EmotiW 2020: Driver Gaze, Group Emotion, Student Engagement and Physiological Signal based Challenges

Abhinav Dhall
Monash University/
Indian Institute of Technology Ropar
abhinav.dhall@monash.edu

Garima Sharma
Monash University
garima.sharma1@monash.edu

Roland Goecke
University of Canberra
roland.goecke@canberra.edu.au

Tom Gedeon
Australian National University
tom.gedeon@anu.edu.au

ABSTRACT

This paper introduces the Eighth Emotion Recognition in the Wild (EmotiW) challenge. EmotiW is a benchmarking effort run as a grand challenge of the 22nd ACM International Conference on Multimodal Interaction 2020. It comprises of four tasks related to automatic human behavior analysis: a) driver gaze prediction; b) audio-visual group-level emotion recognition; c) engagement prediction in the wild; and d) physiological signal based emotion recognition. The motivation of EmotiW is to bring researchers in affective computing, computer vision, speech processing and machine learning to a common platform for evaluating techniques on a test data. We discuss the challenge protocols, databases and their associated baselines.

CCS CONCEPTS

• **Computing methodologies** → *Machine learning*; **Computer vision**.

KEYWORDS

Affective computing, automatic human behavior analysis, group emotions, driver gaze prediction, student engagement

ACM Reference Format:
Abhinav Dhall, Garima Sharma, Roland Goecke, and Tom Gedeon. 2020. EmotiW 2020: Driver Gaze, Group Emotion, Student Engagement and Physiological Signal based Challenges. In *Proceedings of the 2020 International Conference on Multimodal Interaction (ICMI '20), October 25–29, 2020, Virtual event, Netherlands.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3382507.3417973

ICMI '20, October 25–29, 2020, Virtual event, Netherlands

ACM ISBN 978-1-4503-7581-8/20/10…$15.00
https://doi.org/10.1145/3382507.3417973

1 INTRODUCTION

The Eighth Emotion Recognition in the Wild[1] (EmotiW) is a grand challenge in the 22nd ACM International Conference on Multimodal Interaction (ICMI) 2020. EmotiW is a benchmarking resource for researchers to evaluate their human behavior analysis systems on a common data accompanied with a fixed evaluation protocol. This year EmotiW challenge consists of four sub-challenges: a) driver gaze prediction; b) audio-visual group-level emotion recognition; c) engagement prediction in the wild, and d) physiological signal based emotion recognition.

With newer databases being introduced in the research community and the astonishing progress in deep learning techniques, it is important that affect analysis systems are compared across each other for assessing the current state-of-the-art progress in the community. To this end, the first EmotiW challenge [5] was organised in 2013 as part of ACM ICMI. The task here was video-level emotion recognition in the wild on the Acted Facial Expressions in the Wild (AFEW) database [6]. Here, 'In the wild' means the varied environments in the data due to different illumination, occlusion and subjects from a large age range. Recently, there have been other important benchmarking efforts in the community too. The audio/visual emotion challenge [22] focuses mainly on depression analysis. The pain estimation challenge [3] focuses on analysing data for chronic pain. The other two relevant challenges are the micro-expression detection and localisation [12] and the sentiment analysis challenge [20].

In EmotiW 2015 [8], a new sub-challenge task of image-level facial expression classification was added. Further, in 2016, the image-level sub-challenge was superseded by group-level emotion recognition task in images. The motivation of this new sub-challenge stemmed from images and videos of social events, which are uploaded on the social network platforms. In many cases, these images and videos contain multiple subjects and hence, it is of interest to predict the perceived emotion of a group of people. The Group AFfect Database (GAF) was used in this challenge. This year, the sub-challenge of group-level emotion recognition has been updated to audio-visual analysis of group of people. The

[1] https://sites.google.com/view/emotiw2020

Figure 1: VGAF database [19] - the three rows show class-wise sample frames from videos.

Video based Group AFfect (VGAF) database is used for this sub-challenge.

Online education has become prominent in the past few years. One of the major challenges for an instructor in online lecture delivery is the difficulty in assessing students' engagement due to the lack of face to face interaction. To this end, in EmotiW 2018, a new sub-challenge of student engagement prediction was introduced. This task has become even more important in the current COVID19 times as most of the teaching is now online. EmotiW 2020 also has engagement prediction as a sub-challenge. The database for this task is the EngageWild [15].

