Facial Hair Area in Face Recognition Across Demographics: Small Size, Big Effect

Haiyu Wu[1] Sicong Tian[3] Aman Bhatta[1] Kağan Öztürk[1]
Karl Ricanek Jr.[2] Kevin W. Bowyer[1]

[1]University of Notre Dame
[2]University of North Carolina Wilmington
[3]Indiana University South Bend

Abstract

Observed variations in face recognition accuracy across demographics, often viewed as "bias", have motivated research into the causes of such variations. Variations in facial hairstyle are an important potential cause of accuracy differences for males. In this work, we first explore how face recognition accuracy is affected by the facial hair region - clean-shaven, mustache, chin-area beard, side-to-side beard. Results show that mustache area facial hair has a greater effect on accuracy than either chin-area beard or side-to-side beard. We then employ a synthetic facial hair method to verify the consistency of the observation across five synthetic facial hair colors and three face matchers. Results of these experiments indicate that, the larger the difference in brightness between facial hair region and skin region, the larger impact of the mustache area. To reduce accuracy differences caused by facial hairstyle, quantified by $\Delta d'$, we adjust the training dataset distribution to have increased representation of facial hair, resulting in an over 40% reduction in accuracy difference.

1. Introduction

Face recognition technologies [15, 20, 23, 34] have demonstrated robust accuracy across a range of conditions. However, the issue of accuracy disparity, or "bias," has become controversial [5, 10, 11, 16, 21, 29, 30, 40]. To better understand the causes of accuracy disparity, researchers have explored factors such as face region brightness [35], hairstyle [3, 9], facial morphology [7, 37], gender ratio in the training sets [6], and beard area [36]. [36] showed that the fraction of the face covered by facial hair has strong differences across demographics. This motivates the importance of understanding the effect of facial hair on accuracy in order to understand demographic differences. This paper explores the effect of the location and size of the facial hair region across different demographics on state-of-the-art face matchers. The rest of the paper is organized as follows. Section 2 provides related literature about facial hair effect in face recognition and synthetic facial hair generation. Section 3 investigates the effect of three facial hair areas in face recognition based on real image pairs. Section 4 uses synthetic facial hair to explore effects of facial hairstyles that occur rarely in existing test sets. Section 5 mitigates the accuracy disparity between facial hair areas by manipulating the facial hair distribution in the training sets. Contributions of this work include the following:

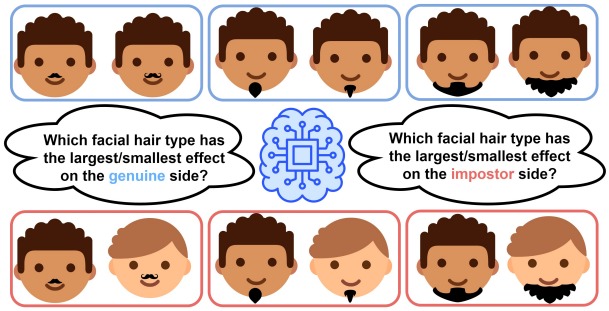


Figure 1. How is the similarity score for a pair of face images affected by facial hairstyles? This work explores this question for mustache, chin-area beard and side-to-side beard. We demonstrate that, despite mustache occupying the smallest area on face, it has the largest influence in face recognition accuracy. In addition, the impact of facial hair can be reduced by increasing the fraction of facial hair images in the training set.

- We demonstrate that, for both real and synthetic facial hair samples, when the facial hair pixel distribution is not close to the face skin pixel distribution, a clean shaven with mustache isolated (CS-MI) has greater impact on the similarity value of image pairs than does

a chin area with no mustache (CA-MN) beard or a side-to-side with no mustache (S2S-MN) beard, even though the CS-MI occludes less than half as many pixels on the face.

- A systematic evaluation is conducted by superimposing different pixel values across varying amounts of typical facial hair locations to explore the generality of the observed pattern. The consistent observations from this study suggest that the position of the pixels carries more impact than either their value or the total area of the face involved.

- Experiments manipulating the fraction of mustache samples per identity in the training data show that appropriate design of the training set can mitigate accuracy differences due to facial hair.

2. Literature Review

Effect of Facial Hair on Face Recognition. Bhatta et al. [9] investigated the effect of beards on face recognition across demographics. They fused results of the Microsoft Face API and Amazon Rekognition to obtain a binary classification of beard / no beard. They found that matching beard vs. no beard images decreases the similarity score, and both images having no beard results in a higher similarity score. However, facial hairstyle can vary significantly in area, length, and connectedness, motivating a more detailed analysis. Wu et al. [36] proposed a face attribute scheme which provides more descriptive facial hair attributes. They reported that a larger difference in beard area type decreases the similarity score and the same beard area attribute affects accuracy differently for each race. This also motivates a more detailed investigation of the effect of mustache, connectedness, and facial hair length in face recognition. Ozturk et al. [28] assembled a facial hair segmentation dataset in order to investigate the effect of the facial hair area size in face recognition accuracy. Terhörst et al. [33] reported that having facial hair enhances the performance of face recognition model. Due to the poor description of facial hair (i.e. Mustache, 5 O'clock Shadow, Goatee, No Beard) in CelebA attributes, this conclusion cannot represent the effect of facial hair area in face recognition. Our results show the opposite conclusion on both bias-factor controlled datasets.

Synthetic Facial Hair. Synthetic beard editing is achieved in three ways: statistical formulation, Generative Adversarial Networks (GANs), and language-enhanced multi-modal approaches.

