Debiasing Multimodal Emotion Detection Algorithms

Parth Vipul Shah*
University of Southern California
Los Angeles, USA
pvshah@usc.edu

Slava Zinevich*
University of Southern California
Los Angeles, USA
zinevich@usc.edu

ABSTRACT

Emotion detection from multimodal data is a prominent research problem in the field of machine learning and human cognition. Past research has shown certain models show bias against gender, ethnicity or race. In this work, we aim to leverage adversarial learning techniques to counter potential biases. The IEMOCAP dataset will explore debiasing against gender, and the SEWA database against ethnicity. We aim to develop more reliable emotion recognition models free from biases. We find that this technique of debiasing a model works as our fairness metrics see an improvement, indicating parity.

KEYWORDS

Emotion Detection, Bias, Fairness in Machine Learning, Adversarial Learning

1 PROBLEM DEFINITION

Emotion detection from audiovisual data is a key research problem in machine learning research of human cognition [1] [26]. There are numerous databases as well as models that aim to tackle this problem, and do so with relatively great success. However, numerous researchers have shown that some prominent models in the field tend to show bias against gender, race or ethnicity [24] [4] [19]. We aim to leverage novel techniques that aim to counter bias in the learning process of machine learning models and apply them in the field of emotion recognition. Emotion detection is essential for the advancement of Human Computer Interaction (HCI). Many systems such as virtual therapists or crowd sourced public safety rely on these [5] [7].

2 LITERATURE REVIEW

To mitigate bias, we must first evaluate existing models on 'fairness.' And for such an evaluation we require metrics. [9] A few such commonly used metrics are Demographic Parity and Equalized Odds. Demographic parity states that predictions must be independent of membership of the protected variable. Equalized odds states that the predictions should be independent of the protected variables and groups must have the same false positive and true positive rates.

Early efforts in the evaluation of human emotional expressions aimed to examine primarily the face of the human subject [7]. In specific, facial action units were established, and FACS (Facial action coding system)[18] has been set as the flagship codification. Using this system, facial expressions can be categorized and then associated with certain emotions as seen in [11]. Thus, some methods aim to classify facial action units, and use that information to detect emotional responses in subjects [25].

However, more modern methods of emotion recognition tend to resort to more complete end-to-end recognition methods[2]. Using deep learning models, multimodal data can be fused to generate classification of emotions. This can be done with early or late fusion.

[16] proves a useful survey of bias in emotion recognition and detection. They note many types of possible biases in the process of constructing these models, from the procurement of data all the way through training and analysis. These can be broadly categorized as 'data obtainment bias' and 'algorithmic bias.'

In brief, 'data obtainment bias' regards issues with obtaining data that is representative of the population, as well as the process of accurate labeling. Those assigned to label the data all have inherent biases, and thus can and do apply those biases to the labeling process. Meanwhile, 'algorithmic bias' refers to biases that arise from the learning process of the algorithm based on the construction of the model.

[19] looked at this issue specifically with respect to the IEMOCAP [3] dataset. This dataset was procured at USC and has been used in many emotion recognition tasks. They noted that existing models do not achieve the desired statistical parity between genders. In order to mitigate this, they re-sampled and filtered the dataset. They then trained a BERT-based model [8] on the new dataset to achieve better results. The model was based on [12] which is also trained and evaluated on the IEMOCAP dataset.

There has been much focused on enhancing the capabilities of emotion recognition models by advancing the embedding extraction process for each modality. [23] These embedding techniques, however, largely did not aim to mitigate the bias that is learned by the models. There are numerous ways to guide embeddings. [20] for instance used adversarial generative loss to enforce embeddings that separately capture the content and the style of text-to-speech models. [15] on the other hand, delves into integrating reinforcement learning policies to edit embeddings of seq2seq models.

3 DATA

To capture diverse applications and two types of biases, two datasets were chosen that contain multimodal data, including video, audio, and text data. The two biases we are studying are gender and ethnic bias. First, the IEMOCAP dataset[3]. This dataset contains videos and contains labeling data for the emotion expressed in the form of 9 labels: angry, excited, fear, sad, surprised, frustrated, happy, disappointed and neutral. For model training, videos are cut into short snippets, typically a few seconds, based on continuous speech from a speaker. Audio data is extracted and used separately. This dataset is valuable due to its extensive size, as well as accurate labeling. It can be used as a good benchmark for evaluating the

*Both authors contributed equally to this research.

