Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification*

Joy Buolamwini JOYAB@MIT.EDU
MIT Media Lab 75 Amherst St. Cambridge, MA 02139

Timnit Gebru TIMNIT.GEBRU@MICROSOFT.COM
Microsoft Research 641 Avenue of the Americas, New York, NY 10011

Editors: Sorelle A. Friedler and Christo Wilson

Abstract

Recent studies demonstrate that machine learning algorithms can discriminate based on classes like race and gender. In this work, we present an approach to evaluate bias present in automated facial analysis algorithms and datasets with respect to phenotypic subgroups. Using the dermatologist approved Fitzpatrick Skin Type classification system, we characterize the gender and skin type distribution of two facial analysis benchmarks, IJB-A and Adience. We find that these datasets are overwhelmingly composed of lighter-skinned subjects (79.6% for IJB-A and 86.2% for Adience) and introduce a new facial analysis dataset which is balanced by gender and skin type. We evaluate 3 commercial gender classification systems using our dataset and show that darker-skinned females are the most misclassified group (with error rates of up to 34.7%). The maximum error rate for lighter-skinned males is 0.8%. The substantial disparities in the accuracy of classifying darker females, lighter females, darker males, and lighter males in gender classification systems require urgent attention if commercial companies are to build genuinely fair, transparent and accountable facial analysis algorithms.

Keywords: Computer Vision, Algorithmic Audit, Gender Classification

1. Introduction

Artificial Intelligence (AI) is rapidly infiltrating every aspect of society. From helping determine who is hired, fired, granted a loan, or how long an individual spends in prison, decisions that have traditionally been performed by humans are rapidly made by algorithms (O'Neil, 2017; Citron and Pasquale, 2014). Even AI-based technologies that are not specifically trained to perform high-stakes tasks (such as determining how long someone spends in prison) can be used in a pipeline that performs such tasks. For example, while face recognition software by itself should not be trained to determine the fate of an individual in the criminal justice system, it is very likely that such software is used to identify suspects. Thus, an error in the output of a face recognition algorithm used as input for other tasks can have serious consequences. For example, someone could be wrongfully accused of a crime based on erroneous but confident misidentification of the perpetrator from security video footage analysis.

Many AI systems, e.g. face recognition tools, rely on machine learning algorithms that are trained with labeled data. It has recently been shown that algorithms trained with biased data have resulted in algorithmic discrimination (Bolukbasi et al., 2016; Caliskan et al., 2017). Bolukbasi et al. even showed that the popular word embedding space, Word2Vec, encodes societal gender biases. The authors used Word2Vec to train an analogy generator that fills in missing words in analogies. The analogy man is to computer programmer as woman is to "X" was completed with "homemaker", conforming to the stereotype that programming is associated with men and homemaking with women. The biases in Word2Vec are thus likely to be propagated throughout any system that uses this embedding.

* Download our gender and skin type balanced PPB dataset at gendershades.org

Although many works have studied how to create fairer algorithms, and benchmarked discrimination in various contexts (Kilbertus et al., 2017; Hardt et al., 2016b,a), only a handful of works have done this analysis for computer vision. However, computer vision systems with inferior performance across demographics can have serious implications. Esteva et al. showed that simple convolutional neural networks can be trained to detect melanoma from images, with accuracies as high as experts (Esteva et al., 2017). However, without a dataset that has labels for various skin characteristics such as color, thickness, and the amount of hair, one cannot measure the accuracy of such automated skin cancer detection systems for individuals with different skin types. Similar to the well documented detrimental effects of biased clinical trials (Popejoy and Fullerton, 2016; Melloni et al., 2010), biased samples in AI for health care can result in treatments that do not work well for many segments of the population.

In other contexts, a demographic group that is underrepresented in benchmark datasets can nonetheless be subjected to frequent targeting. The use of automated face recognition by law enforcement provides such an example. At least 117 million Americans are included in law enforcement face recognition networks. A year-long research investigation across 100 police departments revealed that African-American individuals are more likely to be stopped by law enforcement and be subjected to face recognition searches than individuals of other ethnicities (Garvie et al., 2016). False positives and unwarranted searches pose a threat to civil liberties. Some face recognition systems have been shown to misidentify people of color, women, and young people at high rates (Klare et al., 2012). Monitoring phenotypic and demographic accuracy of these systems as well as their use is necessary to protect citizens' rights and keep vendors and law enforcement accountable to the public.

We take a step in this direction by making two contributions. First, our work advances gender classification benchmarking by introducing a new face dataset composed of 1270 unique individuals that is more phenotypically balanced on the basis of skin type than existing benchmarks. To our knowledge this is the first gender classification benchmark labeled by the Fitzpatrick (TB, 1988) six-point skin type scale, allowing us to benchmark the performance of gender classification algorithms by skin type. Second, this work introduces the first intersectional demographic and phenotypic evaluation of face-based gender classification accuracy. Instead of evaluating accuracy by gender or skin type alone, accuracy is also examined on 4 intersectional subgroups: darker females, darker males, lighter females, and lighter males. The 3 evaluated commercial gender classifiers have the lowest accuracy on darker females. Since computer vision technology is being utilized in high-stakes sectors such as healthcare and law enforcement, more work needs to be done in benchmarking vision algorithms for various demographic and phenotypic groups.

2. Related Work

Automated Facial Analysis. Automated facial image analysis describes a range of face perception tasks including, but not limited to, face detection (Zafeiriou et al., 2015; Mathias et al., 2014; Bai and Ghanem, 2017), face classification (Reid et al., 2013; Levi and Hassner, 2015a; Rothe et al., 2016) and face recognition (Parkhi et al., 2015; Wen et al., 2016; Ranjan et al., 2017). Face recognition software is now built into most smart phones and companies such as Google, IBM, Microsoft and Face++ have released commercial software that perform automated facial analysis (IBM; Microsoft; Face++; Google).

