– Independent Work Paper Spring 2020 –

Evaluating the Accuracy of Automated Facial Emotion Detection on Intersectional Inputs

Dorothy Zhao
Adviser: Professor Olga Russakovsky

Abstract

The use of artificial intelligence (AI) for decision-making like choosing who to hire or determining recidivism rates is becoming more commonplace. However, ubiquity does not mean that AI is a perfected technology. In particular, this study will focus on one controversial application of AI— facial recognition. While there have been studies showing the intersectional discrepancies of facial recognition, current work has not yet extended to auditing emotional detection, a subarea of study within facial recognition. To fill this gap, we will be modifying the phenotypical features of faces from the AffectNet dataset in this study. Using this augmented dataset, we will then audit the performance of commercial systems in determining the emotions of these faces. In particular, we are looking to see if there are intersectional discrepancies between our four key target groups (lighter-skinned males, lighter-skinned females, darker-skinned males, and darker-skinned females). The aim of this project is to see whether there are any differences, and, if so, what biases these commercial companies might be upholding through their technologies.

1. Introduction

Over a decade ago Desi Cryer posted a video on YouTube demonstrating how a Hewlett-Packard face-tracking webcam did not register Black people while readily identifying white individuals [8]. In 2015, Google Photos came under fire for labelling Black individuals as "gorillas" in photos [8]. Recently, the University of California Los Angeles dropped their plan for using facial recognition technologies to monitor the students after concerns raised by the campus community [18]. These are only a few motivating examples to illustrate

facial recognition technology's already controversial history. As artificial intelligence (AI) technologies become more ubiquitous, we see it being applied to a wider variety of services– from making hiring decisions in HR to determining recidivism rates in law enforcement [17] [16]. However, AI technologies are not objective. As seen in the examples given, they often encode and replicate existing biases that people have.

Facial emotion recognition (FER), a subcategory of facial recognition, falls under the broader branch of affective computing– a field of artificial intelligence that is exploring how computers can detect and express human emotions. Emotion recognition has already been deployed to a variety of societal purposes, ranging from judging consumers' reactions toward an advertisement to helping children with Asperger's understand facial expressions [26]. The desire for emotion recognition technology is only expected to increase with the market for these solutions projected to be worth over 56 billion dollars by 2024 [1]. Yet, the popularity does not come without scrutiny. To start, there are debates about whether these technologies can actually infer any information about emotion. For example, just because a person is scowling does not always mean that they are angry [26]. Furthermore, the expressions that people use to show emotions can vary across cultures and situations. There are also concerns about bias in FER, especially if these technologies are employed for security or surveillance purposes. Given the black-box nature of facial recognition technologies offered by companies, it is necessary to conducts audits in order to preserve their accountability to the public.

This paper aims to make these technologies more transparent by conducting an intersectional analysis of facial emotion recognition services. First, this study advances our knowledge of existing FER technologies. Since many government agencies or other companies use services provided by large technology companies (e.g. Amazon, Google, etc.) rather than creating their own, it is more applicable to audit these existing APIs. While there have been studies surveying the accuracy of these services, the studies have not looked at it with respect to gender and race. Second, this work is building upon Buolamwini and Gebru's "Gender Shades" paper [6]. In "Gender Shades," Buolamwini and Gebru conducted an audit on 3

commercial gender classifiers, looking at differences in accuracy rates based on a confluence of skin color and gender. Similarly, we are using an intersectional approach and looking at the combined factors of both skin color and gender on FER classifiers. Instead of considering these factors in isolation, we will look at the accuracy of the 3 commercial classifiers on four subgroups: lighter-skinned males, lighter-skinned females, darker-skinned males, and darker-skinned females. This provides a greater insight as to how these services work on different groups.

2. Problem Background and Related Work

2.1. Emotion Detection

Facial emotion recognition is the automated interpretation of visual facial expressions to predict human emotion. In social interaction, facial expressions along with vocal intonation are key nonverbal communicators for gauging human emotion. As computer vision and artificial intelligence becomes more advanced and widespread, it is unsurprising that companies, such as Face++ or Affectiva, and academics alike are searching for an automated way to detect human facial emotion. At the moment, much of the research is built upon Ekman et al.'s [9] defined "universal face expressions," which include happiness, sadness, anger, fear, surprise, and disgust as the basic units of emotion. These six facial expressions, as well as a seventh "neutral" emotion, are hypothesized to be universally applicable, regardless of culture. Generally speaking, automated FER algorithms will take in facial images and return a confidence value or values corresponding to these basic emotions.

In computer vision, the current state-of-the-art FER classifiers are built using Convolutional Neural Networks (CNNs). CNNs work by convolving the input image with a series of filters in the convolutional layers to create a feature map. The fully combined layer then uses this feature map to help classify the likelihood the input belongs to a certain facial expression class. In Saravanan et al.'s [24] paper, they compared the accuracy rate of three different types of models– decision trees, feedforward neural networks, and CNNs– to find that CNNs were

the best-performing model out of the three. The CNN that Saravanan et al. [24] proposed included six convolutional, two max pooling, and two fully connected layers to achieve an accuracy of 0.60 on the FER-2013 dataset [7].

As FER algorithms have become increasingly popular, they have proven to be applicable to a wide variety of purposes. In Zhan et al.'s [27] paper, they proposed an automated facial emotion system for multiplayer online games. The facial emotion recognition could be used to manipulate the expressions of the avatars in the game. Also within the realm of computer gaming, Blom et al. [5] explored using facial emotion recognition for a more personalized gaming experience. In the study, researchers would tailor the level that a user was playing based on affect analysis. In a completely different application, Kalantarian et al. [19] used FER technologies to assist children with autism. Given the varied applications for emotion detection, the societal effects of these technologies will be profound. Thus, ensuring that these algorithms are held to a standard of accountability and transparency, which can be done through intersectional audits, is essential.

2.2. Classifier Audits

There have been prior studies evaluating the performances of commercial AI services. The National Institute of Standards and Technology's Face Recognition Vendor Test (FRVT) [15] found that commercial gender classifiers were more accurate on male images than female. Moreover, all of the algorithms were less accurate when predicting the gender for females over the age of 50. Furthering this research, Buolamwini and Gebru's [6] 2018 paper was a particularly influential piece of work that evaluated gender classifiers for intersectional disparities. Buolamwini and Gebru created an intersectional benchmark, looking at four target groups: lighter-skinned males, lighter-skinned females, darker-skinned females, and darker-skinned males. They found that the commercial classifiers (IBM, Face++, and Microsoft) studied were most likely to incorrectly classify images of darker-skinned females. In addition, they released the Pilot Parliaments Benchmark, a more balanced dataset in terms of gender

and skin color (using Fitzpatrick Skin Type).

