

DO I HAVE YOUR ATTENTION: A LARGE SCALE ENGAGEMENT PREDICTION DATASET AND BASELINES

Monisha Singh, Ximi Hoque, Donghuo Zeng, Yanan Wang, Kazushi Ikeda, Abhinav Dhall

ABSTRACT

The degree of concentration, enthusiasm, optimism, and passion displayed by individual(s) while interacting with a machine is referred to as ‘user engagement’. Engagement comprises of behavioural, cognitive, and affect related cues [1]. To create engagement predictions systems, which can work in real-world conditions it is quintessential to learn from rich diverse datasets. To this end, a large scale multi-faceted engagement in the wild dataset *EngageNet* is proposed. 31 hours duration data of 127 participants representing different illumination conditions is recorded. Thorough experiments are performed exploring applicability of different features action units, eye gaze and head pose and transformers. To further validate the rich nature of the dataset, evaluation is also performed on the EngageWild dataset [2]. The experiments show the usefulness of the proposed dataset. The code, models and dataset will be made publicly available.

Index Terms— Engagement, behavior, cognitive

1. INTRODUCTION

COVID-19 has accelerated adoption of online meetings, education classes and business activities. From online classes to online product launches virtual meetings have been adopted and accepted rapidly. Although these online meetings have numerous advantages, such as accessibility and affordability, it suffers from few challenges in the form of engagement. Lets take the example of a classroom, for an instructor it is easier to gauge a student’s interest in the didactic content taught in a classroom-based setting by looking for signs like head nodding and yawning. However, observing such a change in behavior in online classroom settings is challenging. It is necessary to develop automatic mechanisms for detecting student engagement to increase student participation and decrease academic ennui. Teachers and instructors can adjust their content and methods to cater to the interests of their students with the use of engagement detection technologies.

The authors, Monisha Singh, Ximi Hoque, and Abhinav Dhall are with the Computer Science and Engineering Department, IIT Ropar, Punjab-140001, India (e-mail: monisha.21csz0023@iitrpr.ac.in; 2021aim1015@iitrpr.ac.in; abhinav@iitrpr.ac.in). The Authors, Donghuo Zeng, Yanan Wang, and Kazushi Ikeda are with the KDDI Research, Japan(e-mail: do-zeng@kddi-research.jp; wa-yanan@kddi-research.jp; kz-ikeda@kddi.com).

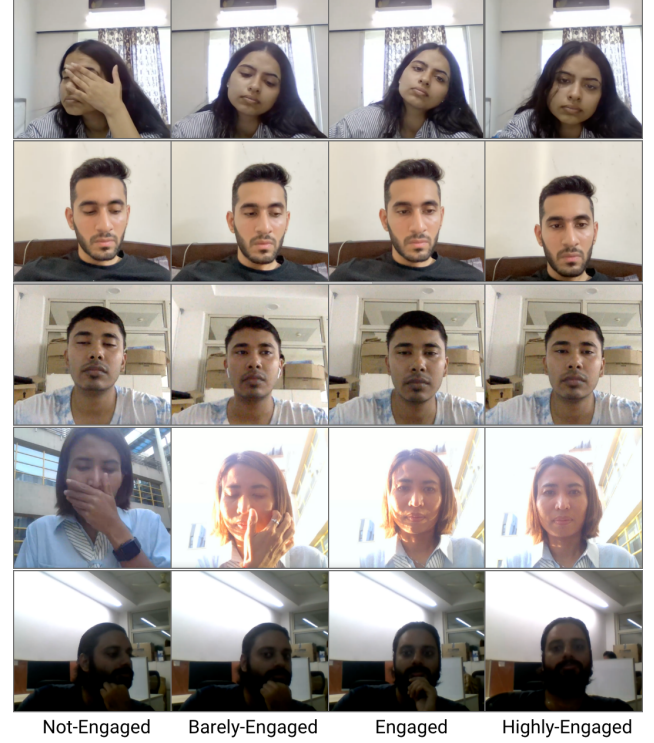


Fig. 1. Frames from the proposed *EngageNet* dataset. The engagement categories are mentioned at the bottom. Notice the ‘in-the-wild’ aspect of the data.

