Race bias in Machine Learning: an emotion detection approach

Ignacio Montes Álvarez, *0957161*, Florian Gaeremynck, *7006705,* Sander Kaspers, *5744512,* Massimiliano Garzoni di Adorgnano, *4866169,* German Savchenko, *2547953,* and Diana Gyurjiyan, *6770908*

Abstract—The present work sets out to examine ethnicity bias in a facial expression classification task. To account for this, we retrieved two sets of images displaying faces in relation to 7 basic emotions. We considered two sets of data that depicted images of people of two distinct ethnicities: western-Caucasian and eastern-Asian. To model this inference task we constructed a convolutional neural network (CNN) architecture to fit the image data. This choice was motivated by the grid-like topology of images, which CNNs resemble closely in their computations. We thus evaluated the classification performance of this model on the multi-label task (emotion detection) across the different sets of image data, which resulted in four main evaluations. The underlying hypothesis was that the Caucasian-only trained model would generalize poorly when applied to the Asian test set, and the same for the Asian-only trained model when applied to the Caucasian test set. Indeed, the results show that both these performances were rather low. Interestingly though, the Asian-trained model performed better on the Caucasian test compared to the Caucasian-trained model on the Asian test set. Finally, directions for future research are suggested.

Index Terms—Ethnicity bias, Emotion detection, Facial expressions, Machine Learning.

1 Introduction

Mental health problems are increasing during the pandemic, and are difficult to identify. Severe mental health problems also increase the likelihood of early death. Emotion detection based on facial impressions can help to identify depression or can be used for safety and security purposes (e.g. early violence detection). Emotion detection is a new and ongoing research field in nowadays computer science. However, emotion detection based on facial expressions is based on object recognition. The data sets used for this purpose therefore should be unbiased, given that a biased data set is highly prone to produce a biased output. For example, a set containing only images of people of Caucasian descent will not represent the world population and therefore the performance of classifiers will decrease when tested on a larger set of images displaying people of other varying ethnicities.

For the work reported in this paper, we construct and train different classifiers on two different data sets: one containing images of western-Caucasian people, while the other comprised of eastern-Asian faces. Based on these two different sets, we examine the performance of the classifiers and their biases given the different input collections. The motivation behind this experiment is to shed light on a better understanding of the classification biases in facial expressions and thus emotional recognition, while trying to mitigate the problem of bias in computationally learning features indicative of faces.

It is important to clarify that in this project we refer to bias as in the general "social" terminology, hereby specifically with concerns to ethnicity. This is not to be confused with the notion of bias commonly reported in Machine Learning, i.e. as with the bias-variance trade-off in the context of model complexity.

2 Related work

Due to the effect of globalisation, human societies become connected with each other. Therefore, communication between these societies is of increasing value to maintain the relationships between cultures [1]. A key aspect of communication between humans are emotions. When communicating, certain facial expressions lead to an understanding of someone's emotions and state of mind [2] [3]. However, facial expressions depend on contextual information [4]. For example, sarcasm can be of influence in the interpretation of certain facial expressions. Also, a smile can be ambiguous because of its multiple possible interpretations [5] [6] and indeed previous research has found that smiles do not express similar emotions between cultures [7] [8]. For instance, social hierarchy for East Asians has more influence on the reason of smiling, whereas social interaction influences Caucasians more [9]. Also, the impact of contextual information on the interpretation of facial expressions varies between cultures [4]. To illustrate, Caucasians are less affected by contextual information compared to East Asians in determining emotions from facial expressions [10] [11]. Furthermore, preceding emotion has, in general, impact on determining emotions from facial expressions. For instance, a neutral facial expression is judged as sad when introduced after a happy facial expression [12]. Especially Caucasians are less likely to be affected by preceding facial expressions or temporal emotional context [4].

The previous research stated above has showed that basic facial expressions are not universal across cultures [13].

