DAiSEE: Towards User Engagement Recognition in the Wild

Abhay Gupta, Arjun D'Cunha, Kamal Awasthi and Vineeth Balasubramanian

Abstract—The difference between real and virtual worlds is shrinking at an astounding pace. With more and more users working on computers to perform a myriad of tasks from online learning to shopping, interaction with such systems is an integral part of life. In such cases, recognizing a user's engagement level with the system (s)he is interacting with can change the way the system interacts back with the user. This will lead not only to better engagement with the system but also pave the way for better human-computer interaction. Hence, recognizing user engagement can play a crucial role in several contemporary vision applications including advertising, healthcare, autonomous vehicles, and e-learning. However, the lack of any publicly available dataset to recognize user engagement severely limits the development of methodologies that can address this problem. To facilitate this, we introduce DAiSEE, the first multi-label video classification dataset comprising of 9068 video snippets captured from 112 users for recognizing the user affective states of boredom, confusion, engagement, and frustration "in the wild". The dataset has four levels of labels namely - very low, low, high, and very high for each of the affective states, which are crowd annotated and correlated with a gold standard annotation created using a team of expert psychologists. We have also established benchmark results on this dataset using state-of-the-art video classification methods that are available today. We believe that DAiSEE will provide the research community with challenges in feature extraction, context-based inference, and development of suitable machine learning methods for related tasks, thus providing a springboard for further research.

Index Terms—Affect Recognition, User Engagement in the Wild, DAiSEE, E-Learning Environments

✦

1 INTRODUCTION

THE progress from research to consumer technologies in classical recognition problems in computer vision have been possible over the last decade due to the availability of large-scale datasets which are made available to researchers and industry practitioners alike. The ImageNet [1] and PASCAL VOC [2] challenges spearheaded the object recognition revolution, while the availability of datasets such as Microsoft COCO [3] and Cityscapes [4] in recent years has fueled the development of newer methods for semantic segmentation and vision-language joint understanding. Needless to say, the availability of more datasets for various problems and subproblems in computer vision will allow for a greater impact in translating research methodologies to businesses and, eventually, a positive influence on users' lives. This work is an effort in this direction - to provide a dataset (and benchmark results) for a vision problem of contemporary relevance: user engagement recognition.

The recognition of user engagement is increasingly relevant to a digital world that floods users with various kinds of content, and it is useful for systems to be "aware" of the user's engagement while providing content. For example, the recognition of user engagement is critical to various domains including advertising - web or television (is the viewer really watching this advertisement on television or on YouTube? is (s)he bored?); healthcare - as pertains to detection and interventions for children with learning or cognitive disabilities (does a child's engagement levels indi-
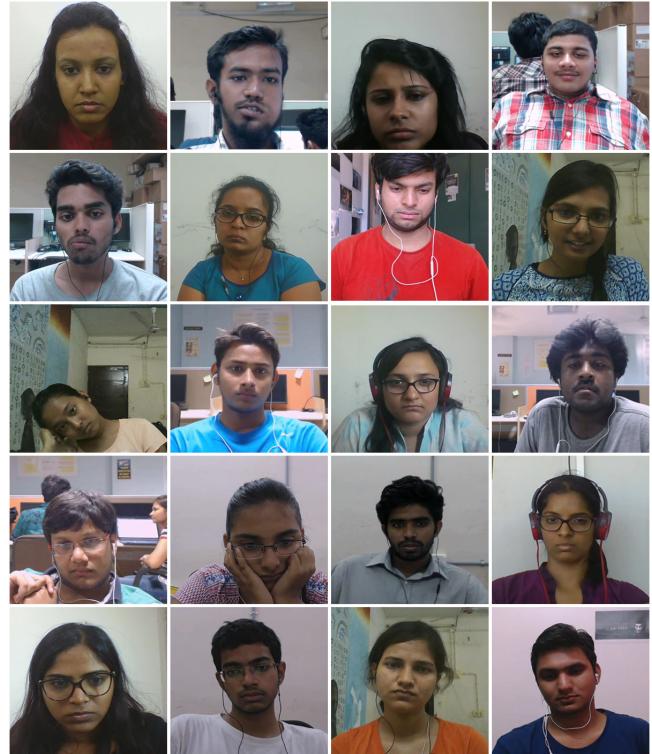
- *A. Gupta abhgup@microsoft.com*
- *A. D'Cunha cs14btech11039@iith.ac.in*
- *K. Awasthi 201451072@iiitvadodara.ac.in*
- *V. Balasubramanian vineethnb@iith.ac.in*

Fig. 1: Examples of video frames from DAiSEE: The dataset captures real-world challenges of recognizing user engagement in natural settings.

cate a tendency for autism or ADHD? are there particular events or objects that interest or frustrate the child?); e-learning (which parts of a lecture are confusing for most students who watch it? how engaged are students in a video?); and autonomous vehicles (Is the driver distracted or is he paying attention? Can we predict his actions based on his confusion/frustration levels?). There have been recent research efforts to develop methods for user engagement recognition [5] [6], but the resulting datasets are very small and not publicly available. Further, existing commercially available affective recognition systems that attempt to track user engagement work within constrained settings and have limited use in real-world environments (illustrated in Section 2). This work introduces DAiSEE, a dataset that aims to facilitate research and development towards user engagement recognition in the wild.

Understanding affective states of a user, an important subarea of computer vision, has for a long time focused on datasets pertaining to the seven basic expressions: neutral, happiness, sadness, anger, disgust, surprise and contempt [7]. While recent efforts in the last few years have expanded datasets in this domain to cover affective states in terms of dimensional representations [8] [9], the vast subtleties in affective states necessitate the development of datasets for specific objectives. Recent trends, including [10] [9], corroborate this approach to help progress towards tangible outcomes.

E-learning environments provide one of the best use cases for studying user engagement in settings where a user interacts with a computer screen. With the accelerated growth of Massive Open Online Courses (MOOCs), there is a need to design intelligent interfaces which simulate the interactions that occur between a teacher and students in a class. The main drawback of existing e-learning systems is that they do not provide real-time interactive feedback to students (or instructors) during the content delivery process, as compared to traditional classroom learning. Surprisingly, MOOCs have a dropout rate of 91-93% [11] [12], with the completion rate for the first assignment being around 45%. An online survey [13] lists the top ten reasons for dropouts from such platforms to include: "poor course design", which included components such as lack of proper feedback, "lecture fatigue" in courses that had only video lectures, "lack of proper course introductions" and "student frustration". Such reasons motivate the need to improve feedback mechanisms to make these platforms more interactive. Understanding user engagement at various junctures of the e-learning experience can help design intuitive interfaces that support better knowledge absorption by students, help decrease dropout rates, and personalize the learning experience. This paper seeks to address the aforementioned issues, by making available a dataset that captures user engagement in the wild during e-learning sessions (which is relevant to any other application domain involving user interaction with a computer screen including advertising, healthcare, and autonomous vehicles).

