RACIAL BIAS IN FACE CLASSIFICATION METHODS

May Jiang
Princeton University
mayjiang@princeton.edu

Claire Du
Princeton University
clairedu@princeton.edu

Juliet Oh
Princeton University
jjoh@princeton.edu

January 14, 2020

Code and data for project available at https://github.com/mjiang9/cos429project

Contents

1 Motivation

Recent studies indicate that automated facial analysis algorithms have inherent bias with respect to factors such as race and gender. Researchers at MIT and Microsoft tested various commercial classification systems and found that darker-skinned females are misclassified with an error rate of up to 34.7%, while lighter-skinned males have a maximum error rate of only 0.8% [1]. Such algorithmic bias often originates from biased training datasets, and has important implications for applications ranging from law enforcement to autofocus capabilities in digital cameras.

Some work has been done to investigate the sources of the discrepancy in performance across different groups. However, most of these studies tend to focus on large, complex commercial facial classification systems and do not fully explore how and why the algorithms themselves might exhibit differing robustness to variations in race and gender. We are interested in comparing the sensitivity of different approaches to facial classification with respect to race, and analyzing the reasons why they might exhibit differing performance. Having a more comprehensive understanding of the behavior of basic facial classification algorithms with respect to racial bias will help inform the design of more robust facial analysis systems.

2 Goal

The goal of our project is to analyze the performance of different algorithmic approaches to face classification on training datasets with varying levels of racial bias. Some work has been done in this area before, particularly on comparing different commercial classification systems, but we would like to investigate more closely how and why different algorithms respond differently to biased datasets. We will focus on three classification methods, logistic regression, random forest, and eigenvalue decomposition (eigenfaces), and two classification features, Haar-like features and HOG feature descriptors.

3 Background and Related Work

A number of studies have investigated race and gender bias in facial classification systems. Many commonly used training datasets and testing benchmarks are highly imbalanced with respect to race and gender, resulting in classifiers that in turn also have biased performance. Buolamwini evaluated

the accuracy of three commercial facial gender classifiers on a gender and skin-color balanced dataset and found that all three classifiers performed significantly worse on darker-skinned females [1]. Muthukumar et al. explored how specific facial features drive algorithmic bias, and concluded that facial structure, not skin type, is the main driver of algorithmic bias in gender classification tasks [2].

Some work has also been done in designing facial analysis systems that are more accurate across gender and race subgroups. InclusiveFaceNet is a face attribute detection model that utilizes learned demographic representations to achieve exceptional accuracy on downstream detection tasks [3]. By providing more insight into the behavior of basic classification algorithms with respect to racial bias, our work will assist continued efforts to develop more inclusive facial analysis systems.

4 Approach

4.1 Data and Preprocessing

For this project, we used the UTKFace and SUN datasets. The UTKFace dataset contains images of cropped, aligned, and generally forward-facing faces with filenames labeled by race, gender, and age [4]. SUN dataset consists of environmental images that served as our non-face set. Images from UTKFace database were made grey-scale and resized to be 36 by 36 pixels, while patches of at least 36 by 36 pixels were randomly sampled from the SUN dataset. Figure 1 shows sample images taken from the two datasets.

For our experiments, five training sets of 4020 face images were created, with $0/100$, $25/75$, $50/50$, $75/25$, and $100/0$ ratios of black, or darker-skinned faces, to white, or lighter-skinned, faces. While the ratio of darker-skinned to lighter-skinned varies across the training sets, image files from the original dataset were parsed and selected from the training sets such that the distribution of age and gender are balanced. The held-out testing set consisted of 300 faces with an even distribution of race, age, and gender, and 300 non-faces from the SUN dataset.

Figure 1: Sample Images from UTKFace (left) and SUN (right) Datasets

4.2 Classification Features

4.2.1 HOG

Histogram of Oriented Gradients (HOG) feature descriptors are normalized local histograms of image gradient orientations and are often a good representation of local object shapes and features.

Dalal and Triggs developed a pedestrian detector using HOG descriptors and demonstrated very good performance on pedestrian datasets [5].

4.2.2 Haar

Haar-like features are simple rectangle filters that are applied to regions of an image by subtracting the pixel intensities of certain rectangular portions of the region from others. They can be quickly computed using integral images. Viola and Jones boosted Haar-like features in an attentional cascade to create a high-performing object detection framework [6]. Examples of some of the most predictive two- and three-rectangle Haar features computed on face images from our datasets are shown in Figure 2. We hypothesized that Haar features would be less robust to racial bias compared to HOG feature descriptors due to their direct dependence on shading and light intensity values.

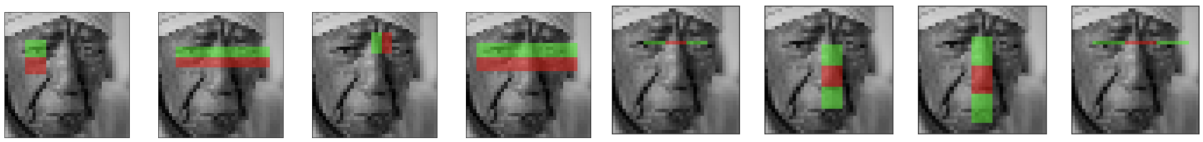

Figure 2: Two and Three rectangle Haar features

4.3 Classification Methods

We test three different classification methods: regularized logistic regression, random forest classification, and eigenvector decomposition.

4.3.1 Regularized Logistic Regression

Regularized logistic regression is a learning method that uses a generalized linear model to perform binary classification. The loss function applies a nonlinearity to the system of variables:

$$L = \sum (z_i - (a + bx_{i1} + cx_{i2} + dx_{i3} + \ldots)^2 \tag{1}$$

The nonlinearity used is the logistic function:

$$s(z) = \frac{1}{1 + e^{-z}} \tag{2}$$

We implement regularized logistic regression using the Gauss-Newton iterative method.

4.3.2 Random Forest Classification

Random forest classification is an ensemble classification method that uses decision trees. The algorithm runs in the following steps [7]:

1. Selecting bootstrap samples from the training data set
2. For each of the bootstrap samples, grow a decision tree, choosing the best split among a random sample of the predictors
   - A decision tree is built by recursively splitting the source set into subsets, with splitting rules based on classification features [8]
3. Make predictions by aggregating the predictions of the decision trees; since our project deals with classification, these predictions are formed by taking majority votes over the trees

An advantage of decision tree algorithms is that there is a clear and straightforward method of ranking feature importance, based on how much a feature decreases the weighted Gini impurity in a tree. Gini impurity is a measure of how often a randomly selected element of the set would be misclassified, calculated as the probability of each element being chosen multiplied by the probability that it is misclassified [8].

This ordering of feature importances enables us to perform feature selection, by selecting the top k features that are most predictive. Moreover, it allows us to inspect the most important features to shed light on the features - in this project, the characteristics of an image - that best determine whether an image is a face.

A disadvantage of random forest classification is that it is highly complex, because a decision tree can have many branches, and with a large number of features the tree depth can be very high. While this enables decision tree algorithms such as random forest to achieve strong performance by capturing complex properties and feature relationships, it can also result in overfitting.

In this project we use the implementation of random forest classification from the scikit-learn library [9].

4.3.3 Eigenvector Decomposition (Eigenfaces)

The eigenface algorithm was developed as an approach to face recognition that essentially depends on template matching, using eigenvector decomposition to obtain a low-dimensional representation of faces in the "face space". The recognition algorithm works by:

1. Finding a low dimensional representation of a face by computing the normalized covariance matrix of all the faces and keeping the top k eigenvectors corresponding to the highest eigenvalues
2. Representing faces as a combination of these eigenvectors by:
    - Subtracting the mean to normalize
    - Projecting onto the low dimensional face eigenspace. The projection of the face onto the eigenspace is computed as the dot product of the face vector with each eigenvector, since the eigenvectors have unit norm.

