Contextual Emotion Learning Challenge

Jainendra Shukla[1] Puneet Gupta[2] Aniket Bera[3] Arka Sarkar[4] Prakhar Goel[1] Shubhangi Butta[5] Anup Kumar Gupta[2] Snehil Sanyal[6] Debanga Raj Neog[7] M K Bhuyan[8] Kalyani Marathe[9] Linda Shapiro[9] Alex Colburn[9] Varchita Lalwani[10]

[1]Department of Computer Science and Engineering and Human-Centered Design, Indraprastha Institute of Information Technology Delhi, India

[2]Department of Computer Science and Engineering, Indian Institute of Technology Indore, India

[3]Department of Computer Science and Engineering, University of Maryland at College Park

[4]Department of Computer Science and Applied Mathematics, Indraprastha Institute of Information Technology Delhi, India

[5]Department of Computer Science and Social Sciences, Indraprastha Institute of Information Technology Delhi, India

[6]Research Scholar, Indian Institute of Technology Guwahati, India

[7]Mehta Family School of Data Science and Artificial Intelligence, Indian Institute of Technology Guwahati, India

[8]Department of Electronics and Electrical Engineering, Indian Institute of Technology Guwahati, India

[9]University of Washington, Seattle, United States of America

[10]Independent Researcher

2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021) | 978-1-6654-3176-7/21/$31.00 ©2021 IEEE | DOI: 10.1109/FG52635.2021.9667034

***Abstract*— Emotion recognition via vision has been deeply associated with facial expressions, and the inference of emotions has, more often than not, been based on the same. However, context, both environmental and social, plays an imperative role in emotion recognition but has not been incorporated widely so far. The meaning of emotion might entirely switch when shifted from one setting to another if only facial expressions are taken into account. Moreover, there exists no study in the Indian context about the same. To cater to this issue, we generate and introduce the Indian Contextual Emotion Recognition (ICER) dataset based on the multi-ethnic Indian context. This paper summarises the Contextual Emotion Learning Challenge (CELC 2021) organized in conjunction with the 16th IEEE Conference on Automatic Face and Gesture Recognition (FG) 2021. We outline the tasks posed in the challenge, the novel dataset, along with its challenges and the evaluation method. Lastly, we conclude by discussing the possible future directions.**

I. Introduction

With society racing towards a technology-driven world, marrying the field of social cognition with technological sciences has been gaining much importance to address the gap between humans and the tech world. This process is usually envisioned as being translated into building systems that are empathetic. However, this translation is not always the most mindful as the intricate and implicit aspects tend to get neglected while the focus is entirely on the more intuitive aspects. To cater to this, there is, first and foremost, a need for a closer dissection of the operation of human emotions.

Emotions are ingrained in every human being and inevitably play a significant role in every action they take. Every human choice is influenced by one or more emotions, whether consciously or unconsciously [42]. Given this control of emotions over humans, it has been identified that an understanding of human emotions will have to be instilled in artificial systems to actualize their relevance in society. Emotion recognition has been gaining a growing significance in technology, as industries are looking for ways to make their products more interactive, which requires them to understand the user emotions. Emotion recognition has also come to be a vital part of emerging monitoring systems in the limelight, such as driver status monitoring [35], [43] to help judge the state of the drivers to determine whether or not they are drowsy; and engagement monitoring [37] wherein interactive systems can recommend alternate activities if users are disinterested.

