The Impact of Racial Distribution in Training Data on Face Recognition Bias: A Closer Look

Manideep Kolla
HyperVerge Inc.
manideep@hyperverge.co

Aravinth Savadamuthu
HyperVerge Inc.
aravinth.muthu@hyperverge.co

Abstract

Face recognition algorithms, when used in the real world, can be very useful, but they can also be dangerous when biased toward certain demographics. So, it is essential to understand how these algorithms are trained and what factors affect their accuracy and fairness to build better ones. In this study, we shed some light on the effect of racial distribution in the training data on the performance of face recognition models. We conduct 16 different experiments with varying racial distributions of faces in the training data. We analyze these trained models using accuracy metrics, clustering metrics, UMAP projections, face quality, and decision thresholds. We show that a uniform distribution of races in the training datasets alone does not guarantee bias-free face recognition algorithms and how factors like face image quality play a crucial role. We also study the correlation between the clustering metrics and bias to understand whether clustering is a good indicator of bias. Finally, we introduce a metric called racial gradation to study the inter and intra race correlation in facial features and how they affect the learning ability of the face recognition models. With this study, we try to bring more understanding to an essential element of face recognition training, the data. A better understanding of the impact of training data on the bias of face recognition algorithms will aid in creating better datasets and, in turn, better face recognition systems.

1. Introduction

Racial bias in face recognition (FR) systems is a widely acknowledged problem. Multiple studies have shown the racial inequity of these systems and the social injustice it translates into. In January 2020, Detroit police arrested Robert Williams in front of his family. It soon turned out that Mr. Williams was innocent and had been wrongfully arrested because an FR system had wrongly matched him with a suspect's face, which started the devastating series of events that happened [38, 16].

Studies in the past have shown that there have been multiple orders of magnitude of improvement in the performance of face recognition algorithms over the last two decades [13]. Despite these improvements, demographic bias in these algorithms is still a critical problem and affects real people. A recent study from NIST [12] analyzed hundreds of algorithms and documented the difference in accuracies across different races. The study has shown that the false match rates (FMR) vary between 10 to 100 times across different demographics, and the false match differentials are much higher than that of false non-match, which vary by a factor of at most 3. The study reports that the East African, West African, and East Asian cohorts have the highest false match rates, and the Eastern European cohort has the lowest. This study has also shown that these algorithms falsely identified Native Americans more often than people from other demographics from U.S. law enforcement images. One interesting finding is that many algorithms built by developers from China have shown different characteristics compared to other algorithms with lower false match rates on East Asian faces, indicating how the algorithms are built, and the sourcing of the training data plays an essential role in determining their performance.

The causes of bias in face recognition systems are multidimensional, with statistical, human, and systemic biases. Many modern face recognition systems rely on large-scale data, which are generally scrapped from the internet. This has led to availability as the prime criterion taking precedence over suitability to the task [28].

American Civil Liberties Union said in a statement that "One false match can lead to missed flights, lengthy interrogations, watch list placements, tense police encounters, false arrests or worse" [30]. Facial recognition technology has a particularly high potential to cause harm when it targets children because it is less accurate at identifying children [2, 31]. To build robust algorithms that are fair and work for everyone, we need to curate more real-world datasets and examine the existing ones.

In recent years, numerous studies have empirically analyzed the bias in face recognition and face attribute analysis systems, with insights on what contributes to bias and ways

to measure it [19, 23, 4, 26, 34, 1, 18, 17]. Research has also shown promising results in mitigating bias in these systems [10, 35, 11, 20].

Given the societal implications of biased face recognition algorithms, it becomes critical to study the datasets used for training these models and how they affect recognition performance. To bring more focus and resolve to bias in face recognition, we extensively study the effects of using different racial distributions in the training datasets on the face recognition models in terms of accuracy, clustering metrics, intra and inter race similarities, face quality, UMAP projections, and decision thresholds.

2. Related Work

With the advancements in deep learning, the new age face recognition (FR) algorithms have come a long way reaching near human-level performance and performing better than humans in tasks like large-scale face search systems. Methods proposed in [33, 32, 25, 27, 37, 7] use different variations of neural networks, mainly convolutional neural networks, to embed the input faces into d-dimensional feature vectors, which would serve as a low-dimensional representation of faces, encoding different features of the face. There are many ways to train a neural network to encode the faces into these vector representations in such a way that these feature vectors can be used to compare the similarity between faces using metrics like euclidean and cosine distance.

Taigman *et al.* [33] and Sun *et al.* [32] propose to learn the vector representations by training a convolutional network using a softmax-based cross-entropy loss with identities as classes. In [25] and [27], the vector representations of the faces are learned by training a convolutional network to minimize the triplet loss between an anchor, positive (mated), and a negative (non-mate) triplet of faces. Later, angular margin-based loss functions were proposed, which proved to be far more discriminative and accurately learn the face embeddings than the existing triplet-based and softmax-based losses without angular margin [37, 7].

Although the current FR algorithms are very accurate and cross human-level performance, many recent studies have pointed out the biases these FR algorithms learn and how they affect people from different demographics.

