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]
[1]University of California, San Diego [2]University of Tübingen
{rraina,mmonares,m6xu,xix068,l8li,wsumerfi,j6gan,desa}@ucsd.edu
sarah.fabi@uni-tuebingen.de

Abstract

Automated facial expression recognition is useful for many applications, but models are often subject to racial biases. These racial biases may be hard to reveal due to the complexity and opacity of the complex networks needed for state-of-the-art performance. Racial biases are also hard to demonstrate due to the inability to fully match facial expressions across real people. In this paper we use artificially created faces where facial expression can be carefully manipulated and matched across artificial faces with different skin colors and different facial shapes. We show that several public facial expression models appear to have racial biases. This work is an important step towards the eventual goal of understanding the basis of these biases and removing them from facial expression models.

1 Introduction

Automated facial expression recognition (FER) is useful for many applications including monitoring: student engagement and boredom, consumer interest and happiness, customer distress, and pain levels in patients who are unable to communicate either due to age, motor disabilities, or cognitive state.

It is important that recognition of facial expressions be fair and consistent in recognizing expressions in people of different races and skin colors. One challenge in achieving model fairness in deep learning FER models is the complexity of deep learning processes. Due to the black-box nature of these models, it is difficult to understand how the FER models digest and interpret facial features to score emotion. Therefore, researchers often struggle to identify the weaknesses of FER models that prompt disparate performance due to racial biases. Such biases may arise due to the under-representation of specific populations or due to difficulties in detecting relevant features that vary with different skin color or facial features.

Another challenge in achieving model fairness is the difficulty of obtaining facial expression datasets with sufficient quality, quantity, and diversity. The traditional means of obtaining data for facial expression datasets is expensive and cumbersome. Also, when using pictures or videos of real people making either natural or posed facial expressions, it is not possible to exactly control all aspects to ensure, for example, that a White face is making exactly the same expression (and thus should be scored in the same way) as a Black face.

Recently, Fabi et al. [6] used artificially generated faces to explore racial bias in pain-related facial expressions in a specific computer vision pain-estimation model [16]. They revealed that, while the network was able to respond largely monotonically and fairly linearly to increases in activation of facial AUs, it was subject to different biases, and gains for different skin colors and races. Importantly, the biases and gains were not solely better for the faces of the majority race and skin color.

In this work, we generate and use new artificially generated faces made by the FaceGen Modeller as a means to explore racial biases in several publicly available computer vision facial expression

NeurIPS 2022 Workshop on Synthetic Data for Empowering ML Research.

models. The advantages of synthetic face images are that we can control facial parameters (such as race and skin color), while holding modeled facial muscle activation constant, to isolate how different manipulations affect the different models. This work serves as an elucidation of the problem and some of the dependencies which will enable future work towards removing these biases from facial expression models. This work may also help inform any sensory components to human bias in FER when observing people of color.

2 Methods

In this section, we describe our synthetic face dataset and explore what results facial expression models produce with controlled manipulations to skin color, facial morphology, and facial muscle activation.

2.1 FaceGen Synthetic Image Dataset

Our work uses artificially generated facial images made through the FaceGen Modeller software. FaceGen Modeller is a programming tool that can generate synthetic faces based on different manipulations you provide such as race, facial expression, coloring, gender, etc. Our dataset consists of four sets of "races": "European", "European features with dark skin color" or "European Black", "African", and "African features with light skin color" or "African White". We used the FaceGen Modeller to generate random male faces using the FaceGen "African" and "European" racial settings. (Male faces were used as they appeared to be more realistic than the female FaceGen faces). For the original "African" and "European" faces, we did not change any parameters under the category "modify". For the other two sets, the skin shade of the faces were set between 2.5 (lighten) to -3.0 (darken). For the set of "European Black", we took each European face and set the skin shade parameter towards -3 adjusting the brightness until the face looked most natural. Similarly, for the set of "African White", we took each African face and set the skin shade parameter toward 2.5 until it looked as natural as possible. "African White" and "European Black" are thus two artificial categories we created to observe how coloring alone (with no change in face shape) affected the models. For the remainder of the paper, we refer to the four categories as "races".

The expressions of the synthetic faces were constructed via manipulating the facial action unit (AU) activation levels. The Facial Action Coding System (FACS) [5] codified the AUs as anatomically-based components of facial muscle movements underlying perceivable facial movements. After modifying the races and colors, we manipulated the facial expression with ten AUs (4: Brow Lowerer, 6: Cheek Raiser, 7: Lid Tightener, 9: Nose Wrinkler, 10: Upper Lip Raiser, 12: Lip Corner Puller, 20: Lip stretcher, 25: Lips part, 26: Jaw Drop, 43: Eyes Closed), each of which was activated to six levels (0%, 20%, 40%, 60%, 80%, 100%) separately (except for AU43: Eyes Closed which was only set to 0% and 100%). We chose to manipulate these AUs individually in order to isolate the effects that would be made on each of our models. This step was applied to the four racial sets in the exact same way, leading to 224 images in total. Fig.1 shows rows of synthetic faces set with the same expressions, except having differing facial morphologies and skin color.

