2021 IEEE Winter Conference on Applications of Computer Vision (WACV) | 978-1-6654-0477-8/20/$31.00 ©2021 IEEE | DOI: 10.1109/WACV48630.2021.00159

FairFace: Face Attribute Dataset for Balanced Race, Gender, and Age for Bias Measurement and Mitigation

Kimmo Kärkkäinen
UCLA
kimmo@cs.ucla.edu

Jungseock Joo
UCLA
jjoo@comm.ucla.edu

Abstract

Existing public face image datasets are strongly biased toward Caucasian faces, and other races (e.g., Latino) are significantly underrepresented. The models trained from such datasets suffer from inconsistent classification accuracy, which limits the applicability of face analytic systems to non-White race groups. To mitigate the race bias problem in these datasets, we constructed a novel face image dataset containing 108,501 images which is balanced on race. We define 7 race groups: White, Black, Indian, East Asian, Southeast Asian, Middle Eastern, and Latino. Images were collected from the YFCC-100M Flickr dataset and labeled with race, gender, and age groups. Evaluations were performed on existing face attribute datasets as well as novel image datasets to measure the generalization performance. We find that the model trained from our dataset is substantially more accurate on novel datasets and the accuracy is consistent across race and gender groups. We also compare several commercial computer vision APIs and report their balanced accuracy across gender, race, and age groups. Our code, data, and models are available at `https://github.com/joojs/fairface`.

1. Introduction

To date, numerous large scale face image datasets [21, 31, 13, 70, 37, 23, 43, 69, 14, 26, 48, 8, 40] have been proposed and fostered research and development for automated face detection [35, 20], alignment [67, 46], recognition [57, 51], generation [68, 5, 25, 58], modification [3, 32, 18], and attribute classification [31, 37]. These systems have been successfully translated into many areas including security, medicine, education, and social sciences.

Despite the sheer amount of available data, existing public face datasets are strongly biased toward Caucasian faces, and other races (e.g., Latino) are significantly underrepresented. A recent study shows that most existing large scale face databases are biased towards "lighter skin" faces (around 80%), e.g. White, compared to "darker" faces, e.g. Black [40]. This means the model may not apply to some subpopulations and its results may not be compared across different groups without calibration. Biased data will produce biased models trained from it. This will raise ethical concerns about fairness of automated systems, which has emerged as a critical topic of study in the recent machine learning and AI literature [16, 11].

For example, several commercial computer vision systems (Microsoft, IBM, Face++) have been criticized due to their asymmetric accuracy across sub-demographics in recent studies [7, 44]. These studies found that the commercial face gender classification systems all perform better on male and on light faces. This can be caused by the biases in their training data. Various unwanted biases in image datasets can easily occur due to biased selection, capture, and negative sets [60]. Most public large scale face datasets have been collected from popular online media – newspapers, Wikipedia, or web search– and these platforms are more frequently used by or showing White people.

To mitigate the race bias in the existing face datasets, we propose a novel face dataset with an emphasis on balanced race composition. Our dataset contains 108,501 facial images collected primarily from the YFCC-100M Flickr dataset [59], which can be freely shared for a research purpose, and also includes examples from other sources such as Twitter and online newspaper outlets. We define 7 race groups: White, Black, Indian, East Asian, Southeast Asian, Middle Eastern, and Latino. Our dataset is well-balanced on these 7 groups (See Figures 1 and 2)

Our paper makes three main contributions. First, we emprically show that existing face attribute datasets and models learned from them do not generalize well to unseen data in which more non-White faces are present. Second, we show that our new dataset performs better on novel data, not only on average, but also across racial groups, i.e. more consistently. Third, to the best of our knowledge, our dataset is the first large scale face attribute dataset in the wild which includes Latino and Middle Eastern and differentiates East Asian and Southeast Asian. Computer vision

978-1-6654-0477-8/21/$31.00 ©2021 IEEE
DOI 10.1109/WACV48630.2021.00159

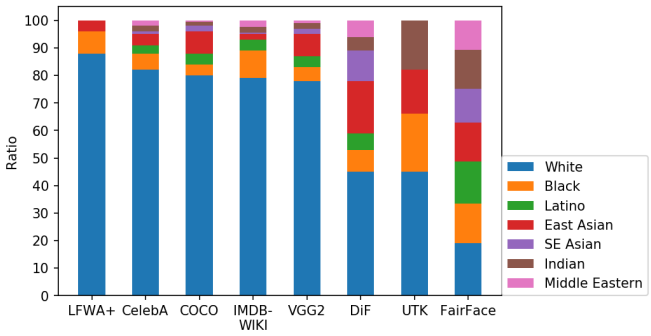


Figure 1: Racial compositions in face datasets.

has been rapidly transferred into other fields such as economics or social sciences, where researchers want to analyze different demographics using image data. The inclusion of major racial groups, which have been missing in existing datasets, therefore significantly enlarges the applicability of computer vision methods to these fields.

2. Related Work

2.1. Face Attribute Recognition

The goal of face attribute recognition is to classify various human attributes such as gender, race, age, emotions, expressions or other facial traits from facial appearance [31, 24, 75, 37]. Table 1 summarizes the statistics of existing large-scale **public** and **in-the-wild** face attribute datasets including our new dataset. As stated earlier, most of these datasets were constructed from online sources and are typically dominated by the White race.

Face attribute recognition has been applied as a subcomponent to other computer vision tasks such as face verification [31] and person re-idenfication [33, 34, 55]. It is imperative to ensure that these systems perform evenly well on different gender and race groups. Failing to do so can be detrimental to the reputations of individual service providers and the public trust about the machine learning and computer vision research community. Most notable incidents regarding the racial bias include Google Photos recognizing African American faces as Gorilla and Nikon's digital cameras prompting a message asking "did someone blink?" to Asian users [74]. These incidents, regardless of whether the models were trained improperly or how much they actually affected the users, often result in the termination of the service or features (e.g. dropping sensitive output categories). For this reason, most commercial service providers have stopped providing a race classifier.