A number of works have gone further than solely performing tasks like face detection, recognition and classification that are easy for humans to perform. For example, companies such as Affectiva (Affectiva) and researchers in academia attempt to identify emotions from images of people's faces (Dehghan et al., 2017; Srinivasan et al., 2016; Fabian Benitez-Quiroz et al., 2016). Some works have also used automated facial analysis to understand and help those with autism (Leo et al., 2015; Palestra et al., 2016). Controversial papers such as (Kosinski and Wang, 2017) claim to determine the sexuality of Caucasian males whose profile pictures are on Facebook or dating sites. And others such as (Wu and Zhang, 2016) and Israeli based company Faception (Faception) have developed software that purports to determine an individual's characteristics (e.g. propensity towards crime, IQ, terrorism) solely from

their faces. The clients of such software include governments. An article by (Aguera Y Arcas et al., 2017) details the dangers and errors propagated by some of these aforementioned works.

Face detection and classification algorithms are also used by US-based law enforcement for surveillance and crime prevention purposes. In "The Perpetual Lineup", Garvie and colleagues provide an in-depth analysis of the unregulated police use of face recognition and call for rigorous standards of automated facial analysis, racial accuracy testing, and regularly informing the public about the use of such technology (Garvie et al., 2016). Past research has also shown that the accuracies of face recognition systems used by US-based law enforcement are systematically lower for people labeled female, Black, or between the ages of 18—30 than for other demographic cohorts (Klare et al., 2012). The latest gender classification report from the National Institute for Standards and Technology (NIST) also shows that algorithms NIST evaluated performed worse for female-labeled faces than male-labeled faces (Ngan et al., 2015).

The lack of datasets that are labeled by ethnicity limits the generalizability of research exploring the impact of ethnicity on gender classification accuracy. While the NIST gender report explored the impact of ethnicity on gender classification through the use of an ethnic proxy (country of origin), none of the 10 locations used in the study were in Africa or the Caribbean where there are significant Black populations. On the other hand, Farinella and Dugelay claimed that ethnicity has no effect on gender classification, but they used a binary ethnic categorization scheme: Caucasian and non-Caucasian (Farinella and Dugelay, 2012). To address the underrepresentation of people of African-descent in previous studies, our work explores gender classification on African faces to further scholarship on the impact of phenotype on gender classification.

Benchmarks. Most large-scale attempts to collect visual face datasets rely on face detection algorithms to first detect faces (Huang et al., 2007; Kemelmacher-Shlizerman et al., 2016). Megaface, which to date is the largest publicly available set of facial images, was composed utilizing Head Hunter (Mathias et al., 2014) to select one million images from the Yahoo Flicker 100M image dataset (Thomee et al., 2015; Kemelmacher-Shlizerman et al., 2016). Any systematic error found in face detectors will inevitably affect the composition of the benchmark. Some datasets collected in this manner have already been documented to contain significant demographic bias. For example, LFW, a dataset composed of celebrity faces which has served as a gold standard benchmark for face recognition, was estimated to be 77.5% male and 83.5% White (Han and Jain, 2014). Although (Taigman et al., 2014)'s face recognition system recently reported 97.35% accuracy on the LFW dataset, its performance is not broken down by race or gender. Given these skews in the LFW dataset, it is not clear that the high reported accuracy is applicable to people who are not well represented in the LFW benchmark. In response to these limitations, Intelligence Advanced Research Projects Activity (IARPA) released the IJB-A dataset as the most geographically diverse set of collected faces (Klare et al., 2015). In order to limit bias, no face detector was used to select images containing faces. In comparison to face recognition, less work has been done to benchmark performance on gender classification. In 2015, the Adience gender and age classification benchmark was released (Levi and Hassner, 2015b). As of 2017, The National Institute of Standards and Technology is starting another challenge to spur improvement in face gender classification by expanding on the 2014-15 study.

3. Intersectional Benchmark

An evaluation of gender classification performance currently requires reducing the construct of gender into defined classes. In this work we use the sex labels of "male" and "female" to define gender classes since the evaluated benchmarks and classification systems use these binary labels. An intersectional evaluation further requires a dataset representing the defined genders with a range of phenotypes that enable subgroup accuracy analysis. To assess the suitability of existing datasets for intersectional benchmarking, we provided skin type annotations for unique subjects within two selected datasets, and compared the distribution of darker females, darker males, lighter females, and lighter males. Due to phenotypic imbalances in existing benchmarks, we

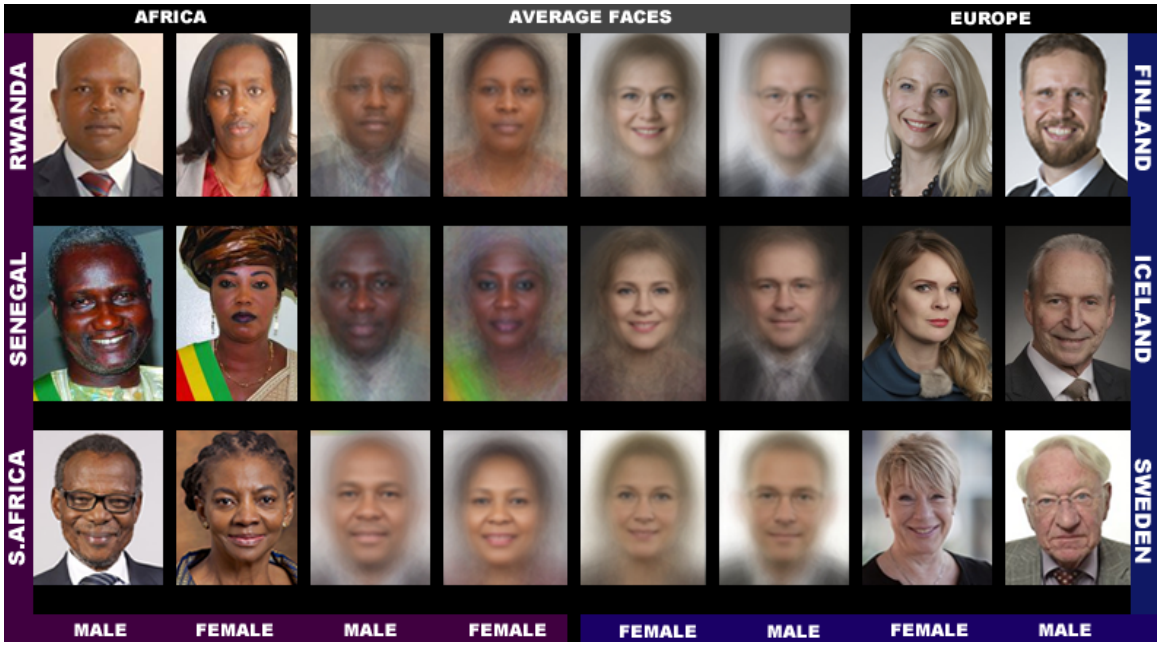

Figure 1: Example images and average faces from the new Pilot Parliaments Benchmark (PPB). As the examples show, the images are constrained with relatively little variation in pose. The subjects are composed of male and female parliamentarians from 6 countries. On average, Senegalese subjects are the darkest skinned while those from Finland and Iceland are the lightest skinned.

created a new dataset with more balanced skin type and gender representations.