Nguyen et al. [26] treated beards as outliers of the non-beard subspace to extract layers and utilize them to add and remove beards from images. Mohammed et al. [24] considered face images as samples from a texture with spatially varying statistics and described this texture with a local non-parametric model. They combined the local and global models to add facial hair. These two pre-deep-learning papers formulated facial hair as a statistical problem.

Brock et al. [12] proposed an Introspective Adversarial Network which integrates both GANs and Variational Autoencoders (VAEs) to leverage the power of adversarial learning. Olszewski et al. [27] introduced an algorithm that enables flexibility in controlling styles of added facial hair by simply drawing masks and strokes. Yao et al. [38] proposed a latent transformation network and a novel loss function that disentangles features while preserving identity. They measured identity preservation by calculating the similarity value between the original image and edited image, rather than using a face matching similarity score. Moreover, while GAN models can generate visually realistic beards, their expressive power is constrained by the variations present in the training dataset.

Thanks to the contributions of CelebA-HQ [22] and CelebA-Dialog [19], it is possible to use multi-modal models to edit facial attributes. Jiang et al. [19] added text descriptions of the facial attributes of each image in CelebA-HQ, which enables training large-scale networks. Moreover, they used the CelebA-Dialog dataset to achieve fine-grained face attribute editing. Huang et al. [18] combined pre-existing uni-modal diffusion models, thereby facilitating multi-modal face generation and editing without necessitating further training. Multi-modal models provide a fine-grained level of face attribute editing, but there is no guarantee of identity preservation.

In this paper, due to different social habits across demographics, and the uniqueness of the attribute, the number of images of Asians with side-to-side beard does not support a conclusive observation. Hence, we add synthetic facial hair to no-facial-hair images for each race. Note that, although face attribute editing models achieve good visual quality, there is no available method that can generate synthetic facial hair to a specific region only. GAN approaches to adding facial hair change more of the image than only the facial hair region. Therefore, we use face landmarks to add different types of synthetic facial hair regions to images, to efficiently gain an overview of the effects.

3. Analysis Based On Natural Facial Hair

To ensure observations that apply across datasets and matchers, we conducted experiments with a popular controlled-acquisition dataset, MORPH3 [2, 31] and also a dataset, BA-test [37], selected from a popular in-the-wild dataset, and with two face matchers. MORPH3 images are frontal, neural expression, consistent indoor lighting and plain background. The version of MORPH3 used has 56,245 images of 8,839 African-American males (AAM) and 35,276 images of 8,835 Caucasian males (CM). Faces