Krishnapriya *et al.* [19] have shown that the false match rate (FMR) is higher in the African-American cohort, and the false non-match rate (FNMR) is higher in the Caucasian cohort at a fixed decision threshold. Apart from this, they have not found any clear evidence for the presumption that darker skin tone causes a higher false match rate. Nagpal *et al.* [23] show that upon limited exposure to other races, face recognition algorithms mimic the human inclination of own-race bias. They also show that the networks trained on faces from a specific race encode the race-specific regions of interest. Robinson *et al.* [26] show a notable boost in the overall performance of face recognition algorithms by learning subgroup-specific thresholds instead of using a global threshold. The experiments in [34] shed light on face quality estimation and its effect on face recognition accuracy across demographics. Bar-Haim *et al.* [1] show that differences in morphological features have a more significant role than skin color in causing bias in face recognition in humans. A similar skin tone between a pair of images increases the likelihood of false matches, but a darker skin tone in itself is not responsible for the observation. Kortylewski *et al.* [18] study the effect of dataset biases in the form of features like lighting and pose using controllable synthetically generated faces and show that these biases significantly impact the generalization of the FR algorithms.

Recently, several methods have been proposed to mitigate demographic bias in face recognition algorithms. Although the research on debiasing deep learning based FR algorithms is still nascent, these methods have shown great promise. Gong *et al.* [10] propose an adversarial training setup that extracts disentangled feature representations to address bias in face recognition. Wang *et al.* [35] propose a reinforcement learning based race balance network to select appropriate margins for use in the large margin losses [37, 7] for non-Caucasians to learn balanced performance for different races and published BUPT-GlobalFace and BUPT-Balancedface datasets to facilitate studies into bias in face recognition algorithms. In [11], a training methodology with adaptive convolution kernels and attention mechanisms is proposed that adapt based on the demographic attributes of the faces to mitigate bias. Li *et al.* [20] formulate debiasing as a signal-denoising problem and propose a progressive cross-transformer architecture to denoise the identity-unrelated components induced by race from the identity-related components for fair face recognition.

Apart from algorithmic and architectural factors, one of the main reasons for biased face recognition algorithms is the presence of a non-uniform distribution of demographic classes, like racial distribution in the training datasets, which leads to unfair performance across different racial groups. In this study, to understand the effect of racial distribution in training datasets on the bias, we 1) train a face recognition model with a fixed architecture on different training datasets with different combinations of races, 2) compute the accuracy and clustering metrics of these trained models across all races to infer their bias, 3) study the correlation between bias and the clustering of faces based on race, 4) study the effect of face quality on the bias, 5) analyze the intra and inter race similarities, 6) visualize the clustering of faces from different races using a dimensionality reduction technique, and 7) study the difference in the decision thresholds of these trained models for different races and their correlation to bias.

3. Methodology

3.1. Network Architecture and Loss Functions

We use the ResNet-50 network [15, 7] as the backbone for our experiments. The backbone consists of 43.6 million parameters. The output of the backbone network is a 512-dimensional (512-D) vector which serves as the embedding vector of an input face. The backbone is followed by a classification layer that outputs classification logits. We use Additive Angular Margin Loss (ArcFace) [7] to modify the logits and softmax Cross-Entropy loss to calculate the final loss from the modified logits.

3.2. Training Data Preparation

Our goal is to understand the contribution of the faces from four available races in the BUPT-BalancedFace dataset [35] during training on bias in face recognition models. To this extent, we prepare 15 distinct training datasets from the BUPT-BalancedFace dataset by choosing combinations of all faces from one race at a time, two races at a time, three races at a time, and all four races at a time, as detailed in Tab. 1. We also train the model on the MS1MV3 dataset [14, 8], which consists of highly imbalanced data with respect to its racial distribution.

4. Experimental Settings

4.1. Implementation Details

Datasets: For training, we mainly use two datasets. First, the BUPT-BalancedFace dataset [35] contains face images with both identity and race labels. The dataset contains faces of people from four ethnic demographics: African, Asian, Caucasian, and Indian, with 7000 identities in each, with around 300,000 unique images in each and 1.25 Million images in total. Second, we use the MS1MV3 dataset [14, 8] as a general-purpose large-scale dataset for training. MS1MV3 contains around 5 Million images with approximately 91,000 identities. MS1MV3 is highly imbalanced with respect to its racial distribution and comprises 76.3% Caucasian, 14.5% African, 6.6% Asian, and 2.6% Indian [36]. For testing, we use the Racial Faces in-the-Wild (RFW) dataset [36] because it contains a uniform 6000 mated and 6000 non-mated pre-defined face pairs from each of the four ethnic demographics: African, Asian, Caucasian, and Indian. For both training and testing datasets, the faces are cropped and aligned using the RetinaFace face detector [6] to produce face crops of size 112×112.

Training Settings: We use Stochastic gradient descent (SGD) with an initial learning rate of 0.1, a momentum of 0.9, and a weight decay of 5×10^{-4}. We train all the experiments with a batch size of 512 for 35 epochs with a learning rate scheduler that decreases the learning rate tenfold at the 18th, 25th, 30th, and 33rd epochs.

4.2. Evaluation Metrics and Protocols

Accuracy metrics: We use verification accuracy as the performance metric similar to [10, 11, 35]. We use the predefined mated and non-mated pairs from the RFW dataset to calculate the accuracy of the trained models on all four races separately. We use 10-fold cross-validation to produce ten decision thresholds at which we attain the highest accuracy for each fold, and the final accuracies are calculated using the average of the 10-folds. Tab. 1 contains the accuracy metrics on different races in the RFW dataset. We also calculate accuracy on the combined RFW dataset with all races combined. This is indicated by "All" in Tab. 1. We use standard deviation between the accuracies across the four races to measure the magnitude of accuracy differentials between races across the models, which we use as one of the metrics to gauge bias [10, 11, 35].