The apparent first feature of detecting an emotion might justifiably come across as facial expressions [2]. They are one of the most prominent indicators of emotions and cause intuitive inference for humans [21]. For this reason, in the present scenario, most works delving into emotion recognition systems focus on using facial expression information to categorize the emotion a person is displaying [33], and that is where the mindlessness seems to surface. Owing to the complexity of human cognition and emotion system, it seems rather reductionist to base an entire emotion recognition system only on facial expressions. Since for an emotion recognition system, aspects of the human emotion interpretation are being fed into the system, it becomes essential to acknowledge environmental context as an implicit yet extremely instrumental aspect in which the respective emotion might be placed [18]. Be it the dynamic aspect, wherein the context is temporal, or the static aspect, wherein the context is environmental, it composes the social context,

This work was not supported by any organization

which plays a central role in identifying the emotions. In the absence of a suggested context, facial expressions can often be misleading in determining a person's emotion. The facial expression might be kept constant, but once the temporal and environmental factors about it are switched, there is a high possibility of a drastic change in the emotion being detected [3]. For example, an open mouth expression can be interpreted as surprise, fear, or even happiness. However, the association of some context to the expression can add a definite meaning to the situation with respect to the emotion at play.

To address both the temporal and the environmental context, this work proposes the Indian Contextual Emotion Recognition (ICER) dataset of video clips that consist of information regarding the spatio-temporal context as well as the facial expressions. We employ Ekman's basic emotions (happiness, sadness, anger, disgust, fear, surprise, neutral) [11], [12] as the categories for classification, with valence and arousal being the continuous measures of the emotions.

Furthermore, several recent studies have shown that the existing face-based deep neural networks have a strong bias against demographic sub-groups [38], [15]. Such biases are prominently introduced due to biased and skewed datasets [22]. Acknowledging the scarcity of such work in the Indian context, the data set is based on the Indian context to enable emotion recognition in the Indian cultural context, which, as expected, will vary across different cultures [13].

II. Related Work

A. Background

The initial works used manual handcrafted features such as local binary patterns for performing emotion recognition [36], [44]. The handcrafted features were later replaced by deep neural networks [14], [26], [27], resulting in a significant performance improvement. Nevertheless, these traditional methods failed to achieve satisfactory results in the presence of ambiguous and indistinguishable faces. One of the prominent reasons for this failure is that these methods relied solely on facial expressions to detect emotions, believing that facial expressions are one of the most prominent indicators of emotions and cause intuitive inference for humans. However, we humans not only consider the facial expressions but heavily rely on the situational and surrounding contexts, such as the place where the event is taking place or the subject's gait and actions for identifying emotion[4], [1]. Hence, it becomes necessary to incorporate context-awareness in the emotion recognition models. In [30], two different models are employed, one solely for recognizing facial expressions and the other for analyzing the hand and the body posture. The resulting features from these models are fed to a third model, which performs emotion recognition based on both facial and posture information. In [7], event, object and scene are recognized using deep learning detectors, which serve as the context information and result in improvement in the emotion recognition task. This work was one of the first to include events as contextual information for emotion recognition. In [25], a network with two-stream architecture is proposed to perform face and context encoding. Context encoding is done by hiding human faces and utilizing the relevant context regions based on attention mechanism, which assists the network through reduced ambiguity and improved accuracy in emotion recognition. Facial encoding is performed using an architecture based on convolutional neural networks (CNNs) that are well suited for spatiotemporal feature representation. Several recent works have used attention to improve model efficacy. For instance, a CNN model with an attention mechanism is proposed in [27] that successfully performs emotion recognition, despite the presence of occlusions over the face.

