Age Bias in Emotion Detection: *An Analysis of Facial Emotion Recognition Performance on Young, Middle-Aged, and Older Adults*

Eugenia Kim, De'Aira Bryant, Deepak Srikanth, and Ayanna Howard
Human-Automation Systems (HumAnS) Lab, School of Interactive Computing

Background & Motivation

The practical and known benefits of understanding emotion have translated into a desire for facial emotion recognition (FER) technology to be developed quickly without careful validation. As commercial FER systems are often embedded into complex applications that have use cases in fields ranging from education to security to healthcare, these algorithms are often used on an unaware public.

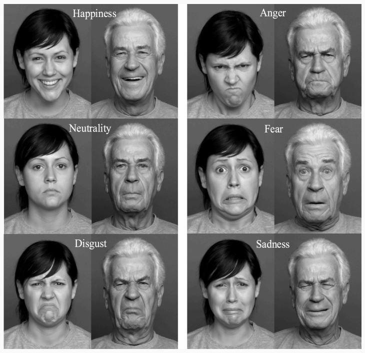
Additional research has revealed that FER systems perform differently within varying sectors of society. Commercially available gender classification software, which utilizes face detection algorithms, found darker-skinned females to be the most misclassified. Simultaneously, fair-skinned, male images showed lower error rates and higher accuracy when they considered groups that were not equally represented in training datasets (Buolamwini and Gebru 2018). However, demographic factors (i.e gender, race, age, etc.) have also been shown to impact emotion detection performance (Klare et al. 2012; Kyriakou et al. 2020).

Although bias is a known limitation of FER systems, research on both dataset changes and attribute-aware algorithms have not led to a generally accepted solution. Additionally, another approach to combat bias is using algorithmic auditing as a strategy to keep commercial systems accountable for their outputs. However, biases that receive less attention may become amplified in digital platforms. As corporate values and user interests influence commercial systems, Rosales and Fernández-Ardèvol showed that ageism is consistently ignored and limits older adults from using digital platforms.

These research findings reveal a gap in FER systems' bias mitigation efforts and the necessity of exposing ageism's effects when creating digital platforms widely in use today. This research highlights how the trend to address only exposed demographic biases is a band-aid solution for a gaping wound, particularly when target users of society belong to subgroups that are not mutually exclusive. It further provides a compelling call for inclusive, intersectional algorithmic developmental and benchmarking practices.

Methodology

A dataset with standardized images was used to evaluate the performance of four different commercial FER systems on three different age groups. The evaluation was conducted on algorithmic performance in 2019 and 2020.



The FACES database of facial expressions contains high quality color photographs of 171 adults, each displaying six different emotions: **anger, disgust, fear, happiness, neutrality, and sadness** (Ebner, Riediger, and Lindenberger 2010). The images are categorized by gender and into three age groups of young (ages 19 - 31), middle-aged (ages 39 - 55), and older adults (ages 69 - 80).

System Name	Identifiable Expressions	Confidence Output
Amazon Rekognition	Calm, Fear, Happy, Angry, Disgusted, Sad, Surprised, Confused	0-100
Face++	Neutral, Fear, Happiness, Anger, Disgust, Sadness, Surprise	0-100
Microsoft Face	Neutral, Fear, Happiness, Anger, Disgust, Sadness, Surprise, Contempt	0-1
Sighthound	Neutral, Fear, Happiness, Anger, Disgust, Sadness, Surprise	0-1

Four FER systems were selected from a review of currently available algorithms that included either an API or SDK. This enables each of the systems with the capability of being embedded into other technology, potentially affecting many diverse groups.

Procedure

Amazon Rekognition, Face++, Microsoft Face, and Sighthound were chosen for our analysis. These systems return confidence values as a percentage for each emotion, and the highest confidence value was recorded as the predicted emotion value.

For each FER system, we conduct a black box test as the explicit details of each algorithm are are not accessible to the public. The true emotion label of each image is stored as well as the equivalent prediction label per algorithm. For example, Amazon's 'happy' label maps to the FACES dataset's 'happiness' label. The different emotion recognition systems then process each image, and all outputs are normalized to probability values between 0.0 and 1.0. The maximum rule was applied to determine the predicted emotion label with the highest confidence and stored to compare to the dataset's true label with these probabilities.

Results

Using the dataset images, we analyze each facial emotion recognition system's outputs by classification accuracy, positive predictive values (PPV), and other standard performance evaluation metrics. Each result was also evaluated one year later with its updated algorithm to measure any documented changes. Some algorithms suggest improved performance for certain subgroups, specifically gender subgroups with their latest versions.