The necessity of measuring engagement is not limited to education settings but also in UX research during scenarios such as A/B testing and even in human robot interaction. From the perspective of social robots, how can a robot understand if the user is engaged in an activity and an interrupt from the robot may disturb the user? Similarly, if user is not engaged in the conversation the robot can stop speaking. For such circumstances, it is important to predict user engagement.

Several aspects of user engagement have been canopied in the literature. Three predominantly identified characteristics, behavioral, cognitive, and affective, consequently bear a considerable impact on user engagement [3]. Behavioral engagement is user’s visible participation in their learning; head pose, eye gaze, and facial expression are some of the visual cues of behavioral engagement [3]. Cognitive engagement

is the degree to which students are prepared and competent to tackle their learning tasks. Affective engagement refers to how the user feels in a particular situation (for eg amused, joyful, or proud etc). Designing and creating student engagement technologies requires access to statistics that capture the three aspects of engagement. A few publicly accessible datasets (DAiSEE [4], HBCU [2], EmotiW [5]) that include behavioral, cognitive, or affective components of engagement partially fill this demand. However, there are not any large sized publicly available datasets that are currently accessible to the general public that records multiple facets of engagement. By creating a large dataset, which captures both the behavioral and cognitive aspects of engagement, we try to address this deficiency in our paper. We refer to the proposed dataset at *EngageNet*. The dataset is generated by designing a study in which participants watch several stimuli videos and answer questions based on the stimuli videos over a web-based interactive platform. A personality-based questionnaire, three video stimuli, and three questionnaires based on the stimuli make up the core of our study design. The main contributions of the paper are described below:

- Large scale dataset EngageNet 31 hours of data of 127 subjects, and over 11.3K 10-sec clips is presented.
- To capture multiple facets of engagement, we included behavioral as well as cognitive components of engagement in dataset recording.
- To establish baselines for evaluating behavioral engagement in multimodal setups, we conducted extensive experiments.
- In the video stimuli, we introduced digital teachers (virtual avatar) to study effect of digital avatars on engagement in future.

For contributing to the research in the community the code, evaluation protocols, features and data will be made publicly available.

2. RELATED WORK

Engagement prediction is a data driven problem, there are a handful of datasets available for public use [5, 4]. There have been several initiatives to address the requirement for freely accessible datasets for training engagement detection models. In this section, we first discuss the standard engagement labelled datasets.

The DAiSEE dataset, created by Gupta et al. [4], includes 112 subjects of which 80 are males, and 32 are females. Videos in the dataset were recorded under three different lighting conditions—light, dark, and neutral—in unconstrained locations, such as dormitories, labs, and libraries. The video of the participants watching a video lecture was recorded by a webcam that was mounted on a Computer.

The 10-second-long videos in the dataset have been annotated by 10 unskilled crowdsourcing workers into four different levels: engaged, bored, confused, and frustrated.

The HBCU dataset [2] was gathered from undergraduate students who took part in a "Cognitive Skills Training Project" that ran between 2010 and 2011. The experimental data from 34 subjects (9 males and 25 females) were split into two groups: (a) 26 subjects who participated in the Cognitive Skills Training study, in spring 2011. (b) 8 subjects participated in the Cognitive Skills Training project, in the summer of 2011 at the University of California (UC). The test subjects either sat by themselves or with the experimenter while the cognitive skills training software was being played in a private space. A group of labelers were assembled from HBCU and UC computer science, cognitive science, and psychology departments, including both undergraduate and graduate students.

