Investigating Bias in Facial Expression Recognition

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

MOTIVATION

- Investigate bias in facial expression recognition (FER)
- Analyze existing in-the-wild FER data sets RAF-DB[1] and ExpW[2]
- Assemble benchmark approximately balanced by *gender*, *race* and *expression* using samples from both ExpW and RAF-DB
- Analyze intersectional accuracy disparities in benchmark performance using FER ensemble classifier ESR-9[3]

RAF-DB DISTRIBUTION ANALYSIS

- 29.672 images from Flickr annotated with *expression, age*, *race*, and *gender*

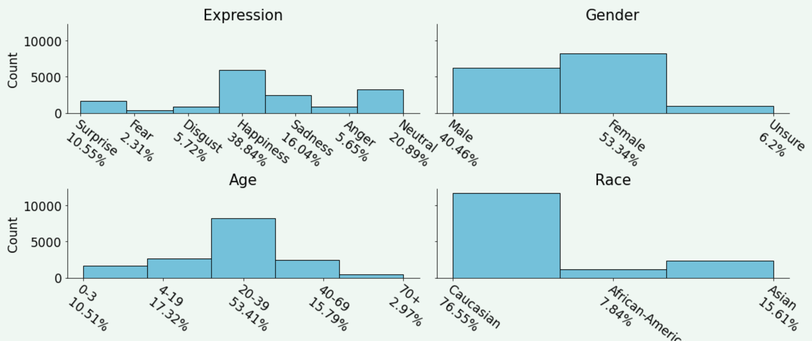

Figure 1) Histograms of RAF-DB annotations

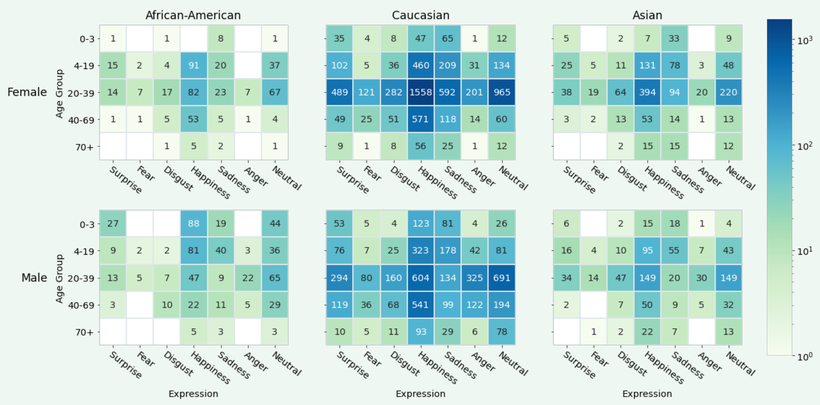

Figure 2) Heat map plots for RAF-DB subgroup sample counts

- RAF-DB highly skewed in the *race*, *age* and *expression* categories (see Fig. 1)
- Intersectional analysis in Fig. 2 reveals…
  - Some subgroups contain no images
  - *Race* category Caucasian and middle *age* group are overrepresented

EXPW DISTRIBUTION ANALYSIS

- 91.793 pictures from image search engine with 7 categorical expression annotations
- We obtained demographic labels automatically using the FairFace[4] attribute classification model