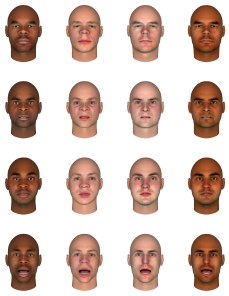

Figure 1: Synthetic face images used for experimental evaluations. Each row shows the same expression for each of African, African White, European, and European Black settings respectively. (First row) This row involves the activation of AU7-Lid Tightener (100% activation) with rest of the AU activations to be set at 0%. (Second row) In this row AU10-Upper Lip Raiser is activated to 100% for all face settings. (Third row) In this row AU25-Lips part is activated to 100% for all face settings. (Fourth row) In this row AU26-Jaw Drop is activated to 100% for all face settings.

2.2 Extended MTL Model for Pain-Estimation

The first model we investigate, follows on the work of [6] to investigate the first-stage of the extended multi-task learning (MTL) pain estimation neural network of Xu et al. [16], referred as the Pain-Estimation model from here onward. This model was trained and achieved state-of-the-art accuracy on the UNBC-McMaster Shoulder Pain dataset [9]. The UNBC-McMaster dataset is publicly available and comprises of 200 face videos from 25 patients with shoulder pain as they performed movements of their shoulder resulting in differing levels of pain. From these videos, each frame was labelled with 11 facial action unit (AU) intensities, and a corresponding PSPI value, and 66 AAM landmarks. The PSPI score [12] is a pain metric computed from a distinct set of AU intensities that are pain-related. The first stage of this three-staged model is a VGGFace network [11] pre-trained on classifying 2622 faces of celebrity icons, most of which are racially white. This network was then additionally trained on the UNBC-McMaster Shoulder Pain dataset to detect pain-related AUs (4, 6, 7, 10, 12, 20, 25, 26, 43) and the PSPI intensity of a facial image. The extended MTL model had two other stages that took the output of stage one and predicted whole-video segment pain scores, but our experiments only utilize the first stage of this model. This model is publicly available.

2.3 Multi-task EfficientNet-B2 Emotion Classification

The second model that we evaluate, Multi-task EfficientNet-B2, follows the work of [13] on the emotion and engagement classification in an online learning setting. This state-of-the-art model follows a two-stage training structure of pre-training on face detection and fine-tuning on emotion classification. The face detection stage of this model is a pre-trained VGGFace network and uses tracking & clustering techniques to extract facial sequences. Then, the model is fine-tuned on the largest publicly available facial emotion dataset, AffectNet [10] - a collection of 287,651 images annotated with eight basic emotions (anger, contempt, disgust, fear, happiness, neutral, sadness, and surprise). The training set of AffectNet is known to be biased, as 64.4% of the faces are White (of 7 total races) [3]. We included this model in our evaluation because it achieved the greatest performance for facial emotion recognition on AffectNet, obtaining the highest known accuracy on both 8-emotion and 7-emotion classification. This model is publicly available.

2.4 Intel OpenVINO Emotion Recognition

Intel's Open Visual Inference and Neural Network Optimization (OpenVINO) library is a series of toolkits that provides effective deep learning neural networks on various tasks such as object, detection, segmentation, and estimation [1]. Their emotion recognition model performs on top of their underlying face detection model which produces coordinates over the bounding box surrounding the detected face. The underlying network features a MobileNet backbone [7] that utilizes depthwise separable convolutions as a measure for running neural networks efficiently on mobile devices. The emotion recognition has been trained and evaluated with only five emotions ('Neutral', 'Happy', 'Sad', 'Surprise', 'Anger') from the AffectNet dataset. The architecture of the emotion recognition model is that of a fully connected neural network that outputs a softmax probability across the five emotions detected. This model is publicly available.

3 Experiments

3.1 Experiment 1: Racial biases in facial action unit estimation.

Synthetic image transformations for AU estimation evaluation. In this experiment, we explored the task of facial action unit estimation on the synthetic face dataset we generated. For evaluating on the Pain-Estimation model [16], we utilize the MTCNN (multitask CNN) [18] face detector to detect the faces off the synthetic images. This aspect was troublesome for the dark-skinned faces as it would sometimes not detect them at all, or mistake the facial bounding box coordinates; unlike the correct detections made on the light-skinned faces. In order to mitigate this issue, we cropped the dark-skinned faces with the complementing light-skinned face bounding box coordinates. The bounding box surrounding the detected face is extended by a factor of 0.3 before cropping. The input images are reshaped to 256 256, center-cropped to 224 224 and their color channels were adjusted to that of the VGGFace network.