March 2021

Recognizing Emotion in the Wild Using Multimodal Data

Shivam Srivastava
University of South Florida

Follow this and additional works at: <https://digitalcommons.usf.edu/etd>

Part of the [Computer Sciences Commons](#)

Scholar Commons Citation

Srivastava, Shivam, "Recognizing Emotion in the Wild Using Multimodal Data" (2021). *USF Tampa Graduate Theses and Dissertations*.
<https://digitalcommons.usf.edu/etd/8873>

This Thesis is brought to you for free and open access by the USF Graduate Theses and Dissertations at Digital Commons @ University of South Florida. It has been accepted for inclusion in USF Tampa Graduate Theses and Dissertations by an authorized administrator of Digital Commons @ University of South Florida. For more information, please contact digitalcommons@usf.edu.

Recognizing Emotion in the Wild Using Multimodal Data

by

Shivam Srivastava

A thesis submitted in partial fulfillment
of the requirements for the degree of
Master of Science in Computer Science
Department of Computer Science and Engineering
College of Engineering
University of South Florida

Major Professor: Shaun Canavan, Ph.D.
Dmitry B. Goldgof, Ph.D.
Srinivas Katkoori, Ph.D.

Date of Approval:
March 6, 2021

Keywords: Gaze Detection, Engagement Prediction, Physiological Signals

Copyright © 2021, Shivam Srivastava

Dedication

I would like to dedicate this thesis to my parents for their endless love, support, guidance and encouragement. I would also like to dedicate this thesis to my friend and mentor, Saurabh Hinduja, for believing in me and providing his continuous support and help.

Acknowledgments

I would first like to thank my thesis advisor, Dr. Shaun Canavan, for the continuous support of my research, for his patience, motivation, enthusiasm, and immense knowledge. Prof. Canavan was always available whenever I ran into a trouble spot or had a question about my research or writing. He consistently allowed this paper to be my work but steered me in the right direction whenever he thought I needed it.

My completion of this research could not have been accomplished without the support of my mentor, Saurabh Hinduja. I am grateful for his consistent encouragement, guidance, insightful comments and ideas.

Finally, I must express my very profound gratitude to my friends for providing me with unfailing support and continuous encouragement throughout my years of study and through the process of researching and writing this thesis. This accomplishment would not have been possible without them.

Table of Contents

List of Tables	iii
List of Figures	iv
Abstract	v
Chapter 1: Introduction	1
1.1 Motivation	1
1.2 Contributions	2
Chapter 2: Audio-Video Group Emotion Recognition	3
2.1 Dataset	3
2.2 Related Works	3
2.3 Group Emotion Recognition Features	4
2.3.1 Features from Video	4
2.3.2 Features from Audio	5
2.4 Experimental Design and Results	5
Chapter 3: Driver Gaze Prediction	8
3.1 Dataset	8
3.2 Related Works	8
3.3 Feature Extraction	9
3.3.1 Pixel Space Features	9
3.3.2 Non-Pixel Space Features	10
3.4 Experimental Design and Results	10
Chapter 4: Engagement Prediction the Wild	16
4.1 Dataset	16
4.2 Related Works	16
4.3 Engagement Prediction Features	18
4.3.1 Facial Action Units	19
4.3.2 Facial Attributes	19
4.3.3 Body Landmarks	20
4.4 Experimental Design and Results	20

Chapter 5: Physiological Signal Based Emotion Recognition	24
5.1 Dataset	24
5.2 Related Works	24
5.3 Feature Extraction	25
5.4 Experimental Design and Results	25
5.4.1 GAN Based Training	25
5.4.1.1 Independent GAN Training	27
5.4.1.2 Ensemble of GAN Discriminators	27
5.4.2 Hand-crafted Features Training	28
5.4.3 Results	28
5.4.3.1 GAN-based Ensemble Results	28
5.4.3.2 Hand-crafted Features Results	28
Chapter 6: Discussion and Future Work	30
References	32
Appendix A: Copyright Permissions	39
About the Author	End Page