The resulting vector of k dimensions represents the image in face space as a weighted combination of the top face eigenvectors that comprise the basis. Face recognition for a given image is then achieved by finding the face in the training set, represented in this same space, that minimizes the distance (Euclidean norm or Mahalanobis distance) to that image [10].

For our purposes, we adapted this face recognition algorithm for face classification, which shares the same underlying principle of using distances between images in a lower dimensional face space. In their paper on using eigenfaces for face recognition, Turk and Pentland write that eigenfaces can be used to recognize a new face by verifying that the distance of an image from the face space is sufficiently small. That is, to classify rather than recognize or match a face, instead of computing distances between a given image and other faces in the set, we compute the distance from a face in n dimensions to its reconstruction using its representation in the k dimensional space and the k eigenvectors. If this "distance from face space" is small enough we classify the image as a face, otherwise as a non-face. To define "small enough" we examined the distribution of distance from face space for all the images in a set and used a threshold distance that would maximize accuracy in predicting face versus non-face.

4.4 Evaluation

To evaluate our classifiers we examine the accuracy (number of correctly classified images out of all images) for the test set overall (in plot legends denoted by "o"), dark-skinned faces (in plot legends denoted by "b"), and light-skinned faces (in plot legends denoted by "w"). For our HOG and Haar feature face classifiers we ran 10 trials, regenerating the non-faces and retraining the classifier in each trial, and plotted the mean accuracy values and one standard deviation above and below.

5 Testing Face Classification Performance

5.1 HOG Feature Descriptors

We compute the HOG descriptors for the images in each training set, train a face/nonface classifier, and evaluate the overall performance accuracy on a balanced test dataset of 300 faces (150 light-skinned, 150 dark-skinned) and 300 non-faces. We also evaluate the separate accuracies on light and dark-skinned images. Results are displayed in Figure 3.

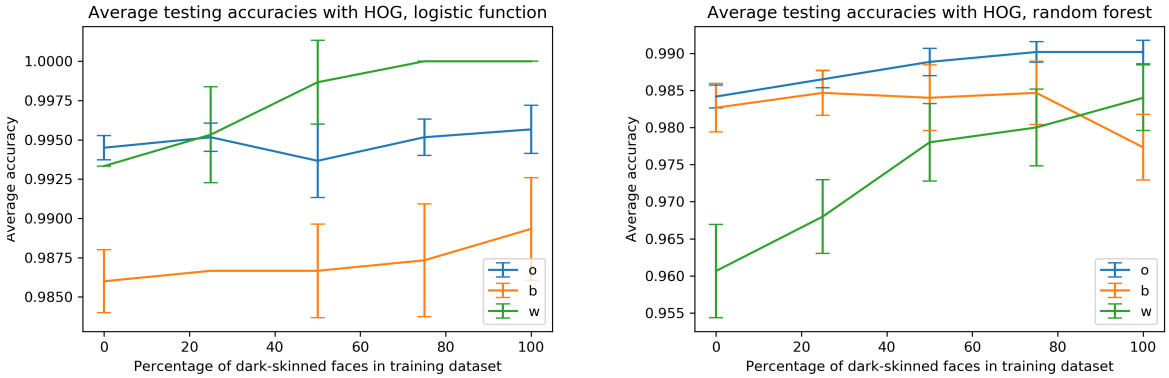

Figure 3: Testing accuracies of face/nonface classifier trained on datasets with varying percentages of dark-skinned faces, averaged over 10 trials. Training datasets consisted of 4020 faces and 4020 non-faces. 'o' denotes overall performance, 'b' performance on dark-skinned faces, and 'w' performance on light-skinned faces.

Overall testing accuracy was higher for classification with logistic function than for random forest. Logistic function performed marginally better on light-skinned faces than dark-skinned faces across all training sets, but did not show sensitivity to different training set compositions. Performance on dark-skinned faces was similar for both logistic function and random forest. Interestingly, random forest performed better on dark-skinned faces than light-skinned faces for most training sets, and performance on light-skinned faces increased significantly with increasing percentages of dark-skinned training examples.

Figure 4 shows a visualization of average HOG descriptors computed on training sets with 0%, 50%, and 100% dark-skinned faces. As expected, the descriptors are largely similar, with slight differences in the nose and mouth regions; intensities are higher in these regions for the 100% dark-skinned faces training set. This might explain why random forest performance improved with increasing proportions of dark-skinned training examples. HOG descriptors for light-skinned faces emphasize the eye region, while descriptors for dark-skinned faces emphasize other facial features such as the nose and mouth, so increasing the proportion of dark-skinned training faces improves the ability of the classifier to recognize faces in general.

These results suggest that although HOG descriptors show some sensitivity to race, the classification method used has a more significant impact on performance.

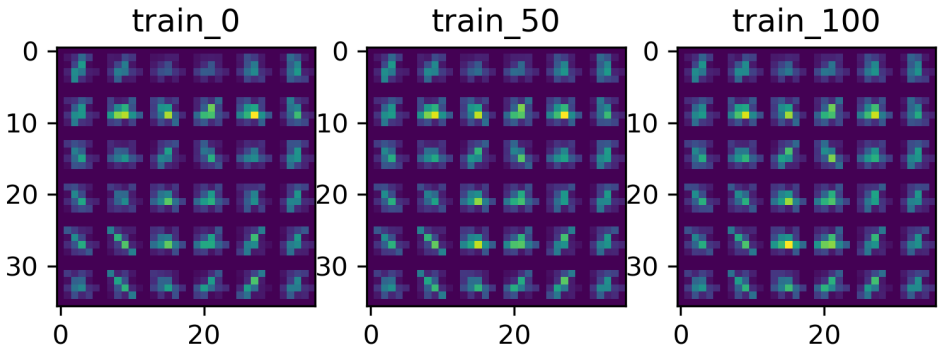

Figure 4: Average HOG features computed across three training sets of 4020 images, with 0%, 50%, and 100% dark-skinned faces, respectively.

5.2 Haar-like Features

We compute two- and three-rectangle Haar features for the images in each training set, select top features and train a face/nonface classifier, and evaluate performance on a balanced testing set.

5.2.1 Feature Selection

The number of possible features consisting of two rectangles in a 36 by 36 image is 403,920, and the number of possible features consisting of two or three rectangles in a 36 by 36 image is over 700,000. As a result, training the classifier became infeasible. To address this, we performed feature selection using the random forest classifier. On a set of 300 faces and 300 non-faces, balanced across race, gender, and age, we trained a random forest classifier with all 700,000+ rectangle Haar features. We then examined the feature importances produced by the random forest classifier. Figure 5 depicts the feature importances - we found that 70% of the branch points in the random forest could be explained by approximately the first 967 features, so we decided to use only the 900 most important features for our classifiers, as our HOG classifiers also used 900 features.

5.2.2 Classifiers

Using the reduced feature set of 900 Haar-like features, we trained face classifiers using both logistic regression and random forest classification. The results are displayed in Figures 6 and 7 for logistic

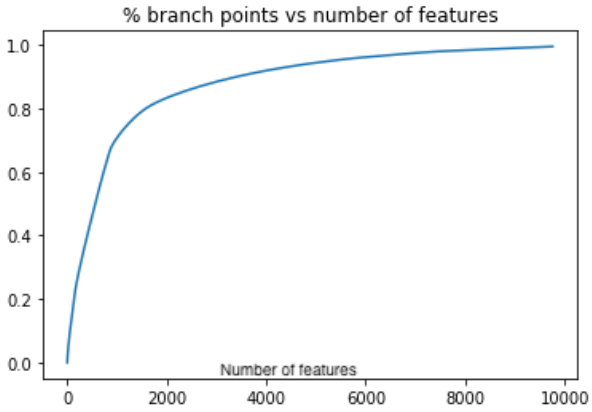

Figure 5: Plot of the percentage of covered branch points against the number of Haar features. A small number of features explains most of the branch points.

regression and random forest, respectively. We found that both models visibly exhibited much sensitivity to the biased datasets, performing more poorly on dark-skinned faces overall, and with accuracy increasing significantly as the proportion of dark-skinned faces represented in the training set increased [11]. In particular, we note that the accuracy in classifying dark-skinned faces was lower than that in classifying light-skinned faces in every training set, even when the training data was composed of a majority or even a unanimity of dark-skinned faces. We reasoned that this could potentially be ascribed to the higher contrast in intensities between light and dark areas on light-skinned faces compared to dark-skinned faces, and we investigate the top features and the effect of sharpening the images to evaluate this hypothesis in later sections.