In this work, we introduce DAiSEE (**D**ataset for **A**ffective **S**tates in **E-E**nvironments) (Figure 1 shows sample images), where e-environment means any environment with applications like e-shopping, e-healthcare, e-learning but not limited to them. DAiSEE will be made publicly available to the community for further research. In particular, we focus on engagement, frustration, confusion, and boredom as the affective states for this work, all of which are relevant to user engagement and concomitant applications. Considering that these affective states are subtle, the annotations for this dataset are crowd-sourced, strengthened using a gold standard created using expert psychologists and all annotations (including each individual crowd annotation) are provided along with this dataset. We also benchmark the performance of standard video classification and deep learning-based models on this dataset to provide a baseline for further research. DAiSEE has the potential to be applied to the following domains, but is not limited to them:

- In e-learning, to support personalized learning for an individual user, and thus, increase retention rate.
- In advertising, to gain insights into what piques a customer's interest, to see how engaging an advertisement is, and to then provide personalized advertising to customers.
- In e-shopping, to understand user preferences in terms of specific items, or a larger domain (clothing, jewelry, electronics, etc.) of interest to a user, and thus allow personalization in the shopping experience.
- In healthcare, to understand and detect early signs of ADHD or autism, or to study fatigue in radiologists viewing images on a screen.
- In autonomous vehicles, to capture the driver's engagement level, or to predict the driver's future actions based on his/her confusion or frustration levels.

The rest of the paper is organized as follows: we discuss the background and related work in Section 2. In Section 3, we introduce the dataset, its components, and details of the process involved in collecting it. In Section 4, we provide a baseline performance on all labels in this dataset using recent state-of-the-art deep learning models for video classification. Section 5 discusses the challenges that DAiSEE poses in establishing a benchmark for the dataset and Section 6 discusses the release and distribution of the dataset. Lastly, in Section 7, we summarize our work and suggest future directions with our dataset.

2 BACKGROUND AND RELATED WORK

Determining the affective state of a user using computer vision and machine learning methods has been studied for over two decades now (please see [14] [15] for surveys).

2.1 Affect Recognition Datasets

Early databases of facial expressions such as CK+ [16] [7], MMI [17], and MultiPie [18] were captured in a lab-controlled environment where the subjects portrayed different facial expressions, which resulted in a clean and high-quality dataset of posed facial expressions. However, posed expressions may differ from real-life spontaneous expressions. Thus, datasets like DISFA [19], AM-FED [20] which captured facial responses to stimuli or Belfast [21] which captured emotions while performing laboratory-based tasks became a trend in the affective computing community. These datasets captured multi-model affects such as voice, biological signals and often worked with a series of frames

to allow for both temporal and dynamic expression recognition. However, the diversity of the datasets is limited due to the number of subjects, head pose variation, and environmental conditions.

To develop systems that are based on natural, unposed facial expressions - datasets in the wild became important. The *Acted Facial Expressions in the Wild (AFEW)* released by Dhall et al. [22] addressed the issue of temporal facial expressions by capturing videos of subjects. The dataset contains 330 subjects aged 1-77 years and is annotated with six basic expressions along with neutral. A static subset (SFEW) [23] of 700 images and 95 subjects was created by selecting some frames from AFEW and it contains unconstrained facial expressions, different head poses and close to real-world scenarios like occlusions and unconstrained illumination.

The *Facial Expression Recognition 2013 (FER-2013)* dataset [24] was created using the Google Search API that matched a set of 184 emotion-related keywords to capture the six basic expressions and the neutral expression. The dataset contains 35,887 images, most of which are in the wild. The images are grayscale, 48x48 pixels and have been cropped to retain only the face and the displayed emotions. FER-2013 is currently the biggest publicly available facial expression database in the wild settings. However, the resolution and quality of the images make it difficult for facial landmark detectors to extract landmarks and only categorical models of affect are provided with FER-2013.

The *Affective-MIT Facial Expression Dataset (AM-FED)* dataset [20] contains 242 facial videos (160K frames) of people watching commercials using their webcam. The recordings were taken in "wild" settings with varied contrast and illumination. The dataset has a frame-by-frame annotation for 14 FACS action units, head movements, and automatically detected landmark points. However, there is not much variation in head pose and the dataset has limited number of subjects.

The FER-Wild [25] dataset contains 24,000 images which match emotion-related terms queried from search engines. The dataset contains annotations for the six basic expressions and neutral from two human labelers. Compared to FER-2013 [24], FER-Wild only has higher resolution with facial landmarks points. It still does not address limitations like categorical modeling of affect and portraying more expressions.

EmotioNet [26] consists of one million images of facial expressions downloaded from the Internet by selecting all words derived from the word "feeling" in WordNet [27]. The images are automatically annotated with AUs and AU intensities after a face detector [28] was used to detect faces. The images are labeled one of 23 (basic or compound) emotion categories defined in [29] based on AUs. Experienced coders were used to annotate 100,000 images manually with AUs. EmotioNet is a "in-the-wild" dataset with a large amount of subject variation and is used to study the FACS model. However, it lacks the dimensional model of affect, and the emotion categories are not manually labeled, rather, defined automatically based on annotated AUs.

Most of the datasets discussed above lacked the dimensional model of affect and were not in the continuous domain. Some datasets like Belfast [21], RELOCA [30], Aff-Wild Database [31] and AffectNet [32] are some examples of these datasets. These datasets are limited in number since the annotation of continuous dimensions is more expensive and requires trained annotators.

The *RELOCA* [30] dataset consists of 46 participants that participated in a video conference that required collaboration for completing a task. Multi-modal data such as audio, video, ECG and EDA were collected and six annotators measured arousal and valence. The dataset was one of the first attempts to model dimensional model of affect with multiple cues and modalities. However, it had only 46 participants and the videos were captured in a lab controlled experiment.

The *Aff-Wild* dataset [31] is the largest dataset that measures continuous affect in the valence-arousal space "in-the-wild". The dataset consists of 500 videos taken from YouTube displaying emotions while watching videos, performing activities and reacting to jokes and three annotators gave frame-by-frame annotations for valence and arousal. While the dataset modeled the temporal variance of affect, it is limited because of 500 subjects.

The *AffectNet* dataset [32] is currently the largest database of categorical and dimensional models of affect in the wild. The dataset has 1,000,000 images queried from the web using emotion-related keywords with words corresponding to gender, age, and ethnicity. The dataset has labels for the six basic expressions and it contains a "None" type that is used for expressions like bored, tired, confused, focused etc. The annotations are focused on the correctness of the affect type rather than intensity. The images are also annotated for valence and arousal in the continuous domain and the dataset contains 450,000 subjects.

The Belfast Natural Induced Emotion Database [21] provides examples of mild to moderately strong natural emotions in response to a series of laboratory-induced tasks. Each example is of varied length (between 5 seconds to 60 seconds in length) and the examples are labeled by information on self-report of emotion, the gender of the subject, and valence in the continuous domain. The dataset contains a total 1400 videos split into three sets of 570, 650 and 180 videos respectively. Of this, the first set contains labels for frustration, disgust, surprise, fear, and amusement. A task was designed to induce frustration/irritation/annoyance and the first 30 seconds of the subject's emotions were recorded. While natural emotions are portrayed in the dataset, an artificial setting of a laboratory is used where illumination, head-poses are controlled.