CSCI 535, Spring 2023, USC

2023. ACM ISBN 978-x-xxxx-xxxx-x/YY/MM. . . $15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

results. This dataset does not contain many speakers, but provides information about the gender. Since it has only 10 unique speakers, it is not suitable for evaluating multifaceted biases such as ethnicity, but it is sufficient for evaluating gender bias. We use 1600 instances from this dataset.

The second dataset is the SEWA dataset [14] which contains a more diverse collection of speakers from various geographical and cultural backgrounds, as well as a diverse range of ages. In detail, SEWA has multimodal data linked to the recordings of the participants, including continuous labeling of traditional measures of human emotional expression such as facial action units triggers, as well as the corresponding valence and arousal. As noted in [11] these labels can be directly linked to classic human emotion categorization if needed. This dataset, while extensive, does not contain direct emotion labels such as those in the IEMOCAP dataset. Additionally, the speakers speak in multiple languages. evaluation can be done based on instances along four continuous dimensions namely amusement, empathy, liking and boredom. We use 600 instances from this dataset and restrict our analysis to the 'amusement' dimension.

4 METHODOLOGY

To evaluate the fairness of a model, we must first define metrics that capture the same. Hence, we use Demographic Parity and Equalized Odds. These two metrics are calculated for the three outputs of the model. This is detailed in the Experiment section. Additionally, we analyze the two modalities separately and together (acoustic, visual and multi-modal) to understand the individual effects of our techniques. Mathematically, the fairness metrics can be defined as follows:

For Demographic Parity, a classifier h satisfies DP under a distribution over (X, A, Y) if its prediction h(x) is statistically independent of the sensitive feature A.

$$E[h(X)A = a] = E[h(X)] \quad \forall a \tag{1}$$

For Equalized Odds, a classifier h satisfies EO under a distribution over (X, A, Y) if its prediction h(X) us conditionally independent of the sensitive feature A given the label Y.

$$E[h(X)A = a, Y = y] = E[h(X)Y = y] \quad \forall a, y \tag{2}$$

As noted above, current methods of emotion recognition are based on generating embeddings for the input data, and then processing them through a deep neural network. As audio, video, and text are temporal inputs, various temporal networks such as RNNs and LSTMs are typically employed to generate the latent state of each modality, and those are then fused to generate the final classification [27] [28].

Our goal is to guide the embeddings to represent a more generalized form. For audio embeddings, temporal audio data can be encoded through the use of an autoencoder architecture. In order to mitigate bias, our conjecture is that prioritizing the content of speech is more important than the pitch. The pitch, or style, can inherently bias against genders if the annotations contain bias, since men and women have distinctly different tonalities.

Training a complete audio encoder on this dataset is impractical. Thus, an existing, general purpose encoder is employed to extract initial embeddings. In this case VGGish [22] is employed to process the input audio into a temporal embedding sequence. To achieve the desired, style-neutral embeddings, a fully connected (FC) network is added to transform the original embedding into a new embedding. This embedding is trained to prioritize the content and emotional expression of the speaker, with the standing hypothesis that it will result in the mitigation of encoding gender or age specific information in the embeddings. In essence, this FC transformation will be policed by another neural network, based on an adversarial policy. This discriminator tries to identify the bias label of the embedding. In conjunction, the main deep temporal neural network receives the new embeddings to classify the emotions expressed by the subjects in the dataset. This is detailed in Figure 1.

For video embeddings, a similar method is used. First, video frames are encoded using Resnet50 [10]. Then, an analogous method to audio is used, where metadata about the bias, e.g. race or gender is be provided to a discriminator, and adversarial loss is assigned based on whether the discriminator can distinguish the bias. This should be more effective for multi-class label data such as race as opposed to gender, as the discriminator will have a stronger influence, compared with a random discriminator. The goal is to steer the generated embedding to contain more structural information about the subject's facial expressions and emotional state, rather than interweave information relating to the subject's personhood or complexion, such as gender, ethnicity or age.

