Representational bias in expression and annotation of emotions in audiovisual databases

William Saakyan[1]*, Olya Hakobyan[2]*, Hanna Drimalla[3]
{wsaakyan@uni-bielefeld.de[1], ohakobyan@uni-bielefeld.de[2], drimalla@techfak.uni-bielefeld.de[3] }
Center for Cognitive Interaction Technology (CITEC), Bielefeld University, Germany[1 2 3]

Abstract. Emotion recognition models can be confounded by representation bias, where populations of certain gender, age or ethnoracial characteristics are not sufficiently represented in the training data. This may result in erroneous predictions with consequences of personal relevance in sensitive contexts. We systematically examined 130 emotion (audio, visual and audio-visual) datasets and found that age and ethnoracial background are the most affected dimensions, while gender is largely balanced in emotion datasets. The observed disparities between age and ethnoracial groups are compounded by scarce and inconsistent reports of demographic information. Finally, we observed a lack of information about the annotators of emotion datasets, another potential source of bias.

Keywords: datasets, emotion recognition, machine learning, bias

1 Introduction

Automated emotion recognition is used in a variety of applications such as personal assistants that are responsive to users' emotions [1], hiring tools that assess applicants' emotional state during the interview process [2] and even tools for psychiatric assessment [3]. Such uses of automated emotion recognition are subject to considerations of fairness as some populations (e.g. of certain gender, age or ethnoracial background) may not be well-represented in the training dataset, potentially leading to erroneous predictions and compromising the opportunities of these people [4, 5, 6]. While representational bias is a well-known issue in the machine learning literature [7, 8], its impact on

*These authors contributed equally

automated decisions also depends on domain-specific confounds related to the different groups. In the context of emotion recognition, one reason for the relevance of balanced representation is the evidence that emotion *expression* might be different between populations. For instance, there are some reports of varying emotion expressivity (on average) in women vs men [9, 10], young vs old [11, 12] and cross-cultural or cross-racial comparisons [13]. This means that over-represented populations might bias the models leading to less accurate performance for the under-represented groups. Furthermore, even under the assumption that the differences in emotion expression are small between the aforementioned groups, the physical characteristics associated with some populations might be identified by machine learning models and bias their performance. For instance, facial analysis tools have been shown to perform better for men and White people and worse for women and people with darker skin [4, 5, 6]. Similarly, depending on the application, facial analysis tools perform differently for people of older [5] or younger age [6]. In addition to expression, there is evidence that emotion *recognition* may be affected by demographic characteristics. For example, women might have higher sensitivity for recognizing emotions at least under some circumstances (for a review, see [14] but also [15]). Another example is decreased sensitivity to negative emotions in older people [16, 11]. Further, there is evidence that emotion recognition is more accurate for people that are perceived as members of one's own ethnoracial group [17, 18]. As human annotations are often used as the ground truth, the aforementioned demographic characteristics of these annotators may have a crucial, and yet, under-appreciated role in the accuracy of emotion recognition algorithms. Hence, sensitive applications, such as emotion recognition, should address representational bias in the data of emotion expressors and annotators to ensure similar model performance for all target groups. In this paper, we explore the landscape of emotion datasets with respect to representational bias (gender, age, ethnoracial background) considering datasets of audio, visual and audio-visual modalities.

2 Methods

To identify relevant emotion datasets, we adapted established guidelines for conducting systematic reviews [19] (see the detailed protocol in Appendix). In brief, we conducted a search on the Web of Science and PubMed platforms using the search query *(emotion OR affect) AND (recognition OR detection) AND (dataset OR database OR corpus) AND (video OR image OR audiovisual OR audio OR speech)*. We then scanned both review papers (full-text) and research articles (abstracts) to identify emotion datasets containing visual, audio or audio-visual data. We only searched datasets that were labeled with either emotion categories, affect dimensions or emotion-related facial action units, i.e., smallest visually discernible facial movements [20]. For the review papers, the relevance of each paper was assessed by two researchers independently and was discussed until an agreement rate of 90% was achieved (calculated automatically based on researcher annotations). The relevance was determined based on the review's likeliness to discuss datasets of the pre-defined inclusion and exclusion criteria (see the detailed protocol in Appendix). The full texts of the relevant reviews were then scanned for datasets. To complement this approach, we also developed a tool for automatic

parsing of the abstracts in the research articles of the search query. In particular, we took advantage of the fact that most dataset names contained at least two uppercase letters and were written in brackets following their full-phrase explanations. Hence, the automatic parsing tool extracted every sentence from an abstract with words that 1) contained at least two uppercase letters and 2) were written in brackets. Some of the common capitalized abbreviations that were not dataset names (e.g., EEG, ECG, etc.) were included in a blacklist (see Appendix). After filtering out words with less than four occurrences (88th percentile) across the scanned papers, we inspected the sentences associated with each term (containing the uppercase letters) and extracted all dataset names. The final list of the datasets is available in the dedicated repository [21]. Next, we inspected the publications associated with the datasets chosen in the previous step and extracted bias-relevant information, such as the modality of the datasets (audio, visual, audio-visual), sample size, age, gender and ethnoracial background of the participants and the annotators/raters (henceforth annotators for simplicity).