We can see that no dataset addresses the four affective states of boredom, confusion, engagement and frustration which are needed for applications such as e-learning, e-advertising etc. While the Belfast [21] dataset does address the affective state of frustration, it is not "in-the-wild" and is stimuli-driven. Also, while AffectNet [32] does try to address focus, boredom, confusion, it combines them into a single "None" class which is also not intensity annotated, not allowing for practical applications in the above-mentioned systems.

A comparison of DAiSEE with several state-of-the-art datasets in the domain of facial expression/affective state recognition is presented in Table 1.

TABLE 1: Summary and Characteristics of Several Datasets in Affect Recognition

Database	Database Information	# of Subjects	Condition	Affect Modelling
FERG DB [33]	- 55,767 images - Frontal view	6	- Posed	- 7 emotion categories
Oulu-CASIA-NIR-VIS [34]	- 2,880 videos - Frontal view	80	- Posed	- 7 emotion categories
AR Face Database [35]	- 4,004 images - Frontal view	154	- Posed	- 7 emotion categories
JAFFE Database [36]	- 213 images	10	- Posed	- 7 emotion categories
CK+ [7]	- 593 images - Frontal & 30 degree images	123	- Controlled - Posed	- 30 AUs - 7 emotion categories
MultiPie [18]	- 750,000 images - Multipl viewpoints & illuminations	337	- Controlled - Posed	- 7 emotion categories
MMI [17]	- 2900 videos - Frontal and side views	25	- Controlled - Posed - Spontaneous	- 31 AUs - 6 Basic Expressions
DISFA [19]	- 130,000 video frames - Frontal views	27	- Controlled - Spontaneous	- 12 AUs
SALDB [37] [38]	- 30,000 video frames - Frontal views	4	- Controlled - Spontaneous	- Valence - Quantized - Continuous
RELOCA [30]	- Multi-modal - Audio, Video, ECG and EDA	46	- Controlled - Spontaneous	- Valence and arousal (continuous) - Self-Assesment
AM-FED [20]	- 242 facial videos	242	Spontaneous	- 14 AUs
DEAP [39]	- 40 one-minute videos - EEG signals recorded	32	- Controlled - Spontaneous	- Valence and arousal (continuous) - Self-Assesment
GFT [40]	- 172,800 video frames	96	- Controlled - Spontaneous	- 20 AUs - Facial Landmarks - Head Pose
B4PD [41]	- 368,036 video frames	41	- Controlled - Spontaneous	- AUs - 2D/3D Facial Landmarks
B4PD+ [42]	- 1,400,000 video frames	140	- Controlled - Spontaneous	- 34 AUs - Head Pose - 2D/3D/IR Facial Landmarks
4DFAB [43]	- 1,800,000 3D meshes	180	- Posed - Spontaneous	- 6 Basic Emotions - Facial Landmarks
AFEW [22]	- 1832 videos	330	- Wild	- 7 emotion categories
FER-2013 [24]	- 35,887 images - Queried from the web	35,887	- Wild	- 7 emotion categories
EmotioNet [26]	- 100,00 images annotated manually - 900,000 images annotated automatically - Queried from the web	100,000	- Wild	- 12 AUs - 23 emotion categories based on AUs
Aff-Wild [31]	- 500 YouTube videos	500	- Wild	- Valence and arousal (continuous)
FER-Wild [25]	- 24,000 images - Queried from the web	24,000	- Wild	- 7 emotion categories
AffectNet [32]	- 1,000,000 images with facial landmarks - 450,000 images annotated manually - Queried from the web	450,000	- Wild	- 8 emotion categories - Valence and arousal (continuous)
Belfast Database [21]	- 1,400 videos	256	- Wild	- Disgust, Fear, Anger, Surprise - Frustration
DAiSEE (This Work)	**- 9,068 video sequences**	**112**	**- Wild**	**- Engagement - Boredom - Confusion - Frustration**

2.2 User Engagement

Hernandez et al. [6] were the first to attempt the problem of recognizing user engagement. They modeled the problem of determining engagement of a TV viewer as a binary classification task, using facial geometric features and SVMs for classification. They created a very small custom dataset (which is not publicly available), labeled by a single coder for the presence of engagement. Similarly, Whitehill et al. [5] attempted to automatically understand engagement in learning environments. They developed a custom dataset labeled by a few coders, which again is not publicly available. They experimented with feature extraction methods and classifiers and concluded that Support Vector Machines with Gabor features gave the best accuracy on recognizing user engagement on a scale of four levels (highest accuracy being ≈ 69%). Besides, the dataset was captured under

constrained settings and did not capture the "in-the-wild" needs of real-world user engagement recognition.

2.3 Commerical Software

The relevance of affective state recognition in real-world applications can be gauged by the rising number of commercial applications that attempt to address this challenge. Applications such as Emotient [44], Emovu [45], and Sightcorp [46] provide an estimation of comparable affective states (called attentiveness, for instance, in SightCorp) in their frameworks. All these applications are firstly constrained only to attentiveness/engagement and do not consider other related affective states, such as boredom or confusion (which we seek to capture in this work). More importantly, our studies with these applications showed that their performance on real-world videos is far below

satisfactory, thus highlighting the need for a dataset that captures real-world conditions for further research. Figure 2 shows an example of the performance of Affdex [47] on videos from our dataset and we see that the software shows a user to be attentive even if the user's eyes are closed or the user is looking away from the screen. We note from the image that the software tracks facial key points and correlates them with emotional and cognitive states. Other applications such as SightCorp [46] use the eye gaze of the subject as the sole determinant of the engagement level. While these are good starting points, most of them re-purpose existing methods to detect engagement, and the availability of a larger dataset will promote better methods for reliable engagement recognition.

2.4 Why E-Learning?

Subject to the exponential growth of MOOCs over the last few years, e-learning has received significant attention from several research groups. While machine learning methods have been used to personalize educational modules, diversify assessment methods and make personalized recommendations based on learner preferences and browsing patterns [48] [49] [50] [51], limited efforts have been attempted in understanding user engagement in the e-learning environment. Understanding how engaged a student is, is an important task that can increase the learning intake of a student. More importantly, e-learning provides a well-defined context for developing this dataset, as well as studying the impact of effective recognition methods, thereby motivating us to choose this context for this dataset, without any loss of generality. This dataset, however, is not restricted to this domain and can be easily extended to advertising, e-shopping, health-care and autonomous driving to name a few.

We believe DAiSEE fulfills this need in the community. In particular, it provides a large enough dataset to train state-of-the-art deep learning models. It provides a significant jump in the nature of affective states and the number of video frames (2,723,882 video frames from 9068 snippets and a total of 25 hours of video recording) from other datasets in this area, as described further in Section 3.

3 THE DAISEE DATASET

We now present the DAiSEE dataset that contains video recordings of subjects in an e-learning environment, annotated with crowdsourced labels for engagement, frustration, confusion, and boredom. The dataset captures "in the wild" settings typically seen in the real-world, and will be made publicly available, along with the individual annotations of the crowd, to facilitate open research. We will now discuss the data collection, annotation, and vote aggregation strategies used to create the dataset.

