Color-Theoretic Experiments to Understand Unequal Gender Classification Accuracy from Face Images

Vidya Muthukumar,[1,2] Tejaswini Pedapati,[1] Nalini Ratha,[1] Prasanna Sattigeri,[1] Chai-Wah Wu,[1]
Brian Kingsbury,[1] Abhishek Kumar,[*1] Samuel Thomas,[1] Aleksandra Mojsilović,[1] and Kush R. Varshney[1]

[1]IBM Research [2]University of California, Berkeley

Abstract

Recent work shows unequal performance of commercial face classification services in the gender classification task across intersectional groups defined by skin type and gender. Accuracy on dark-skinned females is significantly worse than on any other group. We provide initial evidence that skin type alone is not the driver for this disparity by conducting novel stability experiments that vary an image's skin type via color-theoretic methods, namely luminance mode-shift and optimal transport. We evaluate the effect of skin type change on the gender classification decision of a pair of state-of-the-art commercial and open-source gender classifiers. The results raise the possibility that broader differences in ethnicity, as opposed to the skin type alone, are what contribute to unequal gender classification accuracy in face images.

1. Introduction

The problem of unequal accuracy rates across groups has recently been highlighted in gender classification from face images. A study by NIST shows that automated gender classification algorithms are more accurate for males than females [29]. Going further, Buolamwini and Gebru created a dataset of parliament members from three European and three African countries — the Pilot Parliaments Benchmark (PPB), balanced across two attributes: gender and Fitzpatrick skin type [15], and evaluated the accuracy of three commercial facial gender classifiers [4]. All three achieved much lower accuracy on dark-skinned females (Fitzpatrick skin types IV–VI) than light-skinned females, dark-skinned males, and light-skinned males. (Note that gender classification is a distinct task from race classification [16].)

The discrepancy at a high level is largely due to imbalanced training datasets and test benchmarks. Commonly used training datasets such as CelebA [26] and IMDb-Face [38] are made up of celebrities, who are overwhelmingly light-skinned people. Test benchmarks such as Labeled Faces in the Wild [20] and Adience [12] are also imbalanced across skin type [4], so high overall accuracies achieved on these test datasets obfuscate the inequality issue. The IJB-A dataset purports to be geographically diverse [23], but a close examination reveals that only 8 percent of the faces are of African descent, whereas more than 50 percent of the faces are of European descent. The PPB dataset is the first of its kind to be balanced by gender *and* balanced between African and European descent [4].[1]

These works, however, do not investigate the underlying causes of the unequal misclassification rates in gender classification. This understanding is important to investigate whether improvements in the algorithms, or the data itself, are what are required to mitigate the issue. For gender classification, since the partition in [4] is *phenotypic* into different skin type categories but the dark-skinned people are predominantly of African descent, it may be that other features, such as hairstyle, facial structure, cosmetics or clothing are the reason for disparity, rather than skin type alone [5]. A study of unequal gender classification accuracy, conducted using images with different parts of the face masked out, points to the nose region as important, but does little to disentangle the various aspects of identity [32]. Buolamwini points to several shortcomings of that study and calls for "further scholarship that attends to the impact of phenotypic characteristic on gender classification that extends beyond skin type" [5].

In this paper we rigorously analyze the influence of the skin type on gender classification accuracy as a first step towards understanding the reasons for unequal gender classification accuracy on face images. We test *stability* to skin type by varying the skin type of a face keeping all other features fixed, and statistically show that the effect of skin type on classification outcome is minimal. Thus, the un-

[*]Work done at IBM Research. Author is now at Google Brain.

[1]Of course, the PPB dataset *does not* represent, e.g. individuals of Asian or South American descent, or younger children. But it does not claim to represent these populations.

equal accuracy observed in [4] likely arises not specifically because of the skin type, but other correlated features of identity [1]. Our methodology for varying the skin type involves a novel application of principles from color theory and optimal transport, and our experiments are applicable to any state-of-the-art gender classifier with access to confidence scores (output probabilities). We acknowledge imperfections in our methodology, but consider it an important starting point for research in this direction.

The paper is organized as follows. In Section 2 we discuss related work. In Section 3 we describe our experimental setup based on the PPB dataset. We describe our experiments to test the stability of gender classification algorithms in Section 4. Empirical results are provided in Section 5. Limitations of the study are presented in Section 6. The paper concludes with a discussion on our findings and future research directions.

2. Related Work

Computer vision tasks like face recognition and classification have been researched for decades. Automated facial analysis tasks include face detection [28, 45, 2], face classification [37, 24, 38] and face recognition [33, 43, 36]. Some facial recognition systems have been shown to misidentify people of color, women, and young people at high rates [22]. More recently, gender classification has been shown to have unequal performance both across gender itself [29] and the combination of gender and skin type [4, 35]. Unequal accuracies across these intersectional groups have important ramifications in face recognition software used in law enforcement applications [17] as well as safety-critical applications like object detection by self-driving cars [44].