In response to these audits on commercial classifiers, Muthukumar et al. [22] tried to isolate what could be the cause of these discrepancies. They manipulated the perceived phenotypic skin color and gender of the subjects through luminance mode-shift, color transfer, and image cropping. From these manipulations, they found that the discrepancies persisted, suggesting that it is not skin color or gender indicators (e.g. hair length) alone causing the differences in performance. Rather instead of finding an invariant classifier that performed equally across demographic groups, Muthukumar et al. suggested that we should focus on creating representative training datasets.

Facial emotion recognition algorithms' performance have not evaluated as closely compared to gender classifiers. There has been a baseline evaluation established for the accuracy of different commercial solutions [2]. Furthermore, a paper by Rhue [23] used a dataset of NBA player headshots to analyze Microsoft and Face++'s emotion detection algorithms. She found that, on average, Black players' emotions were more likely to be interpreted as negative (e.g. angry or contemptuous) than white players. For Face++, this occurred regardless of the player's expression whereas for Microsoft, the discrepancy only occurred when the player was not smiling. Nonetheless, the question of race and gender has not been studied in regards to commercial FER classifiers. This study addresses this question through using an intersectional output and furthers our understanding by conducting experiments to see whether race or gender in isolation has any effect on the classifiers' accuracy.

3. Approach

In this paper, we evaluated the performance of 3 commercial FER classifiers– Microsoft, Amazon, and Google. We used a testing set drawn from Mollahosseini et al.'s [21] AffectNet dataset as the input to the classifiers. Looking at the effects of race and gender on accuracy rates, we found that all three classifiers performed the worst on darker-skinned males. We then modified the input images to cosmetically adjust the perceived race and gender. Manipulating

the images provides better insight into how these features in isolation result in different accuracy rates for the FER classifiers, which has not been previously studied.

The key premise behind this study is to evaluate the performance of commercial systems on intersectional inputs. The existing technologies are treated as a black-box system that are being tested against a set of varied inputs to see whether the outputs are biased. In particular, this study is using techniques that have been proven to be successful in previous studies of facial recognition and applying them to an unexplored subset. In addition, this study is deepening the knowledge of emotion detection technologies. While other audits looked at the effect of race, they were not using datasets that were annotated with the different facial expressions and were also only done by race rather than phenotypic skin color. Overall, this study will provide a guideline for not only how to conduct intersectional evaluations of commercial systems but also how to find possible explanations that might account for these differences.

4. Implementation

4.1. Data Acquisition

This study uses Mollahosseini et al.'s [21] AffectNet as the testing set for the commercial classifiers. AffectNet includes approximately 420K facial images that have been manually annotated by twelve human experts for eight emotions: neutral, happiness, sadness, anger, surprise, fear, disgust, and none. The "none" category includes images for which the emotion could not be assigned to one of the other seven other emotions listed. The dataset has a larger distribution of happy and neutral faces, followed by sadness and anger - see Table 1. In addition to the affect annotation, AffectNet also includes the valence and arousal of the input image on a scale from [-1, 1] as well as the OpenCV face boundary boxes.

This study chose to use AffectNet as the main dataset because it has the largest annotated dataset of facial expressions in the wild. Compared to other annotated datasets, AffectNet has a greater degree of granularity for emotions– offering eight emotions while most offer

Emotion	Count	Percentage
Happy	134,915	32.1%
Neutral	75,374	17.9%
None	33,588	8.0%
Sad	25,959	6.2%
Anger	25,382	6.0%
Surprise	14,590	3.6%
Fear	6,878	1.6%
Disgust	4,303	1.0%
Contempt	4,250	1.0%
Uncertain	12,145	2.9%
Non-face	82,915	19.7%
Total	**420,299**	**100.0%**

Table 1: AffectNet manually annotated images

Figure 1: Examples of emotions in AffectNet

seven. Also, running the commercial classifiers on in the wild images rather than controlled or posed images is more similar to actual applications of the technologies, giving a better sense of what the accuracy rates might be in the real world.

4.2. Data Labelling

The input images from the AffectNet dataset needed to be labelled for the intersectional analysis. The first 1000 images of the dataset were manually annotated for both gender and phenotypic skin color.

4.2.1. Gender Labelling For gender, we labelled using a binary system based on perceived gender expression. This was the only option considering that no other identifying information

was provided. While we acknowledge that this labelling system ignores the full spectrum of gender identity by reducing it to a binary scale, the 3 classifiers we are analyzing also return gender as a binary, although there are inconsistencies amongst classifiers if they are measuring gender expression or identity [25]. Images that we were unsure of for gender, such as infants, were marked as unknown and later left out of the testing set.

4.2.2. Skin Color Labelling This study uses phenotypic skin color (e.g. lighter-skinned and darker-skinned) rather than race. This decision was influenced by Buolamwini and Gebru's [6] rationale. Even though race is a protected class and is often used in algorithmic audits, this study acknowledges that there is a wide diversity of phenotypic features and skin color within a racial group. Furthermore, given the social construction of race, we thought it would be more accurate to have an objective measure, like skin color, to use when evaluating the classifiers.

To label skin color, we used the Fitzpatrick Skin Type classification system, a commonly used method for ranking phenotypic skin color based on response to UV exposure. The Fitzpatrick skin type scale ranks skin color on a six-point basis with I representing pale and extremely sensitive skin to VI representing "deeply pigmented" skin [11]. Using the Fitzpatrick skin type, all images that were labelled IV, V, or VI were in the "darker-skinned" category and those rated I, II, or III were in the "lighter-skinned" category.

We chose to use the Fitzpatrick skin type scale since it is widely used amongst dermatologists and has also been used in other intersectional audits involving race. As was acknowledged in the Buolamwini and Gebru paper [6], the Fitzpatrick skin type scale does skew towards representing White or light-skinned individuals. Nonetheless, as illustrated from Figure 2 which uses images from the AffectNet dataset to illustrate the spectrum of the Fitzpatrick skin type scale, we believe the categories were broad enough to capture the range of phenotypic skin color present in the dataset.

