Review of *Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification*

Grace Yi Chen

This paper evaluated bias in automated facial analysis algorithms and datasets with respect to skin type and gender. Currently, the benchmark datasets IJB-A and Adience are overwhelmingly composed of lighter-skinned subjects and algorithms trained with these biased datasets resulted in discriminations. The authors introduced a new face dataset Pilot Parliaments Benchmark (PPB) which is more phenotypically balanced to achieve better intersectional representation based on gender and skin type. Accuracy, positive predictive values (PPV), true positive rate (TPR) and false positive rate (FPR) of 3 commercial gender classification algorithms are examined on 4 intersectional subgroups: dark or light skin type in male or female using the PPB dataset. All classifiers performed best for lighter individuals and males and worst for darker females. This shows there exists substantial disparities in the algorithm and we need to build facial analysis algorithms that are more fair, transparent, and accountable.

I think the authors showed strong evidence that there is bias in automated facial analysis algorithms and datasets with respect to skin type and gender. Ethics in AI technologies is now an increasingly important topic in machine learning with big data. Since face detection algorithms are incorporated in the surveillance and crime prevention in the US, false positives and unnecessary searches will lead to big social problems. One major cause of disparity is the biased training data of the machine learning algorithms. The training sample is very important in building machine learning models. As the authors mentioned, face recognition systems developed in Western nations and Asian nations perform better on their populations respectively. Here in this article, the authors demonstrated how algorithms trained by biased data could result in gender and race discriminations. Misclassification on minorities might not be prioritized since the current evaluation metrics of the machine learning algorithms emphasize the misclassification on majority group. Thus, by constructed a more balanced benchmark training dataset, we might be able to have more accountable machine learning algorithms.

Question:
Other than the biased training dataset, what else would lead to disparities in the face classification algorithm?

PPB is highly constrained as the poses in the face photos are similar. How much influence is the pose on the face classification algorithm?