1. Summary

Our project analyzes the accuracy of facial recognition software that estimates the ages of various groups of men and women (white, black, Asian, Indian, and other). In the past few years, life insurance companies have begun to adopt similar age-identifying technology, with the goal of quickly generating policy offers without requiring their clients to undergo medical examinations[1]. However, there are concerns about how accurate this software is and whether there are disparities in its accuracy for particular groups of people (especially groups that have been historically underrepresented by facial recognition training data, such as black women[2]). Our results show that our current model is, on average, approximately ±15 years off when predicting someone's age, and that there are some small disparities in the accuracy of our model when it is predicting the age of someone who is not Black, Asian, or Indian (white or "other"). There are even larger disparities in the model's accuracy between genders within the context of race, such as the difference between white women (±18 years) and Indian men (±12 years). Given that our training data contained over 11,000 images with a good representation of all races and genders, these results are not surprising but still concerning. Due to the inaccuracy of our model on such a large dataset and the noticeable disparities among race and gender, we strongly recommended that this software not be implemented for generating cursory life insurance policy offers. This report will highlight the background of the technology surrounding our project, our methodology, the findings from the results of our model, and our recommendations on how to pursue this technology going forward.

2. Introduction

Our project uses facial recognition software to estimate the age of a person. Only recently have life insurance companies adopted this technology as a way to quickly generate policy offers without requiring their clients to undergo medical examinations[2]. Purportedly, the technology these companies are piloting estimates lifespans using only a single image of a client's face. However, facial recognition software is already known to produce disproportionately inaccurate results for specific race/age groups. For example, based on a study by NYU's AI Now Institute, facial recognition inaccuracy is particularly common for black, female users[2]. This raises the general question of how accurate facial recognition software truly is at predicting age, and further, whether or not it should even be used for this purpose. Also, is facial recognition software equally accurate at predicting age between racial groups and genders? Our project explores these questions and ultimately provides quantitative results with interpretation.

It is important to highlight the fact that our question is not related to accuracy alone. If the model is generally inaccurate regardless of one's race or gender, then the software would undoubtedly affect people's access to fair life insurance policies. On the other hand, if this software is relatively accurate and disproportionately inaccurate for particular groups of people, this indicates a more sinister issue. While some race/gender/age groups may benefit from shorter wait times for an accurate life insurance policy offer without feeling the harmful effects of a facial recognition model's inaccuracy, white people (as shown by our model) would have a statistically higher chance of getting inaccurate results, causing frustration. Perhaps they might be more inclined to take the results and accept an expensive policy.

In order to investigate our questions, we first recreated age-identifying software similar to the software currently in testing by life insurance companies. Our first step was to collect a variety of images for our testing and training data[4]. From this collection of images, we separated our dataset between testing images (used to test the accuracy of the model) and training images (used to associate particular pixels from an image with a face's age, race, and gender). For our training data, we created a 4096x11854 matrix that displayed each image as a vector so that we could calculate mean-centered data and a covariance matrix (comparing correlation between pixels). Then, we used EVD (eigenvalue decomposition) to determine the ten most significant features (the largest eigenvectors) for predicting someone's age. These "eigenfaces" were projected onto our test data and showed how much of our test data matched with each of the most significant features. Using the K-Nearest Neighbor (KNN) algorithm, we then determined the approximate age of each person in our test data based on how the pixels in each test image correlated with the pixels in the images in the training data. The output of the KNN search is the index of the five closest images with which the test image shares the highest correlation.

3. *Methods*

Our model uses two major concepts to collect data. The first is Principle Component Analysis (PCA), a method for finding the eigenvalues of a correlation or covariance matrix with the purpose of finding the "directions" of variation in a dataset. Our data consists of "training faces" (people whose ages are known) and their personal attributes (age, race, and gender). While projecting the test data onto the principal components of the covariance matrix of our training data does result in loss of some data (lossy compression) and therefore does not capture all of the original values, it does capture the major ways in which the faces vary (hence the most important parts of the data for age recognition).

