NEURAL INFORMATION PROCESSING SYSTEMS

Exploring Biases in Facial Expression Analysis using Synthetic Faces

Ritik Raina[1], Miguel Monares[1], Mingze Xu[1], Sarah Fabi[2], Xiaojing Xu[1], Lehan Li[1], William Sumerfield[1], Jin Gan[1], Virginia R. de Sa[1]

{rraina,mmonares,m6xu,xix068,l8li,wsumerfi,j6gan,desa}@ucsd.edu
sarah.fabi@uni-tuebingen.de

[1] UC San Diego

[2] EBERHARD KARLS UNIVERSITÄT TÜBINGEN

Facial expression recognition

Facial expression recognition (FER) has numerous practical applications in the field including:

- healthcare
- human-computer interaction
- student engagement
- consumer interest and happiness
- communication

However, FER models are often subject to racial biases.

Facial expression recognition

Facial expression recognition (FER) has numerous practical applications in the field including:

- healthcare
- human-computer interaction
- student engagement
- consumer interest and happiness
- communication

However, FER models are often subject to racial biases. **Why?**

1. Deep learning processes are complex and opaque.
2. High-quality facial expression dataset are difficult to obtain.

Motivation

(Fabi et al. 2022) used *artificially generated faces* to explore racial biases in pain-related facial expressions in a specific computer vision pain estimation model (Xu et al. 2020).

Results revealed:

- Different biases and gains in expression analysis for different skin colors and races.
- Biases and gains were not solely better for the faces of the majority race and skin color.

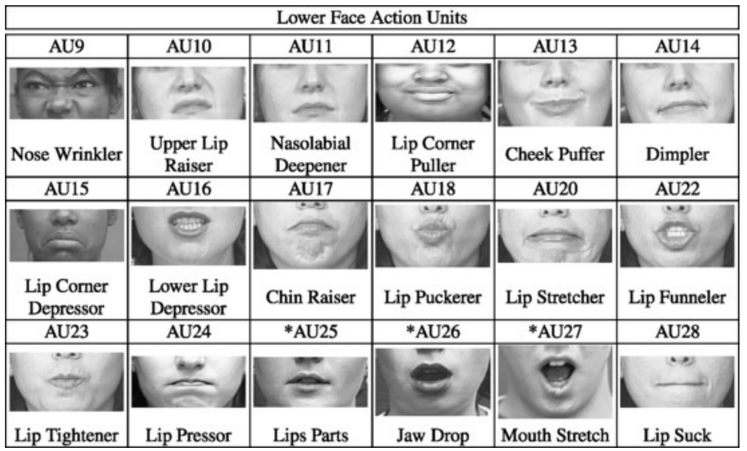
In our work, we artificially generate a facial image dataset as a means for exploring the racial biases in several publicly available computer vision facial expression models.

Synthetic facial image dataset

We used FaceGen Modeller to generate synthetic faces based on various manipulations to race, action unit activation levels, skin color.

Our dataset contains four sets of races: **African**, **African White** (African features with light skin color)**, European**, and **European Black** (European features with dark skin color).

Expressions of the synthetic faces were constructed via manipulations to the facial action unit (AU) activation levels (Ekman and Friesen, 1976). We manipulated the facial expressions with ten AUs individually to isolate how different manipulations affect the different models.



[1]

1) Pybeebee. "Breaking down the Facial Action Coding System (FACS)." Fresh Topics, 20 Oct. 2018, https://pybeebee.com/blogEN/2016/12/03/breaking-down-the-facial-action-coding-system-facs/index.html.

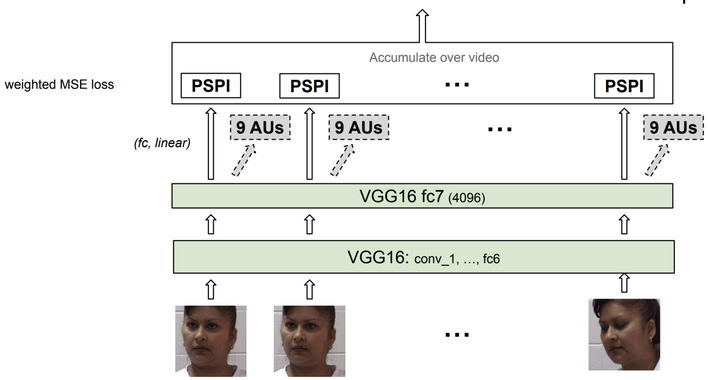
Public facial expression models

Extended MTL Model for Pain-Estimation (Xu et al. 2019)