Between the two classification methods, the logistic regression model resulted in higher variance, visualized as the standard deviation error lines (dotted) in the figures, as well as higher accuracy in identifying dark and light skinned faces as faces, although consistently lower accuracy on the test set overall - indicating a general propensity towards classifying any given image as a face. This may be attributed to logistic regression being a generalized linear model, which always depends on a sum of inputs and parameters, compared to the random forest classifier which has the ability to capture more complex and nonlinear relationships because of the many branches in its decision trees. Accordingly, given an important feature such as a Haar-like rectangle feature representing one eye, the logistic regression model might weight the feature heavily and be more likely to classify any image with that single feature as a face. On the other hand, a random forest model would use that feature as only one branch split, and would therefore be able to capture nonlinear relationships such as conditionals. For the eye example, a random forest might only classify an image as a face if there are two eyes present. We see this reflected in the stronger performance of the random forest classifier in classifying non-face images, as there are many such nonlinear relationships in determining whether a given image is a face.

Because the random forest model had lower variance, higher performance classifying non-faces, and also provided a means of inspecting the top features by feature importance, we used the random forest model for further analysis of the sensitivity of Haar features to racial bias from the imbalanced datasets.

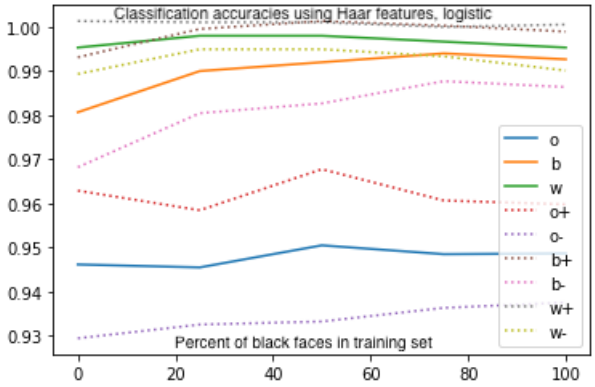

Figure 6: Average testing accuracies of logistic Haar classifier trained on datasets with varying percentages of dark-skinned faces. 'o' denotes overall performance, 'b' performance on dark-skinned faces, and 'w' performance on light-skinned faces. Dashed lines indicate +/- 1 standard deviation.

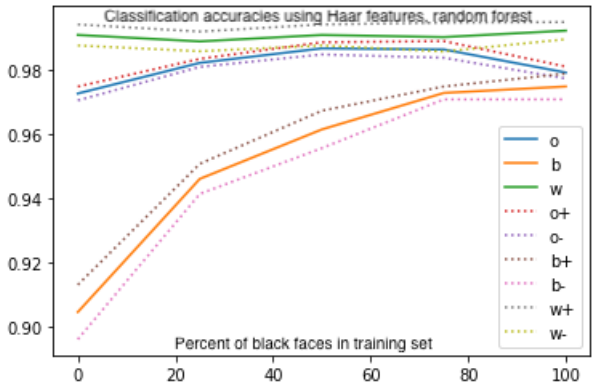

Figure 7: Average testing accuracies of random forest Haar classifier trained on datasets with varying percentages of dark-skinned faces.

5.2.3 Most Predictive Features

To gain insight into the characteristics of dark-skinned faces and light-skinned faces that result in differences in performance on classification, we examined the most predictive features for determining face/non-face for the classifiers trained on the 0%, 25%, 50%, 75%, and 100% ratio data sets, depicted in Figures 8, 9, 10, 11, and 12 respectively.

As the composition of dark-skinned faces in the training set changes, the most important rectangle filters also change. For a training set composed of only light-skinned faces, the top eight features are all horizontal filters that emphasize the horizontal eye region. However, as the proportion of dark-skinned faces increases, the most important features include more vertical filters that emphasize the eye and nose region. This suggests that for distinguishing light-skinned faces from non-faces, the contrast between the eyes and cheek area was most useful, but that this was not easily generalized to classifying dark-skinned faces. The most determinant features in dark-skinned faces, on the other hand, were more universal facial features such as the eyes and nose. This discrepancy explains why the performance on dark-skinned faces depends positively on the proportion of dark-skinned faces while the performance on light-skinned faces did not suffer when the proportion of light-skinned faces decreased accordingly.