List of Tables

Table 2.1	Group Emotion Recognition Results - Validation and Test	5
Table 3.1	Driver Zone Prediction Results - Validation and Test	15
Table 4.1	Student Engagement Prediction Results - Validation	21
Table 4.2	Student Engagement Prediction Results - Test	22
Table 5.1	Physiological Signal Based Emotion Recognition Results	29
Table 5.2	Physiological Signal Emotion Wise Accuracy	29

List of Figures

Figure 2.1	Group Emotion Recognition Architecture	7
Figure 3.1	Driver Gaze Prediction Data Distribution	9
Figure 3.2	Comparison of incorrect gaze vectors	11
Figure 3.3	Driver Gaze Prediction Architecture	13
Figure 4.1	Level-wise distribution of Student Engagement Prediction Dataset .	17
Figure 4.2	Ensemble-based Student Engagement Prediction Architecture	23
Figure 5.1	Physiological Signals of a subject in training set	26

Abstract

In this work, I will present our approach of using multi-modal data for recognizing human emotion and behavior in the wild. The study is divided into four tasks: group emotion recognition, driver gaze prediction, student engagement prediction, 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).

Chapter 1: Introduction

1.1 Motivation

¹ To study the impact of multiple modalities in recognising human emotion and behavior, we used four different tracks: (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). The datasets of these tracks were shared in the eighth Emotion Recognition in the Wild Grand Challenge (EmotiW 2020). Each of the tracks is an important and timely topic that has 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 than 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 [30]. Along with helping with safe driving, gaze can also be used to help assess engagement in students in classrooms (i.e. wild setting). Considering this, Thomas and Jayagopi [39] 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 other 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. [44] used the fusion of physiological signals along with other features such as crying and body movement, to assess neonatal pain. Their results suggest that important features can be extracted from these signals for an accurate assessment.

¹This chapter was published in ©2020 ACM International Conference on Multimodal Interaction [37]. Reprinted with permission included in Appendix A.

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, 31]. Van Kleef and Fischer [41] 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.

1.2 Contributions

Considering each of the important, broad applications that each of these 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 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.

Chapter 2: Audio-Video Group Emotion Recognition

2.1 Dataset

² The dataset [32] 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.

2.2 Related Works

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. [42]. They proposed the use of edge detection to extract features for group emotion recognition. Using these features to train a support vector machine, they

²This chapter was published in ©2020 ACM International Conference on Multimodal Interaction [37]. Reprinted with permission included in Appendix A.

showed promising results when the group’s emotion was captured while they watched sports videos.

2.3 Group Emotion Recognition Features

2.3.1 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 [29, 34]. This results in a 32 dimension feature vector $F_{good} = [x_1, y_1, \dots, x_{16}, y_{16}]$, 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 $M_{final} = [x_{mean_1}, y_{mean_1}, o_{mean_1}, o_{mean_1}, \dots, x_{mean_{256}}, y_{mean_{256}}, o_{mean_{256}}, o_{mean_{256}}]$. We then construct similar 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 $32 \times 32 \times 3$. 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 approach allows for a new temporal representation of the flow in videos of group emotion (Figure 2.1).

2.3.2 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 [33], using 128 bins for 384 timestamps. This results in a 49,152-dimension feature vector $spec = [ms_1, \dots, ms_{49152}]$, 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 .

2.4 Experimental Design and Results

The images of visual (optical flow) and audio (mel spectrogram) 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 $DF = [dfa_1, \dots, dfa_{2048}, dfv_1, \dots, dfv_{2048}]$, 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 2.1).

As can be seen in Table 2.1, 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

Table 2.1: Group Emotion Recognition Results - Validation and Test[37]

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

features have been shown to boost accuracy over unimodal features [45]. 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 $\tilde{12\%}$ –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.

Figure 2.1: 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.[37]

Chapter 3: Driver Gaze Prediction

3.1 Dataset

³ 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 3.1 shows the distribution of the training and validation sets.