Clustering metrics: We also compute the Calinski-Harabasz (CH) index [5] to understand the clustering ability and, in turn, the discriminative power of each model across races. CH index helps us understand the bias in face recognition models by gauging inter-race and intra-race cluster distance [9]. The intuition is to quantify how well the faces are clustered with respect to a specific characteristic like race. A good clustering of faces with respect to a specific characteristic like race suggests that the model has learned race-specific features apart from the identity-specific ones, potentially making the model susceptible to discrimination against people from different races. A higher value of the CH index indicates good clustering and, in turn, more bias. We computed the CH index in three scenarios for each model in Tab. 1. First, the index is computed using faces from all four races of the RFW dataset, followed by using only the races present during the training for a particular experiment, and last, the races that are not a part of the training. This will help us understand the clustering power of the models on the faces of ethnicities that are part of the training and those that are not. We discuss this in more detail in Sec. 5.2 and study whether clustering metrics like the CH index are an indicator of bias in FR algorithms. The clustering metrics are computed and detailed in Tab. 1.

Racial Gradation: We introduce the Racial Gradation metric to study the intra and inter race correlation in facial features across races. Gradation in facial features can be defined as the extent of similarity between faces from different races. To quantitatively measure the gradation, we calculate the intra and inter race average non-mated pair cosine distance across all the races by randomly sampling 100,000 non-mated pairs from the RFW test set for each intra and inter race setting. For a particular race A, if the inter-race average non-mated cosine distance is the lowest with a certain other race B, intuitively, the face recognition algorithm should be able to learn to recognize certain features from one race when it is trained only with the faces from the race

Training Data	Clustering Metrics ↑			Accuracy Metrics (in %)					
	CH-All	CH-T	CH-NT	African	Asian	Cauc.	Indian	All	STD
African+Asian+Caucasian+Indian	293.5	293.5	-	94.85	94.37	96.97	95.48	95.13	1.13
MS1MV3	199.5	199.5	-	96.75	96.42	99.02	97.32	97.07	1.16
African+Asian+Caucasian	329.1	121.1	-	94.23	94.65	96.55	92.05	93.42	1.85
African+Asian+Indian	511.1	365.9	-	93.57	93.92	92.08	94.72	92.88	1.10
African+Caucasian+Indian	796.0	348.4	-	93.75	85.60	96.32	94.45	90.39	4.74
Asian+Caucasian+Indian	863.0	218.3	-	83.05	93.30	96.03	94.23	86.56	5.85
African+Asian	567.2	158.2	823.8	92.50	92.52	90.87	89.38	90.05	1.50
African+Caucasian	867.0	145.8	1389.4	92.78	82.48	95.35	90.37	87.95	5.56
African+Indian	922.8	610.9	1263.3	92.25	83.38	90.43	92.93	88.57	4.37
Asian+Caucasian	989.4	25.4	1436.3	79.68	93.20	95.17	88.57	84.40	6.89
Asian+Indian	1065.4	70.8	2167.2	79.77	92.17	88.27	92.48	84.34	5.92
Caucasian+Indian	1250.0	317.5	2323.3	81.28	84.20	94.67	92.97	84.73	6.54
African	1042.1	-	1240.8	89.83	78.75	86.77	86.15	84.31	4.70
Asian	1377.6	-	1513.5	71.32	89.23	81.22	79.73	76.92	7.34
Caucasian	1332.7	-	1474.4	74.50	77.73	92.38	84.33	79.38	7.91
Indian	1418.3	-	1850.2	73.88	78.42	83.92	88.40	79.47	6.34

Table 1: This table depicts the 16 experiments conducted with different racial distributions in the training data along with the Calinski-Harabasz (CH) index, accuracies, and standard deviation of accuracies across the four racial cohorts available. Here, CH-All, CH-T, and CH-NT indicate the CH index calculated on the RFW test set with faces from all four races, from the races that are part of the training, and from the races that are not a part of the training, respectively. CH index is not calculated when only faces from one race are present. ↑ indicates that the higher the CH index, the better the clustering.

with which it has lowest average distance. These intra and inter race average non-mated pair cosine distances across different races are depicted in Fig. 4.

UMAP projections: We use the Uniform Manifold Approximation and Projection (UMAP) [21] to visualize how the faces from different races are clustered by different models trained with varying racial distribution by projecting higher dimensional face embedding vectors to lower dimensions. To do this, we extract the 512-dimensional embedding vectors of the faces from the RFW test set and use the UMAP algorithm to project these embeddings onto a 2-dimensional plane. This particularly helps us analyze how the clustering of faces differs from model to model depending on the training data settings as depicted in Fig. 2.