3 Results

The data extracted in the systematic review process was analyzed to identify potential biases in gender, age and ethnoracial distributions. Importantly, these characteristics were analyzed for both the participants from whom the data were acquired and annotators, whenever the information was available. In total, we have extracted information from 130 datasets published in the time period of $1996-2020$. 55 of the datasets contained purely audio, 48 of the datasets contained purely visual data, while 27 datasets were comprised of audio-visual data. The number of datasets more than doubled at the turn of the century with increasing involvement of visual and audio-visual data and a decrease of purely audio data in the following decades (Fig.A1a). There was high variability in the sample size (images, video or audio snippets) of the datasets (Fig.A1b) and although less datasets were created more recently ($2010-2014$ vs $2015-2019$), the dataset sizes are larger in the latter time period (Fig.A1b). 61 datasets shared information on whether the dataset was validated or annotated by self-ratings ($N = 7$), external annotators ($N = 47$) or both ($N = 7$).

From 130 datasets, 94 datasets reported quantitative gender information for participants and 18 for annotators, all following a binary gender classification. To investigate gender bias in emotion datasets, we calculated the male ratio by dividing the number of male participants by the sum of both genders. Hence, a perfectly balanced dataset would score 0.5 by this metric, while 1 denotes exclusively male and 0 exclusively female datasets. Most datasets exhibited values close to 0.5 (Fig.1a, Fig.1b), indicating high share of balanced datasets for both participants and annotators. To test whether the normality assumption applies for these distributions, we inspected the histograms and performed the D'Agostino-Pearson normality test recommended for distributions containing several identical values [22] and observed a slight violation of normality for participants (K^2 = 6.25, p = 0.04) and none for annotators (K^2 = 5.52, p = 0.06). Accordingly, a one-sample signed-rank Wilcoxon test for gender ratio was applied for the participants, showing no evidence that the median of the male ratio was statistically different from the expected value of 0.5 (participants: T = 1111.00, p = 0.70). Similarly, a one-sample t-test did not show significant difference from a

mean of 0.5 for the annotators ($t = 0.07$, $p = 0.93$). To see if the male ratio changes over time, we calculated the Spearman's correlation between publication years and the male ratio metric. There was no significant correlation for either group (participants: $r_s = 0.18$, $p = 0.06$; annotators: $r_s = 0.03$, $p = 0.89$), providing no evidence for strong systematic shifts in terms of gender bias over time (Fig.A1c).

