Evaluating the Impact of Bias in Training Datasets on Facial Recognition Software

**Abstract**

This research seeks to determine if biased training datasets might be the cause of the noted difference in accuracy with which facial recognition software identifies Black people when compared to White people. This is done by examining facial detection, the basis for facial recognition software, and training the same facial detection software on an unbiased dataset, a majority Black dataset, and a majority White dataset. This research found that further study on the topic is needed as, while the unbiased dataset did not return completely even results when identifying both Black and White people, the biased datasets did show results which included an improved accuracy with which they detected the faces of the race which made up the majority of the training dataset.

**Introduction**

Facial detection is the first step in facial recognition software. Facial recognition software is currently being used by the Federal Bureau of Investigation (FBI) and at least 25% of state and local police departments [1]. In approximately 26-30 states this facial recognition software is being used in conjunction with drivers license and ID photo databases putting at least 117 million people in their searches [1]. Most of the software being used is made by private companies. There is little oversight monitoring them, their use, or their efficacy [1].

Prior research into facial recognition software has found that there is a variation in its accuracy based on the race of the subject it is identifying. This research found that the accuracy with which facial recognition software identified White faces was higher than the accuracy with which it identified Black faces [2]. Other research has found that the training dataset for image recognition technology has an impact on the accuracy and that, when image recognition software is tested on a testing dataset which varies from its training dataset, the accuracy is reduced [3]. This could potentially be a factor in the disparity with which facial recognition software identifies White and Black faces as many popular and readily available datasets with which facial recognition systems are trained and tested are largely White, like the Face Recognition Grand Challenge dataset which is 70% White [4].

This research seeks to investigate this issue by examining the impact that datasets have on facial recognition software. If the training dataset is the source of the disparities with which facial recognition software identifies people of different races than a facial detection software trained on a racially balanced training dataset will show less variation in accuracy based on the race of subjects in testing images than a facial detection software trained on biased datasets.

**Methodology**

First three testing datasets were created. These each contained 100 images of faces of both Black and White people. The images used were of celebrities found online through google images and imdb who had publicly identified or talked about their race. Of these datasets one (referred to as Dataset A) was a racially balanced dataset with 50 of the images being of White faces and 50 of the images being of Black faces. The other datasets were each biased, one with 75 images of Black faces and 25 images of White faces (Dataset B) and the other with 25 images of Black faces and 75 images of White faces (Dataset C). These datasets can be found at the github repository: https://github.com/juliagracem/SeniorResearch.

The facial detection software was then created using Haar Cascades. The first step in doing was creating a cascade xml file which was done using the software Cascade Trainer GUI Version 3.3.1 (<http://amin-ahmadi.com/cascade-trainer-gui/>) was used to create a cascade trained on each dataset. To use this software positive images (the training dataset of faces) are saved as a folder p and a second gallery of images which contain no faces is created as a negative training set saved as a folder n in the same location. The Cascade Trainer GUI is then run. This requires the size of the images and their location to be uploaded to the software for a cascade xml file to be created. This can then be used in a Python script which calls the cascade to compare against an image to attempt to detect a face then draws a box around any detected face. The process of creating a cascade file was repeated for each of the datasets. These files can also be found at https://github.com/juliagracem/SeniorResearch.

Next testing dataset was created. This dataset included 30 images of faces, 15 White and 15 Black, as well as 15 images which did not contain faces. The photos without faces were similar to those found in the negative datasets; these were product photos, mannequins, clothes, and empty spaces found on google images. Three testing folders were then created, one for each version of the facial detection software created. They were labelled according to their training dataset and each contained the following: a copy of the same Python facial detection script, the training dataset, an empty folder for outputs and their unique cascade. These facial detection testing folders can be found at https://github.com/juliagracem/SeniorResearch. The Python Script was then called to test each image, the outputs (which were copies of the images where the facial detection script drew red rectangles around any detected faces) were saved and the accuracy (including true positive, true negative, false positive and false negative results) was manually recorded.

**Results**

The overall results for the facial detection software created using the unbiased training dataset can be seen in **Table 1**.



Predicted

|  |  |  |
| --- | --- | --- |
|  | Positive | Negative |
| Positive | 17 | 13 |
| Negative | 23 | 12 |

Actual

Table : A confusion matrix showing the results of running the facial detection software developed using Dataset A

The facial detection software trained on Dataset A correctly identified 7 of the 15 White faces in the testing dataset and 10 of the 15 Black faces in the testing dataset. There were 23 instances where it identified something that was not a face. This happened both in images without any faces and in images with faces where the software highlighted an area without the face instead of or in addition to the face in the image. The overall true positive rate for the software trained on Dataset A is 56.667%. The true positive rate for White faces is 46.667% and the true positive rate for Black faces is 66.667% .The overall precision is 42.5%.

The overall results for the facial detection software created using Dataset B (the 75% Black, 25% White dataset) can be seen in **Table 2**.