Kaur et al. [5] provided a database for the creation of their proposed method for detecting student engagement. Videos of 5-minute duration were captured from 78 individuals — 25 females and 53 males. To gather data, they have employed a variety of settings, including computer labs, dormitories, the outdoors, etc. A group of five annotators rated the subjects' level of participation in the videos. Not engaged, barely engaged, engaged, and highly engaged were the four possible engagement levels that were used to annotate the data. They used weighted Cohen's K with quadratic weights as the performance parameter to evaluate the reliability of annotators. Additionally, they have successfully addressed the issue of subjective bias and low annotator agreement. A brief comparison of the existing datasets is presented in Table 1.

Approaches for detecting engagement based on computer vision gather data from gestures, eye gaze, head movements, and facial expressions. This gathered data is then fed into the machine or deep learning models. Here we will go over a couple of the methods that have been suggested for engagement detection. One of the seminal work in this direction is by Whitehill et al. [6]. To predict the subject's engagement level, Kaur et al. [5] proposed a segment-based random forest model and LSTM-based deep sequence network baselines. Whitehill et al. [2] employed binary classifiers for the classification task. Three classifier combinations have been compared: GentleBoost with Box Filter features (Boost(BF)), Support vector machines with Gabor features, and Multinomial logistic regression with expression outputs from the Computer Expression Recognition Toolbox [7]. In the final stage, a regressor was fed the outputs of the binary classifiers to automatically detect engagement. For the classification job, Thomas et al. [8] used machine learning methods including Support Vector Machine (SVM) and Logistic Regression (LR). In a different paper, Thomas et al. [9] provided a system for measuring student engagement levels using the Temporal Convolutional Network (TCN). For video classification, Gupta et al. [4] used a 3D CNN model for full training and employed a long-term

Recurrent Convolutional Network for unifying visual and sequence learning. On the DAiSEE dataset Ali et al. [10] accurately identified disengaged videos and achieved better classification accuracy by employing a variety of temporal(LSTM and TCN) and Non-temporal(fully connected neural network, SVM and RF) models.

Between our EngageNet dataset and past efforts, we discovered significant variances. First off, very few of the earlier studies address multiple aspects of engagement. Our dataset, in turn, records engagement’s behavioral and cognitive facets. Second, to the best of our knowledge, our dataset is the largest dataset captured in the wild in varied illumination settings. Third, in addition to the labels that expert labelers annotated, we also collected the respondents’ retrospective self-labels. Finally, we have obtained personality data from participants that can help us understand how engagement and personality relate to one another in future works.

3. ENGAGENET DATASET

3.1. Study Design

A web based platform was designed for the data collection study. The participants were between the age range of 18 to 37 years. The usage of a computer or laptop with a quality webcam and a steady internet connection was advised to the participants. The participant could use the web based interface anywhere they felt comfortable. This enabled collection of data in diverse environments. We acquired demographic data including age and gender. Participants’ consent was also requested for the scope of usage of the data. Participants were also required to complete a pre-study personality-based questionnaire. This information will be shared as part of the meta-data with the EngageNet dataset.

To account for the subjective bias of the annotators, we presented three stimuli videos that were randomly shuffled at the start of each session. The average length of the study session was about 25 to 30 minutes, and each video stimulus was about 5 minutes long. To gauge the subject’s level of cognitive engagement, a brief questionnaire based on the content of each video stimulus was presented afterward. The participants were asked to self-identify their perceived level of engagement after completing the brief questionnaire. The participants were also questioned if they had previously watched the video and whether they would recommend it to others.

3.2. Data Collection, Pre-processing and Annotation

3.2.1. Data Collection

In this section, we discuss the study’s data recording paradigm. Broadly two sets of information are captured during the experiment: (i) capturing the participants on camera when they watch the stimulus video and (ii) a brief questionnaire is answered by the participant based on the con-