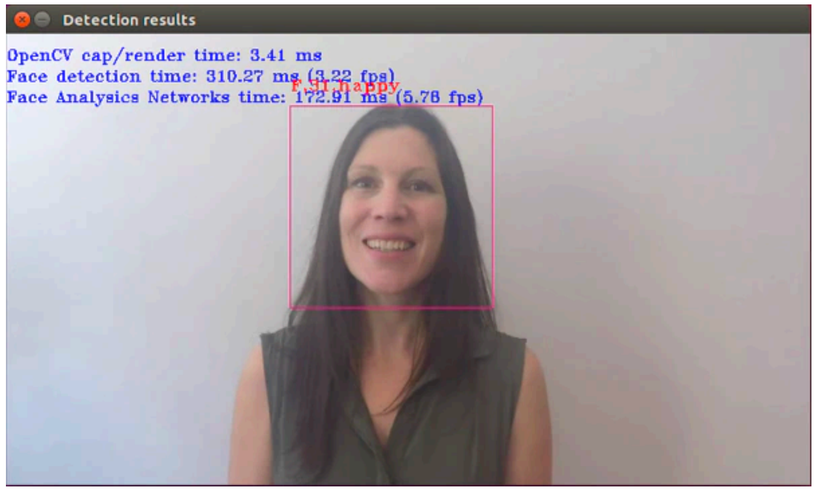
Intel OpenVINO Emotion Recognition

Multi-task EfficientNet-B2 Emotion Classification (Savchenko et al. 2022)

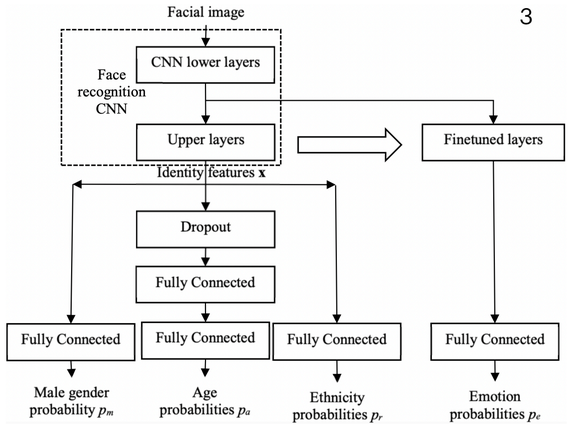
1



2



3



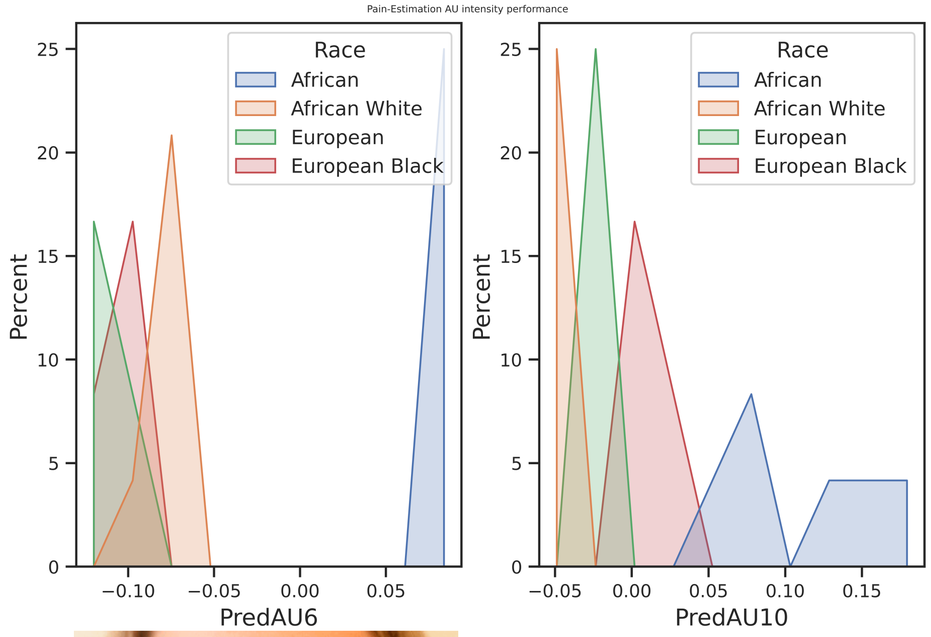
1) (Xu et al., 2019)
2) "Use the Deep Learning Recognition Models in the Intel® Distribution..." Intel, https://www.intel.com/content/www/us/en/developer/articles/technical/use-the-deep-learning-recognition-models-in-the-intel-distribution-of-openvino-toolkit.html.
3) (Savchenko et al., 2022)