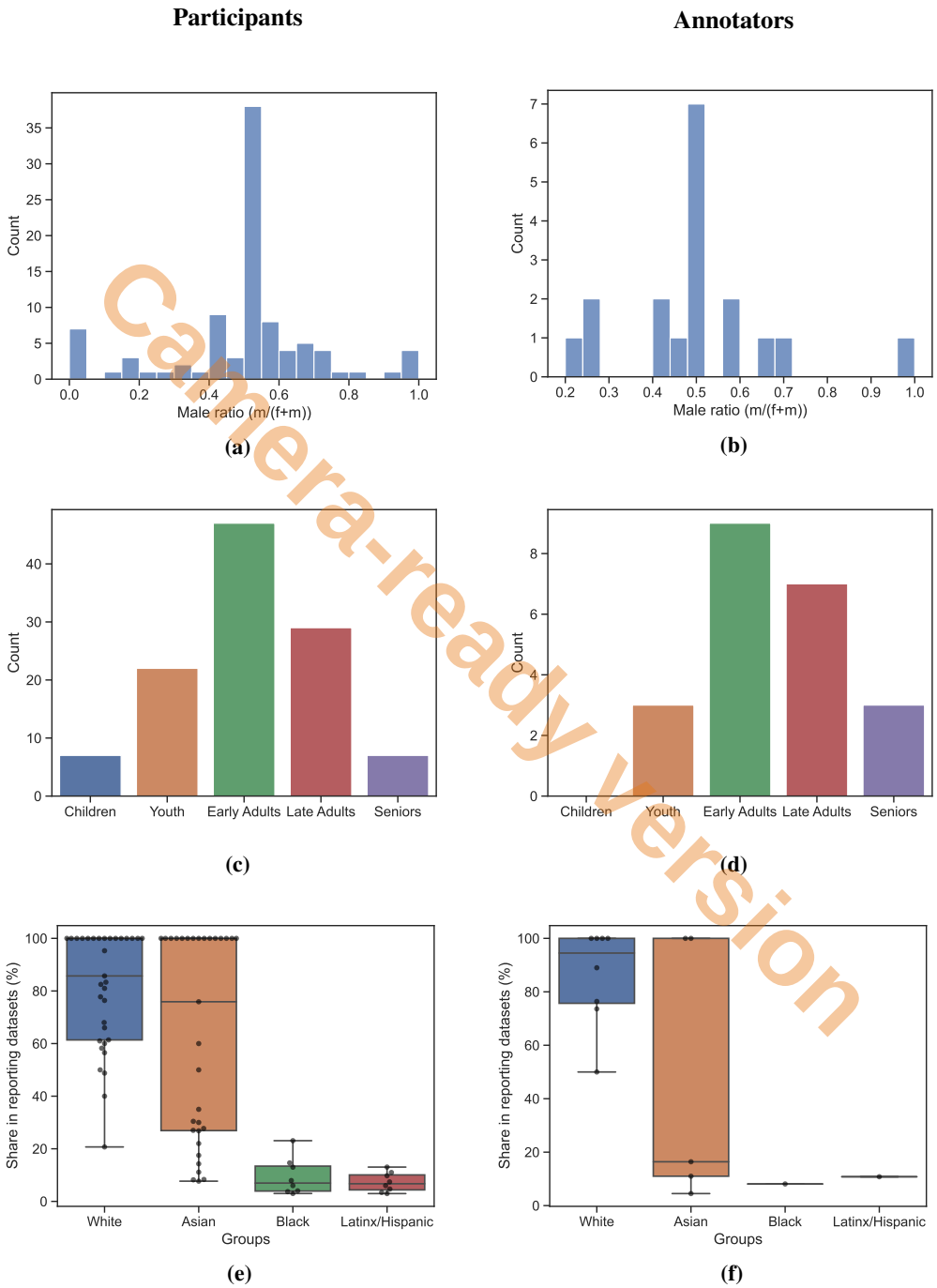


Fig. 1. Representational bias in emotion datasets. The male ratio is close to 0.5 for participants (a) and annotators (b) indicating gender balance. Children and older adults are the least represented group in both participant (c) and annotator (d) data. (e-f) Most datasets represent White people as a majority. Asians are well-represented in some and poorly represented in other datasets, while Black and Latinx/Hispanic people are consistently under-represented.

Quantitative information about the age range was reported in less than half of the datasets for the participants ($N = 49$) and even less frequently for the annotators ($N = 9$). Datasets typically mentioned the age range of the involved people (varying between datasets) and did not specify the underlying distributions, impeding a direct comparison. To overcome this issue, we defined age groups as children [0-12), youth [12-20), early adults [20-40), late adults [40-65) and seniors (65+), following established age group definitions [23]. For each age group, we calculated the number of datasets, in which it was represented. For instance, if a dataset reported the range 19-48, the age groups youth, early adults and late adults each received one count. As shown in Fig.1c and Fig.1d, early adults were the most represented groups both among participants and annotators, while fewer datasets contained data from children and senior adults. The minima (participants: M= 19.96, SD = 8.92; annotators: M= 20.77, SD = 3.49) and maxima (participants: M= 44.56, SD = 17.22; annotators: M= 57.11, SD = 20.14) of the age ranges also pointed to a bias towards younger individuals. We also calculated the span of each dataset, showing that on average, the datasets cover around two to three decades of age (participants: M= 24.08, SD = 17.66; annotators: M= 36.33, SD = 21.51). As shown in Fig.A1d and confirmed by Spearman's correlation (r_s = 0.09, p = 0.49), no strong systematic changes of age span were observed over time for the participants. This test was omitted for the annotators due to the scarceness of age reports.

Finally, only 53 datasets provided any information (quantitative, e.g. 60% Caucasian or qualitative, e.g. "mostly Caucasian") about the ethnicity or race of the participants and only 13 reported this information about the annotators. We combined the ethnoracial information into four large groups: White, Black, Asian and Latinx/Hispanic, following accepted group naming conventions from [24]. We are aware that such classifications are extreme oversimplifications of human ethnic and racial diversity, however, the labels used in the datasets themselves (e.g. Caucasian, Asian etc) were so vague that a broad classification was required to accommodate all reports. Fig.1e shows the percentage that each group comprised in datasets that reported the ethnoracial makeup *quantitatively* (participants: N = 46, annotators N = 10). Most common groups are White and Asian. They are not only represented in most datasets reporting ethnoracial characteristics (Whites in 33 and Asians in 31 datasets), but also their share is high (White: M= 80.99%, SD = 22.17; Asian: M= 62.97% , SD = 38.96). The White group is the majority in 84.38% of datasets in which it was represented. Black and Latinx/Hispanic people, by contrast, are represented in only 8 datasets and they are always in minority (Black: M= 9.42%, SD = 7.01; Latinx/Hispanic: M= 7.28%, SD = 3.68). A similar trend was also observed for the ethnoracial distribution of the annotators. Black and Latinx/Hispanic people were only represented in 1 of the 13 datasets, each, while White and Asian people were represented in 8 and 5 datasets, respectively. For annotators, too, White people were always in the majority with the exception of one dataset, where White people comprised 50% and the ethnoracial information of the other half was not specified. For speech datasets, we also examined the languages represented and observed that 42 out of 60 reported languages were of Indo-European origin, 11 were denoted as Mandarin or Chinese, there were 2 instances of Japonic and Semitic, each, 3 were instances of Turkic and Dravidian languages, each, and only 1 Uralic language.