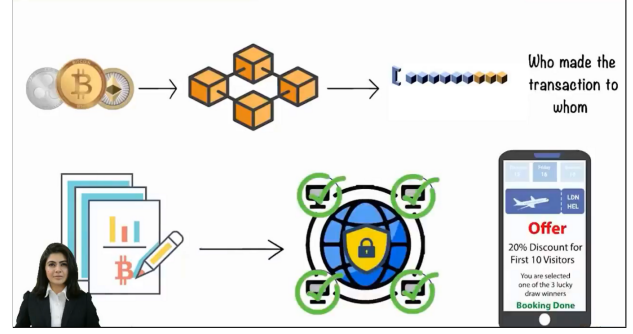


Fig. 2. A frame from one of the stimuli video used in data collection. Note the digital avatar on the bottom left.

tent of the viewed video stimulus. We used three educational videos as stimulus: (i) *Schrödinger’s cat: A thought experiment in quantum mechanics* (ii) *What is cryptocurrency?* (iii) *Where did English come from?* To appeal to a wider audience it was important to select videos with different themes. The audio part was orated by a digital avatar added to the videos. The reason to add a digital avatar (teacher) is to check for the engagement in videos with a digital teacher and without it. Note these experiments will be a future work. The digital teacher has been added using ¹. A screenshot of the stimuli is shown in Figure 2.

There are 127 participants in the dataset (83 males and 44 females). The participants are between the ages of 18 and 37. A total of 474 videos, each lasting 5 to 10 minutes, were gathered. The dataset is gathered in an unconstrained environment at various sites, including computer labs, dorm rooms, open spaces, etc. The videos were recorded in at least 640x480 resolution.

3.2.2. Data Pre-processing

Overall for the EngageNet dataset, the video duration against stimuli and question response varied between 5 to 10 minutes. The videos were further divided into 10 seconds duration. This was performed as the engagement level varies over time in long videos [12]. The final data contains videos of 10 seconds in duration.

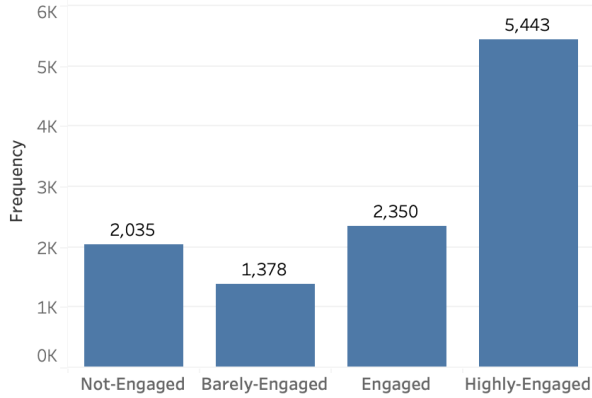
3.2.3. Annotation

The participants in the videos were assigned to different engagement classes by a group of three annotators. We employed the four-engagement class definitions that Kaur et al. [5] stated. The following are the engagement classes: *“Not Engaged”* denotes a wholly disengaged subject, such as one who frequently glances away from the screen and appears disinterested. *“Barely Engaged”* refers to being only minimally attentive, like when the individual fidgets restlessly in the chair or hardly opens his or her eyes. The term *“Engaged”*

¹The Artiste AI Studio - www.kroop.ai

Table 1. Comparison of engagement labelled datasets.

Dataset	Setting	#Participants	Annotators	Duration	Total Hours	Level of Abstraction-Combination
DAiSEE [4]	Virtual learning, lecture, non-interactive, in the wild	112	Observers (10 Non-Experts)	10 sec	25	Affective dimension
HBCU [2]	Cognitive Skills Training [11], interactive	34	Observers (10 sec-7 labelers 60 sec-2 labelers)	10, 60 sec	25.5	Behavioral, Cognitive dimension
EngageWild [5]	Computer based, lecture, non interactive, in the wild	78	Observers (5 experts)	5 min	16.5	Behavioral dimension
EngageNet (Ours)	Computer based, talk, interactive, in the wild, digital teacher avatar	127	Observers (3 experts)	10 sec	31	Behavioral, Cognitive dimension

**Fig. 3.** Distribution of Engagement Classes in EngageNet.

refers to participants who are interested in the content, such as when they interact with the video and appear to like it. *“Highly Engaged”* refers to a subject who appears to be entirely attentive and glued to the screen. Additionally, we used weighted Cohen’s K [13] with quadratic weights as the performance parameter to evaluate the reliability of annotators.

