Honours Project Facial Recognition Bias

Ethan Leider

December 2021

1 Project and Motivation:

Facial Recognition, a form of biometric identification, permeates daily life. It is used to access computers, phones and even through security in some places. The technology intends to match a person's face from an image against a database of faces with the goal of giving access to information. It typically works by pinpointing and measuring facial features or landmarks. Facial recognition provides a simplified, convenient and sometimes more secure way to function on day-to-day popular applications. Unfortunately, current facial recognition programs have a flaw to them: they contain bias. These fundamental flaws are not news and the media has recently acknowledged and revealed that Twitter's facial recognition software was ageist, racist and ableist [5].

The reason I am creating this program is that people use and rely on facial recognition systems every day. Every industry has been focused on digitizing data and creating technology that optimizes that data. Even big company's facial recognition systems have been shown to contain bias which results in inaccurate data. Disturbingly current facial recognition technology accuracy is not consistent and involves significant bias specifically against minorities. These flawed algorithms influence decisions that dramatically impact and change the trajectory of lives. It is my goal to try and reduce the bias (improving accuracy) that comes through by identifying bias in the photos first. I want to minimize the unintended consequences of the application of facial recognition software.

Facial recognition has been a dream of many in the science fiction field. Authors created worlds where robots or machines leveraged their artificial intelligence (AI) to recognize the hero and give the character access to secure areas. Those who weren't on the pre-programmed allowed list of people would never get access to those areas. This is a simple description of facial recognition software's process.

The first huge step made in Machine learning was in the 1960s by Woody Bledsoe, Helen Chan Wolf and Charles Bisson who worked for a US military agency.[9] They created a device that would allow for the manual classification of photos by recording facial feature location data.[6] This would then be saved into a database and when a photo was entered it would retrieve the photo with the closest resemblance based on measuring the distances between landmarks on a face.[6, 16]

The next big step in facial recognition development was the change from manual computation to the automated computer model. This occurred in the late 1980s to the early 1990s. This change was called Eigenfaces and does not depend on three-dimensional information or detailed geometry. Eigenfaces decompose faces into a small set of characteristic feature images. Facial recognition happens by projecting a face (or image) into the subspace spanned by the eigenfaces and then classifying the face by comparing it.[20] Eigenfaces expanded the technology by incorporating the use of linear algebra in the process of facial recognition.[9]

Since then, facial recognition techniques and algorithms have evolved and there is widespread adoption for different purposes. The U.S. government-sponsored an initiative to encourage facial recognition algorithms, the FERET (Face Recognition Technology) program (1990-2000s) to be used in security, intelligence and law enforcement.[6, 17] During the 2002 Super Bowl, facial recognition was used to detect petty crime.[6] Understanding the functionality and limitations was critical to the development of the technology. Its application and implementation have since been popularized in law enforcement. By 2009, a law enforcement forensic database was created to let officers take pictures to cross-check against a database of criminals.[6] Expansion of facial recognition continues to improve the software and scale options for its incorporation into day-to-day life. Today, facial recognition can be found everywhere from getting into buildings, accessing phones and the social media we use.

A backlash to the progress is the common awareness of the built-in bias.[4] The technology often misidentifies people with non-white faces, which in turn affects communities of colour. The use of inaccurate data has sent innocent people to jail.

Before technology, people relied on a human form of facial recognition or perception of a person. This is how humans recognized and acknowledged the difference between people. Naturally, humans also would put people into groups depending on their preferences and unconscious/conscious biases. People also recognize faces of those that appear most like them easier. This means that despite the opportunity that facial recognition potentially bypasses human error, the data used to create the algorithms reflect racial, gender and other human biases.[7] This bias can appear as a result of not having enough training data in a specific race or sex.

In 2018, Gender Shades a landmark study that occurred on four gender classification algorithms.[15] The researchers found that females with darker skin performed the worst with error rates that were 34% higher than that of lighter-skinned males.[15] Even now corporations are having issues with facial recognition and bias. Twitter's facial recognition software has been revealed to have many issues over the years.[5] First, in 2020 it was revealed that Twitter's facial recognition software would prioritize lighter colour faces over darker colour faces.[12] Then this year it was later revealed that Twitter's facial recognition algorithm was also ageist, ableist and Islamophobic.[5]

Thankfully there is significant work being done by several large groups of people who are trying to unbias facial recognition programs. One example is a group at Beijing University of Posts and Telecommunications, Canon Information Technology (Beijing) Co., Ltd trying to reduce bias in facial recognition programs.[21] Another example is the DebFace Model mentioned in a paper called the "Jointly De-biasing Face Recognition and Demographic Attribute Estimation." It seeks to address the bias that comes from gender, race and age.[1]

Despite the controversy, investors have directed hundreds of millions of dollars into facial recognition start-ups. Crunchbase data found that there was a dramatic increase in venture funding: $500 million in July 2021 a rise from the $622 million for all of 2020.[22]

2 Background:

It is important to understand the impact of the bias that people have issues with. One of the concerns is the perception that white people seem to get preferred treatment by facial recognition detectors. The algorithm when presented with a picture with a white person and a person of another race program will first detect the white person's face. In life situations, if a computer program is relied on guiding the hiring or lending money, this could bias the results towards a white person providing them with more employment opportunities or capital.