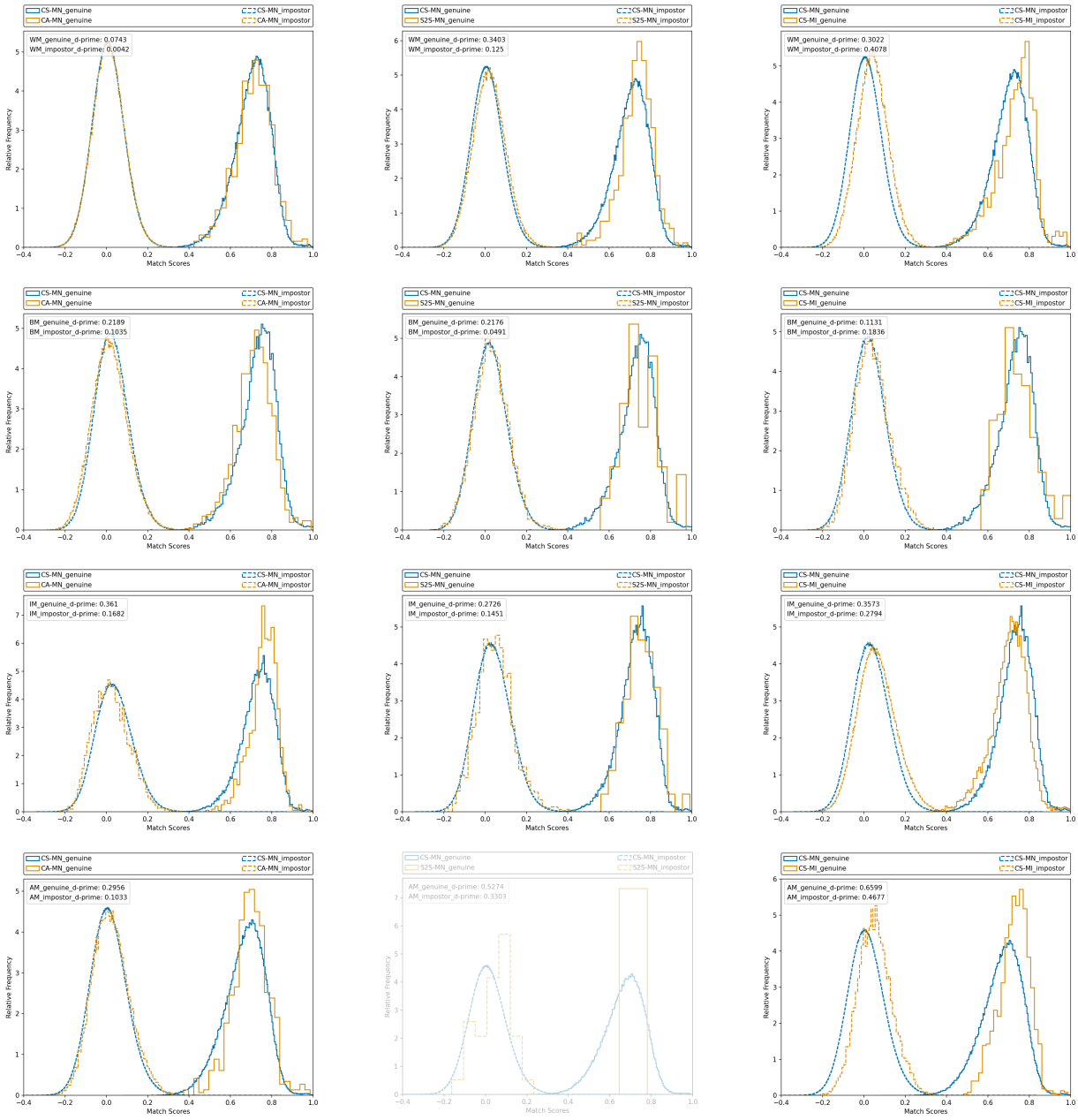


Figure 2. The genuine and impostor distributions comparisons between CS-MN and {CA-MN, S2S-MN, and CS-MI} in BA-test datasets. The face matcher is ArcFace-R100 trained with MS1MV2. Note that results derived from insufficient data are rendered in a semi-transparent format.

were detected and aligned using img2pose [4]. Different from MORPH3, BA-test is an in-the-wild dataset assembled from VGGFace2 [14], with images selected based on head pose, brightness, and image quality, and with mitigation of label noise. BA-test has 8 demographic groups: 45,642 images of 3,631 White males (WM), 13,311 images of 288 Asian males (AM), 10,610 images of 577 Black males (BM), 11,091 images of 244 Indian males (IM). Since facial hair is a gender-related attribute, we only use male images from both datasets. We use two versions of ArcFace (R100 backbone) for real facial hair analysis, one trained with MS1MV2 [15] and one trained with Glint360K [8]. These two datasets are in-the-wild, identity-cleaned, with millions of images.

We evaluate four facial hair conditions: (1) no facial hair, (2) mustache with no beard, (3) chin-area beard, and (4) side-to-side beard. In the terminology of [36], these correspond to attributes of clean-shaven and mustache-none (CS-

Figure 3. Examples of synthetic facial hair pixel value and pixel area manipulation. **Pixel area**: *chin area beard + mustache none* attribute (Top two rows), *side-to-side + mustache none* attribute (Middle two rows), and *clean shaven + mustache isolated* attribute (Bottom two rows). **Pixel value**: *black, average value of hair, average value of skin, white*, and *cropped beard* (from left to right).

MN), clean-shaven beard and mustache-isolated (CS-MI), chin-area beard and mustache-none (CA-MN) and side-to-side beard and mustache-none (S2S-MN), respectively. Table 1 shows the number of images for each attribute, for each race, for MORPH3 and BA-test. We use the attribute classifier from [1, 36] with a threshold value of 0.9 to select images. The backbone of this classifier is SE-ResNeXt101. Note that some races do not have sufficient samples of some attributes. For example, there are only 9 S2S-MN AM samples selected from BA-test, 60 S2S-MN IM samples selected from BA-test, and 60 S2S-MN AAM samples selected from MORPH3. Results for this groups are inherently less reliable due to the small number of images.

3.1. Chin Area, Side to Side, Mustache Isolated

The difference, measured by d′ between CS-MN and {CA-MN, CS-MI, S2S-MN} across four races within the BA-test dataset is shown in Figure 2. Due to s2s beard with no mustached being a rare hairstyle, we disregard results not supported by a sufficient amount of data. There are facial hair effects for impostor image pairs that are consistent across across three races and two face matchers. CS-MI impostor pairs have the highest similarity, S2S-MN the next highest, and CA-MN the lowest similarity. For impostor image pairs, lower similarity means better accuracy. So, impostor image pairs with mustache-only facial hairstyle are at increased risk of a false match. The other results are in Figure 1 and Figure 2 of the Supplementary Material. Note that, AAM and BM have inconsistent pattern than the other race. We speculate that it is caused by the close pixel distribution between facial hair and face skin for dark skin people. To evaluate this speculation and find the potential trend of the rarely appeared facial hair attributes, we conducted additional experiments with synthetic facial hair regions.

Dataset	Race	CS*	CA*	CS†	S2S*
BA-test	BM	1,899	492	111	112
	WM	18,968	757	426	533
	AM	9,102	282	168	9
	IM	2,598	144	1,060	60
MORPH	AA-M	2,267	931	1,531	60
	C-M	4,958	2,314	978	324

Table 1. Number of images selected from MORPH and BA-test for each race and attribute. * does not have mustache (MN), † has isolated mustache (MI). CS, CA and S2S are clean shaven, chin area beard, and side-to-side beard.

4. Effects of Synthetic Facial Hair Regions

Current GAN tools for face attribute editing are not strongly identity preserving and do not enable facial hair edits to the specific regions analyzed here. Therefore, we use face landmark points to target synthetic facial hair edits to specific regions, without changing pixel values in the rest of the face image. Since this approach does not produce as photo-realistic facial hair results as a GAN might, we conducted experiments with five different brightness/color values for the synthetic facial hair region, in order to ensure the patterns of results are consistent across five synthetic facial hair edits and three matchers. We add MagFace matcher results in this section, using R100 backbone trained with MS1MV2.