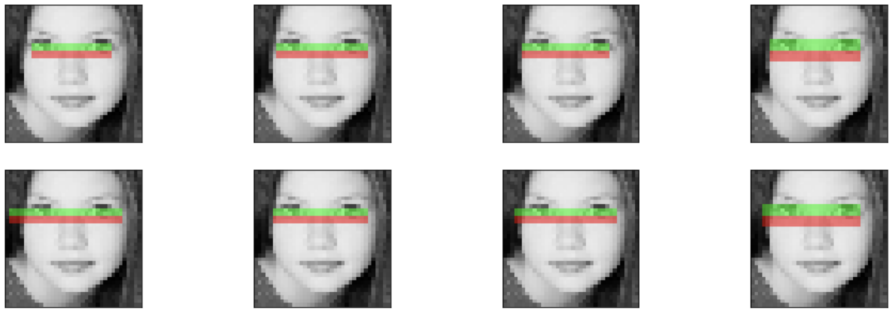

Figure 8: Top Haar features for 0% dark-skinned faces set

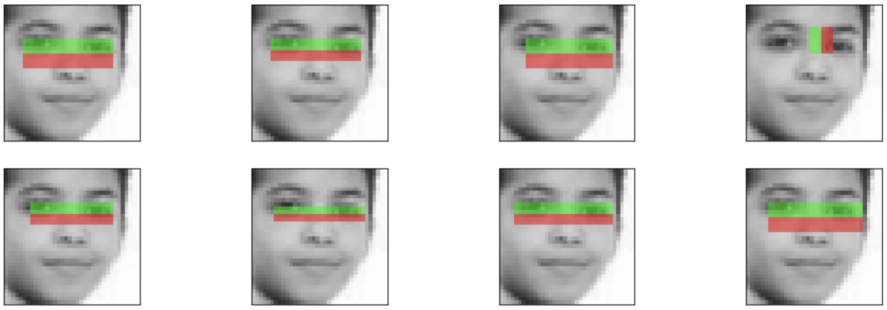

Figure 9: Top Haar features for 25% dark-skinned faces set

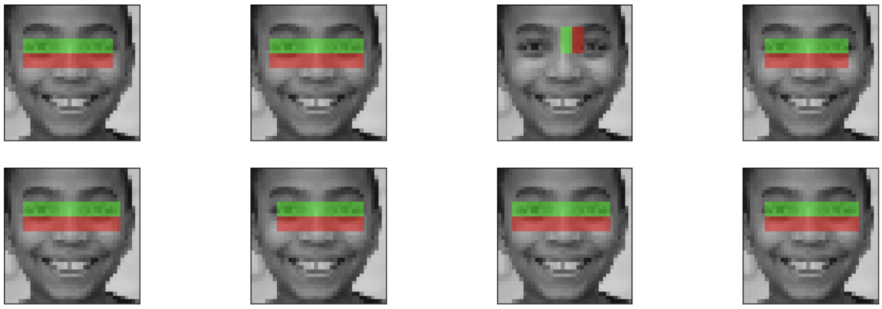

Figure 10: Top Haar features for 50% dark-skinned face set

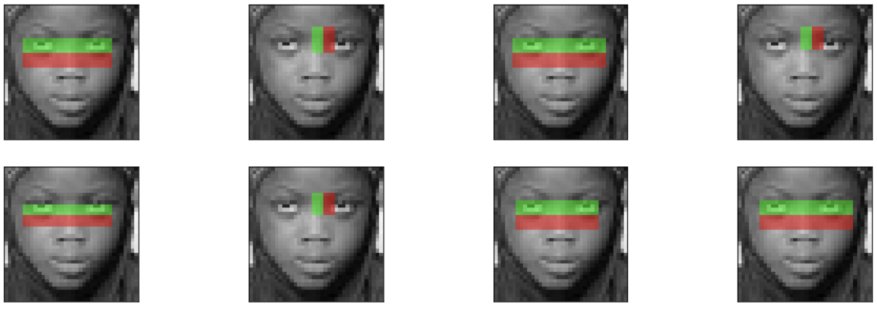

Figure 11: Top Haar features for 75% dark-skinned face set

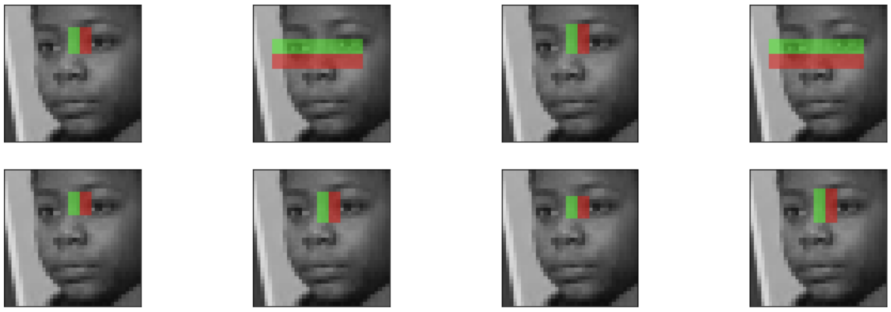

Figure 12: Top Haar features for 100% dark-skinned face set

5.3 Comparison of Methods

Between face classifiers trained with HOG feature descriptors and Haar-like features, we found that HOG feature descriptors were significantly more robust to racial bias than Haar-like features. This was supported by the observation that performance patterns of HOG classifiers trained with logistic function versus random forest varied greatly, whereas for Haar features both classifiers exhibited similar patterns, with performance on dark-skinned faces improving dramatically with increasing proportions of dark-skinned training examples. This difference in robustness could be explained by the types of features themselves; whereas HOG is based on gradient orientations and overall face shape, Haar rectangle features directly depend on the pixel intensity of different regions of the face and are consequently more sensitive to factors such as skin tone and contrast of facial features.

6 Understanding Bias and Performance

We conduct a number of analyses to explore the differences in race sensitivity between HOG and Haar and the origins of the sensitivity.

6.1 Intra- and Inter-class Distances

We compute average intra- and inter-class Euclidean distances for dark-skinned and light-skinned faces, for both types of features, to see if there is a difference between features. The results are shown below in Table 1.

Table 1: Average Intra- and Inter-class Distances of Feature Descriptors

	Hog	Haar-like
Intraclass Light-Skinned Faces	4.581	37700.129
Intraclass Dark-Skinned Faces	4.743	38354.826
Interclass Dark-Skinned and Light-Skinned faces	4.734	37920.252

For both Haar and HOG features, the average intra-class Euclidean distance for dark-skinned faces is greater than that for light-skinned faces, suggesting that for both features there is more variation between dark-skinned faces. Such variation could present an obstacle to classifying faces within the same class and could help explain the discrepancy in performance. However, with our conclusion that Haar features are more sensitive to racial bias, we don't see the expected result of inter-class

distance being greater than both intra-class distances for Haar features. This could be because Euclidean distance does not fully capture the differences between classes. We next conduct principal component analysis to try to reveal more insight into the distinction between HOG and Haar features for the face classes.

6.2 Principal Component Analysis

To further investigate the underlying differences between Haar-like features and HOG feature descriptors for dark-skinned faces and light-skinned faces, we perform principal component analysis to inspect and compare the top principal components (PC) from the face images. Dimensionality reduction was performed on all feature vectors and plotted to compare two principal components at a time for Hog and Haar-like features. Figure 13 shows the PCA plot for principal components 1 and 2. For HOG feature descriptors there is almost no distinction between dark-skinned and light-skinned faces, whereas for Haar features there is a significant distinction; we see a clear pattern of separation between and clustering within the classes. Principal components 1 and 4 also exhibit this pattern (Figure 14). Principal components 2 and 3 still distinguish between dark-skinned and light-skinned examples, but the difference is less dramatic (Figure 15).

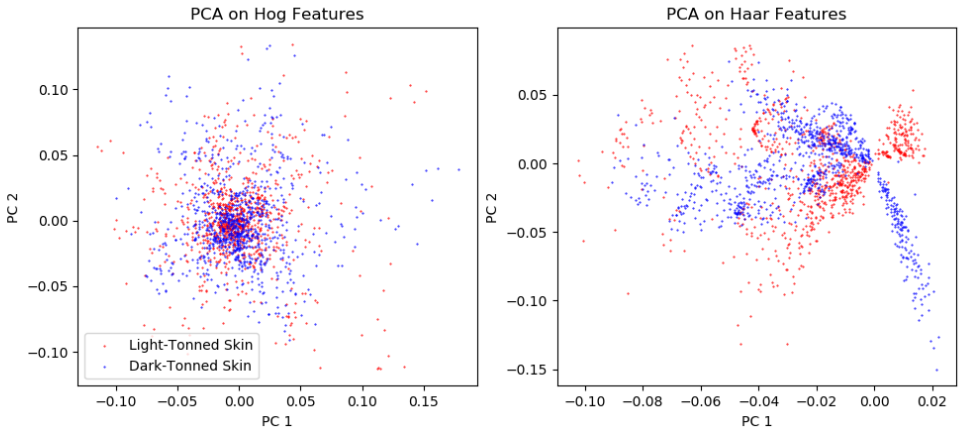

Figure 13: Principal Components 1 v. 2 on Hog and Haar Features

6.3 Predicting Race from Features

As a final comparison between Haar and Hog features, we built a classifier using each feature type to predict whether a given face image was a light-skinned face or a dark-skinned face. The classifier was trained with random forest on the 50% training set of 4020 face images, with 2010 light-skinned and 2010 dark-skinned faces balanced across gender and age, to predict the race label of the image. Using Haar features, the classifier attained 83.6% accuracy in predicting the race label on the balanced test set of 300 images, while HOG features attained only 49% accuracy, which is no better than chance. This supports our conclusion that Haar features are more sensitive to race.

Figure 16 shows the most predictive Haar-like rectangle features in determining whether a face was dark-skinned or light-skinned. These top features, primarily reflecting the shapes, sizes, and positions of the eye and nose area of the faces, indicate that the difference in sensitivity to bias