Text sequences do not contain obvious tells of people and therefore possible biases. There could and do exist some differences that could imply one's gender, race, or age based on manner of speech learned through cultural mediums and personal experiences. [4] However, we do not believe that this modality is ideal for this particular work. It is possible to employ tools created in other research efforts [21] but this is not an item of our focus presently.

For multimodal classification, a multimodal late fusion method is used to get the final classifications. The weights for each modality are assigned using hyperparameter tuning. The auditory modality, especially in the case of the IEMOCAP dataset, works better for emotional classification than the visual modality, hence a larger weight is placed on it. For SEWA, the converse is true.

5 EXPERIMENT

The model is designed in Python using the Pytorch machine learning library [17]. The model is designed with 3 major components: a full-connected network, designed to transform the embedding of an individual modality, a discriminator network, to identify bias in the generated embeddings, and temporal network for classification of emotional labels. The loss for both label classification as well as the bias classification is cross entropy loss, and the optimization method is ADAM [13] with a learning rate of 0.005.

The fully connected network consists of two hidden layers, and an output layer, which outputs embeddings of the same size as the input. The audio embedding is of size 128 [22], and the hidden layers are of the same size. The visual embedding is of size 2048 [10], and the hidden layers are each of size 512. The Adversarial network is a simple fully connected network with an equivalent implementation, except the output is the number of bias labels, which is two for gender bias and 6 for racial bias. The temporal network is a 2-layer GRU [6] with hidden layer size of 128. The

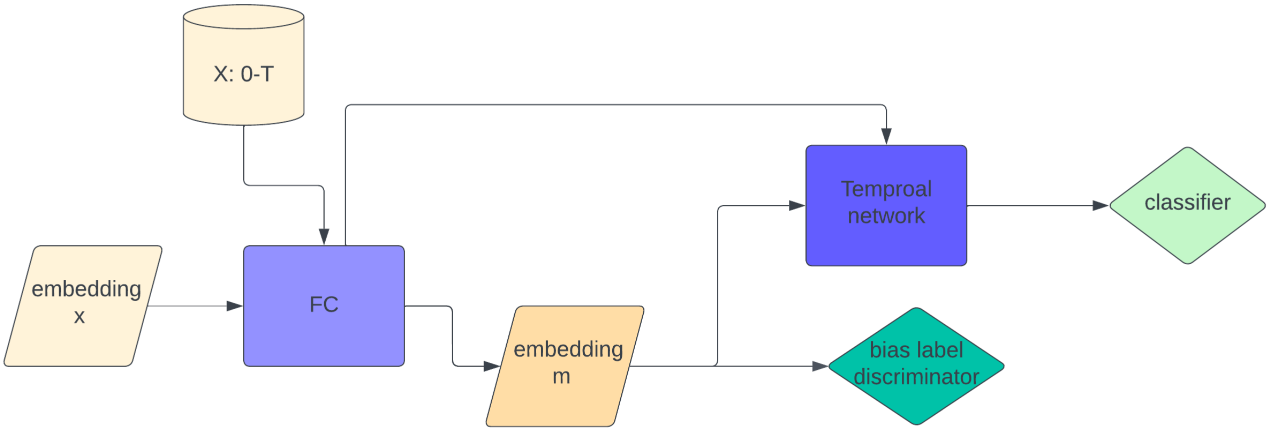

Figure 1: Model architecture to debias a modality using adversarial learning.

output size is conditioned on the number of labels the model is evaluating against.

For the abalation study, three model varieties are trained for 300 epochs a piece. The first is a simple temporal classifier with the above configuration that takes the audio and visual embeddings generated by the third-party encoders. The second features the embedding transformer network mentioned above in addition to the temporal classifier. The third is the proposed model, which has the additional adversarial policy. The temporal network, in all three cases, is trained on a single cross entropy loss based on the emotional classification label

$$L_c = -\sum_{i=1}^{C_1} \phi(y, \hat{y}_i) \log(p_i) \tag{3}$$

where ϕ is the indicator function whether the true label is equal to the predicted label. The adversarial network is trained similarly on cross entropy loss, but conditioned on the bias label

$$L_a = -\sum_{i=1}^{C_2} \phi(y, \hat{y}_i) \log(p_i) \tag{4}$$

The embedding transformation network is trained on the combined loss of the two mentioned above.

$$L_{fc} = L_c - \lambda L_a \tag{5}$$

For the second model there is no adversarial policy, so λ is 0, and for the third model $\lambda = 0.015$.

