Recognizing Emotion in the Wild using Multimodal Data

Shivam Srivastava
shivam4@usf.edu
University of South Florida
Tampa, Florida

Saandeep Aathreya Sidhapur Lakshminarayan
saandeepaath@usf.edu
University of South Florida
Tampa, Florida

Saurabh Hinduja
saurabhh@usf.edu
University of South Florida
Tampa, Florida

Sk Rahatul Jannat
jannat@usf.edu
University of South Florida
Tampa, Florida

Hamza Elhamdadi
hme1@usf.edu
University of South Florida
Tampa, Florida

Shaun Canavan
scanavan@usf.edu
University of South Florida
Tampa, Florida

ABSTRACT

In this work, we present our approach for all four tracks of the eighth Emotion Recognition in the Wild Challenge (EmotiW 2020). The four tasks are group emotion recognition, driver gaze prediction, predicting engagement in the wild, and emotion recognition using physiological signals. We explore multiple approaches including classical machine learning tools such as random forests, state of the art deep neural networks, and multiple fusion and ensemble-based approaches. We also show that similar approaches can be used across tracks as many of the features generalize well to the different problems (e.g. facial features). We detail evaluation results that are either comparable to or outperform the baseline results for both the validation and testing for most of the tracks.

CCS CONCEPTS

• **Computing methodologies** → **Artificial intelligence**; *Machine learning*; Biometrics.

KEYWORDS

Gaze Detection, Group Emotion, Engagement Prediction, Physiological Signals

ACM Reference Format:
Shivam Srivastava, Saandeep Aathreya Sidhapur Lakshminarayan, Saurabh Hinduja, Sk Rahatul Jannat, Hamza Elhamdadi, and Shaun Canavan. 2020. Recognizing Emotion in the Wild using Multimodal Data. In *Proceedings of the 2020 International Conference on Multimodal Interaction (ICMI '20), October 25–29, 2020, Virtual event, Netherlands.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3382507.3417970

1 INTRODUCTION

We present our approaches to the four tracks of the eighth Emotion Recognition in the Wild Challenge (EmotiW 2020): (1) Audio-video based group emotion recognition (AV); (2) Driver gaze prediction (GC); (3) Engagement prediction in the wild (EW); and (4) Physiological signal based emotion recognition (PER). Each of the tracks are important and timely topics that have significant real-world applications. Predicting driver gaze is important for the safety of the driver, passengers, and other people on the road, as well as walking on sidewalks. If the driver's attention is diverted from the road for longer then the recommended time from the National Highway Traffic Safety Administration, the driver could lose control of the vehicle. This diverted attention could cause injuries, as well as fatal accidents [28]. Along with helping with safe driving, gaze can also be used to help assess engagement in students in classrooms (i.e. wild setting). This is an important task for teachers and professors as it can help improve their teaching, as well as improve the students' learning experience. Considering this, Thomas and Jayagopi [35] used eye gaze along with other behavior cues such as head pose and facial expressions for this task. They found that the fusion of gaze along with the others modalities, can perform better than a baseline evaluator. Another important modality is physiological data, as it has been used in important applications in the medical field. Zamzmi et al. [41] used the fusion of physiological signals along with others features such as crying and body movement, to assess neonatal pain. Their results suggest that important features can be extracted from these signals for accurate assessment.

Although these modalities and applications are important, they are largely in the context of a single person (e.g. pain assessment in one neonate). Group emotion recognition is also an important topic, as social interactions have a large influence on the elicitation of emotions [2, 29]. Van Kleef and Fischer [37] investigated the role of emotions in groups. More specifically, they look at how group-level events shape emotion, how emotional expressions are recognized by the group, and how those expressions influence the behavior of group members. Considering each of the important, broad applications that each of the EmotiW tracks facilitates, we propose new approaches for all four tracks. The contributions of this work, across all tracks, can be summarized as follows.

(1) We propose an approach to group emotion recognition that fuses optical flow and mel spectrogram features from video and audio. We investigate the positive impact of the proposed fusion approach by also conducting unimodal (e.g. audio only) experiments on the validation set.
(2) For driver zone classification, we investigate a range of modalities including, but not limited to, gaze, head pose, and

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

ACM ISBN 978-1-4503-7581-8/20/10…$15.00
https://doi.org/10.1145/3382507.3417970

facial landmarks. We conducted an ablation study showing the utility of individual features along with their fusion.

(3) We show that similar approaches can be used for multiple tracks as many features (e.g. facial features) generalize well to different problems. Specifically, we use similar features as used for driver zone classification for engagement prediction (e.g. gaze and head pose).
(4) A GAN-based discriminator ensemble is proposed for emotion recognition using physiological signals. We also compare the proposed approach with hand-crafted features used to train a random forest.

2 RELATED WORK

2.1 Group Emotion Recognition

In recent years, there have been encouraging results for group emotion recognition. Gupta et al. [18] consider both global and local information from images. The global information is captured from the entire image (i.e. group emotion) and the local information is captured from individual faces. The global information is learned through a convolutional neural network, while the local information is learned through an attention mechanism. The two branches (global and local) are finally fused, through concatenation, for the final group emotion recognition. Another interesting approach to group emotion recognition was proposed by Veenendaal et al. [38]. They proposed the use of edge detection to extract features for group emotion recognition. Using these features to train a support vector machine, they showed promising results when the group's emotion was captured while they watched sports videos.

2.2 Driver Gaze Prediction

In recent years, there has been promising work predicting driver eye gaze. Wang et al. [39] proposed an appearance-based, head pose free approach to predicting driver gaze. Using a combination of a 3D facial model, landmarks, and features extracted from eye regions they trained a random forest for this task. They showed promising results for predicting driver gaze in a real-world driving environment. Another interesting approach is the recent work from Deng et al. [9]. They proposed a Convolutional-Deconvolutional Neural Network (CDNN) for predicting drivers' gaze fixations. They showed that the proposed CDNN can predict major fixation locations, as well as detect important objects in the scene that should not be ignored (e.g. person riding a bicycle in the street).

2.3 Engagement in the Wild

In the baseline paper for the engagement in the wild track, Kaur et al. [21] detail a new dataset for student engagement and localization. To perform the engagement detection, they propose a multimodal approach that makes use of LBP-TOP [42] for spatio-temporal facial features, as well as gaze and head pose to give cues about attention and areas of interest. Using these features, they train a deep multi-instance network for engagement prediction and localization. Using this approach they detail encouraging baseline results for this track. Al-Alwani [1] also makes use of facial features, where they are used to detect the mood and subsequent engagement of students in an e-learning environment. Using facial features, a neural network was trained to detect the mood from facial expressions. Subject self-report on their engagement and mood was collected, showing a positive correlation between the mood of the students and their engagement in the material.