Yet, software designed for automatically identifying emotions from facial expressions do not take these dissimilarities into account [14]. Therefore, psychological insights about the reasons that underlie these differences across cultures are required to obtain a robust system for emotion detection. Psychological research has gained insight in the origins of these dissimilarities and found that the differences across cultures are systematically confused [15]. Especially fear and disgust happen to be hard facial expressions that are not culturally universal [13]. A possible solution for the problem of culture specific emotions is to take variation between certain facial regions into account [16]. Multiple machine learning techniques attempted to overcome the problem of culture specific facial expressions [17] [18] [19]. However, due to the large amount of variations in faces these techniques could not yet succeed in effectively identifying emotions from facial expressions between cultures [20] [21]. Artificial neural networks are often used to enhance the performance of identifying emotions from facial expressions. Specifically, convolutional neural networks are capable of obtaining very good performance on this task [22].

A drawback in classifying emotions from facial expressions is the fact that other factors also have influence on the classification. For instance, the classification of darker skins is less accurate compared to lighter skins and similarly, females are more difficult to classify compared to males [23]. Besides ethnicity and gender, age is also a factor that can influence emotion detection. However, this does not imply that emotion detection is more difficult for females. Instead, compared to ethnicity or age bias, gender bias is the smallest [24]. Regarding gender bias, it goes both ways. To illustrate, detecting the emotion of 'surprise' is easier for male subjects and 'upset' emotions are easier to identify in females [24].

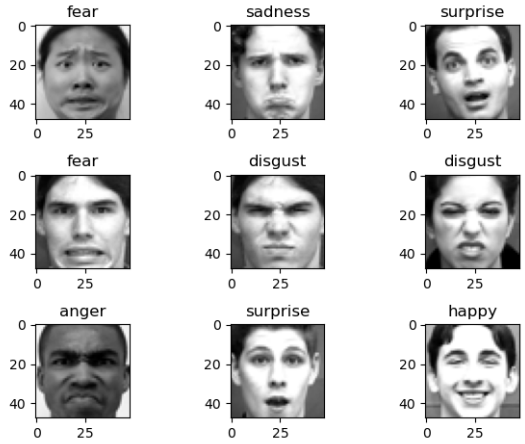

Fig. 1. Sample of images in from the CK+ data set.

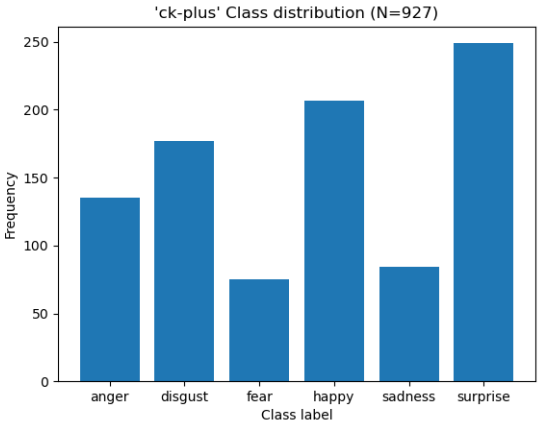

Fig. 2. Distribution of labels in the CK+ data set.

3 Data

For the purpose of this project we considered different sources of facial expression data. We defined the main difference between these as being the culturally different physiological characteristics of the faces. More specifically, we considered two different physiological traits in faces: western-Caucasian and eastern-Asian. To account for the modeling and assessment of cultural bias we focused on only these two face configurations, narrowing it down to these two main divergent classes.

Concerning the set of input images that present western-Caucasian faces, we decided to employ the Cohn-Kanade (CK+) data set for facial expression recognition. This included a total of 981 images displaying individuals of Caucasian origins. The person in each image exhibited a facial expression in relation to the following seven emotions: *anger*, *disgust*, *fear*, *happiness*, *sadness*, *surprise* and *contempt*. A sample of this data is presented in figure 1. In figure 2 the distribution of the labels in the CK+ data set is shown. Based on this graph, we can conclude that the labels are not evenly distributed, the label *surprise* is the most frequent label in the data set, while *fear* and *sadness* are the least frequent label in the data set.