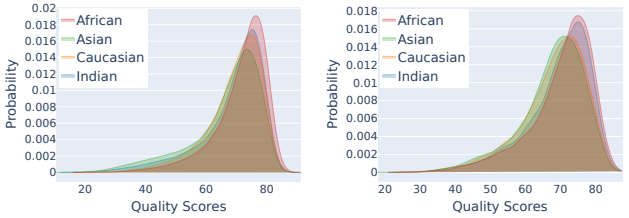
Face quality: To further understand whether face image quality plays a role in inducing bias, we calculate the face quality scores for training and test data using the face image quality assessment (FIQA) method proposed in [24]. This gives a positive quality score for each face; the higher the score, the better the face quality. Although face quality is an important factor to study, we should also note that this FIQA model might be biased in itself due to the data it is trained on. For train and test sets, Fig. 1 contains the distribution of these quality scores, and Tab. 2 contains the median and mean face quality scores.

Decision Thresholds: Finally, graphs in Fig. 5 show the cosine distance decision thresholds for each model at a false match rate (FMR) of 0.1% for each race in the RFW test set.

5. Results and Discussion

5.1. On Accuracy

We can observe an obvious pattern from Tab. 1 that, on average, the standard deviation is the highest for the models trained on a single race and lowest in the models trained on all races. Within each data setting in Tab. 1, even when the settings concerning the racial distribution are similar, the standard deviations still vary a lot. In particular, the models trained with African faces in the training set have lower standard deviations, followed by Indian faces, and models trained with Caucasian faces have some of the highest standard deviations. This is true for all the data settings, from single-race training to training with three races. In particular, among the models trained on three races, the standard deviation is the lowest when trained with African and Indian faces, followed by the models trained with African



(a) Training Data (b) Test Data

Figure 1: Face quality scores distribution

Race	Training Set		Test Set	
	Mean	Median	Mean	Median
African	71.59	73.66	70.10	72.33
Asian	66.59	69.74	67.50	69.18
Caucasian	69.26	71.26	67.92	69.62
Indian	68.73	71.69	69.33	71.36

Table 2: Mean and median face quality scores

and Asian faces. The models trained with Caucasian and without African faces have the highest standard deviation. This phenomenon can be attributed to two factors. First, the quality of the face images in training and test sets which we discuss in Sec. 5.3. Second, the gradation in facial features between different races is discussed in Sec. 5.4.

The standard deviations between the model trained on the full BUPT-BalancedFace dataset and MS1MV3 are similar, even though the MS1MV3 is racially imbalanced. This is because MS1MV3 is a much larger dataset than BUPT-BalancedFace, and the model trained on the former inherently has lower error rates than when trained on the latter and, in turn, lower standard deviation. As observed here, the model trained on a smaller dataset, such as BUPT-BalancedFace face but with balanced racial distribution, has a standard deviation similar to that of the MS1MV3 model, even though the latter has lower error rates across all races. This indicates that the models trained on larger sets have lower absolute differences in accuracies between races, but this does not inherently mean that they are less biased.