2.4 Physiological Signal Based Emotion Recognition

Recent works have shown promising results on physiological signal-based emotion recognition, such as the work from Fabiano et al. [13]. They proposed a new weighted approach for fusing physiological signals that outperformed state of the art for 10 elicited emotional categories in BP4D+ [43], and arousal, valence, liking, and dominance in DEAP [24]. Promising work has also been done in more specialized areas such as pain recognition. Hinduja et al. [19] looked at the physiological signals that corresponded with the most expressive parts of the face during a video sequence. This was done using facial images that were captured at the same time as the physiological signals. In their investigation, they successfully recognized the context that was used to elicit a painful response (i.e. cold compression). Along with this, their results suggest that there is a high correlation between the physiological signals and highly expressive faces, however, the correlation decreases when low expressive faces are occurring.

3 AUDIO-VIDEO GROUP EMOTION RECOGNITION

3.1 Dataset

The dataset [30] for the audio-video group emotion recognition track contains 2661 videos for the training set, 756 videos for the validation set and 756 videos for the test set. The dataset consists of many types of group activity that include videos of talk show hosts and their guest(s), crowded groups of people talking, and groups of people being interviewed. The data is divided into 3 classes – Positive, Neutral and Negative corresponding to the 3 emotion categories. The dataset is challenging due to obstructions like head and body pose variations of the people, occlusions in faces, and various indoor and outdoor environment settings.

3.2 Group Emotion Recognition Features

Features From Video: Given a group video, we extract 256 frames to represent the image-space. We then track 16 features, in each frame, using the KLT features tracker [27, 32]. This results in a 32 dimension feature vector $F_{good} = [x_1, y_1, \ldots, x_{16}, y_{16}]$, where x_i and y_i are the (x, y) coordinates of the i^{th} tracked feature. Given F_{good}, we then calculate the optical flow between consecutive frames by calculating the Δ between each corresponding set of (x, y) coordinates of the features. To account for the first frame, we calculate the Δ with itself (i.e. 0). Given the Δ for each feature between frames, we then calculate the mean across the 16 Δ values. This is done per axis (x. y), as well as overall across both x and y. We then construct a 4-dimension feature vector, per frame, $M = [x_{mean}, y_{mean}, o_{mean}, o_{mean}]$, where x_{mean} and y_{mean} are the per-axis mean, and o_{mean} is the overall mean across both x and y. We then concatenate each per-frame M in a 1024-dimension (256 × 4) feature vector $M_{final} = [x_{mean_1}, y_{mean_1}, o_{mean_1}, o_{mean_1}, \ldots, x_{mean_{256}}, y_{mean_{256}}, o_{mean_{256}}, o_{mean_{256}}]$. We then construct similar

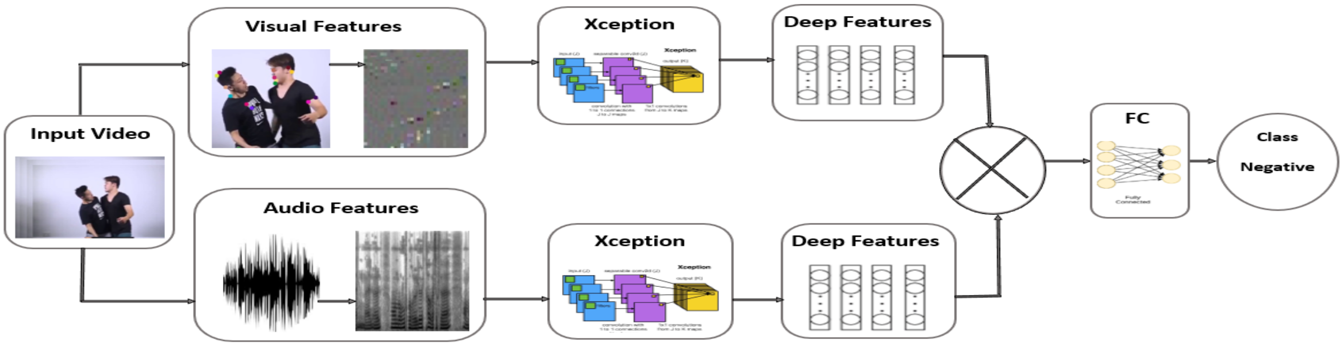


Figure 1: Group emotion recognition architecture. Given an input video, optical flow and mel spectrogram features are converted into image representations. Each modality (audio/video) is then used to train a separate Xception network, where deep features are extracted and fused for final recognition.

Table 1: Group emotion recognition results.

Model	Validation Accuracy	Test Accuracy
Baseline	**50.05%**	**47.88%**
Fusion	37.00%	35.00%
Xception Visual	24.00%	-
Xception Audio	35.00%	-

vectors for the median and standard deviation across the 16 Δ values. This results in three feature vectors (mean, median, and standard deviation) of size 1024, for a total of 3072 features to represent each video. It is important to note that we repeat the o_{mean} value in our feature vectors, as we will transform our vectors into the image space. This allows us to keep our size consistent which results in a final image of size 32 × 32 × 3. Given 3 feature vectors of size 1024, we then map the i^{th} value in each vector to the i^{th} pixel in our new image. We construct an RGB image, therefore we map the mean value to red, median value to green, and standard deviation value to blue. The final image is then resized to 256 × 256 to be consistent with the image of mel spectrogram audio features, as we detail next. This approaches allows for a new temporal representation of the flow in videos of group emotion (Figure 1).

Features From Audio: Given a video, we first extract an MP3 of the audio which we then convert to waveform. Given the waveform audio, we then compute the mel spectrogram [31], using 128 bins for 384 timestamps. This results in a 49,152-dimension feature vector $spec = [ms_1, \ldots, ms_{49152}]$, where ms_i is the i^{th} mel spectrogram feature computed. Similar to the video features, we map the i^{th} mel spectrogram feature to the i^{th} pixel in a new image of size 128 × 384, which we then resize to 256 × 256 (Figure 1).