Face attribute recognition is also used for demographic surveys performed in marketing or social science research, aimed at understanding human social behaviors and their relations to demographic backgrounds of individuals. Using off-the-shelf tools [2, 4] and commercial services, social scientists have begun to use images of people to infer their demographic attributes and analyze their behaviors. Notable examples are demographic analyses of social media users using their photographs [9, 45, 65, 66, 63]. The cost of unfair classification is huge as it can over- or under-estimate specific sub-populations in their analysis, which may have policy implications.

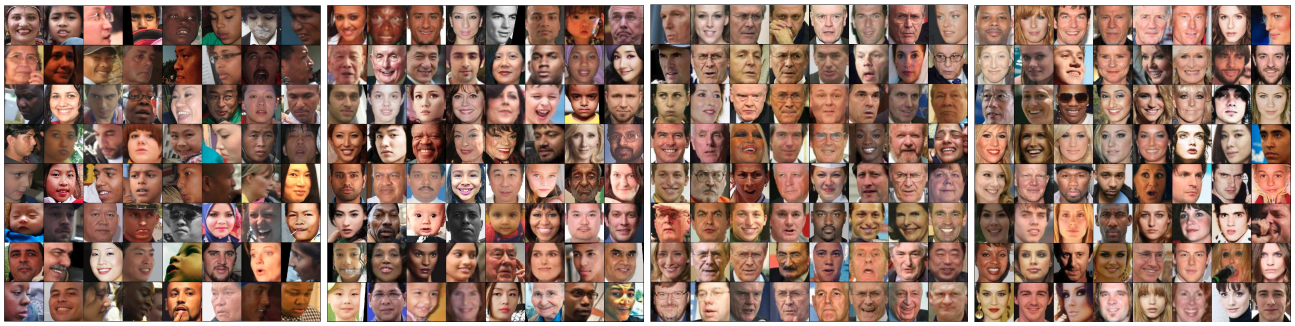
2.2. Fair Classification and Dataset Bias

AI and machine learning communities have increasingly paid attention to algorithmic fairness and dataset and model biases [72, 11, 77, 73]. There exist many different definitions of fairness used in the literature [61]. In this paper, we focus on balanced accuracy–whether the attribute classification accuracy is independent of race and gender. More generally, research in fairness is concerned with a model's ability to produce fair outcomes (e.g. loan approval) independent of protected or sensitive attributes such as race or gender.

Studies in algorithmic fairness have focused on either 1) discovering (auditing) existing bias in datasets or systems [52, 7, 30, 39, 22], 2) making a better dataset [40, 1], or 3) designing a better algorithm or model [12, 1, 49, 72, 71, 62, 41, 27], typically by learning representations invariant to sensitive attributes. Our work falls into the first two categories. While our paper does not propose a new method, we believe the contribution of our new dataset is still significant for the growing topic of bias studies. This is because 1) model biases are mainly caused by dataset biases and a balanced dataset can mitigate the problem and 2) our dataset can also be used to evaluate models and methods on fairness, which will facilitate the progress in the field.

The main task of interest in our paper is (balanced) gender classification from facial images. [7] demonstrated many commercial gender classification systems are biased and least accurate on dark-skinned females. The biased results may be caused by biased datasets, such as skewed image origins (45% of images are from the U.S. in Imagenet) [56] or biased underlying associations between scene and race in images [54]. It is, however, "infeasible to balance across all possible co-occurrences" of attributes [19], except in a lab-controlled setting.

Therefore, the contribution of our paper is to mitigate, not entirely solve, the current limitations and biases of existing databases by collecting more diverse face images from non-White race groups. We empirically show this significantly improves the generalization performance to novel image datasets whose racial compositions are not dominated by the White race. Furthermore, as shown in Table 1, our dataset is the first large scale in-the-wild face image dataset which includes Southeast Asian and Middle Eastern races. While their faces share similarity with East Asian



(a) FairFace (b) UTKFace (c) LFWA+ (d) CelebA

Figure 2: Random samples from face attribute datasets.

Table 1: Summary Statistics of Various Public Face Datasets

Name	Source	# of faces	In-the-wild?	Age	Gender	Race Annotation							
						White*		Asian*		Black	Indian	Latino	Balanced?
						W	ME	E	SE				
PPB [7]	Gov. Official Profiles	1K		✓	✓	**Skin color prediction							
MORPH [47]	Public Data	55K		✓	✓	merged				✓		✓	no
PubFig [31]	Celebrity	13K	✓	Model generated predictions									no
IMDB-WIKI [48]	IMDB, WIKI	500K	✓	✓	✓								no
FotW [13]	Flickr	25K	✓	✓	✓								yes
CACD [10]	celebrity	160K	✓	✓									no
DiF [40]	Flickr	1M	✓	✓	✓	**Skin color prediction							
†CelebA [37]	CelebFace LFW	200K	✓	✓	✓								no
LFW+ [15]	LFW (Newspapers)	15K	✓	✓	✓	merged		merged					no
†LFWA+ [37]	LFW (Newspapers)	13K	✓		✓	merged		merged		✓	✓		no
†UTKFace [76]	MORPH, CACD Web	20K	✓	✓	✓	merged		merged		✓	✓		yes
FairFace (Ours)	Flickr, Twitter Newspapers, Web	108K	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	yes

*FairFace (Ours) also defines East (E) Asian, Southeast (SE) Asian, Middle Eastern (ME), and Western (W) White.
**PPB and DiF do not provide race annotations but skin color annotated or automatically computed as a proxy to race.
†denotes datasets used in our experiments.

and White groups, we argue that not having these major race groups in datasets is a strong form of discrimination.

3. Dataset Construction

3.1. Race Taxonomy

Our dataset defines 7 race groups: White, Black, Indian, East Asian, Southeast Asian, Middle Eastern, and Latino. Race and ethnicity are different categorizations of humans. Race is defined based on physical traits and ethnicity is based on cultural similarities [50]. For example, Asian immigrants in Latin America can be of Latino ethnicity. In practice, these two terms are often used interchangeably.