3.1. Rationale for Phenotypic Labeling

Though demographic labels for protected classes like race and ethnicity have been used for performing algorithmic audits (Friedler et al., 2016; Angwin et al., 2016) and assessing dataset diversity (Han and Jain, 2014), phenotypic labels are seldom used for these purposes. While race labels are suitable for assessing potential algorithmic discrimination in some forms of data (e.g. those used to predict criminal recidivism rates), they face two key limitations when used on visual images. First, subjects' phenotypic features can vary widely within a racial or ethnic category. For example, the skin types of individuals identifying as Black in the US can represent many hues. Thus, facial analysis benchmarks consisting of lighter-skinned Black individuals would not adequately represent darker-skinned ones. Second, racial and ethnic categories are not consistent across geographies: even within countries these categories change over time.

Since race and ethnic labels are unstable, we decided to use skin type as a more visually precise label to measure dataset diversity. Skin type is one phenotypic attribute that can be used to more objectively characterize datasets along with eye and nose shapes. Furthermore, skin type was chosen as a phenotypic factor of interest because default camera settings are calibrated to expose lighter-skinned individuals (Roth, 2009). Poorly exposed images that result from sensor optimizations for lighter-skinned subjects or poor illumination can prove challenging for automated facial analysis. By labeling faces with skin type, we can increase our understanding of performance on this important phenotypic attribute.

3.2. Existing Benchmark Selection Rationale

IJB-A is a US government benchmark released by the National Institute of Standards and Tech-

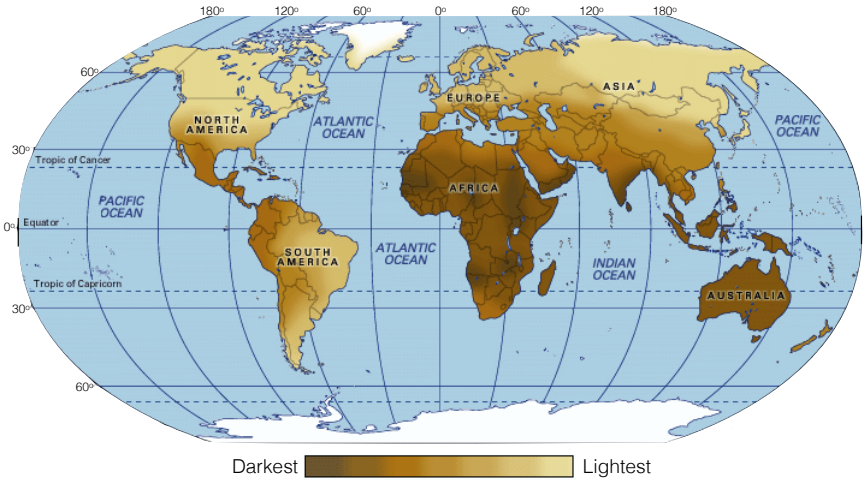

Figure 2: The global distribution of skin color. Most Africans have darker skin while those from Nordic countries are lighter-skinned. Image from (Encyclopedia Britannica) ©Copyright 2012 Encyclopedia Britannica.

nology (NIST) in 2015. We chose to evaluate this dataset given the government's involvement and the explicit development of the benchmark to be geographically diverse (as mentioned in Sec. 2). At the time of assessment in April and May of 2017, the dataset consisted of 500 unique subjects who are public figures. One image of each unique subject was manually labeled with one of six Fitzpatrick skin types (TB, 1988).

Adience is a gender classification benchmark released in 2014 and was selected due to its recency and unconstrained nature. The Adience benchmark contains 2,284 unique individual subjects. 2,194 of those subjects had reference images that were discernible enough to be labeled by skin type and gender. Like the IJB-A dataset, only one image of each subject was labeled for skin type.

3.3. Creation of Pilot Parliaments Benchmark

Preliminary analysis of the IJB-A and Adience benchmarks revealed overrepresentation of lighter males, underrepresentation of darker females, and underrepresentation of darker individuals in general. We developed the Pilot Parliaments Benchmark (PPB) to achieve better intersectional representation on the basis of gender and skin type. PPB consists of 1270 individuals from three African countries (Rwanda, Senegal, South Africa) and three European countries (Iceland, Finland, Sweden) selected for gender parity in the national parliaments.

Property	PPB	IJB-A	Adience
Release Year	2017	2015	2014
#Subjects	1270	500	2284
Avg. IPD	63 pixels	-	-
BBox Size	141 (avg)	≥36	-
IM Width	160-590	-	816
IM Height	213-886	-	816

Table 1: Various image characteristics of the Pilot Parliaments Benchmark compared with prior datasets. #Subjects denotes the number of unique subjects, the average bounding box size is given in pixels, and IM stands for image.

Figure 1 shows example images from PPB as well as average faces of males and females in each country represented in the datasets. We decided to use images of parliamentarians since they are public figures with known identities and photos available under non-restrictive licenses posted on government websites. To add skin

type diversity to the dataset, we chose parliamentarians from African and European countries. Fig. 2 shows an approximated distribution of average skin types around the world. As seen in the map, African countries typically have darker-skinned individuals whereas Nordic countries tend to have lighter-skinned citizens. Colonization and migration patterns nonetheless influence the phenotypic distribution of skin type and not all Africans are darker-skinned. Similarly, not all citizens of Nordic countries can be classified as lighter-skinned.

The specific African and European countries were selected based on their ranking for gender parity as assessed by the Inter Parliamentary Union (Inter Parliamentary Union Ranking). Of all the countries in the world, Rwanda has the highest proportion of women in parliament. Nordic countries were also well represented in the top 10 nations. Given the gender parity and prevalence of lighter skin in the region, Iceland, Finland, and Sweden were chosen. To balance for darker skin, the next two highest-ranking African nations, Senegal and South Africa, were also added.