5.2. On Clustering Metrics

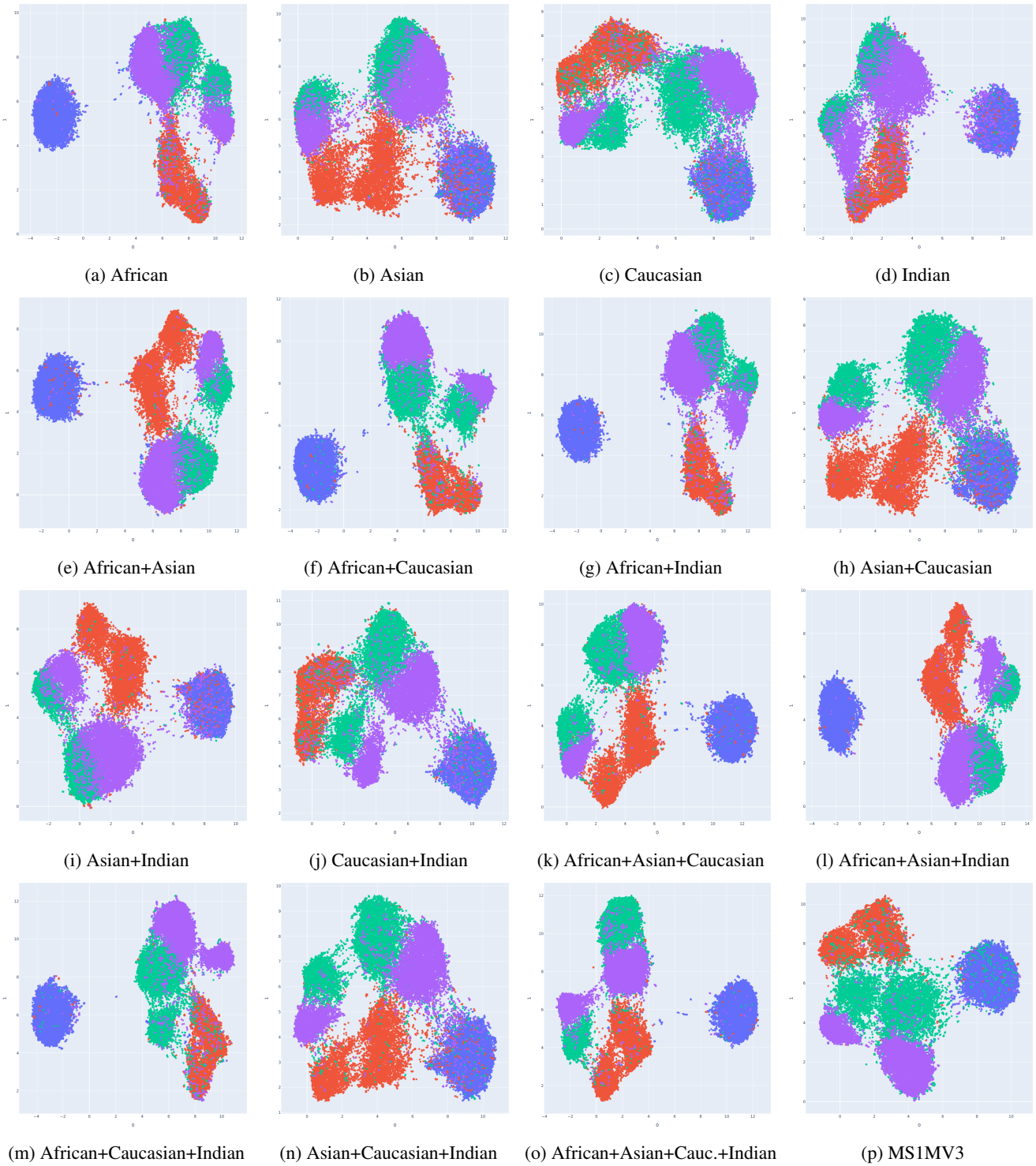
From Tab. 1, we can observe that the CH index is lower for in-train races and higher for out-of-training races. This indicates that the out-of-training races are getting clustered better than those in training, which contradicts the hypotheses made in Sec. 4.2. We can observe from Tab. 1 that the CH index and standard deviation do not have a monotonic relationship similar to the finding in [9]. In addition to this, to understand the extent of correlation between the CH index and standard deviation, we attempt to calculate Pearson's Correlation Coefficient [3] between these two metrics within each of the training data setting and across training data settings of Tab. 1. Pearson's Correlation Coefficient is 0.921 across training data settings, indicating a very high positive correlation between the two metrics. Whereas within each training data setting, the correlation is inconsistent; however, has a high correlation for the most part. Within the four-race setting, Pearson's Correlation is -1, indicating a perfect negative correlation as the number of settings for the four-race scenario is two, i.e, the sample size is two which are inversely correlated for calculating the Pearson's Correlation. Pearson's Correlation is 0.902 and 0.868 within the three and two race settings respectively indicating a high positive correlation between the two metrics. The positive correlation is the lowest in the one-race setting, with a Pearson's Correlation of 0.775. These Pearson's Correlation Coefficients indicate that except for the four race setting, there is a high correlation between the CH index and the standard deviation and particularly very high across training data settings from top to bottom of Tab. 1 compared to within each setting. This finding is in contrast to the findings of [9] which reports no link between clustering metrics like CH index and standard deviation. Also, with an increasing CH index from four to single race settings, the overall accuracy on the full RFW test set decreases, indicating that models with better clustering ability might not be more accurate.

5.3. On Dependence on Face Quality

We can observe from Fig. 1 and Tab. 2 that African faces have the highest quality and Asian have the lowest quality in both train and test sets. This correlates with the observations made in Sec. 5.1 that the models trained with African faces have significantly lower standard deviations than other races. This is because the models would be able to learn facial features better when the quality of the faces is better and will help in recognizing the faces from other races better than when the models are trained on faces of lower quality. We can particularly observe this in the single-race training setting, where the model trained with African faces also has the highest accuracies on other races after the models trained with the faces from those respective races.

5.4. On Gradation in Facial Features

An observation from Tab. 1, when the training data consists of African faces, the accuracy on the African test set is high and is comparable to the accuracy on the other three races when their respective data is used during training. However, when the African faces are not in the training set, the accuracy on the African test set is very low and is far lower than any other race in the test set in all training settings. The distinctly low accuracy on the African set in the



(a) African (b) Asian (c) Caucasian (d) Indian

(e) African+Asian (f) African+Caucasian (g) African+Indian (h) Asian+Caucasian

(i) Asian+Indian (j) Caucasian+Indian (k) African+Asian+Caucasian (l) African+Asian+Indian

(m) African+Caucasian+Indian (n) Asian+Caucasian+Indian (o) African+Asian+Cauc.+Indian (p) MS1MV3

- **Blue** indicate the African faces
- **Red** indicate the Asian faces
- **Green** indicate the Caucasian faces
- **Violet** indicate the Indian faces.

Figure 2: Two-dimensional UMAP projections of the face embeddings from the RFW test set computed using all the trained models.

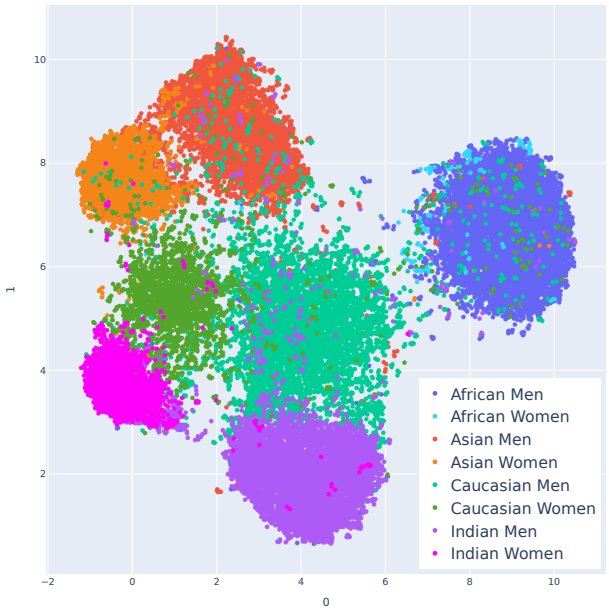


Figure 3: UMAP projections for the model trained on MS1MV3 data visualizing the gender sub-clusters.