We first adopted a commonly accepted race classification from the U.S. Census Bureau (White, Black, Asian, Hawaiian and Pacific Islanders, Native Americans, and Latino). Latino is often treated as an ethnicity, but we consider Latino a race, which can be judged from the facial appearance. We then further divided subgroups such as Middle Eastern, East Asian, Southeast Asian, and Indian, as they look clearly distinct. During the data collection, we found very few examples for Hawaiian and Pacific Islanders and Native Americans and discarded these categories. All the experiments conducted in this paper were therefore based

on 7 race classification.

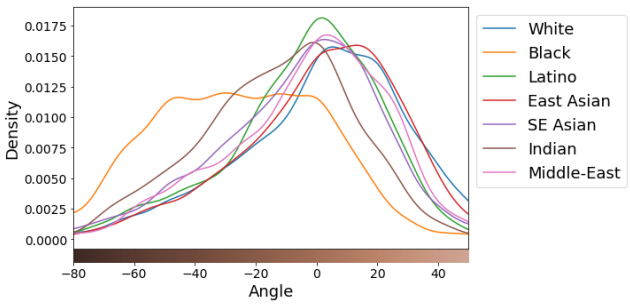


Figure 3: Individual Typology Angle (ITA), i.e. skin color, distribution of different races measured in our dataset.

An important criterion to measure dataset bias is on which basis the bias should be measured: **skin color or race?** A few recent studies [7, 40] use skin color as a proxy to racial or ethnicity grouping. While skin color can be easily computed without subjective annotations, it has limitations. First, skin color is heavily affected by illumination and light conditions. The Pilot Parliaments Benchmark (PPB) dataset [7] only used profile photographs of government officials taken in well controlled lighting, which makes it non-in-the-wild. Second, within-group variations of skin color are huge. Even same individuals can show different skin colors over time. Third, most importantly, race is a multidimensional concept whereas skin color (i.e. brightness) is one dimensional. Figure 3 shows the distributions of the skin color of multiple race groups, measured by Individual Typology Angle (ITA) [64]. As shown here, the skin color provides no information to differentiate many groups such as East Asian and White. Therefore, we explicitly use race and annotate the physical race by human annotators' judgments. A potential drawback of using the annotated race (as well as gender and age groups) comes from the subjectivity of annotators. To complement the limitation of race categorization, we also use skin color, measured by ITA, following the same procedure used by [40].

3.2. Image Collection and Annotation

Many existing face datasets have been sourced from photographs of public figures such as politicians or celebrities [31, 21, 23, 48, 37]. Despite the easiness of collecting images and ground truth attributes, the selection of these populations may be biased. For example, politicians may be older and actors may be more attractive than typical faces. Their images are usually taken by professional photographers in limited situations, leading to the quality bias. Some datasets were collected via web search using keywords such as "Asian boy" [76]. These queries may return only stereotypical faces or prioritize celebrities in those categories rather than diverse individuals among general public.

Our goal is to minimize the selection bias introduced by such filtering and maximize the diversity and coverage of the dataset. We started from a huge public image dataset, Yahoo YFCC100M dataset [59], and detected faces from the images without any preselection. A recent work also used the same dataset to construct a huge unfiltered face dataset (Diversity in Faces, DiF) [40]. Our dataset is smaller but more balanced on race (See Figure 1).

For an efficient collection, we incrementally increased the dataset size. We first detected and annotated 7,125 faces randomly sampled from the entire YFCC100M dataset ignoring the locations of images. After obtaining annotations on this initial set, we estimated demographic compositions of each country. Based on this statistic, we adaptively adjusted the number of images for each country sampled from the dataset such that the dataset is not dominated by the White race. Consequently, we excluded the U.S. and European countries in the later stage of data collection after we sampled enough White faces from those countries. The minimum size of a detected face was set to 50 by 50 pixels. This is a relatively smaller size compared to other datasets, but we find the attributes are still recognizable and these examples can actually make the classifiers more robust against noisy data. We only used images with "Attribution" and "Share Alike" Creative Commons licenses, which allow derivative work and commercial usages.

We used Amazon Mechanical Turk to annotate the race, gender and age group for each face. We assigned three workers for each image. If two or three workers agreed on their judgements, we took the values as ground-truth. If all three workers produced different responses, we republished the image to another 3 workers and subsequently discarded the image if the new annotators did not agree. These annotations at this stage were still noisy. We further refined the annotations by training a model from the initial ground truth annotations and applying back to the dataset. We then manually re-verified the annotations for images whose annotations differed from model predictions.

4. Experiments

4.1. Measuring Bias in Datasets

We first measure how skewed each dataset is in terms of its race composition. For the datasets with race annotations, we use the reported statistics. For the other datasets, we annotated the race labels for 3,000 random samples drawn from each dataset. See Figure 1 for the result. As expected, most existing face attribute datasets, especially the ones focusing on celebrities or politicians, are biased toward the White race. Unlike race, we find that most datasets are relatively more balanced on gender ranging from 40%-60% male ratio.

4.2. Model and Cross-Dataset Performance

To compare model performance of different datasets, we used an identical model architecture, ResNet-34 [17], to be trained from each dataset. We used ADAM optimization [29] with a learning rate of 0.0001. Given an image, we detected faces using the dlib's (dlib.net) CNN-based face detector [28] and ran the attribute classifier on each face. The experiment was done in PyTorch.

Throughout the evaluations, we compare our dataset with three other datasets: UTKFace [76], LFWA+, and CelebA [37]. Both UTKFace and LFWA+ have race annotations, and thus, are suitable for comparison with our dataset. CelebA does not have race annotations, so we only use it for gender classification. See Table 1 for more detailed dataset characteristics.

Using models trained from these datasets, we first performed cross-dataset classifications, by alternating training sets and test sets. Note that FairFace is the only dataset with 7 races. To make it compatible with other datasets, we merged our fine racial groups when tested on other datasets. CelebA does not have race annotations but was included for gender classification.

