hair and hairstyles – can affect recognition accuracy. Section 6 shows how gender-balancing the test data based on how hair occludes the face changes the initial gender gap in accuracy. Section 7 shows experimentally that the results drawn in section 6 are not just by chance. Section 8 draws conclusions from the experiments done and discusses few future directions.

2. Related Work

In recent years, the topic of demographic accuracy differences in face recognition has attracted attention from news media [18, 27, 36, 39] and researchers. For a broad overview of research in the area, see recent surveys [6, 20]. Here we briefly touch on selected related works.

The earliest work we are aware of to report lower accuracy for females is the 2002 Face Recognition Vendor Test (FRVT) [33]. Evaluating ten algorithms of the pre-deep learning era, identification rates of the top systems are 6% to 9% lower for females. Klare et al. [28] analyzed demographic accuracy differences using multiple matchers (commercial, nontrainable, and trainable; all pre deep learning) and reported that "The female, Black, and younger cohorts have worse ROC curves and thus have lower accuracy". They also showed impostor and genuine distributions with the same relation across gender as in Figure 2.

There is relatively little work that attempts to identify the cause of the gender gap in recognition accuracy. For example, even the most extensive study in the area, the NIST report [32] on demographic effects, lists "analyze cause and effect" under the heading of "what we did not do". Past researchers have speculated causes such as the use of cosmetics [15, 28, 30], more varied hairstyles [10], or shorter height for women, leading to non-optimal camera angle [15, 24]. Since the advent of deep learning, imbalanced training data is often suggested as the cause [20, 31, 34]. Few works have made any experimental analysis to attempt to determine the size of any speculated cause.

Imbalance in the training data is explored as a possible cause of the the gap in recognition accuracy in [11]. Experiments with VGGFace2 [14] and MS-Celeb [25], various loss functions, and multiple test sets did not reveal evidence to support the speculation that gender-balanced training data results in balanced accuracy on test data. Experiments examining different pose, expression, makeup use and forehead occlusion by hair between females and males in the MORPH dataset were reported in [10]. Differences were found between females and males in each of these factors, but balancing for the factors individually or together did not equalize female/male accuracy in the test data.

Research on how facial hair and hairstyles affect recognition accuracy is limited. Studies before the deep-learning era [22, 23] reported that accuracy is better if there is facial hair in one of the images. However, Lu et al. [30] used deep learning matchers to study effects of facial hair and reported that facial hair does not change the key features of faces, and state-of-the-art deep learning models can handle most facial hair variations. Terhörst et al. [37] investigate the influence of 47 attributes on the verification accuracy using the MAADface datset. However, this work does not evaluate hair-related attributes across demographics. Also, MAADFace is based on VGGFace2 [14], which is known to have issues with accuracy of identity labels in VGGFace2 [8], which may carry over to affect MAADFace.

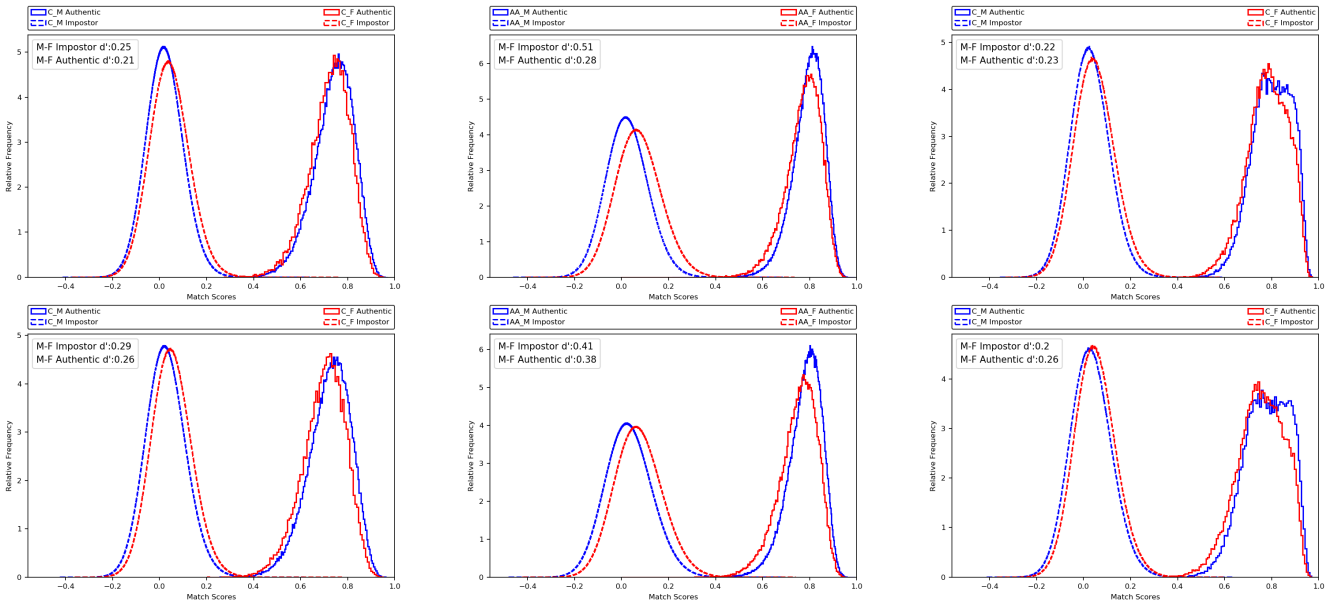
The closest related works [7, 13] attempt to explain the difference in female/male genuine distributions by balancing the test image sets on fraction of the image occupied by the face. However, they do not distinguish between male face images with and without facial hair. Also, they do not clearly identify a cause for the difference in female/male impostor distributions, and speculate that it is due to biological differences in appearance. Our work uses a more detailed and complete analysis of facial hair and hairstyle to show that the observed female/male differences in both impostor and genuine distributions are largely accounted for by hairstyle differences.