3.1 Data Collection

To model real-life settings, we used a full HD web camera (1920x1080, 30 fps, focal length 3.6mm, 78° field of view) mounted on a computer focusing on student users watching videos was used. To simulate the e-learning environment,

a custom application was created that presented a subject with 2 different videos (20 minutes total in length), one educational and one recreational to capture both focused and relaxed settings, which allow natural variations in users engagement levels. To model unconstrained settings, the subjects had the option to scroll through the videos.

There are 112 subjects in the dataset belonging to the age group of 18-30, all of whom are currently enrolled students. The race of the subjects is Asian, with 32 female and 80 male subjects. A total of 12583 video snippets are collected, each 10 seconds long. This duration is chosen as Jacob Whitehill [5] observed that 10-second labeling tasks are more intuitive. After data cleaning, we end up with a dataset that has 9068 video snippets varying across 6 different locations such as dorm rooms, crowded lab spaces, library etc and 3 different illumination settings (light, dark and neutral); with a male-to-female ratio of 2.13:1. Each video snippet is given a unique identification number to help differentiate between the settings of video snippets.

The Hawthorne effect [52] [53] also referred to as the observer effect, is a type of reactivity where individuals modify an aspect of their behavior in response to their awareness of being observed. This is a critical aspect of such a data capture setting and it is highly probable that the subjects may adapt their behavior to suit the objectives of the experiment. To limit the occurrence of such circumstances, the subjects are recorded without being trained or setting any parameters for the experiment.

3.1.1 Subject Privacy

All participants who appear in the video snippets have given signed consent for the recorded videos to be distributed for use by the wider research community. In the event consent is declined, the captured videos are deleted. Further, the anonymity of every subject is ensured by giving him/her a uniquely generated 3-digit id whose correspondence with the identity is not recorded anywhere.

3.2 Data Annotation

Motivated by recent work in intelligent tutoring systems [54], our dataset consists of labels for four affective states related to user engagement, viz., engagement, frustration, confusion, and boredom. Recent work [55] has shown that the six basic expressions: anger, disgust, fear, joy, sadness, and surprise [56] are not reliable in prolonged learning situations, as they are prone to rapid changes. Each of the affective states is defined at four levels: (1) very low (2) low (3) high and (4) very high (similar to [5]). We followed this labeling strategy to avoid the "neutral" state since early experiments showed that crowd annotators often preferred to choose "neutral" as a state when unsure. The possible levels for labels ensured that the annotators "took a stand" on the affective state level, which is essential for a robust dataset. We also note that as in [5], the above 4-level annotation can be trivially changed to 2-levels (*high* and *low*) when required for a given application setting.

Subtle affective states such as user engagement are subjective and vary based on the viewer's discretion. Hence, we rely on "wisdom-of-the-crowd" for our annotations in the dataset. The easy availability of annotators on crowdsourcing platforms has resulted in many new computer

(a) High Engagement (b) Zero Engagement (c) Minimal Engagement
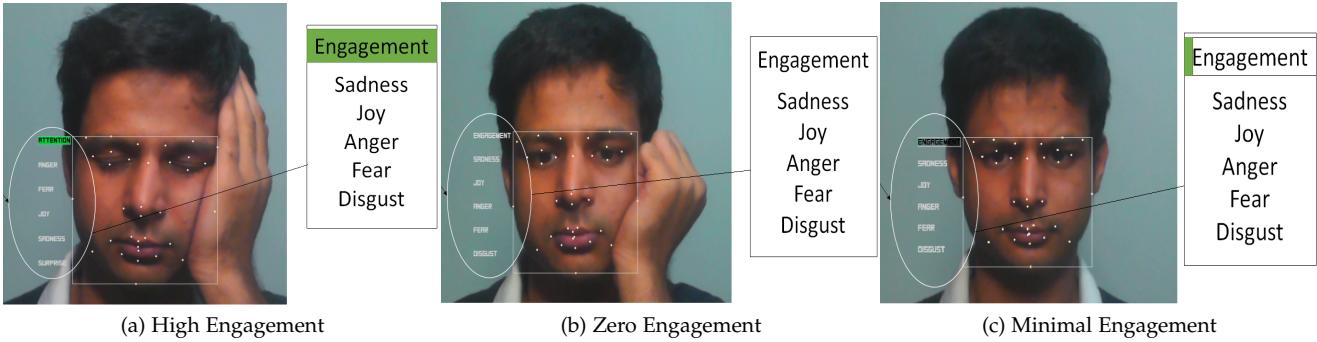
Fig. 2: (Best viewed in color) Results of Affdex on videos from our dataset. The topmost entry on the text-inset in each figure shows the level of engagement. In the image (a), the software detects that the subject is engaged (full green bar), even though the subject is not. In subfigure (b), the software detects zero engagement (zero green bar), even though the subject is engaged. In subfigure (c), the software detects a highly engaged person as having minimal engagement (low green bar). All levels of engagement are determined by psychologists

vision datasets such as [1], [3], [57] which tap into the crowd for annotating large data. Although the annotators can be non-experts, it has been shown that repeated labeling of examples by multiple annotators produces high-quality labels [58], [59], [60]. In this work, we used CrowdFlower, for the annotations, similar to [61]. CrowdFlower provides advanced quality control mechanisms, worker targeting and detailed reports on the final annotation results obtained. Other features of CrowdFlower that we used in this work are mitigation of bot labeling, priming of annotator to the specific task using reasoned test questions, and flagging labels of underperforming annotators.

To obtain the votes, each annotator is presented with a video snippet and asked to vote. Annotators are presented with instructions on how to perform the task and illustrative examples to facilitate the process. Additionally, each annotator answers a standardized test question, that helps us remove the votes of under-performing annotators. For each video snippet, we get votes from 10 different annotators (which is comparable to other standard crowdsourced datasets such as [62] [63]).

3.3 Vote Aggregation

Vote aggregation is used to assign a label for each affective state in a given video snippet, using the annotations from the crowd. We use the Dawid-Skene [64] vote aggregation algorithm to obtain the ground truth label for each snippet, since this is often considered 'gold standard' for aggregation in practice. Dawid-Skene is an unsupervised inference algorithm that gives the Maximum Likelihood Estimate of observer error rates using the EM algorithm. However, before vote aggregation, we removed faulty annotations and noise in order to increase the robustness of votes and reliability of annotators.

3.3.1 Removing Bad Annotators

The annotators on CrowdFlower are not experts and noisy annotations cause the aggregation algorithm to output false labels for a video snippet. DAiSEE has a total of 1690 unique annotators who voted on 12583 video snippets, with each snippet having 10 annotations. To remove noisy labels, we

created a gold standard for a subset of 1157 unique video snippets, which included every annotator. Two teams of experts, each consisting of a social, a clinical and a behavioral psychologist from the department of psychology at our institute were formed. Each group worked on mutually exclusive videos from the subset and consensus from each of the three experts used to create the gold standard. This is used to identify and remove faulty annotators.

To estimate an annotator's reliability, for each set of annotations, we use the gold standard to establish the inter-coder agreement. Since the labels for any video snippet are ordinal (1,2,3,4) in nature, we use a weighted Cohen's κ (score between 0-1) with quadratic weights to penalize label disagreement between the annotator and the gold standard. Any annotator whose agreement coefficient is less than 0.5 is marked as unreliable, their annotations as noisy, and removed from the dataset.