6 RESULTS

The IEMOCAP dataset was used to test our model's ability to detect emotions (multi-class classification) while being apathetic to the gender of the subject. Similarily, the SEWA dataset was used to test our model's ability to detect emotional arousal (binary classification) while being apathetic to the ethnicity of the subject. The results for IEMOCAP are detailed in Table 1, 2 and 3. The results for SEWA are detailed in Table 4, 5 and 6. The tables are divided on the basis of the modality that was being tested. Each table contains results for baseline, embedding transformation and adversarial learning variants of our model. The Demographic Parity and Equalized Odds measures are presented in the same table. A lower numerical value of the difference is desired and a higher numerical value of the ratio is desired to indicate parity. This is represented by the arrows in the table header. The bold font indicates the direction of increase in parity. All values are averaged over all emotion classes. We also report the corresponding F1 score.

Over half of the results display this increase/decrease in one or more of the measures of fairness we evaluated the three model varieties on. This promising result indicates success in our debiasing technique. The overall accuracy suffers marginally.

Table 1: IEMOCAP - Acoustic Modality

	Difference ↓	Ratio ↑	F1
Demographic Parity			
Baseline	0.0189	0.8970	0.9998
Embedding Transformation	0.0371	0.7697	0.8044
Adversarial Learning	**0.0169**	0.8768	0.8007
Equalized Odds			
Baseline	0.0473	0.8291	0.9998
Embedding Transformation	0.0812	0.6967	0.8044
Adversarial Learning	**0.0322**	**0.8301**	0.8007

7 CONCLUSION

Research in mitigating the bias of emotion recognition tasks specifically and all other tasks generally is increasingly important as these models get deployed into more and more applications. Machine learning developers should think critically about how a fair model would operate and conduct experiments to understand the types of biases their system may exhibit. Implementing adversarial networks for debiasing is an alluring technique as there is no modification to

Table 2: IEMOCAP - Visual Modality

	Difference ↓	Ratio ↑	F1
Demographic Parity			
Baseline	0.0007	0.7487	0.6907
Embedding Transformation	0.0113	0.9416	0.7812
Adversarial Learning	0.0078	**0.9231**	0.7131
Equalized Odds			
Baseline	0.0051	0.4867	0.6907
Embedding Transformation	0.0326	0.8724	0.7812
Adversarial Learning	0.0248	**0.8297**	0.7131

Table 3: IEMOCAP - Multiple Modalities (Acoustic and Visual)

	Difference ↓	Ratio ↑	F1
Demographic Parity			
Baseline	0.0077	0.8944	0.7715
Embedding Transformation	0.0089	0.9501	0.8261
Adversarial Learning	**0.0014**	**0.9970**	0.4354
Equalized Odds			
Baseline	0.0376	0.8162	0.7715
Embedding Transformation	0.0343	0.8774	0.8261
Adversarial Learning	**0.0026**	**0.9966**	0.4354

Table 4: SEWA - Acoustic Modality

	Difference ↓	Ratio ↑	F1
Demographic Parity			
Baseline	0.2599	0.6285	0.9750
Embedding Transformation	0.1099	0.8589	0.6900
Adversarial Learning	**0.1799**	**0.7906**	0.6516
Equalized Odds			
Baseline	0.1111	0.000	0.9750
Embedding Transformation	0.2363	0.6380	0.6900
Adversarial Learning	0.3040	**0.6090**	0.6516

the raw data. Following the results, we conclude that this technique of debiasing models for emotion detection works. Further work will focus on the accuracy aspect of these classifications and extending this for other demographic biases.