3.2 Related Works

In recent years, there has been promising work predicting driver eye gaze. Wang et al. [43] 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).

³This chapter was published in ©2020 ACM International Conference on Multimodal Interaction [37]. Reprinted with permission included in Appendix A.

Figure 3.1: Driver Gaze Prediction per-zone data distribution.[37]

3.3 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.

3.3.1 Pixel Space Features

Given an image of a driver, we extracted and cropped the face and eye regions. We used Dlib[22] library to extract the face region. We then used the pre-trained facial landmark detector of the Dlib library to first estimate the location of 68 (x,y) coordinates that map to facial structures on the face and then use the landmarks corresponding to eye region to extract the ROI.

3.3.2 Non-Pixel Space Features

We extracted the following non-pixel space feature vectors, from each frame using the open source library Openface [3].

- *HOG Features*[8]: We extracted HOG features over the eye regions and constructed the 500-dimension feature vector $E_{HOG} = [H_1, H_2, \dots, H_{500}]$, 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 $FL = [x_1, y_1, x_2, y_2, \dots, x_{68}, y_{68}]$, where x_i and y_i are the i^{th} extracted x and y coordinate, respectively.

3.4 Experimental Design and Results

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 [38]. Each model in the ensemble receives a weight depending on their accuracy during the training phase, making

(a) Gaze direction misalignment

(b) HOG features

Figure 3.2: Comparison of incorrect gaze vectors, due to lighting, along with lighting-invariant HOG features.[37]

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, $\mathbf{w} = [w_1 \dots w_m]^\top$, where $\mathbf{w}$ 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: $\hat{\mathbf{y}} = f(\Phi; \mathbf{w})$. The model, in addition to being a function of the input variable Φ , captures a special aspect: the nature of the domain $\mathbf{w}$. The output vector $\hat{\mathbf{y}}$ 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:

$$\hat{\mathbf{y}} = \arg \max \sum_{i=1}^m \hat{\mathbf{y}}_c. \quad (3.1)$$

Here, $c \in [1, 2, \dots, 9]$, which are the driver zones (e.g. windshield). See Figure 3.3 for an overview of the proposed architecture.

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 [36], and the models are trained for 200 epochs.

On validation data, 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 3.2 shows an instance where the extractor inaccurately estimates eye gaze with uneven 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 [45]. For example, when gaze and HOG features were combined in a multimodal fashion, the accuracy increased to 85.38%. See Table 3.1, for details on all conducted unimodal and multimodal experiments.

For the test data, 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;

Figure 3.3: Driver Gaze Prediction Architecture - ensemble of random forests and InceptionV3 networks via voted prediction.[37]

(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 3.1). 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%.

Table 3.1: Driver Zone Prediction Results - Validation and Test[37]

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

Chapter 4: Engagement Prediction the Wild

4.1 Dataset

⁴ The engagement prediction in the wild dataset consists of 148 training and 48 validation videos of students watching educational videos. 78 subjects (25 female and 53 male) participated in the data collection. These 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 recorded at a resolution of 640 x 480 pixels at 30fps 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[46] features which include the subject's high- and low-level features.

4.2 Related Works

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 [46] 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.

⁴This chapter was published in ©2020 ACM International Conference on Multimodal Interaction [37]. Reprinted with permission included in Appendix A.

Figure 4.1: Distribution of Student Engagement Prediction Dataset. Level 0 = not engaged; level 1 = moderately engaged; level 2 = sufficiently engaged, and level 4 = high engagement.[37]

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.

As can be seen in Figure 4.1, 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 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].

4.3 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 [26] for extracting these features as we observed that using this gave us a better estimate in cases of occlusion. See Figure 4.2 for an overview of our proposed approach.

4.3.1 Facial Action Units

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 $FAU = [mean_1, mean_2, \dots, mean_{17}, std_1, std_2, \dots, std_{17}]$, where $mean_i$ and std_i are the mean and standard deviation of the i^{th} action unit across all video frames.