3. Dataset and Matchers

We use the MORPH dataset and the Notre Dame Male/Female Accuracy Dataset. (MFAD). MORPH [5, 35] was originally collected for research in face aging and has become widely used in the study of demographic variation in accuracy [7, 10, 11, 19, 21, 29]. MORPH is appropriate for for demographic studies, as noted by Drozdowski et al. [19], "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 version of MORPH used in [13], larger than the version used in [19], with 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. MORPH faces were detected and aligned using img2pose [9] (which in our experience slightly outperforms RetinaFace[16]) for face detection and alignment.

MFAD is drawn from images previously acquired at Notre Dame, with human subjects approval allowing release of the images. Using a second dataset other than MORPH guards against results being dependent on a particular dataset. MFAD images were acquired indoors, with roughly frontal pose and neutral expression. There are 5,444 images of 575 Caucasian females and 7,003 images of 687 Caucasian males.

The two matchers used are ArcFace [17] and a



(a) MORPH Caucasian (b) MORPH African-American (c) MFAD Caucasian

Figure 2: The "gender gap" in face recognition accuracy. C_M, C_F, AA_M and, AA_F stands for Caucasian Male, Caucasian Female, African-American Male and African-American Female respectively. The observation that females have worse impostor and genuine distribution was made by Klare et al. [28] a decade ago. The same qualitative pattern is seen here. The top row compares impostor and genuine distributions for the MORPH and MFAD dataset using standard ArcFace matcher; bottom row shows comparisons for a version of ArcFace trained on a smaller, explicitly gender-balanced dataset.

gender-balanced matcher [11]. The ArcFace used is the R100(mxnet) version available at [3]. Input to ArcFace is an aligned 112x112 face, and output is a 512-d feature vector that is matched using cosine similarity. The gender-balanced matcher [11] is a ResNet-based [26] matcher whose training data is explicitly balanced on number of female and male identities and images, available at [12]. The input is a 112x112 face image, and the output is a 512-d feature vector that is matched using cosine similarity.

4. Baseline Gender Gap In Accuracy

The impostor and genuine distributions in Figure 2 are representative of the gender gap observed by various researchers. The top row results are from ArcFace and the bottom row from the gender-balanced matcher. The first column is for the Caucasian subset of MORPH [5, 35], the second column for the African-American subset of MORPH, and the last column for MFAD Caucasian images. In all six instances of (different matcher $\times$ different racial group), the female impostor distribution (and so the FMR) and the female genuine distribution (and so the FNMR) are worse. This is the baseline "gender gap". To quantify the gap, the d' difference between the corresponding female and male distribution is given. A larger d' indicates a larger gap between the female and male distributions. The ArcFace results show that the gender gap exists for the best-known open-source state-of-the-art matcher, when trained on imbalanced training data. *The results from the gender-balanced matcher emphasize that simply balancing number of identities and images in the training data is no guarantee of balanced accuracy on test data.* Also, accuracy is overall lower for the gender-balanced matcher.

5. Gender-Based Differences In Hairstyle

This section discusses three dimensions of hairstyle: bald hairstyle, facial hair (e.g., beard) and "size" of hairstyle as measured by the fraction of the 112x112 face image that represents hair. Results document that female and male face images, as groups, differ greatly on each of these dimensions, and also that a difference in any of these dimensions can cause a noticeable difference in the impostor and / or genuine distribution.

