

The Sensitivity of Facial Analysis Algorithms to Race and Gender

MOHAMMED ZEERAK



Motivation

- Face detection is used in many sectors and industries
- Bias introduces many unfair consequences to individuals
- Growing area of research

Aims

The aims of the project were to:

- Provide quantifiable results
- Evaluate a set of face detection algorithms which differ in structure
- Analyse results to understand if and why there exists a bias

Algorithms

The algorithms chosen for this project were:

- Viola-Jones
- Histogram of Oriented Gradients (HOG)
- Multi-Task Cascaded Convolutional Neural Networks (MTCNN)
- RetinaFace

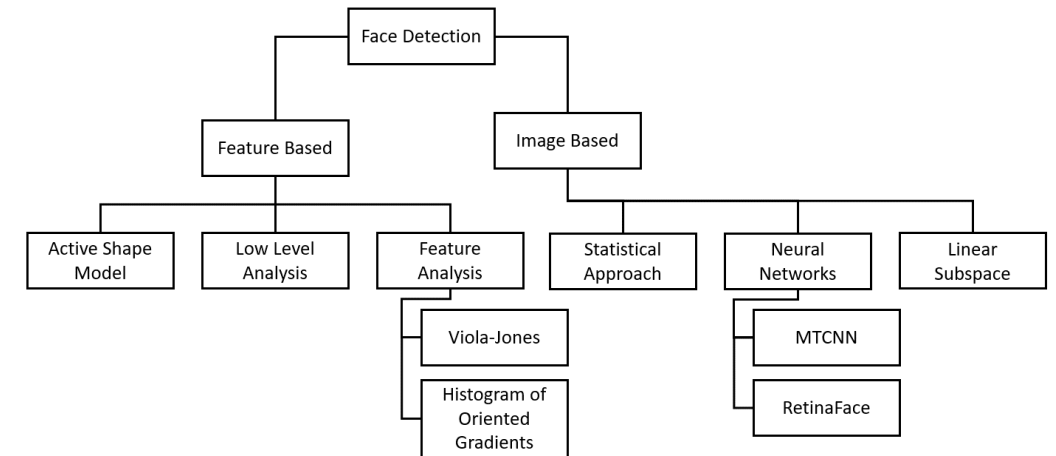


Figure 1: Algorithms categorised

5025 Dataset

- Motivation

- Compile a set of images that the algorithm could evaluate to see
assess if there was difference in accuracy between races and
genders

- Design choices

- | | |
|-------------------------|---------------------|
| ◦ Pre-existing or Novel | ◦ Technical Details |
| ◦ Dataset diversity | ◦ Annotations |
| ◦ Collation of Images | ◦ Limitations |

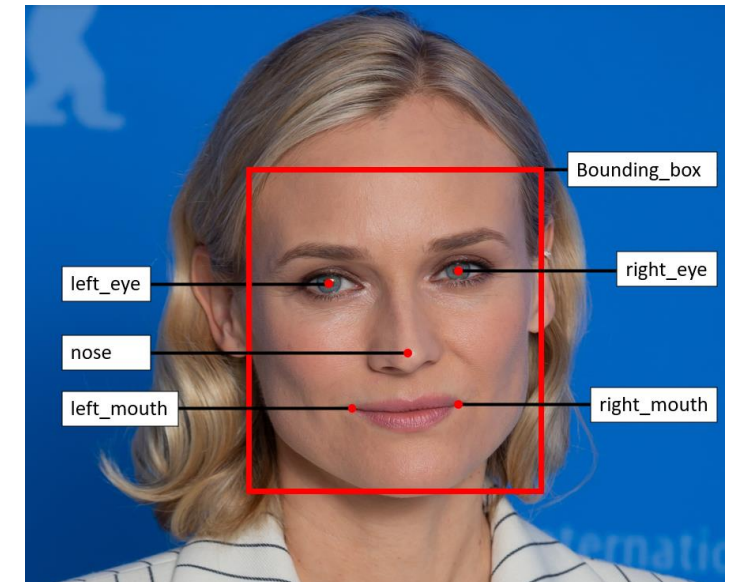


Figure 2: Bounding Box Annotations

Evaluation

- Evaluation Strategy
 - Execute algorithm against dataset of underrepresented and represented faces
 - Compare predicted coordinates to ground truth coordinates
 - Normalise to produce accuracy metric called “Error Difference”