4.3.2 Facial Attributes

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

- *Eye Gaze*: 3D eye gaze vector, from both eyes, was used to estimate the point where one is looking at. We created 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*: As the name suggests, Head Pose Estimation refers to the pose or angle of a human head in an image. 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*: The Point Distribution Model (PDM) is a technique for describing features that have well understood general shape, but which cannot be easily described by a rigid model. It relies on 2D landmark points. We used 34 non-rigid PDM parameters to make a 34-dimension feature vector $f_{pdm} = [p_1, p_2, \dots, p_{34}]$, 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 $FA = [mean_1, mean_2, \dots, mean_{46}, std_1, std_2, \dots, std_{46}]$, where $mean_i$ and std_i are the mean and standard deviation, across all frames of a video, for each element in F .

4.3.3 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 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 $BL = [mean_1, mean_2, \dots, mean_{12}, std_1, std_2, \dots, std_{12}]$, 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 .

4.4 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 4.3: (1) Action units (FAU); (2) Facial Attributes (FA); and (3) Body Landmarks (BL). We evaluated

Table 4.1: Student engagement prediction Results - Validation. [37]

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	-	-	-	-	-

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

$$w_{total} = \sum_{i=1}^N \frac{1}{MSE_{model_i}}, \quad (4.1)$$

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

$$w_{model_i} = \frac{\frac{1}{MSE_{model_i}}}{w_{total}}. \quad (4.2)$$

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

$$pred_{final} = \sum_{i=1}^N pred_{model_i} \times w_{model_i}. \quad (4.3)$$

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

Table 4.2: Student engagement prediction test Results - Test [37]

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	-	-	-	-

As can be seen in Table 4.1, the model trained on AUs (FAU) had the lowest MSE with 0.054, on the validation set. However, as can be seen in Table 4.2, 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 4.1. Using the merged+balanced training data, we achieved our lowest MSE of 0.0659 (3rd place in challenge) with our weighted average ensemble approach (Table 4.2). 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.

Figure 4.2: Ensemble-based Student 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 were then used with a weighted average to make the final prediction of engagement.[37]

Chapter 5: Physiological Signal Based Emotion Recognition

5.1 Dataset

⁵ 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 5.1a, however when the shorter, emotion-specific segments of the sequences are extracted, the variance decreases and multiple emotions look similar (Figure 5.1b - 5.1h).

5.2 Related Works

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+ [47], 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

⁵This chapter was published in ©2020 ACM International Conference on Multimodal Interaction [37]. Reprinted with permission included in Appendix A.

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.

5.3 Feature Extraction

To reduce the differences of observed EDA signals between participant, we apply min-max normalization to the original physiological signal data. We then extracted 8 features from these signals corresponding to each video. These first four features are max, min, mean and variance of all the observed values of a sequence. Next, we have mean abs diff and second mean abs diff which refers to the mean of the absolute value of First-Order and Second-order differences respectively. Finally, we have age and gender of the subject.

5.4 Experimental Design and Results

5.4.1 GAN Based Training

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, 40]. 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.

We used two Fully Connected Neural Networks (FNN), where one is a generator and the other is a discriminator. 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

(a) Full Signal

(b) Happy Signal

(c) Surprise Signal

(d) Disgust Signal

(e) Angry Signal

(f) Fear Signal

(g) Sad Signal

(h) Neutral Signal

Figure 5.1: Physiological Signals of a subject in training set.[37]

discriminator has 7 layers - 5 dense layers and 2 dropout layers. The first layer is the 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.

5.4.1.1 *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.

5.4.1.2 *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 $\mathbf{G}_{prob} = [D_1, D_2, \dots, D_7]$, where D_i is the probability output from the i^{th} GAN discriminator. Given $\mathbf{G}_{prob}$, we then train a random forest to recognize emotions.