5.1. Bald Hairstyle

A bald hairstyle is one with little to no visible hair on the top of the head. To detect baldness, we used a fusion of the modified Bilateral Segmentation network ("BiSeNet") algorithm [2, 40] results and Microsoft Face API [4] results. A pre-trained version of BiSeNet segments a face image into semantic regions, with region 17 corresponding to hair flowing from the top of the head, not including facial hair such as beard or mustache. A 112x112 cropped (frontal) face image with less than 2% of its pixels labeled as hair by BiSeNet generally corresponds to a bald hairstyle. The Mi-

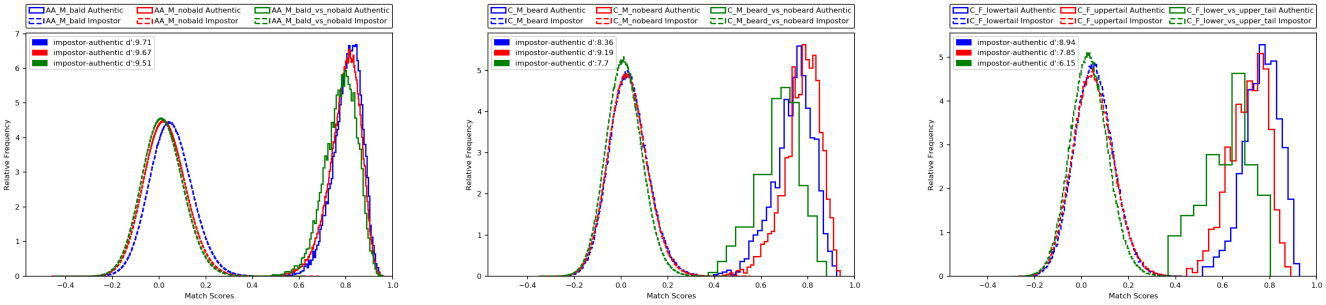
Facial Hair / No Facial Hair Classification Using MS Face + Rekognition Fusion

Dataset	Prediction	Caucasian Males	Caucasian Females	African-American Males	African-American Females
MORPH	**Facial Hair**	24,958(70%)	2(0.1%)	50,570(90%)	132(0.5%)
	No Facial Hair	10,646(30%)	10,938(99.9%)	5,640(10%)	24,721(99.5%)
MFAD	**Facial Hair**	1710(25%)	0 (0%)	-	-
	No Facial Hair	5293(75%)	5444(100%)	-	-

Bald/ Not-bald Classification Using BiSeNet Hair Ratio + MS Face

Dataset	Prediction	Caucasian Males	Caucasian Females	African-American Males	African-American Females
MORPH	**Bald**	1,371(4%)	3(0.1%)	5,868(10.5%)	30(0.2%)
	Not Bald	33,873(96%)	10,937(99.9%)	50,342(89.5%)	24,823(99.8%)
MFAD	**Bald**	30(0.4%)	0(0%)	-	-
	Not Bald	6973(99.6%)	5444(100%)	-	-

Table 1: Females and males, Caucasians and African-Americans, differ strongly in facial hair and baldness.



(a) Impact of Bald Hairstyle On Accuracy (b) Impact of Facial Hair on Accuracy (c) Impact of Fullness of Hairstyle

Figure 3: Bald hairstyle, facial hair, and "fullness" of hairstyle impact the genuine and impostor distributions.

crosoft Face API predicts baldness with a confidence ranging from 0 to 1. We found that a threshold of 0.97 results in high confidence for a bald hairstyle. We label an image as bald if (a) less than 2% of pixels are labeled as hair in the BiSeNet segmentation and (b) Microsoft Face API baldness prediction is ≥ 0.97. The fraction of the datasets labeled as bald is given in Table 1. For MORPH, just 0.1% of Caucasian female and 0.2% of African-American female images are labeled as bald. In contrast, 4% of Caucasian male and over 10% of African-American male images are labeled as bald. For MFAD, no female images and 0.4% of male images were labeled as bald.

As an example of how frequency of bald hairstyle can impact accuracy, Figure 3a shows the impostor and genuine distributions for MORPH African-American male broken out by (a) pairs of images both labeled bald, (b) both not bald, and (c) bald/not-bald pairs. The impostor distribution for bald/not-bald image pairs shows the lowest similarity, followed by not-bald pairs, and then bald pairs. On average, images of two different persons, one bald and one not, look less similar (have lower FMR) than images of different persons both bald or both not bald.

5.2. Facial Hair

Beard, mustache, sideburns, five o'clock shadow and related facial hair are generally limited to male images. We used a three-part fusion of results from the Microsoft Face API and Amazon Rekognition [1] to classify images as clean-shaven or facial hair. Microsoft Face predicts presence of beard, mustache, and sideburns, individually, each with confidence score values of 0, 0.1, 0.4, 0.6 and 0.9. In our experience, a score of 0.6 or 0.9 is generally accurate for presence of facial hair, but some instances of facial hair still occur at lower confidence values. For this reason, images with Microsoft Face confidence less than 0.6 are filtered with results from Amazon Rekognition. Amazon Rekognition gives a True/False for facial hair along with a confidence score from 50 to 100. An Amazon Rekognition result of True with a confidence greater than 85 is taken as indicating facial hair. Lastly, an image with a Microsoft Face confidence of 0.4 and an Amazon Rekognition True with confidence greater than 55 or an Amazon Rekognition False with confidence less than 65 is taken as indicating facial hair. This fusion approach is reasonably accurate in classifying images for facial hair / clean-shaven, but greater accuracy would be desirable.