Results

Frequency of Errors at Thresholds

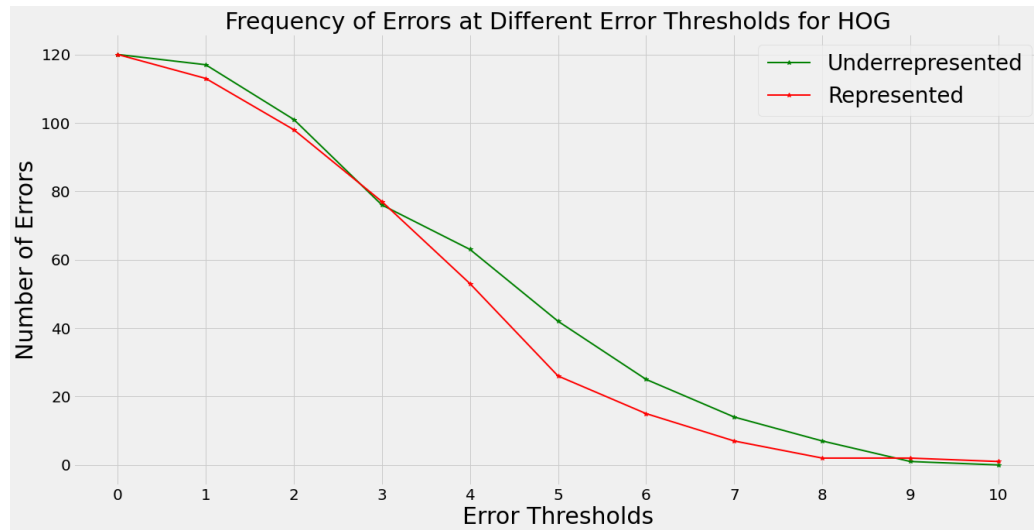


Figure 3: Frequency of Errors at Error Thresholds for HOG

- There is a similar performance between both groups up to the 3 error difference threshold.
- Underrepresented faces perform worst after 3 error threshold.
- Evidence of bias.

Frequency of Errors at Thresholds

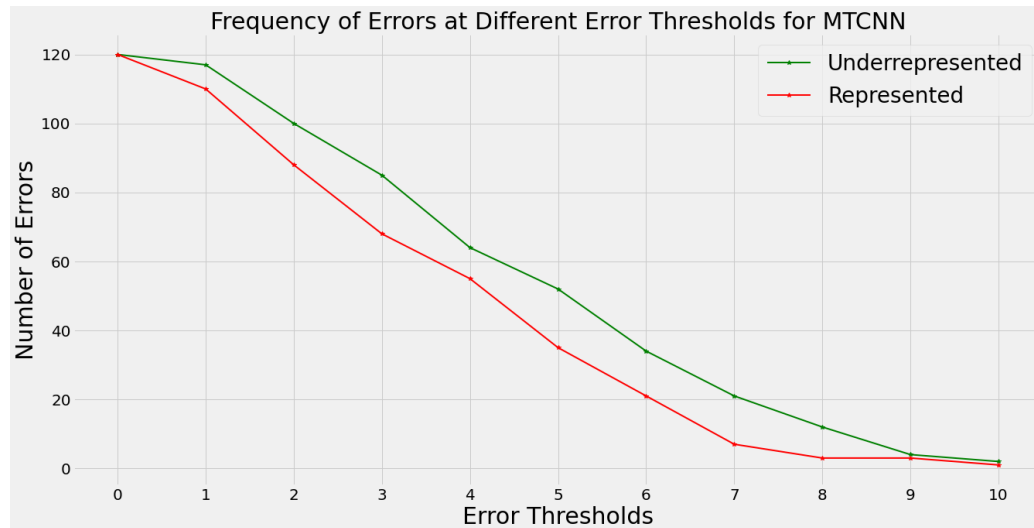


Figure 4: Frequency of Errors at Error Thresholds for MTCNN

- Underrepresented faces perform worst at every threshold.
- Represented faces have consistently lower errors.
- Difference of around 15 errors at 5 threshold
- Evidence of bias.