EmotiW 2020 has two more new sub-challenges: Driver Gaze prediction in the wild and the other is Physiological Signal based Emotion Recognition. The sub-challenges, their data and baselines are discussed below in detail.

2 CHALLENGE PROTOCOL

The data in the four sub-challenges is divided into *Train*, *Validation* and *Test* sets. During February/March 2020, the *Train* and *Validation* sets and corresponding labels were shared with the participants. In total, EmotiW witnessed over 90 team registrations. During June 2020, the unlabelled *Test* sets were shared with the participants. Each team could submit upto five sets of labels, per challenge, for evaluation. At the end of the *Test* phase, the best performance across the labels sets submitted by a team was assigned to be the team's final entry. The top three teams in each sub-challenge were asked to share their code/library for evaluation with us. All teams were invited to submit papers describing their techniques. The papers were thoroughly reviewed and accepted as a factor of the proposed method's relative performance in the challenge and method's novelty.

3 AUDIO-VISUAL GROUP-LEVEL EMOTION

Group-level affect analysis can be performed either on images or on videos. One of the early image-based group affect analysis work focused on finding the happiness intensity of a group of people [4]. The introduction of Group AFfect database (GAF) [7] in a series of EmotiW challenges has provided an opportunity for researchers to work on this task. Image based group-level affect recognition methods usually combine the face-level information of each individual with the contextual/background information [1, 18, 23].

Facial expressions are temporal in nature and hence, videos can better represent the emotional change of each individual in a group across time. Video-based group affect recognition also allows one to use the audio features which provides an additional modality along with the frame-level information. To leverage this temporal information, Sharma et al. [19] proposed the Video-level Group AFect (VGAF) database. VGAF was curated by keyword based searching for videos on YouTube with creative commons license. The keywords such as 'interview', 'festival', 'party', 'silent protest', 'violence', 'argument', 'birthday', 'wedding', 'meeting', and 'fighting' were used in the search corresponding to social events, which generally contain group of people. The downloaded videos were cropped to segments of ~5 seconds. The data was labelled with three emotion classes - *Positive*, *Neutral* and *Negative* (corresponding to the Valence axis). These segments gave us a total of 4183 samples (*Train* - 2661, *Validation* - 766 and *Test* - 756). Figure 1 shows three video frames from the VGAF database.

Baseline: The baseline on the VGAF database is computed by combining the audio and visual information. First, an Inception V3 network was trained on the image-based GAF database [10]. The database contains 15K images with the

Figure 2: DGW database [11] - sample frames with subjects looking at different zones inside the car.

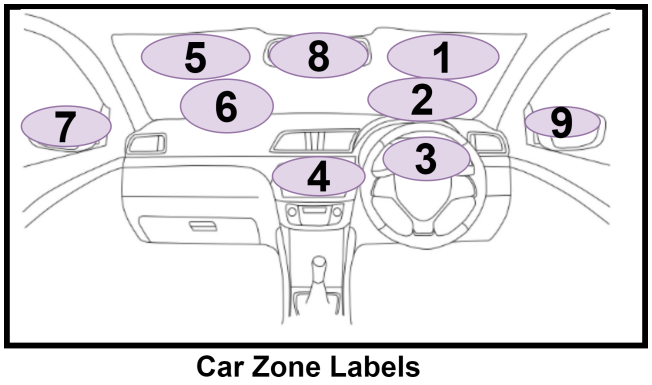

Figure 3: The figure shows the location of the nine gaze classes in the DGW database.

same three emotion labels as the VGAF database. The choice of using the Inception V3 network was based on it's good performance on image-level group emotion recognition. Each sample in VGAF contains ~150 frames. For each frame, 4096 dimension feature representation was extracted using the Inception V3 network. These extracted features representing the frame were then used for training on a four layered Long Short Term Memory (LSTM) network [13] containing 256, 512, 1024 and 2048 kernels, respectively.