We use the same attribute classifier with a threshold value of 0.9 to select no-facial-hair images. There results in 2,267 AAM and 4,958 CM images selected from MORPH3 dataset. We use [13] to obtain 68 landmark points on each face, and use a subset of the landmark points to anchor edits for synthetic facial hair regions. Pixels in the synthetic facial hair region are assigned one of five color values: black, white, average RGB of skin region of the particular image, average RGB of scalp hair region of the particular image, and a sample of real beard pixel values. These variations are illustrated in Figure 3. To get the value of hair and skin, we use BiSeNet [39] to segment the face skin and scalp hair regions, then average the pixel values for those regions.

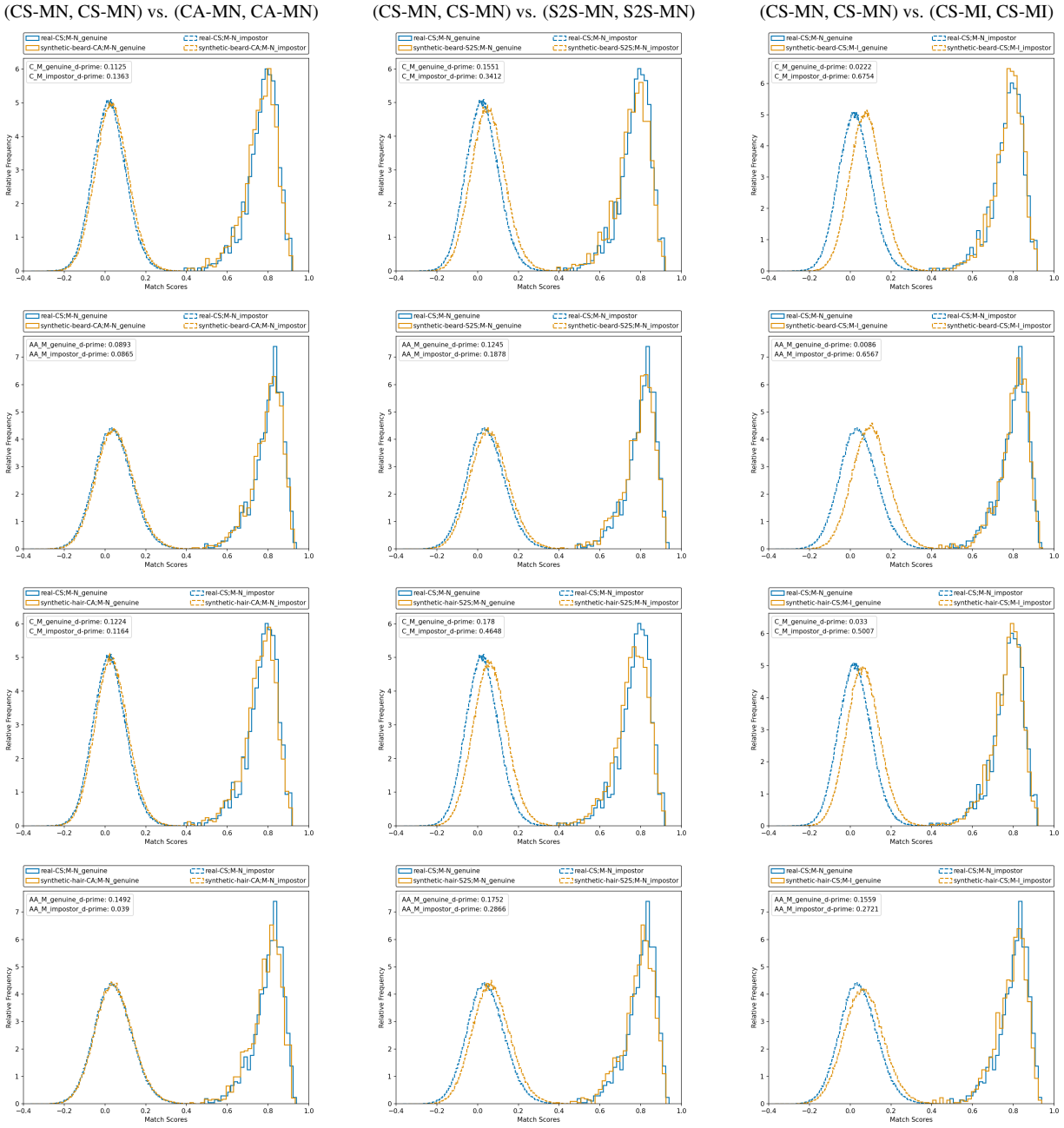


Figure 4. How Does Facial Hair Change Impostor and Genuine Distributions? The baseline impostor and genuine distributions for no facial hair are the same across the three plots in the 1st and 3rd rows for Caucasian, and across the three plots in the 2nd and 4th rows for African-American. Comparison plots in the {left, middle, right} column are for the same images with a {chin-area beard region, mustache region, side-to-side beard region} edited into each image. The beard pixels (top two rows) and hair pixels (bottom two rows) are added. The face matcher is ArcFace-R100, trained on the MS1MV2 dataset.

Effects of facial hair on impostor and genuine distributions. We evaluate the effect of facial hair on the impostor and genuine distributions by comparison to the distributions for images with no facial hair. Figure 3 shows that CS-MI occupies fewer pixels than CA-MN, which occupies fewer pixels than S2S-MN. On average, the number of pixels added in the chin area beard is **more than twice** the number for mustache area, and the side-to-side area is **over 4.5 times** the number of pixels for the mustache area. A reasonable initial expectation is that the size of effect on the impostor and genuine distributions would follow the size of the facial hair region.