Using this fusion algorithm, the fraction of images labeled as facial hair / clean-shaven for each demographic is given in Table 1. For MORPH, almost no female images are labeled as facial hair, but 90% of African-American male and 70% of Caucasian male images are. For MFAD, no female face images were labeled as facial hair, and 25% of Caucasian male images were labeled as facial hair.

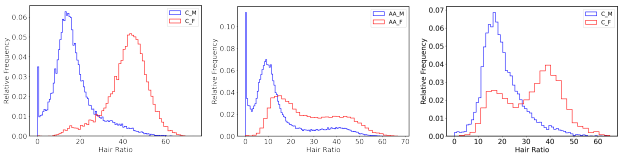
As an example of how frequency of facial hair impacts

Figure 4: Increased hair ratio means greater face occlusion. Top row has ≈10% of the 112x112 image occupied by pixels representing hair, middle row ≈30%, and bottom row hair ≈60%. Different distribution of hair ratio can cause differences in the impostor and genuine distribution.

accuracy, Figure 3b shows the impostor and genuine differences for MORPH Caucasian male images broken out by pairs with both images classified as clean-shaven, both classified as facial hair, and (clean-shaven, facial hair) mix. Image pairs with one image having facial hair and one clean-shaven have an impostor distribution and a genuine distribution with lower average similarity than pairs with both images having facial hair, or with both being clean-shaven.

5.2.1 Misclassification in Facial Hair Prediction

As explained above, we use a fusion of Microsoft Face and Amazon Rekognition results to classify an image as having facial hair or clean-shaven. To check the accuracy of this approach, we randomly selected 300 images for MORPH African-American male and 300 for MORPH Caucasian male. Each group of 300 had 100 with prominent beard and facial hair, 100 with less prominent facial hair, and 100 clean-shaven. For Caucasian male, all 100 of the prominent facial hair group were classified as facial hair, 97 of 100 with less prominent facial hair were classified as facial hair, and 81 of 100 with no facial hair were classified as clean-shaven. For African-American male, all 100 of the prominent facial hair group were classified as facial hair, 99 of 100 with less prominent facial hair were classified as facial hair, and only 32 of 100 with no facial hair were classified as clean-shaven. These results point to two limitations in our current ability to classify face images as clean-shaven / facial hair. One is that more clean-shaven images are incorrectly classified as facial hair, than facial hair images incorrectly classified as clean-shaven. The second is that the accuracy of classifying clean-shaven is lower for African-American than for Caucasian. Facial hair classification with higher accuracy and balanced accuracy across demographics is a topic for future research.



(a) MORPH C (b) MORPH AA (c) MFAD C

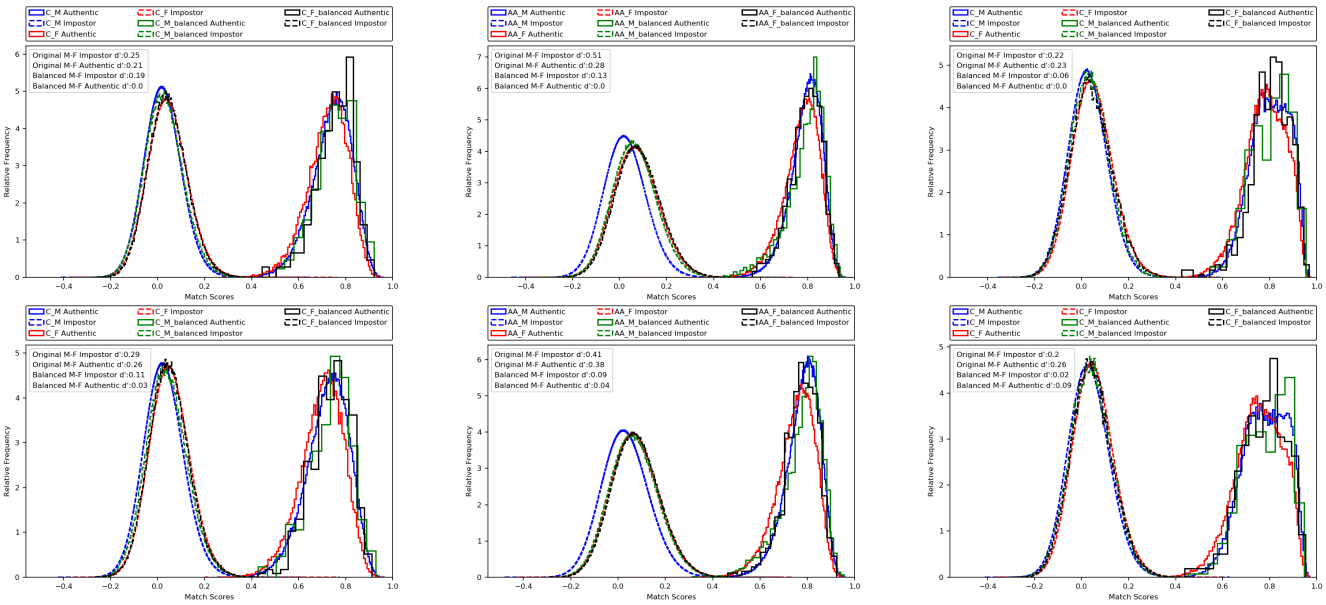
Figure 5: Distributions of fraction of cropped face image containing hair. Female images have, on average, a much larger of the face occluded by hair.