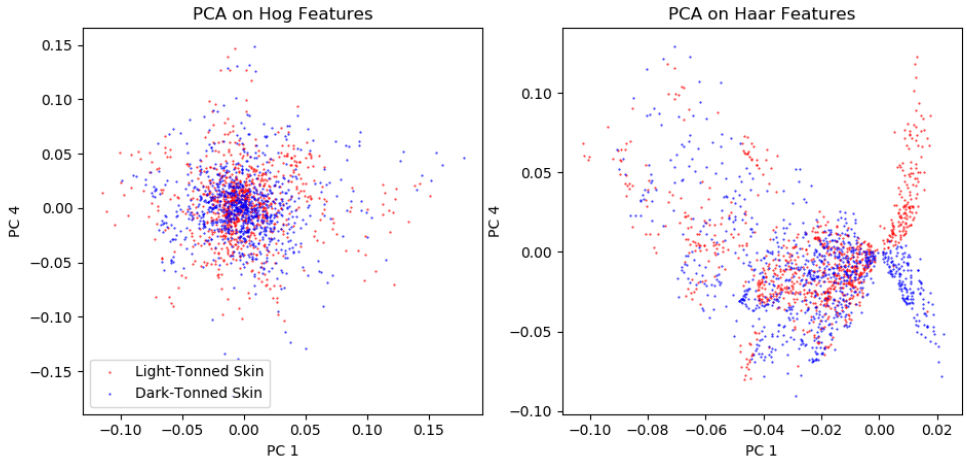

Figure 14: Principal Components 1 v. 4 on Hog and Haar Features

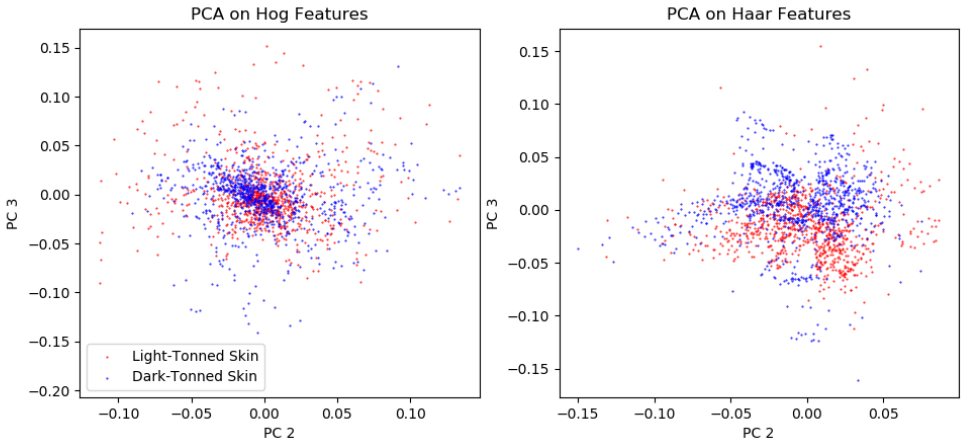

Figure 15: Principal Components 2 v. 3 on Hog and Haar Features

cannot be explained by skin tone alone, but rather is also heavily influenced by facial structure characteristics.

6.4 False Positives and False Negatives

Studying the false positive and false negative images also reveals the characteristics and factors that determine faces and non-faces. Figure 17 shows examples of false negative images (faces that were not correctly classified as faces) and false positive images (non-faces that were incorrectly classified as faces) for a Haar classifier trained on a balanced dataset of 4020 images, and Figure 18 shows false negatives and positives for the similarly trained HOG classifier.

The false negatives for the Haar classifier included many dark-skinned faces, typically in low-lighting, suggesting that the performance may have been negatively impacted by lower contrast between dark areas and light areas that characterize a face. The false positives for the Haar classifier consisted of images with regions resembling eyes or noses, illustrating that Haar features focus on

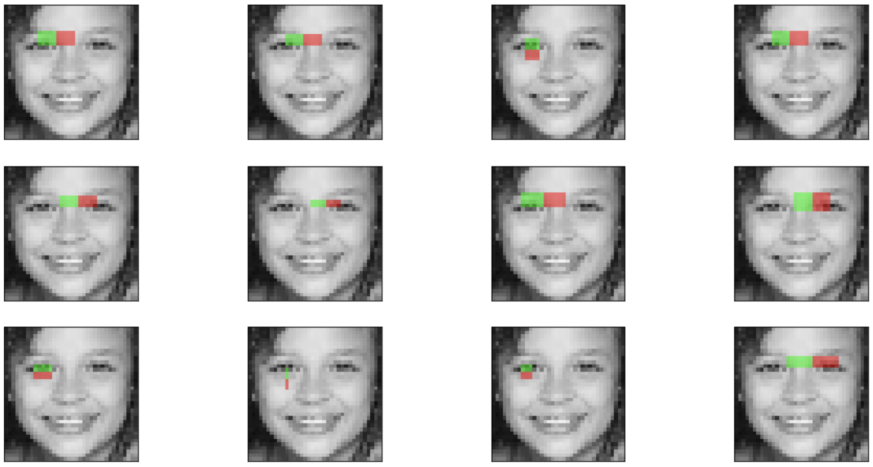

Figure 16: Top Haar features from race classifier.

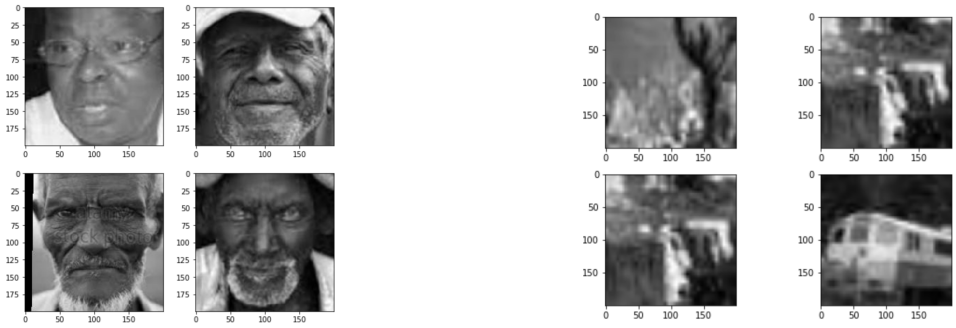

Figure 17: False negatives and false positives by Haar classifier trained on balanced dataset of 4020 examples.

individual facial features rather than overall shape, as expected based on our analysis and results in previous sections.

The false negatives for the HOG classifier also include mostly dark-skinned faces, but in contrast to Haar most of these faces were not front facing but rather at an angle. The false positives for the HOG classifier consisted mostly of circular shapes. This illustrates that HOG descriptors capture the overall shape of the face while Haar features focus on individual features, which explains the greater race sensitivity of Haar features.

6.5 Eigenfaces

To better understand the characteristics that distinguish between faces and non-faces, as well as between light-skinned faces and dark-skinned faces, we implemented the eigenface algorithm for face classification based on eigenvector decomposition and analyzed the face images as vectors. Figure 19 shows the values of the eigenvalues of the eigenvectors in decreasing order; this scree plot indicates that there is a much lower dimensional representation of the training data, as the elbow shows that the first ten or so eigenvectors have higher eigenvalues but by around the 20th eigenvector there is a significant drop in the eigenvalue. Accordingly, we use the top 30 eigenvectors with the highest eigenvalues to form our lower dimensional representation of the face space. Figure 20 shows the top eight eigenfaces that represent the basis of the face space. We see that these eigenfaces,

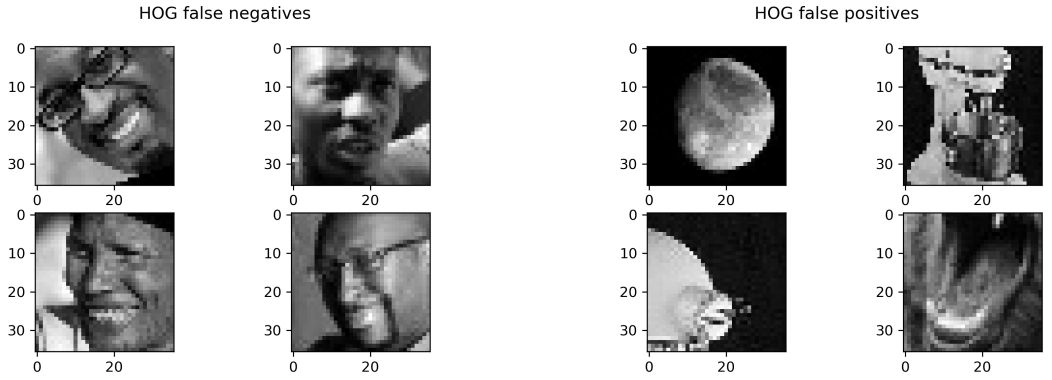

Figure 18: False negatives and false positives by HOG classifier trained on balanced dataset of 4020 examples.

which represent dimensions of the face space, are all immediately recognizable as face-like, with varying degrees of light intensity, facial structure, and facial feature emphases.