B. Emotion Recognition Datasets

The traditional emotion recognition datasets such as the GENKI [40] and the University of California, Davis, Set of Emotion Expressions (UCD-SEE) [39] were collected in lab settings. AffectNet [33] is an image dataset that is collected from web searches by querying emotion-related keywords in different languages. The datasets Acted Facial Expressions In The Wild (AFEW) [23], Static Facial Expression in the Wild (SFEW) [9], and HAPPEI [8], [10] were collected as a part of the Emotion Recognition in the Wild (EmotiW) challenges. The HAPPEI dataset primarily focused on emotion recognition on a group scale. The inceptions of works that have started using context for emotion recognition tasks have generated a dire need for datasets with context. The Interactive Emotional Dyadic Motion Capture (IEMOCAP) database [6] is one of the first few emotion recognition datasets with context. IEMOCAP database was built by the recording of both scripted and spontaneous conversations of ten actors. The dataset consists of a lower number of emotion classes (only four, which were: angry, happy, neutral, and sad) and has limited context information [32]. Emotions In Context (EMOTIC) [24] is a database of images collected from pre-existing datasets for image captioning, segmentation and detection such as Microsoft Common Objects in Context (MS-COCO) [29] and ADE20K [45]. It also includes the manual web searched images. EMOTIC dataset consists of twenty-six emotion classes. Context-Aware Emotion Recognition (CAER) [25] is a dataset that consists of video clips from 79 TV shows. Similarly, CAER-S [25] is a collection of static images and is a subset of static images of the CAER dataset. Both CAER and CAER-S datasets consist of seven emotion classes. The recently proposed GroupWalk dataset [32] is a collection of videos of people walking in the crowd. It consists of four emotion categories. It is one of the very few datasets that have videos from uncontrolled settings and emotion class labeling for both faces and gaits. The information of the datasets is summarized in Table I.

There is a dearth of dynamic datasets, especially in the Indian social context, which makes it difficult to generalize the already existing models on data from the Indian ethnicity as the cultural differences are significant, and these disparities will not do justice to the performance of the available models. To address these existing limitations, we

Context	Dataset	Data Type	Dataset Size (Number of images or videos)	Number of Classes
No	AffectNet [33]	Image	450,000	8
	AFEW [23]	Video	1,809	7
Yes	IEMOCAP [6]	Video	-	4
	EMOTIC [24]	Image	18,316	26
	CAER-S [25]	Image	70,000	7
	CAER-S [25]	Video	13,201	7
	GroupWalk dataset [32]	Video	45	4
	ICER (ours)	Video	(Labeled) 4,533	7
			(Un-labeled) 7,913	
			(Total) 12,446	

TABLE I. Comparison of existing datasets for emotion recognition tasks.

propose the Indian Contextual Emotion Recognition (ICER) dataset as the first dynamic dataset from the Indian ethnicity. Moreover, there is a dire need for generalizable models that work on various subjects, irrespective of ethnic background. We believe that training and testing the models on datasets with ethnic diversity will help build better performing and more generalist emotion recognition models. To this end, the proposed dataset will be a step forward in building a generalized dataset consisting of subjects from various backgrounds and ethnicities.

III. Challenge Tasks

A. Task 1: Classification using strongly annotated videos

The first task was to devise a supervised approach to perform classification solely on the strongly annotated set of videos using the training set. This task required the participants to devise a supervised approach to predict the three labels of the clips of the ICER dataset (refer section V for the label information). The dataset contains 2892 training samples, 620 validation samples and 620 testing samples. The first task is depicted pictorially in 1.

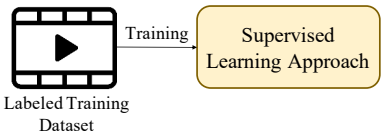

Fig. 1. Overview of our the Task 1: Classification using strongly annotated videos

B. Task 2: Semi-supervised classification using both strongly annotated and unlabeled videos

The second task was to use the training set along with all of the unlabeled data and create a semi-supervised classifier to predict the labels of the test set. This task required the participants to use the training set and the unlabeled set to devise a semi-supervised approach for the prediction of the three labels of the unlabeled dataset. The dataset has 7913 unlabeled clips. The second task is depicted pictorially in 2.