The fairness problem is prevalent in many applications of machine learning other than computer vision. Theoretical approaches that have been designed to solve the fairness problem range from defining new fairness metrics such as *demographic parity* [6, 46] and *equality of odds/opportunity* [18], to deciding how to actually design fair ML algorithms according to these metrics [3]. To conclusively resolve this problem, we need to improve ML algorithms, improve the quality of training data, *or both*. Approaches to improve algorithms could involve trying to achieve invariance in an optimized pre-processing step on the data [7, 40], or *being aware of the protected attributes* and using them to train decoupled classifiers on different demographic groups [10, 11]. Both types of approaches have their pros and cons: pre-processing for invariance could lead to the loss of useful information and suboptimal overall accuracy, while a decoupled classifier increases the data requirement multi-fold. The correct approach is often application-dependent [25]. For the task of gender classification in computer vision, the metric is clear – equal accuracy rates across all four groups, which is exactly the "equality of odds" metric [18]. For a task like gender classification from face images in which there are known imbalances in training data, it is unclear whether achieving uniform accuracy across demographic groups is best done by pre-processing for invariance (which could throw away useful information) or decoupled classification (which requires much more data). Answering this question requires understanding both our data and our models better.

Table 1: Gender and skin type composition of PPB*/PPB dataset.

Set	Number	Female	Male
All subjects	1204/1270	42.1/44.6%	57.9/55.4%
Dark-skinned	507/589	41.8/45.9%	58.2/54.1%
Light-skinned	697/681	42.4/43.4%	57.6/56.6%

3. Setup

3.1. Pilot Parliaments Benchmark Dataset

The PPB dataset is the first benchmark dataset that is balanced across gender and Fitzpatrick skin type; the methodology of its collection is detailed in [4]. The creators intentionally chose countries with majority populations at opposite ends of the skin type scale to make the lighter/darker dichotomy more distinct. The images are uniform in (high) resolution quality, pose, illumination and expression, reducing the possibility of attributing differences in performance to variations in these quantities, all of which are known to be significant technical challenges [42].

We use an approximation of the PPB dataset for the experiments in this paper. This dataset contains images of parliament members from the six countries identified in [4] and were manually labeled by us into the categories dark-skinned and light-skinned.[2] Our approximation to the PPB dataset, which we call PPB*, is very similar to PPB and satisfies the relevant characteristics for the study we perform. Table 1 compares the composition of the original PPB dataset and our PPB* approximation according to skin type and gender.

3.2. Classification Models

We evaluate the IBM Watson gender classifier service available in August 2018, which achieves 99% accuracy on several test benchmarks, as well as 99% accuracy on the light male, light female and dark male groups of the PPB* dataset. We access the gender classifier using the

[2]The images were accessed in January 2018. We do not work with the PPB dataset directly due to its terms and conditions of use.

Table 2: Accuracies on dark females (DF), light females (LF), dark males (DM), and light males (LM) on PPB*.

Classifier	DF	DM	LF	LM
Watson	82.5%	99.3%	98.5%	99.5%
DEX	78.0%	99.5%	96.7%	99.5%

API, which takes in an input image of variable size and returns (in the event that a face is detected) a score $s \in [0, 1]$ that the image is of a male person. Values $s \leq 0.5$ are classified female and values $s > 0.5$ are classified male. The accessibility of scores from the IBM Watson API make it a good choice of commercial classifier for carrying out stability experiments. (Scores are not available from the other two commercial classifiers studied in [4].) In addition to the IBM Watson gender classifier, which is commercial, we evaluate the state-of-the-art open source convolutional-neural-network gender classifier DEX developed by researchers from ETH Zurich [38]. Accuracies on the PPB* dataset for both models are presented in Table 2. Both models obtain high accuracy on light males, light females and dark males, making them attractive baselines.

4. Methodology for Stability Experiments

We now describe the methodology for our experiments to test the stability of gender classification algorithms to variation in skin type. We systematically isolate the skin type and test the gender classification outcome for significant changes as a function of varying skin type. Isolating a latent facial attribute, and thus changing it, is in general known to be a challenging computer vision task. Likelihood based generative models [21] and conditional generative adversarial networks (GANs) [8, 34] have made recent progress in varying attributes like hair color and facial expressions. However, these tools themselves are trained on imbalanced celebrity datasets. Moreover, these approaches are not effective in varying one attribute *in isolation*, leaving other attributes unchanged. We empirically show the existence of an approximately low-dimensional structure in color space that describes the group of human skin types. Leveraging this structure, we provide simple but mathematically grounded rules to change the skin type of a face.