Conversely, for the set of images of eastern-asian facial expressions we considered the Japanese Female Facial Expressions (JAFFE) database. In this case, the total amount of images was 213, each showing the face of a female Japanese (of Asian descent). Similarly as to the CK+ dataset, each person displayed a facial expression related to the same six emotions (*anger*, *disgust*, *fear*, *happiness*, *sadness*, *surprise*), with the exception of the seven class being *neutral* instead of *contempt*. Refer to 3 for a sample of this images. In figure 4 the distribution of the labels in the JAFFE data set is shown. Based on this graph, we can conclude that the labels are approximately evenly distributed.

It is worth to mention that the discovery and retrieval of eastern-Asian facial expression data was particularly cumbersome. This is because additional data proved to be either not accessible for retrieval (i.e. original authors not granting access) or simply non-existent. Therefore, in order to keep the balance between the data distributions, we decided to augment the eastern-Asian dataset by adding 702 images. For this purpose, we generated synthetic images by rotating and flipping each image by varying degrees of angulation. Specifically, we used `OpenCV` to perform left and right flips of the images defined as: 90º clockwise, 90º counter-clockwise and 180º rotations. These modifications can be seen in the sample of data reported in figure 3.

Another pre-processing step was done on the images coming from the JAFFE data set. We resized each image

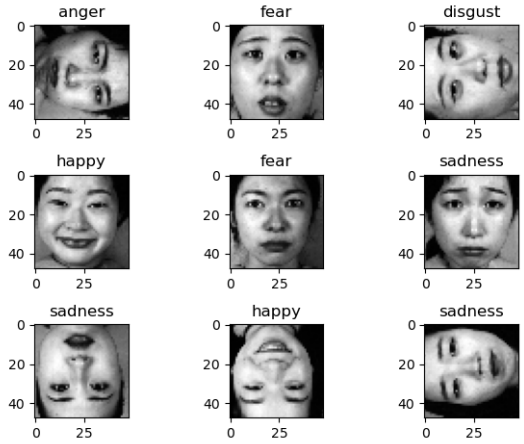

Fig. 3. Sample of images in from the JAFFE data set.

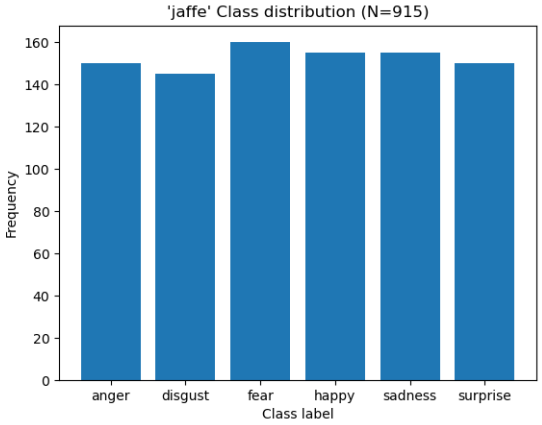

Fig. 4. Distribution of labels in the JAFFE data set.

(of size 256 by 256 pixels) to have width ranging from pixel 80 until 230 and height ranging from pixel 50 until 200, in order to crop out redundant visual information that was present in the pictures, such as people's necks or higher segments of clothing (e.g. shirt collars). These portions of the images are not informative of facial expression and even loss of ethnicity, and thus were removed as such could be confounds for the learning task at hand.

Finally, we resized the dimensions of all the images from every data set to result in square images of 48 by 48 pixels, for the purpose of consistency in the input data to feed to the learning models. Moreover, given that the classes of emotions included in the data sets were not consistent, we decided to only consider images that related to the following six basic facial expressions: *anger*, *disgust*, *fear*, *happiness*, *sadness*, *surprise*. By doing this, we discarded information about the other expressions such as *contempt* or *neutralness*. Although possibly considered a shortcoming, this was not an aspect of key relevance for the present work since the goal was not to discriminate between emotions.

