# Gender Shades:
# Intersectional Accuracy Disparities in Commercial Gender Classification

王泓霏

中国传媒大学 媒介音视频教育部重点实验室

2022.8.24

# Abstract

- Machine learning algorithms can discriminate based on classes like race and gender.
- present an approach to evaluate bias present in automated facial analysis algorithms and datasets with respect to phenotypic subgroups.
- Introduce a new facial analysis dataset which is balanced by gender and skin type
- Evaluate 3 commercial gender classification systems using our dataset and show that darker-skinned females are the most misclassified group (with error rates of up to 34.7%). The maximum error rate for lighter-skinned males is 0.8%.
- Require urgent attention if commercial companies are to build genuinely fair, transparent and accountable facial analysis algorithms.

# Introduction

- Artificial Intelligence (AI) is rapidly infiltrating every aspect of society
- Even AI-based technologies that are not specifically trained to perform high-stakes tasks can be used in a pipeline that performs such tasks.
- Thus, an error in the output of a face recognition algorithm used as input for other tasks can have serious consequences.
- Recently it have been shown that algorithms trained with biased data have resulted in algorithmic discrimination. Word2Vec, encodes societal gender biases.
- Without a dataset that has labels for various skin characteristics such as color, thickness, and the amount of hair, one cannot measure the accuracy of such automated skin cancer detection systems for individuals with different skin types.

# Related Work

- Automated Facial Analysis
  - face detection
  - face classifification
  - face recognition
  - The lack of datasets that are labeled by ethnicity limits the generalizability of research exploring the impact of ethnicity on gender classifification accuracy
- Benchmarks
  - Any systematic error found in face detectors will in evitably affect the composition of the benchmark
  - Some datasets collected in this manner have already been documented to contain significant demographic bias

# Intersectional Benchmark

- Rationale for Phenotypic Labeling
- Existing Benchmark Selection Rationale
- Creation of Pilot Parliaments Benchmark
- Intersectional Labeling Methodology Skin Type Labels
    - Skin Type Labels
    - Gender Labels
    - Labeling Process
    - Creation of Pilot Parliaments Benchmark
- Fitzpatrick Skin Type Comparison

# Experiments

- Key Findings on Evaluated Classififiers
- Commercial Gender Classififier Selection: Microsoft, IBM, Face++
- Evaluation Methodology
- Audit Results
- Analysis of Results
- Accuracy Metrics
- Data Quality and Sensors

# Conclusion

- measured the accuracy of 3 commercial gender classifification algorithms on the new Pilot Parliaments Benchmark which is balanced by gender and skin type.
- annotated the dataset with the Fitzpatrick skin classifification system and tested gender classifification performance on 4 subgroups: darker females, darker males, lighter females and lighter males.
- found that all classififiers performed best for lighter individuals and males overall. The classififiers performed worst for darker females.
- the findings from this work concerning benchmark representation and intersectional auditing provide empirical support for increased demographic and phenotypic transparency and accountability in artifificial intelligence.

# Thanks for listening!

王泓霏

faywang@cuc.edu.cn