Frequency of Errors at Thresholds

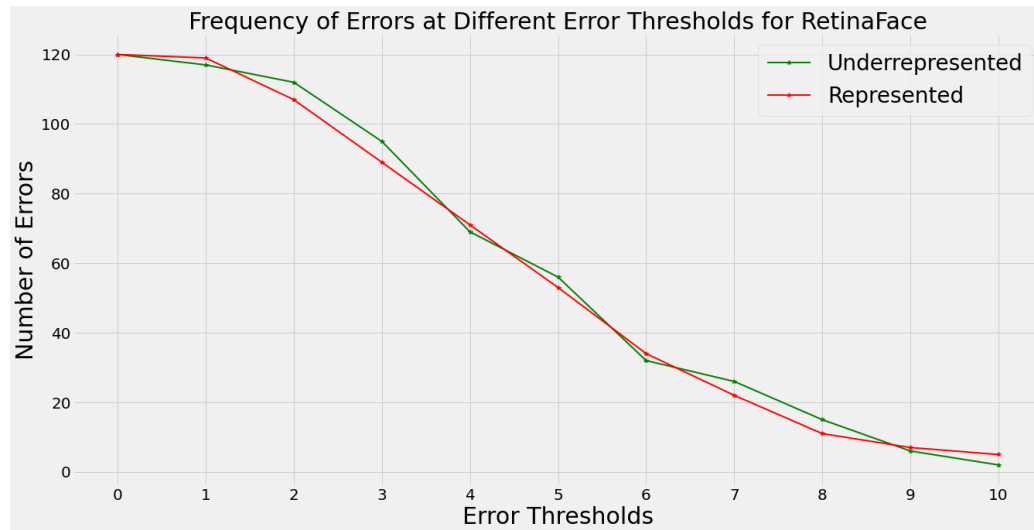


Figure 5: Frequency of Errors at Error Thresholds for RetinaFace

- Underrepresented faces and represented faces perform similarly.
- No outstanding difference present.
- Evidence of no bias.