4 Modeling Approach and Architecture

4.1 Architecture

As we are dealing with images, we decided to use a convolutional neural network method given its affinity with grid-like data. Our network architecture was developed in Python and using `Tensorflow` 2.3 with `Keras`; it is sequential and defined as follows:

Feature learning

1) **Rescaling** layer where the input size was declared as 48x48x1 (height, width and the gray scale channel) and normalized by diving each pixel value by 255.
2) **Conv2D** layer with 32 filters of size 3x3.
3) **Conv2D** layer with 64 filters of size 3x3. The amount of filters was increased in order to produce more feature maps and capture higher level characteristics from the images.

It is worth to note that two additional layers followed each one of the convolutional layers, a **MaxPooling2D** layer of size 2x2 and a **dropout** layer.

Classification

1) **Flatten** layer of size 9216.
2) **Dense** layer of size 64.
3) **Dropout** layer.
4) **Dense** layer of size 6 (the total number of classes to be predicted).
5) A softmax **activation** layer to enable multi-class classification.

Finally, this whole setup led to a total 609,094 parameters (weights) to be estimated, distributed across just three of the layers: 320 for the first Conv2D layer, 18,496 for the second one and 589,888 for the first dense layer.

4.2 Hyperparameter tuning

Besides the architecture definition, an optimal hyperparameter search was performed. We used the `Keras Tuner` to execute a random search (with ten different trials and two executions per trial to account for the variability incurred from training the network) over the following set of spaces:

- Activation: after both Conv2D layers and the first dense layer, *ReLU*, *tanh* and *sigmoid* activations were tried.
- Dropout rate: each dropout layer was configured to look for a value ranging from 0 to 0.5.
- Optimizer: *adam*, *adagrad*, *rmsprop* and *sgd* were considered to compile the model.
- Learning rate: to be applied to the chosen optimizer and ranging from 0.0001 to 0.01.

In addition to this, as we encoded our target variable (emotion) as an integer ranging from 0 to 5 (number of classes - 1), we used the sparse categorical cross-entropy as the loss function. Accordingly, we used as evaluation metric the sparse categorical accuracy. For training, we defined a maximum of 15 epochs to find the best combination of hyper-parameters. Once these were estimated, another model was constructed with these values and trained for a maximum of 50 epochs.

4.3 Other considerations

The data was split in a 70/20/10 fashion, each portion for training/testing/validation accordingly (by using the validation set, we ensured that the model had a metric to adjust the hyper-parameters during training). Furthermore, we used two techniques to approach the issue of model over-fitting: the addition of dropout layers and early stopping. The former were placed in the architecture right after the pooling layers as specified earlier. The latter one was set to monitor the validation loss and wait a maximum of three epochs to stop training if the value has not improved by at least 0.075. Also, the experiment[1] was conducted on an system with an AMD Ryzen 3600X CPU, 16 GB of RAM and a Nvidia GTX 980 GPU (Nvidia CUDA Toolkit 10.1 and cuDNN 8.0.5); the computations were performed on the GPU and the training elapsed for 3 minutes (due to the different trials for hyper-parameter combinations).

5 Results

The best values for the hyper-parameters are shown in table 1. The results of both models can be found in table 2 and 3. In the first table the results of the training and testing on the same race data set are shown, while in the latter the results of the mixed-race data set are shown. The model is performing better when trained and tested on the CK+ data set. However, when the test set is a mixed-race data set, the performance of the model trained on the CK+ drops drastically.