5.4.2 Hand-crafted Features Training

Motivated by the work from Liu et al. [28], we also created an 8-dimension vector of hand-crafted features $G_{hc} = [\textit{min}, \textit{max}, \textit{mean}, \textit{variance}, \textit{mean abs diff}, \textit{second mean abs diff}, \textit{age}, \textit{gender}]$, 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.

5.4.3 Results

5.4.3.1 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.

5.4.3.2 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

Table 5.1: Physiological Signal Based Emotion Recognition Results - Validation and Test

Network	Validation	Test
GAN	14.58%	15.18%
Random Forest	15.44%	19.17%
Ensemble	-	17.48%

Table 5.2: Physiological Signal Based Emotion Recognition - Emotion Wise Accuracy [37]

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%

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 5.2, 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).

Chapter 6: Discussion and Future Work

We detailed our proposed multi-modal approaches to 4 tracks of emotion in the wild:

1. Group emotion recognition
2. Drive gaze estimation
3. Engagement prediction
4. Physiological signal-based emotion recognition

More specifically, We showed how deep features from extracted from deep neural networks, trained on optical flow and mel spectrogram 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. Finally, using a challenging physiological signal dataset, we proposed the use of a GAN-based ensemble for recognizing emotion.

Recent studies[27, 48] have shown that analyzing facial expression and body pose of a person along with the context of the video can contribute significantly in recognizing group-level emotions. We also observed in our track, Driver Gaze Prediction, that facial features like head pose, eye gaze, or landmarks can get miscalculated under varying illumination conditions, camera angles, or facial occlusions (Figure 3.2). Further research will focus on studying the impact of these modalities in recognizing emotion and developing a model invariant to discussed limitations. For analyzing the engagement level, we considered a 1-D vector of hand-crafted feature’s mean and standard deviation over the time subject is watching a video. Next, we further propose to utilize the temporal correlation of these

features as well. A lot of efforts have also been made to establish relationships between physiological signals and emotion recently[35]. However, there is still active research ongoing to figure out what physiological signals are most significantly relevant to emotion changes. Another issue is obtaining genuine emotions which are dependent heavily on stimulating tasks or materials. As our next step, we propose to establish a relationship between brain's activity and emotions using the dataset BU-EEG[25] which provides 128-channel EEG signals and both posed facial actions and spontaneous expression face videos.

References

- [1] Abdulkareem Al-Alwani. Mood extraction using facial features to improve learning curves of students in e-learning systems. *International Journal of Advanced Computer Science and Applications*, 7(11), 2016.
- [2] Peter A Anderson and Laura K Guerrero. Handbook of communication and emotion: Research, theory, applications, and contexts, 1998.
- [3] Tadas Baltrušaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. Openface 2.0: Facial behavior analysis toolkit. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 59–66. IEEE, 2018.
- [4] Tadas Baltrušaitis, Peter Robinson, and Louis-Philippe Morency. Openface:an open source facial behavior analysis toolkit. *Workshop on Applications of Computer Vision (WACV), IEEE Winter Conference on. IEEE, pp. 1–10, 2016*, 2016.
- [5] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.
- [6] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.
- [7] F. Chollet. Xception: Deep learning with depthwise separable convolutions. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1800–1807, 2017.