4.1. A Low-Dimensional Skin Type Group in YCrCb Space

Recall that image pixels can be represented in the 3-dimensional vector space $[0, 255]^3$. Multiple bases for the color space such as the standard RGB [13], HSV [31] and YCrCb [19], have been used to create skin type detection rules, as well as more recently proposed hybrid rules that also work under complex lighting conditions [30, 27]. We

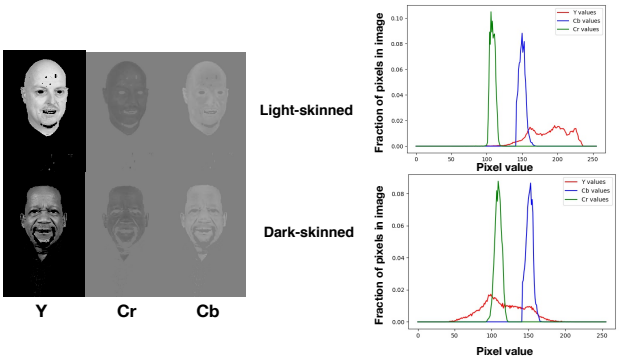


Figure 1: Example of a light-skinned and dark-skinned image in the PPB* dataset. Observe that the Cr and Cb channels are similar across both images. Practically all variation in the skin type is captured in the Y component.

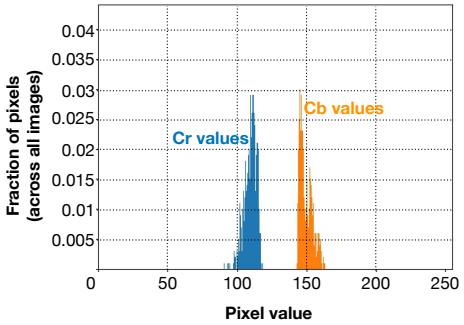


Figure 2: Frequencies of Cr and Cb values across all skin type pixels across all images.

use the following skin type detection rule [19] based on the YCrCb space, where Y stands for luminance and Cr, Cb stand for chrominance values.

$$\text{pixel} = \begin{cases} \text{skin if } 90 \leq Cr \leq 115 \text{ and } 140 \leq Cb \leq 195 \\ \text{not skin otherwise .} \end{cases} \quad (1)$$

We employ this rule for its simplicity and fairly good performance in skin type detection across ethnicities under the favorable lighting conditions in the PPB* dataset.

We also plotted histograms of the YCrCb values of skin type pixels detected for each face image and observed that the Cr and Cb values fall into an even narrower range than described in (1). As the illustration in Figure 1 depicts, the chrominance values do not appear different for individuals with light or dark skin type. More rigorously, Figure 2 plots the histogram of Cr and Cb values across all 1204 images in PPB*; we observe that the chrominance values are stable.

Practically all the variation in skin type is captured by the Y channel alone.[3]

4.2. Methods to Change the Skin Type

Based on the low-dimensional structure described in the previous subsection, we describe two rules that we employ to change the skin type of a face. Both are carried out in the YCrCb color space. We represent an image in YCrCb space by $I_{\mathsf{YCrCb}} \in [0, 255]^{w \times h \times 3}$, where (w, h) represents the width and height of the image.

Procedure 1 (Luminance mode-shift) *We shift the skin type luminance mode of an image in the following sequence of steps:*

1. *Determine old Y mode* $=$ *Y mode*$(\{I_{\mathsf{YCrCb}}(i, j) \in$ *skin types*$\})$.
2. *Calculate the mode-shift-value* $\delta =$ *new Y mode* $-$ *old Y mode.*
3. *Shift the luminance values, i.e.* $I_{\mathsf{Y}} = I_{\mathsf{Y}} + \delta$.
4. *Clip luminance values to* $[0, 255]$.

Procedure 1 is attractive for its simplicity and quick computation ($\mathcal{O}(1)$ time), but the results of skin type change according to luminance mode shift are not always visually attractive, as demonstrated in Figure 3. Perhaps the luminance mode of skin type pixels is not sufficiently descriptive, and we would rather consider a transform between *skin type histograms*. Motivated by this, we next consider a skin type operation based on optimal transport, which has recently shown to be effective in color transfer in RGB space [14].

Procedure 2 (Optimal transport [14]) *This procedure takes as input a target skin type distribution over Y values. We denote the skin type distribution of a grayscale image by $\mu(I)$ and the target skin type distribution by μ'_Y. Then, the optimally transported image is defined as follows:*

$$I^*_{\mathsf{Y}} := \arg\min ||I_{\mathsf{Y}} - I'_{\mathsf{Y}}||_2 \text{ subject to } \mu_Y(I'_{\mathsf{Y}}) = \mu'_Y.$$

Figure 4 shows that the results of optimally transported skin type are visually more realistic. However, the computational cost of using this operation is more; the optimal transport operation has complexity $\mathcal{O}((w \times h)^3)$.[4]

[3] We expect this phenomenon will hold for any face image with high resolution quality and uniform illumination. We eschewed more complex skin type detection rules that are robust to more challenging lighting conditions in favor of simplicity [30, 27].

[4] The minimum size of images that we work with is 128×128, and in practice it takes 30 seconds to a minute to optimally transport an image, compared to milliseconds to luminance mode shift an image. For future work, we could utilize the computational reductions in computing the optimal transport using Sinkhorn regularization [9].

5. Results

We consider the following ensemble of skin-type changes on the PPB* dataset:

1. Dark females/dark males: Evaluate the score on the original image. Evaluate the average new score on the set of lightened images.

2. Light females/light males: Evaluate the score on the original image. Evaluate the average new score on the set of darkened images.