To extract the audio level information, INTERSPEECH 2013 ComParE challenge features were extracted for each video using the OpenSMILE toolkit [9]. These features extract low level descriptors from a given audio. The 6373 dimensional GeMPAS features were then used as an input to a fully connected network with 128, 256, 512 and 1024 kernels. The 2048 and 1024 dimensional output from the LSTM (visual) and fully connected layers (audio), respectively, was concatenated together. The training was performed using categorical cross entropy loss function with SGD optimizer and 0.01 learning rate. Finally, softmax activation function was used for final emotion prediction. Classification accuracy is used as the evaluation metric. The baseline method achieved 51.30% and 47.88% accuracy on the validation and test set, respectively.

4 DRIVER GAZE PREDICTION

Estimation of a driver's gaze is an important task for in-cabin monitoring in a smart car. The car can use this information for initiating a hand over from the driver and/or warn a driver in case he/she is not attentive. Recent relevant works [14, 16, 21] have used visual analysis of head pose, eyes and face of driver with traditional machine learning and/or deep learning techniques for predicting driver gaze. Driver gaze prediction is a new sub-challenge in EmotiW 2020. The database used in the sub-challenge is the Driver Gaze in the Wild[2] (DGW) database [11].

Figure 2 shows the sample images from the database. The data has been collected in a Hyundai car with different subjects at the driver's position. The inside of the car cabin was divided into nine zones, each corresponding to a class (as shown in Figure 3). Stickers representing different zones in the car were posted in different locations in car. The nine car zones represent areas of back mirror, side mirrors, radio, speedometer and windshield. The driver of the car can look at one of these zone and the task of this sub-challenge is to predict such zone.

The recording sensor used was a Microsoft Lifecam RGB camera, which also contains a microphone. The data contains 330 subjects (247 males and 91 females) within the age range of [18-63] years. Subjects were asked to look at the sticky notes pasted on the car console and read the zone number on it. Speech to text converter was used to automatically compute the initial labels. Frequency and energy components of speech were analysed to prune the labels. For finer details, please refer to the DGW paper [11]. The data is challenging as the recording was performed at different times of the day. There are samples, which have different illumination sources such as street lights at different locations. The database is divided into *Train, Validation* and *Test* sets containing 203, 83 and 52 different subjects, respectively. This gives a total of 50484 images (*Train* - 29,448, *Validation* - 9995 and *Test* - 11041).

Baseline: The baseline for the database is computed by training an Inception V1 network using the SGD optimizer with 0.01 learning rate with $1 \times e^{-6}$ decay per epoch. Classification accuracy is used as the evaluation metric. The method gives 60.10% validation accuracy and 60.98% test accuracy for driver gaze prediction in the wild sub-challenge.

[2]https://sites.google.com/view/drivergazeprediction/

Figure 4: EngageWild database [15] - sample frames from videos showing different levels of engagement.

5 STUDENT ENGAGEMENT PREDICTION

Engagement prediction has become an active research problem in recent years. The need to move classes and meetings online has made this problem even more important in COVID19 times. Some of the relevant works [15, 24] in this direction have been using machine learning techniques for analysis of visual data. In the engagement prediction in the wild sub-challenge, the task is to predict the engagement intensity of subject in a given video. The database used for this task is EngageWild [15], which was collected by showing 'new language learning' videos to subjects. The videos were collected across different subjects and recordings were conducted at the different time of the day at different locations (sample frames - Figure 4). While watching any learning video, the engagement of a participant may change, which is observed in the form of visual markers such as yawning, too much or too less change in the eye gaze and too much change in the head pose. Hence, one can use visual feature descriptors to estimate the engagement level of the participant. The videos in EngageWild database are ~5 minutes long and are annotated for the intensity range [0-3] representing engagement mapped to [*disengaged*, *barely engaged*, *engaged* and *highly engaged*]. The database has total 264 videos which are divided as 148 for *Training*, 48 for *Validation* and 67 for the *Testing*.

Baseline: The baseline for this sub-challenge is computed by combining head pose and eye gaze information of the subjects. OpenFace library [2] is used to extract these facial features. The videos are converted to segments and each segment is represented by the standard deviation of the head movements across the frames of the given segment. These extracted features are trained using a LSTM layer and 3 fully connected layers followed by average pooling. The evaluation metric is mean squared error with respect to the ground truth engagement intensity. The baseline network achieved 0.10 and 0.15 Mean Square Error (MSE) on the validation and test set, respectively.