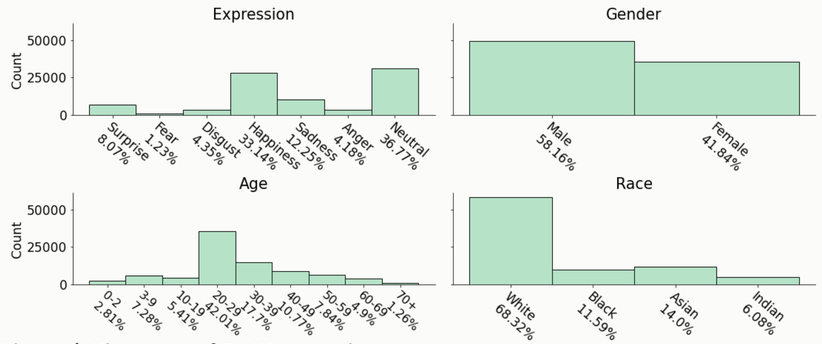

Figure 3) Histograms of ExpW annotations

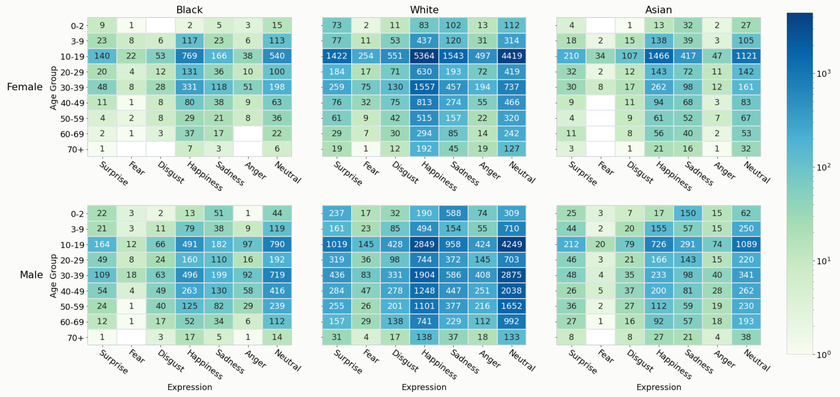

Figure 4) Heat map plots for ExpW subgroup sample counts

BENCHMARK DATA SET CREATION

- 42 subgroups from 3 *race*, 2 *gender* and 7 *expression* categories
- Target sample count of 50 images per subgroup
- Benchmark creation:
  1. RAF-DB: Filter & assign samples to demographic subgroups
  2. ExpW: Annotate images with demographic labels using the FairFace model
  3. Annotated ExpW: Filter & assign samples to demographic subgroups

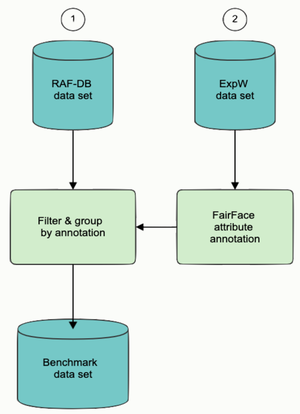

Figure 5) Benchmark creation

RESULTS

- Classification results obtained with ESR-9 (8 *expression* labels)
- Confusion matrix (Fig. 6) shows *Disgust* and *Fear* were often misclassified
- Heat map plot (Fig. 7) displays subgroup accuracies:
  - Overall bad performance for *Fear* and *Disgust*
  - *Happiness* and *Sadness* have highest discrepancy between subgroups (see also Table 1)

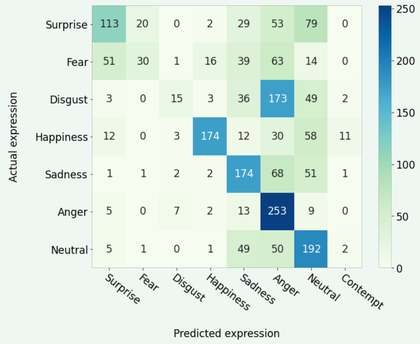

Figure 6) ESR-9 confusion matrix

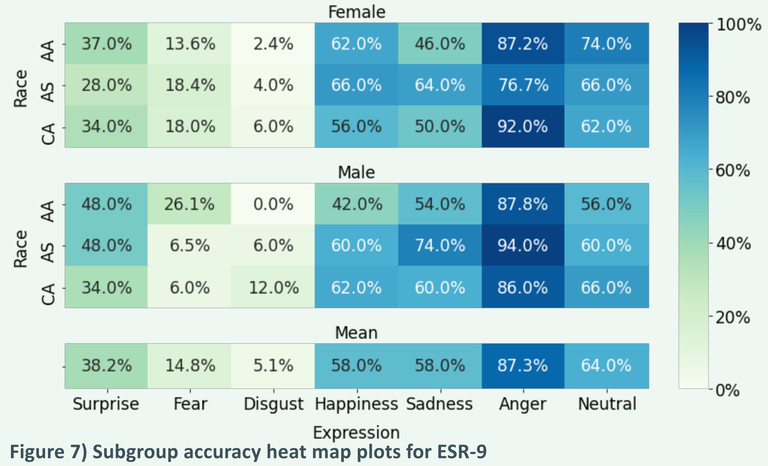

Figure 7) Subgroup accuracy heat map plots for ESR-9

Surprise	Fear	Disgust	Happiness	Sadness	Anger	Neutral
20.0%	20.1%	12.0%	24.0%	28.0%	17.3%	18.0%

Table 1) Largest accuracy differences between subgroups in one *expression* category

CONCLUSION

- We analyzed two existing FER data sets and used them as a basis for our benchmark approximately balanced by *gender*, *race* and *expression*
- Our findings suggest in-the-wild FER data set distributions are highly skewed
- ESR-9's benchmark performance showed large discrepancies between subgroups within the same *expression* category
- We encourage more rigorous investigations into bias in FER

- [1] Shan Li, Weihong Deng, and JunPing Du. Reliable Crowd- sourcing 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.
- [2] 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.
- [3] 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.
- [4] Kimmo Kärkkäinen 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.