The set of darkened/lightened mode-shifted images represents all luminance-mode-shifts with negative/positive δ. Owing to the computational expense of optimal transport, we pick ten images on varying ends of the skin type spectrum. Before proceeding to the overall results, we would like to mention an important detail in investigating the performance of DEX that involves pre-processing of the input face images using face detection and eye alignment. We used the standard Viola-Jones face detectors implemented in OpenCV and dlib, and observed that 25 out of 296 (8.4%) light female faces were not detected; and 35 out of 212 (16.5%) dark female faces were not detected. Thus, our results for DEX are reported for 271 and 177 images of light and dark females respectively.[5]

Figure 5 and 6 show the distribution of affected differences in prediction on lightening the set of dark females in the PPB dataset, either using mode-shift (Figure 5a and Figure 6a) or optimal transport (Figure 5b and Figure 6b) for IBM Watson and DEX, respectively.[6] We observe that most images' scores do not change meaningfully after lightening/darkening. In the case of dark females, 86.6% of the images' scores on IBM Watson, and 80.6% on DEX do not change by more than 0.1 on lightening using mode-shift. 76.6% of the images' scores on IBM Watson, and 70.4% on DEX do not change by more than 0.1 on lightening using optimal transport. In the case of light females, 96.3% of the images' scores on IBM Watson, and 94.4% on DEX, do not change by more than 0.1 on darkening using mode-shift. 92.1% of the images' scores on IBM Watson, and 86.7% on DEX, do not change by more than 0.1 on darkening using optimal transport. We conducted one-sample t-tests to test the null hypothesis that the mean of differences in scores

[5] These numbers suggest a potential discrepancy in the quality of face detection across skin type, a phenomenon which deserves further independent study and may be related to the recently observed discrepancies in object detection [44].

[6] The quality of the experiment itself is better with the optimal transport method as the lightened images are more realistic, but owing to computational complexity of optimal transport, we also have fewer lightened samples to average over. On the other hand, the mode-shift operation generates images that are not as realistic, but the experiment itself is statistically more robust as we can quickly generate many lightened samples. Thus, observing similar conclusions for the two methods strengthens our result.

Figure 3: Examples of light-skinned and dark-skinned faces whose luminance modes are shifted.

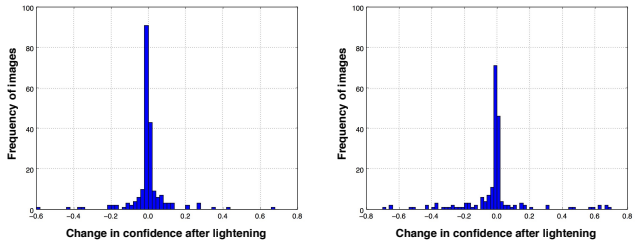
Figure 4: Examples of light-skinned and dark-skinned faces that are optimally transported to new skin types, either darkened or lightened.

Table 3: Results of one sample t-test on mean of differences in scores with respect to 0 after skin type change. CI stands for confidence interval.

Category	95% CI (IBM)	95% CI (DEX)
DF, mode-shift	$[-0.013, 0.015]$	$[-0.031, -0.0005]$
DF, OT	$[-0.071, -0.003]$	$[-0.051, -0.007]$
LF, mode-shift	$[-0.010, 0.001]$	$[-0.002, 0.016]$
LF, OT	$[-0.035, 0.016]$	$[-0.0139, 0.043]$

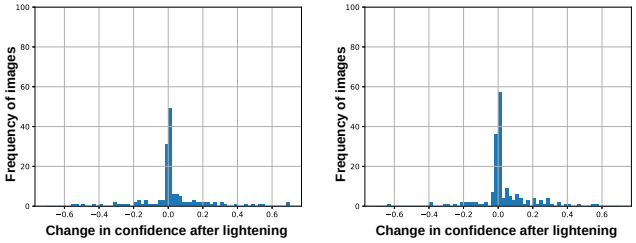
is equal to 0. The results in terms of the 95% confidence intervals are presented in Table 3.

Figures 9, 11, 10 and 12 shed insight into the *relative difference in predictions*: in particular, the fraction of images whose average classification decision changes after lightening/darkening. In the scatterplots of original score vs. score after change in skin type (cf. Figures 9b, 9c, 11b and 11c for IBM Watson; 10b, 10c, 12b and 12c for DEX), we highlight the points that fall in the red-shaded region as representing dark females that are correctly classified *only after lightening*, or light females that are incorrectly classified *only after darkening*. Very few images fall into these categories: 5 and 9 dark females (out of 212) are correctly classified by IBM Watson only after lightening using mode-shift and optimal transport respectively. 8 and 10 dark females (out of 177) are correctly classified by DEX only after lightening using mode shift and optimal transport respectively. The ef-



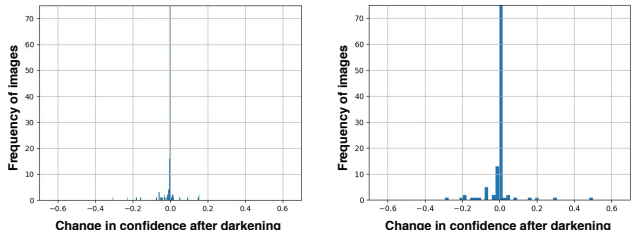
(a) Luminance-mode-shift. (b) Optimal transport.