4 Discussion

We analyzed 130 emotion recognition datasets with respect to representational bias in gender, age and ethnoracial background. Our findings show that emotion datasets do not seem to suffer from a prominent gender bias, while age and ethnoracial background exhibit large disparities. With respect to age, most datasets cover a narrow age span, mostly including younger people and underrepresenting children and older adults. In terms of ethnoracial background, White people are the most widely represented group, followed by Asians, while Black people and Latinx/Hispanic people are severely under-represented. Both age and ethnoracial imbalance confirm earlier reports from other domains [5, 25]. Despite most emotion datasets being gender-balanced, models of emotion recognition may not be entirely bias-free with respect to gender. Specifically, bias can be propagated from earlier stages of emotion detection pipelines as they often include automatic face detection, an area in which gender imbalances are a known issue [4, 5, 6]. Hence, in addition to circumventing representational bias in emotion databases, it is recommended to employ established tests of algorithmic fairness [26, 27, 28].

One critical issue in bias assessment is the frequency of reported demographic information as all three demographic characteristics were omitted for a considerable number of datasets. Intriguingly, age and ethnoracial background, the characteristics with the highest representational disparity, were the most poorly reported. Furthermore, datasets which did report demographic information lacked consistent and unambiguous descriptions. For instance, most reports on age included either an average value or the age range instead of the distribution itself. Similarly, there was no standard classification for ethnoracial information, often employing controversial terms (e.g. Caucasian) or masking the ethnoracial diversity in broad categories. The observed scarceness and insufficient detail of age and ethnoracial reporting point to an undervalued role of these characteristics in emotion recognition models. This is especially problematic for applications in sensitive domains, such as psychiatric assessment, in which ageism and ethnoracial bias have been reported [29, 25]. Naturally, diversity in the representation of groups *per dataset* need not necessarily be a desired feature in all applications, as some products may be targeted to specific groups only (e.g. a local product with regional importance). Nevertheless, efforts should be made towards ensuring that emotion datasets *collectively* cover all groups at least for applications targeted to broader populations (e.g. personal assistants in a medical context).

Importantly, the vast majority of datasets do not report demographic information for data annotators. Human annotations serve as the ground truth in many models. Therefore, demographic information on annotators is needed as it might influence the emotion recognition accuracy. The relation between annotators and the people expressing emotions, influenced by attitudes or group membership [18, 30, 31], for example, may further confound the ground truth ratings. This overlooked aspect of representational bias deserves further consideration, as human biases can be cemented and amplified in AI applications with critical implications, ranging from hiring decisions to mental health assessment. While representational bias in the data of participants may be handled with existing debiasing techniques [32], issues stemming from annotation bias may require different solutions.

Based on our findings, we identified the following recommendations for future emotion database collection and use. First, data collectors should aim for diverse datasets, especially by including more older participants and non-White people. Second, datasets should carefully report demographic information of the people expressing the emotions. Third, much more information is needed about the individuals annotating the emotions and the possible confounds that influence their emotion recognition. These strategies would allow dataset users to choose appropriate datasets matching their target group and implement suiting debiasing methods, resulting in more fair emotion recognition.

Acknowledgements

We would like to thank Tatjana Maskaljunas and Bhargav Acharya for their assistance in the data extraction process. Furthermore, we thank Dr. David Johnson, Matthias Norden, Alina Deriyeva and Steffen Johannknecht for stimulating discussions and their valuable feedback.

This study was funded by the Federal Ministry of Education and Research Germany (BMBF; 01IS20046). The BMBF had no involvement in the study design, the collection, analysis and interpretation of data, the writing of the report and the decision to submit the article for publication.

References

[1] Castillo JC, Castro-González Á, Alonso-Martín F, Fernández-Caballero A, Salichs MÁ. Emotion Detection and Regulation from Personal Assistant Robot in Smart Environment. In: Costa A, Julian V, Novais P, editors. Personal Assistants: Emerging Computational Technologies. Intelligent Systems Reference Library. Cham: Springer International Publishing; 2018. p. 179–195.

[2] Adepu Y, Boga VR, U S. Interviewee Performance Analyzer Using Facial Emotion Recognition and Speech Fluency Recognition. In: 2020 IEEE International Conference for Innovation in Technology (INOCON); 2020. p. 1–5.

[3] Harati S, Crowell A, Mayberg H, Nemati S. Depression Severity Classification from Speech Emotion. In: 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC); 2018. p. 5763–5766.

[4] Buolamwini J, Gebru T. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In: Proceedings of the 1st Conference on Fairness, Accountability and Transparency. PMLR; 2018. p. 77–91.

[5] Dooley S, Goldstein T, Dickerson JP. Robustness Disparities in Commercial Face Detection. arXiv:210812508 [cs]. 2021 Aug.