Predicted

|  |  |  |
| --- | --- | --- |
|  | Positive | Negative |
| Positive | 17 | 13 |
| Negative | 22 | 12 |

Actual

Table 2: A confusion matrix showing the results of running the facial detection software developed using Dataset B

The software trained on Dataset B correctly found 8 of the 15 White faces and 9 of the 15 Black faces. It had similar false positives to the software trained on Dataset A where the software highlighted areas without faces both in images with and in images without faces. The overall true positive rate for the software trained on Dataset B is also 56.667%. The true positive rate for White faces is 53.333% and the true positive rate for Black faces is 60%. The precision is 43.59%.

Table 3 shows the overall results for the facial detection software using trained on Dataset C (the 75% White 25% Black Dataset).

Predicted

|  |  |  |
| --- | --- | --- |
|  | Positive | Negative |
| Positive | 16 | 12 |
| Negative | 5 | 14 |

Actual

Table 2: A confusion matrix showing the results of running the facial detection software developed using the Dataset C

The software trained on Dataset C correctly found 9 of the 15 White faces and 7 out of 15 of the Black faces. It had an overall true positive rate of 53.333%. For White faces the true positive rate was 60%. For Black faces the true positive rate was 46.667%. The overall precision was 76.190%.

The true positive identifications for the software is shown below, broken down by race, is shown in **Table 4**.

Table 4: Showing the number of true positive identifications made by all three of the facial recognition softwares

Table 5 shows the false negative rate for each of the softwares broken down by race.

Table 5: Showing the false negative results of each of the softwares broken down by race and training dataset.

**Discussion**

The results of this research are mixed. The facial detection software trained on the unbiased Dataset A had skewed results where the software performed more accurately on Black Faces than White faces despite the balanced dataset. This would point toward a rejection of the hypothesis however the two softwares created using the biased datasets show a different story. The facial detection software trained on Dataset B, the majority Black dataset, had an increased true positive rate when detecting Black people rather than White people. Furthermore, the facial detection software trained on Dataset C, the majority White dataset, had an increased true positive rate when detecting White people rather than Black people. This supports the hypotheses. This means that, further research is needed to examine the impact of training dataset bias on facial detection software.

This study was limited by the availability of images. Initially this research was going to be done using images from existing datasets of faces. These were the Labelled Faces in the Wild Facial Recognition Dataset and the Large-scale CelebFaces Attributes (CelebA) Dataset. The two datasets were the largest readily available datasets of images of faces. Both datasets had many images of celebrities which was useful as, though they were not indicated in the dataset, the information about the race of the subjects was readily available online. However, there were not enough images of Black men to create the required datasets for this research. So, new datasets had to be created. Due to limited access the new datasets had to be comprised of images found online. This meant that control over variations in the photos such as lighting and posture was extremely limited (though both previously existing datasets which were considered had a variety of lighting and posture as well). Such variations have been found to have a negative impact on facial recognition [5]. Age could also have been a factor as previous research shows that facial recognition software is more accurate with people 30-70 compared to people 18-30 and, since the age of the subjects at the time of their photographing was not possible to determine for all photos so age could not be controlled for [2]. A third dataset was found but not used. This was the NIST Mugshot Identification Database (MID) which was not a viable source for images for this research as information about the race of the subjects was not available.

Additionally, the size of the datasets it was possible to build was limited. This was because of the time limits on this research as well as the time-consuming process of manually searching for individuals who fit the categories of the study and their photos. Thus, it would be beneficial to recreate this research with larger datasets of images taken in a more controlled environment in order to find more conclusive results.

**References**

[1]

“The Perpetual Line-Up,” *Perpetual Line Up*. [Online]. Available: <https://www.perpetuallineup.org/>. [Accessed: 28-Feb-2020].

[2]

B. F. Klare, M. J. Burge, J. C. Klontz, R. W. Vorder Bruegge, and A. K. Jain, “Face Recognition Performance: Role of Demographic Information,” *IEEE Trans.Inform.Forensic Secur.*, vol. 7, no. 6, pp. 1789–1801, Dec. 2012, doi: [10.1109/TIFS.2012.2214212](https://doi.org/10.1109/TIFS.2012.2214212).

[3]

A. Torralba and A. A. Efros, “Efros. Unbiased look at dataset bias,” in *In Proc. CVPR*, pp. 1521–1528.

[4]

P. J. Phillips, F. Jiang, A. Narvekar, J. Ayyad, and A. J. O’Toole, “An other-race effect for face recognition algorithms,” *ACM Trans. Appl. Percept.*, vol. 8, no. 2, pp. 14:1–14:11, Feb. 2011, doi: [10.1145/1870076.1870082](https://doi.org/10.1145/1870076.1870082).

[5]

A. Kortylewski, B. Egger, A. Schneider, T. Gerig, A. Morel-Forster, and T. Vetter, “Empirically Analyzing the Effect of Dataset Biases on Deep Face Recognition Systems,” in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Salt Lake City, UT, USA, 2018, pp. 2174–217409, doi: [10.1109/CVPRW.2018.00283](https://doi.org/10.1109/CVPRW.2018.00283).