Our model first calculates the mean of each column of data (each picture) in our matrix and subtracts that mean from each column. This so-called "mean-centered" data matrix is then multiplied by the transpose of itself, thus creating a "covariance matrix".

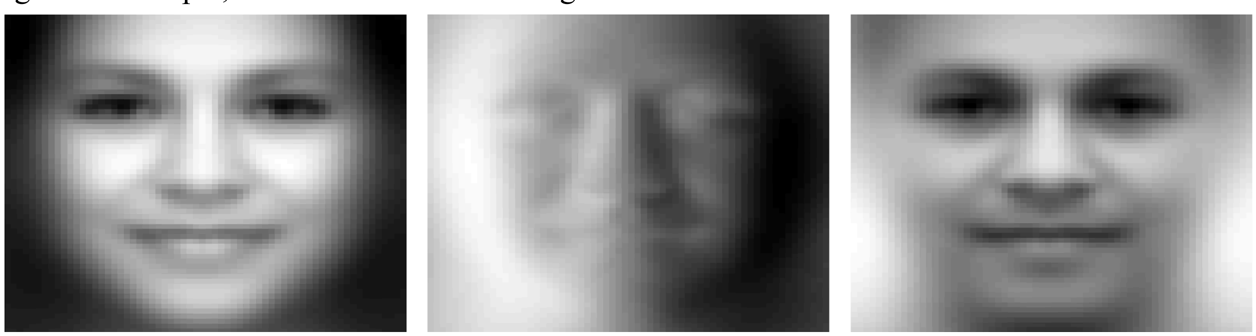
$$\frac{1}{(n-1)} * \sum_{i=1}^{n} (x_i - \mu_x)(y_i - \mu_y)$$

The equation used to calculate the covariance between each pixel for pixels x and y.

After calculating a covariance matrix, we then found its eigenvectors. Using the *eigs* function in MATLAB allows us to do EVD (finding eigenvectors) more quickly than it could be done by hand. The fundamental principle of EVD uses the equation below, where a matrix Q represents the amount of n eigenvectors for a given n-by-n square matrix (in this case our covariance matrix, shown as A). Λ represents a diagonal matrix where each diagonal element represents a corresponding eigenvalue for the vector (how much the eigenvector is scaled). Multiplying Q by Λ then Q^{-1} gives the directions of the eigenvectors of the matrix A.

$$A = Q\Lambda Q^{-1}$$

Using the *eigs* function, we extracted all the eigenvectors of A and chose the first 50 as the eigenvectors we wanted to project onto our test data. Projecting these eigenvectors onto our test data created "eigenfaces" that represent the most prominent visual features in determining one's age. For example, these are the first three eigenfaces we extracted:

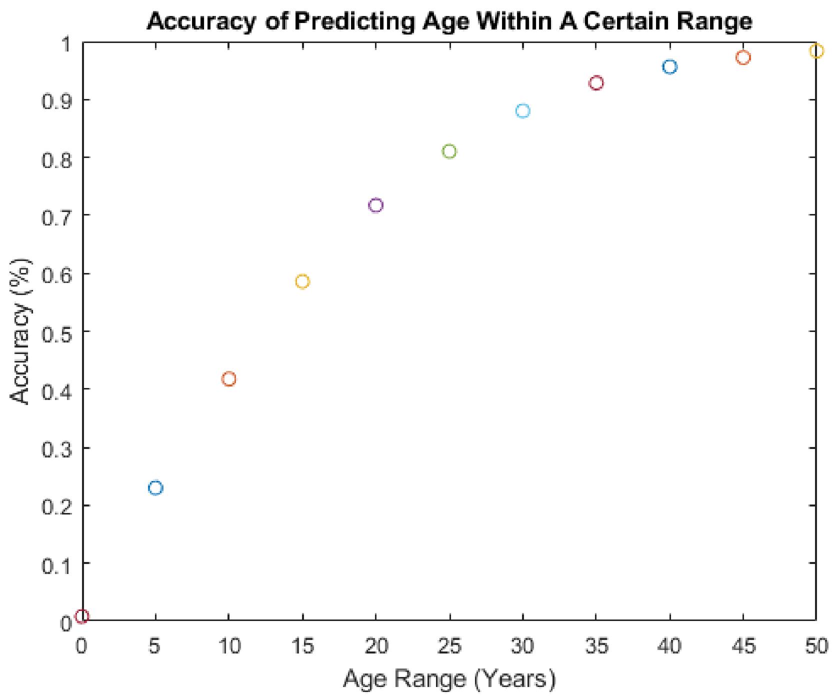


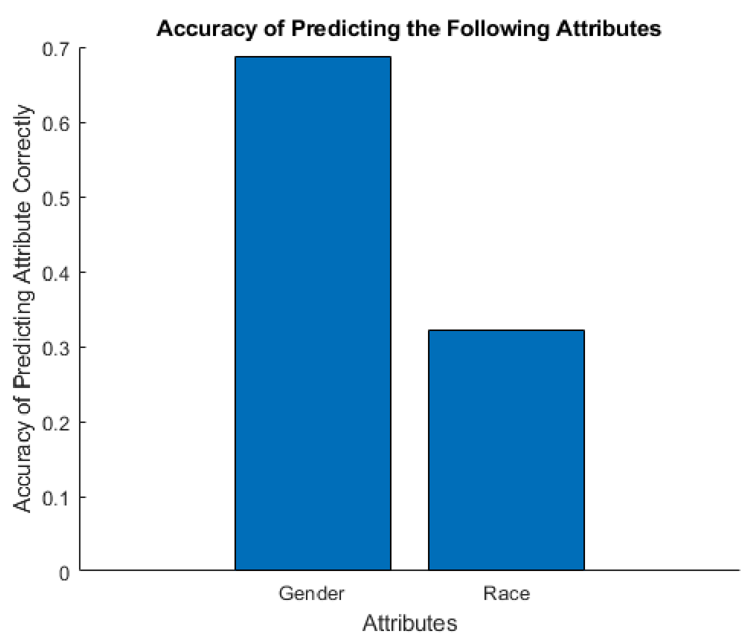
After extracting the eigenfaces, we used the K-Nearest Neighbor algorithm to determine the five original training faces that had the most correlation between each pixel with the test images. Then, we determined the approximate age of each person in our test data by basing it on the average value of each age of our matched training faces. This algorithm determines the five closest neighboring images using the Euclidean Distance formula (a way to calculate the distance between each point, representing an image, on a graph):

$$d(p, q) = \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2}$$

The distance formula for a KNN search.