Another fear is that the accuracy of the artificial intelligence is flawed and the wrong person is identified thereby mislabeling or even worse imposing a history of behaviour, such as criminal activity. These recognition inaccuracies threaten different minority communities.

Facial recognition bias is an ongoing issue. Bias relates to outcomes that are systematically less favourable to individuals within a particular group and where there is no relevant difference between groups that justifies such harms.[2] Facial recognition must be sensitive to those whom it is being used on and those who use it. There have been multiple proposed solutions to the problem. The one that was tested was using a machine-learning algorithm to unbias the database before the database is used in facial recognition software. The thought process is that if the database itself is unbias the resulting facial recognition would be less biased.

Before there can be a conversation about detecting bias one has to know how faces are detected. To a computer, faces are simply a collection of pixels on a screen. What facial recognition programs do is that they detect the face, identify specific features on the face and then generate face encodings of 128 values.[18] Using this encoding we can then measure the distance between faces to tell how similar faces are to each other. This process to detect faces is done by machine learning as it is the process in which we get the encodings that a program can detect. Doing facial recognition can also identify facial landmarks. These are points on the face of significance such as lips, eyes, ears, etc. These landmarks can be used for manipulation of the faces in the photo but give no direct help for facial distance. For this program, a face-recognition program was used so that I would not create a facial recognition machine learning algorithm as well.[8]

Finally, we have machine learning for grouping faces together. This program used the K-means clustering an unsupervised machine learning algorithm. The method clusters face features and its analysis process incorporates light, expression and facial scarring. K-means functions as an unsupervised machine learning algorithm whereby the data that is fed into the program has no labels and is grouped without the intervention of a human.[11] As opposed to supervised machine learning where a human provides the labels for the machine-learning algorithm.

K-means is intended to develop clusters by randomly assigning K points on the grid. Then each object figures out which of those K points is the closest. This object will then join the point that is the closest and be considered part of that K points cluster. This continues until the program has gone through all the objects. Then the middle point of that cluster becomes the new K point and this process is repeated until there is minimal movement of the middle point.

To improve the clustering results and help solve the issue of trying to visualize the encodings a program called Principal Component Analysis (PCA) is often applied. What this program does is that it reduce the number of dimensions or features to a lower amount. This is very useful when it came to encodings as they are given in 128 dimensions and PCA was very useful in visualizing by reducing that amount to 2 dimensions.

3 Thought Process:

There were a variety of facial databases to choose from as a basis for my research. I was looking for a facial database that contained a balance of races, sexes and ages. The UTKFace (In-the-wild Faces) photos featured only one person, containing 13835 usable photos and fulfilling the previous criteria.[23] The library of pictures had its limitation, there was a disproportionate less picture of people that identified as the following races: Hispanic, Latino, Middle Eastern and other races. However, the UTK database did have a robust representation of gender, age and the following races: White, Black, Asian, Indian and others. The other upside to using UTK faces the images were labelled according to age, gender and race (0 is White, 1 is Black, 2 is Asian, 3 is Indian, 4 is other) which would allow me to understand the groupings of my machine learning algorithm.[23]

Initially, I attempted to remove photos from a cluster that were too far from the middle of the cluster. The original attempt and expectation were to develop a way to possibly detect outliers to the cluster. During my effort, I realized that my assumptions were inaccurate and the detection of outliers only meant that the faces were more dissimilar from that point. The process was a critical step, as I was able to discount focusing on that strategy and recognize it as a warning that a group was too big. However, it still affected my thought process, my effort and my learning.

The thought process prompted me to rigorously evaluate and determine which machine learning algorithm to use. I was undecided between K-means and Hierarchical clustering both unsupervised machine learning algorithms. Each had its advantages and disadvantages.

K-means would allow me to cluster algorithms easily as it is simply the distance from the closest K point over and over.[14] However, the disadvantage to using this method was that I would have to figure out the amount of K which would absorb a significant amount of time and would potentially result in significant delays. As mentioned earlier, my decision was influenced by believing that objects that were too far away were separated and must be considered outliers. I knew that with K-means that could be easily achieved by creating an algorithm going into the cluster and removing them and putting them into an outlier group.