3.3. Data Split and Statistics

We created subject independent data split leading to 90 participants in *Train*, 11 participants in *Validation*, and 26 participants in *Test*. Total videos for *Train*, *Validation*, and *Test* sets were 7906, 1071, and 2257, respectively. The study included 127 participants, of which 65.35% were male, and 34.65% were female. The participants were between the ages of 18 and 37. The dataset consists of approximately 11.3K clips lasting 10 seconds, or roughly 31 hours of data. The class distribution of labels is as follows: *Highly Engaged* (48.57%), *Engaged* (20.97%), *Barely Engaged* (12.3%), and *Not Engaged* (18.16%).

4. EXPERIMENTS

4.1. Feature Extraction

Face and facial landmark positions were identified in each frame using the OpenFace toolkit [14]. A window of frames’ worth of features is taken from the subsampled video. The following batch of frames is then extracted to compute features before the window slides. We computed the mean and standard deviation of head location, head rotation, gaze vector and gaze angle of both eyes, facial action units to capture the changing head pose, gaze, and facial expression in each segment. The final dimensions of eye gaze, head pose, and facial AUs features were 16, 12, and 70, respectively, after calculating the mean and standard deviation.

4.2. Engagement Prediction Baselines

To predict the engagement class the engineered features are fed into several deep learning models. The details of these models are described below.

4.2.1. Long Short-Term Memory Networks

Deep networks built on LSTMs [15] were chosen because they are effective at identifying long-term dependencies in sequential data. The network is made up of two LSTM layers, which are followed by a Dropout layer to stop the model from overfitting and a Flatten layer to get a single activation vector for all segments of the video. The resulting feature vector is fed into three fully connected layers, two of which have ReLU activation, while the third layer has softmax activation. Empirically, we observed 10 cells (segments) as optimum setting.

4.2.2. CNN-LSTM

The CNN-LSTM [16] based model was chosen since CNN has demonstrated ground-breaking breakthroughs in the field of computer vision and LSTM is widely recognized for learning dependencies between time steps in time series

	LSTM					CNN-LSTM						TCN			
	G	HP	AU	G+HP	G+HP+AU	G	HP	AU	G+AU	HP+AU	G+HP+AU	G	HP	AU	G+HP+AU
Val.	61.34	67.41	62.75	65.17	66.67	60.78	67.88	62.00	62.46	65.17	67.13	63.03	67.51	64.24	68.72
Test	60.33	59.97	59.80	62.85	62.77	59.62	61.75	60.64	60.77	63.30	64.27	60.15	60.28	59.00	62.90

Table 2. Comparison (%) of different feature combinations. Here, G:Eye Gaze, HP:Head Pose, AU:Facial Action Units.

Features	Validation	Test
Gaze	55.45	54.87
Gaze + Head Pose	64.45	62.19
Gaze + Head Pose + AU	65.40	63.56

Table 3. Performance comparison (%) of Transformer model on different combination of different features.

and sequential data. Two convolutional layers, MaxPooling, TimeDistributed Flatten, two successive LSTM layers, Dropout, Flatten, and ultimately a fully connected layer with softmax activation make up the model.

4.2.3. Temporal Convolutional Network

The TCN [17] based model was chosen since it has proven to be more effective than other sequential models like LSTMs and can capture long-term dependencies. To produce a single vector of activations, the TCN-based deep model comprises of two TCN layers placed one after the other, followed by a Dropout layer and a Flatten layer. The resultant vector is then passed into a series of fully connected layers, the last layer of which is activated using softmax.

4.2.4. Transformer

Recent works [18] in computer vision have witnessed increase in performance due to the Transformers architectures [19]. As a result, we developed a model based on transformers with positional encoding with an attention layer stacked above and a MLP to predict engagement. The tokens were created by dividing the 10 second samples into 20 segments. Each segment's final feature is mean and standard deviation of features from all frames belonging to a segment. 20 tokens was choosen after emperical evaluation.