Table 1 compares image characteristics of PPB with IJB-A and Adience. PPB is highly constrained since it is composed of official profile photos of parliamentarians. These profile photos are taken under conditions with cooperative subjects where pose is relatively fixed, illumination is constant, and expressions are neutral or smiling. Conversely, the images in the IJB-A and Adience benchmarks are unconstrained and subject pose, illumination, and expression by construction have more variation.

3.4. Intersectional Labeling Methodology

Skin Type Labels. We chose the Fitzpatrick six-point labeling system to determine skin type labels given its scientific origins. Dermatologists use this scale as the gold standard for skin classification and determining risk for skin cancer (TB, 1988).

The six-point Fitzpatrick classification system which labels skin as Type I to Type VI is skewed towards lighter skin and has three categories that can be applied to people perceived as White (Figure 2). Yet when it comes to fully representing the sepia spectrum that characterizes the rest of the world, the categorizations are fairly coarse. Nonetheless, the scale provides a scientifically based starting point for auditing algorithms and datasets by skin type.

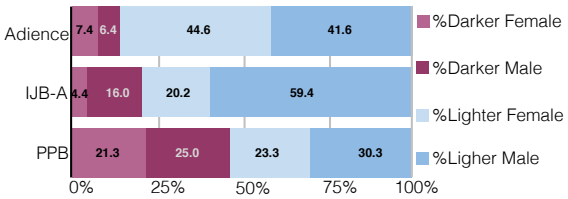

Figure 3: The percentage of darker female, lighter female, darker male, and lighter male subjects in PPB, IJB-A and Adience. Only 4.4% of subjects in Adience are darker-skinned and female in comparison to 21.3% in PPB.

Gender Labels. All evaluated companies provided a "gender classification" feature that uses the binary sex labels of female and male. This reductionist view of gender does not adequately capture the complexities of gender or address transgender identities. The companies provide no documentation to clarify if their gender classification systems which provide sex labels are classifying gender identity or biological sex. To label the PPB data, we use female and male labels to indicate subjects perceived as women or men respectively.

Labeling Process. For existing benchmarks, one author labeled each image with one of six Fitzpatrick skin types and provided gender annotations for the IJB-A dataset. The Adience benchmark was already annotated for gender. These preliminary skin type annotations on existing datasets were used to determine if a new benchmark was needed.

More annotation resources were used to label PPB. For the new parliamentarian benchmark, 3 annotators including the authors provided gender and Fitzpatrick labels. A board-certified surgical dermatologist provided the definitive labels for the Fitzpatrick skin type. Gender labels were determined based on the name of the parliamentarian, gendered title, prefixes such as Mr or Ms, and the appearance of the photo.

Set	n	F	M	Darker	Lighter	DF	DM	LF	LM
All Subjects	1270	44.6%	55.4%	46.4%	53.6%	21.3%	25.0%	23.3%	30.3%
Africa	661	43.9%	56.1%	86.2%	13.8%	39.8%	46.4%	4.1%	9.7%
South Africa	437	41.4%	58.6%	79.2%	20.8%	35.2%	43.9%	6.2%	14.6%
Senegal	149	43.0%	57.0%	100.0%	0.0%	43.0%	57.0%	0.0%	0.0%
Rwanda	75	60.0%	40.0%	100.0%	0.0%	60.0%	40.0%	0.0%	0.0%
Europe	609	45.5%	54.5%	3.1%	96.9%	1.3%	1.8%	44.2%	52.7%
Sweden	349	46.7%	53.3%	4.9%	95.1%	2.0%	2.9%	44.7%	50.4%
Finland	197	42.6%	57.4%	1.0%	99.0%	0.5%	0.5%	42.1%	56.9%
Iceland	63	47.6%	52.4%	0.0%	100.0%	0.0%	0.0%	47.6%	52.4%

Table 2: Pilot Parliaments Benchmark decomposition by the total number of female subjects denoted as F, total number of male subjects (M), total number of darker and lighter subjects, as well as female darker/lighter (DF/LF) and male darker/lighter subjects (DM/LM). The group compositions are shown for all unique subjects, Africa, Europe and the countries in our dataset located in each of these continents.

Dataset	Lighter (I,II,III)		Darker (IV, V, VI)		Total
PPB	53.6%	681	46.4%	589	1270
IJB-A	79.6%	398	20.4%	102	500
Adience	86.2%	1892	13.8%	302	2194

Table 3: The distributions of lighter and darker-skinned subjects (according to the Fitzpatrick classification system) in PPB, IJB-A, and Adience datasets. Adience has the most skewed distribution with 86.2% of the subjects consisting of lighter-skinned individuals whereas PPB is more evenly distributed between lighter (53.6%) and darker (46.4%) subjects.

3.5. Fitzpatrick Skin Type Comparison

For the purposes of our analysis, lighter subjects will refer to faces with a Fitzpatrick skin type of I,II, or III. Darker subjects will refer to faces labeled with a Fitzpatrick skin type of IV,V, or VI. We intentionally choose countries with majority populations at opposite ends of the skin type scale to make the lighter/darker dichotomy more distinct. The skin types are aggregated to account for potential off-by-one errors since the skin type is estimated using images instead of employing a standard spectrophotometer and Fitzpatrick questionnaire.

Table 2 presents the gender, skin type, and intersectional gender by skin type composition of PPB. And Figure 3 compares the percentage of images from darker female, darker male, lighter female and lighter male subjects from Adience, IJB-A, and PBB. PPB provides the most balanced representation of all four groups whereas IJB-A has the least balanced distribution.

Darker females are the least represented in IJB-A (4.4%) and darker males are the least represented in Adience (6.4%). Lighter males are the most represented unique subjects in all datasets. IJB-A is composed of 59.4% unique lighter males whereas this percentage is reduced to 41.6% in Adience and 30.3% in PPB. As seen in Table 3, Adience has the most skewed distribution by skin type.

While all the datasets have more lighter-skinned unique individuals, PPB is around half light at 53.6% whereas the proportion of lighter-skinned unique subjects in IJB-A and Adience

is 79.6% and 86.2% respectively. PPB provides substantially more darker-skinned unique subjects than IJB-A and Adience. Even though Adience has 2194 labeled unique subjects, which is nearly twice that of the 1270 subjects in PPB, it has 302 darker subjects, nearly half the 589 darker subjects in PPB. Overall, PPB has a more balanced representation of lighter and darker subjects as compared to the IJB-A and Adience datasets.