4.2.3. Testing Set To get our final testing dataset, we filtered the annotated images. Some images we could not label with a gender or race. For example, the dataset included images of

Figure 2: Fitzpatrick skin type scale using AffectNet images

emojis and cartoon characters. These were removed from the testing set. After filtering the dataset, we had a total of 930 images in our testing set. The testing set is slightly skewed towards more females than males at 61.2% versus 38.8%, differing from many datasets which typically have more images of males - see Table 2. AffectNet is heavily skewed towards images of lighter-skinned individuals. Of the four categories, darker-skinned males make up only 6.9% of the total images in the testing set. As compared to Table 1, the testing set is more imbalanced in terms of emotions– see Figure 3. Overall, the emotion distribution of the testing set is biased towards happy emotions (68.5%) compared to the entire AffectNet dataset (32.1%).

n	F	M	L	D	LF	LM	DF	DM
930	61.2	38.8	80.3	19.7	48.4	31.9	12.8	6.9

Table 2: Breakdown of the AffectNet testing set showing the percentage of females (F), males (M), darker-skinned individuals (D), lighter-skinned individuals (L), darker / lighter-skinned females (DF / LF), and darker / lighter-skinner males (DM, LM)

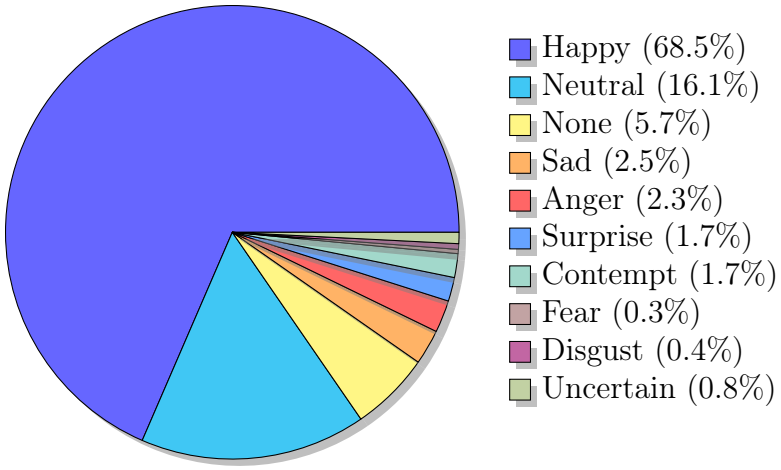

Figure 3: Distribution of emotions in the testing set

4.3. Classifiers

This study audits 3 popular commercial classifiers– Microsoft Azure Cognitive Services, Amazon Rekognition, and the Google Cloud Vision. Rather than using academic FER classifiers, this study focuses on commercial offerings since they are more likely to be used by government agencies or private corporations. For example, Microsoft Cognitive Services lists KPMG, Volkswagen, Uber, Jet.com, and Seattle Against Slavery as successful customer stories on their website [20]. Amazon Rekognition also lists clients in wide range sectors from retail (Daniel Wellington and Popsugar) to public services (Washington County Sheriff Office)[3]. Thus, these 3 solutions were chosen since they come from big tech companies that offer popular AI bundles which included facial emotion recognition.

4.3.1. Microsoft Azure Cognitive Services. Microsoft Azure Cognitive Service's Face API offers face attribute detection and recognition algorithms, including perceived emotion recognition. Microsoft outputs eight basic emotions: happy, sad, angry, confused, disgusted, surprised, fear, and neutral. For each detected face, the API returns a confidence score in [0, 1], and the highest confidence score represents the dominant emotion detected. Microsoft makes the disclaimer in its documentation that the classifier returns the perceived emotion, which does not always reflect the individual's internal state [20].

4.3.2. Amazon Rekognition. Similar to Microsoft Azure, the Amazon Rekognition API can detect eight basic emotions: happy, sad, angry, confused, disgusted, surprised, fear, and calm. The calm emotion corresponds to neutral in the AffectNet dataset. The API returns a confidence score in [0, 100] for each of the eight emotions. The highest confidence score denotes the dominant emotion showcased. At the time of evaluation, Amazon Rekognition stated that they used "deep neural networks" to detect and label faces but did not give any specificity on the expected performance or training data of the API. To note, in the developer documentation, Amazon did preface the classifier by stating the API is only judging perceived emotion and not the subject's actual internal state [3].

4.3.3. Google Cloud Vision. Even though Google does not offer "specific individual facial recognition" technologies, based on privacy concerns, it does offer solutions that analyze facial images for gender, age, and emotions. Compared to Microsoft and Amazon's algorithms, Google Cloud Vision API is limited in its output and granularity, only returning four possible emotions: joy, sorrow, anger, and surprise. Furthermore, unlike Microsoft and Amazon that both provided continuous confidence scores, Google Cloud Vision ranked each image on a discrete scale of very likely, likely, possible, unlikely, very unlikely, and unknown. Google did not provide any specifics on the algorithms behind their API or any disclaimers about accuracy for the users [14].

4.3.4. Performance While companies do not provide any information about the accuracy of the classifiers or performance on benchmark datasets, Al-Omair and Huang [2] conducted an audit on the performance of these services. Across all 3 commercial classifiers, they found that the APIs were most accurate at detecting happiness, which also had the highest average confidence scores. Google had the highest overall recognition accuracy score at 85%, but this might have been due to the fact that the classifier recognizes less emotions compared to the others. When Al-Omair and Huang eliminated the "uncommon emotions" (e.g. fear and disgust), they found that Microsoft had the highest recognition accuracy at 97%. Given the large proportion of happy emotions in our testing set (Figure 3), we expect that the accuracy

rates for the classifier will be higher than those reported in Al-Omair and Huang's study.

4.4. Experiments

Using the techniques from Muthukumar et al [22], we conducted a series of experiments on the input images. These experiments were a computationally inexpensive way of manipulating the phenotypic skin color and perceived gender expression. In this study, we used the luminance mode-shift and color transfer to adjust the phenotypic skin color and image cropping to adjust gender expression.