4.3. Baseline Results

The following experiments were performed using the extracted features and proposed deep learning baseline models.

4.3.1. LSTM, CNN-LSTM, and TCN Baseline Evaluations

Eye gaze, head position, and facial AUs features were input to LSTM, CNN-LSTM, and TCN models, separately. Later, the LSTM model was input a combination of gaze and head pose

Train Dataset	Test Dataset	Test MSE
EmotiW [12]	EmotiW	0.082
EngageNet (ours)	EmotiW	0.162

Table 4. Cross dataset learning results (%).

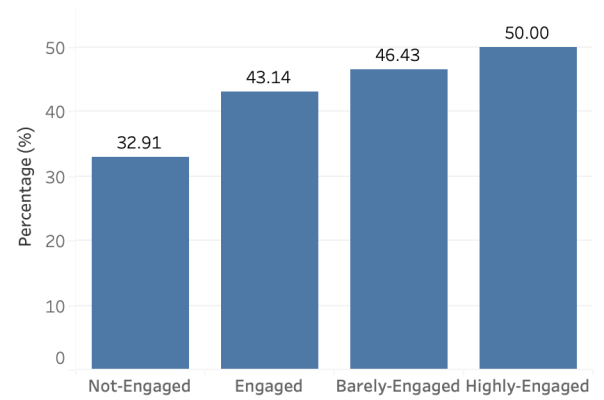


Fig. 4. Engagement Labels vs percentage of correct responses per label category. Note that not-Engaged class has least correct answers.

features, and the CNN-LSTM was input first a combination of gaze and facial AUs, and then a combination of head pose and facial AUs. Finally, the three models were input a combination of gaze, head pose, and facial AUs. The outcome of this experiment is shown in Table 2. We observe that multi-modal features outperformed individual features on the Test set. Particularly, the combination of eye gaze, head pose, and facial AUs features performs remarkably well in predicting engagement, which is suggestive of the fact that gaze, head pose, and facial expression were all taken into account when annotating the video clips.

4.3.2. Transformer Baseline Evaluations

The transformer model was first evaluated with the gaze and followed by a combination of gaze and head pose, gaze, head pose, and facial AU features. Table 3 shows the outcome of this experiment. It is observed that the fusion of features as tokens performs better than gaze only and gaze + head pose features.

4.3.3. Cross Dataset Analysis

For evaluating the effectiveness and diversity of EngageNet, we evaluated the performance of method trained on EngageNet and evaluated on the EmotiW dataset [5]. For fair comparison, we follow the protocol and label normalisation mentioned in [5]. The transformer model was trained in a regression setting to compare the performance of our model with the baseline model of EmotiW [12]. In the series of experiments, first, the transformer model was trained and tested on the EmotiW dataset, then the model was trained and tested on our dataset, and finally, the model was trained on our dataset and tested on the EmotiW dataset. The performance comparison is shown in Table 4. The comparable results show the effectiveness of the proposed EngageNet dataset.

4.4. Cognitive Engagement Analysis

We analyzed the annotated labels and the performance of participants in the video based questionnaire to understand the cognitive engagement of the participants. We computed the mode of engagement labels of 10 second segments in order to identify the engagement label for the unsegmented videos, which were broken into 10 second segments for annotation. The analysis is shown in Figure 4. We find that *Not Engaged* category has the least correct answers and *Highly Engaged* has the highest percentage of correct answers among the engagement class itself. This also indicates that higher user engagement can lead to better learning.

5. CONCLUSION

We present a large scale user engagement dataset EngageNet. Both behavioral and cognitive engagement is considered during the dataset creation. We perform extensive baseline analysis with facial features and recent machine learning techniques. The result show the richness and complexity of the data. We observe that highly engaged user answer the largest number of questions correctly as compared with other engagement classes. We also evaluate the data quality by evaluating on another engagement dataset. We believe the dataset, the extensive baselines, and the cognitive analysis will serve as useful benchmarking resource in the research community.