Histogram of Error Difference for each Landmark

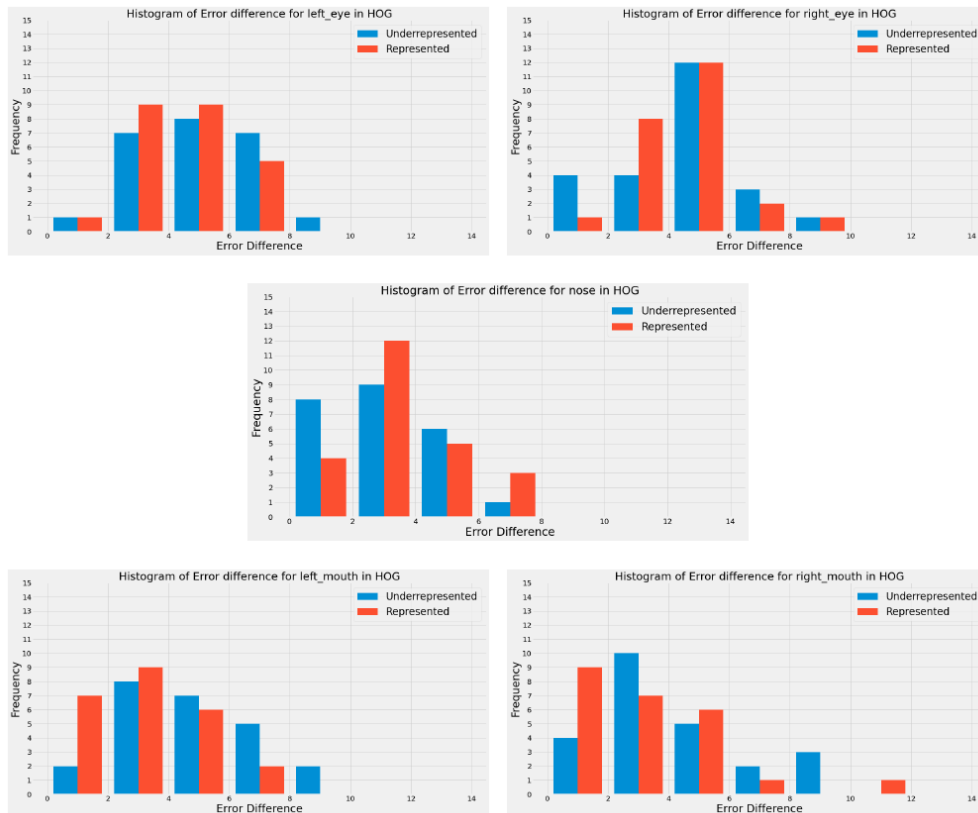


Figure 6: Histograms of Error Difference for HOG

- Left and right mouth show higher number of small errors for represented faces and a higher number of large errors for underrepresented faces.
- Represented faces perform slightly better for eye landmarks with more errors of a small magnitude.
- Underrepresented faces perform better for the nose landmark with more small errors.

Histogram of Error Difference for each Landmark

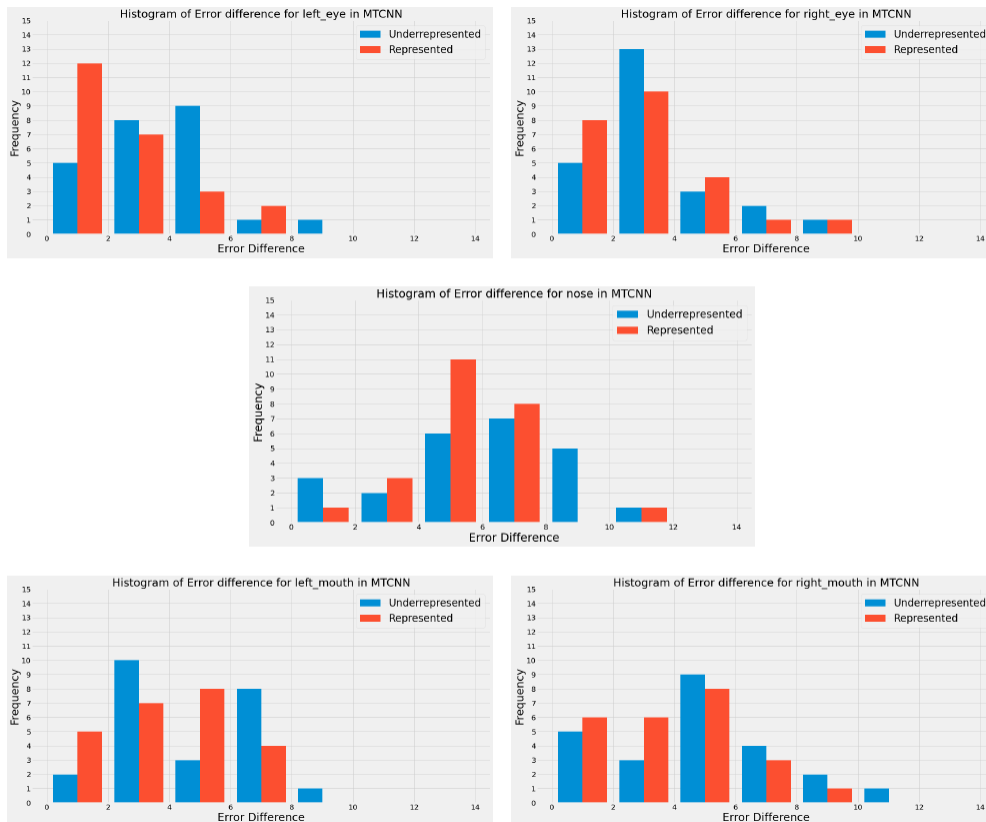


Figure 7: Histograms of Error Difference for MTCNN

- Each landmark has more errors of a larger magnitude for underrepresented than represented faces.
- Both represented and underrepresented perform poorly for the nose landmark.
- The mouth landmarks are where the underrepresented group performs the worst.