Figure 5: Histograms of differences in scores of dark females in PPB* dataset on IBM Watson commercial classifier after lightening the skin type.



(a) Luminance-mode-shift. (b) Optimal transport.

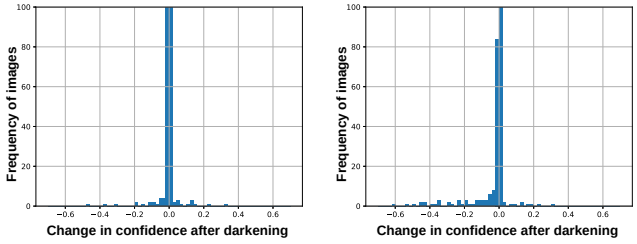
Figure 6: Histograms of differences in scores of dark females in PPB* dataset on the open-source DEX classifier after lightening the skin type.



(a) Luminance-mode-shift. (b) Optimal transport.

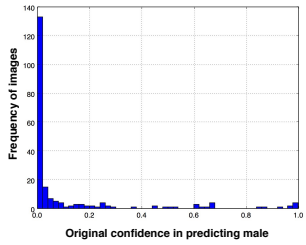
Figure 7: Histograms of differences in scores of light females in PPB* dataset on IBM Watson commercial classifier after darkening the skin type.

fect of darkening is even less pronounced for light females – after mode-shift and optimal transport, 0 and 2 females (out of 296) respectively become incorrectly classified by IBM Watson model. For DEX, 2 and 6 females (out of 271) become incorrectly classified after darkening according to mode-shift and optimal transport respectively. Looking at the distribution on original scores of dark and light females

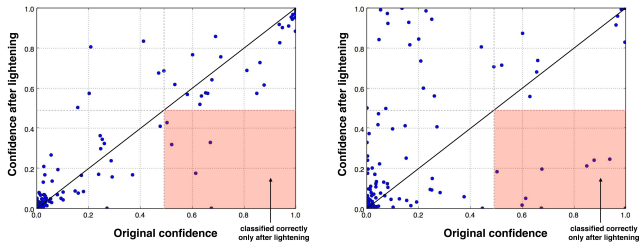


(a) Luminance-mode-shift. (b) Optimal transport.

Figure 8: Histograms of differences in scores of light females in PPB* dataset on the open-source DEX classifier after darkening the skin type.



(a) Scores on IBM Watson commercial classifier.

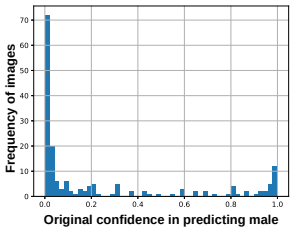


(b) Mode-shift. (c) OT.

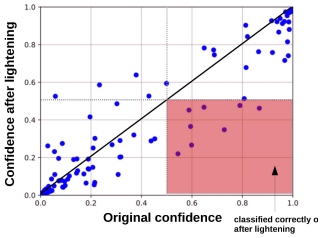
Figure 9: Scatterplots of original prediction vs prediction after lightening for dark females, on the IBM Watson commercial classifier. Shaded region represents dark females correctly classified after lightening.

(Figures 5, 6, 7 and 8), we see that almost all light females are classified as female with extremely high score, and almost all dark females are classified as either female or male with extremely high score. The dark females that are classified as male with extremely high score, say above 0.9, do not change significantly in score or classification decision on lightening.

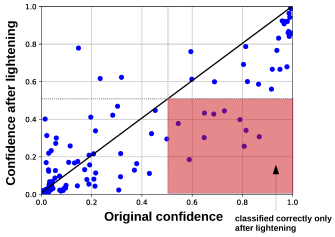
All of these results, together, lead us to conclude that *the skin type by itself has a minimal effect on classification decisions*.



(a) Scores on the open-source DEX classifier.

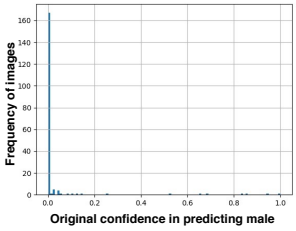


(b) Mode-shift.

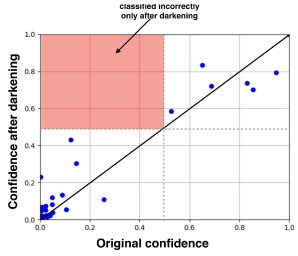


(c) OT.

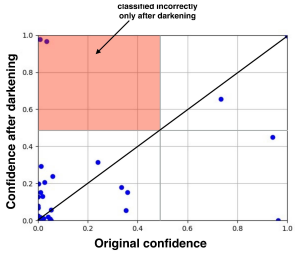
Figure 10: Scatterplots of original prediction vs prediction after lightening for dark females, on the open-source DEX classifier. Shaded region represents dark females correctly classified after lightening.



(a) Scores.