TABLE 1
Best hyper-parameters chosen (same for both models, CK+ and JAFFE)

Parameter	Value
activation_conv1	tanh
dropout_conv1	0.3
activation_conv2	relu
dropout_conv2	0.3
activation_dense	sigmoid
dropout_dense	0.1
optimizer	rmsprop
learning_rate	0.001

TABLE 2
Results for both models using non-mixed data (test images are from same data set as training ones)

	CK+	JAFFE
Training	97.8	84
Testing	96.2	66.1

TABLE 3
Results for both models using mixed data (test images do not pertain to the same data-set from which the model was trained, i.e CK+ header means that the CK+ trained model was tested with JAFFE images and the other way around for the other header)

	CK+	JAFFE
Accuracy	25.1	46.2

1. All the code is available at https://github.com/Nacho888/pr2021-facial-emotion

The results of the training and validation accuracy of all experiments are shown in figures 5 and 6. Based on these graphs we can conclude that training on the CK+ data sets gives better results and a less "spikey" slope compared to the JAFFE data set. The validation accuracy of the CK+ trained model is also slightly higher than the training accuracy, which is an indication of over-fitting of the model. Taking this into consideration, the overall results can be easily explained: because of the lack of the ability to generalize the model does not recognize the right features of the images, but merely memorized the training set. This could be an explanation of the low performance of the CK+ trained model on the JAFFE test set. Moreover, look at the confusion matrix of this particular model in figure 7, this conclusion is confirmed: due to the high frequency of the label *surprise*, this particular label is the most frequent predicted label in the confusion matrix.

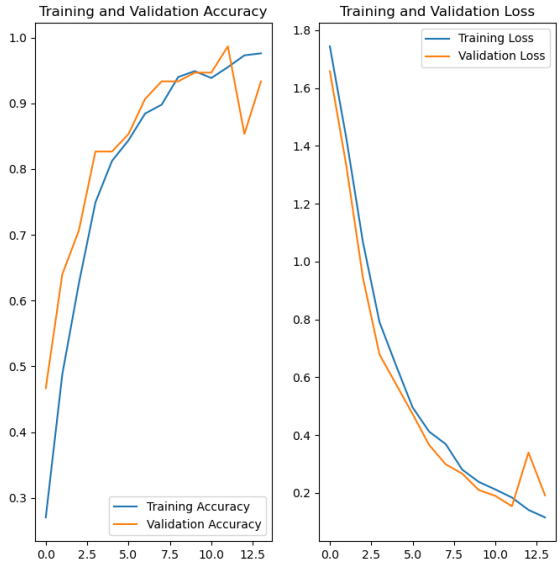

Fig. 5. Training loss and accuracy - CK+ data set

In figures 8, 7, 9 and 10 the confusion matrices of the different experiments are displayed. Based on these figures we can conclude that model trained and tested on solely CK+ data set gives the best results. Using the CK+ as a training set and the JAFFE as a test set, as seen in figure 7, gives a surprisingly effect, namely that the most frequently predicted label is the emotion *surprise*. One reason for this could be the class distribution of the CK+ data set, where *surprise* is the most frequent label. Looking at figure 9, we conclude that the results are not as good as the results of solely training and testing on the CK+ data set, but still performing better than the mixed race experiment. *Anger* is mostly mistaken for the emotion *sadness*, which is similar to the results of the figure 8. Lastly, when we test the JAFFE trained model on the CK+ test set, we get the results as seen in figure 10. The model still has difficulty with recognizing *surprise* and the difficulty of distinguishing *anger* from *sadness* remains. However, *surprise* is not the most frequent

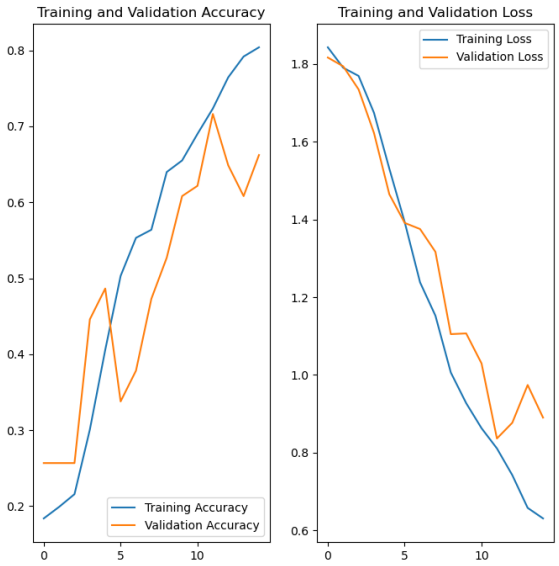

Fig. 6. Training loss and accuracy - JAFFE data set

predicted label, as we have seen in figure 7.

