Evaluating the Accuracy of Automated Facial Emotion Detection on Intersectional Inputs

Dora Zhao, Advised by Professor Olga Russakovsky

INTRODUCTION

Facial Emotion Recognition (FER)

• FER is a growing subset of facial recognition that automatically interprets visual facial expressions to predict human emotion

• Classifiers predict emotion using basic emotion units as defined by Ekman et al.'s [2] "universal face expressions"

Classifier Audits

• Previous audits of commercial gender classifiers, including NIST'S Facial Recognition Vendor Test (FRVT) [3]

• FRVT [3] found classifiers performed worse on female input images, especially for female over 50

• Buolamwini and Gebru [1] conducted an intersectional audit of gender classifiers and found that these algorithms had the highest error rates for darker-skinned females

• Rhue [5] found that emotion detection classifiers tended to categorize Black male faces as more negative than their white counterparts

APPROACH

• FER classifiers have not been audited using intersectional inputs along the axes of phenotypic skin color and gender expression

• **Goal 1:** Evaluate the accuracy of 3 commercial emotion detection classifiers along intersectional inputs

• **Goal 2**: Manipulate phenotypic skin color and gender expression to isolate these factors

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%

Breakdown of the AffectNet testing set along the intersectionl subgroups

• General Strategy 1: Divide the inputs into 4 subgroups: lighter-skinned females (LF), lighter-skinned males (LM), darker-skinned females (DF), and darker-skinned males (DM)

• General Strategy 2: Conduct a series of experiments changing skin color and gender expression and comparing results to baseline

DATASET

• AffectNet dataset consists of 420K facial images that have been manually annotated for 8 different emotions

• Further annotated for phenotypic skin color using the Fitzpatrick Skin Type scale and perceived gender expression (binary scale)

Fitzpatrick Skin Type Scale

CLASSIFIER AUDIT

Baseline Testing Set

• Microsoft and Amazon return confidence scores-- the highest confidence score is the dominant emotion

• Google returns on a 5-point discrete scale with 5 indicating the dominant emotion

Experiments

Luminance Mode-Shift **Color Transfer**

• Luminance mode-shift and color transfer are used to manipulate phenotypic skin color

• The methods are used to create a lightened and darkened version of each testing set

• Images are cropped to the OpenCV boundary box dimensions as a means of removing hair length, a potential gender indicator

Cropped Images

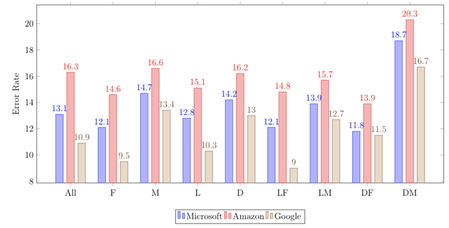
RESULTS

Baseline Results

• Classifiers have a skewed accuracy rate for different emotions, performing better when detecting happy and neutral emotions.

• Contrary to expectation, Microsoft and Amazon had lowest error rates on darker-skinned females

• All classifiers performed the worst on darker-skinned males (error rates between 16.7% and 20.3%)



Baseline Error Rates

• The misclassified darker-skinned male images tended to be higher on the Fitzpatrick Skin Type Scale

Average face of incorrectly predicted darker-skinned male

Experiment Results

• The change in accuracy for the experimental results compare to the baseline was evaluating using chi-square test and difference in proportions t-test

• Luminance mode-shift had a significant change in accuracy rates for Microsoft and Google in lightened images. However, this could be due an overall decrease in the number of images identified, not a change in accuracy rate

• Color transfer did not lead to significant changes in accuracy rates

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

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%	

One-tailed difference in proportions t-test for Amazon cropped results

CONCLUSIONS & FUTURE WORK

Conclusion

• Conducted an intersectional audit of 3 commercial classifiers (Microsoft, Amazon, and Google)

• Examined the isolated effects of manipulating phenotypic skin color and gender expression

• Emotion detection classifiers before better on female images than male images and better on lighter-skinned images than darker-skinned images

• Darker-skinned males have the highest error rate across all 3 commercial classifiers and all experiments

• Cannot say that one factor in isolation leads to discrepancies in accuracy rates, so it is important to have diverse training sets for these classifiers (not limited to just gender and racial diversity)

Future Work

• Look at "minimal sufficient explanation" or minimum facial features necessary in order for a classifier to make a prediction

Example of "minimal sufficient explanation" analysis from Muthukumar et al. [4]

• Imperative to continue auditing "black-box" commercial algorithms as a means of increasing accountability and transparency

REFERENCES

[1] J. Buolamwini and T. Gebru, "Gender shades: Intersectional accuracy disparities in commercialgender 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.

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

[3] 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.

[4] V. Muthukumar *et al.*, "Understanding unequal gender classification accuracy from face images," *CoRR*, vol. abs/1812.00099, 2018

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