4.4.1. Luminance Mode-Shift To implement luminance mode-shift, there needs to be a rule for differentiating between skin color pixels and non-skin color pixels. In Muthukumar et al.'s [22] study, they implemented the following simple skin detection rule based in the YCbCr space. The Y dimension represents the luminance and Cr, Cb represent the chrominance values.

$$\text{skin} = \begin{cases} \text{true} = 90 \leq Cr \leq 115 \text{ and } 140 \leq Cb \leq 195 \\ \text{false} = \text{otherwise} \end{cases}$$

However, this skin detection rule did not work on the images in AffectNet. Instead, we used the skin detection rule by Basilio et al [4]. This technique also is implemented in the YCbCr color space but with different bounds for Cb and Cr. We found the following worked better given the varying skin colors and image lighting across the AffectNet dataset.

$$\text{skin} = \begin{cases} \text{true} = 85 < Cr < 135 \text{ and } 135 < Cb \leq 180 \\ \text{false} = \text{otherwise} \end{cases}$$

Using this skin detection rule, we implemented the luminance mode-shift using the following procedure for an inputted new Y mode, Y_{new}:

1. Convert image from input color space to YCbCr color space
2. Find the Y mode, Y_{old} for all pixels $\in$ {skin color}

3. Calculate the luminance mode shift $\delta = Y_{\text{new}} - Y_{\text{old}}$
4. Shift the luminance values for all pixels such that $I'_Y = I_Y + \delta$
5. Clamp all pixels so that $I'_Y \in [0, 255]$

Luminance mode-shift approximately changes the skin color of the input subject - see Figure 4. This method does not produce the most visually realistic results; however, luminance mode shift is a quick computation, especially when compared to color transfer. For the experiment, we used one set of images where the skin color had been approximately lightened ($Y_{new} = 100$) and another set that had been approximately darkened ($Y_{new} = 30$).

4.4.2. Color Transfer The goal of using color transfer was to create a more visually realistic output image. Rather than adjusting luminance values as we did for luminance mode-shift, color transfer instead tries to match the color histogram of one image to a reference histogram. Color transfer can also be represented in terms of an optimal transport (OT) problem. In this implementation, we used the open-source Python library, Python Optimal Transport (POT) for color transfer [12].

POT's implementation of color transfer is based on Ferradan et al.'s [10] paper, which introduces regularized discrete OT for color transfer. The regularized OT method improves the resulting image by reducing the color artifacts that arise from noise amplification. POT gives two solvers for optimal transport: the Earth Mover's Distance and the Sinkhorn-Knopp algorithm. While we did find that the Sinkhorn-Knopp implementation reduced the color artifact on the image, it also flattened it. Thus, in this study, we decided to use the output images from the Earth Mover's Distance transport algorithm.

Using POT, we implement color transfer. In order to do this, we must have two images: an input image and a reference image. The goal is to match the color histogram of the input image to the reference image. After fitting the transport algorithm to the input and reference image, we use POT to generate the OT matrix. This is then used to correspond the pixel colors in the input image to that of the reference. To darken and lighten the skin color of the test set images, an image with Fitzpatrick Skin Type VI and I respectively were used as the

reference image - see Figure 5.

Figure 4: Example images from AffectNet which have had their luminances mode-shifted. From left to right, images go from darkened to lightened

Figure 5: Example images from AffectNet which have been color transferred. Reference images used for color transfer are labelled "Lighter Reference" and "Darker Reference"

4.4.3. Cropped Images The justification for cropping the images was to eliminate hair length as a possible indicator for gender expression. When visually analyzing Buolamwini

and Gebru's results, Muthukumar et al. [22] found that female subjects with short hair were more likely to be inaccurately identified. Thus, they hypothesized that these classifiers take simplistic features like hair length to be indicators of binary gender expression.

To account for this in the experiments, we cropped the images to only remove differences in hair length. The input images from AffectNet are a 15% expansion of the OpenCV boundary boxes. We cropped these images according to the OpenCV boundary boxes provided in the AffectNet dataset. As illustrated in Figure 6, this method does not remove all indicators of hairstyle, such as facial hair, which also is used in society as a gender indicator. While removing facial hair or other gender indicators could be done using more complicated methods, like a GAN, these are not only more computationally expensive than the simplistic cropping method but also not guaranteed to give convincing results, especially when applied to a diverse set of input images [22].

Figure 6: Examples images from the AffectNet dataset with the cropped input images in row A compared to the original image in row B

4.5. Emotion Detection Audit

Both the the original testing set and generated images are inputted into the commercial APIs for emotion recognition. The three APIs all return their responses in a JSON format. The responses include the confidence score of each basic emotion for every detected face. Given that our input images are all portraits, we assume that the APIs will detect at most one face

per image. From the returned confidence scores, the highest confidence score represents the perceived dominant emotion. Only for Google API which returns text rankings (e.g. very likely, unlikely, etc.) do we convert the confidence score to a numeric scale [0, 4]. We record the dominant and the corresponding confidence score for each of the input images.

5. Results

We evaluated the 3 commercial facial emotion detection classifiers– Microsoft, Amazon, and Google– using the AffectNet dataset. Besides evaluating the baseline dataset, we also conducted 3 experiments, editing the phenotypic skin color and hair length. Overall, the classifiers performed better on female than males and better on lighter-skinned inputs than darker-skinned. All three commercial classifiers performed the worst on darker-skinned males.

5.1. Baseline Results

The key insights from evaluating the results for the baseline dataset are as follows:

- Classifiers have a skewed accuracy rate for different emotions
- Microsoft and Amazon performed the best on darker females with error rates of 11.8% and 13.9%
- All classifiers performed the worst on darker-skinned males (error rates between 16.7% and 20.3%)

Following Buolamwini and Gebru [6] as well as NIST's [15] evaluations of gender classifiers, we used positive predictive value (PPV) to measure the accuracy of the classifiers. We found the error rate for the classifiers by subtracting the PPV from 1. Contrary to expectation, we found that darker-skinned females were the best-performing group for two of the three classifiers. In addition, all three classifiers performed better on females than males with differences in error rates ranging 2% to 3.9%. This contradicts the findings from NIST[15] and Buolamwini and Gebru [6]'s studies conducted on gender classifiers, which tend to perform better on male images and the worst on darker-skinned females.