IV. Dataset and Annotations

We generate the Indian Contextual Emotion Recognition (ICER) dataset, consisting of 12,446 video clips. Each video

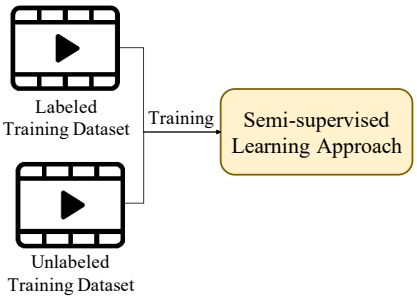

Fig. 2. Task 2: Semi-supervised classification using both strongly annotated and unlabeled videos

in the dataset has a length between 2 to 10 seconds and comprises subjects of Indian ethnicity. The videos have been taken from 327 Indian movies, web series and short films freely available on official YouTube channels with creative commons license. The data was collected by 12 recruits, who were incentivized appropriately with subscriptions for video streaming platforms according to the number of videos collected by them. The recruits were given a detailed protocol document listing the inclusion and exclusion criteria for the video clips. Only those videos were collected that adhered to the following:

- Were available on official YouTube channels with creative commons license.
- Belonged solely to Indian movies / Hindi TV shows / Indian web series.
- Length of each video was within the range of 2-10 seconds.
- A single subject was in focus; that is, throughout the clip, there should be only one target person whose face is visible, other people may be present, but their faces should not be visible in the clip.
- Did not contain unclear face shots, but low light shots were allowed.
- Could include side shots that did not exceed 45 degrees.
- Preferably included the whole scene to amplify the contextual information.
- Subjects emoting via the body but not the face were also accepted.
- Minimal music in the background, which meant no music videos were allowed.

- Did not contain profanity words.

We use YouTube Trimmer[1], an online application for clipping video intervals from the videos and providing the required links for the same during the annotation process. We provided sample clips to further clarify the protocol. The videos were downloaded using the *pafy* library[2]. The first 5000 videos were manually verified and annotated. The verification of the data, carried out by us, ensured that the protocol for data collection was followed.

Set	Number of Samples
Task 1	
Training Set	2893
Validation Set	620
Testing Set	620
Task 2	
Unlabeled Set	7913

TABLE II. The number of videos in training, validation and testing sets for both the tasks.

Class	Emotion	Train set	Validation set
1	Anger	683	136
2	Fear	154	40
3	Disgust	45	8
4	Surprise	131	30
5	Neutral	499	107
6	Sadness	649	144
7	Happiness	731	155
Total	–	2892	620

TABLE III. Dataset Distribution for the Annotated Dataset. Data points for Emotion class, consisting of seven emotion classes.

Class	Train set	Validation set
1	496	110
2	1035	198
3	717	175
4	487	107
5	157	30
Total	2892	620

TABLE IV. Dataset Distribution for the Annotated Dataset. Data points for Valence class. It has five classes, where a higher class value denotes a higher positivity of a given. For instance, class 1 denotes a video is negative, 5 denotes a video is positive and 3 denotes a neutral video.

A. Annotations

Post verification, there were 4533 labeled videos and 7913 weakly annotated videos (kindly refer Table II). The strong annotation surveys were carried out using Google forms, and the clips were randomized within pages and sets to remove any bias.

Each of the strongly annotated clips was annotated on Ekman's basic emotion categories [11], [12], which are "happiness", "sadness", "anger", "fear", "surprise", and "disgust", in addition to "neutral", along with valence and arousal. The emotions, valence and arousal, were annotated on the Likert scale of 1 to 5, with 1 and 5 being the lowest and the highest intensity, respectively [28]. The valence scale indicates whether the video is "negative" or "positive", with 1 as "negative", 3 as "neutral" and 5 as "positive". The arousal scale indicates the extent of excitation the participant experienced as a result of watching the video, with 1 being the lowest excitation and 5 being the highest. A total of 16 annotators took part in the survey with consent. Kindly refer Table III, Table IV and Table V, respectively, for emotion class-wise, valence class-wise and arousal class-wise dataset distribution.

The annotators were asked to take into consideration: (i) the expression of the subject, (ii) any contextual information from the surrounding environment, the subject's body actions, or (iii) any other audio-visual cues within the video for annotation. Please note that each video was annotated by three annotators. The annotators filled out their responses independently and were unaware of the responses of other annotators.