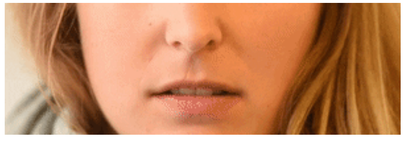
Racial biases in facial AU estimation

A bias in dark-skinned vs light-skinned facial AU activation estimation is prevalent.

The Pain-Estimation model produced higher AU6/AU10 prediction levels for African faces than African-White faces.

We did not observe that one skin color necessarily performed better or worse than the other.



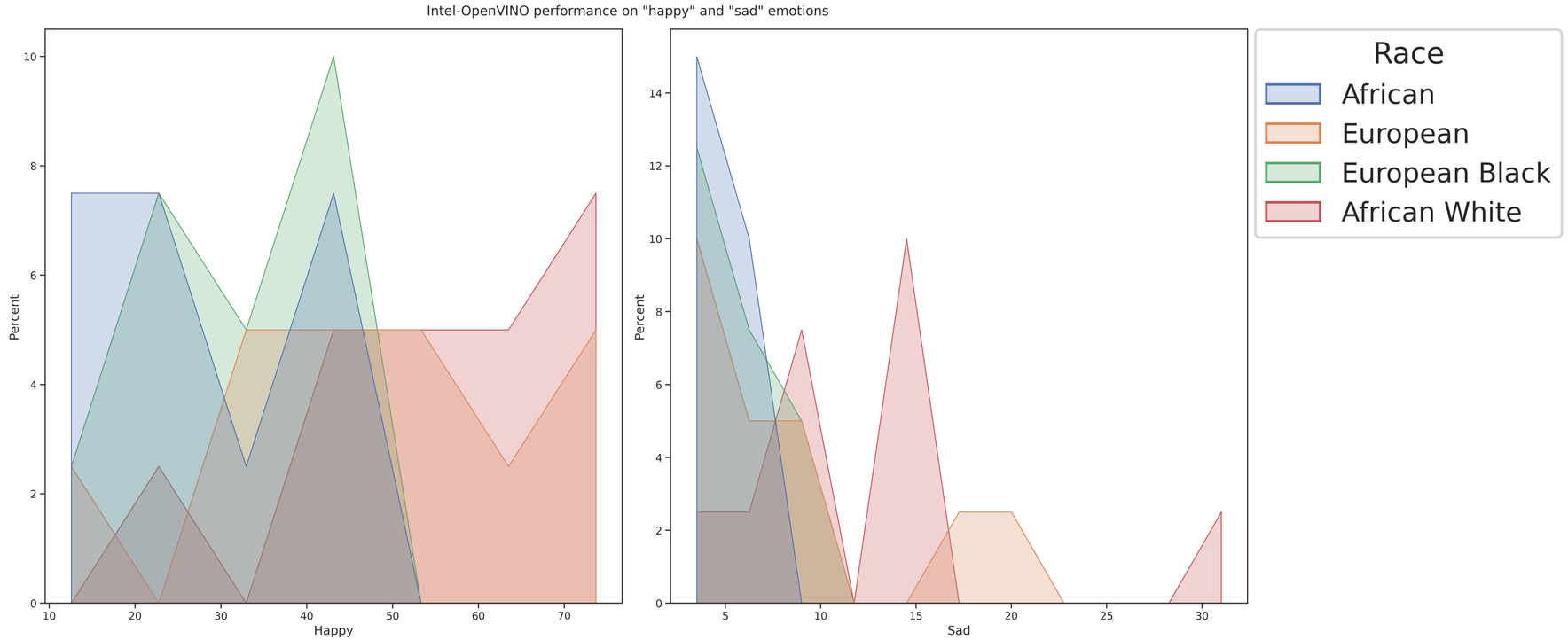


1) "Facial Action Coding System (FACS) - A Visual Guidebook." Imotions, 18 Oct. 2022, https://imotions.com/blog/facial-action-coding-system/.

Racial biases in facial emotion classification

Classification intensities showed polarizing biases between negative and positive emotions.