6 PHYSIOLOGICAL SIGNAL BASED EMOTION RECOGNITION

This is another new sub-challenge in EmotiW 2020. The task is to predict emotion of a person in a video, which is viewed by subjects. The data is in the form of physiological signals (ElectroDermal Activity (EDA)), which were collected while observers watched short clips from the AFEW database [6]. The resulting database containing physiological signals is called the Physiological AFEW (PAFEW). The labels in PAFEW are the same as that of the corresponding clips watched by the observers. The task is to predict an emotion label for each physiological signal series from seven universal emotion classes: *Anger, Disgust, Fear, Happy, Neutral, Sad* and *Surprise*. The baseline for this sub-challenge is based on the feature extraction and a three-layered network. Classification accuracy is employed as the evaluation metric. For each EDA sequence, six features corresponding to basic statistical variables (max, min, mean and variance), mean absolute difference and mean second absolute difference are extracted. These features are then trained by using a three layer fully connected neural network which achieves an accuracy of 42.08% and 27.96% on the *Validation* and *Test* sets, respectively. Further details are provided in the report [17].

7 RESULTS

In this section, we discuss the baseline results and the participating methods' performance comparison for the four sub-challenges.

Audio-Visual Group Emotion Recognition: For this task, the audio-visual baseline achieved 51.30% on the *Validation* and 47.88% on the *Test* set. The class-wise accuracy are presented in Table 1. In total 16 teams submitted labels generated by their methods for evaluation. Table 2 shows the leader-board for the sub-challenge. The progress on the task is visible as the top performing method outperforms the baseline by a margin of 28.97%.

Driver Gaze Prediction: In the driver gaze prediction task, the Inception V3 baseline gave 60.10% on the *Validation* set and 60.98% on the *Test* set. The class-wise accuracy are presented in Table 3. Total of six teams submitted *Test* set labels for evaluation and Table 4 shows the leaderboard for

Table 1: Class-wise and overall baseline accuracy (%) for the audio-visual group emotion sub-challenge.

Class	Positive	Neutral	Negative	**Overall**
Val. acc.	38.74	54.64	66.84	**51.30**
Test acc.	49.30	37.54	60.43	**47.88**

Table 2: Leaderboard for audio-visual group emotion recognition sub-challenge (accuracy in %).

Rank	Team Name	Institute	Accuracy
1	SituTech	Situ Vision Tech.	76.85
2	DD_VISION	DD Vision	70.76
3	DeepBlueAI	DeepBlueAI	69.44
4	ZBTlab	-	66.93
5	KDDIResearch	KDDI Research	66.40
6	LosEmotibos	INESC TEC	65.74
7	Stanford231	Stanford University	63.88
8	AugsBurger	Uni. of Augsburg	62.69
9	KDDI lab & NAIST	KDDI lab and NAIST	59.65
10	GlobalFeatures Only	Uni. Grenoble Alpes LIG, Inria	59.12
11	BNU	Beijing Normal Uni.	58.06
12	Cognitive_Systems	Korea University	52.51
-	Baseline	-	47.88
13	ISIA-Lab-UMONS	Uni. of Mons	46.82
14	SCUT	South China Univ. of Technology	46.82
15	UoE	Uni. of Edinburgh	45.23
16	USF_Affective _Bulls	Uni. South Florida	35.44
17	GauriD	TCS, Uni. of Augsburg	31.21

this sub-challenge. The winning method outperforms the baseline by a margin of ~21.54%.

Engagement Prediction: In this sub-challenge, the LSTM based baseline achieves 0.10 on *Validation* and 0.15 on the *Test* set. A total of 5 teams participated during the *Test* phase and submitted labels for evaluation. In EmotiW 2019, 8 teams had participated during the *Test* phase. Table 5

Table 3: Class-wise and overall baseline accuracy (%) for driver gaze prediction sub-challenge.

Zone	1	2	3	4	5	6	7	8	9	Overall
Val.	68.24	61.48	74.37	61.55	51.15	44.78	19.90	57.70	77.96	**60.10**
Test	73.07	76.65	82.52	74.23	51.79	45.52	26.61	60.35	66.73	**60.98**

Table 4: Leaderboard for driver gaze prediction sub-challenge (accuracy in %).