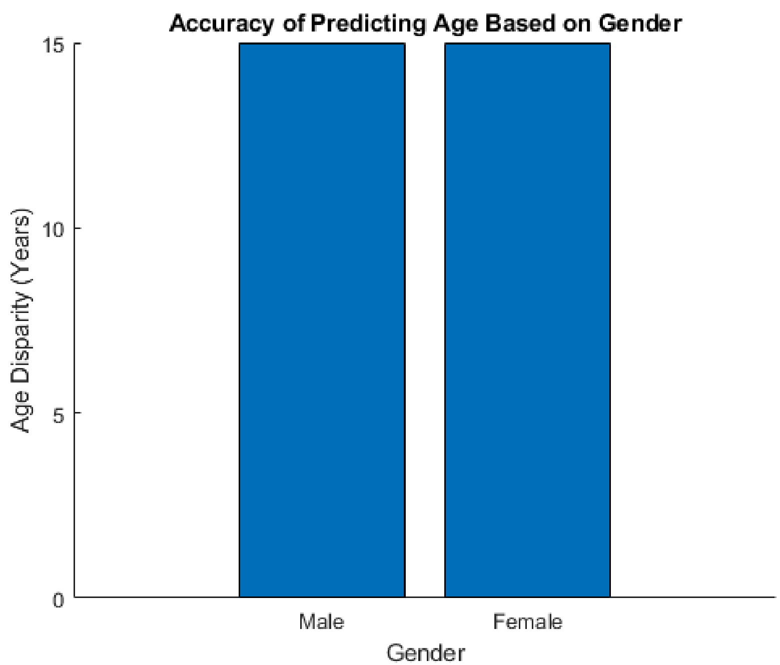
4. Detailed Findings

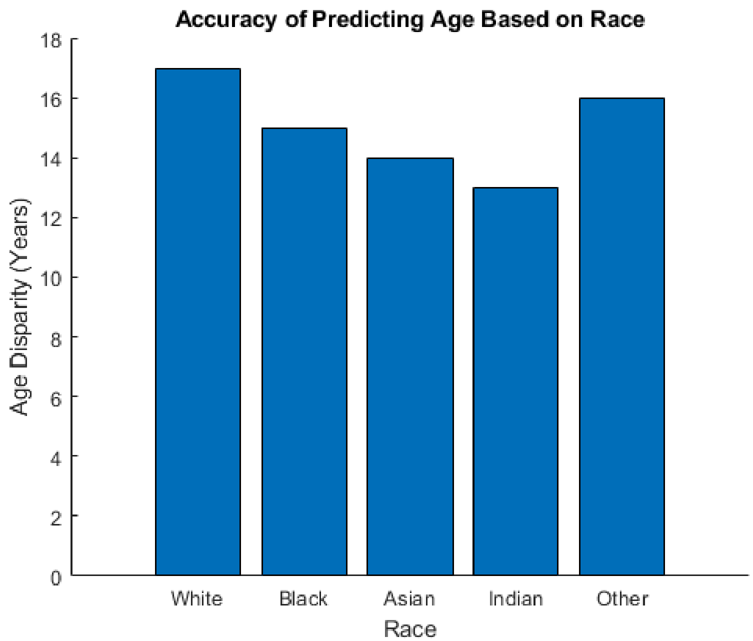


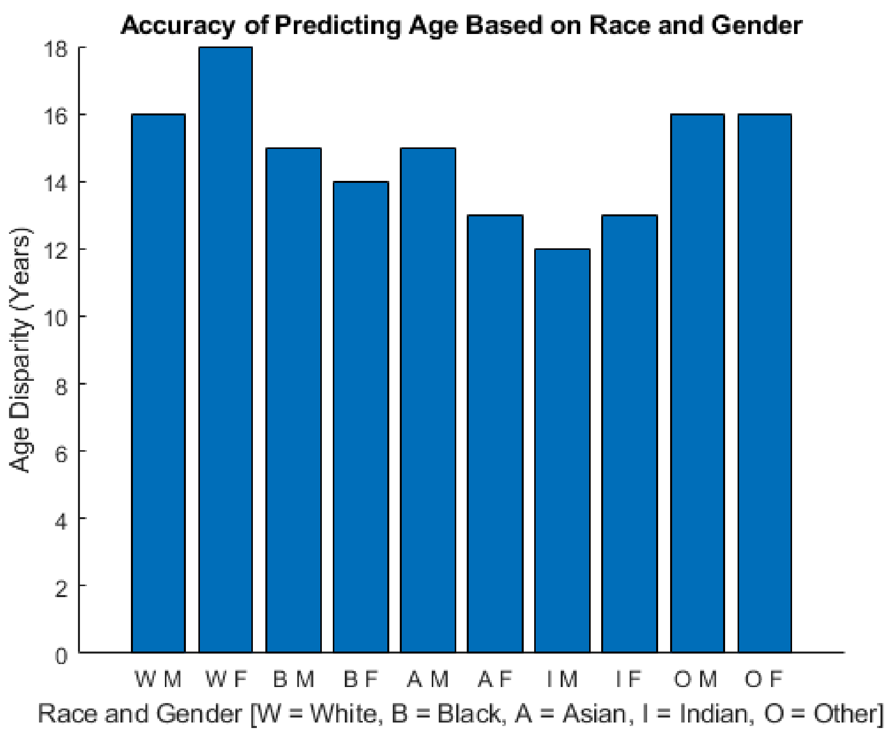


Ultimately, our model is generally accurate at predicting someone's age within 30 years (at ~89%). On average, the model's guessed age is ±15 years off from the person's actual age. In terms of predicting gender and race (additional features in our model), our model is significantly more accurate (68%) at predicting gender than race (32%). This is likely because there are fewer options when tagging gender (2) in comparison to race (5). Even though these percentages are theoretically better than randomly guessing, in our opinion it is still not as accurate as the model should be in order to be considered for commercial use.

With our current model, we also were able to find small disparities in the accuracy of our model when tested on particular racial groups (and genders within racial groups). There is no significant disparities in the accuracy of our model between genders alone. The following graphs below describe the disparities in accuracy of our model between gender, racial groups, and gender within the context of race:







We found that for race alone, white people or people classified as "other" under our test data had the highest inaccuracies in our model (the average age guessed was off by ±17 years). The model performed slightly better on Indian and Asian people, with the average age guessed being off by ±13 and ±14 years respectively. Gender within race also had its own disparities, with Indian males being easiest to recognize by age (their predicted age was off by ±12 years on average, in comparison to white women whose predicted age was off by ±18 years on average). Ultimately, the disparities in our algorithm were not as prominent as we thought they could have been. This is partially due to the dataset that we had, in which each race was significantly represented except for the "other" category (there were 15126, 6768, 5139, 5946, and 2523 images for these categories respectively). Even though we had a large amount of white subjects, however, the model still produced some significant disparities in accuracy (in addition to it being largely inaccurate).