Dataset	CS	CA	S2S	MN	MI	M-CTB
MSV2	54.8	10.9	9.7	70.7	9.8	9.9
Glint	50.7	10.1	13.4	64.4	10.5	12.9

Table 2. Fraction of images classified with various facial hair attributes, MS1MV2 and Glint360K. CS = clean shaven, CA = china area beard, S2S = side-to-side beard, MN = no mustache, MI = mustached isolated from beard, M-CTB = mustache connected to beard.

Figure 4 contrasts the genuine and impostor distributions resulting from three facial hair conditions with the baseline no-facial-hair condition. The impact of each facial hair condition is measured by $\Delta d'$ for each of the impostor and the genuine distributions. The results indicate that, for both Caucasian and African-American, adding synthetic facial hair has negligible effect on the genuine distribution. That is, it makes little to no difference in the similarity score for two images of the same person whether the person is clean-shaven in both, or has the same facial hairstyle in both. However, for the impostor distribution, CS-MI and S2S-MN facial hairstyles substantially degrade the impostor distribution by shifting it to higher similarity, while CA-MN does not have as strong of an impact as CS-MI and S2S-MN with two different pixel distributions added. It echos back to the conclusions in the real facial hair experiments.

With the respect to the filled pixels {real beard, hair}, for Caucasian males, S2S-MN area type has a {150% and 299%} $\Delta d'$ increase compared to CA-MN, and mustache isolated has a {396%, 330%} increase compared to CA-MN. For African American, S2S-MN has a {117%, 635%} increase and mustache isolated has a {659%, 598} increase of $\Delta d'$ compared to CA-MN, respectively. More results are in Figure 3 to Figure 8 in the Supplementary Material. Note that, the inconsistent pattern when adding average skin pixels is because of the low pixel distribution difference between the facial hair area and face skin. It is consistent with the observations on AAM and BM in real facial hair experiments.

General conclusions from these results are as follows. One, similar facial hairstyle in a pair of images affects impostor comparisons more than genuine comparisons. Two, as the difference between face skin area pixel distribution and facial hair area pixel distribution going larger, the effect of position becomes larger than the area size. These two conclusions are consistent with the observations in the real facial hair experiments and it provides a potential trend of how pixel area and value affect the face recognition model.

5. Accuracy Disparity Mitigation

To investigate the observed patterns' cause, we collected facial hair predictions from two training sets. Images with logically inconsistent predictions, such as both clean-shaven and beard-at-chin-area predicted as True, are not included in the analysis. However, logically inconsistent predictions occurred for only around 1.5% of the images, so any effect from dropping them should be small. Table 2 illustrates the fraction of each attribute in MS1MV2 and Glint360K. For both datasets, CA, S2S, and MI images occur with very similar frequency, in both datasets. *This suggests that the accuracy differences observed for these facial hairstyles likely cannot be attributed to an imbalance in their occurrence in the training data.*

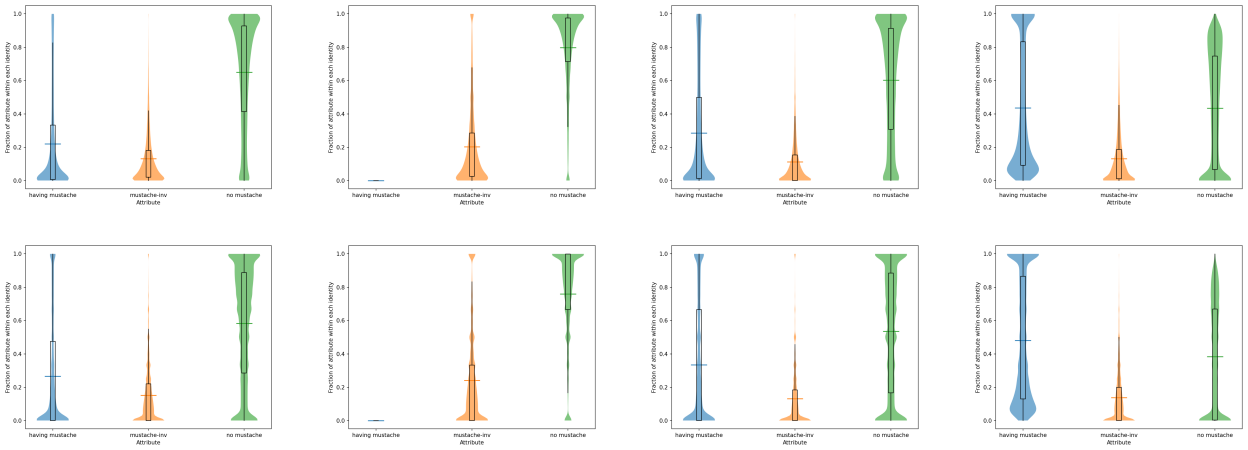
The difference effect of CA-MN and S2S-MN on accuracy could be explained by the size of the face region involved in each, and the explanation of the reduced accuracy for CS-MI image pairs is that central part is more important than peripheral part for face recognition models. Subsequently, we focus on the CS-MI hairstyle in this section and try to answer *Can the accuracy discrepancy be mitigated by frequency of facial hairstyles in the training data?* To explore this possibility, we create three subsets of MS1MV2 with the same number of images but different distributions of facial hairstyle (mustache), and the same with the Glint360K. The we train matchers with the training sets that have equal numbers of images but different frequencies of mustache, and compare the accuracies achieved. Matchers are trained on four RTX6000 GPUs. Due to GPU memory size, we utilized an R100 backbone for ArcFace loss training with MS1MV2, but an R50 backbone for the larger Glint360K-derived training sets. We use the same package[1] in [6].