The resulting dataset has video snippets with annotations varying from 0-10 in number, with a median of 4 annotations. All video snippets with less than 4 annotations are removed, resulting in a dataset with 9,068 snippets and 2,723,882 frames.

3.3.2 Dawid-Skene Aggregation

We use the Dawid-Skene [64] vote aggregation strategy to obtain the ground truth label for each snippet. Dawid-Skene is an unsupervised inference algorithm that gives the Maximum Likelihood Estimate of observer error rates using the EM algorithm.

1) Using the labels given by multiple annotators, estimate the most likely "correct" label for each video snippet.
2) Based on the estimated correct answer for each object, compute the error rates for each annotator.
3) Taking into consideration the error rates for each annotator, recompute the most likely "correct" label for each object.
4) Repeat steps 2 and 3 until one of the termination criteria is met (error rates are below a pre-specified threshold or a pre-specified number of iterations are completed).

The distribution of labels after removing bad annotators is depicted in Table 2

TABLE 2: Distribution of labels in DAiSEE across its affective states

Affective State	Very Low	Low	High	Very High
Boredom	3869	2931	1934	334
Confusion	6024	2191	752	101
Engagement	61	459	4477	4071
Frustration	6986	1649	346	87

3.4 How good is DAiSEE?

Examples of different levels of each affective state is depicted in Figures 3, 4, 5 and 6.

Fig. 3: From left to right we can see engagement vary from very low to very high

Fig. 4: From left to right we can see boredom vary from very low to very high

Fig. 5: From left to right we can see confusion vary from very low to very high

The "in the wild" settings of DAiSEE can be seen in Figure 7 where we see how different illumination settings change across images and how different people express very high engagement through different poses.

DAiSEE contains high-fidelity annotations for all videos. To establish this, we compared the labels obtained from Dawid-Skene after removal of bad annotators and the labels from experts on the gold standard and observe that for 70% of the videos, the labels match with the gold standard. To further study the correlation of DAiSEE's final aggregated labels with layman users, we conducted a user study with 10 randomly chosen subjects on 100 videos and asked each user to agree or disagree with the label associated with a video snippet. The users were not given any further information about the nature or background of this work. Our results of this study showed that on 84 of the 100 videos, the majority aggregation of the users' opinions agreed with the

Fig. 6: From left to right we can see frustration vary from very low to very high

Fig. 7: Variety in DAiSEE : All subjects have Very High Engagement but express it through various poses. Illumination changes are visible from light to dark in the above images as we move from left to right

aggregated labels from the crowd. Among the remaining 16 videos, based on subjective feedback gathered, we found that the disagreement was about the level of intensity of the label (low vs very low engagement), rather than of labels with opposite polarities (very high vs very low engagement) in 10 videos, and of the affective state with opposite polarity in 6 videos.

The unique features of DAiSEE are summarized below:

- It is the first publicly available dataset for studying user engagement and related affective states (in the wild).
- While there has been substantial work in recognition of the seven basic emotions, DAiSEE presents a dataset to understand more subtle affective states, such as engagement and boredom, which are often not exhibited explicitly on a human. We believe that this will help advance the field in a new direction, with potential applications in several fields (as described earlier).
- It contains videos, thus allowing researchers to use the temporal information for effective recognition.
- DAiSEE will be publicly available and will include both raw crowd annotations, as well as the high-fidelity aggregated annotations obtained through the process described above, for all video snippets. The crowd annotations could be used to develop better vote aggregation methods.

DAiSEE presents a platform for advancing applications related to e-learning, healthcare, advertising and autonomous vehicles which benefit from recognition of user engagement and associated affective states. We now present the benchmark results we obtained on DAiSEE using many state-of-the-art video classification methods, as described in Section 4.

4 BENCHMARK RESULTS

Majority of the videos in DAiSEE are captured with the frame of capture extending up to the bust of the user. This

allows interested researchers to study the relevance of non-facial cues such as upper body postures on user engagement (for example, a laidback posture could indicate higher levels of boredom, while an upright posture could indicate high levels of engagement). We use a number of state-of-the-art deep learning models in order to assess how well they are working on DAiSEE. We run experiments using two types of models; *Static* (Frame classification/prediction) and *Dynamic* (temporal video classification). In order to be model-agnostic while experimenting, minimal hyper-parameter tuning is done. All experimental results are based on vanilla implementations of models. We now discuss the models, their performance on DAiSEE and the final benchmarking numbers.

4.1 Dataset Split

To prepare the dataset for benchmarking, we create a data-split into train, validation and test sets. The following three principles are used:

- For training deep learning models, follow the general Kaggle [65] practice of 60:20:20 for the train:validation:test sets
- All splits are mutually exclusive and exhaustive with respect to subjects
- The same ratio of males:females is maintained across the splits

4.2 Models

We experiment with two types of models, temporal and static. All models are based on Convolutional Neural Networks (CNNs) [66] since they have shown ground-breaking results in the area of computer vision in the past years. Since we used vanilla implementations with minimal hyper-parameter tuning, we used bi-linear interpolation to re-shape the image and fed it to the model. Each affective state is benchmarked individually using the given models:

- **Single Frame Classification**: We did frame level classification and accuracy measurement. We used a InceptionNet V3 model, pre-trained using ImageNet [67]; restricting the output softmax layer to 4. We fine-tuned the top dense layers and then trained the top two inception blocks. The images were reshaped to (229x229x3) and the data splits shuffled to not introduce sequencing before being used as input to the model.
- **Single Frame Prediction, Video Classification**: We did frame level predictions and video level accuracy measurements. No shuffling is done on the input, and all the frames for a given snippet are fed to the model, their outputs aggregated for a video snippet and measured for accuracy. We used the same model as single frame classification. This is done to observe if temporal models are needed or pooled results from static networks would be sufficient.
- **Video Classification, Full training**: We use C3D [68] which is a modified version of BVLC Caffe [69] to train a 3D CNN. The model is fully trained using our data splits, an output softmax layer for four classes and accuracy measured against the label for a video.
- **Video Classification, Transfer learning**: We use the pre-trained C3D model on Sports-1M [66], and fine-tuned it

for our dataset by modifying the output softmax layer for four classes and measuring accuracy against the label for a video.
- **Video Classification, Sequence learning**: We used a Long-Term Recurrent Convolutional Network [70] for unifying visual and sequence learning. We did two modifications: 1) As subjects are mostly static, optical flow does not provide any useful information and hence it is not computed; 2) We recreated the video by considering every alternate frame as continual affective states such as engagement do not vary in less than 30 ms. After modifying the output softmax layer for four classes, we did a full training for our data splits and measured accuracy against the label for a video.

4.3 Performance Metrics

All the classification models were tested on Top-1 accuracy. We ran each model described in the previous section 3 times to avoid any randomness bias (due to factors such as weight initialization in the models). We present our results and analysis of the dataset below.