Similar color biases exhibited by the Intel OpenVINO model. African-White faces rated with higher “happy” and “sad” emotion intensities than African faces. However, no significant difference between the European/European-Black faces.

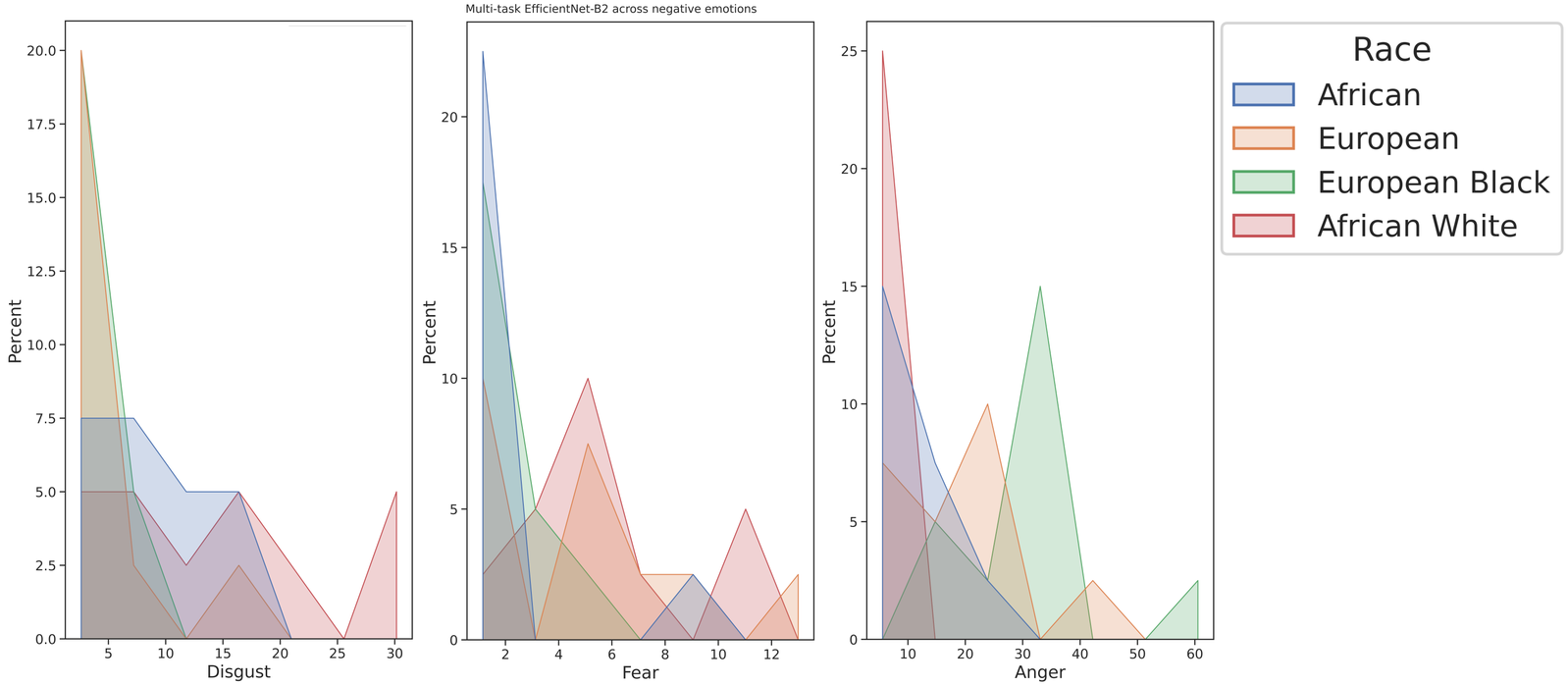


Racial biases in facial emotion classification

Classification intensities showed polarizing biases between negative and positive emotions.

Multi-task EfficientNet-B2 emotion classification model showed a greater bias surrounding negative emotion classification, than positive.

We observed that differing facial morphologies prompted differences in emotion intensity predictions, most importantly for the “disgust” emotion.



Conclusion and future directions

Synthetic face images were advantageous in our endeavor for exposing racial biases in FER models due to their **modular generation infrastructure**.

Model biases were not solely representative of dataset representations; racial imbalances were set due to the difficulties in understanding the **different appearances of facial features**.

In the future; we hope to leverage synthetic images to help **identify and mitigate** the root of racial biases in FER models.