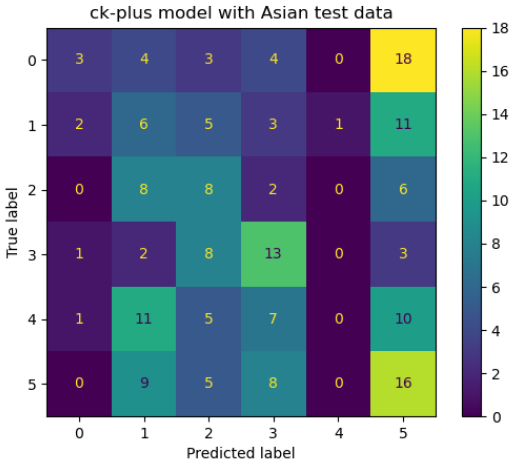

Fig. 7. Model trained on CK+ and tested on JAFFE

6 Conclusions

Based on the results in this experiment, we can conclude the following. First of all, training on the CK+ and testing on the CK+ data set (Figure 8), resulted in better performance compared to the JAFFE data set (Figure 9). Using the CK+ data set on the JAFFE testing set (Figure 7), resulted in lower performance, namely 25.1% accuracy. One surprising result is that the emotion *surprise* is the most frequent predicted label when the CK+ data set is used as training set. Looking at the distribution of the labels in the CK+ data set, we can conclude that this is due to the fact that *surprise* is the most frequent occurring label in the set.

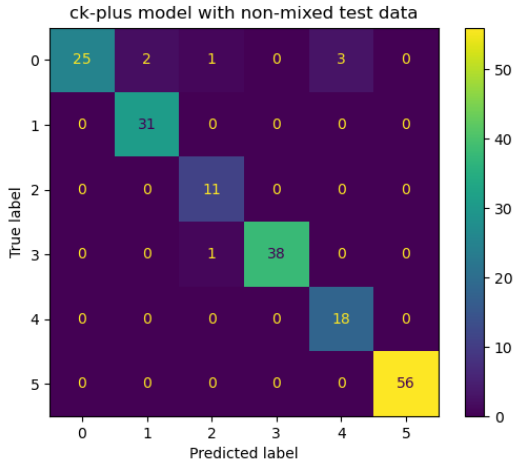

Fig. 8. Model trained and tested on CK+

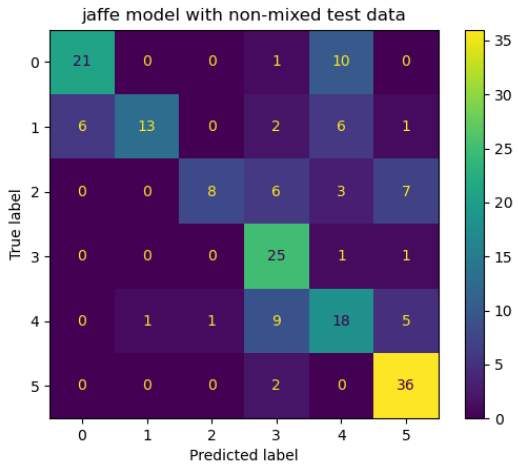

Fig. 9. Model trained and tested on JAFFE

However, when using the JAFFE as training set, this resulted in bad performance compared to the CK+ experiment, but a slightly better performance when a Caucasian data set was used as testing set. Training the model on JAFFE and testing the model on CK+ (Figure 10) resulted in 46.2% accuracy. Another important result here is that training on the JAFFE data set did not result in high prediction of the label surprise as we have seen with the CK+ data set. Contrarily, *Surprise* is the best predicted label with the lowest errors in the confusion matrix.