[1]Application available at : www.youtubetrimmer.com

[2]Implementation available at: `https://github.com/mps-youtube/pafy`

B. Data Pre-processing

Fleiss' Kappa score [41] was used as an indicator for the inter-annotator agreement. Given the range of -1 to 1 for Kappa scores, any video scoring a Kappa value less than 0 was rejected, as it indicated poor agreement between the annotators for that particular video sample. Moreover, in case a video had less than three annotations, it was removed from the dataset. For each of the remaining strongly annotated videos, the intensity of emotion categories, valence and arousal were replaced by a single vector containing the median values of the three annotations. Finally, 4533 strongly annotated videos remained, which combinedly had 496125 frames. All videos were then converted to a uniform resolution of 144x80 pixels (16:9 aspect ratio) and post padded with zeros to make each video the same size.

The entire unlabeled set and the training labeled set were made available in the first phase. The validation set without the labels were made available from the first phase as well. We provided the labels for the validation set and the test set (without the labels) in the second phase. The database is available on the platform used to host the competition[3].

C. Challenges and Complexity of ICER

In this section, we discuss the complexity of the ICER dataset. We believe that it is a challenging dataset to work with. Following are the salient features and challenges of the dataset:

1) Although the sentiment or overall effect of the video is a particular emotion, the frames at different time instances depict different emotions. For instance, in

[3]Dataset available at : Contextual Emotion Learning Challenge

Class	Train set	Validation set
1	1318	272
2	823	374
3	494	97
4	197	45
5	60	19
Total	2892	620

TABLE V. Dataset Distribution for the Annotated Dataset. Data points for Arousal class. It has five classes. The arousal scale indicates the extent of excitation experienced as a result of watching the video. That is, class 1 denotes the videos with the lowest excitation, whereas class 5 denotes the videos with the highest excitation.

video ID 1461, where at the beginning, the subject is happy (smiling), and then at the end shows anger.

2) The dataset is different from both "in the wild" and "in the lab" settings. In the case of movies or television shows, actors are asked to perform in a certain way so that the emotion of the scene is intensified.
3) In movies, perceived emotion is often a combined result of:
   a) background music,
   b) camera angle and position, and
   c) language.

   Hence it is challenging to predict emotion, based only on visual information.
4) In a cinema setting, actors and actresses are often talking. In that case, an open mouth pose does not necessarily mean surprise or fear.
5) The overall emotion perceived from videos can be dependent on only a particular part of the video, which means the rest of the video frames are irrelevant or will add noise if fed to a deep neural network.
6) The dynamics of the subject also have a significant impact on perceived emotion. A sudden movement causes intense emotion, which may not be captured very well by deep learning models.
7) There is a class imbalance in the dataset: video clips with disgust annotation are less frequent than those of the happy class.
8) In some videos, different subjects are present at different time instances. It could be difficult for deep models to learn temporal information in these cases.
9) A few videos have multiple subjects in a single frame, which adds complexity to the problem.
10) Annotations are subjective. Perceived emotion and its intensity may differ from person to person.
11) The videos in the dataset vary considerably in length. This imposes an additional challenge to developing deep learning models on this dataset.
12) In some cases, actions, rather than facial expressions, convey the emotion.

V. Evaluation

The submissions will be judged based on the following three outputs: (i) Emotion, (ii) Valence, and (iii) Arousal. Each participating team is required to submit a comma-separated values (CSV) file named 'predictions.csv'. The prescribed format for the submission file was as follows: (i) The file should contain a header row with four columns. (i) The header row should be as follows: [Video ID, Emotion, Arousal, Valence]. (iii) Each subsequent row in the file should contain four values, with the Video ID and the corresponding predictions of the emotion, the valence rating, and the arousal rating. We use the average $\overline{F_1}$ score as the evaluation criteria. The F_1 for each output is calculated as:

$$\begin{aligned} F_1 &= \frac{2}{\text{recall}^{-1} + \text{precision}^{-1}} \\ &= 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} \\ &= \frac{\text{TP}}{\text{TP} + \frac{1}{2}(\text{FP} + \text{FN})} \end{aligned}$$

where TP, FP and TN are number of True Positive, False Positive and True Negative samples, respectively. The average $\overline{F_1}$, is then calculated as:

$$\overline{F_1} = \frac{F_1^{\text{emotion}} + F_1^{\text{arousal}} + F_1^{\text{valence}}}{3}$$

where $F_1^{\text{emotion}}, F_1^{\text{arousal}}, F_1^{\text{valence}}$ are individual F_1 scores of the outputs Emotion, Arousal, and Valence respectively.

VI. Conclusion and Future Work

In this work, we propose an Indian ethnicity-based dataset for context-aware emotion tasks. We believe that our dataset will open new research paths in building context-aware and generalized emotion recognition models. It will also help future research works perform cross dataset testing to improve the generalizability of the emotion recognition models. The scores obtained by the participants point out that the proposed ICER dataset is highly challenging. Moreover, we also observed that the scores obtained by the participants in Task 2 (refer III-B) were not up to the mark. As a future work, it would be interesting to explore the utility of unlabeled data in the ICER dataset using semi-supervised to make emotion recognition models better. Even though extensively used, there are disagreements regarding the universality of Ekman's classification [20]. It would be interesting to explore this direction.

Our work hypothesizes that emotions are contextual and hold little meaning in the absence of a context they are set in. With this hypothesis, we hope to contribute to future research works by making the understanding and hence the process of emotion recognition more mindful. Since this work is limited to visual information of the data used, we also hope to focus on the auditory information of the data in the future to create a more holistic system. It has been seen audio and visual modalities complement each other for similar tasks. For instance, audiovisual speech recognition [31] where we have both the modalities provides better efficacy compared to audio speech recognition [16] and visual speech recognition [17] where only one modality is present. It would

be interesting to study the application of biological signals such as respiration rate [5], remote photoplethysmography (RPPG) [34] and facial micro-expressions [19] for emotion recognition. The size of the videos in the dataset is short as of now, but we plan to extend this to have a larger clip size for each dataset. This will enable us to have multiple annotations for a single long clip. The multiple annotations can contribute to having a deeper contextual meaning of the clips. We can consider the context of all the previous frames to get the results for the current frame.

We use F_1 score as an evaluation metric for both the tasks. As a future work we will explore other suitable metrics for evaluating such tasks. One such score is Unweighted Average Recall (UAR), which is also known as the balanced accuracy of the system. This metric proves to be more useful compared to F_1 score and vanilla accuracy, especially in the case of unbalanced datasets (such as proposed dataset ICER) [18]. It is given as

$$UAR = \frac{1}{C} \cdot \sum_{c=1}^{C} \frac{\mathrm{TP}_c}{N_c}$$

where C is the total number of classes, TP_c denotes number of True Positives for class c and N_c denotes the total number of samples present in class c.

The sincere motivation behind this project is to not only making emotion recognition a wider field of research but also to contribute to the larger picture of acknowledging the immediate need for social sciences in technological advancements. Blending the proper means of both the fields can be foreseen, giving rise to a truly productive end.

VII. Acknowledgments

The authors gratefully acknowledge the contribution of the annotators for their hard-work and dedication towards building the dataset.

References

[1] E. M. Aminoff, K. Kveraga, and M. Bar. The role of the parahippocampal cortex in cognition. *Trends in cognitive sciences*, 17(8):379–390, 2013.

[2] P. Babajee, G. Suddul, S. Armoogum, and R. Foogooa. Identifying human emotions from facial expressions with deep learning. In *2020 Zooming Innovation in Consumer Technologies Conference*, pages 36–39, 2020.