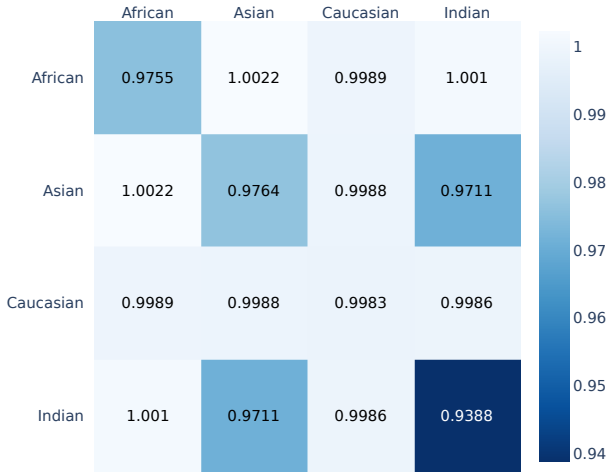


Figure 4: Gradation Matrix: Average cosine distances between non-mated face pairs across all races.

absence of African faces from training results in a high standard deviation. Apart from face quality, we can attribute this to the fact that the African faces have much lower similarity to faces from other races, unlike other races, which have at least one neighboring race, as observed in Fig. 4.

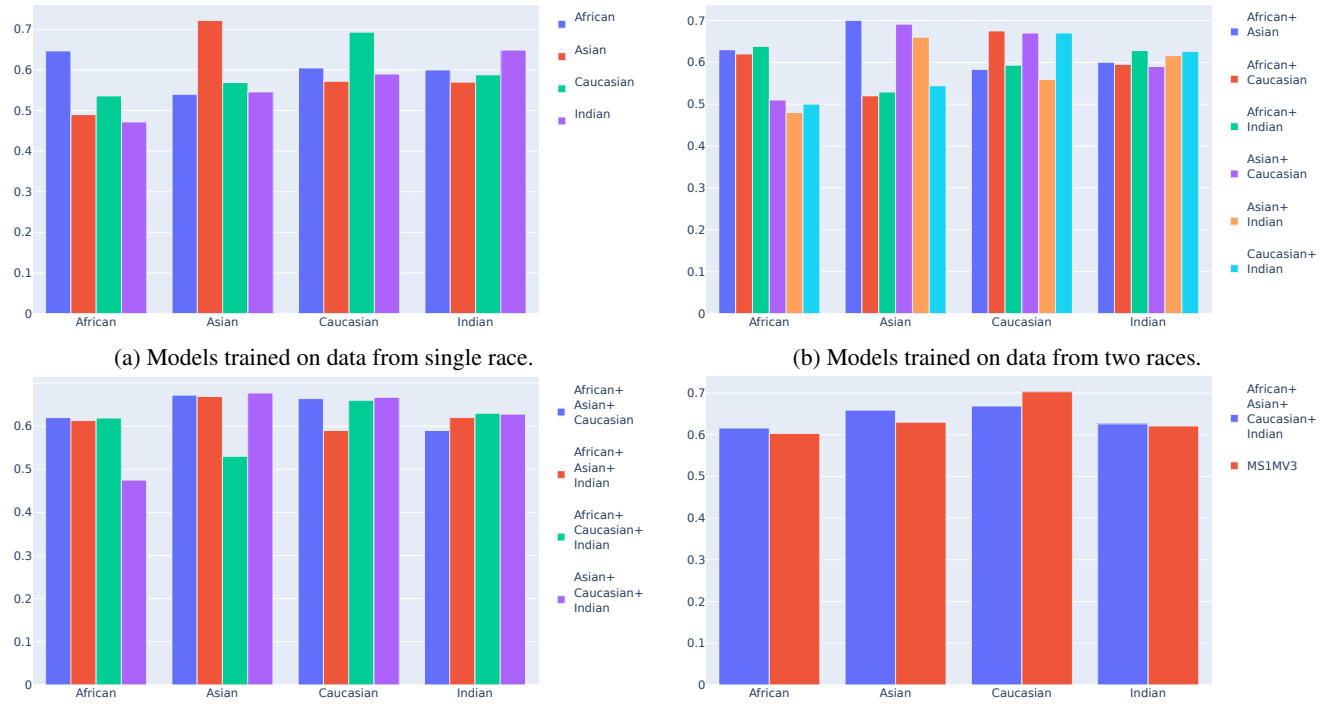
5.5. On Correlation Between Gradation and Face Quality

If two races are nearer, according to Fig. 4, the accuracies on those races should be the highest when both the races are present together during training, compared to all other cases. Even if a particular race is absent in the training set, a nearer race from Fig. 4 would provide reasonable information to learn the former race. In two race settings of Tab. 1, when the training data consists of both Asian and Indian faces, the accuracy on both Asian and Indian test data is expected to be higher compared to all other settings, but this is not the observation from Tab. 1.