Across all classifiers, we found that darker-skinned males had the lowest accuracy rate for

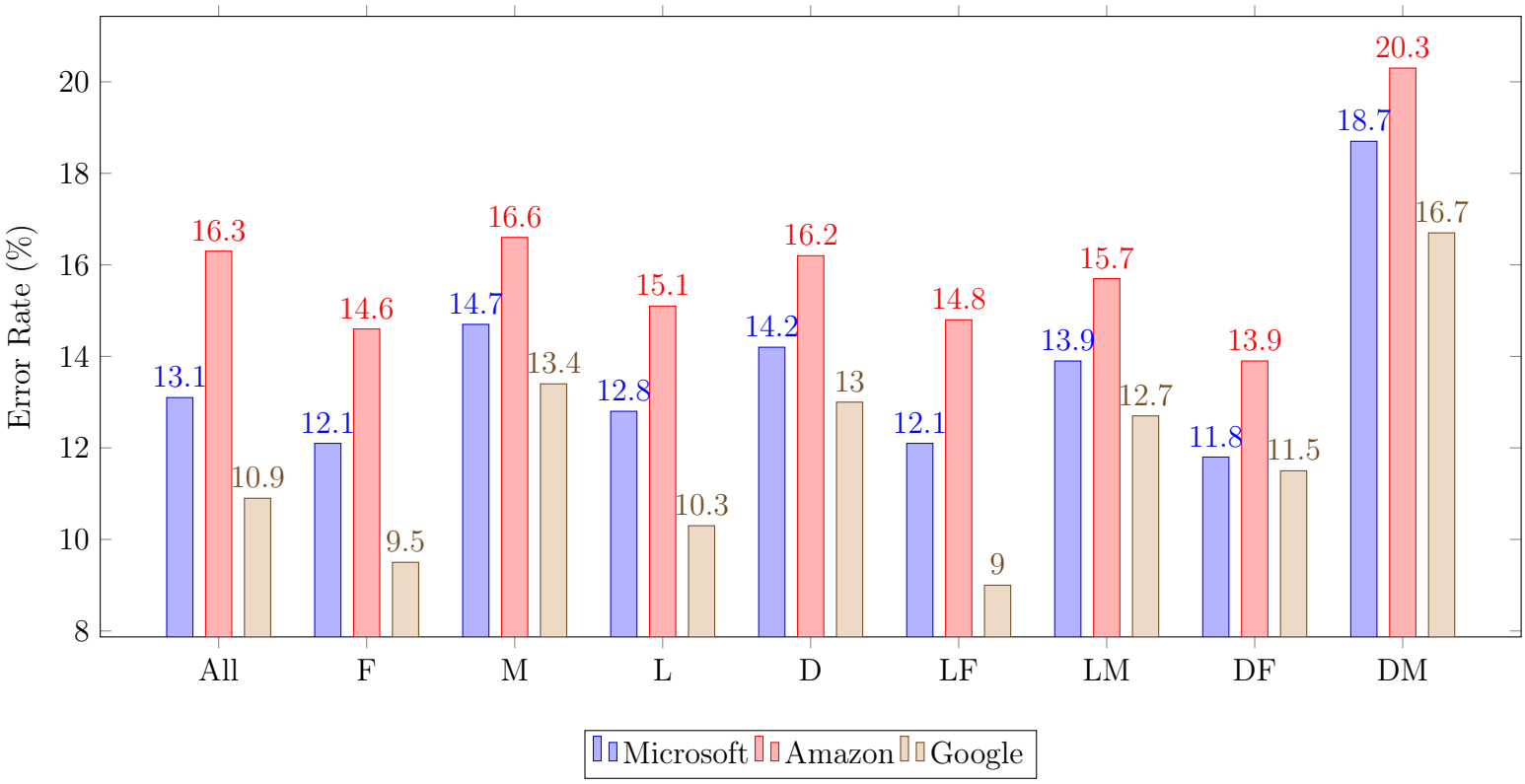

Figure 7: Error rates (1 - Positive Predictive Value) for each of the classifiers

emotion prediction. Darker males have an error rate ranging from 16.7% to 20.3%. Compared to the overall error rate, this is 4% to 5.7% greater than the average error rate across all groups - see Figure 7. We did a deeper exploration of the input images for darker-skinned males that the classifiers were incorrectly identifying. The commercial classifiers are more likely to incorrectly identify darker-skinned males who are higher on the Fitzpatrick Skin Type scale. Microsoft and Azure performed perfectly on males with Fitzpatrick Skin Type IV (see Table 3). The average skin color rating for darker-skinned males is a 5.1 from the baseline testing compared to an average rating of 5.4, 5.6, and 5.3 for the inaccurately predicted male images from Microsoft, Amazon, and Google respectively. This suggests that even within our intersectional analysis there is further gradation in performance stemming from variations in skin color even within the "lighter-skinned" or "darker-skinned" category.

Since the emotion detection had multiple classes, unlike gender classifiers which tend to be binary, we included mean average precision (mAP) and mean accuracy per class (mean accuracy) to get a better insight on how the classifiers performed for different emotions. We

	IV (%)	V (%)	VI (%)
Testing Set	22.0	42.4	35.6
Microsoft	0.0	71.4	28.5
Amazon	0.0	41.7	58.3
Google	9.1	45.5	45.5

Table 3: Percent distribution of Fitzpatrick Skin Type ratings for darker-skinned males. Testing set gives the total distribution of darker skin males. Distributions for Microsoft, Amazon, and Google represent incorrectly identified images

Microsoft Amazon Google

Figure 8: Average face of incorrectly predicted darker-skinned males [13]

see that the mAP and mean accuracy values tend to be similar; however, when comparing mAP values to PPV, the difference is much larger, ranging between 19.5% and 55.6%. This suggests that there might be discrepancies in accuracy rates between the emotions. Following the results from Al-Omair [2], we found that all three classifiers had high accuracy rates when detecting happy and neutral emotions with an average. But for sad, surprised, angry, and contemptuous faces, the classifiers had much lower accuracy rates. Given the skewed distribution of emotions for our testing set -see Table 3, this can contribute to the difference in PPV and mAP / Mean Accuracy.