5.3. Fullness of Hairstyle

In general, an increasing fraction of the image containing hair means increasing occlusion of the face, as illustrated in Figure 4. The distribution of the fraction of the image that is labeled as hair in the BiSeNet segmentation, the "hair ratio", is shown in Figure 5. Note that for MORPH, both African-American and Caucasian males have a spike at 0%, representing bald, and then another broader peak under 20%, representing hairstyles with low face occlusion. In contrast, Caucasian females have a broad peak in the 40% to 50% range, implying substantially more occlusion of the face by hair. And African-American females have a broad plateau in the 10% to 50% range, indicating a varied range of hairstyles that occlude different amounts of the face. For MFAD, there is a peak for males at slightly above 20%, whereas for women there is peak in the range 35% to 45%. It is clear from the distributions in Figure 5 that females have a broader range of hairstyles, and that on average a female image has more of the face occluded by hair.

As an example of how different distributions of "hair ratio" translate into occlusion that impacts accuracy, we divide the MORPH Caucasian female distribution in Figure 3c into a lower tail (below 25% hair ratio) and an upper tail (above 50% hair ratio). Figure 3c shows the impostor and genuine distributions for image pairs in the lower tail (less face occlusion by hair), the upper tail (more face occlusion by hair), and across the lower and upper tail (different patterns of occlusion). Image pairs from across the tails result in an impostor distribution and a genuine distribution centered at lower similarity than the distributions from image pairs with similar face occlusion by hair.

There are three important points from the results in this section. One, female and male face images, as groups, exhibit major differences in how hairstyle occludes the face. Two, each of the gendered hairstyle differences explored can substantially affect recognition accuracy. Three, previous research that has observed a gender gap in recognition accuracy has generally made no attempt to control for differences in hairstyle. We take up this third point next.



(a) MORPH Caucasian (b) MORPH African-American (c) MFAD Caucasian

Figure 6: Impostor and genuine distributions for image sets that are hairstyle-balanced across female / male. Top row results are for ArcFace, bottom row results are for the gender-balanced matcher.

6. Hairstyle-Balanced Accuracy Comparison

How does recognition accuracy compare for females and males when test data is "fair" in the sense of being balanced on hairstyle? To approach this question, we first define what it means to be "balanced on hairstyle".

6.1. Hairstyle Balancing

Bald is more frequent for males, and mixed (bald/not-bald) image pairs have different impostor and genuine distributions than (not-bald/not-bald) pairs. Therefore, to get a hairstyle-balanced comparison across gender, we drop images with bald hairstyle. Based on Table 1, this obviously reduces the number of male images far more than it does the number of female images.

Facial hair is common for males and basically non-existent for females, and mixed (facial-hair/clean-shaven) image pairs have different impostor and genuine distributions than (clean-shaven/clean-shaven) pairs. Therefore, to get a hairstyle-balanced comparison of across gender, we also drop images with facial hair. This step also reduces the number of male images.

Changes in the distribution of fraction of the image representing hair impact the impostor and genuine distribution. Therefore, we want to balance the female and male image sets based on the portion of the 112x112 cropped face image that represents hair. This is done by establishing a correspondence between female and male images based on the intersection-over-union (IoU) of the pixels in the hair regions of the images. For each female image, select the male image with the highest IoU of the hair regions, and if this IoU is above a threshold of 0.8, the images are kept for the hairstyle-balanced accuracy evaluation. Using IoU to balance hairstyle ensures that both images not only have equal percentage of face visible but also have approximately the same regions of the face visible.

The resulting hairstyle-balanced comparison of female / male impostor and genuine distributions is in Figure 6. The changes in the d' differences are tabulated in Table 2. Please refer to Section 4 of the supplementary material for the summarized steps for HairStyle Balancing.