4. Commercial Gender Classification Audit

We evaluated 3 commercial gender classifiers. Overall, male subjects were more accurately classified than female subjects replicating previous findings (Ngan et al., 2015), and lighter subjects were more accurately classified than darker individuals. An intersectional breakdown reveals that all classifiers performed worst on darker female subjects.

4.1. Key Findings on Evaluated Classifiers

- All classifiers perform better on male faces than female faces (8.1% − 20.6% difference in error rate)
- All classifiers perform better on lighter faces than darker faces (11.8% − 19.2% difference in error rate)
- All classifiers perform worst on darker female faces (20.8% − 34.7% error rate)
- Microsoft and IBM classifiers perform best on lighter male faces (error rates of 0.0% and 0.3% respectively)
- Face++ classifiers perform best on darker male faces (0.7% error rate)
- The maximum difference in error rate between the best and worst classified groups is 34.4%

4.2. Commercial Gender Classifier Selection: Microsoft, IBM, Face++

We focus on gender classifiers sold in API bundles made available by Microsoft, IBM, and Face++ (Microsoft; IBM; Face++). Microsoft's Cognitive Services Face API and IBM's Watson Visual Recognition API were chosen since both companies have made large investments in artificial intelligence, capture significant market shares in the machine learning services domain, and provide public demonstrations of their facial analysis technology. At the time of evaluation, Google did not provide a publicly available gender classifier. Previous studies have shown that face recognition systems developed in Western nations and those developed in Asian nations tend to perform better on their respective populations (Phillips et al., 2011). Face++, a computer vision company headquartered in China with facial analysis technology previously integrated with some Lenovo computers, was thus chosen to see if this observation holds for gender classification. Like Microsoft and IBM, Face++ also provided a publicly available demonstration of their gender classification capabilities at the time of evaluation(April and May 2017).

All of the companies offered gender classification as a component of a set of proprietary facial analysis API services (Microsoft; IBM; Face++). The description of classification methodology lacked detail and there was no mention of what training data was used. At the time of evaluation, Microsoft's Face Detect service was described as using advanced statistical algorithms that "may not always be 100% precise" (Microsoft API Reference). IBM Watson Visual Recognition and Face++ services were said to use deep learning-based algorithms (IBM API Reference; Face++ Terms of Service). None of the commercial gender classifiers chosen for this analysis reported performance metrics on existing gender estimation benchmarks in their provided documentation. The Face++ terms of use explicitly disclaim any warranties of accuracy. Only IBM provided confidence scores (between 0 and 1) for face-based gender classification labels. But it did not report how any metrics like true positive rates (TPR) or false positive rates (FPR) were balanced.

4.3. Evaluation Methodology

In following the gender classification evaluation precedent established by the National Institute for Standards and Technology (NIST), we assess

Classifier	Metric	All	F	M	Darker	Lighter	DF	DM	LF	LM
MSFT	TPR(%)	93.7	89.3	97.4	87.1	99.3	79.2	94.0	98.3	**100**
	Error Rate(%)	6.3	10.7	2.6	12.9	0.7	**20.8**	6.0	1.7	0.0
	PPV (%)	93.7	96.5	91.7	87.1	99.3	92.1	83.7	**100**	98.7
	FPR (%)	6.3	2.6	10.7	12.9	0.7	6.0	**20.8**	0.0	1.7
Face++	TPR(%)	90.0	78.7	99.3	83.5	95.3	65.5	**99.3**	90.2	99.2
	Error Rate(%)	10.0	21.3	0.7	16.5	4.7	**34.5**	0.7	9.8	0.8
	PPV (%)	90.0	98.9	85.1	83.5	95.3	98.8	76.6	**98.9**	92.9
	FPR (%)	10.0	0.7	21.3	16.5	4.7	0.7	**34.5**	0.8	9.8
IBM	TPR(%)	87.9	79.7	94.4	77.6	96.8	65.3	88.0	92.9	**99.7**
	Error Rate(%)	12.1	20.3	5.6	22.4	3.2	**34.7**	12.0	7.1	0.3
	PPV (%)	87.9	92.1	85.2	77.6	96.8	82.3	74.8	**99.6**	94.8
	FPR (%)	12.1	5.6	20.3	22.4	3.2	12.0	**34.7**	0.3	7.1

Table 4: Gender classification performance as measured by the positive predictive value (PPV), error rate (1-TPR), true positive rate (TPR), and false positive rate (FPR) of the 3 evaluated commercial classifiers on the PPB dataset. All classifiers have the highest error rates for darker-skinned females (ranging from 20.8% for Microsoft to 34.7% for IBM).

Classifier	Metric	DF	DM	LF	LM
MSFT	TPR(%)	76.2	**100**	**100**	**100**
	Error Rate(%)	**23.8**	0.0	0.0	0.0
	PPV(%)	**100**	84.2	**100**	**100**
	FPR(%)	0.0	**23.8**	0.0	0.0
Face++	TPR(%)	64.0	99.5	92.6	**100**
	Error Rate(%)	**36.0**	0.5	7.4	0.0
	PPV(%)	99.0	77.8	**100**	96.9
	FPR(%)	0.5	**36.0**	0.0	7.4
IBM	TPR(%)	66.9	94.3	**100**	98.4
	Error Rate(%)	**33.1**	5.7	0.0	1.6
	PPV(%)	90.4	78.0	96.4	**100**
	FPR(%)	5.7	**33.1**	1.6	0.0

Table 5: Gender classification performance as measured by the positive predictive value (PPV), error rate (1-TPR), true positive rate (TPR), and false positive rate (FPR) of the 3 evaluated commercial classifiers on the South African subset of the PPB dataset. Results for South Africa follow the overall trend with the highest error rates seen on darker-skinned females.

the overall classification accuracy, male classification accuracy, and female classification accuracy as measured by the true positive rate (TPR). Extending beyond the NIST methodology we also evaluate the positive predictive value, false positive rate, and error rate (1-TPR) of the following groups: all subjects, male subjects, female subjects, lighter subjects, darker subjects, darker females, darker males, lighter females, and lighter males. See Table 2 in supplementary materials for results disaggregated by gender and each Fitzpatrick Skin Type.

