Investigating Bias in Facial Expression Recognition

Stefanie Stoppel and Simone Frintrop
Department of Informatics, University of Hamburg, Germany
{8stoppel,frintrop}@informatik.uni-hamburg.de

Motivation: While several studies have investigated demographic bias in fields such as facial recognition, little work on this topic has been done in the area of facial expression recognition (FER). This work aims at taking a first step in this direction by analyzing the annotation distributions of RAF-DB [4], an in-the-wild FER data set. We further assemble a benchmark data set approximately balanced by *expression*, *gender* and *race* labels by combining samples from two existing data sets. The result is then used to benchmark a state of the art FER classifier, thereby focusing specifically on intersectional accuracy disparities between subgroups. To conclude, we suggest directions for future research based on our findings.

Method: We create a benchmark from samples of both the RAF-DB [4] and ExpW [7] data sets, whereby the former contains demographic labels for *age*, *race* and *gender* in addition to categorical *expression* annotations. A distribution analysis of RAF-DB showed that some *expression* $\times$ *gender* $\times$ *race* $\times$ *age* subgroups contain no or very few samples. Consequently, the benchmark was set to contain 42 subgroups combined from 3 *race*, 2 *gender* and 7 *expression* categories. The target sample count per subgroup was set to 50. Next, images in RAF-DB were assigned into their respective subgroups. Due to the low resulting subgroup sample counts, ExpW images were annotated with demogaphic labels using the FairFace [3] model for facial attribute annotation. The annotated images were then manually filtered and assigned to their respective subgroup. In the final benchmark, 7 out of 42 subgroups (e.g. "Fear, Asian" and "Disgust, African-American" for both genders) contain less than 50 samples. We use ESR-9 [5], an ensemble model for FER, as base model for our evaluation of the benchmark.

Experimental results: ESR-9 achieved a mean classification accuracy of 46.5% on our benchmark. The accuracies of 47.1% for "Male" and 45.9% for "Female" gender categories, as well as the ones for race categories with 45.5% for "African-American", 48.0% for "Asian", and 46.0% for "Caucasian", were close to the overall mean. Conversely, we observed large differences across expression categories (see bottom row of Fig. 1), with the highest mean accuracy for "Anger" (87.3%), and the lowest for "Disgust" (5.1%). These divergences hint at inter data set bias, possibly resulting from differences in expression annotation practices.

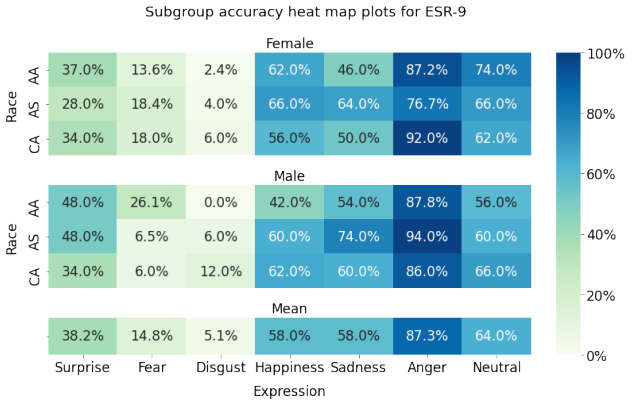

Figure 1. Subgroup accuracies of ESR-9 on the benchmark data set. AA = African-American; AS = Asian; CA = Caucasian;

Moreover, Fig. 1 highlights the intersectional accuracy disparities between subgroups within the same expression category, with the largest accuracy difference of 28% between the "Sadness, Asian, Male" (74.0%) and "Sadness, African-American, Female" (46.0%) subgroups.

Discussion and conclusion: We assembled an approximately balanced benchmark from samples of two existing FER data sets. Recent analyses [6] and our examination of RAF-DB found that the data set is highly skewed towards race label "Caucasian", as well as expression categories "Happiness" and "Neutral", with very few samples in "Disgust" and "Fear". While these facts plus the sparse data for some age groups prevented us from assembling a completely balanced benchmark, this work can be seen as starting point disclosing the need for more rigorous investigations. Finally, we argue that further research in adjacent fields is needed, e.g. regarding the reliability of mapping facial expressions to categorical emotions [1], as well as current questionable data set collection and curation practices [2].

References

[1] Lisa Feldman Barrett, Ralph Adolphs, Stacy Marsella, Aleix M Martinez, and Seth D Pollak. Emotional expressions reconsidered: Challenges to inferring emotion from human facial movements. *Psychological science in the public interest*, 20(1):1–68, 2019. 1

[2] Abeba Birhane and Vinay Uday Prabhu. Large Image Datasets: A Pyrrhic Win for Computer Vision? In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1537–1547, January 2021. 1

[3] Kimmo Karkkainen and Jungseock Joo. FairFace: Face Attribute Dataset for Balanced Race, Gender, and Age for Bias Measurement and Mitigation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1548–1558, 2021. 1

[4] Shan Li, Weihong Deng, and JunPing Du. Reliable Crowdsourcing and Deep Locality-Preserving Learning for Expression Recognition in the Wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 1

[5] Henrique Siqueira, Sven Magg, and Stefan Wermter. Efficient Facial Feature Learning with Wide Ensemble-Based Convolutional Neural Networks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(04):5800–5809, 2020. 1

[6] Tian Xu, Jennifer White, Sinan Kalkan, and Hatice Gunes. Investigating Bias and Fairness in Facial Expression Recognition. In Adrien Bartoli and Andrea Fusiello, editors, *Computer Vision – ECCV 2020 Workshops*, pages 506–523. Springer International Publishing, 2020. 1

[7] Zhanpeng Zhang, Ping Luo, Chen Change Loy, and Xiaoou Tang. From Facial Expression Recognition to Interpersonal Relation Prediction. *International Journal of Computer Vision*, 126(5):550–569, 2018. 1