5.2. Experiment Results

Looking at the three experiments we conducted– luminance mode-shift, color transfer, and image cropping, we found the following results:

- Luminance mode-shift had a significant change in accuracy rates for Microsoft and Google in lightened images
- Color transfer did not lead to significant changes in accuracy rates

Classifier	Metric	All	F	M	L	D	LF	LM	DF	DM
Microsoft	PPV (%)	86.9	87.9	85.2	87.2	85.8	87.9	86.1	**88.2**	81.4
	mAP (%)	42.8	50.2	44.2	46.9	37.0	**48.7**	48.4	33.2	36.9
	Mean Accuracy (%)	40.0	46.0	40.1	40.4	34.5	38.1	**41.1**	30.8	33.8
Amazon	PPV (%)	83.7	85.4	83.4	84.9	83.8	85.2	84.3	**86.1**	79.7
	mAP (%)	55.8	53.7	60.0	60.7	57.1	50.5	**64.8**	58.1	52.5
	Mean Accuracy (%)	55.0	49.6	57.0	54.7	49.4	46.3	**58.5**	44.1	43.8
Google	PPV (%)	89.1	90.5	86.6	89.7	87.0	**91.0**	87.3	88.5	83.3
	mAP (%)	41.7	42.3	43.7	46.0	49.3	41.4	49.7	**49.8**	36.6
	Mean Accuracy (%)	40.2	37.2	40.4	40.6	38.4	35.3	**42.6**	36.6	28.6

Table 4: Emotion prediction classifier performance measured by PPV, mAP, and mean accuracy for the 3 classifiers

- Cropping the image results in a significant change in accuracy rates for Amazon, especially for darker-skinned inputs

See Appendix for results disaggregated by gender, skin-color, and subgroup.

5.2.1. Phenotypic Skin Color Experiments We conducted two experiments for changing phenotypic skin color: luminance mode-shift and color transfer. For both experiments, we produced a version of the testing dataset with the phenotypic skin color lightened and another darkened. Based on observations from the baseline tests, we hypothesized that lightening the skin color should improve accuracy rates since the classifiers performed better on lighter skin inputs. Conversely, the classifiers should decrease in accuracy on the darkened images using this same intuition.

To evaluate the significance of the changes, we conducted chi-squared tests on the experimental results to see whether there is a statistically significant difference between the expected frequencies and the actual frequencies. For our chi-squared tests, we took the count of accurately identified images from the baseline results for each subgroup (LF, LM, DF, DM) to be the expected and the count of accurately identified from the experimental results to be our actual frequencies. Our null hypothesis was that there was no difference between the expected and the actual results. To find the p-value based on our chi-squared statistic, we used 3 degrees of freedom in our calculations given that we had $k = 4$ parameters.

The darkened skin inputs show that darkening phenotypic skin color did not have a

Classifier	Type	χ^2	p-value
Microsoft	Dark	0.24	0.97
	Light	17.4	5.8e-4*
Amazon	Dark	0.99	0.80
	Light	0.0	1.0
Google	Dark	0.06	0.99
	Light	37.6	3.4e-8

* significant at 1%

Table 5: Test statistic (χ^2) and p-value of chi-squared test on PPV of the four subgroups (LF, LM, DF, DM) from luminance mode-shift images compared to baseline images. Dark refers to darkened images and light to lightened

Classifier	Type	χ^2	p-value
Microsoft	Dark	0.34	0.95
	Light	0.80	0.85
Amazon	Dark	1.23	0.74
	Light	6.21	0.10
Google	Dark	0.69	0.88
	Light	1.48	0.69

Table 6: Results of chi-squared test on PPV of the four subgroups from the color transfer images compared to baseline images

significant effect on the accuracy rates. There was no pattern for the changes in PPV. Furthemore, from the chi-squared test, the p-value > 0.01 for all three classifiers given 3 degrees of freedom - see Tables 5 and 6, meaning we cannot reject our null hypothesis. This suggests that any discrepancies that we saw between the baseline and the darkened testing sets is most likely due to chance rather than related to the phenotypic changes.

Looking at the lightened skin datasets, we find significant results for luminance mode-shift. For Microsoft and Google, the chi-squared test returns p-values of 0.0005 and $3.4e{-}8$ respectively. We reject the null hypothesis, stating that the difference between the baseline and luminance mode-shift lightened images is not due to chance. However, it is important to note that chi-square test is used on counts rather than proportions. The significance of the results might be due to the fact that the total number of identified images decreased as the classifier could not detect faces in some images rather than a change in the accuracy rate for the different subgroups.

5.2.2. Cropped Image Experiment Based on the results from the baseline testing set indicating that the commercial classifiers perform worse on male images compared to females, we hypothesize that cropping the images will decrease the accuracy rate of commercial classifiers. For Microsoft and Google, we found that cropping the images did not have a significant effect on the accuracy rates. In fact, for darker-skinned females, cropping the images increased the PPV by 2.6% and 1.2 % respectively - see Table 7.

On the other hand, for Amazon, the classifier performed significantly worse on the cropped images than on the baseline images. The p-value from the chi-square test is 4.2e−7. The difference is most pronounced for darker-skinned females and males with differences in PPV equal to 14.2% and 7.2%. In case this discrepancy was due to a decrease in the total overall recognized images rather than a difference in accuracies, we also conducted one-tailed difference in proportions t-tests to confirm these findings.

The t-test helps us evaluate whether there is statistically significant difference in the proportion of accurately identified images between the baseline and the cropped image sets. The metric we are comparing is the PPV for darker-skinned females / males. P_1 denotes the PPV from baseline images and P_2 denotes from cropped. Using our observations from the baseline results that classifiers performed the worst on darker-skinned males, we state the following null and alternate hypotheses:

$$H_0 : P_1 - P_2 \leq 0$$

$$H_a : P_1 - P_2 > 0$$

Using a threshold of < 0.01, we can reject the null hypothesis for darker-skinned females but not for darker-skinned males - see Table 8. For darker females, we get a p-value of 0.006, indicating that the PPV is significantly lower for the cropped images than the baseline. Since the cropping images removed the gender indicator of hair length, this decrease in performance on darker-skinned females is in line with our previous observation that the classifiers had the

Classifier	Type	LF	LM	DF	DM	Test Statistic
Microsoft	Baseline	86.1	87.9	**88.2**	81.4	$\chi^2 = 0.23$,
	Crop	87.1	85.4	**91.8**	79.7	p-value = 0.97
Amazon	Baseline	84.3	85.2	**86.1**	79.7	$\chi^2 = 32.5$,
	Crop	**83.8**	82.2	71.9	72.5	p-value = 4.2 e-7*
Google	Baseline	87.3	**91.0**	88.5	83.3	$\chi^2 = 1.46$,
	Crop	**90.1**	88.1	89.7	78.0	p-value = 0.69

* significant at 1%

Table 7: Emotion classification performance measured by the PPV of the 3 commercial classifiers for both the baseline testing set and the cropped testing set

Group	Type	n	Proportion	p-value
DF	Baseline	108	86.1%	0.006*
	Crop	96	71.9%	
DM	Baseline	59	79.7%	0.19
	Crop	51	72.5%	

* significant at 1%

Table 8: One-tailed difference in proportions t-test for the Amazon classifier comparing the PPV for the baseline testing set and the cropped testing set

highest error rates on darker-skinned males.