4.4. Audit Results

Male and Female Error Rates

To conduct a demographic performance analysis, the differences in male and female error rates for each gender classifier are compared first in aggregate (Table 4) and then for South Africa (Table 5). The NIST Evaluation of Automated Gender Classification Algorithms report revealed that gender classification performance on female faces was 1.8% to 12.5% lower than performance on male faces for the nine evaluated algorithms (Ngan et al., 2015). The gender misclassification rates on the Pilot Parliaments Benchmark replicate this trend across all classifiers. The differences between female and male classification error rates range from 8.1% to 20.6%. The relatively high positive predictive value for females indicate that when a face is predicted to be female the estimation is more likely to be correct than when a face is predicted to be male. For the Microsoft and IBM classifiers, the false positive rates (FPR) for males are triple or more than the FPR for females. The FPR for males is more than 30 times that of females with the Face++ classifier.

Darker and Lighter Error Rates

To conduct a phenotypic performance analysis, the differences in darker and lighter skin type error rates for each gender classifier are compared first in aggregate (Table 4) and then for South Africa (Table 5). All classifiers perform better on lighter subjects than darker subjects in PPB. Microsoft achieves the best result with error rates of 12.9% on darker subjects and 0.7% on lighter individuals. On darker subjects, IBM achieves the worst classification accuracy with an error rate of 22.4%. This rate is nearly 7 times higher than the IBM error rate on lighter faces.

Intersectional Error Rates

To conduct an intersectional demographic and phenotypic analysis, the error rates for four intersectional groups (darker females, darker males, lighter females and lighter males) are compared in aggregate and then for South Africa.

Across the board, darker females account for the largest proportion of misclassified subjects. Even though darker females make up 21.3% of the PPB benchmark, they constitute between 61.0% to 72.4% of the classification error. Lighter males who make up 30.3% of the benchmark contribute only 0.0% to 2.4% of the total errors from these classifiers (See Table 1 in supplementary materials).

We present a deeper look at images from South Africa to see if differences in algorithmic performance are mainly due to image quality from each parliament. In PPB, the European parliamentary images tend to be of higher resolution with less pose variation when compared to images from African parliaments. The South African parliament, however, has comparable image resolution and has the largest skin type spread of all the parliaments. Lighter subjects makeup 20.8% (n=91) of the images, and darker subjects make up the remaining 79.2% (n=346) of images. Table 5 shows that all algorithms perform worse on female and darker subjects when compared to their counterpart male and lighter subjects. The Microsoft gender classifier performs the best, with zero errors on classifying all males and lighter females.

On the South African subset of the PPB benchmark, all the error for Microsoft arises from misclassifying images of darker females. Table 5 also shows that all classifiers perform worse on darker females. Face++ is flawless on lighter males. IBM performs best on lighter females with 0.0% error rate. Examining classification performance on the South African subset of PPB reveals trends that closely match the algorithmic performance on the entire dataset. Thus, we conclude that variation in performance due to the image characteristics of each country does not fully account for the differences in misclassification rates between intersectional subgroups. In other words, the presence of more darker individuals is a better explanation for error rates than a deviation in how images of parliamentarians are composed and produced. However, darker skin alone may not be fully responsible for misclassification. Instead, darker skin may be highly correlated with facial geometries or gender display norms that were less represented in the training data of the evaluated classifiers.

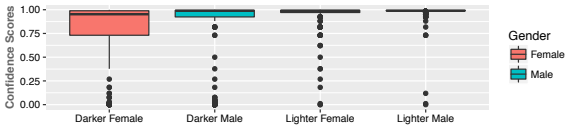

Figure 4: Gender classification confidence scores from IBM (IBM). Scores are near 1 for lighter male and female subjects while they range from $\sim 0.75 - 1$ for darker females.

4.5. Analysis of Results

The overall gender classification accuracy results show the obfuscating nature of single performance metrics. Taken at face value, gender classification accuracies ranging from 87.9% to 93.7% on the PPB dataset, suggest that these classifiers can be used for all populations represented by the benchmark. A company might justify the market readiness of a classifier by presenting performance results in aggregate. Yet a gender and phenotypic breakdown of the results shows that performance differs substantially for distinct subgroups. Classification is $8.1\% - 20.6\%$ worse on female than male subjects and $11.8\% - 19.2\%$ worse on darker than lighter subjects.

Though helpful in seeing systematic error, gender and skin type analysis by themselves do not present the whole story. Is misclassification distributed evenly amongst all females? Are there other factors at play? Likewise, is the misclassification of darker skin uniform across gender?

The intersectional error analysis that targets gender classification performance on darker female, lighter female, darker male, and lighter male subgroups provides more answers. Darker females have the highest error rates for all gender classifiers ranging from $20.8\% - 34.7\%$. For Microsoft and IBM classifiers lighter males are the best classified group with 0.0% and 0.3% error rates respectively. Face++ classifies darker males best with an error rate of 0.7%. When examining the gap in lighter and darker skin classification, we see that even though darker females are most impacted, darker males are still more misclassified than lighter males for IBM and Microsoft. The most improvement is needed on darker females specifically. More broadly, the error gaps between male and female classification along with lighter and darker classification should be closed.

4.6. Accuracy Metrics

Microsoft and Face++ APIs solely output single labels indicating whether the face was classified as female or male. IBM's API outputs an additional number which indicates the confidence with which the classification was made. Figure 4 plots the distribution of confidence values for each of the subgroups we evaluate (i.e. darker females, darker males, lighter females and lighter males). Numbers near 0 indicate low confidence whereas those close to 1 denote high confidence in classifying gender. As shown in the box plots, the API is most confident in classifying lighter males and least confident in classifying darker females.

While confidence values give users more information, commercial classifiers should provide additional metrics. All 3 evaluated APIs only provide gender classifications, they do not output probabilities associated with the likelihood of being a particular gender. This indicates that companies are choosing a threshold which determines the classification: if the prediction probability is greater than this threshold, the image is determined to be that of a male (or female) subject, and viceversa if the probability is less than this number. This does not give users the ability to analyze true positive (TPR) and false positive (FPR) rates for various subgroups if different thresholds were to be chosen. The commercial classifiers have picked thresholds that result in specific TPR and FPR rates for each subgroup. And the FPR for some groups can be much higher than those for others. By having APIs that fail to provide the ability to adjust these thresholds, they are limiting users' ability to pick their own TPR/FPR trade-off.