Tables 2 and 3 show the classification results for race, gender, and age on the datasets across subpopulations. As expected, each model tends to perform better on the same dataset on which it was trained. However, the accuracy of our model was highest on some variables on the LFWA+ dataset and also very close to the leader in other cases. This is partly because LFWA+ is the most biased dataset and ours is the most diverse, and thus more generalizable dataset.

4.3. Generalization Performance

4.3.1 Datasets

To test the generalization performance of the models, we consider three novel datasets. Note that these datasets were collected from completely different sources than our data from Flickr and not used in training. Since we want to measure the effectiveness of the model on diverse races, we chose the test datasets that contain people in different locations as follows.

Geo-tagged Tweets. First we consider images uploaded by Twitter users whose locations are identified by geotags (longitude and latitude), provided by [53]. From this set, we chose four countries (France, Iraq, Philippines, and Venezuela) and randomly sampled 5,000 faces.

Media Photographs. Next, we also use photographs posted by 500 online professional media outlets. Specifically, we use a public dataset of tweet IDs [36] posted by 4,000 known media accounts, e.g. @nytimes. Note that although we use Twitter to access the photographs, these tweets are simply external links to pages in the main newspaper sites. Therefore this data is considered as media photographs and different from general tweet images mostly uploaded by ordinary users. We randomly sampled 8,000 faces from the set.

Protest Dataset. Lastly, we also use a public image dataset collected for a recent protest activity study [65]. The authors collected the majority of data from Google Image search by using keywords such as "Venezuela protest" or "football game" (for hard negatives). The dataset exhibits a wide range of diverse race and gender groups engaging in different activities in various countries. We randomly sampled 8,000 faces from the set.

These faces were annotated for gender, race, and age by Amazon Mechanical Turk workers.

4.3.2 Result

Table 8 in Supplementary Material shows the classification accuracy of different models. Because our dataset is larger than LFWA+ and UTKFace, we report the three variants of the FairFace model by limiting the size of a training set (9k, 18k, and Full) for fair comparisons.

Improved Accuracy. As clearly shown in the result, the model trained by FairFace outperforms all the other models for race, gender, and age, on the novel datasets, which have never been used in training and also come from different data sources. The models trained with fewer training images (9k and 18k) still outperform other datasets including CelebA which is larger than FairFace. This suggests that the dataset size is not the only reason for the performance improvement.

Balanced Accuracy. Our model also produces more consistent results – for race, gender, age classification – across different race groups compared to other datasets. We measure the model consistency by standard deviations of classification accuracy measured on different subpopulations, as shown in Table 5. More formally, one can consider conditional use accuracy equality [6] or equalized odds [16] as the measure of fair classification. For gender classification:

$$P(\widehat{Y} = i|Y = i, A = j) = P(\widehat{Y} = i|Y = i, A = k), \\ i \in \{\text{male, female}\}, \forall j, k \in \text{D}, \quad (1)$$

where $\widehat{Y}$ is the predicted gender, Y is the true gender, A refers to the demographic group, and D is the set of different demographic groups being considered (*i.e.* race). When we consider different gender groups for A, this needs to be modified to measure accuracy equality [6]:

$$P(\widehat{Y} = Y|A = j) = P(\widehat{Y} = Y|A = k), \forall j, k \in \text{D}. \quad (2)$$

We therefore define the maximum accuracy disparity of a

Table 2: Cross-Dataset Classification Accuracy on White Race.

		Tested on								
		Race			Gender				Age	
		FairFace	UTKFace	LFWA+	FairFace	UTKFace	LFWA+	CelebA*	FairFace	UTKFace
Trained on	FairFace	**.937**	.936	**.970**	**.942**	**.940**	.920	**.981**	**.597**	.565
	UTKFace	.800	.918	.925	.860	.935	.916	.962	.413	**.576**
	LFWA+	.879	**.947**	.961	.761	.842	**.930**	.940	-	-
	CelebA	-	-	-	.812	.880	.905	.971	-	-

* CelebA doesn't provide race annotations. The result was obtained from the whole set (white and non-white).

Table 3: Cross-Dataset Classification Accuracy on non-White Races.

		Tested on								
		Race†			Gender				Age	
		FairFace	UTKFace	LFWA+	FairFace	UTKFace	LFWA+	CelebA*	FairFace	UTKFace
Trained on	FairFace	**.754**	.801	**.960**	**.944**	**.939**	**.930**	**.981**	**.607**	.616
	UTKFace	.693	**.839**	.887	.823	.925	.908	.962	.418	**.617**
	LFWA+	.541	.380	.866	.738	.833	.894	.940	-	-
	CelebA	-	-	-	.781	.886	.901	.971	-	-

* CelebA doesn't provide race annotations. The result was obtained from the whole set (white and non-white).
† FairFace defines 7 race categories but only 4 races (White, Black, Asian, and Indian) were used in this result to make it comparable to UTKFace.

classifier as follows:

$$\epsilon(\widehat{Y}) = \max_{\forall j,k \in \mathrm{D}} \left(\log \frac{P(\widehat{Y} = Y | A = j)}{P(\widehat{Y} = Y | A = k)} \right). \quad (3)$$

Table 4 shows the gender classification accuracy of different models measured on the external validation datasets for each race and gender group. The FairFace model achieves the lowest maximum accuracy disparity. The LFWA+ model yields the highest disparity, strongly biased toward the male category. The CelebA model tends to exhibit a bias toward the female category as the dataset contains more female images than male.

The FairFace model achieves less than 1% accuracy discrepancy between male $\leftrightarrow$ female and White $\leftrightarrow$ non-White for gender classification (Table 8). All the other models show a strong bias toward the male class, yielding much lower accuracy on the female group, and perform more inaccurately on the non-White group. The gender performance gap was the biggest in LFWA+ (32%), which is the smallest among the datasets used in the experiment. Recent work has also reported asymmetric gender biases in commercial computer vision services [7], and our result further suggests the cause is likely due to the unbalanced representation in training data.