Histogram of Error Difference for each Landmark

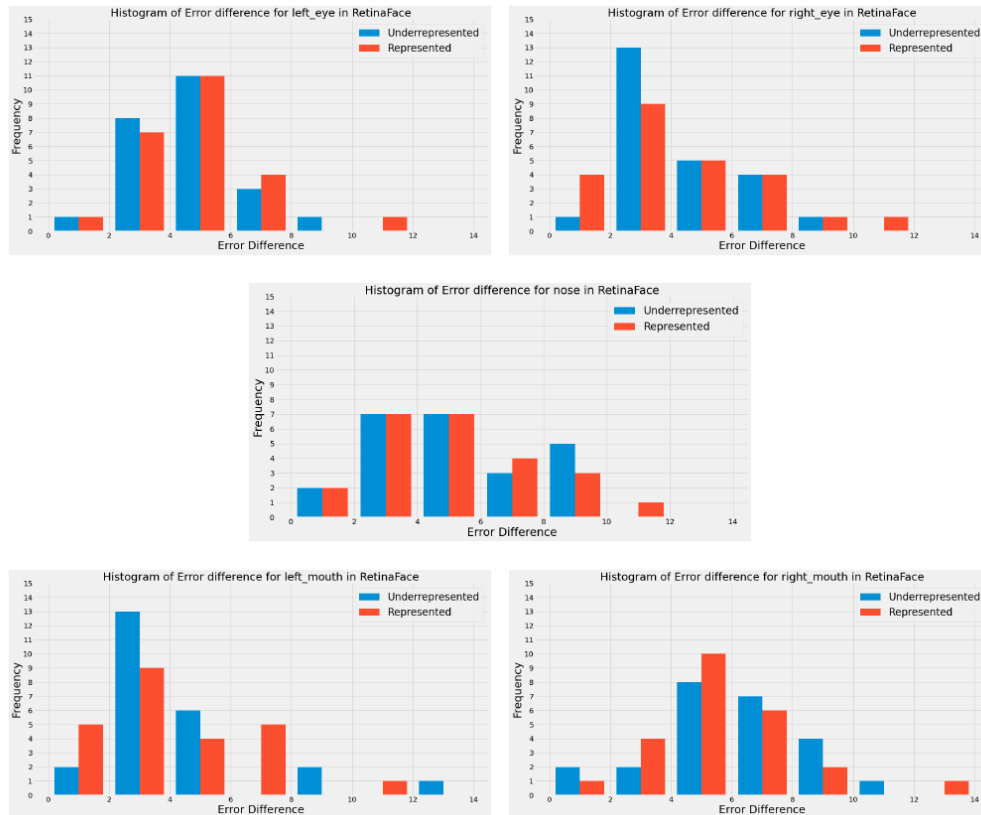


Figure 8: Histograms of Error Difference for RetinaFace

- Underrepresented faces and represented faces perform similarly across all landmarks.
- Largest difference between groups at left mouth landmark under the 2 threshold.
- No obvious difference between the groups in the landmarks.

Average Difference

DLIB	Underrepresented	Represented	Uncertainty (+/-)	Difference
left_eye_distance	5.048	4.411	0.246	0.637
right_eye_distance	4.282	4.434	0.131	-0.152
nose_distance	3.068	3.636	0.151	-0.568
left_mouth_distance	4.479	3.115	0.093	1.364
right_mouth_distance	4.19	3.186	0.436	1.004
total average	4.214	3.757	0.547	0.457

MTCNN	Underrepresented	Represented	Uncertainty (+/-)	Difference
left_eye_distance	3.778	2.574	0.196	1.204
right_eye_distance	3.606	3.045	0.132	0.561
nose_distance	5.935	5.472	0.392	0.463
left_mouth_distance	4.698	3.889	0.249	0.809
right_mouth_distance	5.053	3.716	0.42	1.337
total average	4.614	3.739	0.669	0.875