6. Conclusion and Future Work

In this study, we evaluated the accuracy for the commercial emotion detection classifiers from Microsoft, Amazon, and Google on a sampling of the AffectNet dataset. The images in the testing set were already manually annotated by experts for perceived emotion, and we further labelled them with the phenotypic skin color using the Fitzpatrick Skin Type Scale and perceived gender expression. We analyzed the classifier outputs using an intersectional approach along the axes of gender and race. From the initial baseline testing set, we found that the classifiers had higher accuracy rates for happy and neutral emotions compared to the other basic emotion units that they detected. Furthermore, all three classifiers performed better on females than males and on lighter-skinned individuals than darker-skinned. Across all three commercial classifiers, darker-skinned males had the lowest accuracy.

Using the baseline data, we then conducted a set of experiments to look at features in isolation. We used luminance mode-shift and color transfer to adjust the phenotypic skin

color. Changing the perceived skin color did not significantly change the accuracy of the classifiers. This might be because racial perception is not solely dependent on phenotypic skin color and the underrepresentation of certain features or facial geometries in the training set can influence the accuracy rate. To isolate gender expression, we cropped the hair from the photos since hair length can be interpreted as a gender indicator. For Amazon, we found a significant decrease in accuracy, especially for darker-skinned individuals. While we acknowledge that there are obviously other gender indicators besides hair, this does echo our previous finding that the classifiers perform particularly poorly on images of darker-skinned males.

Future work should look into how these classifiers are interpreting the images. For example, in Muthukumar et al. [22], they conducted an experiment to find the "minimal sufficient explanation," meaning they tried to find the minimum features necessary for the classifier to make a gender prediction. This can be applied to emotion detection classifiers and also analyzed with an intersectional approach to see if there are differences depending on the gender and racial indicators in the image. Future work can also include conducting audits using an unbiased dataset. Our sample testing set from AffectNet was skewed toward happy emotions. Emotion detection classifiers tend to perform better on neutral and happy emotions. To get a more precise understanding of these classifier's accuracy across different emotions, we need to test on a dataset that is more balanced as well.

Looking at this experiment within a broader scope, it illustrates the importance of algorithmic audits. Given the wide applications of facial emotion detection and the great social footprint these technologies will inevitably have, it is essential to conduct audits as a means of promoting transparency and keeping these companies accountable. As this study shows, there is not one specific feature in isolation that attributes to discrepancies in intersectional accuracies. Part of the means to create more equitable technology is thus to make sure that our training data is robust and more representative, including, but of course not limited to, race and gender.

7. Acknowledgments

I would like to acknowledge my advisor, Professor Olga Russakovsky, for her continued support and guidance throughout the semester. I would also like to thank Princeton University's School of Engineering and Applied Sciences for provided the funding for this project. Finally, I want to thank Mollahosseini et al. for creating the AffectNet dataset, Flamary and Courty for developing the Python Optimal Transport library, and George Gach for creating the average face generator.

8. Honor Code

I pledge my honor that this paper represents my own work in accordance with University regulation.

References

[1] "Emotion detection and recognition market worth $56.0 billion by 2024," PR Newswire, Feb 2020.

[2] O. M. Al-Omair and S. Huang, "A comparative study on detection accuracy of cloud-based emotion recognition services," in *Proceedings of the 2018 International Conference on Signal Processing and Machine Learning*, ser. SPML '18. New York, NY, USA: Association for Computing Machinery, 2018, p. 142–148. Available: https://doi.org/10.1145/3297067.3297079

[3] *Amazon Rekognition Documentation*, Amazon, 2020.

[4] J. A. M. Basilio *et al.*, "Explicit image detection using ycbcr space color model as skin detection," *Applications of Mathematics and Computer Engineering*, pp. 123–128, 2011.

[5] P. M. Blom *et al.*, "Towards personalised gaming via facial expression recognition," in *Tenth Artificial Intelligence and Interactive Digital Entertainment Conference*, 2014.

[6] J. Buolamwini and T. Gebru, "Gender shades: Intersectional accuracy disparities in commercial gender classification," in *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, ser. Proceedings of Machine Learning Research, S. A. Friedler and C. Wilson, Eds., vol. 81. New York, NY, USA: PMLR, 23–24 Feb 2018, pp. 77–91. Available: http://proceedings.mlr.press/v81/buolamwini18a.html

[7] P.-L. Carrier and A. Courville, "Fer-2013," 2013.

[8] N. Chokshi, "Facial recognition's many controversies, from stadium surveillance to racist software," The New York Times, May 2019. Available: https://www.nytimes.com/2019/05/15/business/facial-recognition-software-controversy.html

[9] P. E. Ekman and R. J. Davidson, *The nature of emotion: Fundamental questions*. Oxford University Press, 1994.

[10] S. Ferradans *et al.*, "Regularized discrete optimal transport," *CoRR*, vol. abs/1307.5551, 2013. Available: http://arxiv.org/abs/1307.5551

[11] T. B. Fitzpatrick, "The validity and practicality of sun-reactive skin types i through vi," *Archives of dermatology*, vol. 124, no. 6, pp. 869–871, 1988.

[12] R. Flamary and N. Courty, "Pot python optimal transport library," 2017. Available: https://github.com/rflamary/POT

[13] G. Gach, "face-average," Oct 2018. Available: https://github.com/georgegach/face-average

[14] *Detect Faces*, Google, Apr 2020.

[15] P. J. Grother, P. J. Grother, and M. Ngan, *Face recognition vendor test (FRVT)*. US Department of Commerce, National Institute of Standards and Technology, 2014.

[16] K. Hao, "Ai is sending people to jail – and getting it wrong," MIT Technology Review, Jan 2019.

[17] D. Harwell, "A face-scanning algorithm increasingly decides whether you deserve the job," Washington Post, Nov 2019.