Data Coverage and Diversity. We further investigate dataset characteristics to measure the data diversity in our dataset. We first visualize randomly sampled faces in 2D space using t-SNE [38] as shown in Figure 4. We used the facial embedding based on ResNet-34 from dlib, which was trained from the FaceScrub dataset [42], the VGG-Face dataset [43] and other online sources, which are likely dominated by the White faces. The faces in FairFace are well spread in the space, and the race groups are loosely separated from each other. This is in part because the embedding was trained from biased datasets, but it also suggests that the dataset contains many non-typical examples. LFWA+ was derived from LFW, which was developed for face recognition, and therefore contains multiple images of the same individuals, i.e. clusters. UTKFace also tends to focus more on local clusters compared to FairFace.

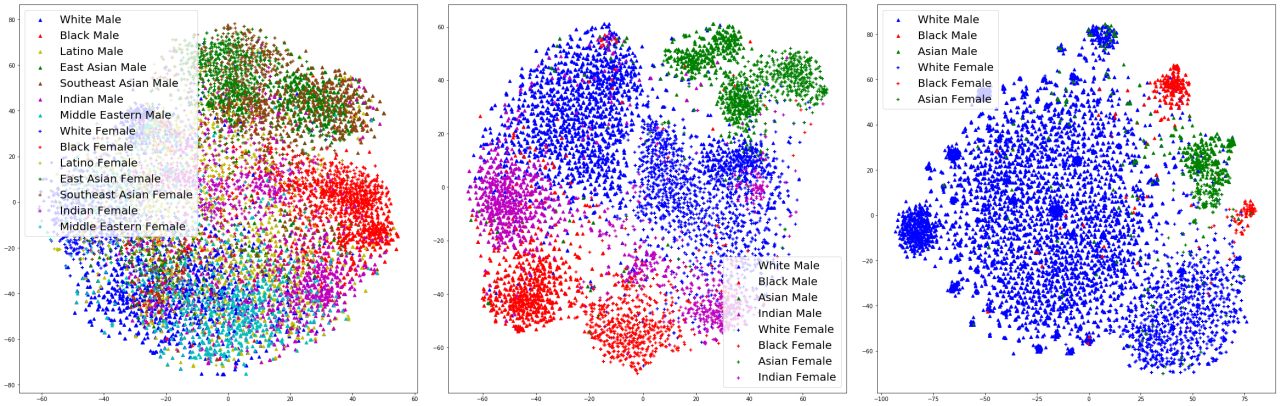
To explicitly measure the diversity of faces in these datasets, we examine the distributions of pairwise distance between faces (Figure 5). On the random subsets, we first obtained the same 128-dimensional facial embedding from dlib and measured pair-wise distance. Figure 5 shows the CDF functions for 3 datasets. As conjectured, UTKFace had more faces that are tightly clustered together and very similar to each other, compared to our dataset. Surprisingly, the faces in LFWA+ were shown very diverse and far from each other, even though the majority of the examples contained a white face. We believe this is mostly due to the fact that the face embedding was also trained on a very similar white-oriented dataset which will be effective in separating white faces, not because the appearance of their faces is actually diverse. (See Figure 2)

4.4. Evaluating Commercial Gender Classifiers

Previous studies have reported that popular commercial face analytic models show inconsistent classification accuracies across different demographic groups [7, 44]. We used the FairFace images to test several online APIs for gender classification: Microsoft Face API, Amazon Rekognition, IBM Watson Visual Recognition, and Face++. Compared to prior work using politicians' faces, our dataset is much

Race	White		Black		East Asian		SE Asian		Latino		Indian		Middle Eastern						
Gender	M	F	M	F	M	F	M	F	M	F	M	F	M	F	Max	Min	AVG	STDV	ϵ
FairFace	.967	.954	.958	.917	.873	.939	.909	.906	.977	.960	.966	.947	.991	.946	**.991**	**.873**	**.944**	**.032**	**.055**
UTK	.926	.864	.909	.795	.841	.824	.906	.795	.939	.821	.978	.742	.949	.730	.978	.730	.859	.078	.127
LFWA+	.946	.680	.974	.432	.826	.684	.938	.574	.951	.613	.968	.518	.988	.635	.988	.432	.766	.196	.359
CelebA	.829	.958	.819	.919	.653	.939	.768	.923	.843	.955	.866	.856	.924	.874	.958	.653	.866	.083	.166

Table 4: Gender classification accuracy measured on external validation datasets across gender-race groups.



(a) FairFace (b) UTKFace (c) LFWA+

Figure 4: t-SNE visualizations [38] of faces in datasets.

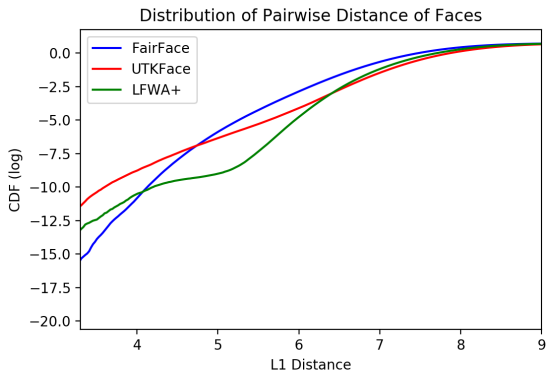


Figure 5: Distribution of pairwise distances of faces in 3 datasets measured by L1 distance on face embedding.

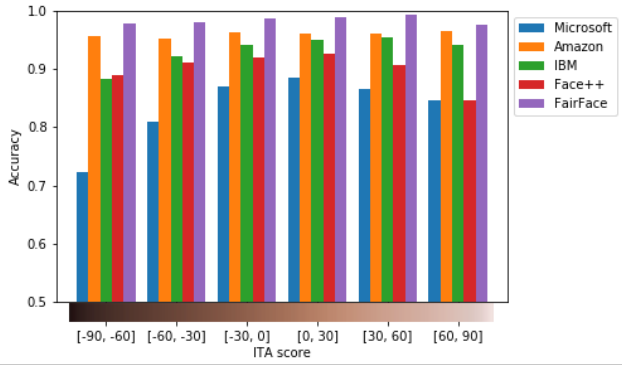


Figure 6: Classification accuracy based on Individual Typology Angle (ITA), i.e. skin color.

more diverse in terms of race, age, expressions, head orientation, and photographic conditions, and thus serves as a much better benchmark for bias measurement. We used 7,476 random samples from FairFace such that it contains an equal number of faces from each race, gender, and age group. We left out children under the age of 20, as these pictures were often ambiguous and the gender could not be determined for certain. The experiments were conducted on August 13th - 16th, 2019.