Rank	Team Name	Institute	Accuracy
1	DD_Vision	Didi	82.52
2	SituAlgorithm	Situ Vision Tech.	81.51
3	Overfit	Southeast Uni.	78.87
4	DeepBlueAI	DeepBlueAI	75.88
5	UDECE	Uni. of Delaware	74.57
6	X-AWARE	Uni. Augsburg	71.62
-	Baseline	-	60.98
7	USF_AFFECTIVE _BULLS	Uni. of South Florida	57.31

Table 5: Leader board for engagement prediction in the wild sub-challenge.

Rank	Team Name	Institute	MSE
1	UDECE	Univ. of Delaware	0.054
2	KDDIResearch	KDDI Research	0.061
3	USF_Affective_Bulls	Univ. of South Florida	0.065
4	DSI@UTS	Univ. of Technology Sydney	0.070
5	DDL	Didi Chuxing, Huazhong Agricultural Univ.	0.100
-	Baseline	-	0.150

shows the comparison of the participating methods and the baseline. The progress in engagement prediction techniques is evident from the large increase in the performance of the top performing method as compared to the baseline.

Physiological Signal based Emotion Recognition: For this sub-challenge, EDA based features were used to compute the baseline, which gave an accuracy of 42.96% on *Validation* set and 27.96% on *Test* set. During the *Test* phase only one team submitted the labels for evaluation and the performance accuracy was 17.44%.

8 CONCLUSION

This paper presents the details of eighth EmotiW challenge. This year EmotiW introduced three new sub-challenges: driver gaze prediction, audio-visual group emotion recognition and physiological signal based emotion recognition. In addition, engagement prediction in the wild is also included due to significance of the task in online learning. In total 29 methods were submitted for evaluation during the *Test* phase and 12 papers were accepted for the publication after peer review. The results of the participating teams show that transfer learning and deep learning based techniques are achieving state-of-the-art results.

9 ACKNOWLEDGEMENT

We are grateful to all the brave frontline workers who are working hard during this difficult COVID19 situation. We appreciate the help of the EmotiW program committee members and the reviewers. Thanks to Yang Liu (Australian National University) and Shreya Ghosh (Monash University) for help with the data. Tom Gedeon and Abhinav Dhall's research is partially supported by the Australian Research Council's grant DP190102919.

REFERENCES

[1] Asad Abbas and Stephan K Chalup. 2017. Group emotion recognition in the wild by combining deep neural networks for facial expression classification and scene-context analysis. In *International Conference on Multimodal Interaction*. ACM, 561–568.

[2] Tadas Baltrušaitis, Peter Robinson, and Louis-Philippe Morency. 2016. Openface: an open source facial behavior analysis toolkit. In *Applications of Computer Vision (WACV), 2016 IEEE Winter Conference on*. IEEE, 1–10.

[3] Nadia Berthouze, Michel Valstar, Amanda Williams, Joy Egede, Temitayo Olugbade, Chongyang Wang, Hongyin Meng, Min Aung, Nicholas Lane, and Siyang Song. 2020. Emopain challenge 2020:

Multimodal pain evaluation from facial and bodily expressions. *arXiv preprint arXiv:2001.07739* (2020).

[4] Abhinav Dhall, Roland Goecke, and Tom Gedeon. 2015. Automatic group happiness intensity analysis. *IEEE Transactions on Affective Computing* (2015), 13–26.

[5] Abhinav Dhall, Roland Goecke, Jyoti Joshi, Michael Wagner, and Tom Gedeon. 2013. Emotion recognition in the wild challenge 2013. In *Proceedings of the 15th ACM on International conference on multimodal interaction*. ACM, 509–516.

[6] Abhinav Dhall, Roland Goecke, Simon Lucey, Tom Gedeon, et al. 2012. Collecting large, richly annotated facial-expression databases from movies. *IEEE multimedia* 19, 3 (2012), 34–41.

[7] Abhinav Dhall, Jyoti Joshi, Karan Sikka, Roland Goecke, and Nicu Sebe. 2015. The more the merrier: Analysing the affect of a group of people in images. In *International Conference and Workshops on Automatic Face and Gesture Recognition*, Vol. 1. IEEE, 1–8.