RetinaFace	Underrepresented	Represented	Uncertainty (+/-)	Difference
left_eye_distance	4.586	4.767	0.181	-0.181
right_eye_distance	4.19	4.422	0.311	-0.232
nose_distance	5.248	5.131	0.415	0.117
left_mouth_distance	4.164	4.194	0.319	-0.03
right_mouth_distance	6.148	5.705	0.344	0.443
total average	4.867	4.844	0.722	0.023

- HOG

- Mouth landmarks show largest difference between groups with underrepresented performing worst.
- Nose and right eye landmark perform slightly worst for represented faces.

- MTCNN

- All landmarks perform worst for underrepresented faces, with biggest difference at left eye, and mouth landmarks.

- RetinaFace:

- Represented group performs worst for left eye, right eye and left mouth landmarks.
- Underrepresented group performs worst for nose and right mouth landmark, with right mouth showing the largest difference between all landmarks.

Figure 8: Average Error Difference for Groups

Bounding Box Overlap Histogram

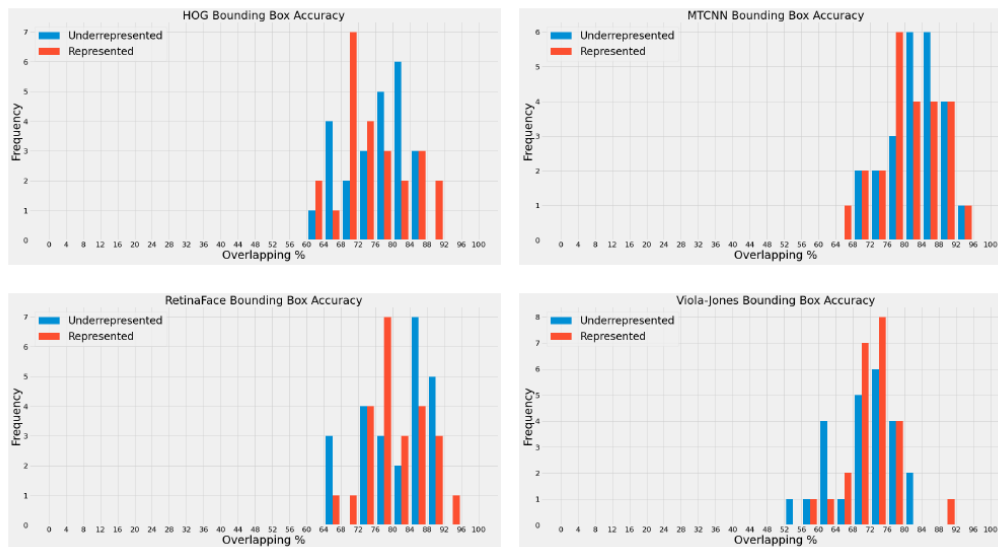


Figure 9: Bounding Box Accuracy Histogram for Algorithms

- HOG
 - Represented faces perform worst in general, however larger number of underrepresented faces at low accuracy extreme.
- MTCNN
 - Performance between both groups is very accurate with little difference.
- RetinaFace
 - Underrepresented faces perform better overall but the difference is small and the performance is average for both groups.
- Viola-Jones
 - Underrepresented group performs with a larger number of low accuracy overlap.
 - Larger number of high accuracy represented faces than underrepresented faces.

Discussion

Viola-Jones and HOG

Performance is worst for underrepresented faces than represented faces, indicating slight bias.

- Feature extraction
 - The difference in intensity between faces.
- Lack of underrepresented training data
 - Dataset wasn't fair.
 - 119 out of the 135 images for the HOG predictor were of represented faces.
 - Resulted in lack of training with underrepresented faces.

MTCNN

Show worst performance across all landmarks in threshold histograms for underrepresented faces.

- No handcrafted filters
- Dataset Bias
 - MTCNN uses WIDER FACE for face classification and CelebA for landmark localisation.
 - As evident from bounding boxes, face classification is strong, but landmark localisation isn't.
- CelebA dataset
 - Sampling Bias present which results in inaccurate predictions.