3.3 Experimental Design and Results

The images of visual (optical flow) and audio (mel spectrogam) features are used to train a separate Xception network [7] which is a convolutional neural network with 71 layers. From each of the networks (audio and video), we extract 2048 deep features from the fully connected layers and then fuse them by concatenating them which results in a new 4096-dimension vector of deep features $DF = [dfa_1, \ldots, dfa_{2048}, dfv_1, \ldots, dfv_{2048}]$, where dfa_i and dfv_i are the i^{th} deep audio and video features, respectively. We then add one dense and one fully connected layer to recognize the emotion from DF (Figure 1).

As can be seen in Table 1, the proposed fusion approach boosts the overall accuracy, on the validation set, for group emotion recognition by 2% and 13% over using audio and video, respectively, in a unimodal fashion. This can be explained, in part, as multimodal features have been shown to boost accuracy over unimodal features [40]. For our evaluation on the test set, we submitted our fusion approach resulting in an accuracy of 35%. While the fusion improved the overall accuracy, compared to a unimodal approach, it is important to note that the overall accuracy is relatively low as it is $\tilde{1}$2%–13% lower than the baseline on the validation and test sets. This can be explained in part by the difficulty in tracking features for some of the videos. This had two direct impacts (1) reduced training data; and (2) invalid results on validation/test data. For the training data, we removed 28% of the training videos due to lack of tracked features. For the validation/test data, when the features were not tracked, to ensure a classification occurred, we set the default value of the features to 0. While this can partially explain the relatively low accuracy, this can also explain the audio features having an 11% higher accuracy compared to the visual features due to a smaller amount of training data for video.

4 DRIVER GAZE PREDICTION

4.1 Dataset

The driver gaze zone in the Wild(DGW) dataset [16] has 338 different subjects. The training set consists of 29,928 frames; validation set consists of 10,295 frames; test set consists of 11,041 frames. The entire dataset is classified into nine different zones within the car, represented by back mirror, side mirror, radio, speedometer and windshield. The data has been recorded under various illumination conditions and during different times of the day. Figure 2 shows the distribution of the training and validation sets.

4.2 Driver Gaze Prediction

For predicting driver gaze, we propose an architecture that consists of n feature extraction functions and an ensemble of m modeling

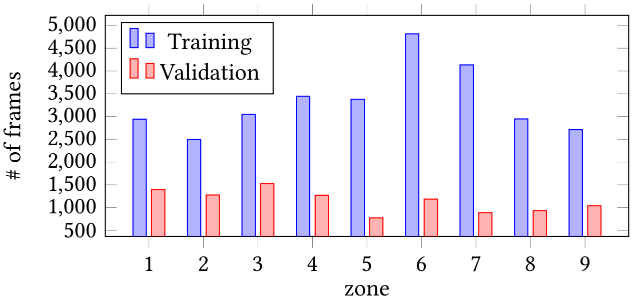


Figure 2: Per-zone data distribution.

functions. Each extraction function can output more than one feature for a given input frame. Thus, the extraction assembly results in p feature vectors such that $p \geq n$. In our experimental design $n = 2$ and $m = 6$, however, the proposed architecture can be extended to any number of feature extraction and modeling functions.

Feature Extraction. The feature extraction phase in the proposed architecture extracts patterns to be analyzed by a mapping function to the target label. In our experimental design, we have two categories of features - non-pixel, and pixel space. These features are extracted by the publicly available Dlib [22] and Openface [3] libraries. For pixel space features, we extracted and cropped the face and eye regions; we extract the following non-pixel space feature vectors, from each frame.

- **HOG Features[8]:** We extracted HOG features over the eye regions and constructed the 500-dimension feature vector $E_{HOG} = [H_1, H_2, \ldots, H_{500}]$, where H_i is the i^{th} HOG feature.
- **Eye Gaze:** We extracted 2 3D eye gaze vectors (left and right eyes) along with the average gaze angle over both x and y. We then constructed the 8-dimension feature vector $EG = [l_x, l_y, l_z, r_x, r_y, r_z, avg_x, avg_y]$, where $l_{\{x,y,z\}}$ and $r_{\{x,y,z\}}$ are the left and right gaze vectors, and avg_x and avg_y are the average gaze angles over x and y, respectively.
- **Head Pose:** We extracted the translation and orientation of the head, both as 3D vectors. We then constructed the 6-dimension feature vector $HP = [t_x, t_y, t_z, o_x, o_y, o_z]$, where $t_{\{x,y,z\}}$ and $o_{\{x,y,z\}}$ are the (x, y, z) coordinates of the translation and orientation vectors, respectively.
- **Facial Landmarks:** We extracted 68 2D facial landmarks resulting in the 136-dimension feature vector $FL = [x_1, y_1, x_2, y_2, \ldots, x_{68}, y_{68}]$, where x_i and y_i are the i^{th} extracted x and y coordinate, respectively.

Ensemble of Modeling Functions. An ensemble of models is trained with the features as input and target labels as output. In our experiments, non-pixel-space features (e.g. HOG, gaze vector) are each modeled by random forests [5] and mapped to the classification label. The pixel-space features (cropped face and eye regions) are each processed by InceptionV3 networks [34]. Each model in the ensemble receives a weight depending on their accuracy during the training phase, making the accuracy values a parameter. The final output maximizes the weighted sum of probability predictions.

In order to minimize the influence of noise on the model, we propose to use a weight vector, $w = [w_1 \ldots w_m]^\top$, where w is a collective measure of the accuracy of each model in the ensemble. Estimates of models that are more accurate, notwithstanding statistical imbalance in the features, are given more weight than others, allowing the ensemble to pick the best model decision. Thus, given a set of feature vectors, Φ, the ensemble of models may be viewed as a singular modeling function: $\hat{y} = f(\Phi; w)$. The model, in addition to being a function of the input variable Φ, captures a special aspect: the nature of the domain w. The output vector $\hat{y}$ is a set of confidence levels expressed as probability values by each model in the ensemble. In order to maximize the classification result by confidence level, we add confidence levels of each target variable corresponding to each model's classification and select the variable with the most confidence:

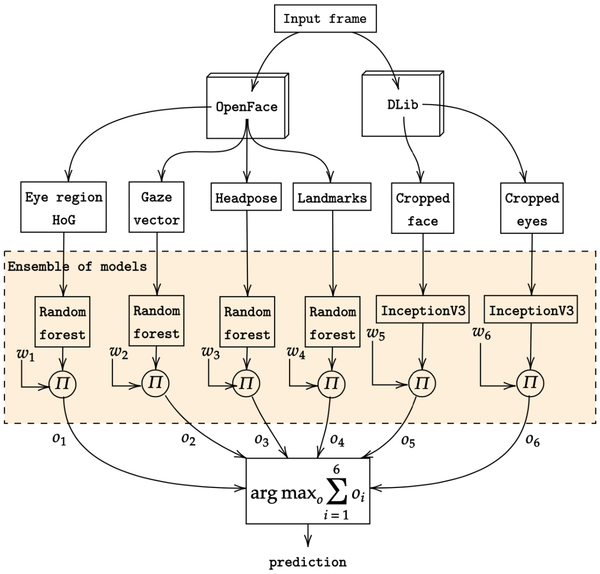


Figure 3: Overview of driver gaze prediction architecture - ensemble of random forests and InceptionV3 networks via voted prediction.

$$\hat{y} = \arg\max \sum_{i=1}^{m} \hat{y}_c. \quad (1)$$

Here, $c \in [1, 2, \ldots, 9]$, which are the driver zones (e.g. windshield). See Figure 3 for an overview of the proposed architecture.

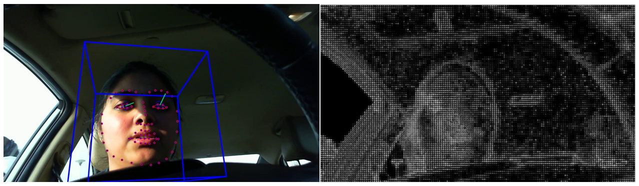
4.3 Experimental Design and Results

Experimental Setup. For our experimental design, we used a combination of random forests and InceptionV3 architecture to combine the effect of statistical and deep learning models. We empirically found the optimal number of trees to be $[250, 300]$ based on the specific feature vector (e.g. gaze vectors). For the InceptionV3 architecture, we use Adam optimizer [23] with a batch size of 32, learning rate of 0.001, 30% dropout [33], and the models are trained for 200 epochs.

Validation results. We investigated multiple combinations of features to train our models. Our experiments indicate that models trained on features obtained from several extraction functions are optimized better than those that rely on homogeneous and restricted sets of features. For example, Figure 4 shows an instance where the extractor inaccurately estimates eye gaze with uneven

Table 2: Driver zone classification results on the validation and test sets.

Feature type	Modality/Model	Feature size	Val accuracy	Test Accuracy	Method
Openface	Gaze	6	55.64	27.78	Random forest
	Headpose	8	90.04	-	Random forest
	Landmarks	136	91.34	-	Random forest
	HOG	500	75.30	26.37	Random forest
	Gaze + Headpose + Landmark	(6+8+136)	92.36	-	Random forest
	Gaze_RF + Headpose_RF + Landmark_RF	(6, 8, 136)	94.43	-	Voted probability predictions
	Headpose_RF + Landmark_RF	(8, 136)	95.56	-	Voted probability predictions
	Gaze_RF + HOG_RF	(6, 500)	85.38	-	voted probability predictions
	Headpose_RF + Landmark_RF + HOG_RF	(8, 136, 500)	96.58	-	Voted probability predictions
Face/Eye bounding box	Eyes	(224, 224)	57.9	-	InceptionV3
	Face	**(224, 224)**	66.45	**57.32**	**InceptionV3**
	Eyes + Face	([224 × 224], [224 × 224])	67.46	-	Voted probability prediction
Openface + Bounding box	Face + Gaze_RF + Landmark_RF + Headpose_RF	([224×224], 6, 8, 136)	96.50	-	Voted probability predictions
	Eyes + Landmark_RF + Headpose_RF	([224 × 224], 136, 8)	96.12	-	Voted probability predictions
	Eyes + Face + Gaze_RF	([224 × 224], [224 × 224], 8)	77.03	-	Voted probability predictions
	Face + Landmarks_RF + Headpose_RF + HOG_RF	([224 × 224], 136, 8, 500)	97.60	33.98	Voted probability predictions
	Face + Eyes + Gaze_RF + Landmark_RF + Headpose_RF	**([224 × 224], [224 × 224], 6, 136, 8)**	**97.8**	-	**Voted probability predictions**
Baseline	Face	-	60	-	InceptionV1



(a) Gaze direction misalignment **(b) HOG features**

Figure 4: Comparison of incorrect gaze vectors, due to lighting, along with lighting-invariant HOG features.

light illumination on the subject's face. To overcome this and similar challenges, we investigate a multimodal approach to the problem. In this example, as there are lighting issues, we investigate the use of illumination-invariant HOG features. Our experiments suggest this approach results in a positive impact on accuracy, as a random forest trained on the eye gaze vectors alone achieved an accuracy of 55.65%, while a random forest trained on HOG achieved an accuracy of 75.3% on the validation set. Certain features are significantly more accurate than others in cases where the extractor has learnt the sub-problem well. Conversely, features of poorly-learned sub-problems result in inferior inferences. Therefore, with a combination of several modalities, our model can rely on features that are generated with the most confidence and still output accurate results even if other features do not contribute as much to learning a given sub-problem. Considering this, a wide range of features obtained from multiple models (i.e. multimodal), can result in a rich set of features for training [40]. For example, when gaze and HOG features were combined in a multimodal fashion, the accuracy increased to 85.38%. See Table 2, for details on all conducted unimodal and multimodal experiments.

Test Results. We performed 4 evaluations; 2 unimodal non-pixel space, 1 unimodal pixel space, and 1 multimodal pixel and non-pixel space. More specifically, (1) gaze features; (2) HOG features; (3) cropped face regions; and (4) cropped face regions, facial landmarks, head pose, and HOG features. In these evaluations, we see the trend of the voted predictions performing better than the individual models, however, it is important to note our predictions involving random forests performed poorly on the test set (Table 2). This can be explained, at least partially, from high variation in the data distribution of the test set compared to the train and validation sets, as well as miscalculated features (i.e. error propagation from feature extractors to modeling functions). Considering this, the cropped face regions trained under the InceptionV3 architecture achieved the highest accuracy of 57.32%.