5.1. Varying Mustache Representation in Training Data

First, images in MS1MV2 and in Glint360K are categorized into three categories - *having mustache*, *no mustache*, and *mustache information not visible (mustache-inv)*. Figure 5 shows the distribution of identity densities (width) at each fraction (y-axis) for each group (x-axis). Figure 5a indicates that only a small fraction of identities in the training set have images with mustaches. This under-representation of mustache images in the training data is a factor that could affect the learned model's accuracy of handling mustache images.

To explore this, we created three same-image-number subsets of each of MS1MV2 and Glint360K. For the first and second of the three subsets, we dropped half of the no mustache and mustache-inv images (i.e. 2,315,779 from MS1MV2 and 6,472,392 from Glint360K). This gives us a baseline matcher trained with higher mustache representation ratio in the training data. For the third of the three subsets, we drop the same number of images including all *having mustache* images and part of *no mustache* and *mustache-inv* images to make sure that all three subsets have

[1] https://github.com/vitoralbiero/face_analysis_pytorch



(a) Original (b) ZM (c) HMR-Random (d) HMR-ID

Figure 5. Density of identities corresponding to the fraction of each mustache attribute across three dataset manipulation strategies. **ZM** means the dataset does not have mustache images. **HMR-Random** means the dataset has higher ratio of mustache images by randomly dropping no-mustache and mustache-inv images from the original dataset, and **HMR-ID** means the dataset has higher ratio of mustache images by preferentially dropping no-mustache and mustache-inv images from the identities that have large fraction of no-mustache and mustache-inv images. MS1MV2 (top), Glint360K (bottom).

$\Delta d'$	Dataset	Strategies	C_M		AA_M	
			CA-MN	S2S-MN	CA-MN	S2S-MN
Genuine	MS1MV2	Original	0.3675	0.5563	0.0494	-
		ZM	0.4356	0.6044	0.0407	-
		HMR-Random	0.3407	0.4659	0.0379	-
		HMR-ID	0.3613	0.4297	0.0296	-
	Glint360K	Original	0.2793	0.4376	0.0155	-
		ZM	0.2904	0.4106	0.048	-
		HMR-Random	0.1909	0.3724	0.0405	-
		HMR-ID	0.2025	0.3537	0.0012	-
Impostor	MS1MV2	Original	0.3379	0.1125	0.0086	-
		ZM	0.2631	0.0488	0.1235	-
		HMR-Random	0.2054	0.0228	0.0018	-
		HMR-ID	0.2416	0.0491	0.042	-
	Glint360K	Original	0.3089	0.1587	0.0446	-
		ZM	0.4948	0.1491	0.1558	-
		HMR-Random	0.3844	0.148	0.0426	-
		HMR-ID	0.3905	0.1463	0.0061	-

Table 3. $\Delta d'$ measurement between *CS-MI* and *listed attributes* on both genuine and impostor distributions of the MORPH dataset. Eight models trained with three training set manipulation strategies and two training sets are used to calculate the values. The worst and the best performances are shown in Red and Green.

the same number of images. We use different strategies for dropping *no mustache* images for the first and second subsets. For one, *no mustache* and *mustache-inv* images are randomly selected and dropped. Randomly selecting images leaves the number of identities that have and that do not have mustache images relatively unaffected. For the second, *no mustache* and *mustache-inv* images are selected preferentially from identities that have the highest frequency of *no mustache* and *mustache-inv* images. There are many identities that have 100% *no mustache* images, and some of these identities are then no longer represented in the dataset. The result is that the first and second training sets have the same number of mustache images, but the third dataset also has a higher ratio of identities with mustache images. We refer to the three training sets as high mustache ratio - random (HMR-Random), high mustache ratio - iden-

Dataset	LFW	CFP-FP	AgeDB-30	Acc
Original	99.81	98.40	98.05	98.75
ZM	99.81	97.89	98.12	98.61
HMR-Random	99.77	97.89	97.98	98.55
HMR-ID	99.70	98.14	97.97	98.60
Original	99.80	98.61	98.33	98.91
ZM	99.82	98.63	98.08	98.84
HMR-Random	99.78	98.63	98.20	98.87
HMR-ID	99.77	98.61	98.17	98.85

Table 4. Accuracy of the models trained with manipulated datasets. MS1MV2 (top) and its variations are used to train ArcFace-R100. Glint360K (bottom) and its variations are used to train ArcFace-R50.

tity (HMR-ID) and zero mustache (ZM). Figure 5c shows that HMR-Random has a similar distribution pattern compared to the full original dataset in terms of identity density at each fraction. However, Figure 5d shows that HMR-ID has a noticeably altered distribution pattern compared to the full original dataset.

Table 4 shows the performance of the two face matchers (ArcFace-R100 trained with MS1MV2 and MS1MV2-derived trainings sets, and , ArcFace-R50 trained with Glint360K and Glint360K-derived training sets. We use LFW [17], CFP-FP [32], and AgeDB-30 [25] to measure the performance. The largest accuracy model difference for ArcFace-R100 and ArcFace-R50 in going from the full dataset to the mustache-manipulated subsets is 0.2% and 0.07%, respectively. This suggests that reducing the number of images by the amount done here does not impact the general performance of the face matcher.