[6] Klare BF, Burge MJ, Klontz JC, Vorder Bruegge RW, Jain AK. Face Recognition Performance: Role of Demographic Information. IEEE Transactions on Information Forensics and Security. 2012 Dec;7(6):1789–1801.

[7] Suresh H, Guttag JV. A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle. arXiv:190110002 [cs, stat]. 2021 Jun.

[8] Mehrabi N, Morstatter F, Saxena N, Lerman K, Galstyan A. A Survey on Bias and Fairness in Machine Learning. arXiv:190809635 [cs]. 2019 Sep.

[9] Deng Y, Chang L, Yang M, Huo M, Zhou R. Gender Differences in Emotional Response: Inconsistency between Experience and Expressivity. PLOS ONE. 2016 Jun;11(6):e0158666.

[10] Chaplin TM. Gender and Emotion Expression: A Developmental Contextual Perspective. Emotion review : journal of the International Society for Research on Emotion. 2015 Jan;7(1):14–21.

[11] Riediger M, Voelkle MC, Ebner NC, Lindenberger U. Beyond "Happy, Angry, or Sad?": Age-of-Poser and Age-of-Rater Effects on Multi-Dimensional Emotion Perception. Cognition and Emotion. 2011 Sep;25(6):968–982.

[12] Fölster M, Hess U, Werheid K. Facial Age Affects Emotional Expression Decoding. Frontiers in Psychology. 2014;5:30.

[13] de Gelder B, Huis in 't Veld E. Cultural Differences in Emotional Expressions and Body Language. In: The Oxford Handbook of Cultural Neuroscience. Oxford Library of Psychology. New York, NY, US: Oxford University Press; 2016. p. 223–234.

[14] Forni-Santos L, Osório FL. Influence of Gender in the Recognition of Basic Facial Expressions: A Critical Literature Review. World Journal of Psychiatry. 2015 Sep;5(3):342–351.

[15] Fischer AH, Kret ME, Broekens J. Gender Differences in Emotion Perception and Self-Reported Emotional Intelligence: A Test of the Emotion Sensitivity Hypothesis. PLOS ONE. 2018 Jan;13(1):e0190712.

[16] Di Domenico A, Palumbo R, Mammarella N, Fairfield B. Aging and Emotional Expressions: Is There a Positivity Bias during Dynamic Emotion Recognition? Frontiers in Psychology. 2015;6:1130.

[17] Kang SM, Lau AS. Revisiting the Out-Group Advantage in Emotion Recognition in a Multicultural Society: Further Evidence for the in-Group Advantage. Emotion (Washington, DC). 2013 Apr;13(2):203–215.

[18] Elfenbein HA, Ambady N. On the Universality and Cultural Specificity of Emotion Recognition: A Meta-Analysis. Psychological Bulletin. 2002 Mar;128(2):203–235.

[19] Moher D, Liberati A, Tetzlaff J, Altman DG, Group TP. Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. PLOS Medicine. 2009 Jul;6(7):e1000097.

[20] Ekman P, Friesen W. Facial Action Coding System: A Technique for the Measurement of Facial Movement. undefined. 1978.

[21] Saakyan W, Hakobyan O, Drimalla H. Assessment of Representational Bias in Emotion Datasets; 2021. DOI: 10.5281/zenodo.5578290, https://github.com/mbp-lab/caip2021_bias_emotions.

[22] Motulsky HJ. Choosing a Normality Test; 2016. https://www.graphpad.com/guides/prism/latest/statistics/stat_choosing_a_normality_test.htm.

[23] Sigelman C, Rider E. Life-Span Human Development. Cengage Learning; 2008.

[24] Association AP. Publication Manual of the American Psychological Association 2020: The Official Guide to APA Style. Seventh ed. American Psychological Association; 2019.

[25] Hitczenko K, Cowan HR, Goldrick M, Mittal VA. Racial and Ethnic Biases in Computational Approaches to Psychopathology. Schizophrenia Bulletin. 2021 Nov:sbab131.

[26] Bellamy RKE, Dey K, Hind M, Hoffman SC, Houde S, Kannan K, et al. AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias. arXiv:181001943 [cs]. 2018 Oct.

[27] Saleiro P, Kuester B, Hinkson L, London J, Stevens A, Anisfeld A, et al. Aequitas: A Bias and Fairness Audit Toolkit. arXiv:181105577 [cs]. 2019 Apr.

[28] Bird S, Dudík M, Edgar R, Horn B, Lutz R, Milan V, et al. Fairlearn: A Toolkit for Assessing and Improving Fairness in AI. 2020 May.

[29] Bodner E, Palgi Y, Wyman MF. Ageism in Mental Health Assessment and Treatment of Older Adults. In: Ayalon L, Tesch-Römer C, editors. Contemporary Perspectives on Ageism. International Perspectives on Aging. Cham: Springer International Publishing; 2018. p. 241–262.