One change we wish we could have implemented was having more accurate data in our model. Some images were not centered, while some images had completely different backgrounds and lighting altogether. This contributed to some significant issues when trying to find someone's race and age in particular. Also, our model only uses 50 eigenvectors to calculate these features. While this speeds up the computing process, it makes our model less accurate when identifying key facial features than using more eigenvectors would. Implementing this strategy would theoretically improve the accuracy of our model by capturing more of the variance in our data.

5. Recommendations

The disparity in accuracy among races are significant, but the larger issue is how inaccurate the model is in general. Given that our model is, on average, about ±15 years off from the actual ages of the test faces, this software seems unreasonable to implement on a commercial scale. It is worth investigating if there are disparities in the accuracy of our model among gender (and gender within the context of race) with better data, since already it seems like there are significant disparities in accuracy within our large dataset.

6. References

- [1] Marquand, Barbara. "How Your Selfie Could Affect Your Life Insurance." *USA Today*, Gannett Satellite Information Network, 25 Apr. 2017, www.usatoday.com/story/money/personalfinance/2017/04/25/how-your-selfie-could-affect-your-life-insurance/100716704/. Retrieved 2/27/2020.
- [2] Guynn, Jessica. "The Problem with AI? Study Says It's Too White and Male, Calls for More Women, Minorities." *USA Today*, Gannett Satellite Information Network, 17 Apr. 2019, www.usatoday.com/story/tech/2019/04/17/ai-too-white-male-more-women-minorities-needed-facial-recognition/3451932002/. Retrieved 2/27/2020.
- [3] Marquand, ibid.
- [4] Training/test data for this project was provided by The University of Tennessee Knoxville under a non-commercial research license. It can be freely obtained here: http://aicip.eecs.utk.edu/wiki/UTKFace. To download the exact data we used, click on the "[2]" next to "Aligend&Cropped [sic] Faces" under "Datasets". Then, right click on "UTKFace.tar.gz", click "Download" and extract the data with the program of your choosing. We successfully used WinRAR (https://www.rarlab.com/).