4.7. Data Quality and Sensors

It is well established that pose, illumination, and expression (PIE) can impact the accuracy of automated facial analysis. Techniques to create robust systems that are invariant to pose, illumination, expression, occlusions, and background have received substantial attention in computer vision research (Kakadiaris et al., 2017; Ganguly

et al., 2015; Ahmad Radzi et al., 2014). Illumination is of particular importance when doing an evaluation based on skin type. Default camera settings are often optimized to expose lighter skin better than darker skin (Roth, 2009). Underexposed or overexposed images that present significant information loss can make accurate classification challenging.

With full awareness of the challenges that arise due to pose and illumination, we intentionally chose an optimistic sample of constrained images that were taken from the parliamentarian websites. Each country had its peculiarities. Images from Rwanda and Senegal had more pose and illumination variation than images from other countries (Figure 1). The Swedish parliamentarians all had photos that were taken with a shadow on the face. The South African images had the most consistent pose and illumination. The South African subset was also composed of a substantial number of lighter and darker subjects. Given the diversity of the subset, the high image resolution, and the consistency of illumination and pose, our finding that classification accuracy varied by gender, skin type, and the intersection of gender with skin type do not appear to be confounded by the quality of sensor readings. The disparities presented with such a constrained dataset do suggest that error rates would be higher on more challenging unconstrained datasets. Future work should explore gender classification on an inclusive benchmark composed of unconstrained images.

5. Conclusion

We measured the accuracy of 3 commercial gender classification algorithms on the new Pilot Parliaments Benchmark which is balanced by gender and skin type. We annotated the dataset with the Fitzpatrick skin classification system and tested gender classification performance on 4 subgroups: darker females, darker males, lighter females and lighter males. We found that all classifiers performed best for lighter individuals and males overall. The classifiers performed worst for darker females. Further work is needed to see if the substantial error rate gaps on the basis of gender, skin type and intersectional subgroup revealed in this study of gender classification persist in other human-based computer vision tasks. Future work should explore intersectional error analysis of facial detection, identification and verification. Intersectional phenotypic and demographic error analysis can help inform methods to improve dataset composition, feature selection, and neural network architectures.

Because algorithmic fairness is based on different contextual assumptions and optimizations for accuracy, this work aimed to show why we need rigorous reporting on the performance metrics on which algorithmic fairness debates center. The work focuses on increasing phenotypic and demographic representation in face datasets and algorithmic evaluation. Inclusive benchmark datasets and subgroup accuracy reports will be necessary to increase transparency and accountability in artificial intelligence. For human-centered computer vision, we define transparency as providing information on the demographic and phenotypic composition of training and benchmark datasets. We define accountability as reporting algorithmic performance on demographic and phenotypic subgroups and actively working to close performance gaps where they arise. Algorithmic transparency and accountability reach beyond technical reports and should include mechanisms for consent and redress which we do not focus on here. Nonetheless, the findings from this work concerning benchmark representation and intersectional auditing provide empirical support for increased demographic and phenotypic transparency and accountability in artificial intelligence.

Acknowledgments

We thank board-certified surgical dermatologist Dr. Helen Raynham for providing the official Fitzpatrick annotations for the Pilot Parliaments Benchmark.

References

Face++ API. http://old.faceplusplus.com/demo-detect/. Accessed: 2017-10-06.

Face, Google APIs for Android, Google Developers. https://developers.google.com/android/reference/com/google/android/gms/vision/face/Face. Accessed: 2017-10-06.

Watson Visual Recognition. https://www.ibm.com/watson/services/visual-recognition/. Accessed: 2017-10-06.

Microsoft Face API. https://www.microsoft.com/cognitive-services/en-us/faceapi. Accessed: 2017-10-06.

Affectiva Emotion Recognition Software and Analysis. https://www.affectiva.com/. Accessed: 2017-10-06.

Physiognomys New Clothes. https://medium.com/@blaisea/physiognomys-new-clothes-f2d4b59fdd6a. Accessed: 2017-10-06.

Face++ Terms of Use. a. Accessed: 2018-12-13.

Faception, Facial Personality Analytics. https://www.faception.com/, b. Accessed: 2017-10-06.

Visual Recognition API Reference. Accessed: 2018-12-13.

How to Detect Faces in Image. Accessed: 2018-12-13.

Proportion of seats held by women in national parliaments. https://data.worldbank.org/indicator/SG.GEN.PARL.ZS?year_high_desc=true. Accessed: 2017-10-06.

Syafeeza Ahmad Radzi, Khalil-Hani Mohamad, Shan Sung Liew, and Rabia Bakhteri. Convolutional neural network for face recognition with pose and illumination variation. *International Journal of Engineering and Technology (IJET)*, 6(1):44–57, 2014.

Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias: Theres software used across the country to predict future criminals. and its biased against blacks. *ProPublica, May*, 23, 2016.

Yancheng Bai and Bernard Ghanem. Multi-scale fully convolutional network for face detection in the wild. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2017 IEEE Conference on*, pages 2078–2087. IEEE, 2017.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 4349–4357. Curran Associates, Inc., 2016. URL http://papers.nips.cc/paper/6228-man-is-to-computer-programmer-as-woman-is-to-homemaker-debiasing-word-embeddings.pdf.

Encyclopedia Britannica. Skin distribution map. https://media1.britannica.com/eb-media/59/61759-004-9A507F1C.gif, 2012. Accessed: 2017-12-17.

Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186, 2017.

Danielle Keats Citron and Frank A Pasquale. The scored society: due process for automated predictions. 2014.

Afshin Dehghan, Enrique G Ortiz, Guang Shu, and Syed Zain Masood. Dager: Deep age, gender and emotion recognition using convolutional neural network. *arXiv preprint arXiv:1702.04280*, 2017.

Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639):115–118, 2017.

C Fabian Benitez-Quiroz, Ramprakash Srinivasan, and Aleix M Martinez. Emotionet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5562–5570, 2016.

Giovanna Farinella and Jean-Luc Dugelay. Demographic classification: Do gender and ethnicity affect each other? In *Informatics, Electronics & Vision (ICIEV), 2012 International Conference on*, pages 383–390. IEEE, 2012.