[30] Elfenbein HA. In-Group Advantage and Other-Group Bias in Facial Emotion Recognition. In: Understanding Facial Expressions in Communication: Cross-Cultural and Multidisciplinary Perspectives. New York, NY, US: Springer Science + Business Media; 2015. p. 57–71.

[31] Chronaki G, Wigelsworth M, Pell MD, Kotz SA. The Development of Cross-Cultural Recognition of Vocal Emotion during Childhood and Adolescence. Scientific Reports. 2018 Jun;8(1):8659.

[32] Li Y, Vasconcelos N. REPAIR: Removing Representation Bias by Dataset Resampling. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2019. p. 9564–9573.

A Appendix

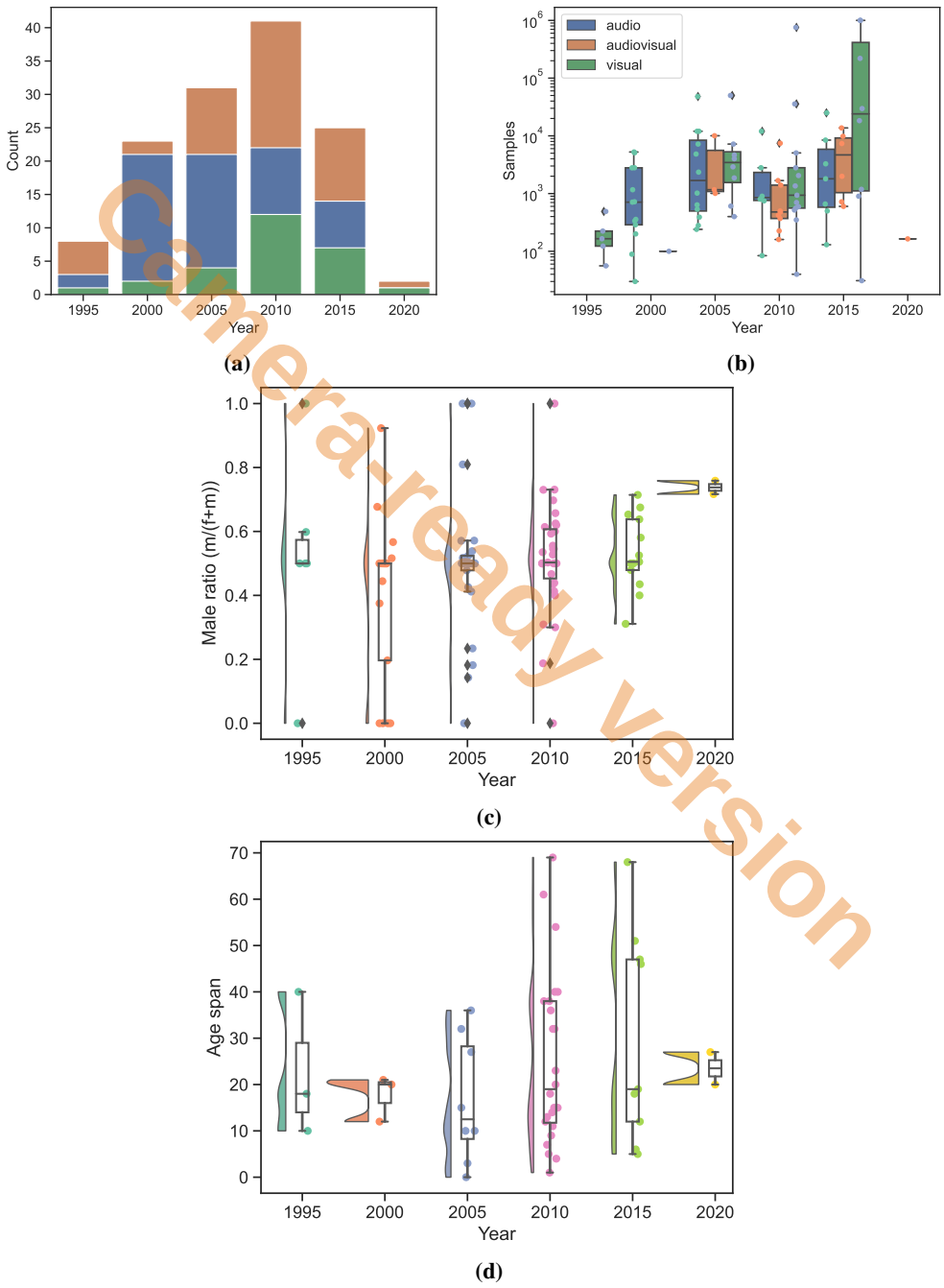


Fig. A1. (a) Datasets count for different modalities over time. (b): Datasets sample size over time for different modalities. (c) Participant male ratio over 5-year time periods. (d) Participant age span (max age - min age) over 5-year time periods.

Review protocol

Study question

How biased are the existing emotion recognition datasets with respect to gender, age and ethnicity?