Figure 21 shows the distribution of distances from the face space among images in the test set. We see no clear pattern among the test set to distinguish between faces and non-faces, as there is no clear split between the distances, and therefore use a threshold close to the center of the distribution that maximizes the prediction accuracy locally to predict face versus nonface.

Figure 22 shows the accuracies on the held-out test set resulting from using the eigenfaces for face classification. We see that the classification accuracies are lower than the HOG and Haar feature classifiers in every category and that this face classifier is also subject to the racial bias sensitivity in the imbalanced training data. Whereas HOG features represented the shape of a face and the orientations of its defining features and Haar features represented the patterns of dark and light regions in a face, the distance from face space in the eigenfaces approach represents the more fundamental underlying structure of a face and patterns in the lighting, shape, and form of a face. However, because the "training" portion of the classifier does not involve non-faces, eigenface classification draws only from closeness to the face eigenbasis rather than what distinguishes a face from a non-face. In other words, the weakness of the eigenface algorithm for face classification may stem from its reliance on similarity to a face rather than dissimilarity to a non-face. Thus, the eigenface algorithm is ultimately not optimal for the task of classification because distance from face space is not sufficiently representative of whether a given image is a face or non-face; it is more suited for recognition of an image already identified as a face, because in this case the comparison portion of the task is constrained to a set of faces and the algorithm needs only to identify the closest face.

Nonetheless, while the eigenface algorithm proved weak as an approach to face classification, analysis of the classes of images as vectors revealed insight into both the weakness of the eigenface classifier and the differences between dark and light skinned faces that could lead to biased performance. In particular, we found that the mean distance between dark-skinned face images and the mean of all face images, was 1962.2 whereas the mean distance between light-skinned face images and the mean of all faces was 1869.6. This would help to explain the discrepancy in the classifier performance between dark and light-skinned faces; because the dark-skinned faces are in general further from the average face in vector distance terms, they are more difficult to classify. Moreover, the mean distance between dark-skinned face images and the mean of all dark-skinned

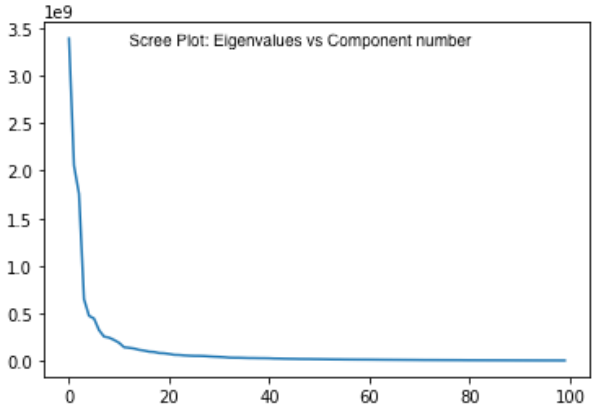

Figure 19: Scree Plot of Eigenvalues

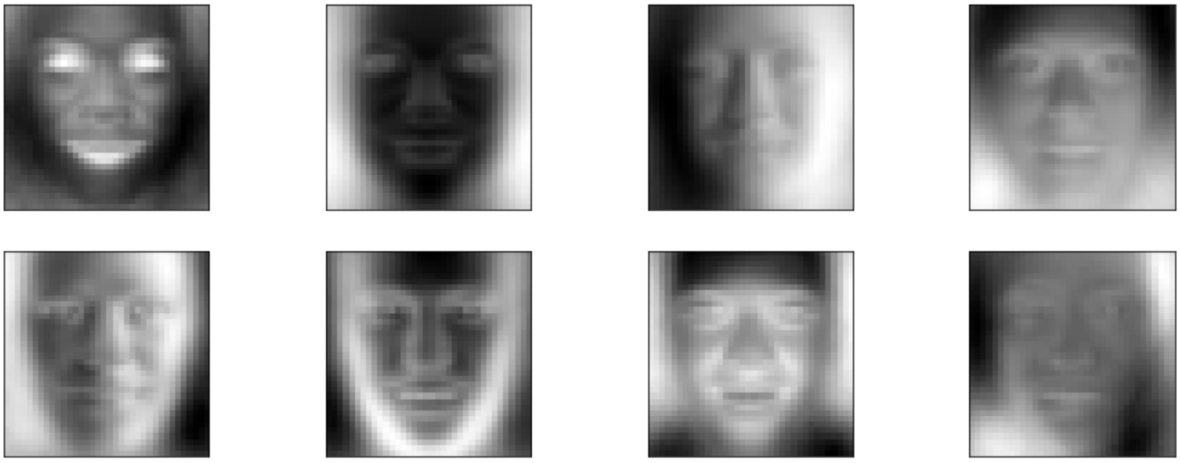

Figure 20: Top eigenfaces that represent the basis of the lower-dimensional face space.

faces was also higher than the mean distance between light-skinned face images and the mean of all light-skinned faces, 1891.9 compared to 1807.5. This suggests that there may be more variation in vector distance terms among the dark-skinned faces, which also contributes to the weakness in classifying dark-skinned faces we found in various classification methods.

Figure 23 visualizes the mean vector of light-skinned and dark-skinned faces. We see that they are similar but with slight differences in overall face shape and width, lighting, structural patterns of the forehead and chin, along with relative position, orientation, shapes and sizes of features such as the eyebrows, eyes, and nose.

7 Investigating Robustness to Bias

We investigated several techniques to try to improve the robustness of Haar-based classifiers to biased training datasets.

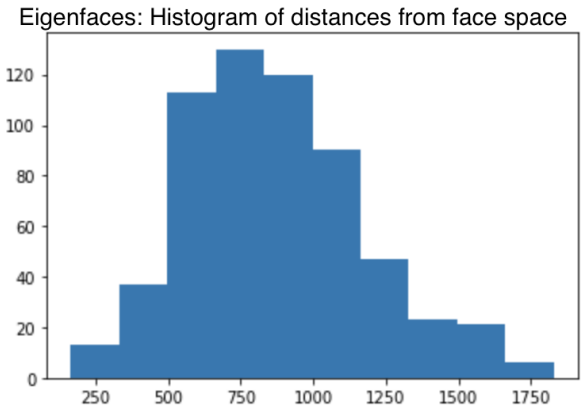

Figure 21: Distribution of distances from the face space among images in the test set (composed of an equal number of faces and non-faces balanced across race, gender, and age), computed as distances between faces and reconstructed faces using the projection of the face onto the face eigenbasis.

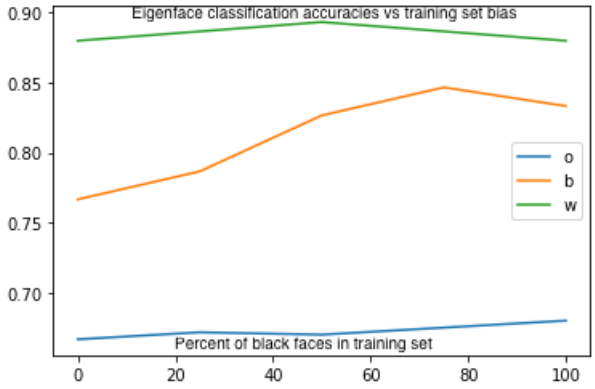

Figure 22: Accuracies on the test set using the eigenfaces for face classification. Training face data was used to obtain a face eigenbasis, and test images were then classified by determining whether the distance of the image vector from the face space was above or below our chosen threshold.