Sorelle A Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. On the (im) possibility of fairness. *arXiv preprint arXiv:1609.07236*, 2016.

Suranjan Ganguly, Debotosh Bhattacharjee, and Mita Nasipuri. Illumination, pose and occlusion invariant face recognition from range images using erfi model. *International Journal of System Dynamics Applications (IJSDA)*, 4(2): 1–20, 2015.

Clare Garvie, Alvaro Bedoya, and Jonathan Frankle. *The Perpetual Line-Up: Unregulated Police Face Recognition in America*. Georgetown Law, Center on Privacy & Technology, 2016.

Hu Han and Anil K Jain. Age, gender and race estimation from unconstrained face images. *Dept. Comput. Sci. Eng., Michigan State Univ., East Lansing, MI, USA, MSU Tech. Rep.(MSU-CSE-14-5)*, 2014.

Moritz Hardt, Eric Price, Nati Srebro, et al. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems*, pages 3315–3323, 2016a.

Moritz Hardt, Eric Price, Nati Srebro, et al. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems*, pages 3315–3323, 2016b.

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, Technical Report 07-49, University of Massachusetts, Amherst, 2007.

Ioannis A Kakadiaris, George Toderici, Georgios Evangelopoulos, Georgios Passalis, Dat Chu, Xi Zhao, Shishir K Shah, and Theoharis Theoharis. 3d-2d face recognition with pose and illumination normalization. *Computer Vision and Image Understanding*, 154:137–151, 2017.

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.

Niki Kilbertus, Mateo Rojas-Carulla, Giambattista Parascandolo, Moritz Hardt, Dominik Janzing, and Bernhard Schölkopf. Avoiding discrimination through causal reasoning. *arXiv preprint arXiv:1706.02744*, 2017.

Brendan F Klare, Mark J Burge, Joshua C Klontz, Richard W Vorder Bruegge, and Anil K Jain. Face recognition performance: Role of demographic information. *IEEE Transactions on Information Forensics and Security*, 7(6):1789–1801, 2012.

Brendan F Klare, Ben Klein, Emma Taborsky, Austin Blanton, Jordan Cheney, Kristen Allen, Patrick Grother, Alan Mah, and Anil K Jain. Pushing the frontiers of unconstrained face detection and recognition: Iarpa janus benchmark a. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1931–1939, 2015.

Michal Kosinski and Yilun Wang. Deep neural networks are more accurate than humans at detecting sexual orientation from facial images. 2017.

Marco Leo, Marco Del Coco, Pierluigi Carcagni, Cosimo Distante, Massimo Bernava, Giovanni Pioggia, and Giuseppe Palestra. Automatic emotion recognition in robot-children interaction for asd treatment. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 145–153, 2015.

Gil Levi and Tal Hassner. Age and gender classification using convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 34–42, 2015a.

Gil Levi and Tal Hassner. Age and gender classification using convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 34–42, 2015b.

Markus Mathias, Rodrigo Benenson, Marco Pedersoli, and Luc Van Gool. Face detection without bells and whistles. In *European Conference on Computer Vision*, pages 720–735. Springer, 2014.

Chiara Melloni, Jeffrey S Berger, Tracy Y Wang, Funda Gunes, Amanda Stebbins, Karen S Pieper, Rowena J Dolor, Pamela S Douglas, Daniel B Mark, and L Kristin Newby. Representation of women in randomized clinical trials of cardiovascular disease prevention. *Circu-*

lation: Cardiovascular Quality and Outcomes, 3(2):135–142, 2010.

Mei Ngan, Mei Ngan, and Patrick Grother. *Face recognition vendor test (FRVT) performance of automated gender classification algorithms*. US Department of Commerce, National Institute of Standards and Technology, 2015.

Cathy O'Neil. *Weapons of math destruction: How big data increases inequality and threatens democracy*. Broadway Books, 2017.

Giuseppe Palestra, Giovanna Varni, Mohamed Chetouani, and Floriana Esposito. A multimodal and multilevel system for robotics treatment of autism in children. In *Proceedings of the International Workshop on Social Learning and Multimodal Interaction for Designing Artificial Agents*, page 3. ACM, 2016.

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

P Jonathon Phillips, Fang Jiang, Abhijit Narvekar, Julianne Ayyad, and Alice J O'Toole. An other-race effect for face recognition algorithms. *ACM Transactions on Applied Perception (TAP)*, 8(2):14, 2011.

Alice B Popejoy and Stephanie M Fullerton. Genomics is failing on diversity. *Nature*, 538 (7624):161, 2016.

Rajeev Ranjan, Swami Sankaranarayanan, Carlos D Castillo, and Rama Chellappa. An all-in-one convolutional neural network for face analysis. In *Automatic Face & Gesture Recognition (FG 2017), 2017 12th IEEE International Conference on*, pages 17–24. IEEE, 2017.

Daniel Reid, Sina Samangooei, Cunjian Chen, Mark Nixon, and Arun Ross. Soft biometrics for surveillance: an overview. *Machine learning: theory and applications. Elsevier*, pages 327–352, 2013.

Lorna Roth. Looking at shirley, the ultimate norm: Colour balance, image technologies, and cognitive equity. *Canadian Journal of Communication*, 34(1):111, 2009.

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*, pages 1–14, 2016.

Ramprakash Srinivasan, Julie D Golomb, and Aleix M Martinez. A neural basis of facial action recognition in humans. *Journal of Neuroscience*, 36(16):4434–4442, 2016.

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.

Fitzpatrick TB. The validity and practicality of sun-reactive skin types i through vi. *Archives of Dermatology*, 124(6):869–871, 1988. doi: 10.1001 / archderm.1988.01670060015008. URL +http : / / dx.doi.org / 10.1001 / archderm.1988.01670060015008.

Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. The new data and new challenges in multimedia research. *arXiv preprint arXiv:1503.01817*, 1(8), 2015.

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.

Xiaolin Wu and Xi Zhang. Automated inference on criminality using face images. *arXiv preprint arXiv:1611.04135*, 2016.

Stefanos Zafeiriou, Cha Zhang, and Zhengyou Zhang. A survey on face detection in the wild: past, present and future. *Computer Vision and Image Understanding*, 138:1–24, 2015.