Information sources for reviewed papers

Included: Web of Science (WoS), PubMed
Reasons: high compatibility of search and export options, wide coverage of scientific papers (~33000 Journals in each platform)

Excluded: Google Scholar, Research Gate
Reasons: no full export options

Search query

(emotion OR affect) AND (recognition OR detection) AND (dataset OR database OR corpus) AND (video OR image OR audiovisual OR audio OR speech)

Filters

Language: English
Fields:
WoS: Title, Abstract, Author Keywords, Keyword+
PubMed: all fields as the selection of multiple fields was not possible. PubMed fields included (Affiliation, Author, Author-Corporate, Author-First, Author-Identifier, Author-Last, Book, Conflict of Interest Statements, Date-Completion, Date-Create, Date-Entry, Date-MeSH, Date-Modification, Date-Publication, EC/RN Number, Editor, Filter, Grant Number, ISBN, Investigator, Issue, Journal, Language, Location ID, MeSH Major Topic, MeSH Subheading, MeSH Terms, Other Term, Pagination, Pharmacological Action, Publication Type, Publisher, Secondary Source ID, Subject - Personal Name, Supplementary Concept, Text Word, Title, Title/Abstract, Transliterated Title, Volume)

Dataset selection strategy

Our goal was to identify as many emotion/affect datasets as possible fulfilling the following criteria:

Inclusion criteria

Modality: audio, video, audio-visual
Labeling: single or multiple emotions/affect dimensions, emotion-related facial action units, i.e., smallest visible movements in the face (Ekman and Friesen 1978). Datasets with indirect estimates of affect (e.g. smiling as proxy of positive emotions) were also included.

Exclusion criteria

Modality: physiology data (EEG, ECG, fMRI etc), body movements without face information
Labeling: datasets focusing on emotion-related phenomena, such as pain or stress.

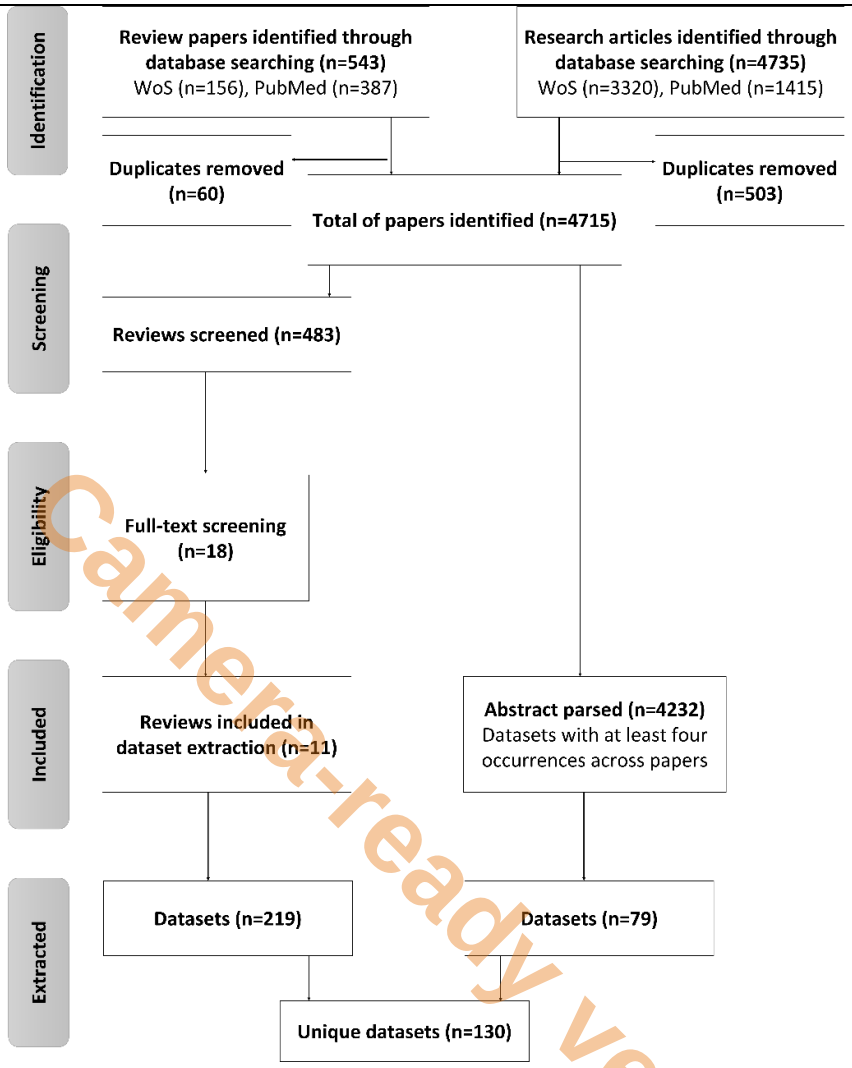
The results of the search on WoS and PubMed were exported in excel and csv files, respectively and combined into a single excel file. The review and research papers were saved into separate files.