Table 6 shows the gender classification accuracies of the tested APIs. These APIs first detect a face from an input image and classify its gender. Not all 7,476 faces were detected by these APIs with the exception of Amazon Rekognition which detected all of them. Table 7 in Appendix reports the detection rate.[1] We report two sets of accuracies: 1) treating mis-detections as mis-classifications and 2) excluding mis-detections. For comparison, we included a model trained with our dataset to provide an upper bound for classification accuracy. Following prior work [40], we also show the classification accuracy as a function of skin

[1]These detection rates should not be interpreted as general face detection performance because we did not measure false detection rates using non-face images.

Table 5: Gender classification accuracy on external validation datasets, across race and age groups.

		Mean across races	SD across races	Mean across ages	SD across ages
Model trained on	FairFace	**94.89%**	**3.03%**	**92.95%**	**6.63%**
	UTKFace	89.54%	3.34%	84.23%	12.83%
	LFWA+	82.46%	5.60%	78.50%	11.51%
	CelebA	86.03%	4.57%	79.53%	17.96%

Table 6: Classification accuracy of commercial services on FairFace dataset. (*Microsoft, *Face++, *IBM indicate accuracies only on the detected faces, ignoring mis-detections.)

	White		Black		East Asian		SE Asian		Latino		Indian		Mid-Eastern			
	F	M	F	M	F	M	F	M	F	M	F	M	F	M	**Mean**	**STD**
Amazon	.923	.966	.901	.955	.925	.949	.918	.914	.921	.987	.951	.979	.906	.983	.941	.030
Microsoft	.822	.777	.766	.717	.824	.775	.852	.794	.843	.848	.863	.790	.839	.772	.806	.042
Face++	.888	.959	.805	.944	.876	.904	.884	.897	.865	.981	.770	.968	.822	.978	.896	.066
IBM	.910	.966	.758	.927	.899	.910	.852	.919	.884	.972	.811	.957	.871	.959	.900	.061
FairFace	.987	.991	.964	.974	.966	.979	.978	.961	.991	.989	.991	.987	.972	.991	.980	.011
*Microsoft	.973	.998	.962	.967	.963	.976	.960	.957	.983	.993	.975	.991	.966	.993	.975	.014
*Face++	.893	.968	.810	.956	.878	.911	.886	.899	.870	.983	.773	.975	.827	.983	.901	.067
*IBM	.914	.981	.761	.956	.909	.920	.852	.926	.892	.977	.819	.975	.881	.979	.910	.066

Table 7: Face detection rates of commercial APIs on FairFace dataset.

	White		Black		East Asian		SE Asian		Latino		Indian		Mid Eastern			
	F	M	F	M	F	M	F	M	F	M	F	M	F	M	**Mean**	**STD**
Amazon	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	.000
Microsoft	.845	.779	.796	.742	.856	.794	.888	.830	.858	.854	.886	.798	.869	.777	.812	.047
Face++	.994	.991	.994	.987	.998	.993	.998	.998	.994	.998	.996	.993	.994	.994	.993	.003
IBM	.996	.985	.996	.970	.989	.989	1.000	.993	.991	.994	.991	.981	.989	.979	.991	.008

color in Figure 6.

The results suggest several findings. First, all tested gender classifiers still favor the **male** category, which is consistent with the previous report [7]. Second, **dark-skinned females** tend to yield higher classification error rates, but there exist many exceptions. For example, Indians have darker skin tones (Figure 3), but some APIs (Amazon and MS) classified them more accurately than Whites. This suggests skin color alone, or any other individual phenotypic feature, is not a sufficient guideline to study model bias. Third, face detection can also introduce significant gender bias. Microsoft's model failed to detect many **male** faces, an opposite direction from the gender classification bias. This was not reported in previous studies which only used clean profile images of frontal faces.

5. Conclusion

This paper proposes a novel face image dataset balanced on race, gender and age. Compared to existing large-scale in-the-wild datasets, our dataset achieves much better generalization classification performance for gender, race, and age on novel image datasets collected from Twitter, international online newspapers, and web search, which contain more non-White faces than typical face datasets. We show that the model trained from our dataset produces balanced accuracy across race, whereas other datasets often lead to asymmetric accuracy on different race groups.

This dataset was derived from the Yahoo YFCC100m dataset [59] for the images with Creative Common Licenses by Attribution and Share Alike, which permit both academic and commercial usage. Our dataset can be used for training a new model and verifying balanced accuracy of existing classifiers.

Algorithmic fairness is an important aspect to consider in designing and developing AI systems, especially because these systems are being translated into many areas in our society and affecting our decision making. Large scale image datasets have contributed to the recent success in computer vision by improving model accuracy; yet the public and media have doubts about its transparency. The novel dataset proposed in this paper will help us discover and mitigate race and gender bias present in computer vision systems such that such systems can be more easily accepted in society.

Acknowledgement. This work was supported by NSF SBE-SMA #1831848.

References

[1] Mohsan Alvi, Andrew Zisserman, and Christoffer Nellaaker. Turning a blind eye: Explicit removal of biases and variation from deep neural network embeddings. In *The European Conference on Computer Vision (ECCV) Workshops*, September 2018.

[2] Brandon Amos, Bartosz Ludwiczuk, Mahadev Satyanarayanan, et al. Openface: A general-purpose face recognition library with mobile applications. *CMU School of Computer Science*, 6, 2016.

[3] Grigory Antipov, Moez Baccouche, and Jean-Luc Dugelay. Face aging with conditional generative adversarial networks. In *2017 IEEE International Conference on Image Processing (ICIP)*, pages 2089–2093. IEEE, 2017.