The data sets used in this experiment however had some limitations. First, the CK+ data set contained other ethnic groups and was therefore not a Caucasian-only data set. Moreover, the emotions in both data sets were extreme, where emotions in reality tend to be more subtle, as well as more complex (mixed emotions like crying and laughing simultaneously adds up for an entirely new category of emotions). Therefore these data sets are not representable and can hardly see any application in real-life situations. Emotions are not always detailed as they are in these data sets. Lastly, the JAFFE data set has female-only images and

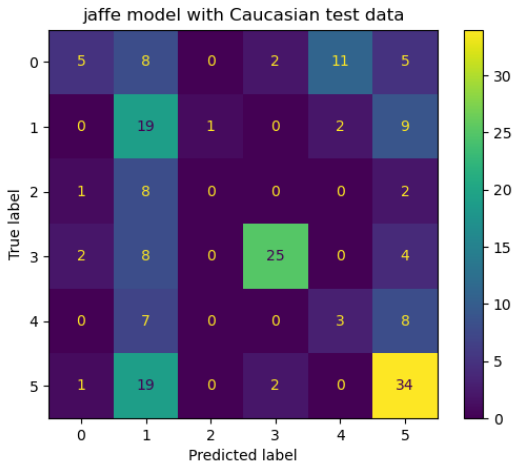

Fig. 10. Model trained on JAFFE and tested on CK+

no male-images were present, which we consider as being a possible confound influencing the results of this experiment. Nevertheless, ethnicity has a larger impact on detecting emotions than gender, and gender bias in emotion detection goes both ways. Therefore, it is more probable that the bias is due to difference in ethnicity than in gender.

For future work we would recommend the use of a mixed gender data set in order to examine the gender or race based as concluded in this experiment. Furthermore, we suggest the use of more complex images, as Neural Networks tend to work better with more extensive data sets compared to limited size collections. In this experiment grey-scaled images were used and for future experiments we would advise high resolution and coloured images.

REFERENCES

[1] W. B. Gudykunst, Y. Matsumoto, S. Ting-Toomey, T. Nishida, K. Kim, and S. Heyman, "The Influence of Cultural Individualism-Collectivism, Self Construals, and Individual Values on Communication Styles Across Cultures," *Human Communication Research*, vol. 22, pp. 510–543, 03 2006.

[2] J. B. Freeman and N. Ambady, "When two become one: Temporally dynamic integration of the face and voice," *Journal of Experimental Social Psychology*, vol. 47, no. 1, pp. 259–263, 2011.

[3] P. M. Niedenthal and M. Brauer, "Social functionality of human emotion," *Annual Review of Psychology*, vol. 63, no. 1, pp. 259–285, 2012. PMID: 22017377.

[4] X. Fang, G. A. van Kleef, K. Kawakami, and D. A. Sauter, "Cultural differences in perceiving transitions in emotional facial expressions: Easterners show greater contrast effects than westerners," *Journal of Experimental Social Psychology*, vol. 95, p. 104143, 2021.

[5] U. Hess, M. G. Beaupré, N. Cheung, *et al.*, "Who to whom and why–cultural differences and similarities in the function of smiles," *An empirical reflection on the smile*, vol. 4, p. 187, 2002.

[6] P. M. Niedenthal, M. Mermillod, M. Maringer, and U. Hess, "The simulation of smiles (sims) model: Embodied simulation and the meaning of facial expression," *Behavioral and Brain Sciences*, vol. 33, no. 6, p. 417–433, 2010.

[7] K. Ishii, Y. Miyamoto, K. Mayama, and P. M. Niedenthal, "When your smile fades away: Cultural differences in sensitivity to the disappearance of smiles," *Social Psychological and Personality Science*, vol. 2, no. 5, pp. 516–522, 2011.