RetinaFace

Similar performance between groups, evidence of no apparent bias present

- Representative dataset
 - WIDER FACE categories.
- Feature Pyramid Network
 - Up-sampling and down-sampling of reconstructed layers could reduce the complexity that introduces bias.
 - Detector can make predictions for images it is better trained for.

Altering Images to Assess Dataset Bias

From the 3 algorithms the MTCNN showed the largest bias.

Goal was to alter poor performing underrepresented faces to be similar to represented faces, to see if algorithm could generate more accurate predictions.

- Altering Contrast
 - The faces chosen were based on worst performance for mouth landmarks (>7 error difference).
 - Contrast was altered as this reduced difference between the brightest and darkest parts of the image, resulting in a lighter skin tone and the shadows becoming less pronounced.



Figure 10: Original Image



Figure 11: Contrast Altered Image

Altering Images to Assess Dataset Bias

Running this back through algorithm generated interesting results:

- The 3 underrepresented male faces showed improvement on both landmarks .
- Out of the 6 underrepresented female faces:
 - 2 showed improvement on both landmarks.
 - 3 showed improvement on a single landmark.
 - 1 performed worst on both landmarks.
- The results conclusive for males were less so for females.












id (gender/group)	face	feature	original	altered	difference
1_m_m		left_mouth	7.92	2.384	5.536
		right_mouth	8.615	3.002	5.613
4_m_m		left_mouth	6.948	2.356	4.592
		right_mouth	8.075	3.077	4.998
6_m_m		left_mouth	7.48	3.214	4.266
		right_mouth	5.682	4.939	0.743
13_m_r		left_mouth	3.507	5.642	-2.135
		right_mouth	9.871	10.8	-0.929
20_m_r		left_mouth	7.417	6.915	0.502
		right_mouth	4.179	4.274	-0.095
27_f_m		left_mouth	6.436	3.196	3.24
		right_mouth	7.292	6.484	0.808
29_f_m		left_mouth	2.001	1.308	0.693
		right_mouth	11.784	11.306	0.478
30_f_m		left_mouth	8.909	8.147	0.762
		right_mouth	0.941	1.186	-0.245
32_f_m		left_mouth	2.409	3.808	-1.399
		right_mouth	7.529	8.607	1.078
33_f_m		left_mouth	5.515	5.81	-0.295
		right_mouth	7.898	7.217	0.681
34_f_m		left_mouth	7.931	16.856	-8.925
		right_mouth	5.789	14.123	-8.334

Figure 12: Altered Contrast Results

Dataset Bias

Knowing that dataset bias is a large factor to why algorithms perform in a biased way, I looked into how dataset bias was introduced and why it hasn't been mitigated.

- Sampling Bias

- Inherent bias in collating images of celebrities.
- Sampling bias is difficult to avoid because it can be introduced indirectly.

- Benchmark Dataset Bias

- Labelled Faces in the Wild.
- Doesn't evaluate race and gender bias.
- Algorithms continue to use biased datasets because of occlusive accuracy metrics.

Summary and Future Work

- This project evaluated 4 algorithms against a novel dataset consisting of represented and underrepresented faces, to investigate if a bias was present. It was found that differing levels of bias were present. Analysing the algorithms, made evident that the bias existed through the datasets the algorithms were trained on, and the design of the algorithms.
- Proposed future work
 - Expand on face detection field by evaluating other algorithms with differing features.
 - Evaluate the algorithms on a larger dataset comprising of an evenly distributed number of represented and underrepresented faces.
 - Retrain the MTCNN algorithm using a fairer dataset to evaluate if there is any improvement against reducing the bias present.

Bias Existence

The results indicated a bias existing within the MTCNN, and to a lesser extent the HOG and Viola-Jones algorithms. Looking at reasons for bias, its clear that it comes from two factors.

- The design of the algorithm
 - Feature descriptors.
- The training data used by the algorithms
 - Modern CNN methods rely entirely on annotated training data to create filter.