Due to different style preferences across demographics and the uniqueness of S2S-MN, the number of samples does not support strong conclusions for AM (9 samples), BM (112 samples), and IM (60 samples) in the BA-test dataset. We conclude the pattern mainly based on the data from MORPH3, shown in Table 3. One general conclusion is that, across both datasets and races, dropping no mustache samples from the dataset reduces the difference between CS-MI and CA-MN and S2S-MN. Also, dropping all mustache samples increases the difference between mustache and other two beard areas. To give a general trend for each strategy, the trend is measured as:

$$\frac{1}{2}\sum_{D\in D'}\frac{\delta d'_{\{s;D\}}-\delta d'_{\{orig.;D\}}}{\delta d'_{\{orig.;D\}}} \quad (1)$$

Where $\delta d_{\{s;D\}}$ is the d' value from the model trained with dataset D manipulated by strategy s.

For genuine pairs, on average, Caucasian males have a 6.24% increase on $\Delta d'$ for the model trained with ZM dataset than trained with original dataset, a 16.43% decrease for the model trained with HMR-Random dataset, and a 17.82% decrease for the model trained with HMR-ID dataset. African American males have a 96% increase on $\Delta d'$ for the model trained with ZM dataset, and a 69.01% increase for the model is trained with a HMR-Random dataset, but a 66.17% decrease for the model trained with HMR-ID dataset. For impostor pairs, on average, Caucasian males, surprisingly, have a 6.16% decrease on $\Delta d'$ when the model trained with ZM dataset than trained with original dataset, a 25.31% decrease for HMR-Random, and a 16.56% decrease for HMR-ID. African American males experience a 7.93% increase on $\Delta d'$ when the model trained with ZM dataset, a 41.78% decrease for HMR-Random, and a 151% increase for HMR-ID.

6. Conclusion and Discussion

This work investigates the impact of the region size and location of facial hair on face recognition accuracy. The results from real facial hair experiments indicate that CS-MI has competitive or larger impact than S2S-MN, which is larger than CA-MN. This trend holds across four races and two face matchers.

Due to the a low number of real facial hair images for some races, and mis-classification on AAM, we adopt a strategy of adding synthetic facial hair regions to images to dig out the potential trend for these cases. The CS-MI facial hair region is **less than half the size** of CA-MN and S2S-MN but causes **over two times** the shift in impostor distribution of the other two. Our results show that the pattern of impact of CS-MI, CA-MN, S2S-MN is consistent with the observation on real facial hair when the added pixels have larger difference from skin color. When the added pixels (i.e. average value of face skin pixels) is similar with the skin color, S2S-MN has the largest impact.

To reduce the accuracy discrepancy among facial hair areas, we adjust the facial hair distributions in the training sets in three ways: 1) dropping all mustache samples, 2) randomly dropping no mustache samples, and 3) targeted dropping no mustache samples. After re-training face recognition models, the general performance does not have a considerable difference. The results show that adjusting frequency of the attribute occurrence can change the accuracy disparity across demographics. Specifically, omitting all mustache samples accentuates the discrepancy, while excluding samples without mustaches decreases this discrepancy by at least 40%. This suggests that adding more images containing mustaches to the training datasets can result in less accuracy disparity among samples with different facial hairstyles.

References

[1] https://github.com/HaiyuWu/LogicalConsistency#testing, last accessed on July 2023. 4

[2] Morph dataset. https://www.faceaginggroup.com/. 2

[3] Vítor Albiero and Kevin W. Bowyer. Is face recognition sexist? no, gendered hairstyles and biology are. In *31st British Machine Vision Conference 2020, BMVC 2020*, 2020. 1

[4] Vitor Albiero, Xingyu Chen, Xi Yin, Guan Pang, and Tal Hassner. img2pose: Face alignment and detection via 6dof, face pose estimation. In *CVPR*, pages 7617–7627, 2021. 3

[5] Vitor Albiero, Krishnapriya Ks, Kushal Vangara, Kai Zhang, Michael C King, and Kevin W Bowyer. Analysis of gender inequality in face recognition accuracy. In *WACVW*, pages 81–89, 2020. 1

[6] Vítor Albiero, Kai Zhang, and Kevin W Bowyer. How does gender balance in training data affect face recognition accuracy? In *2020 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–10. IEEE, 2020. 1, 6

[7] Vítor Albiero, Kai Zhang, Michael C King, and Kevin W Bowyer. Gendered differences in face recognition accuracy explained by hairstyles, makeup, and facial morphology. *TIFS*, 17:127–137, 2021. 1

[8] 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 *CVPR*, pages 1445–1449, 2021. 3

[9] Aman Bhatta, Vítor Albiero, Kevin W Bowyer, and Michael C King. The gender gap in face recognition accuracy is a hairy problem. In *WACVW*, pages 303–312, 2023. 1, 2

[10] Aman Bhatta, Domingo Mery, Haiyu Wu, Joyce Annan, Micheal C King, and Kevin W Bowyer. Our deep cnn face matchers have developed achromatopsia. *arXiv preprint arXiv:2309.05180*, 2023. 1

[11] Aman Bhatta, Gabriella Pangelinan, Micheal C King, and Kevin W Bowyer. Demographic disparities in 1-to-many facial identification. *arXiv preprint arXiv:2309.04447*, 2023. 1

[12] Andrew Brock, Theodore Lim, James M Ritchie, and Nick Weston. Neural photo editing with introspective adversarial networks. *ICLR*, 2017. 2