[3] L. F. Barrett, R. Adolphs, S. Marsella, A. M. Martinez, and S. D. Pollak. Emotional expressions reconsidered: Challenges to inferring emotion from human facial movements. *Psychological science in the public interest*, 20(1):1–68, 2019.

[4] L. F. Barrett, B. Mesquita, and M. Gendron. Context in emotion perception. *Current Directions in Psychological Science*, 20(5):286–290, 2011.

[5] L. Birla and P. Gupta. Patron: Exploring respiratory signal derived from non-contact face videos for face anti-spoofing. *Expert Systems with Applications*, 187:115883, 2022.

[6] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan. IEMOCAP: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42(4):335–359, 2008.

[7] C. Chen, Z. Wu, and Y.-G. Jiang. Emotion in context: Deep semantic feature fusion for video emotion recognition. In *ACM international conference on Multimedia*, pages 127–131, 2016.

[8] A. Dhall, R. Goecke, and T. Gedeon. Automatic group happiness intensity analysis. *IEEE Transactions on Affective Computing*, 6(1):13–26, 2015.

[9] A. Dhall, R. Goecke, S. Lucey, and T. Gedeon. Static facial expression analysis in tough conditions: Data, evaluation protocol and benchmark. In *IEEE International Conference on Computer Vision Workshops*, pages 2106–2112, 2011.

[10] A. Dhall, J. Joshi, I. Radwan, and R. Goecke. Finding happiest moments in a social context. In *Asian Conference on Computer Vision*, pages 613–626. Springer, 2012.

[11] P. Ekman. Are there basic emotions? *Psychological Review*, 99(3):550–553, 1992.

[12] P. Ekman. An argument for basic emotions. *Cognition & emotion*, 6(3-4):169–200, 1992.

[13] H. A. Elfenbein, M. K. Mandal, N. Ambady, S. Harizuka, and S. Kumar. Cross-cultural patterns in emotion recognition: Highlighting design and analytical techniques. *Emotion*, 2(1):75, 2002.

[14] C. Fabian Benitez-Quiroz, R. Srinivasan, and A. M. Martinez. Emotionet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild. In *IEEE conference on computer vision and pattern recognition*, pages 5562–5570, 2016.

[15] R. V. Garcia, L. Wandzik, L. Grabner, and J. Krueger. The harms of demographic bias in deep face recognition research. In *2019 International Conference on Biometrics (ICB)*, pages 1–6, 2019.

[16] A. Graves and N. Jaitly. Towards end-to-end speech recognition with recurrent neural networks. In *International conference on machine learning*, pages 1764–1772. PMLR, 2014.

[17] A. K. Gupta, P. Gupta, and E. Rahtu. FATALRead - Fooling visual speech recognition models. *Applied Intelligence*, 2021.

[18] P. Gupta. MERASTC: Micro-expression recognition using effective feature encodings and 2D convolutional neural network. *IEEE Transactions on Affective Computing*, 2021.

[19] P. Gupta, B. Bhowmick, and A. Pal. Exploring the feasibility of face video based instantaneous heart-rate for micro-expression spotting. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2018.

[20] D. Heaven. Why faces don't always tell the truth about feelings, 2020.

[21] M. Imani and G. A. Montazer. A survey of emotion recognition methods with emphasis on e-learning environments. *Journal of Network and Computer Applications*, 147:102423, 2019.

[22] K. Karkkainen and J. 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.

[23] J. Kossaifi, G. Tzimiropoulos, S. Todorovic, and M. Pantic. Afew-va database for valence and arousal estimation in-the-wild. *Image and Vision Computing*, 65:23–36, 2017.

[24] R. Kosti, J. M. Alvarez, A. Recasens, and A. Lapedriza. Context based emotion recognition using EMOTIC dataset. *IEEE transactions on pattern analysis and machine intelligence*, 42(11):2755–2766, 2019.