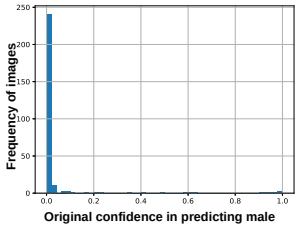


(b) Mode-shift.

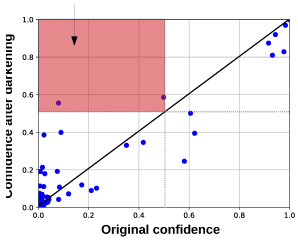


(c) OT.

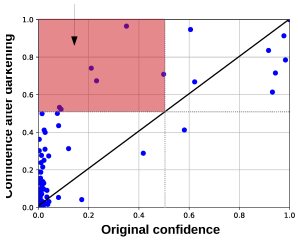
Figure 11: Scatterplots of original prediction vs prediction after darkening for light females, on the IBM Watson commercial classifier. The shaded region represents light females that would be incorrectly classified after darkening.



(a) Scores.



(b) Mode-shift.



(c) OT.

Figure 12: Scatterplots of original prediction vs prediction after darkening for light females, on the open-source DEX classifier. The shaded region represents light females that would be incorrectly classified after darkening.

6. Limitations of Study

In this section, we acknowledge the inherent limitations in our methodology, particularly our method of varying the skin type of an image while keeping other attributes constant. The low dimensionality of skin type has already been observed in the color theory literature [19], and the optimal transport method can be seen as a generalization of histogram equalization (where the histograms in question are the skin type histograms of the two faces). Regardless, the images that are produced by varying the skin type are not always visually realistic, and there are a number of potential reasons for this. For one, our methodology does not currently take into account variations in illumination in images; thus, the effects of illumination could be unintentionally suppressed or enhanced while varying the skin type via luminance-mode-shift or optimal transport.[7] An interesting improvement in our methodology would constitute applying our method to change the skin type *after isolating the illumination* on the image, the latter of which has seen excellent results recently [41].

More generally, the provided methodology for changing the skin type, even leaving aside illumination effects, may not necessarily capture relative differences in skin color

[7]It is, however, worth noting that the PPB dataset was carefully designed to minimize variation in illumination, to strengthen the conclusion that the discrepancies that arose were because of disparity in skin type [4].

across the face in an optimal fashion. While the optimal transport method *does* attempt to equalize the histograms of skin types across the entire face, it does not impose further spatial constraints on the histograms (for example, pixel values of cheeks and nose being close to one another). It is worth noting that while individual skin pixels have a relatively simple structure in color space, their spatial relationships do not and can vary for individuals of different skin types.

In computer vision, characterizing the *stability* to a feature, especially an underrepresented one, in an image is difficult because of the high dimensionality of the data. We purposely eschewed recent learning-based approaches that vary a feature (including skin "paleness") using likelihood based generative models [21] or conditional generative adversarial networks (GANs) [8] as they themselves are trained on unbalanced datasets in which darker skin types are underrepresented. There may be fundamental limitations in the ability of *any* post-processing approach to realistically vary the skin type, although it would be interesting to investigate approaches taken by computer graphics professionals. The ideal methodology would involve *naturally* varying an individual's skin type (from light to dark) using tanning – but the necessity of human participation in such an experiment would introduce a different set of limitations.

In spite of these limitations, we believe our methodology provides a useful starting point for evaluating the influence of skin type on classification decisions, at least at the pixel level. We encourage discussion and future research efforts on this important and complex problem.

7. Discussion and Future Work

We rigorously showed that the result of the gender classification task is relatively stable to variations in skin type and thus the skin type *by itself* has a minimal effect on the classification decision. We began this research with the aim of developing invariant or equivariant face classifiers that would ignore skin type completely and thereby have equal accuracy across groups. Such an approach would preclude the need for a high level of diversification in training datasets. However, our mathematically-oriented analysis using the low-dimensional skin type group revealed that high-performing gender classifiers are already invariant to skin type. In Section 6, we discussed inherent limitations of our methodology: mathematically tractable methods do not necessarily translate into visually realistic results. We emphasize that we view these results as useful initial insights into evaluating the influence of skin type on gender classification, and welcome discussion on improved methodology for doing this.

To solve the problem of unequal performance, it is quite possible that algorithmic approaches will be fundamentally limited. We really need diverse training datasets that represent humanity across many dimensions of identity, starting but not ending with ethnicity. Many questions remain as to how exactly to go about diversifying training data as even ethnicity does not fully encapsulate an individual's identity. This is a parallel issue to the issue of underrepresentation in skin type and correlated attributes: while females and males are balanced in training data, they are *stereotypical* females and males from the celebrity population. Informally speaking, we would expect the appearance of the general population of females and males alike to be quite different. We suggest that a good training dataset should diversify not only across ethnicity, but also across profession, cultural norms, and economic status, to capture a *truly* global population. Collecting such a dataset while controlling for image quality is a difficult, but necessary task.

As a parallel effort, it would also be interesting to examine the potential of decoupled classification on demographic groups, which along with task transfer learning has been shown to mitigate disparities in classification of other facial attributes across race and gender [39].