The other option was Hierarchical clustering.[19] This machine-learning algorithm works by figuring out the smallest distance between any two objects. These two objects then form a cluster. Then you find the next closest point and put them into the same group. This continues until the maximum distance is reached and that will be the groups. The advantage of this method was that at the end of the program there would be perfectly assigned groups at the end within a maximum distance. The disadvantage was that I would need to figure out the distance between every single object with each other and given the facial database ended up with 13835 photos it means that I would have to figure out 95696695 ($\frac{13835*13835-13835}{2}$) unique connections. In the end, I ended up choosing to go with K-means not because it fulfilled my desire to eliminate outliers but simply because it was a simpler method to manage and sort features into clusters.

The next step was determining the actual facial recognition process. My goal was to find the bias in facial recognition. It is important to acknowledge that facial recognition was a complicated machine learning algorithm. Rather than create an algorithm to do the facial recognition, it was valuable to choose an effective program that would allow me to find the bias within a database. The optimal facial recognition program used was called face-recognition.[8] It was chosen because I was working from the fundamental assumption that I relied on K-means clustering algorithm's face landmarks. Face-recognition was the solution that would most easily find the facial landmarks that I found. Later I realized the focus should not be facial landmarks but facial encodings and made the switch very quickly using the same program. The effective number of photos were 13835 as there were a few photos that were taken from UTKFace that would not go through. The next stage was to understand the information that my code revealed. Initially, I plotted the information on a chart to visualize the result. Unfortunately, that was challenging as the data was given in 128 dimensions. It is unfeasible to display an output of 128 dimensions. There were several efforts attempted to easily visualize my results.

First, I decided to try simply using two-dimensional slices to see the output result. That worked to understand a little but I quickly determined I wasn't receiving any information. This is because to figure out anything of value from the data I would need to slice 8128 ($\frac{8128*8128-8128}{2}$) different times to see the full picture of each slice. Even then it wouldn't tell me much as I would be constantly looking through the graphs not finding anything.

I ended up turning to PCA to reduce the dimensions of the encodings to something that could be visualized in 2 dimensions.[3] From there I simply used a scatterplot to get a better idea of my clusters.[10] It is not a perfect system as sometimes I cannot tell how the groupings are occurring but it is better than nothing and provides a visual idea of what I am looking at. This was the point I decided to avoid removing objects from a cluster and realized that with the proper amount of K value it would not be necessary to do.

After being able to visualize my data, I wanted to make sure to have some sort of understanding of the clusters that were made. As a result, I passed along the information containing the person in the picture's age, race and sex so that when I look at a group, I might be able to figure out the reasoning behind the cluster.

The last thing I had to figure out was the value of K. I first tried the Silhouette Score (a way of figuring out what the optimal K value was) and found that it recommended a K value of 3.[13] In the end, I determined that the best process required a guess and check method to figure out the appropriate K value.

4 Results:

The results were quite surprising. As I was using the guess and check method to find my K value, I set up a huge net of my possibilities. I decided that a K being at 4 (Figure 1) would be my low amount K being at 25 (Figure 2) would be my high amount and then I put K being 15 to be what I expected my middle to be. I then took those results and tried to find groups that would simply represent one specific group. An example is a white male in the age range of 21 to 25. While a group like those existed, several groups contain inputs that could be considered outside of those groups. I then tried K value of 50 saw the same thing, then 100.

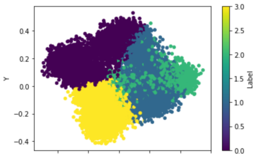
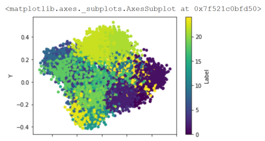


Figure 1: This is when K is equal to 4

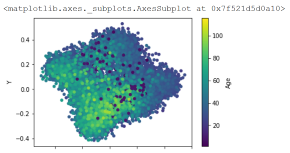


```
<matplotlib.axes._subplots.AxesSubplot at 0x7f521c0bfd50>
```

Figure 2: This is when K is equal to 25

Finally, even though I did not see the result I was expecting I reached a K value of 200 and realized I needed to stop. At this point, with 13835 photos and groups of 200 if there was an equal division of photos there would be around 69 different photos per group. As such, I needed to figure out what my code was saying.

The conclusion that can be drawn from the data is that the pattern of groups containing more than one of the following; ethnicity, sex and/or age range would continue. This can be seen when looking at the



```
<matplotlib.axes._subplots.AxesSubplot at 0x7f521d5d0a10>
```

Figure 3: Graph of Age Breakdown

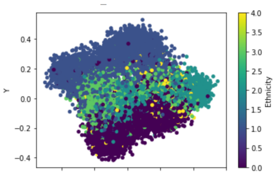
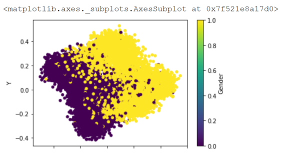


Figure 4: Graph of Ethnicity Breakdown



```
<matplotlib.axes._subplots.AxesSubplot at 0x7f521e8a17d0>
```

Figure 5: Graph of Gender Breakdown

age (Figure 3), ethnicity (Figure 4), and gender (Figure 5) it reveals that there is a lot of overlap in the areas that they fall into.