Table 5: SEWA - Visual Modality

	Difference ↓	Ratio ↑	F1
Demographic Parity			
Baseline	0.3245	0.6172	0.9487
Embedding Transformation	0.2299	0.6714	0.9383
Adversarial Learning	**0.2599**	**0.6338**	0.9150
Equalized Odds			
Baseline	0.1578	0.4673	0.9487
Embedding Transformation	0.1193	0.3636	0.9383
Adversarial Learning	**0.1408**	0.2755	0.9150

Table 6: SEWA - Multiple Modalities (Acoustic and Visual)

	Difference ↓	Ratio ↑	F1
Demographic Parity			
Baseline	0.2100	0.6911	0.9383
Embedding Transformation	0.2199	0.6986	0.9266
Adversarial Learning	**0.1978**	**0.7651**	0.9450
Equalized Odds			
Baseline	0.0862	0.3789	0.9383
Embedding Transformation	0.1505	0.3116	0.9266
Adversarial Learning	0.1161	**0.4839**	0.9450

8 FUTURE WORK

Our next steps involve improving the accuracy of our classifications on both, IEMOCAP and SEWA databases for emotion recognition. We have successfully demonstrated that we can debias predictions using adversarial learning. This approach can be adopted in real-world systems to ensure fairness and equity. This work explored the acoustic, visual and combined modality for two (gender and ethnicity) protected variables. Future work can explore adapting this method for the mitigation of different types of demographic biases like racial. The textual modality can also be incorporated to understand the bias of this modality.

9 CONTRIBUTIONS

The contributions are as follows:

(1) Slava is responsible for implementing the new adversarial learning network. Researched relevant work on adversarial learning.
(2) Parth is responsible for data pre-processing, implementing the fairness metrics and aggregating results. Researched relevant work on fairness and bias in AI.

This work is fully reproducible with the code files provided. Embedded data is available on request.

ACKNOWLEDGMENTS

We would like to thank Professor Mohammad Soleymani and Suu Lei from the Institute for Creative Technologies (ICT) at USC for their insightful discussions and guidance.

REFERENCES

[1] Babak Joze Abbaschian, Daniel Sierra-Sosa, and Adel Elmaghraby. 2021. Deep Learning Techniques for Speech Emotion Recognition, from Databases to Models. *Sensors* 21, 4 (2021). https://doi.org/10.3390/s21041249

[2] Sharmeen M Saleem Abdullah Abdullah, Siddeeq Y Ameen Ameen, Mohammed AM Sadeeq, and Subhi Zeebaree. 2021. Multimodal emotion recognition using deep learning. *Journal of Applied Science and Technology Trends* 2, 02 (2021), 52–58.

[3] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N. Chang, Sungbok Lee, and Shrikanth S. Narayanan. 2008. IEMOCAP: interactive emotional dyadic motion capture database. *Language Resources and Evaluation* 42, 4 (Dec. 2008), 335–359. https://doi.org/10.1007/s10579-008-9076-6

[4] Kai-Wei Chang, Vinodkumar Prabhakaran, and Vicente Ordonez. 2019. Bias and Fairness in Natural Language Processing. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): Tutorial Abstracts*. Association for Computational Linguistics, Hong Kong, China. https://aclanthology.org/D19-2004

[5] M Kalpana Chowdary, Tu N Nguyen, and D Jude Hemanth. 2021. Deep learning-based facial emotion recognition for human–computer interaction applications. *Neural Computing and Applications* (2021), 1–18.

[6] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling. arXiv:1412.3555 [cs.NE]

[7] Roddy Cowie, Ellen Douglas-Cowie, Nicolas Tsapatsoulis, George Votsis, Stefanos Kollias, Winfried Fellenz, and John G Taylor. 2001. Emotion recognition in human-computer interaction. *IEEE Signal processing magazine* 18, 1 (2001), 32–80.