[25] J. Lee, S. Kim, S. Kim, J. Park, and K. Sohn. Context-aware emotion recognition networks. In *IEEE/CVF International Conference on Computer Vision*, pages 10143–10152, 2019.

[26] S. Li, W. Deng, and J. Du. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2852–2861, 2017.

[27] Y. Li, J. Zeng, S. Shan, and X. Chen. Occlusion aware facial expression recognition using cnn with attention mechanism. *IEEE Transactions on Image Processing*, 28(5):2439–2450, 2018.

[28] R. Likert. A technique for the measurement of attitudes. *Archives of Psychology*, 22 140:55–55, 1932.

[29] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft COCO: Common objects in context. In *European Conference on Computer Vision*, pages 740–755. Springer, 2014.

[30] P. Metri, J. Ghorpade-Aher, and A. Butalia. Facial emotion recognition using context based multimodal approach. *International Jorunal on Interactive Multimedia and Artificial Intelligence*, 1:12–15, 2011.

[31] S. Mishra, A. K. Gupta, and P. Gupta. DARE: Deceiving audio–visual speech recognition model. *Knowledge-Based Systems*, 232:107503, 2021.

[32] T. Mittal, P. Guhan, U. Bhattacharya, R. Chandra, A. Bera, and D. Manocha. Emoticon: Context-aware multimodal emotion recognition using frege's principle. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14234–14243, 2020.

[33] A. Mollahosseini, B. Hasani, and M. H. Mahoor. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing*, 10(1):18–31, 2019.

[34] E. M. Nowara, A. Sabharwal, and A. Veeraraghavan. Ppgsecure: Biometric presentation attack detection using photopletysmograms.

In *IEEE International Conference on Automatic Face & Gesture Recognition*, pages 56–62. IEEE, 2017.

[35] M. N. RASTGOO, B. Nakisa, A. Rakotonirainy, V. Chandran, and D. Tjondronegoro. A critical review of proactive detection of driver stress levels based on multimodal measurements. *ACM Computing Surveys*, 51(5):88:1–88:35, 2018.

[36] C. Shan, S. Gong, and P. W. McOwan. Facial expression recognition based on local binary patterns: A comprehensive study. *Image and vision Computing*, 27(6):803–816, 2009.

[37] P. Sharma, S. Joshi, S. Gautam, V. Filipe, and M. J. C. S. Reis. Student engagement detection using emotion analysis, eye tracking and head movement with machine learning. *arXiv:1907.00193 [cs]*, 2019.

[38] P. Terhörst, M. L. Tran, N. Damer, F. Kirchbuchner, and A. Kuijper. Comparison-level mitigation of ethnic bias in face recognition. In *International Workshop on Biometrics and Forensics*, pages 1–6. IEEE, 2020.

[39] J. L. Tracy, R. W. Robins, and R. A. Schriber. Development of a facs-verified set of basic and self-conscious emotion expressions. *Emotion*, 9(4):554, 2009.

[40] `http://mplab.ucsd.edu`. The MPLab GENKI database.

[41] `https://en.wikipedia.org/wiki/Fleiss%27_kappa`. Fleiss' kappa, 2021.

[42] S. Whitener. Council Post: How Your Emotions Influence Your Decisions.

[43] S. Zepf, J. Hernandez, A. Schmitt, W. Minker, and R. W. Picard. Driver emotion recognition for intelligent vehicles: A survey. *ACM Computing Surveys*, 53(3):64:1–64:30, 2020.

[44] L. Zhong, Q. Liu, P. Yang, B. Liu, J. Huang, and D. N. Metaxas. Learning active facial patches for expression analysis. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2562–2569. IEEE, 2012.

[45] B. Zhou, H. Zhao, X. Puig, T. Xiao, S. Fidler, A. Barriuso, and A. Torralba. Semantic understanding of scenes through the ADE20k dataset. *International Journal of Computer Vision*, 127(3):302–321, 2019.