[13] Adrian Bulat and Georgios Tzimiropoulos. How far are we from solving the 2d & 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks). In *ICCV*, 2017. 4

[14] Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, pages 67–74. IEEE, 2018. 3

[15] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *CVPR*, pages 4690–4699, 2019. 1, 3

[16] 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. 1

[17] Gary B Huang, Marwan Mattar, Tamara Berg, and Eric Learned-Miller. Labeled faces in the wild: A database forstudying face recognition in unconstrained environments. In *Workshop on faces in'Real-Life'Images: detection, alignment, and recognition*, 2008. 8

[18] Ziqi Huang, Kelvin CK Chan, Yuming Jiang, and Ziwei Liu. Collaborative diffusion for multi-modal face generation and editing. In *CVPR*, pages 6080–6090, 2023. 2

[19] Yuming Jiang, Ziqi Huang, Xingang Pan, Chen Change Loy, and Ziwei Liu. Talk-to-edit: Fine-grained facial editing via dialog. In *ICCV*, pages 13799–13808, 2021. 2

[20] Minchul Kim, Anil K Jain, and Xiaoming Liu. Adaface: Quality adaptive margin for face recognition. In *CVPR*, pages 18750–18759, 2022. 1

[21] KS Krishnendu. Analysis of recent trends in face recognition systems. 2023. 1

[22] Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. Maskgan: Towards diverse and interactive facial image manipulation. In *CVPR*, pages 5549–5558, 2020. 2

[23] Qiang Meng, Shichao Zhao, Zhida Huang, and Feng Zhou. MagFace: A universal representation for face recognition and quality assessment. In *CVPR*, 2021. 1

[24] Umar Mohammed, Simon JD Prince, and Jan Kautz. Visiolization: generating novel facial images. *ACM Transactions on Graphics (ToG)*, 28(3):1–8, 2009. 2

[25] Stylianos Moschoglou, Athanasios Papaioannou, Christos Sagonas, Jiankang Deng, Irene Kotsia, and Stefanos Zafeiriou. Agedb: the first manually collected, in-the-wild age database. In *CVPR workshops*, pages 51–59, 2017. 8

[26] Minh Hoai Nguyen, Jean-Francois Lalonde, Alexei A Efros, and Fernando De la Torre. Image-based shaving. In *Computer graphics forum*, volume 27, pages 627–635. Wiley Online Library, 2008. 2

[27] Kyle Olszewski, Duygu Ceylan, Jun Xing, Jose Echevarria, Zhili Chen, Weikai Chen, and Hao Li. Intuitive, interactive beard and hair synthesis with generative models. In *CVPR*, pages 7446–7456, 2020. 2

[28] Kagan Ozturk, Grace Bezold, Aman Bhatta, Haiyu Wu, and Kevin Bowyer. Beard segmentation and recognition bias. *arXiv preprint arXiv:2308.15740*, 2023. 2

[29] P Jonathon Phillips, Patrick J Flynn, Todd Scruggs, Kevin W Bowyer, Jin Chang, Kevin Hoffman, Joe Marques, Jaesik Min, and William Worek. Overview of the face recognition grand challenge. In *CVPR*, volume 1, pages 947–954. IEEE, 2005. 1

[30] Christian Rathgeb, Pawel Drozdowski, Dinusha C Frings, Naser Damer, and Christoph Busch. Demographic fairness in biometric systems: What do the experts say? *IEEE Technology and Society Magazine*, 41(4):71–82, 2022. 1

[31] Karl Ricanek and Tamirat Tesafaye. Morph: A longitudinal image database of normal adult age-progression. In *7th International Conference on Automatic Face and Gesture Recognition (FGR06)*, pages 341–345. IEEE, 2006. 2

[32] Soumyadip Sengupta, Jun-Cheng Chen, Carlos Castillo, Vishal M Patel, Rama Chellappa, and David W Jacobs. Frontal to profile face verification in the wild. In *2016 IEEE winter conference on applications of computer vision (WACV)*, pages 1–9. IEEE, 2016. 8

[33] 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. *IEEE Transactions on Technology and Society*, 3(1):16–30, 2021. 2

[34] Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei Liu. Cosface: Large margin cosine loss for deep face recognition. In *CVPR*, pages 5265–5274, 2018. 1

[35] Haiyu Wu, Vítor Albiero, KS Krishnapriya, Michael C King, and Kevin W Bowyer. Face recognition accuracy across demographics: Shining a light into the problem. In *CVPR workshops*, pages 1041–1050, 2023. 1

[36] Haiyu Wu, Grace Bezold, Aman Bhatta, and Kevin W Bowyer. Logical consistency and greater descriptive power for facial hair attribute learning. In *CVPR*, pages 8588–8597, 2023. 1, 2, 3, 4

[37] Haiyu Wu and Kevin W Bowyer. What should be balanced in a "balanced" face recognition dataset? *BMVC*, 2023. 1, 2

[38] Xu Yao, Alasdair Newson, Yann Gousseau, and Pierre Hellier. A latent transformer for disentangled face editing in images and videos. In *ICCV*, pages 13789–13798, 2021. 2

[39] Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang. Bisenet: Bilateral segmentation network for real-time semantic segmentation. In *ECCV*, 2018. 4

[40] Seyma Yucer, Furkan Tektas, Noura Al Moubayed, and Toby P Breckon. Racial bias within face recognition: A survey. *arXiv preprint arXiv:2305.00817*, 2023. 1