[8] Abhinav Dhall, OV Ramana Murthy, Roland Goecke, Jyoti Joshi, and Tom Gedeon. 2015. Video and image based emotion recognition challenges in the wild: Emotiw 2015. In *Proceedings of the 2015 ACM on international conference on multimodal interaction*. 423–426.

[9] Florian Eyben, Felix Weninger, Florian Gross, and Björn Schuller. 2013. Recent developments in opensmile, the munich open-source multimedia feature extractor. In *International conference on Multimedia*. ACM, 835–838.

[10] Shreya Ghosh, Abhinav Dhall, and Nicu Sebe. 2018. Automatic group affect analysis in images via visual attribute and feature networks. In *2018 25th IEEE International Conference on Image Processing (ICIP)*. IEEE, 1967–1971.

[11] Shreya Ghosh, Abhinav Dhall, Garima Sharma, Sarthak Gupta, and Nicu Sebe. 2020. Speak2Label: Using Domain Knowledge for Creating a Large Scale Driver Gaze Zone Estimation Dataset. *arXiv preprint arXiv:2004.05973* (2020).

[12] Ying He, Su-Jing Wang, Jingting Li, and Moi Hoon Yap. 2019. Spotting macro-and micro-expression intervals in long video sequences. *arXiv preprint arXiv:1912.11985* (2019).

[13] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.

[14] S Jha and C Busso. 2018. Probabilistic Estimation of the Gaze Region of the Driver using Dense Classification. In *IEEE International Conference on Intelligent Transportation Systems*. 697–702.

[15] Amanjot Kaur, Aamir Mustafa, Love Mehta, and Abhinav Dhall. 2018. Prediction and localization of student engagement in the wild. In *2018 Digital Image Computing: Techniques and Applications (DICTA)*. IEEE, 1–8.

[16] M Leo, D Cazzato, T De Marco, and C Distante. 2014. Unsupervised Eye Pupil Localization through Differential Geometry and Local Self-Similarity. *Public Library of Science* 9, 8 (2014).

[17] Yang Liu, Tom Gedeon, Sabrina Caldwell, Shouxu Lin, and Zi Jin. 2020. Emotion Recognition Through Observer's Physiological Signals. *arXiv preprint arXiv:2002.08034* (2020).

[18] Alexandr Rassadin, Alexey Gruzdev, and Andrey Savchenko. 2017. Group-level emotion recognition using transfer learning from face identification. In *International Conference on Multimodal Interaction*. ACM, 544–548.

[19] Garima Sharma, Shreya Ghosh, and Abhinav Dhall. 2019. Automatic Group Level Affect and Cohesion Prediction in Videos. In *2019 8th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*. IEEE, 161–167.

[20] Lukas Stappen, Alice Baird, Georgios Rizos, Panagiotis Tzirakis, Xinchen Du, Felix Hafner, Lea Schumann, Adria Mallol-Ragolta, Björn W Schuller, Iulia Lefter, et al. 2020. MuSe 2020–The First International Multimodal Sentiment Analysis in Real-life Media Challenge and Workshop. *arXiv preprint arXiv:2004.14858* (2020).

[21] S Vora, A Rangesh, and M M Trivedi. 2018. Driver Gaze Zone Estimation using Convolutional Neural Networks: A General Framework and Ablative Analysis. *IEEE Transactions on Intelligent Vehicles* (2018), 254–265.

[22] Gou Wei, Li Jian, and Sun Mo. [n.d.]. Multimodal (Audio, Facial and Gesture) Based Emotion Recognition Challenge. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)(FG)*. 902–905.

[23] Qinglan Wei, Yijia Zhao, Qihua Xu, Liandong Li, Jun He, Lejun Yu, and Bo Sun. 2017. A new deep-learning framework for group emotion recognition. In *ACM ICMI*. 587–592.

[24] Jacob Whitehill, Zewelanji Serpell, Yi-Ching Lin, Aysha Foster, and Javier R Movellan. 2014. The faces of engagement: Automatic recognition of student engagementfrom facial expressions. *IEEE Transactions on Affective Computing* 5, 1 (2014), 86–98.