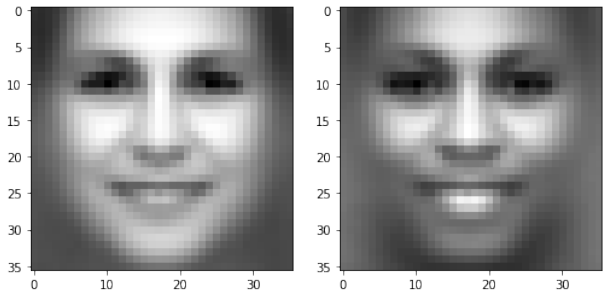

Figure 23: Mean eigenvectors for light-skinned and dark-skinned faces

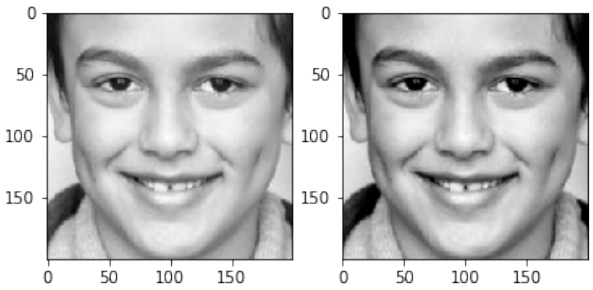

Figure 24: Image before and after normalization

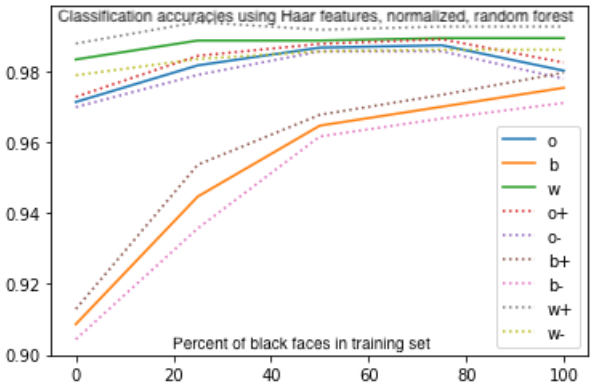

Figure 25: Accuracy of random forest Haar classifier, after normalization of input images.

7.1 Normalization

We found a significant difference in the mean light intensity between dark-skinned face images and light-skinned face images, as expected, with the mean light intensity of the dark-skinned faces averaging 137.76 and the mean light intensity of the light-skinned faces averaging 147, a significant difference on the light intensity scale from 0 to 255.

Due to the differences in light intensity we found between the dark-skinned faces and light-skinned faces, we decided to investigate whether normalizing the light intensity across all photos during training and testing would improve robustness of the models to bias. To normalize an image, we convert the image to a 32-bit integer representation and then use the following approximation to standardize all images to the same mean μ and standard deviation σ:

$$N(im) = \mu + \sigma * (im - \mu_{im})/\sigma_{im}$$

Figure 24 shows an example of a face before and after normalization [12].

We found that performance didn't improve significantly, implying that the differences in overall light intensity of the image alone could not explain the differences in performance between the biased data sets (Figure 25). However, it is important to note that our method normalized across both the face and the background of the image, so the resulting images still retained similar gradients and relative intensity values.

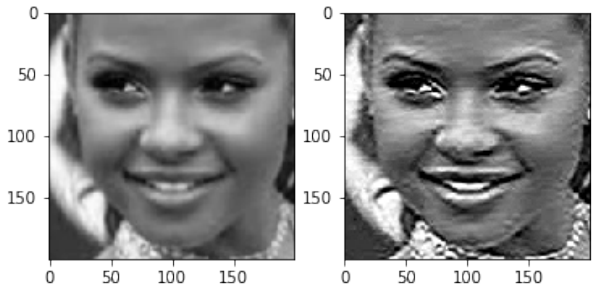

Figure 26: Image before and after sharpening

7.2 Sharpening

Since a Haar-like rectangle feature uses the difference between the pixel light intensities of a region and an adjacent region, it captures the contrast between regions. We also observed that Haar classifiers perform poorly on faces with lower contrast. Therefore, we hypothesize that the differences in performance on the biased data sets could also be attributed to the differences in intensity around important features. For example, there is a large intensity difference between the region inside an eye and the skin region around an eye, and this difference is typically smaller for darker-skinned faces.

To sharpen an image, we convolve the image with a sharpening filter, which sums to 1 and accentuates differences with the local average: $\begin{bmatrix} 0 & -1 & 0 \\ -1 & 5 & -1 \\ 0 & -1 & 0 \end{bmatrix}$

Figure 26 shows a face image before and after sharpening; we see that it emphasizes the features.

Figure 27 shows the performance of a random forest classifier using Haar features after sharpening all the images across faces, non-faces, training, and test sets. Although we still see sensitivity to race - the percentage of dark-skinned faces correctly identified as faces increases as the proportion of dark-skinned faces in the training set increases from 0 to 100% - we see that the rate at which the classifier improves is faster than that of the classifier on the non-sharpened images. For instance, increasing the ratio of dark-skinned faces from 0% to 25% brings the accuracy rate for dark-skinned faces from about 90% to about 96%. On the other hand, in the original classifier and the normalized classifiers, an increase from 0% to 50% ratio of dark-skinned faces was needed to attain the same improvement in performance. These findings support our hypothesis that the performance of a Haar-feature based classifier in predicting faces versus non-faces is linked to the contrast in images near defining features, such as the contrast between the darker light intensity of an eye with the typically lighter surrounding skin area. Moreover, the importance of such contrast in face classification using Haar features can explain the higher performance on light-skinned faces across all proportion ratios.

7.3 Gradients

Since the HOG classifier was more robust to the biased datasets, we trained a random forest classifier using Haar features on gradients of the images, to determine whether it would improve the robustness of the Haar features to racial bias by reducing the number of features characterizing the image to the most accentuated parts - the edges and outlines of key facial features and structure. We

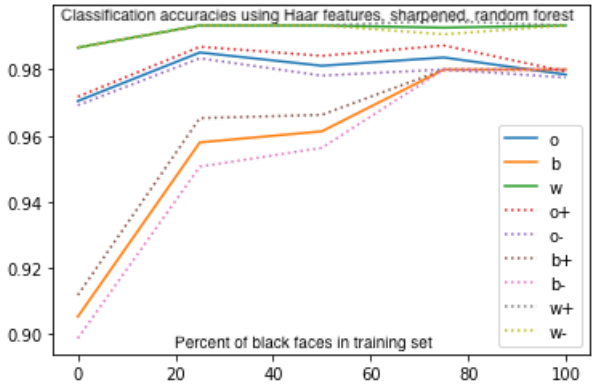

Figure 27: Accuracy of random forest Haar classifier, after sharpening input images.

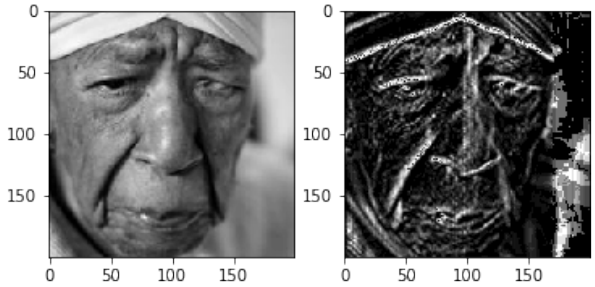

Figure 28: Image before and after computing the gradient

reasoned that such a classifier might be more robust to racial bias because the gradient of an image for two faces of different races does not depend directly on skin color or intensity, and so the edges that define a face should be comparable between different races.

To compute the filtered gradient images, we calculated the x- and y- gradient at each pixel of the image by convolving the image with the derivative of Gaussians - the Gaussian for smoothing from noise and the derivative to produce the gradient. Then the magnitude was taken as $\sqrt{F_x^2 + F_y^2}$, where F_x and F_y are the horizontal and vertical components of the gradient, respectively. A sample image with the filtered gradient applied can be seen in Figure 28.