[18] T. Hatmaker, "Ucla backtracks on plan for campus facial recognition tech," TechCrunch, Feb 2020.

[19] H. Kalantarian *et al.*, "Labeling images with facial emotion and the potential for pediatric healthcare," *Artificial intelligence in medicine*, vol. 98, pp. 77–86, 2019.

[20] *Face Documentation*, Microsoft, 2020.

[21] A. Mollahosseini, B. Hasani, and M. H. Mahoor, "Affectnet: A database for facial expression, valence, and arousal computing in the wild," *IEEE Transactions on Affective Computing*, vol. 10, no. 1, pp. 18–31, 2017.

[22] V. Muthukumar *et al.*, "Understanding unequal gender classification accuracy from face images," *CoRR*, vol. abs/1812.00099, 2018. Available: http://arxiv.org/abs/1812.00099

[23] L. Rhue, "Racial influence on automated perceptions of emotions," *Available at SSRN 3281765*, 2018.

[24] A. Saravanan, G. Perichetla, and D. K. S. Gayathri, "Facial emotion recognition using convolutional neural networks," 2019.

[25] M. K. Scheuerman, J. M. Paul, and J. R. Brubaker, "How computers see gender: An evaluation of gender classification in commercial facial analysis services," *Proc. ACM Hum.-Comput. Interact.*, vol. 3, no. CSCW, Nov. 2019. Available: https://doi.org/10.1145/3359246

[26] O. Schwartz, "Don't look now: why you should be worried about machines reading your emotions," The Guardian, Mar 2019. Available: https://www.theguardian.com/technology/2019/mar/06/facial-recognition-software-emotional-science

[27] C. Zhan *et al.*, "Facial expression recognition for multiplayer online games," in *Proceedings of the 3rd Australasian conference on Interactive entertainment*. Murdoch University, 2006, pp. 52–58.

9. Appendix

9.1. Luminance Mode-Shift

Classifier	Type	All	F	M	L	D	LF	LM	DF	DM
Microsoft	Baseline	86.9	87.9	85.3	87.2	85.8	87.9	86.1	**88.2**	81.4
	Dark	87.2	88.5	85.1	87.1	87.5	87.9	85.9	**90.8**	81.4
Amazon	Baseline	83.7	85.4	83.4	84.9	83.8	85.2	84.3	**86.1**	79.7
	Dark	86.5	86.8	85.9	86.6	85.8	86.5	86.9	**88.1**	81.5
Google	Baseline	89.1	90.5	86.6	89.7	87.0	**91.0**	87.3	88.5	83.3
	Dark	88.9	89.8	87.4	89.3	87.6	**90.1**	87.8	88.5	85.4

Table 9: Emotion classification performance for the baseline testing set and the darkened luminance mode-shifted testing set

Classifier	Type	All	F	M	L	D	LF	LM	DF	DM
Microsoft	Baseline	86.9	87.9	85.3	87.2	85.8	87.9	86.1	**88.2**	81.4
	Light	87.0	89.1	83.3	87.5	84.9	**89.4**	84.6	88.3	77.6
Amazon	Baseline	83.7	85.4	83.4	84.9	83.8	85.2	84.3	**86.1**	79.7
	Light	84.7	85.4	83.4	84.9	83.8	85.2	84.3	**86.1**	79.7
Google	Baseline	89.1	90.5	86.6	89.7	87.0	**91.0**	87.3	88.5	83.3
	Light	97.1	97.7	96.0	97.1	96.9	**97.9**	95.9	97.1	96.6

Table 10: Emotion classification performance measured by the PPV of the 3 commercial classifiers for the baseline testing set and the lightened luminance mode-shifted testing set

9.2. Color Transfer

Classifier	Type	All	F	M	L	D	LF	LM	DF	DM
Microsoft	Baseline	86.9	87.9	85.3	87.2	85.8	87.9	86.1	**88.2**	81.4
	Dark	87.8	89.5	84.9	88.2	85.9	89.5	86.1	**89.5**	79.3
Amazon	Baseline	83.7	85.4	83.4	84.9	83.8	85.2	84.3	**86.1**	79.7
	Dark	85.0	85.1	84.7	85.1	84.2	84.6	86.0	**87.3**	78.6
Google	Baseline	89.1	90.5	86.6	89.7	87.0	**91.0**	87.3	88.5	83.3
	Dark	90.4	91.3	88.9	90.8	88.8	91.2	90.2	**91.3**	83.3

Table 11: Emotion classification performance measured by the PPV of the 3 commercial classifiers for the baseline testing set and darkened color transfer testing set

Classifier	Type	All	F	M	L	D	LF	LM	DF	DM
Microsoft	Baseline	86.9	87.9	85.3	87.2	85.8	87.9	86.1	**88.2**	81.4
	Light	84.1	85.3	82.3	84.4	82.8	**85.4**	82.8	84.5	79.7
Amazon	Baseline	83.7	85.4	83.4	84.9	83.8	85.2	84.3	**86.1**	79.7
	Light	84.7	84.1	85.6	84.8	83.9	83.7	**86.7**	85.7	80.4
Google	Baseline	89.1	90.5	86.6	89.7	87.0	**91.0**	87.3	88.5	83.3
	Light	91.5	91.7	91.1	91.9	90.0	**91.9**	91.8	91.2	87.2

Table 12: Emotion classification performance measured by the PPV of the 3 commercial classifiers for the baseline testing set and lightened color transfer testing set.

9.3. Cropped

Classifier	Type	All	F	M	L	D	LF	LM	DF	DM
Microsoft	Baseline	86.9	87.9	85.2	87.2	85.8	86.1	87.9	**88.2**	81.4
	Crop	86.7	88.1	84.4	86.5	87.6	87.1	85.4	**91.8**	79.7
Amazon	Baseline	83.7	85.4	83.4	84.9	83.8	84.3	85.2	**86.1**	79.7
	Crop	80.8	81.1	80.4	83.2	72.1	**83.8**	82.2	71.9	72.5
Google	Baseline	89.1	90.5	86.6	89.7	87.0	87.3	**91.0**	88.5	83.3
	Crop	88.7	90.0	86.4	89.4	85.9	**90.1**	88.1	89.7	78.0

Table 13: Emotion classification performance measured by the PPV of the 3 commercial classifiers for both the baseline testing set and the cropped testing set