5 ENGAGEMENT PREDICTION THE WILD

5.1 Dataset

The engagement prediction in the wild dataset consists of 148 training and 48 validation videos of students watching educational videos. The videos were recorded at different locations such as computer labs, parks, and dormitories providing varied locations of "in the wild" settings. Each video is approximately 5 minutes long and is provided with an engagement level of $[0, 1]$, where 0 corresponds to disengaged and 1 corresponds to highly engaged. The levels were annotated by five annotators on the basis of video content only (without audio). Along with the videos, the dataset also contains the OpenFace[4] and LBP-TOP[42] features which include the subject's high- and low-level features. As can be seen in Figure 6, there is an imbalance of sample videos across each level. Considering this, we construct a subject independent, modified dataset which we refer to as merged + balanced, to account for the imbalance. In this modified dataset, we partially merged the training and validation videos to to be more balanced across each levels. Due to more videos for level 2 compared to levels, 0, 1, and 3 we merge and balance them by the following criteria: (1) merge all training and validation samples for levels 0, 1, and 3 resulting in 9, 45, and 53 sample videos for each level, respectively; (2) only use training samples for level 2 and reduce overall number of samples

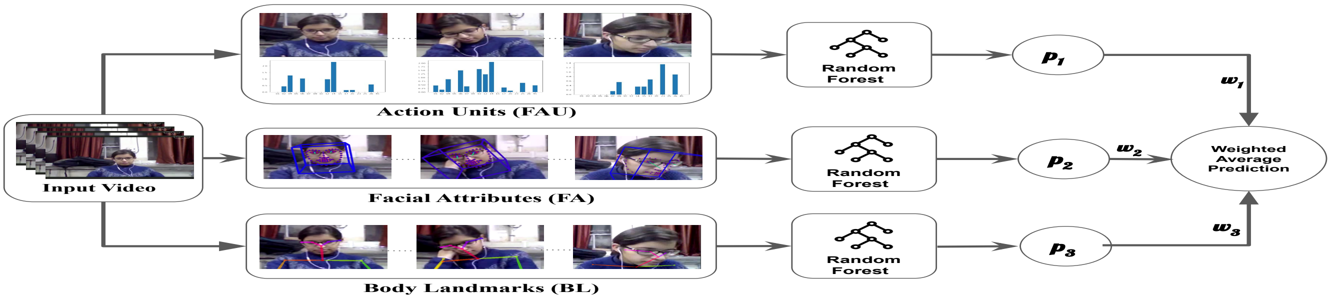


Figure 5: Ensemble-based engagement prediction architecture. AUs, gaze, head pose, facial landmarks, and body movement are extracted from input videos. Each modality is used to train an individual random forest, where the extracted predictions where then used with a weighted average to make the final prediction of engagement.

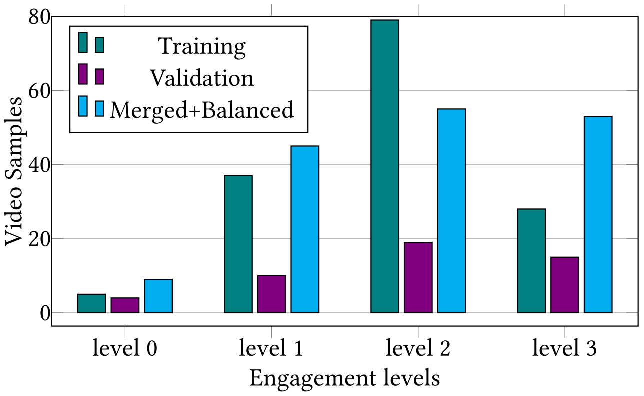


Figure 6: Distribution of Engagement Dataset. Level 0 = not engaged; level 1 = moderately engaged; level 2 = sufficiently engaged, and level 4 = high engagement.

from 79 to 55 (reduction of approximately 30%). It is important to note that this merged + balanced dataset was only used on the test set, validation results are from using the original training set.

We have also empirically found that many of the consecutive frames in the videos exhibit little change. Considering this, we down-sampled the videos to 10 frames per second. There are segments in each video where the students adjust the camera at the beginning and end of the videos which we filter by only considering frames within the timestamp range of $[00:30-4:30]$.

5.2 Engagement Prediction Features

We are motivated by the previous work that has shown the positive impact of facial attributes on predicting user engagement [1, 21]. Considering this, we propose to use AUs, gaze, head pose, and a non-rigid point distribution model (PDM) [4]. Along with these features, we also propose to use body movement as it can give us insight into the relation of body movement to engagement. To do this, we extract the body key points using the publicly available OpenPose[6], which extracts key body landmarks corresponding to face, hands and neck. Along with the default model for OpenPose, we also make use of the Common Object in Context (COCO) model [25] for extracting these features as we observed that using this gave us a better estimate in cases of occlusion. See Figure 5 for an overview of our proposed approach.

Facial Action Units [11] refer to facial muscle movements such as lowering of brow and pulling lip corner. We are motivated to use AUs by recent works that have shown them to be useful for inferring facial expressions [1, 19]. To investigate their positive impact for predicting engagement, we computed the mean and standard deviation for all 17 AUs, that are extracted with OpenFace, over a video sequence. We then construct the 34-dimension feature vector $FAU = [mean_1,\ mean_2,\ \ldots,\ mean_{17},\ std_1,\ std_2,\ \ldots,\ std_{17}]$, where $mean_i$ and std_i are the mean and standard deviation of the i^{th} action unit across all video frames.

Facial Attributes have been successfully used in previous works for predicting engagement [1, 21]. Considering this, we propose to use the following per-frame features in our approach.

- **Eye Gaze:** 3D eye gaze vector, from both eyes, resulting in a 6-dimension feature vector $f_g = [l_x,\ l_y,\ l_z,\ r_x,\ r_y,\ r_z]$, where $l_{\{x,\ y,\ z\}}$ and $r_{\{x,\ y,\ z\}}$ are the (x, y, z) coordinates of the left and right eye gaze direction vector, respectively.
- **Head Pose:** We constructed a 6-dimension feature vector $f_h = [t_x,\ t_y,\ t_z,\ o_x,\ o_y,\ o_z]$, where $t_{\{x,\ y,\ z\}}$ and $o_{\{x,\ y,\ z\}}$ are the (x, y, z) coordinates of the head pose translation and orientation vectors, respectively.
- **Point Distribution Model:** We used 34 non-rigid point distribution model (PDM) parameters to make a 34-dimension feature vector $f_{pdm} = [p_1, p_2, \ldots, p_{34}]$, where p_i is the i^{th} non-rigid PDM parameter.