For the review papers, the relevance of each paper was assessed by two researchers independently and was discussed until an agreement rate of 90% was achieved (calculated automatically based on researcher annotations). The relevance was determined based on the review's likeliness to discuss datasets of the pre-defined inclusion and exclusion criteria. The full texts of the relevant reviews were then scanned for datasets.

To complement this approach, we also developed a tool for automatic parsing of the abstracts in the research articles of the search query. In particular, we took advantage of the fact that most dataset names contained at least two uppercase letters and were written in brackets following their full-phrase explanations. Hence, we developed an automated parsing tool, which extracts every sentence from an abstract with words that 1) contained at least two uppercase letters and 2) were written in brackets. Some of the common capitalized abbreviations that were not dataset names (e.g., EEG, ECG, etc.) were included in a blacklist (see next). After filtering out words with less than four occurrences (88th percentile) across the scanned papers, we inspected the sentences associated with each term (containing the uppercase letters) and extracted all dataset names.

Blacklist

eeg, audiovisual, ser, fmri, cnn, svm, bovw, auc, hci, hog, swift, uk, usa, microexpression, ii, hrv, ecg, hmm, iot, youtube, imdb, softmax, vgg, ai, ml, dnn, dcnn, pca, lda, sift, mri, mfcc, cnns, pubmed, rgb, mlp, covid, methods, results, resnet, roi, knn, unet, shortterm, background, matlab, tv, iou, cca, bow, rnn, gru, fps, eegbased, densenet, svms, objective, db, ann, materials, melfrequency, lidar, mrmri, knearest, violajones, fscore, inceptionv, ssd, gan, bv, cnnbased, yolo, ct, lstm, conclusion, rgbd, alexnet, hmms, mfccs, conclusions, photooptical, adaboost, dl, lbp, dnns, ieee, dr, rcnn, rmse, roc, nlp, yolov, elm, cad, ci, ad, pd, au, modis, dbn, rois, spie, rf, ir, wm, md, asd, map, ga, fau, pet, cc, snr, ar, cd, lr, dti, gmm, ict, fp, eer, mes, me, er, si, hr, nir, surf, glcm, sar, asr, mr, aus, fa, tbi, ms, amd, dwt, mci, sd, oct, dct, us, uav, tm



PRISMA flow chart for the systematic review detailing the database searches, the number of abstracts and full texts screened, and the datasets extracted.

Information extraction from the datasets

In the next step, we searched the descriptions of the identified datasets in the associated publications. The information source for each dataset is documented and can be viewed in our dataset repository. Two researchers and two research assistants participated in the information extraction process using clearly defined guidelines as follows:

Dataset: str, dataset name
Authors: str, authors of the dataset, can be first author et al.
Link: link where you get the information, paper (preferable) or dataset…
Year: int, XXXX, year of the paper publication (preferable) or dataset
Notes: str, optional
Voice: boolean, 1 or 0, is there speech/voice modality or not

Image: boolean, 1 or 0, is there image modality or not. Of course videos consist from images, by definition however if there is video modality image should be coded with 0 UNLESS the videos are short image sequences leading to the emotion (for example, participants are asked to gradually make an angry face)
Video: boolean, 1 or 0, is there video modality or not
Other modalities: str, optional (if easy to extract), delimiter is ‘,’. Example: EEG, EDA, Gestures
Samples: int, number of videos, images, utterances, etc.
Setting: str, lab or wild, dropdown
Setting type: str, posing(make a sad face, raise eyebrows); acting(simulate emotions in context); spontaneous; mix (specify in notes); induced (E.g. participants were recorded when viewing/processing emotional material); dropdown
Posed: str, what they are posing? Emotions, facial action units? Dropdown
Prof. Actors: boolean, 1 or 0, if they are professional actors
Interaction: boolean, 1 or 0, was it an interaction or not?
N of participants: int, how many participants?
N-female: int, how many female. Count, not percentage
N-male: int, how many male. Count, not percentage
P-detailed age info: boolean, 1 or 0. If there is standard deviation or age information for each participant, then 1, otherwise 0.
Age range: int-int. Example: 18-76.
Age mean: float. Mean of the age distribution
Age median: float. Median of the age distribution.
Ethnicity: str. Percentage also if available. For example: Caucasian(80%), East-Asian(10%)
Language: str. If native speakers, L1-French, L2-English
Rater: str, who is rater, [self, external, both]
Total raters: int, number of external raters
RpI: int, raters per item, if they rated only some subset
N-ER-female: int, number of female raters
N-ER-male: int, number of male raters
R-detailed age info: boolean, 1 or 0. If there is standard deviation or age information for each rater, then 1, otherwise 0.
ER-Age range: int-int. Example: 18-76.
ER-Age mean: float. Mean of the age distribution
ER-Age median: float. Median of the age distribution.
ER-ethnicity: str
Affect labeling mode: str, scale, method (SAM)… If it’s range, use #
Emotion labels: str, delimiter is ‘,’. For example: anger, happiness, calm
Emotion Label Mode: str, e.g. FC (forced choice), Likert scale…