[4] Tadas Baltrusaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. Openface 2.0: Facial behavior analysis toolkit. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 59–66. IEEE, 2018.

[5] Jianmin Bao, Dong Chen, Fang Wen, Houqiang Li, and Gang Hua. Cvae-gan: fine-grained image generation through asymmetric training. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2745–2754, 2017.

[6] Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research*, page 0049124118782533.

[7] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on Fairness, Accountability and Transparency*, pages 77–91, 2018.

[8] 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.

[9] Abhijnan Chakraborty, Johnnatan Messias, Fabricio Benevenuto, Saptarshi Ghosh, Niloy Ganguly, and Krishna P Gummadi. Who makes trends? understanding demographic biases in crowdsourced recommendations. In *Eleventh International AAAI Conference on Web and Social Media*, 2017.

[10] Bor-Chun Chen, Chu-Song Chen, and Winston H Hsu. Face recognition and retrieval using cross-age reference coding with cross-age celebrity dataset. *IEEE Transactions on Multimedia*, 17(6):804–815, 2015.

[11] Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 797–806. ACM, 2017.

[12] Abhijit Das, Antitza Dantcheva, and Francois Bremond. Mitigating bias in gender, age and ethnicity classification: a multi-task convolution neural network approach. In *The European Conference on Computer Vision (ECCV) Workshops*, September 2018.

[13] Sergio Escalera, Mercedes Torres Torres, Brais Martinez, Xavier Baró, Hugo Jair Escalante, Isabelle Guyon, Georgios Tzimiropoulos, Ciprian Corneou, Marc Oliu, Mohammad Ali Bagheri, et al. Chalearn looking at people and faces of the world: Face analysis workshop and challenge 2016. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 1–8, 2016.

[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] Hu Han, Anil K Jain, Fang Wang, Shiguang Shan, and Xilin Chen. Heterogeneous face attribute estimation: A deep multi-task learning approach. *IEEE transactions on pattern analysis and machine intelligence*, 40(11):2597–2609, 2018.

[16] Moritz Hardt, Eric Price, Nati Srebro, et al. Equality of opportunity in supervised learning. In *Advances in neural information processing systems*, pages 3315–3323, 2016.

[17] 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.

[18] Zhenliang He, Wangmeng Zuo, Meina Kan, Shiguang Shan, and Xilin Chen. Arbitrary facial attribute editing: Only change what you want. *arXiv preprint arXiv:1711.10678*, 1(3), 2017.

[19] Lisa Anne Hendricks, Kaylee Burns, Kate Saenko, Trevor Darrell, and Anna Rohrbach. Women also snowboard: Overcoming bias in captioning models. In *European Conference on Computer Vision*, pages 793–811. Springer, 2018.

[20] Peiyun Hu and Deva Ramanan. Finding tiny faces. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 951–959, 2017.

[21] Gary B. Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst, October 2007.

[22] Jungseock Joo and Kimmo Kärkkäinen. Gender slopes: Counterfactual fairness for computer vision models by attribute manipulation. *arXiv preprint arXiv:2005.10430*, 2020.

[23] Jungseock Joo, Francis F Steen, and Song-Chun Zhu. Automated facial trait judgment and election outcome prediction: Social dimensions of face. In *Proceedings of the IEEE international conference on computer vision*, pages 3712–3720, 2015.

[24] Jungseock Joo, Shuo Wang, and Song-Chun Zhu. Human attribute recognition by rich appearance dictionary. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 721–728, 2013.

[25] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. *arXiv preprint arXiv:1812.04948*, 2018.

[26] Ira Kemelmacher-Shlizerman, Steven M Seitz, Daniel Miller, and Evan Brossard. The megaface benchmark: 1 million faces for recognition at scale. In *Proceedings of the*

IEEE Conference on Computer Vision and Pattern Recognition, pages 4873–4882, 2016.

[27] Byungju Kim, Hyunwoo Kim, Kyungsu Kim, Sungjin Kim, and Junmo Kim. Learning not to learn: Training deep neural networks with biased data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9012–9020, 2019.

[28] Davis E King. Max-margin object detection. *arXiv preprint arXiv:1502.00046*, 2015.

[29] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[30] Svetlana Kiritchenko and Saif M Mohammad. Examining gender and race bias in two hundred sentiment analysis systems. *arXiv preprint arXiv:1805.04508*, 2018.

[31] Neeraj Kumar, Alexander Berg, Peter N Belhumeur, and Shree Nayar. Describable visual attributes for face verification and image search. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(10):1962–1977, 2011.

[32] Guillaume Lample, Neil Zeghidour, Nicolas Usunier, Antoine Bordes, Ludovic Denoyer, et al. Fader networks: Manipulating images by sliding attributes. In *Advances in Neural Information Processing Systems*, pages 5967–5976, 2017.

[33] Ryan Layne, Timothy M Hospedales, Shaogang Gong, and Q Mary. Person re-identification by attributes. In *Bmvc*, volume 2, page 8, 2012.

[34] Annan Li, Luoqi Liu, Kang Wang, Si Liu, and Shuicheng Yan. Clothing attributes assisted person reidentification. *IEEE Transactions on Circuits and Systems for Video Technology*, 25(5):869–878, 2015.

[35] Haoxiang Li, Zhe Lin, Xiaohui Shen, Jonathan Brandt, and Gang Hua. A convolutional neural network cascade for face detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5325–5334, 2015.

[36] Justin Littman, Laura Wrubel, Daniel Kerchner, and Yonah Bromberg Gaber. News Outlet Tweet Ids, 2017.

[37] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, 2015.

[38] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.

[39] Daniel McDuff, Shuang Ma, Yale Song, and Ashish Kapoor. Characterizing bias in classifiers using generative models. *arXiv preprint arXiv:1906.11891*, 2019.

[40] Michele Merler, Nalini Ratha, Rogerio S Feris, and John R Smith. Diversity in faces. *arXiv preprint arXiv:1901.10436*, 2019.