- [8] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 886–893. IEEE, 2005.
- [9] Tao Deng, Hongmei Yan, Long Qin, Thuyen Ngo, and BS Manjunath. How do drivers allocate their potential attention? driving fixation prediction via convolutional neural networks. *IEEE Transactions on Intelligent Transportation Systems*, 21(5):2146–2154, 2019.
- [10] Abhinav Dhall, Roland Goecke, Simon Lucey, and Tom Gedeon. Acted facial expressions in the wild database. 2011.
- [11] P. Ekman. What the face reveals: Basic and app studies of spon exp using the facial action coding system (facs). *Oxford University Press*, 1997.
- [12] D. Fabiano and S. Canavan. Spontaneous and non-spontaneous 3d facial expression recognition using a statistical model with global and local constraints. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pages 3089–3093, 2018.
- [13] Diego Fabiano and Shaun Canavan. Emotion recognition using fused physiological signals. In *International Conference on Affective Computing and Intelligent Interaction (ACII)*, 2019.
- [14] Maayan Frid-Adar, Idit Diamant, Eyal Klang, Michal Amitai, Jacob Goldberger, and Hayit Greenspan. Gan-based synthetic medical image augmentation for increased cnn performance in liver lesion classification. *Neurocomputing*, 321:321–331, 2018.
- [15] Maayan Frid-Adar, Eyal Klang, Michal Amitai, Jacob Goldberger, and Hayit Greenspan. Synthetic data augmentation using gan for improved liver lesion classification. In *2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018)*, pages 289–293. IEEE, 2018.

- [16] Shreya Ghosh, Abhinav Dhall, Garima Sharma, Sarthak Gupta, and Nicu Sebe. 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. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, pages 2672–2680, 2014.
- [18] Aarush Gupta, Dakshit Agrawal, Hardik Chauhan, Jose Dolz, and Marco Pedersoli. An attention model for group-level emotion recognition. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, pages 611–615, 2018.
- [19] S Hinduja, S Canavan, and G Kaur. Multimodal fusion of physiological signals and facial action units for pain recognition. In *15th IEEE International Conference on Automatic Face and Gesture Recognition*, pages 387–391, 2020.
- [20] L. Jeni et al. Facing imbalanced data—recommendations for the use of performance metrics. In *International Conference on Affective Computing and Intelligent Interaction and workshops*, 2013.
- [21] Amanjot Kaur, Aamir Mustafa, Love Mehta, and Abhinav Dhall. Prediction and localization of student engagement in the wild. In *2018 Digital Image Computing: Techniques and Applications (DICTA)*, pages 1–8. IEEE, 2018.
- [22] Davis E King. Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research*, 10:1755–1758, December 2009.
- [23] Diederik P Kingma et al. 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. Deap: A database for emotion analysis; using physiological signals. *IEEE transactions on affective computing*, 3(1):18–31, 2011.
- [25] X. Li, X. Zhang, H. Yang, W. Duan, W. Dai, and L. Yin. An eeg-based multi-modal emotion database with both posed and authentic facial actions for emotion analysis. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*, pages 336–343, 2020.
- [26] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.
- [27] Chuanhe Liu, Wenqiang Jiang, Minghao Wang, and Tianhao Tang. Group level audio-video emotion recognition using hybrid networks. In *Proceedings of the 2020 International Conference on Multimodal Interaction, ICMI '20*, page 807–812, New York, NY, USA, 2020. Association for Computing Machinery.
- [28] Yang Liu, Tom Gedeon, Sabrina Caldwell, Shouxu Lin, and Zi Jin. Emotion recognition through observer’s physiological signals. *arXiv preprint arXiv:2002.08034*, 2020.
- [29] Bruce D. Lucas and Takeo Kanade. An iterative image registration technique with an application to stereo vision. In *Proceedings of the 7th International Joint Conference on Artificial Intelligence - Volume 2, IJCAI’81*, page 674–679, San Francisco, CA, USA, 1981. Morgan Kaufmann Publishers Inc.
- [30] Cristina Olaverri-Monreal, Ahmed Elsherbiny Hasan, Jonathan Bulut, Moritz Körber, and Klaus Bengler. Impact of in-vehicle displays location preferences on drivers’ performance and gaze. *IEEE Transactions on Intelligent Transportation Systems*, 15(4):1770–1780, 2014.