6.2. Results and Discussions

The results in Figure 6b are for a balanced subset from MORPH, with 2,127 images of African-American males (1024 subjects) and 2,127 images of African-American females(1564 subjects). The impostor and genuine distributions for the hairstyle-balanced image sets show a fundamental change from the original dataset. For ArcFace, the original d' between male and female impostor distributions is 0.509, reduced to 0.129 after hairstyle-balancing. A similar pattern holds for the gender-balanced matcher; the original d' between male and female impostor distributions is 0.410, reduced to 0.085 after hairstyle-balancing. Thus, hairstlye balancing reduces the impostor d' between African-American males and females by $\approx$75% and $\approx$79% for ArcFace and gender-balanced matcher, respectively.

The genuine distribution for males in the balanced subset is relatively unchanged but is slightly better for females in the balanced subset than the original dataset, reducing the gap in genuine distribution and thus, reducing

Matcher	Dataset	Category	Impostor			Genuine		
			d-prime before	d-prime after	delta d-prime	d-prime before	d-prime after	delta d-prime
ArcFace	MORPH	C_M vs C_F	0.246	0.185	-24%	0.208	0.004	-98%
		AA_M vs AA_F	0.509	0.129	-75%	0.283	0.003	-99%
	MFAD	C_M vs C_F	0.224	0.061	-73%	0.228	0.000	-100%
Gender Balanced	MORPH	C_M vs C_F	0.287	0.113	-61%	0.260	0.028	-89%
		AA_M vs AA_F	0.410	0.085	-79%	0.375	0.042	-89%
	MFAD	C_M vs C_F	0.204	0.023	-89%	0.261	0.091	-65%

Table 2: d-prime for female / male impostor and genuine distributions. "d-prime before" is for original test data, and "d-prime after" is for hairstyle-balanced. Balancing on hairstyle decreases the gap between female and male impostor distributions and between female and male genuine distributions. The text in red is to show that the results might not be very reliable due to: (a) very few genuine pairs, and (b) shoulder in the high similarity tail of original genuine distribution.

Matcher	Dataset	Category	Impostor			Genuine		
			d-prime balanced	Mean d-prime random	Std.dev d-prime random	d-prime balanced	Mean d-prime random	Std.dev d-prime random
ArcFace	MORPH	C_M vs C_F	0.185	0.235	0.030	0.004	0.402	0.252
		AA_M vs AA_F	0.129	0.503	0.019	0.003	0.305	0.147
	MFAD	C_M vs C_F	0.061	0.239	0.091	0.000	0.276	0.236
Gender Balanced	MORPH	C_M vs C_F	0.113	0.274	0.003	0.028	0.430	0.263
		AA_M vs AA_F	0.085	0.410	0.018	0.042	0.390	0.151
	MFAD	C_M vs C_F	0.023	0.214	0.085	0.091	0.271	0.230

Table 3: Mean male-female impostor/genuine d-prime and std. dev. for 1000 random samples without replacement. "d-prime balanced" is for hairstyle-balanced subset, "Mean d-prime random" and "Std.dev. d-prime random" are for 1000 random samples. The hairstyle-balanced d-primes are not within one std. dev. of randomly sampled mean. The text in red shows that balanced d-prime doesn't fall within one standard deviation. The likely cause is mentioned in caption of Table 2.

the differences in genuine d' between males and females. For ArcFace, the original d' score between male and female genuine is 0.283, whereas it is 0.003 after balancing hair dimensions. A similar pattern holds for the gender-balanced matcher; the original d' score between male and female genuine is 0.375, whereas it is 0.042 after balancing hair dimensions. Thus, balancing hair dimensions reduces ≈99% and ≈89% in the genuine d' gap gap between African-American males and females for ArcFace and gender-balanced matcher respectively.

After filtering for the hair dimensions, we present the results in Figure 6a for a balanced subset from MORPH with 684 images of Caucasian males (522 Subjects) and 684 images of Caucasian females (481 Subjects). For ArcFace, the original d' score between male and female impostors is 0.246, whereas it is 0.187 after balancing hair dimensions. A similar pattern holds for the gender-balanced matcher, with the original d' between male and female impostors of 0.287 reduced to 0.113 by hairstyle balancing. Thus, hairstyle balancing reduces the impostor d' gap between Caucasian males and females by ≈24% and ≈61% for ArcFace and gender-balanced matcher, respectively.

In addition, the genuine distribution seems to improve for both males and females after hairstyle balancing. This, in turn, causes the genuine d' gap for the hairstyle-balanced subset to be significantly lower than the original dataset. For Arcface, the original d' score between male and female impostors is 0.208, whereas it is 0.004 after balancing hair dimensions. Similar pattern holds for the gender-balanced matcher. The original d' score between male and female genuine is 0.260, whereas it is 0.028 after balancing hair dimensions. Thus, balancing hair dimensions reduces ≈98% and ≈89% in the genuine d' gap gap between African-American males and females for ArcFace and gender-balanced matcher respectively.