Given feature vectors f_g, f_h, and f_{pdm}, we then construct the 46-dimension feature vector $F = [f_g,\ f_h,\ f_{pdm}]$ resulting in a new facial attributes feature vector that consists of gaze, head pose, and non-rigid PDM parameters. We then calculate the mean and standard deviation of each element of F across all frames in each video. We then constructed our final 92-dimension facial attribute feature vector $FA = [mean_1,\ mean_2,\ \ldots,\ mean_{46},\ std_1,\ std_2,\ \ldots,\ std_{46}]$, where $mean_i$ and std_i are the mean and standard deviation, across all frames of a video, for each element in F.

Body Landmarks. We detected 12 key-points around the face, neck and hand to construct a 12-dimension feature vector $B = [nose,\ neck,\ shoulder_{left},\ shoulder_{right},\ elbow_{left},\ elbow_{right},\ wrist_{left},\ wrist_{right},\ eye_{left},\ eye_{right},\ ear_{left},\ ear_{right}]$, where

Table 3: Engagement prediction results - original train data.

Modality	MSE$_{val}$	PCC$_{val}$	MSE$_{level-wise}$			
			Level 0	Level 1	Level 2	Level 1
AUs (FAU)	**0.054**	**0.673**	**0.090**	**0.100**	**0.017**	**0.060**
Facial Attributes (FA)	0.072	0.497	0.149	0.093	0.013	0.110
Body Landmarks (BL)	0.068	0.554	0.201	0.094	0.010	0.088
Average Ensemble	0.062	0.622	0.141	0.094	0.012	0.081
Weighted Avg Ensemble	0.061	0.631	0.136	0.095	0.012	0.079
Baseline	0.10	-	-	-	-	-

each element of B is a key-point from the specific part of the body (e.g. $elbow_{right}$ is a key-point extracted from the right elbow). We then calculated the mean and standard deviation for each key-point, in B across all frames of a video. This results in our final 24-dimension body landmark feature vector $BL = [mean_1,\ mean_2,\ \ldots,\ mean_{12},\ std_1,\ std_2,\ \ldots,\ std_{12}]$, where $mean_i$ and std_i are the mean and standard deviation, across all frames of a video, for the i^{th} element in B.

5.3 Experimental Design and Results

To predict engagement in the wild, we trained 3 random forests [5] with 200 trees each. Each random forest was trained with one of the feature types described in Section 5.2: (1) Action units (FAU); (2) Facial Attributes (FA); and (3) Body Landmarks (BL). We evaluated the accuracy of prediction using each individual model, as well as an average ensemble, and a weighted average ensemble. For our evaluation, we calculated the Mean Squared Error (MSE) along with Pearson's Correlation Coefficient (PCC) for understanding how related our predictions are to the true values. For a better understanding of level-wise performance, engagement level-wise MSEs are also calculated. For the average ensemble, we take the average prediction across the FAU, FA, and BL models as the final prediction. For our weighted ensemble, we find the weights based on the performance of each model on the validation set as

$$w_{total} = \sum_{i=1}^{N} \frac{1}{MSE_{model_i}}, \tag{2}$$

where w_{total} is the total weight across all models, N = total number of models, and MSE_{model_i} is the MSE of the evaluated model. We then calculate the weight for each individual model as

$$w_{model_i} = \frac{\frac{1}{MSE_{model_i}}}{w_{total}}. \tag{3}$$

Given w_{model_i}, we then calculate the final prediction as

$$pred_{final} = \sum_{i=1}^{N} pred_{model_i} \times w_{model_i}. \tag{4}$$

In our experimental design $N = 3$, and $model_i \in [FAU, FA, BL]$.

As can be seen in Table 3, the model trained on AUs (FAU) had the lowest MSE with 0.054, on the validation set. However, as can be seen in Table 4, this is not the same for the testing set when we used the original training set as training data. Both the average and weighted ensemble achieved a lower MSE. This can be explained, at least partially, by the ensemble being able to better handle unseen data in the testing set and the distribution of AUs could have been different in the validation and testing sets causing a decrease in performance [20]. Considering this, we also submitted evaluations on the test set using the average and weighted ensemble trained on the merged+balanced training data, as detailed in Section 5.1. Using the merged+balanced training data, we achieved our lowest MSE of 0.0659 (3^{rd} place in challenge) with our weighted average ensemble approach (Table 4). This overall lower MSE suggests that our weighted average ensemble approach, along with a more balanced training set can help increase the performance of predicting engagement levels in wild settings.

Table 4: Engagement prediction results - merged train data.

Training Data	Modality	MSE$_{test}$	MSE$_{level-wise}$			
			Level 0	Level 1	Level 2	Level 3
Original	Average Ensemble	0.0666	0.281	0.064	0.013	0.073
	Weighted Avg Ensemble	0.0662	0.280	0.063	0.013	0.069
	Action Units	0.0675	0.268	0.060	0.023	0.035
Merged+ Balanced	Average Ensemble	0.0691	0.274	0.063	0.020	0.079
	Weighted Avg Ensemble	**0.0659**	**0.266**	**0.061**	**0.018**	**0.067**
	Baseline	0.15	-	-	-	-

6 PHYSIOLOGICAL SIGNAL BASED EMOTION RECOGNITION

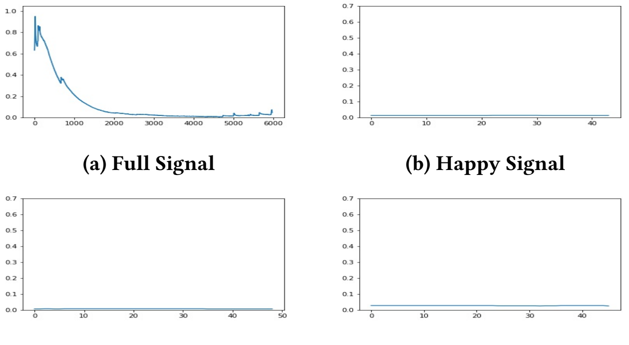
6.1 Dataset

The dataset consists of physiological data collected while subjects watch movie clips from the Acted Facial Expressions in The Wild (AFEW) [10]. The videos in the AFEW include the 7 emotions Happy, Sad, Disgust, Surprise, Fear, Angry and Neutral. Each of the videos are approximately 300ms - 5400ms long. The physiological data is collected at a frequency of 4Hz, resulting in each signal having a range of data points from $[12, 216]$ to represent the video. It is important to note that the subjects watch each of the videos in a consecutive fashion. This can result in significant variance over time as can be seen in Figure 7a, however when the shorter, emotion-specific segments of the sequences are extracted, the variance decreases and multiple emotions look similar (Figure 7b - 7d).