Finally, the perspectives presented here are limited to the problem of binary gender classification from visual data, itself a flawed problem especially when considering various non-binary gendered individuals. The community needs to move beyond the binary gender construct in future work.

Acknowledgment

The authors thank Joy Buolamwini and Karthikeyan Natesan Ramamurthy for their constructive comments on the work. This work was conducted under the auspices of the IBM Science for Social Good initiative.

References

[1] G. A. Akerlof and R. E. Kranton. *Identity Economics: How Our Identities Shape Our Work, Wages, and Well-Being*. Princeton University Press, 2010. 2

[2] Y. Bai and B. Ghanem. Multi-scale fully convolutional network for face detection in the wild. In *Proc. IEEE Conf. Comput. Vision Pattern Recogn. Workshops*, pages 2078–2087, 2017. 2

[3] R. K. E. Bellamy, K. Dey, M. Hind, S. C. Hoffman, S. Houde, K. Kannan, P. K. Lohia, J. Martino, S. Mehta, A. Mojsilović, S. Nagar, K. N. Ramamurthy, J. Richards, D. Saha, P. Sattigeri, M. Singh, K. R. Varshney, and Y. Zhang. AI fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. arXiv:1810.01943, 2018. 2

[4] J. Buolamwini and T. Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Proc. Conf. Fair. Account. Transp.*, pages 77–91, 2018. 1, 2, 3, 7

[5] J. A. Buolamwini. Gender shades: Intersectional phenotypic and demographic evaluation of face datasets and gender classifiers. Master's thesis, Massachusetts Institute of Technology, Sept. 2017. 1

[6] T. Calders, F. Kamiran, and M. Pechenizkiy. Building classifiers with independency constraints. In *Proc. IEEE Int. Conf. Data Min. Workshops*, pages 13–18, 2009. 2

[7] F. P. Calmon, D. Wei, B. Vinzamuri, K. N. Ramamurthy, and K. R. Varshney. Optimized preprocessing for discrimination prevention. In *Adv. Neur. Inf. Process. Syst.*, pages 3992–4001, 2017. 2

[8] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo. StarGAN: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proc. Conf. Comput. Vision Pattern Recogn.*, 2018. 3, 8

[9] M. Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Adv. Neur. Inf. Process. Syst.*, pages 2292–2300, 2013. 4

[10] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel. Fairness through awareness. In *Proc. Innov. Theor. Comput. Sci. Conf.*, pages 214–226, 2012. 2

[11] C. Dwork, N. Immorlica, A. T. Kalai, and M. Leiserson. Decoupled classifiers for group-fair and efficient machine learning. In *Proc. Conf. Fair. Account. Transp.*, pages 119–133, 2018. 2

[12] E. Eidinger, R. Enbar, and T. Hassner. Age and gender estimation of unfiltered faces. *IEEE Trans. Inf. Foren. Sec.*, 9(12):2170–2179, 2014. 1

[13] R. D. F. Feitosa, L. L. G. de Oliveira, D. L. Borges, and M. M. N. Filho. A mathematical model for reducing the likely spectrum of human skin tones in the RGB color space. In *Proc. IEEE Int. Conf. Imag. Syst. Tech.*, pages 329–334, 2014. 3

[14] S. Ferradans, N. Papadakis, G. Peyré, and J.-F. Aujol. Regularized discrete optimal transport. *SIAM J. Imag. Sci.*, 7(3):1853–1882, 2014. 4

[15] T. B. Fitzpatrick. The validity and practicality of sun-reactive skin types I through VI. *Arch. Derm.*, 124(6):869–871, 1988. 1

[16] S. Fu, H. He, and Z.-G. Hou. Learning race from face: A survey. *IEEE Trans. Pattern Anal. Mach. Intel.*, 36(12):2483–2509, 2014. 1

[17] C. Garvie, A. M. Bedoya, and J. Frankle. *The perpetual line-up: Unregulated police face recognition in America*. Georgetown Law, Center on Privacy & Technology, 2016. 2

[18] M. Hardt, E. Price, and N. Srebro. Equality of opportunity in supervised learning. In *Adv. Neur. Inf. Process. Syst.*, pages 3315–3323, 2016. 2

[19] R.-L. Hsu, M. Abdel-Mottaleb, and A. K. Jain. Face detection in color images. *IEEE Trans. Pattern Anal. Mach. Intel.*, 24(5):696–706, 2002. 3, 7

[20] G. B. Huang, M. Mattar, T. Berg, and E. Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. In *ECCV Workshop Faces Real-Life Imag.*, 2008. 1

[21] D. P. Kingma and P. Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. In *Adv. Neur. Inf. Process. Syst.*, pages 10215–10224, 2018. 3, 8

[22] B. F. Klare, M. J. Burge, J. C. Klontz, R. W. V. Bruegge, and A. K. Jain. Face recognition performance: Role of demographic information. *IEEE Trans. Inf. Foren. Sec.*, 7(6):1789–1801, 2012. 2