- [31] Sally Planalp. *Communicating emotion: Social, moral, and cultural processes*. Cambridge University Press, 1999.
- [32] Garima Sharma, Shreya Ghosh, and Abhinav Dhall. Automatic group level affect and cohesion prediction in videos. In *International Conference on Affective Computing and Intelligent Interaction Workshops and Demo*, pages 161–167, 2019.
- [33] Jonathan Shen, Ruoming Pang, Ron J Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, Rj Skerrv-Ryan, et al. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4779–4783. IEEE, 2018.
- [34] Jianbo Shi et al. Good features to track. In *1994 Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 593–600. IEEE, 1994.
- [35] Lin Shu, Jinyan Xie, Mingyue Yang, Ziyi Li, Zhenqi Li, Dan Liao, Xiangmin Xu, and Xinyi Yang. A review of emotion recognition using physiological signals. *Sensors*, 18(7), 2018.
- [36] Ramprakash Srinivasan and Aleix M Martinez. Cross-cultural and cultural-specific production and perception of facial expressions of emotion in the wild. *IEEE Transactions on Affective Computing*, 2018.
- [37] Shivam Srivastava, Saandeep Aathreya Sidhapur Lakshminarayan, Saurabh Hinduja, Sk Rahatul Jannat, Hamza Elhamdadi, and Shaun Canavan. Recognizing emotion in the wild using multimodal data. In *Proceedings of the 2020 International Conference on Multimodal Interaction, ICMI ’20*, page 849–857, New York, NY, USA, 2020. Association for Computing Machinery.

- [38] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.
- [39] Chinchu Thomas and Dinesh Babu Jayagopi. Predicting student engagement in classrooms using facial behavioral cues. In *Proceedings of the 1st ACM SIGCHI International Workshop on Multimodal Interaction for Education*, pages 33–40, 2017.
- [40] Luan Tran, Xi Yin, and Xiaoming Liu. Disentangled representation learning gan for pose-invariant face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1415–1424, 2017.
- [41] Gerben A Van Kleef and Agneta H Fischer. Emotional collectives: How groups shape emotions and emotions shape groups. *Cognition and Emotion*, 30(1):3–19, 2016.
- [42] Anne Veenendaal, Elliot Daly, Eddie Jones, Zhao Gang, Sumalini Vartak, and Rahul S Patwardhan. Group emotion detection using edge detection mesh analysis. *Computer Science and Emerging Research Journal*, 2, 2014.
- [43] Yafei Wang, Tongtong Zhao, Xueyan Ding, Jiming Bian, and Xianping Fu. Head pose-free eye gaze prediction for driver attention study. In *2017 IEEE International Conference on Big Data and Smart Computing (BigComp)*, pages 42–46. IEEE, 2017.
- [44] Ghada Zamzmi, Pai Chih-Yun, Dmitry Goldgof, R Kasturi, Terri Ashmeade, and Yu Sun. A comprehensive and context-sensitive neonatal pain assessment using computer vision. *IEEE Transactions on Affective Computing*, 2019.
- [45] Ghada Zamzmi et al. Machine-based multimodal pain assessment tool for infants: a review. *arXiv preprint arXiv:1607.00331*, 2016.

- [46] Guoying Zhao and Matti Pietikainen. 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.
- [47] Z. Zheng et al. Multimodal spontaneous emotion corpus for human behavior analysis. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.
- [48] Hengshun Zhou, Debin Meng, Yuanyuan Zhang, Xiaojiang Peng, Jun Du, Kai Wang, and Yu Qiao. Exploring emotion features and fusion strategies for audio-video emotion recognition. In *2019 International Conference on Multimodal Interaction, ICMI '19*, page 562–566, New York, NY, USA, 2019. Association for Computing Machinery.

Appendix A: Copyright Permissions

This study was also published at ICMI conference, 2020 as a research paper with the same title[37]. The image below includes the author's copyright information for all the figures and tables reproduced from the research paper.[37]

About the Author

Shivam Srivastava obtained a Bachelor's degree in Computer Science and Engineering in 2016 from JSS Academy of Technical Education, India. After his bachelors, he worked as a software engineer for three years building distributed and scalable systems and web services. During the pursuit of his current Master's degree at University of South Florida, his research focus has been on Computer Vision, Deep Learning and, Affective Computing.