6.2 GAN-based Ensemble

Generative Adversarial Networks [17] have a Generator and Discriminator network and the two models attempt to outperform the other. In many works, GANs are used to generate new synthetic data with the help of the generator [14, 15, 36]. In our approach, we propose to use the discriminator part of the GAN to recognize emotions. We trained 7 GAN architectures, where each one was trained on a specific emotion (e.g. Happy). To facilitate this we divided the training into two stages: (1) Independent GAN training, and (2) Discriminator ensemble.

GAN Architecture. Two Fully Connected Neural Networks (FNN), where one is the generator and one is the discriminator are used. The generator has 3 dense layers with the first layer having 5000 neurons, second layer has 2500 neurons, and the output layer has 1500 output neurons. The first 2 layers use reLu activation and the output layer uses linear activation. The discriminator has 7 layers - 5 dense layers and 2 dropout layers. The first layer is the



(a) Full Signal

(b) Happy Signal

(c) Surprise Signal

(d) Disgust Signal

Figure 7: Psychological signals of one subject in training set.

input layer with 1500 neurons, the second layer has 1000 neurons, third layer is a dropout layer with a 0.2 dropout, fourth layer has 750 neurons, fifth layer is another dropout layer with 0.2 dropout, the sixth layer is a dense layer with 150 neurons and the output layer has just a single neuron. The output layer has sigmoid activation while the other dense layers have reLu activation. We used binary cross entropy as the loss function and adam as the optimizer. We trained each GAN network for 500 epochs.

Independent GAN training. GANs train the generator to confuse the discriminator, and the discriminator tries to identify fake signals from real signals. Motivated by this, we modify the training of the GAN to train the discriminator to identify the emotion signal from noise and other emotion signals (i.e. real emotion vs. fake/other emotion). In original GAN training, the discriminator is trained with the original signal labeled as real and the generated signal labeled as fake. In our training method, we trained the discriminators by labeling the target emotion of the discriminator as real and other emotions along with signals generated by the generators as fake. To facilitate this training, it is important to note that the physiological signals are of different lengths. To create a uniform signal across participants and emotions, we first performed min-max normalization to the entire signal and then re-sampled the signal to 1500 points to ensure consistent signal lengths across all participants and emotions.

Ensemble of GAN Discriminators. After we have trained the GANs, we take the 7 discriminators and concatenate the output layers of the discriminators resulting in the 7-dimension probability vector $G_{prob} = [D_1, D_2, \ldots, D_7]$, where D_i is the probability output from the i^{th} GAN discriminator. Given G_{prob}, we then train a random forest to recognize emotions.

6.3 Experimental Design and Results

Motivated by the work from Liu et al. [26], we also created an 8-dimension vector of handcrafted features $G_{hc} = [min, max, mean, variance, mean\ abs\ diff, second\ mean\ abs\ diff, age, gender]$, which is used to train a random forest to recognize emotion. For our experimental design, we have empirically found that the optimal number of trees for our random forests is 275.

GAN-based Ensemble Results. In the validation setting, we recognized the emotion of 1 subject watching a video. Under this setting, we achieved an overall accuracy of 14.58%. The test setting was modified compared to the validation setting. In this setting, we were required to give one label to the video which was watched by multiple participants. To facilitate this, we used a max voting approach and we broke ties by summing the discriminator probabilities of all participants, who watched the video, and the emotion with the highest summed probability was predicted as the final emotion for the video. Using this experimental design, We achieved a overall accuracy of 15.18% on the challenge testing set.

Table 5: Emotion-level results (physiological) on test set.

Emotion	GAN	Random Forest
Happy	22.22%	22.22%
Surprise	17.86%	3.57%
Disgust	2.50%	12.50%
Angry	10.31%	11.34%
Fear	10.00%	14.29%
Sad	12.50%	12.50%
Neutral	17.62%	29.02%

Hand-crafted Features Results. For training a random forest with hand-crafted features, We conducted a similar experimental design as that of our GAN-based ensemble. While we still implemented max-voting, in this setting we broke ties by randomly selecting the emotion with the max votes. Using this experimental design, we achieved an overall accuracy of 15.44% and 19.17% on the validation and test sets respectively.

It is interesting to note, that the hand-crafted features outperformed the GAN-based ensemble, for overall accuracy, on both the validation and testing sets. This can be explained, in part, by the small variance in signal across emotions. It has been shown that physiological signals, that exhibit high variance, generally have higher performance for emotion recognition [13], while random forests can perform well when the relative number of dimensions is low for the data [12]. As can be seen in Table 5, while the hand-crafted features outperformed our GAN-based ensemble for overall accuracy, this is not true for all emotions, in the test set. For example, the Surprise emotion achieved an average accuracy of 17.86% and 3.57% on our proposed GAN-based ensemble and the hand-crafted features respectively. This could potentially be explained by the changes in the signals that were not captured by the hand-crafted features, however, our proposed ensemble was able to more accurately determine what was real and what was fake (i.e. not Surprise). It is important to note that we are not showing the baseline results for this track due to the baseline paper [26] using different data, and we are the *only* group to submit to this track.

7 CONCLUSION

We detailed our proposed approaches to all 4 tracks of the EmotiW 2020 challenge. We showed how optical flow and mel sprectrogram features can be fused for group emotion recognition. Along with this, we also showed that facial features such as gaze and head pose, can be used for multiple tracks as they generalize well to different problems such as driver gaze estimation and engagement in the wild. Using a challenging physiological signal dataset, we proposed the use of a GAN-based ensemble for recognizing emotion. Through most of the tracks, we achieve results that are comparable to, or outperform the baseline on the validation and test sets.

REFERENCES

[1] Abdulkareem Al-Alwani. 2016. Mood Extraction Using Facial Features to Improve Learning Curves of Students in ELearning Systems. *International Journal of Advanced Computer Science and Applications, vol. 7, no. 11, pp. 444- 453* (2016).

[2] Peter A Anderson and Laura K Guerrero. 1998. The handbook of communication and emotion.

[3] Tadas Baltrusaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. 2018. Openface 2.0: Facial behavior analysis toolkit. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*. IEEE, 59–66.

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

[5] Leo Breiman. 2001. Random forests. *Machine learning* 45, 1 (2001), 5–32.