We found, however, that such a classifier, while perhaps more blind to skin color, was also unable to distinguish between faces and non-faces. Figure 29 displays the low classification accuracies resulting from the Haar-like feature classifier trained and tested on the gradient images. Because both HOG descriptors and Haar-like features proved to be high-performing features for face classification, but the Haar-like features of the gradients of the images were not, we concluded that there is critical information that is lost in the process of computing a Haar-like feature on the gradient of an image. In particular, information about orientation of edges of a face, especially those that may be slanted or of a distinct shape such as a nose, can be lost in a Haar feature which only looks at differences between sums of intensities in adjacent rectangles and consequently cannot capture such orientations or distinct shapes. Therefore information about the skin tone and texture is important for Haar features but is lost after taking the gradient of an image. This analysis led us to next consider building a classifier using both HOG and Haar features.

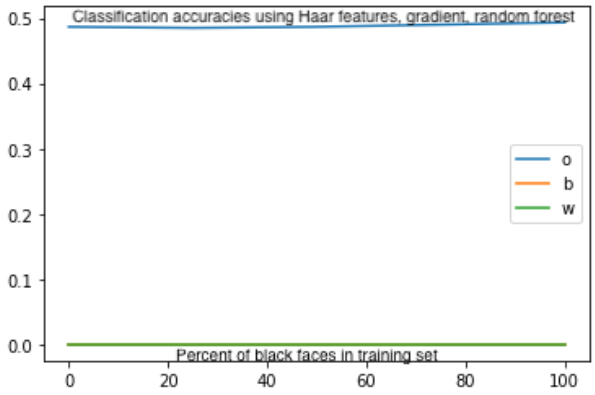

Figure 29: Accuracy of random forest Haar classifier trained on the gradients of input images.

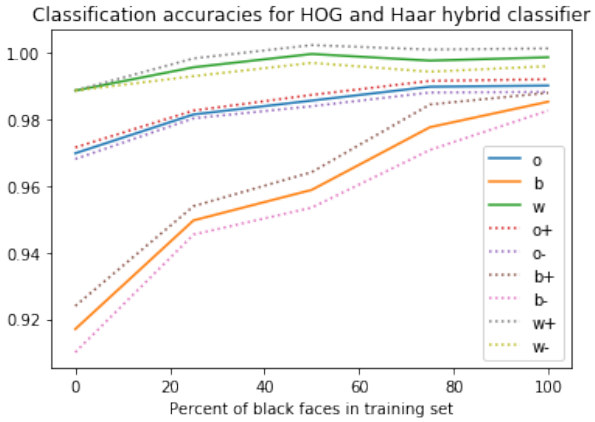

Figure 30: Accuracy of combined HOG and Haar classifier

7.4 Combining HOG and Haar Features

To explore another alternative, we drew from works such as "Hybrid method for hand gesture recognition based on combination of Haar-like and HOG features," that used both HOG feature descriptors and Haar features to build a model to benefit from the unique advantages of both [13]. Since we found that HOG descriptors were robust to race whereas Haar features were not, we hypothesized that combining the two types of features would produce a stronger model more robust to race than the classifier trained with Haar features alone. Figure 30 depicts the performance of this hybrid classifier on the datasets of varying degrees of bias. While this classifier outperformed the classifier with only Haar features in every dataset across every category it still reflected some sensitivity to the bias. We still saw the performance of the classifier on dark-skinned faces improve as we increased the number of dark-skinned faces, although by smaller increments and at higher baseline accuracies.

8 Conclusions and Future Work

In this project, we explored the sensitivity of different face classification methods to racial bias originating from imbalanced training data. To this end, we trained logistic and random forest classifiers on face and non-face data containing varying proportions of dark-skinned to light-skinned

faces using HOG feature descriptors and Haar-like features. Using these classifiers, we examined top features, considered false positive and false negative images from the test results, and assessed the overall accuracy along with accuracy on dark-skinned and light-skinned faces individually of the different classifiers. Additionally, we investigated properties of the feature types and the faces as vectors through PCA and the eigenface algorithm. Finally, we experimented with methods for improving robustness to bias for Haar descriptors through preprocessing the images before training.

Through our exploration and analysis, we found that HOG features were more robust to racial imbalance in the data, whereas Haar-like features demonstrated sensitivity to the racial bias, with performance on dark-skinned faces increasing monotonically as the proportion of dark-skinned training faces increased. Our exploration using principal component analysis and predicting race labels using the feature types also corroborated this conclusion. Further, by inspecting top features and properties of classes as images and vectors, we determined that the sensitivity of the Haar-like features to racial bias could not be explained by the light intensity of the skin tone alone, but that significant factors included contrast between dark and light regions, the shape, size, and location of distinctive facial characteristics such as eyes and nose, as well as higher variation within the dark-skinned faces in intra-class distance and distance from both the mean face and the mean dark-skinned face. Moreover, we found that this bias could be only slightly alleviated by sharpening the images to increase contrast, and that other preprocessing methods such as light intensity normalization had little effect on the robustness of the classifier to the racial bias.

We see a few key ways that future work could build upon and further our analysis. Since we concluded that face classification using Haar-like features was significantly sensitive to racial bias from imbalanced training data, largely due to facial features such as eye and nose shapes, sizes, positions, and orientations, and that as a result preprocessing methods such as sharpening and normalization could only slightly mitigate the bias if at all, to ensure fairness in such a classifier a better approach might be to enforce constraints on the training data set itself. Although ensuring a balanced data set in itself is a challenge, a Haar-based classifier we trained achieved high accuracy in predicting the race label of a face image, and such classifiers could be adapted and utilized to measure the racial diversity of a dataset. Further, methods such as generative adversarial nets, which work by pairing two neural networks against each other - one, the generator, to generate output, and the other, the discriminator, to distinguish between true and synthetic data - have proven successful at generating synthetic images based on input samples. In particular, DC-GAN (Deep convolutional generative adversarial networks) have been able to generate highly realistic human-like images [14]. Such images could be used for data augmentation to ensure diversity in data across not only race but also other dimensions such as gender.

Ultimately, our findings confirm the results of previous studies into bias in face classification and contribute a better understanding of the characteristics of faces and the face classification features and methods themselves that affect robustness to racial bias in imbalanced training data.

References

[1] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. *Proceedings of Machine Learning Research 81:1–15*, 2018.

[2] Vidya Muthukumar, Tejaswini Pedapati, Nalini Ratha, Prasanna Sattigeri, Chai-Wah Wu, Brian Kingsbury, Abhishek Kumar, Samuel Thomas, Aleksandra Mojsilovic, and Kush R. Varshney. Understanding unequal gender classification accuracy from face images. *arXiv:1812.00099v1*, 2018.

[3] Hee Jung Ryu, Hartwig Adam, and Margaret Mitchell. Inclusivefacenet: Improving face attribute detection with race and gender diversity. *arXiv:1712.00193v3*, 2018.

[4] Utkface large scale face dataset. Available from https://susanqq.github.io/UTKFace/.

[5] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2005.

[6] Rapid object detection using a boosted cascade of simple features. In *Accepted Conference on Computer Vision and Pattern Recognition*, 2001.

[7] Andy Liaw, Matthew Wiener, et al. Classification and regression by randomforest. *R news*, 2(3):18–22, 2002.

[8] J. R. Quinlan. Induction of decision trees. *Machine Learning*, 1(1):81–106, Mar 1986.

[9] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[10] Matthew Turk and Alex Pentland. Eigenfaces for recognition. *Journal of cognitive neuroscience*, 3(1):71–86, 1991.

[11] Raphael Couronné, Philipp Probst, and Anne-Laure Boulesteix. Random forest versus logistic regression: a large-scale benchmark experiment. *BMC Bioinformatics*, 19(1):270, 2018.

[12] Mahdi Rezaei and Reinhard Klette. Adaptive haar-like classifier for eye status detection under non-ideal lighting conditions. 11 2012.

[13] S. Ghafouri and H. Seyedarabi. Hybrid method for hand gesture recognition based on combination of haar-like and hog features. In *2013 21st Iranian Conference on Electrical Engineering (ICEE)*, pages 1–4, May 2013.

[14] Zhigang Li and Yupin Luo. Generate identity-preserving faces by generative adversarial networks. *arXiv preprint arXiv:1706.03227*, 2017.