Results for MFAD are in Figure 6c. The balanced subset is 344 images of Caucasian males (178 Subjects), and 344 images of Caucasian females (149 Subjects). A similar pattern of shifts in impostor and genuine distribution is evident for MFAD. For ArcFace, the original d' between male and female impostors is 0.224, whereas it is 0.061 after hairstyle-balancing. A similar pattern holds for the gender-balanced matcher; the original d' between male and female impostors is 0.204, whereas it is 0.023 after hairstyle-balancing. Thus, hairstyle-balancing accounts for ≈73% and ≈89% of the impostor d' gap between males and females for ArcFace and gender-balanced matcher respectively.

The genuine distribution for males in the balanced subset seems to be relatively unchanged but is slightly better for females in the balanced subset than the original dataset, reducing the gap in genuine distribution. In other words, the difference in genuine d' gap between males and females significantly reduces after balancing for hair dimensions.

For ArcFace, the original d' score between male and female genuine is 0.228, whereas it is 0.000 after hairstyle-balancing. A similar pattern holds for the gender-balanced matcher, with the original d' between male and female genuine of 0.261 reduced to 0.091 after hairstyle-balancing. Thus, hairstyle-balancing closes ≈65% of the genuine d' gap between African-American males and females.

7. Bootstrap Confidence Analysis

The number of images in our hairstyle-balanced accuracy comparison is greatly reduced from the original dataset. To analyze whether the results could be due to a random sampling of that amount of data from the original dataset, we randomly selected 1000 times from the original dataset the same number of subjects and images as in the final hairstyle-balanced subset. For both matchers, the d' for male-female impostors and genuine of hair dimensions balanced subset is not within one standard deviation of the mean of the randomly-sampled subsets, suggesting that our hairstyle-balancing results are highly unlikely to be by chance. All the results are shown in Table 3.

8. Conclusions

Cause of observed gender gap in accuracy. One main contribution of this work is to document and explain a cause-and-effect understanding of the gender gap in face recognition accuracy. The gender gap in accuracy that is initially observed with both ArcFace and with a gender-balanced matcher (as shown in Figure 2) largely disappears when the test image set is hairstyle-balanced so that female and male have about the same amount of the image that represents the face (as shown in Figure 6).

Quality of test data, not quantity of training data. One initial reaction to the observed gender gap in face recognition accuracy is that it must be caused by imbalance in the quantity of training data [11]. Table 2 compares d' differences between ArcFace trained on the imbalanced MS1MV2 dataset, and a matcher trained on explicitly balanced training data. For the original test data, the gender-balanced matcher had a smaller d' only for African-American impostor distributions. For the hair-balanced test data, the gender-balanced matcher had smaller d' for the impostor distributions, but larger d' for the genuine distributions. Thus, while composition of training data is in general an important consideration, balancing training data on number of identities and images showed no consistent improvement toward more gender-balanced accuracy on the test data.

Cause identified, possible solutions. Unequal accuracy caused by gendered hairstyle patterns is harder to solve than if unequal accuracy was caused by training data imbalance. In certain controlled image acquisition scenarios, a partial solution might be to ask persons to pull their hair back when the image is taken. But a broader, algorithm-level solution is likely to involve more explicit recognition of and accounting for hairstyle differences between face images. This is a relatively under-studied element of face image analysis, at least compared to issues of pose, illumination, expression and aging.

Other datasets, other possible causes: In-the-wild, celebrity images. Our analysis is done using datasets with relatively controlled image acquisition, as is also the case in [33, 28, 32, 10]. Web-scraped, in-the-wild images have greater variation in pose, illumination, expression and occlusion, not to mention unknown image compression and "photoshopping" effects. Also, celebrity images will likely bring greater use of makeup and other enhancements. Thus for some other types of datasets, gender-based hairstyle difference may or may not have the same level of relative importance.

Accuracy of face attributes. Classification of face attributes such as presence of beard or mustache is an active research area [38, 41]. Our experience suggests there is still substantial room to improve the accuracy of such algorithms, especially for detecting elements of facial hair. Related to demographics studies, it will be important to further explore whether the accuracy of such algorithms varies across demographic groups, as our initial experience suggests that it does.

References

[1] Amazon rekognition. https://aws.amazon.com/rekognition/.

[2] Bisenet. https://github.com/zllrunning/face-parsing.PyTorch.

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

[4] Microsoft face api. https://azure.microsoft.com/en-us/services/cognitive-services/face/.

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

[6] Salem Hamed Abdurrahim, Salina Abdul Samad, and Aqilah Baseri Huddin. Review on the effects of age, gender, and race demographics on automatic face recognition. In *The Visual Computer*, 2018.

[7] Vítor Albiero and Kevin W Bowyer. Is face recognition sexist? no, gendered hairstyles and biology are. *arXiv preprint arXiv:2008.06989*, 2020.

[8] 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.

[9] Vítor Albiero, Xingyu Chen, Xi Yin, Guan Pang, and Tal Hassner. img2pose: Face alignment and detection via 6dof, face pose estimation. In *Proc. Conf. Comput. Vision Pattern Recognition*, 2021.