4.4 Results

Figure 8 and Table 3 show the baseline accuracy results from our studies. We see that LRCN generally performs better than all other models. Additionally, the experiments also show that temporal classifiers tend to outperform the static classifiers illustrating that affective states such as engagement, frustration, boredom, and confusion are prolonged in nature and cannot be estimated by looking at a particular instant in time. Experiment on InceptionNet with frame level prediction/video classification showed us that the accuracy is only slightly better than random guessing for boredom, again illustrating that static models are insufficient for analyzing such affective states.
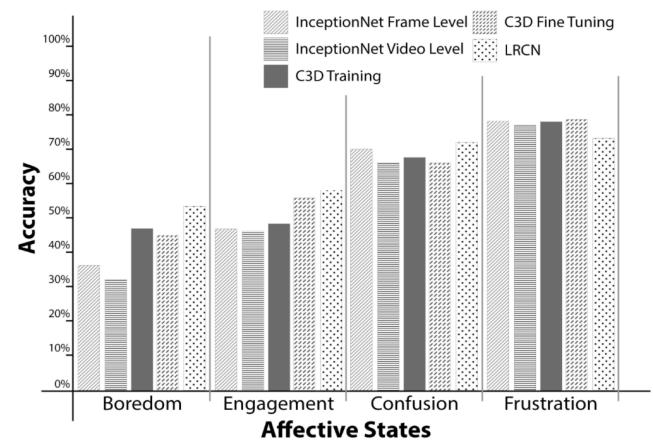

Fig. 8: Results of various experiments on DAiSEE. We report top-1 accuracy with five-fold cross-validation

4.5 Analysis

Effect of removing bad annotators: We did a study to see how many labels improved (came closer to the true level) for

TABLE 3: Benchmark results on DAiSEE. We report Top-1 Accuracy averaged over three runs

Affective State	InceptionNet Frame Level	InceptionNet Video Level	C3D Training	C3D FineTuning	LRCN
Boredom	36.5%	32.3%	47.2%	45.2%	53.7%
Engagement	47.1%	46.4%	48.6%	56.1%	57.9%
Confusion	70.3%	66.3%	67.9%	66.3%	72.3%
Frustration	78.3%	77.3%	78.3%	79.1%	73.5%

the gold standard for each affective state after the removal of bad annotators. We followed a two-step approach to i) run Dawid-Skene without removal of bad annotators and ii) run Dawid-Skene with the removal of bad annotators and then compare them against the annotations by the experts (psychologists) to test for any improvement in label quality. The results are summarized in Table 4.

TABLE 4: Improvement in label quality after removal of bad annotators for different affective states of DAiSEE in the gold standard

Affective State	% of Modified Labels
Engagement	5.78
Boredom	2.46
Confusion	10.82
Frustration	8.12

Using EmotionNet to analyze DAiSEE: To further analyze DAiSEE, we use EmotionNet [26], a CNN used for emotion recognition in photographs of human faces as a benchmarking model. We used the implementation of EmotionNet given by authors [1]. EmotionNet is pre-trained using CK+ [7] and Karolinska Directed Emotional Faces (KDEF) [71] datasets. Before passing the video as input to the model, we used FaceNet [72] to detect and extract faces. In the event that a face was not detected, we used LabelImg [73] to manually extract the face. We, then, fine-tune the model on DAiSEE by sampling every fourth frame from the video (Our experiments show that this does not affect accuracy). The results are summarized in Table 5

TABLE 5: Benchmarking EmotionNet on DAiSEE

Affective State	Accuracy
Engagement	51.07%
Boredom	35.89%
Confusion	57.45%
Frustration	73.09%

Using pre-trained EmotionNet model on CK+: We use the EmotionNet model trained for engagement and pass various images from the CK+ [7] database. For each image, we obtain a label corresponding to the level of engagement the network predicts. Some sample outputs are shown in Figure 9. The model does not output "very-low" engagement for any image. We believe that the reason for this is because the subjects in the CK+ database are well-aware of their environment (the experiments are controlled) and hence are engaged during the course of the experiment.

Considering each label in a binary classification setting: We also considered the problem of binary classification for engagement, by trivially changing the labels from (low, very low) categories to 'not-engaged' and from (high, very

1. https://github.com/co60ca/EmotionNet2

(a) Low Engagement   (b) High Engagement   (c) Very-High Engagement

Fig. 9: Samples of pre-trained EmotionNet on CK+. In the first image, we see "low" level of engagement. In the second image, we see "high" level of engagement and in the third image we see "very-high" level of engagement

high) categories to 'engaged'. This study was done by using multiple subsets of the data (identical in size), with the same ratio of engaged:not-engaged labels as DAiSEE and averaging the results of these multiple subsets. Using LRCN, we achieved a Top-1 accuracy of 94.6%. This shows the challenges when trying to learn the subjective levels for engagement compared to learning whether the person is engaged or not and provides directions for further research in improving techniques to recognize multiple levels for any affective state.

5 BENCHMARKING CHALLENGES OF DAISEE

While DAiSEE shows significant numbers for accuracy, the "in-the-wild" settings of DAiSEE provide some significant challenges in benchmarking the dataset, resulting in possible misclassification of videos. From our analysis, such misclassified videos are shown to have at least one of the following characteristics:

Low Illumination: One of the most common problems was that of poor illumination. In videos with low illumination, the faces are not clearly visible, making it difficult to capture the affective state of the individual, as shown in Figure 10.

Face Occlusion/Lack of Frontal Pose: Another cause for misclassification is when there is occlusion, for example, the hair of a female user covering her face or a user using his/her hands to cover parts of the face. Also, since DAiSEE is captured in the wild, distractions such as opening doors or people walking by caused the subject to turn around, hiding frontal facial features, thus resulting in misclassification. Some samples can be seen in Figure 11.

Variety of Affective States Displayed: There are several video snippets where the subject displays a change in affective states expressed as the snippet progresses, making it difficult for the model to learn essential features to classify
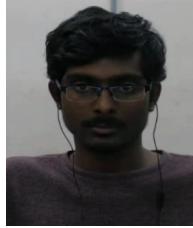
Fig. 10: This figure depicts an instance of low illumination in DAiSEE. The low illumination makes it difficult for the faces to be detected and hence the model cannot learn appropriate features required for correct classification

Fig. 11: Examples displaying the lack of frontal face alignment or facial occlusion

the video correctly. An example of such a video can be seen in Figure 12.

Complementary Labels: While working with the dataset, we noticed how boredom and engagement complimented each other. When engagement is low, boredom is generally high and vice-versa. In the instances that both boredom and engagement were low for a video snippet, the subject displayed high levels of confusion or frustration. This shows the subtleness but disparate nature of different affective states and how well DAiSEE captures these changes. This is another direction for future research where the correlation between labels is used to develop robust models that identify these affective states simultaneously. We can see in Figure 13 that boredom and engagement are complementary to each other. On the other hand, the two affective states are not complementary in Figure 14.

6 DAISEE RELEASE

DAiSEE is available for download on https://iith.ac.in/˜daisee-dataset.