When applying the conclusion in the real world the results make sense. There are variations other than age, sex and ethnic group that influence of person's features. Here are a few: expressions (smiling versus non-smiling), body size, acne, scarring, illumination, and ageing. Other complications include a person's facial shape, geographic intermarriage and level of masculinity and femininity. Given my research, it is understandable why big companies may be having issues with bias. One possibility is that it is because they are using supervised machine learning. They may be training their machine learning algorithms using specific features like age, sex, race, etc. In this project, while there was knowledge of age, sex and race the machine learning algorithm was blind to any of that information. Meanwhile, these companies using supervised machine learning algorithms base their machine learning algorithms around age, sex, race, etc.

Corporations do this because they have access to this data/information. Data is perceived as powerful and influential. Facebook, LinkedIn Netflix and other social media platforms encourage people to subscribe to their service actively encouraging manual input and identification of a person's race, sex and other information. The user then posts several photos of themselves based on how they wish to be perceived in a variety of circumstances, expressions, angles and then tags (manually) their photo. It's a goldmine of accurate data. Facebook now has photos with a race, sex, age, gender, communities that the person identifies etc. and can use that information in their algorithm.

The challenge with the choice to rely on supervised learning may be a cause of bias. When using machine learning, the computer takes information provided by humans and uses that information as the basis of learning. As a result, supervised learning may look at the information about gender, race, ethnicity, etc. and decide to make groups around those features instead of just basing the algorithm on the faces themselves.

This would mean that the large amount of data that the companies get a hold of being age, gender, ethnicity, etc. may be a hindrance to the goal of unbiasing machine learning. That having access to all this information instead of being useful is instead the issue that already exists. Simply adding more information and pictures will not help and will instead propagate the issue.

References

[1] Rucha Apte. Eliminating bias in facial recognition, Dec 2020.

[2] Demand @ASME. Understanding bias in algorithmic design, Apr 2018.

[3] Lars Buitinck, Gilles Louppe, Mathieu Blondel, Fabian Pedregosa, Andreas Mueller, Olivier Grisel, Vlad Niculae, Peter Prettenhofer, Alexandre Gramfort, Jaques Grobler, Robert Layton, Jake VanderPlas, Arnaud Joly, Brian Holt, and Gaël Varoquaux. API design for machine learning software: experiences from the scikit-learn project. In *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, pages 108–122, 2013.

[4] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In Sorelle A. Friedler and Christo Wilson, editors, *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, pages 77–91. PMLR, 23–24 Feb 2018.

[5] Kevin Collier. Twitter's racist algorithm is also ageist, ableist and islamaphobic, researchers find, Aug 2021.

[6] History of face recognition & facial recognition software, May 2019.

[7] Nicholas Furl, P. Jonathon Phillips, and Alice J. OToole. Face recognition algorithms and the other-race effect: computational mechanisms for a developmental contact hypothesis, Feb 2010.

[8] Adam Geitgey. face-recognition, Mar 2017.

[9] Claude Hochreutiner. The history of facial recognition technologies: How image recognition got so advanced, Apr 2020.

[10] J. D. Hunter. Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, 9(3):90–95, 2007.

[11] Unsupervised learning.

[12] Khari Johnson. Twitter's photo crop algorithm favors white faces and women, May 2021.

[13] Khyati Mahendru. How to determine the optimal k for k-means?, Jun 2019.

[14] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *K-means*. Cambridge University Press, 2018.

[15] Alex Najibi. Racial discrimination in face recognition technology, Oct 2020.

[16] NEC. Facial recognition - science fact or science fiction - nec new zealand, Sep 2020.

[17] Patricia.flanagan@nist.gov. Face recognition technology (feret), Jul 2017.

[18] Vikas Solegaonkar. Introduction to face recognition, Jul 2021.

[19] Great Learning Team and Satish Rajendran. What is hierarchical clustering? an introduction to hierarchical clustering, Jul 2020.

[20] M.a. Turk and A.p. Pentland. Face recognition using eigenfaces. *Proceedings. 1991 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 1991.

[21] Mei Wang, Weihong Deng, Jiani Hu, Xunqiang Tao, and Yaohai Huang. Racial faces in the wild: Reducing racial bias by information maximization adaptation network. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

[22] Zack Whittaker. Despite controversies and bans, facial recognition startups are flush with vc cash, Jul 2021.

[23] Zhifei Zhang, Yang Song, and Hairong Qi. Utkface, 2017.