[10] 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.

[11] 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.

[12] Vítor Albiero, Kai Zhang, and Kevin W Bowyer. How Does Gender Balance In Training Data Affect Face Recognition Accuracy? https://github.com/vitoralbiero/gender_balance_training_data, 2020.

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

[14] Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *Automatic Face and Gesture Recognition*, 2018.

[15] Cynthia M Cook, John J Howard, Yevgeniy B Sirotin, and Jerry L Tipton. Fixed and varying effects of demographic factors on the performance of eleven commercial facial recognition systems. *Trans. Biometrics, Behavior, and Identity Science*, 2019.

[16] Jiankang Deng, Jia Guo, Evangelos Ververas, Irene Kotsia, and Stefanos Zafeiriou. Retinaface: Single-shot multi-level face localisation in the wild. In *Proc. Conf. Comput. Vision Pattern Recognition*, pages 5203–5212, 2020.

[17] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proc. Conf. Comput. Vision Pattern Recognition*, 2019.

[18] C. Doctorow. NIST confirms that facial recognition is a racist, sexist dumpster-fire, 2019. https://boingboing.net/2019/12/19/demographics-v-robots.html.

[19] Pawel Drozdowski, Christian Rathgeb, and Christoph Busch. The watchlist imbalance effect in biometric face identification: Comparing theoretical estimates and empiric measurements. In *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 3750–3758, 2021.

[20] Pawel Drozdowski, Christian Rathgeb, Antitza Dantcheva, Naser Damer, and Christoph Busch. Demographic bias in biometrics: A survey on an emerging challenge. *IEEE Transactions on Technology and Society*, 1(2):89–103, 2020.

[21] Markos Georgopoulos, James Oldfield, Mihalis A. Nicolaou, Yannis Panagakis, and Maja Pantic. Mitigating demographic bias in facial datasets with style-based multi-attribute transfer. *Internationl Journal of Computer Vision*, 129:2288–2307, 2021.

[22] Geof Givens, J Ross Beveridge, Bruce A Draper, and David Bolme. A statistical assessment of subject factors in the pca recognition of human faces. In *Proc. Conf. Comput. Vision Pattern Recognition Workshops*, 2003.

[23] Geof Givens, J Ross Beveridge, Bruce A Draper, Patrick Grother, and P Jonathon Phillips. How features of the human face affect recognition: a statistical comparison of three face recognition algorithms. In *Proc. Conf. Comput. Vision Pattern Recognition*, 2004.

[24] Patrick J Grother, Patrick J Grother, P Jonathon Phillips, and George W Quinn. *Report on the evaluation of 2D still-image face recognition algorithms*. Citeseer, 2011.

[25] Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao. MS-celeb-1M: A dataset and benchmark for large-scale face recognition. In *European Conf. Comput. Vision*, 2016.

[26] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proc. Conf. Comput. Vision Pattern Recognition*, pages 770–778, 2016.

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

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

[29] 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. *Trans. Technology and Society*, 2020.

[30] Boyu Lu, Jun-Cheng Chen, Carlos D Castillo, and Rama Chellappa. An experimental evaluation of covariates effects on unconstrained face verification. *Trans. Biometrics, Behavior, and Identity Science*, 2019.

[31] M. Merler, N. Ratha, R. Feris, and J.R. Smith. Diversity in faces. *https://arxiv.org/abs/1901.10436*.

[32] Mei Ngan Patrick Grother and Kayee Hanaoka. Face Recognition Vendor Test (FRVT) Part 3: Demographic Effects. *NIST IR 8280*, 2003.

[33] P.J. Phillips, P.J. Grother, R.J. Micheals, D.M. Blackburn, E. Tabassi, and J.M. Bone. Face Recognition Vendor Test 2002: Evaluation Report. *NIST IR 6965*, 2003.

[34] A. Morales R. Vera-Rodriguez, M. Blazquez. Facegenderid: Exploiting gender information in dcnns face recognition systems. In *Proc. Conf. Comput. Vision Pattern Recognition Workshops*, 2019.

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

[36] E. Santow. Emerging from AI utopia. *Science*, 2020.

[37] Philipp Terhörst, Jan Niklas Kolf, Marco Huber, Florian Kirchbuchner, Naser Damer, Aythami Morales Moreno, Julian Fierrez, and Arjan Kuijper. A comprehensive study on face recognition biases beyond demographics. *Trans. Technology and Society*, 2022.

[38] Nathan Thom and Emily M Hand. Facial attribute recognition: A survey. *Computer Vision: A Reference Guide*, 2020.

[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] Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang. Bisenet: Bilateral segmentation network for real-time semantic segmentation. In *European Conf. Comput. Vision*, 2018.

[41] Xin Zheng, Yanqing Guo, Huaibo Huang, Yi Li, and Ran He. A survey of deep facial attribute analysis. *Int. J. Comput. Vision*, 2020.