Indian accuracy in the Asian+Indian setting (92.48%) is less than in the African+Indian setting (92.93%). Similarly, Indian accuracy in the African+Asian setting (89.38%) is less than in the African+Caucasian setting (90.37%). Although this is an exception to the assumption made above from the gradation point of view, these exceptions can be explained using the face quality scores. Where in the first exception, African data has much better face quality than Asian, and in the second exception, Asian data has the lowest face quality of all. Similarly, Indian accuracy in the Asian+Indian setting (92.48%) is marginally less than in the Caucasian+Indian setting (92.97%) because the quality of Asian faces is lower than that of Caucasian faces. Though African face quality is higher than Caucasian face quality, Indian accuracy in the Caucasian+Indian setting (92.97%) is marginally higher than that of the African+Indian setting (92.93%) because Caucasian race is closer to Indian race compared to African as observed in Tab. 4. Based on this, we conclude that face quality comes into play only when the quality difference is significant, like in the case of the first exception mentioned above, where African faces are of the highest quality, and Asian faces are of the lowest quality.

5.6. On UMAP Projections

With most models in Fig. 2, African faces form a distinct and distant cluster. African faces are best clustered when African faces are present in the training set as expected and discussed in Sec. 5.2. However, also when the training data has Indian faces, the African faces are clustered distantly, as seen in Fig. 2d. This is because the Indian race is the farthest from the African race in Fig. 4. When one of them is learned, the faces from these two races move apart as the features learned by the model from this race are very different from the other race. With most models in Fig. 2, the Caucasian and Indian races are getting clustered together. Also, the faces of Asian, Caucasian, and Indian races are split into two sub-clusters each. To understand if gender plays a role in forming these sub-clusters, we predict the genders of all the faces in the RFW test set using [29], which has a comparable gender prediction accuracy across races. Using this gender information, we plot the UMAP projections of faces in the RFW set with the model trained on the MS1MV3 dataset as shown in Fig. 3. This clearly indicates that these sub-clusters are due to gender, and the sub-



(a) Models trained on data from single race.

(b) Models trained on data from two races.

(c) Models trained on data from three races.

(d) Models trained on data from all four races.

Figure 5: Bar plots with the cosine distance decision thresholds of all the trained models on the four races of the RFW test set. The y-axis indicates the cosine distances, the x-axis indicates the races from the test set on which the thresholds are calculated, and each bar indicates a trained model identified by their training data racial distribution.

clusters of the same gender are grouped together. African faces do not form visible sub-clusters because the representation of women in the African subset of the RFW test set is meager compared to other races to form a visible sub-cluster of women. The proportion of women in the African cohort is 1.62%, whereas it is 31.69%, 28.16%, and 25.53% for the Asian, Caucasian, and Indian cohorts of the RFW set.

5.7. On Decision Thresholds

Fig. 5 depict the decision thresholds for all the models trained with different data settings. From Fig. 5a, we can observe that the decision threshold is the highest for the race using which the model is trained. This indicates that the model trained on faces from a certain race tends to be more confident in recognizing faces from that race and, in turn, has a higher decision threshold. This is true for all the training data settings. Similarly, in Fig. 5d, the model trained on the MS1MV3 dataset has the highest decision threshold for the Caucasian cohort, whereas the model trained on all four races of the BUPT-BalancedFace dataset has comparable decision thresholds across all racial cohorts. This is because the MS1MV3 dataset has disproportionately more Caucasian faces, which is not the case for the BUPT-BalancedFace. This is also called own-race bias [22].

6. Conclusion

In this work, we show how the racial distribution in the training data affects the racial bias of face recognition models. We show that the face recognition models perform worse on the faces belonging to the unseen ethnicities during training, but we find that having an equal representation of faces from different races does not guarantee bias-free algorithms. We empirically show that, apart from non-uniform racial distribution in training data, factors like face quality and gradation in facial features across races play a considerable role in inducing bias in these models. We also analyze whether the clustering of faces based on race is a good indicator of bias. Unlike previous studies, we find a very high correlation between the two across different training data settings and a high correlation within all training data settings except the four-race setting. Finally, we visualize the clustering of faces using UMAP projections, which uncovered the role of gender in clustering. In this study, we lay down various ideas on what to look for in the training data to understand bias in face recognition algorithms apart from just racial distribution. We hope our research will guide future work in understanding the bias in face recognition models through the lens of data and help curate more educated datasets.

References

[1] Yair Bar-Haim, Talia Saidel, and Galit Yovel. The role of skin colour in face recognition. *Perception*, 38(1):145–148, 2009.

[2] Lindsey Barrett. Ban facial recognition technologies for children-and for everyone else. *BUJ Sci. & Tech. L.*, 26:223, 2020.

[3] Jacob Benesty, Jingdong Chen, Yiteng Huang, and Israel Cohen. Pearson correlation coefficient. In *Noise reduction in speech processing*, pages 1–4. Springer, 2009.

[4] Kevin Bowyer and Michael King. Why face recognition accuracy varies due to race. *Biometric Technology Today*, 2019(8):8–11, 2019.

[5] Tadeusz Caliński and Jerzy Harabasz. A dendrite method for cluster analysis. *Communications in Statistics-theory and Methods*, 3(1):1–27, 1974.