		Young	Middle-Aged	Older
Amazon	*2019*	0.86	0.81	0.68
	2020	0.89	0.85	0.68
Face++	*2019*	0.8	0.75	0.64
	2020	0.81	0.76	0.63
Microsoft	*2019*	0.88	0.76	0.65
	2020	0.88	0.76	0.65
Sighthound	*2019*	0.78	0.73	0.65
	2020	0.78	0.74	0.64

		Male	Female
Amazon	*2019*	0.78	0.79
	2020	0.81	0.8
Face++	*2019*	0.73	0.73
	2020	0.72	0.74
Microsoft	*2019*	0.76	0.78
	2020	0.76	0.77
Sighthound	*2019*	0.71	0.73
	2020	0.7	0.73

Classification accuracy, or the fraction of all correct predictions over the total inputs, was used to evaluate how often each algorithm was correct. The accuracy values for each system considering all images for each age group can be found in the above left table. Older adults had the lowest classification accuracy scores for each of the four assessed algorithms while young adults had the highest. While the margins were notably small, Microsoft performed best for images of young adults and Amazon performed best for images of middle-aged adults and older adults. Some systems highlighted improvements made to address gender disparities. To investigate whether any observed age group. disparities could be entangled with gender disparities, we also evaluated the results by gender subgroups show in the table to the right.

		Positive Predictive Value						
		Anger	*Disgust*	*Fear*	*Happiness*	*Neutral*	*Sadness*	***Average PPV***
Amazon	*Young*	96%	89%	88%	100%	95%	84%	***92%***
	Middle	90%	80%	92%	97%	89%	82%	***88%***
	Old	63%	85%	90%	89%	59%	76%	***77%***
Face++	*Young*	84%	86%	89%	93%	74%	84%	***85%***
	Middle	86%	74%	92%	90%	72%	68%	***80%***
	Old	62%	65%	86%	83%	53%	76%	***71%***
Microsoft	*Young*	100%	100%	100%	98%	78%	83%	***93%***
	Middle	88%	94%	97%	93%	62%	73%	***85%***
	Old	58%	92%	100%	88%	47%	73%	***76%***
Sighthound	*Young*	95%	70%	98%	100%	58%	75%	***83%***
	Middle	90%	66%	98%	98%	55%	65%	***79%***
	Old	69%	73%	100%	94%	41%	46%	***71%***

The PPV was calculated to show how trustworthy the recognition system is for perceiving each emotion. While the margins were again small, Microsoft had the highest average PPV for young adults while Amazon had the highest average for middle and older adults.

Discussion / Conclusion

We selected four commercial FER systems and found that each algorithm most accurately perceived emotion in images of young adults. Each system also produced the lowest classification accuracy and PPV scores when used to perceive expressive images of older adults.

Overall, the average PPV of all four recognition systems was highest for images of younger adults and lowest for images of older adults; however, each algorithm had varying PPV trends when recognizing fear within the different age groups. The dataset used was validated with human judges correctly identifying the emotions with at least 95% accuracy for the final set; however, all four algorithms could not accurately predict emotions for older adults beyond 68% mean accuracy.

Our results also showed that gender bias did not influence the notable differences in accuracy and PPV scores between the young and older adult subgroups.

As use cases for FER systems become relevant among larger sectors of the global population, developers of FER technology cannot continue to approach demographic biases retroactively. Our results demonstrate the importance of considering various demographic subgroups during FER system validation and the need for inclusive, intersectional algorithmic developmental practices.

Contact

Eugenia Kim
School of Interactive Computing
Georgia Institute of Technology
ekim317@gatech.edu

References

1. Buolamwini, J.; and Gebru, T. 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. Proceedings of Machine Learning Research 81 1–15. URL http://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf.
2. Klare, B. F.; Burge, M. J.; Klontz, J. C.; Bruegge, R. W. V.; and Jain, A. K. 2012. Face Recognition Performance: Role of Demographic Information. IEEE Transactions on Information Forensics and Security 7: 1789–1801. doi:10.1109/TIFS.2012.2214212. URL https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6327355.
3. Kyriakou, K.; Kleanthous, S.; Otterbacher, J.; and Papadopoulos, G. A. 2020. Emotion-Based Stereotypes in Image Analysis Services. In Adjunct Publication of the 28th ACM Conference on User Modeling, Adaptation and Personalization, UMAP '20 Adjunct, 252–259. New York, NY, USA: Association for Computing Machinery. ISBN 9781450379502. doi:10.1145/3386392.3399567. URL https://doi.org/10.1145/3386392.3399567.
4. Rosales, A.; and Fernandez-Ard´ evol, M. 2020. Ageism in the era` of digital platforms. Convergence: The International Journal of Research into New Media Technologies 26. ISSN 1354-8565. doi: 10.1177/1354856520930905.
5. Ebner, N. C.; Riediger, M.; and Lindenberger, U. 2010. FACES—A database of facial expressions in young, middle-aged, and older women and men: Development and validation. Behavior Research Methods 351–362. doi:10.3758/BRM.42.1.351. URL http://hdl.handle.net/11858/00-001M-0000-0013-3A21-0.