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:1810.04805 [cs.CL]

[9] Pratyush Garg, John D. Villasenor, and Virginia Foggo. 2020. Fairness Metrics: A Comparative Analysis. *CoRR* abs/2001.07864 (2020). arXiv:2001.07864 https://arxiv.org/abs/2001.07864

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep Residual Learning for Image Recognition. *CoRR* abs/1512.03385 (2015). arXiv:1512.03385 http://arxiv.org/abs/1512.03385

[11] Sylwia Hyniewska, Wataru Sato, Susanne Kaiser, and Catherine Pelachaud. 2019. Naturalistic Emotion Decoding From Facial Action Sets. *Frontiers in Psychology* 9 (2019). https://doi.org/10.3389/fpsyg.2018.02678

[12] Taewoon Kim and Piek Vossen. 2021. EmoBERTa: Speaker-Aware Emotion Recognition in Conversation with RoBERTa. arXiv:2108.12009 [cs.CL]

[13] Diederik P. Kingma and Jimmy Ba. 2017. Adam: A Method for Stochastic Optimization. arXiv:1412.6980 [cs.LG]

[14] Jean Kossaifi, Robert Walecki, Yannis Panagakis, Jie Shen, Maximilian Schmitt, Fabien Ringeval, Jing Han, Vedhas Pandit, Björn W. Schuller, Kam Star, Elnar Hajiyev, and Maja Pantic. 2019. SEWA DB: A Rich Database for Audio-Visual Emotion and Sentiment Research in the Wild. *CoRR* abs/1901.02839 (2019). arXiv:1901.02839 http://arxiv.org/abs/1901.02839

[15] Shunfu Mao and Joshua Fan. [n. d.]. Edit Embedding via Reinforcement Learning. http://joshuafan.github.io/files/EditEmbed_Final_Report.pdf. Accessed on March 26, 2023.

[16] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2019. A Survey on Bias and Fairness in Machine Learning. *CoRR* abs/1908.09635 (2019). arXiv:1908.09635 http://arxiv.org/abs/1908.09635

[17] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. *PyTorch: An Imperative Style, High-Performance Deep Learning Library.* Curran Associates Inc., Red Hook, NY, USA.

[18] Erika L Rosenberg and Paul Ekman. 2020. *What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS).* Oxford University Press.

[19] Matheus Schmitz, Rehan Ahmed, and Jim Cao. 2022. Bias and Fairness on Multimodal Emotion Detection Algorithms. *ArXiv* abs/2205.08383 (2022).

[20] ShuangMa, Daniel McDuff, and Yale Song. 2019. Neural TTS Stylization with Adversarial and Collaborative Games. In *International Conference on Learning Representations (ICLR).* https://www.microsoft.com/en-us/research/publication/neural-tts-stylization-with-adversarial-and-collaborative-games/

[21] Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth M. Belding, Kai-Wei Chang, and William Yang Wang. 2019. Mitigating Gender Bias in Natural Language Processing: Literature Review. *CoRR* abs/1906.08976 (2019). arXiv:1906.08976 http://arxiv.org/abs/1906.08976

[22] Tensorflow. [n. d.]. Models for AudioSet: A Large Scale Dataset of Audio Events. https://github.com/tensorflow/models/tree/master/research/audioset. Accessed on March 26, 2023.

[23] Zhongwei Xie, Ling Liu, Lin Li, and Luo Zhong. 2021. Learning Joint Embedding with Modality Alignments for Cross-Modal Retrieval of Recipes and Food Images. In *Proceedings of the 30th ACM International Conference on Information &amp Knowledge Management.* ACM. https://doi.org/10.1145/3459637.3482270

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

[25] Li Yao, Yan Wan, Hongjie Ni, and Bugao Xu. 2021. Action unit classification for facial expression recognition using active learning and SVM. *Multimedia Tools and Applications* (2021). https://doi.org/10.1007/s11042-021-10836-w

[26] Jianhua Zhang, Zhong Yin, Peng Chen, and Stefano Nichele. 2020. Emotion recognition using multi-modal data and machine learning techniques: A tutorial and review. *Information Fusion* 59 (2020), 103–126. https://doi.org/10.1016/j.inffus.2020.01.011

[27] Kaihao Zhang, Yongzhen Huang, Yong Du, and Liang Wang. 2017. Facial expression recognition based on deep evolutional spatial-temporal networks. *IEEE Transactions on Image Processing* 26, 9 (2017), 4193–4203.

[28] Tong Zhang, Wenming Zheng, Zhen Cui, Yuan Zong, and Yang Li. 2018. Spatial–temporal recurrent neural network for emotion recognition. *IEEE transactions on cybernetics* 49, 3 (2018), 839–847.

Received 1 April 2023