[6] Jiankang Deng, Jia Guo, Evangelos Ververas, Irene Kotsia, and Stefanos Zafeiriou. Retinaface: Single-shot multi-level face localisation in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5203–5212, 2020.

[7] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4690–4699, 2019.

[8] Jiankang Deng, Jia Guo, Debing Zhang, Yafeng Deng, Xiangju Lu, and Song Shi. Lightweight face recognition challenge. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019.

[9] Stefan Glüge, Mohammadreza Amirian, Dandolo Flumini, and Thilo Stadelmann. How (not) to measure bias in face recognition networks. In *IAPR Workshop on Artificial Neural Networks in Pattern Recognition*, pages 125–137. Springer, 2020.

[10] Sixue Gong, Xiaoming Liu, and Anil K Jain. Jointly debiasing face recognition and demographic attribute estimation. In *European Conference on Computer Vision*, pages 330–347. Springer, 2020.

[11] Sixue Gong, Xiaoming Liu, and Anil K Jain. Mitigating face recognition bias via group adaptive classifier. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3414–3424, 2021.

[12] Patrick Grother, Mei Ngan, and Kayee Hanaoka. *Face recognition vendor test (fvrt): Part 3, demographic effects*. National Institute of Standards and Technology, 2019.

[13] PJ Grother, GW Quinn, and PJ Phillips. Mbe 2010: Report on the evaluation of 2d still-image face recognition algorithms. *National Institute of Standards and Technology, NISTIR*, 7709:1, 2010.

[14] 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 conference on computer vision*, pages 87–102. Springer, 2016.

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[16] Kashmir Hill. Wrongfully accused by an algorithm. In *Ethics of Data and Analytics*, pages 138–142. Auerbach Publications, 2020.

[17] Kimmo Karkkainen and Jungseock Joo. Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1548–1558, 2021.

[18] Adam Kortylewski, Bernhard Egger, Andreas Schneider, Thomas Gerig, Andreas Morel-Forster, and Thomas Vetter. Empirically analyzing the effect of dataset biases on deep face recognition systems. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 2093–2102, 2018.

[19] KS Krishnapriya, Vítor Albiero, Kushal Vangara, Michael C King, and Kevin W Bowyer. Issues related to face recognition accuracy varying based on race and skin tone. *IEEE Transactions on Technology and Society*, 1(1):8–20, 2020.

[20] Yong Li, Yufei Sun, Zhen Cui, Shiguang Shan, and Jian Yang. Learning fair face representation with progressive cross transformer. *arXiv preprint arXiv:2108.04983*, 2021.

[21] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

[22] Christian A Meissner and John C Brigham. Thirty years of investigating the own-race bias in memory for faces: A meta-analytic review. *Psychology, Public Policy, and Law*, 7(1):3, 2001.

[23] Shruti Nagpal, Maneet Singh, Richa Singh, and Mayank Vatsa. Deep learning for face recognition: Pride or prejudiced? *arXiv preprint arXiv:1904.01219*, 2019.

[24] Fu-Zhao Ou, Xingyu Chen, Ruixin Zhang, Yuge Huang, Shaoxin Li, Jilin Li, Yong Li, Liujuan Cao, and Yuan-Gen Wang. Sdd-fiqa: Unsupervised face image quality assessment with similarity distribution distance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7670–7679, 2021.

[25] Omkar M. Parkhi, Andrea Vedaldi, and Andrew Zisserman. Deep face recognition. In *British Machine Vision Conference*, 2015.

[26] Joseph P Robinson, Gennady Livitz, Yann Henon, Can Qin, Yun Fu, and Samson Timoner. Face recognition: too bias, or not too bias? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–1, 2020.

[27] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015.

[28] Reva Schwartz, Apostol Vassilev, Kristen Greene, Lori Perine, Andrew Burt, Patrick Hall, et al. Towards a standard for identifying and managing bias in artificial intelligence. 2022.

[29] Sefik Ilkin Serengil and Alper Ozpinar. Hyperextended lightface: A facial attribute analysis framework. In *2021 International Conference on Engineering and Emerging Technologies (ICEET)*, pages 1–4. IEEE, 2021.

[30] Natasha Singer and Cade Metz. Many facial-recognition systems are biased, says us study. *The New York Times*, 19, 2019.

[31] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019.

[32] Yi Sun. *Deep learning face representation by joint identification-verification*. The Chinese University of Hong Kong (Hong Kong), 2015.

[33] Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. Deepface: Closing the gap to human-level performance in face verification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1701–1708, 2014.

[34] Philipp Terhörst, Jan Niklas Kolf, Naser Damer, Florian Kirchbuchner, and Arjan Kuijper. Face quality estimation and its correlation to demographic and non-demographic bias in face recognition. In *2020 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–11. IEEE, 2020.

[35] Mei Wang and Weihong Deng. Mitigating bias in face recognition using skewness-aware reinforcement learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9322–9331, 2020.

[36] Mei Wang, Weihong Deng, Jiani Hu, Xunqiang Tao, and Yaohai Huang. Racial faces in the wild: Reducing racial bias by information maximization adaptation network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 692–702, 2019.

[37] Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. A discriminative feature learning approach for deep face recognition. In *European conference on computer vision*, pages 499–515. Springer, 2016.

[38] Robert Williams. I did nothing wrong. i was arrested anyway. *American Civil Liberties Union (ACLU)*, 2021.