[41] Aythami Morales, Julian Fierrez, and Ruben Vera-Rodriguez. Sensitivenets: Learning agnostic representations with application to face recognition. *arXiv preprint arXiv:1902.00334*, 2019.

[42] Hong-Wei Ng and Stefan Winkler. A data-driven approach to cleaning large face datasets. In *2014 IEEE International Conference on Image Processing (ICIP)*, pages 343–347. IEEE, 2014.

[43] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, et al. Deep face recognition. In *bmvc*, volume 1, page 6, 2015.

[44] Inioluwa Deborah Raji and Joy Buolamwini. Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial ai products. In *AAAI/ACM Conf. on AI Ethics and Society*, volume 1, 2019.

[45] Julio Reis, Haewoon Kwak, Jisun An, Johnnatan Messias, and Fabricio Benevenuto. Demographics of news sharing in the us twittersphere. In *Proceedings of the 28th ACM Conference on Hypertext and Social Media*, pages 195–204. ACM, 2017.

[46] Shaoqing Ren, Xudong Cao, Yichen Wei, and Jian Sun. Face alignment at 3000 fps via regressing local binary features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1685–1692, 2014.

[47] 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.

[48] Rasmus Rothe, Radu Timofte, and Luc Van Gool. Deep expectation of real and apparent age from a single image without facial landmarks. *International Journal of Computer Vision (IJCV)*, July 2016.

[49] Hee Jung Ryu, Hartwig Adam, and Margaret Mitchell. Inclusivefacenet: Improving face attribute detection with race and gender diversity. *arXiv preprint arXiv:1712.00193*, 2017.

[50] Richard T Schaefer. *Encyclopedia of race, ethnicity, and society*, volume 1. Sage, 2008.

[51] 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.

[52] Shreya Shankar, Yoni Halpern, Eric Breck, James Atwood, Jimbo Wilson, and D Sculley. No classification without representation: Assessing geodiversity issues in open data sets for the developing world. *arXiv preprint arXiv:1711.08536*, 2017.

[53] Zachary C Steinert-Threlkeld. *Twitter as data*. Cambridge University Press, 2018.

[54] Pierre Stock and Moustapha Cisse. Convnets and imagenet beyond accuracy: Understanding mistakes and uncovering biases. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 498–512, 2018.

[55] Chi Su, Fan Yang, Shiliang Zhang, Qi Tian, Larry Steven Davis, and Wen Gao. Multi-task learning with low rank attribute embedding for multi-camera person re-identification. *IEEE transactions on pattern analysis and machine intelligence*, 40(5):1167–1181, 2018.

[56] Harini Suresh, Jen J Gong, and John V Guttag. Learning tasks for multitask learning: Heterogenous patient populations in the icu. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 802–810. ACM, 2018.

[57] 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.

[58] Christopher Thomas and Adriana Kovashka. Persuasive faces: Generating faces in advertisements. *arXiv preprint arXiv:1807.09882*, 2018.

[59] Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. Yfcc100m: The new data in multimedia research. *Communications of the ACM*, 59(2):64–73.

[60] A Torralba and AA Efros. Unbiased look at dataset bias. In *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1521–1528. IEEE Computer Society, 2011.

[61] Sahil Verma and Julia Rubin. Fairness definitions explained. In *2018 IEEE/ACM International Workshop on Software Fairness (FairWare)*, pages 1–7. IEEE, 2018.

[62] Mei Wang, Weihong Deng, Jiani Hu, Jianteng Peng, Xunqiang Tao, and Yaohai Huang. Racial faces in-the-wild: Reducing racial bias by deep unsupervised domain adaptation. *arXiv preprint arXiv:1812.00194*, 2018.

[63] Yu Wang, Yang Feng, Zhe Hong, Ryan Berger, and Jiebo Luo. How polarized have we become? a multimodal classification of trump followers and clinton followers. In *International Conference on Social Informatics*, pages 440–456. Springer, 2017.

[64] Marcus Wilkes, Caradee Y Wright, Johan L du Plessis, and Anthony Reeder. Fitzpatrick skin type, individual typology angle, and melanin index in an african population: steps toward universally applicable skin photosensitivity assessments. *JAMA dermatology*, 151(8):902–903, 2015.

[65] Donghyeon Won, Zachary C Steinert-Threlkeld, and Jungseock Joo. Protest activity detection and perceived violence estimation from social media images. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 786–794. ACM, 2017.

[66] Nan Xi, Di Ma, Marcus Liou, Zachary C Steinert-Threlkeld, Jason Anastasopoulos, and Jungseock Joo. Understanding the political ideology of legislators from social media images. *arXiv preprint arXiv:1907.09594*, 2019.

[67] Xuehan Xiong and Fernando De la Torre. Supervised descent method and its applications to face alignment. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 532–539, 2013.

[68] Xinchen Yan, Jimei Yang, Kihyuk Sohn, and Honglak Lee. Attribute2image: Conditional image generation from visual attributes. In *European Conference on Computer Vision*, pages 776–791. Springer, 2016.

[69] Shuo Yang, Ping Luo, Chen-Change Loy, and Xiaoou Tang. Wider face: A face detection benchmark. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5525–5533, 2016.

[70] Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z Li. Learning face representation from scratch. *arXiv preprint arXiv:1411.7923*, 2014.

[71] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rogriguez, and Krishna P Gummadi. Fairness constraints: Mechanisms for fair classification. In *Artificial Intelligence and Statistics*, pages 962–970, 2017.

[72] Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning fair representations. In *International Conference on Machine Learning*, pages 325–333, 2013.

[73] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 335–340. ACM, 2018.

[74] Maggie Zhang. Google photos tags two african-americans as gorillas through facial recognition software, Jul 2015.

[75] Zhanpeng Zhang, Ping Luo, Chen-Change Loy, and Xiaoou Tang. Learning social relation traits from face images. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3631–3639, 2015.

[76] Zhifei Zhang, Yang Song, and Hairong Qi. Age progression/regression by conditional adversarial autoencoder. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5810–5818, 2017.

[77] James Zou and Londa Schiebinger. Ai can be sexist and racist—it's time to make it fair, 2018.