[8] D. Matsumoto and T. Kudoh, "American-japanese cultural differences in attributions of personality based on smiles," *Journal of Nonverbal behavior*, vol. 17, no. 4, pp. 231–243, 1993.

[9] M. Rychlowska, Y. Miyamoto, D. Matsumoto, U. Hess, E. Gilboa-Schechtman, S. Kamble, H. Muluk, T. Masuda, and P. M. Niedenthal, "Heterogeneity of long-history migration explains cultural differences in reports of emotional expressivity and the functions of smiles," *Proceedings of the National Academy of Sciences*, vol. 112, no. 19, pp. E2429–E2436, 2015.

[10] T. Masuda, P. C. Ellsworth, B. Mesquita, J. Leu, S. Tanida, and E. Van de Veerdonk, "Placing the face in context: cultural differences in the perception of facial emotion.," *Journal of personality and social psychology*, vol. 94, no. 3, p. 365, 2008.

[11] J. T. Stanley, X. Zhang, H. H. Fung, and D. M. Isaacowitz, "Cultural differences in gaze and emotion recognition: Americans contrast more than chinese.," *Emotion*, vol. 13, no. 1, p. 36, 2013.

[12] J. A. Russell and B. Fehr, "Relativity in the perception of emotion in facial expressions.," *Journal of Experimental Psychology: General*, vol. 116, no. 3, p. 223, 1987.

[13] G. Benitez-Garcie, T. Nakamure, and M. Kaneko, "Multicultural facial expression recognition based on differences of western-caucasian and east-asian facial expressions of emotions," *IEICE Transactions on Information and Systems*, vol. E101.D, no. 5, pp. 1317–1324, 2018.

[14] Y. Tian, T. Kanade, and J. F. Cohn, *Facial Expression Recognition*, pp. 487–519. London: Springer London, 2011.

[15] R. E. Jack, "Culture and facial expressions of emotion," *Visual Cognition*, vol. 21, no. 9-10, pp. 1248–1286, 2013.

[16] X. Zhao and S. Zhang, "Facial expression recognition using local binary patterns and discriminant kernel locally linear embedding," *EURASIP journal on Advances in signal processing*, vol. 2012, no. 1, pp. 1–9, 2012.

[17] G. D. . L. J., "Extreme learning machine ensemble using bagging for facial expression recognition," *Journal of Information Processing Systems*, vol. 10, pp. 443–458, 09 2014.

[18] G. Levi and T. Hassner, "Emotion recognition in the wild via convolutional neural networks and mapped binary patterns," ICMI '15, (New York, NY, USA), Association for Computing Machinery, 2015.

[19] E. Owusu, Y. Zhan, and Q. R. Mao, "A neural-adaboost based facial expression recognition system," *Expert Systems with Applications*, vol. 41, no. 7, pp. 3383–3390, 2014.

[20] G. Ali, A. Ali, F. Ali, U. Draz, F. Majeed, S. Yasin, T. Ali, and N. Haider, "Artificial neural network based ensemble approach for multicultural facial expressions analysis," *IEEE Access*, vol. 8, pp. 134950–134963, 2020.

[21] F. A. M. da Silva and H. Pedrini, "Effects of cultural characteristics on building an emotion classifier through facial expression analysis," *Journal of Electronic Imaging*, vol. 24, no. 2, pp. 1 – 9, 2015.

[22] H. Jung, S. Lee, S. Park, B. Kim, J. Kim, I. Lee, and C. Ahn, "Development of deep learning-based facial expression recognition system," in *2015 21st Korea-Japan Joint Workshop on Frontiers of Computer Vision (FCV)*, pp. 1–4, 2015.

[23] J. A. Buolamwini, *Gender shades: intersectional phenotypic and demographic evaluation of face datasets and gender classifiers*. PhD thesis, Massachusetts Institute of Technology, 2017.

[24] A. Domnich and G. Anbarjafari, "Responsible ai: Gender bias assessment in emotion recognition," *arXiv preprint arXiv:2103.11436*, 2021.