7 CONCLUSION AND FUTURE WORK

In this work, we present DAiSEE, a dataset for user engagement in the wild. The novelty of DAiSEE comes from

Fig. 12: Example of the range of affective states observed in a single 10 second video snippet

(a) Highly Bored   (b) Highly Engaged

Fig. 13: From the figure, we can see that boredom and engagement are complementary in nature to each other. In figure (a) we observe high boredom, but low engagement whereas in figure (b) we observe high engagement, but low boredom
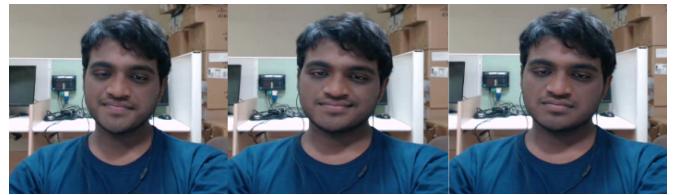

Fig. 14: In the figure we can see that boredom and engagement are not complementary in nature to each other. The labels for boredom and engagement are 0 while that of frustration is 2. This clearly shows that there is no direct correlation between engagement and boredom

the rich information that it has of different affective states such as engagement, boredom, confusion, and frustration. Each affective state is categorized from very low to very high (without neutral) and is annotated using "wisdom-of-the-crowd". The dataset captures the nature of real-world e-learning environments in an organic manner, with varying user poses, positions and background noises typically observed in such settings. It is unique as this is the first publicly available dataset to study these four affective states compared to the seven emotion categories and is the largest available facial emotion/expression dataset for the research community to work on. We also present benchmarking results for DAiSEE and establish a baseline for the research community to build on. To help create a more open community for DAiSEE, we present its raw annotation data for conducting research to improve vote aggregation algorithms or for using these annotations in the training process to improve upon the baselines shared in this work.

Going forward, methods that determine geometric features (such as facial fiducials), facial action units, body and head pose; gaze and gesture can be used as input to models that learn to recognize user engagement. Also, systems that use such mid-level cues, such as pose and gesture, can often be used to develop cognitive models of the subject which include engagement, attentional focus, and intention.

We hope that DAiSEE assists teachers, content creators and students in the domain of e-learning, advertisement makers, medical professionals and autonomous vehicle companies in creating better and more responsive systems to help improve human-computer interaction.

REFERENCES

[1] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on.* IEEE, 2009, pp. 248–255.

[2] M. Everingham, L. V. Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," p. 88(2):303338, 2010.

[3] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *Computer Vision–ECCV 2014.* Springer, 2014, pp. 740–755.

[4] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 3213–3223.

[5] J. Whitehill, Z. Serpell, Y.-C. Lin, A. Foster, and J. R. Movellan, "The faces of engagement: Automatic recognition of student engagementfrom facial expressions," *Affective Computing, IEEE Transactions on*, vol. 5, no. 1, pp. 86–98, 2014.

[6] J. Hernandez, Z. Liu, G. Hulten, D. DeBarr, K. Krum, and Z. Zhang, "Measuring the engagement level of tv viewers," in *Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on.* IEEE, 2013, pp. 1–7.

[7] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression," in *Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on.* IEEE, 2010, pp. 94–101.

[8] B. Schuller, M. Valster, F. Eyben, R. Cowie, and M. Pantic, "Avec 2012: the continuous audio/visual emotion challenge," in *Proceedings of the 14th ACM international conference on Multimodal interaction.* ACM, 2012, pp. 449–456.

[9] J. Rehg, G. Abowd, A. Rozga, M. Romero, M. Clements, S. Sclaroff, I. Essa, O. Ousley, Y. Li, C. Kim, H. Rao, J. Kim, L. L. Presti, J. Zhang, D. Lantsman, J. Bidwell, and Z. Ye, "Decoding children's social behavior," p. 34143421, 2013.

[10] M. Valstar, B. Schuller, K. Smith, T. Almaev, F. Eyben, J. Krajewski, R. Cowie, and M. Pantic, "Avec 2014 3d dimensional affect and depression recognition challenge," p. 310, 2014.

[11] K. Jordan, "Initial trends in enrolment and completion of massive open online courses," *The International Review of Research in Open and Distributed Learning*, vol. 15, no. 1, 2014.

[12] D. Yang, T. Sinha, D. Adamson, and C. P. Rose, "turn on, tune in, drop out: Anticipating student dropouts in massive open online courses," in *Proceedings of the 2013 NIPS Data-driven education workshop.* NIPS, 2013.

[13] D. Colman, "Mooc interrupted: Top 10 reasons our readers didnt finish a massive open online course. 5 april 2013," *Dan Colman: Open Culture*, 2013.

[14] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 31, no. 1, pp. 39–58, 2009.

[15] J. Tao and T. Tan, "Affective computing: A review," in *Affective computing and intelligent interaction.* Springer, 2005, pp. 981–995.

[16] Y.-l. Tian, T. Kanade, and J. F. Cohn, "Recognizing action units for facial expression analysis," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 23, no. 2, pp. 97–115, 2001.

[17] M. Pantic, M. Valstar, R. Rademaker, and L. Maat, "Web-based database for facial expression analysis," in *Multimedia and Expo, 2005. ICME 2005. IEEE International Conference on.* IEEE, 2005, pp. 5–pp.

[18] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker, "Multipie," *Image and Vision Computing*, vol. 28, no. 5, pp. 807–813, 2010.

[19] S. M. Mavadati, M. H. Mahoor, K. Bartlett, P. Trinh, and J. F. Cohn, "Disfa: A spontaneous facial action intensity database," *IEEE Transactions on Affective Computing*, vol. 4, no. 2, pp. 151–160, 2013.

[20] D. McDuff, R. El Kaliouby, T. Senechal, M. Amr, J. F. Cohn, and R. Picard, "Affectiva-mit facial expression dataset (am-fed): Naturalistic and spontaneous facial expressions collected" in-the-wild"," in *Computer Vision and Pattern Recognition Workshops (CVPRW), 2013 IEEE Conference on.* IEEE, 2013, pp. 881–888.

[21] I. Sneddon, M. McRorie, G. McKeown, and J. Hanratty, "The belfast induced natural emotion database," *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 32–41, 2012.

[22] A. Dhall, R. Goecke, J. Joshi, M. Wagner, and T. Gedeon, "Emotion recognition in the wild challenge 2013," in *Proceedings of the 15th ACM on International conference on multimodal interaction.* ACM, 2013, pp. 509–516.

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

[24] I. J. Goodfellow, D. Erhan, P. L. Carrier, A. Courville, M. Mirza, B. Hamner, W. Cukierski, Y. Tang, D. Thaler, D.-H. Lee *et al.*, "Challenges in representation learning: A report on three machine learning contests," in *International Conference on Neural Information Processing.* Springer, 2013, pp. 117–124.

[25] A. Mollahosseini, B. Hassani, M. J. Salvador, H. Abdollahi, D. Chan, and M. H. Mahoor, "Facial expression recognition from world wild web," in *Computer Vision and Pattern Recognition Workshops (CVPRW), 2016 IEEE Conference on.* IEEE, 2016, pp. 1509–1516.

[26] C. F. Benitez-Quiroz, R. Srinivasan, A. M. Martinez *et al.*, "Emotionet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild." in *CVPR*, 2016, pp. 5562–5570.