6. REFERENCES

- [1] Heather L O'Brien and Elaine G Toms, "The development and evaluation of a survey to measure user engagement," *Journal of the American Society for Information Science and Technology*, vol. 61, no. 1, pp. 50–69, 2010.
- [2] Jacob Whitehill, Zewelanji Serpell, Yi-Ching Lin, Aysha Foster, and Javier R. Movellan, "The faces of engagement: Automatic recognition of student engagement from facial expressions," *IEEE Transactions on Affective Computing*, vol. 5, no. 1, pp. 86–98, 2014.
- [3] 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)*. IEEE, 2018, pp. 1–8.
- [4] Abhay Gupta, Arjun D'Cunha, Kamal Awasthi, and Vineeth Balasubramanian, "Daisee: Towards user engagement recognition in the wild," *arXiv preprint arXiv:1609.01885*, 2016.
- [5] 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)*, 2018, pp. 1–8.
- [6] Jacob Whitehill, Zewelanji Serpell, Yi-Ching Lin, Aysha Foster, and Javier R Movellan, "The faces of engagement: Automatic recognition of student engagement from facial expressions," *IEEE Transactions on Affective Computing*, vol. 5, no. 1, pp. 86–98, 2014.
- [7] Gwen Littlewort, Jacob Whitehill, Tingfan Wu, Ian Fasel, Mark Frank, Javier Movellan, and Marian Bartlett, "The computer expression recognition toolbox (cert)," in *2011 IEEE International Conference on Automatic Face & Gesture Recognition (FG)*. IEEE, 2011, pp. 298–305.
- [8] 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*, 2017, pp. 33–40.
- [9] Chinchu Thomas, Nitin Nair, and Dinesh Babu Jayagopi, "Predicting engagement intensity in the wild using temporal convolutional network," in *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, 2018, pp. 604–610.
- [10] Ali Abedi and Shehroz Khan, "Affect-driven engagement measurement from videos," *arXiv preprint arXiv:2106.10882*, 2021.
- [11] Jacob Whitehill, Zewelanji Serpell, Aysha Foster, Yi-Ching Lin, Brittney Pearson, Marian Bartlett, and Javier Movellan, "Towards an optimal affect-sensitive instructional system of cognitive skills," in *CVPR 2011 WORKSHOPS*. IEEE, 2011, pp. 20–25.
- [12] Abhinav Dhall, Garima Sharma, Roland Goecke, and Tom Gedeon, "EmotiW 2020: Driver gaze, group emotion, student engagement and physiological signal based challenges," in *Proceedings of the 2020 International Conference on Multimodal Interaction*, 2020, pp. 784–789.
- [13] Jacob Cohen, "Weighted kappa: nominal scale agreement provision for scaled disagreement or partial credit," *Psychological bulletin*, vol. 70, no. 4, pp. 213, 1968.
- [14] Tadas Baltrušaitis, Peter Robinson, and Louis-Philippe Morency, "Openface: An open source facial behavior analysis toolkit," in *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2016, pp. 1–10.
- [15] Sepp Hochreiter and Jürgen Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
- [16] Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell, "Long-term recurrent convolutional networks for visual recognition and description," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 2625–2634.

- [17] Colin Lea, Michael D Flynn, Rene Vidal, Austin Reiter, and Gregory D Hager, “Temporal convolutional networks for action segmentation and detection,” in *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 156–165.
- [18] Salman Khan, Muzammal Naseer, Munawar Hayat, Syed Waqas Zamir, Fahad Shahbaz Khan, and Mubarak Shah, “Transformers in vision: A survey,” *ACM computing surveys (CSUR)*, vol. 54, no. 10s, pp. 1–41, 2022.
- [19] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin, “Attention is all you need,” *Advances in neural information processing systems*, vol. 30, 2017.