[23] B. F. Klare, B. Klein, E. Taborsky, A. Blanton, J. Cheney, K. Allen, P. Grother, A. Mah, and A. K. Jain. Pushing the frontiers of unconstrained face detection and recognition: IARPA Janus Benchmark A. In *Proc. IEEE Conf. Comput. Vision Pattern Recogn.*, pages 1931–1939, 2015. 1

[24] G. Levi and T. Hassner. Age and gender classification using convolutional neural networks. In *Proc. IEEE Conf. Comput. Vision Pattern Recogn. Workshops*, pages 34–42, 2015. 2

[25] L. T. Liu, S. Dean, E. Rolf, M. Simchowitz, and M. Hardt. Delayed impact of fair machine learning. In *Proc. Int. Conf. Mach. Learn.*, pages 3150–3158, 2018. 2

[26] Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In *Proc. IEEE Int. Conf. Comput. Vision*, pages 3730–3738, 2015. 1

[27] Z. Lu, X. Jiang, and A. Kot. Color space construction by optimizing luminance and chrominance components for face recognition. *Pattern Recogn.*, 83:456–468, 2018. 3, 4

[28] M. Mathias, R. Benenson, M. Pedersoli, and L. Van Gool. Face detection without bells and whistles. In *Proc. Eur. Conf. Comput. Vision*, pages 720–735, 2014. 2

[29] M. Ngan, M. Ngan, and P. Grother. *Face Recognition Vendor Test (FRVT) Performance of Automated Gender Classification Algorithms*. US Department of Commerce, National Institute of Standards and Technology, 2015. 1, 2

[30] M. M. Oghaz, M. A. Maarof, A. Zainal, M. F. Rohani, and S. H. Yaghoubyan. A hybrid color space for skin

detection using genetic algorithm heuristic search and principal component analysis technique. *PLOS One*, 10(8):e0134828, 2015. 3, 4

[31] V. A. Oliveira and A. Conci. Skin detection using HSV color space. In *Brazilian Symp. Comput. Graph. Imag. Process.*, 2009. 3

[32] Özlem Özbudak, M. Kırcı, Y. Çakır, and E. O. Güneş. Effects of the facial and racial features on gender classification. In *Proc. Mediterran. Electrotechnical Conf.*, pages 26–29, 2010. 1

[33] O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. In *Proc. Brit. Mach. Vision Conf.*, page 41, 2015. 2

[34] G. Perarnau, J. van de Weijer, B. Raducanu, and J. M. Álvarez. Invertible conditional GANs for image editing. arXiv:1611.06355, 2016. 3

[35] I. D. Raji and J. Buolamwini. Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial AI products. In *Proc. AAAI/ACM Conf. Artif. Intel. Ethics Society*, 2019. 2

[36] R. Ranjan, S. Sankaranarayanan, C. D. Castillo, and R. Chellappa. An all-in-one convolutional neural network for face analysis. In *Proc. IEEE Int. Conf. Auto. Face Gesture Recogn.*, pages 17–24, 2017. 2

[37] D. A. Reid, S. Samangooei, C. Chen, M. S. Nixon, and A. Ross. Soft biometrics for surveillance: An overview. *Handbook Statist.*, 31:327–352, 2013. 2

[38] R. Rothe, R. Timofte, and L. Van Gool. Deep expectation of real and apparent age from a single image without facial landmarks. *Int. J. Comput. Vis.*, 126(2-4):144–157, 2018. 1, 2, 3

[39] H. J. Ryu, H. Adam, and M. Mitchell. InclusiveFaceNet: Improving face attribute detection with race and gender diversity. In *Proc. Fair. Account. Transp. Mach. Learn. Workshop*, 2018. 8

[40] P. Sattigeri, S. C. Hoffman, V. Chenthamarakshan, and K. R. Varshney. Fairness GAN: Generating datasets with fairness properties using a generative adversarial network. In *Proc. ICLR Workshop Safe Mach. Learn.*, 2019. 2

[41] S. Sengupta, A. Kanazawa, C. D. Castillo, and D. W. Jacobs. SfSNet: Learning shape, reflectance and illuminance of faces 'in the wild'. In *Proc. Conf. Comput. Vision Pattern Recogn.*, pages 6296–6305, 2018. 7

[42] T. Sim, S. Baker, and M. Bsat. The CMU pose, illumination, and expression database. *IEEE Trans. Pattern Anal. Mach. Intel.*, 25(12):1615–1618, 2003. 2

[43] Y. Wen, K. Zhang, Z. Li, and Y. Qiao. A discriminative feature learning approach for deep face recognition. In *Proc. Eur. Conf. Comput. Vision*, pages 499–515, 2016. 2

[44] B. Wilson, J. Hoffman, and J. Morgenstern. Predictive inequity in object detection. arXiv:1902.11097, 2019. 2, 4

[45] S. Zafeiriou, C. Zhang, and Z. Zhang. A survey on face detection in the wild: Past, present and future. *Comput. Vision Image Understand.*, 138:1–24, 2015. 2

[46] I. Žliobaitė. On the relation between accuracy and fairness in binary classification. In *Proc. Fair. Account. Transp. Mach. Learn. Workshop*, 2015. 2