[6] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh. 2019. OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2019).

[7] François Chollet. 2017. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1251–1258.

[8] Navneet Dalal and Bill Triggs. 2005. Histograms of oriented gradients for human detection. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, Vol. 1. IEEE, 886–893.

[9] Tao Deng, Hongmei Yan, Long Qin, Thuyen Ngo, and BS Manjunath. 2019. How do drivers allocate their potential attention? Driving fixation prediction via convolutional neural networks. *IEEE Transactions on Intelligent Transportation Systems* 21, 5 (2019), 2146–2154.

[10] Abhinav Dhall, Roland Goecke, Simon Lucey, and Tom Gedeon. 2011. Acted facial expressions in the wild database. (2011).

[11] P. Ekman. 1997. What the face reveals: Basic and app studies of spon exp using the Facial Action Coding System (FACS). *Ox Uni Press* (1997).

[12] D. Fabiano et al. 2018. Spontaneous and non-spontaneous 3D facial expression recognition using a statistical model with global and local constraints. *ICIP* (2018).

[13] Diego Fabiano and Shaun Canavan. 2019. Emotion Recognition Using Fused Physiological Signals. In *ACII)*.

[14] Maayan Frid-Adar, Idit Diamant, Eyal Klang, Michal Amitai, Jacob Goldberger, and Hayit Greenspan. 2018. GAN-based synthetic medical image augmentation for increased CNN performance in liver lesion classification. *Neurocomputing* 321 (2018), 321–331.

[15] Maayan Frid-Adar, Eyal Klang, Michal Amitai, Jacob Goldberger, and Hayit Greenspan. 2018. Synthetic data augmentation using GAN for improved liver lesion classification. In *2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018)*. IEEE, 289–293.

[16] 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).

[17] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Advances in neural information processing systems*. 2672–2680.

[18] Aarush Gupta, Dakshit Agrawal, Hardik Chauhan, Jose Dolz, and Marco Pedersoli. 2018. An attention model for group-level emotion recognition. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction*. 611–615.

[19] S Hinduja, S Canavan, and G Kaur. 2020. Multimodal Fusion of Physiological Signals and Facial Action Units for Pain Recognition. In *15th IEEE International Conference on Automatic Face and Gesture Recognition*. 387–391.

[20] L. Jeni et al. 2013. Facing imbalanced data–recommendations for the use of performance metrics. In *ACII*.

[21] 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.

[22] Davis E King. 2009. Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research* 10, Jul (2009), 1755–1758.

[23] Diederik P Kingma et al. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[24] Sander Koelstra, Christian Muhl, Mohammad Soleymani, Jong-Seok Lee, Ashkan Yazdani, Touradj Ebrahimi, Thierry Pun, Anton Nijholt, and Ioannis Patras. 2011. Deap: A database for emotion analysis; using physiological signals. *IEEE transactions on affective computing* 3, 1 (2011), 18–31.

[25] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*. Springer, 740–755.

[26] 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).

[27] Bruce D Lucas, Takeo Kanade, et al. 1981. An iterative image registration technique with an application to stereo vision. (1981).

[28] Cristina Olaverri-Monreal, Ahmed Elsherbiny Hasan, Jonathan Bulut, Moritz Körber, and Klaus Bengler. 2014. Impact of in-vehicle displays location preferences on drivers' performance and gaze. *IEEE Transactions on Intelligent Transportation Systems* 15, 4 (2014), 1770–1780.

[29] Sally Planalp. 1999. *Communicating emotion: Social, moral, and cultural processes*. Cambridge University Press.

[30] Garima Sharma, Shreya Ghosh, and Abhinav Dhall. 2019. Automatic Group Level Affect and Cohesion Prediction in Videos. In *International Conference on Affective Computing and Intelligent Interaction Workshops and Demo*. 161–167.

[31] Jonathan Shen, Ruoming Pang, Ron J Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, Rj Skerrv-Ryan, et al. 2018. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 4779–4783.

[32] Jianbo Shi et al. 1994. Good features to track. In *1994 Proceedings of IEEE conference on computer vision and pattern recognition*. IEEE, 593–600.

[33] Ramprakash Srinivasan and Aleix M Martinez. 2018. Cross-Cultural and Cultural-Specific Production and Perception of Facial Expressions of Emotion in the Wild. *IEEE Transactions on Affective Computing* (2018).

[34] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2818–2826.

[35] Chinchu Thomas and Dinesh Babu Jayagopi. 2017. Predicting student engagement in classrooms using facial behavioral cues. In *Proceedings of the 1st ACM SIGCHI international workshop on multimodal interaction for education*. 33–40.

[36] Luan Tran, Xi Yin, and Xiaoming Liu. 2017. Disentangled representation learning gan for pose-invariant face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1415–1424.

[37] Gerben A Van Kleef and Agneta H Fischer. 2016. Emotional collectives: How groups shape emotions and emotions shape groups. *Cognition and Emotion* 30, 1 (2016), 3–19.

[38] Anne Veenendaal, Elliot Daly, Eddie Jones, Zhao Gang, Sumalini Vartak, and Rahul S Patwardhan. 2014. Group emotion detection using edge detecttion mesh analysis. *Computer Science and Emerging Research Journal* 2 (2014).

[39] Yafei Wang, Tongtong Zhao, Xueyan Ding, Jiming Bian, and Xianping Fu. 2017. Head pose-free eye gaze prediction for driver attention study. In *2017 IEEE International Conference on Big Data and Smart Computing (BigComp)*. IEEE, 42–46.

[40] Ghada Zamzmi et al. 2016. Machine-based multimodal pain assessment tool for infants: a review. *arXiv preprint arXiv:1607.00331* (2016).

[41] Ghada Zamzmi, Pai Chih-Yun, Dmitry Goldgof, R Kasturi, Terri Ashmeade, and Yu Sun. 2019. A comprehensive and context-sensitive neonatal pain assessment using computer vision. *IEEE Transactions on Affective Computing* (2019).

[42] Guoying Zhao and Matti Pietikainen. 2007. Dynamic Texture Recognition Using Local Binary Patterns with an Application to Facial Expressions. *IEEE Transaction on Pattern Analysis and Machine Intelligence 29, 6 (2007), 915–928* (2007).

[43] Z. Zheng et al. 2016. Multimodal spontaneous emotion corpus for human behavior analysis. In *CVPR*.