[27] G. A. Miller, "Wordnet: a lexical database for english," *Communications of the ACM*, vol. 38, no. 11, pp. 39–41, 1995.

[28] P. Viola and M. J. Jones, "Robust real-time face detection," *International journal of computer vision*, vol. 57, no. 2, pp. 137–154, 2004.

[29] S. Du, Y. Tao, and A. M. Martinez, "Compound facial expressions of emotion," *Proceedings of the National Academy of Sciences*, vol. 111, no. 15, pp. E1454–E1462, 2014.

[30] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne, "Introducing the recola multimodal corpus of remote collaborative and affective interactions," in *Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on.* IEEE, 2013, pp. 1–8.

[31] S. Zafeiriou, A. Papaioannou, I. Kotsia, M. Nicolaou, and G. Zhao, "Facial affect"in-the-wild"," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2016, pp. 36–47.

[32] A. Mollahosseini, B. Hasani, and M. H. Mahoor, "Affectnet: A database for facial expression, valence, and arousal computing in the wild," *arXiv preprint arXiv:1708.03985*, 2017.

[33] D. Aneja, A. Colburn, G. Faigin, L. Shapiro, and B. Mones, "Modeling stylized character expressions via deep learning," in *Asian Conference on Computer Vision.* Springer, 2016, pp. 136–153.

[34] G. Zhao, X. Huang, M. Taini, S. Z. Li, and M. PietikäInen, "Facial expression recognition from near-infrared videos," *Image and Vision Computing*, vol. 29, no. 9, pp. 607–619, 2011.

[35] A. M. Martinez, "The ar face database," *CVC Technical Report24*, 1998.

[36] M. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba, "Coding facial expressions with gabor wavelets," in *Automatic Face and Gesture Recognition, 1998. Proceedings. Third IEEE International Conference on.* IEEE, 1998, pp. 200–205.

[37] M. A. Nicolaou, H. Gunes, and M. Pantic, "Audio-visual classification and fusion of spontaneous affective data in likelihood space," in *Pattern Recognition (ICPR), 2010 20th International Conference on.* IEEE, 2010, pp. 3695–3699.

[38] ——, "Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space," *IEEE Transactions on Affective Computing*, vol. 2, no. 2, pp. 92–105, 2011.

[39] S. Koelstra, C. Mühl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras, "Deap: A database for emotion analysis; using physiological signals," *Affective Computing, IEEE Transactions on*, vol. 3, no. 1, pp. 18–31, 2012.

[40] J. M. Girard, W.-S. Chu, L. A. Jeni, and J. F. Cohn, "Sayette group formation task (gft) spontaneous facial expression database," in *Automatic Face & Gesture Recognition (FG 2017), 2017 12th IEEE International Conference on.* IEEE, 2017, pp. 581–588.

[41] X. Zhang, L. Yin, J. F. Cohn, S. Canavan, M. Reale, A. Horowitz, P. Liu, and J. M. Girard, "Bp4d-spontaneous: a high-resolution spontaneous 3d dynamic facial expression database," *Image and Vision Computing*, vol. 32, no. 10, pp. 692–706, 2014.

[42] Z. Zhang, J. M. Girard, Y. Wu, X. Zhang, P. Liu, U. Ciftci, S. Canavan, M. Reale, A. Horowitz, H. Yang *et al.*, "Multimodal spontaneous emotion corpus for human behavior analysis," in

Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 3438–3446.

[43] S. Cheng, I. Kotsia, M. Pantic, and S. Zafeiriou, "4dfab: A large scale 4d facial expression database for biometric applications," *arXiv preprint arXiv:1712.01443*, 2017.

[44] Emotient, http://www.emotient.com/, [Last accessed: 11-November-2017].

[45] Emovu, http://www.enovu.com/e/, [Last accessed: 11-November-2017].

[46] SightCorp, "http://www.sigthcorp.com," 2016. [Online]. Available: http://www.sightcorp.com

[47] Affdex, http://www.affectiva.com/solutions/affdex/, [Online; accessed 12-November-2017].

[48] F. Castro, A. Vellido, À. Nebot, and F. Mugica, "Applying data mining techniques to e-learning problems," in *Evolution of teaching and learning paradigms in intelligent environment*. Springer, 2007, pp. 183–221.

[49] M.-J. Huang, H.-S. Huang, and M.-Y. Chen, "Constructing a personalized e-learning system based on genetic algorithm and case-based reasoning approach," *Expert Systems with Applications*, vol. 33, no. 3, pp. 551–564, 2007.

[50] A. Baylari and G. A. Montazer, "Design a personalized e-learning system based on item response theory and artificial neural network approach," *Expert Systems with Applications*, vol. 36, no. 4, pp. 8013–8021, 2009.

[51] M. Pivec, C. Trummer, and J. Pripfl, "Eye-tracking adaptable e-learning and content authoring support," *Informatica*, vol. 30, no. 1, 2006.

[52] R. McCarney, J. Warner, S. Iliffe, R. Van Haselen, M. Griffin, and P. Fisher, "The hawthorne effect: a randomised, controlled trial," *BMC medical research methodology*, vol. 7, no. 1, p. 1, 2007.

[53] T. Monahan and J. A. Fisher, "Benefits of observer effects: lessons from the field," *Qualitative Research*, vol. 10, no. 3, pp. 357–376, 2010.

[54] R. Rajendran, "Enriching the student model in an intelligent tutoring system," Ph.D. dissertation, The IITB-Monash Research Academy, 2014.

[55] R. A. Calvo and S. D'Mello, "Affect detection: An interdisciplinary review of models, methods, and their applications," *Affective Computing, IEEE Transactions on*, vol. 1, no. 1, pp. 18–37, 2010.

[56] P. Eckman, "Universal and cultural differences in facial expression of emotion," in *Nebraska symposium on motivation*, vol. 19. University of Nebraska Press Lincoln, 1972, pp. 207–284.

[57] G. Van Horn, S. Branson, R. Farrell, S. Haber, J. Barry, P. Ipeirotis, P. Perona, and S. Belongie, "Building a bird recognition app and large scale dataset with citizen scientists: The fine print in fine-grained dataset collection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 595–604.

[58] V. S. Sheng, F. Provost, and P. G. Ipeirotis, "Get another label? improving data quality and data mining using multiple, noisy labelers," in *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2008, pp. 614–622.

[59] P. Welinder and P. Perona, "Online crowdsourcing: rating annotators and obtaining cost-effective labels," 2010.

[60] P. G. Ipeirotis, F. Provost, V. S. Sheng, and J. Wang, "Repeated labeling using multiple noisy labelers," *Data Mining and Knowledge Discovery*, vol. 28, no. 2, pp. 402–441, 2014.

[61] A. Burke, "Crowdsourcing scientific progress: how crowdflower's hordes help harvard researchers study tb," *Forbes. October*, vol. 16, 2011.

[62] D. Zhou, S. Basu, Y. Mao, and J. C. Platt, "Learning from the wisdom of crowds by minimax entropy," in *Advances in Neural Information Processing Systems*, 2012, pp. 2195–2203.

[63] H. Han, C. Otto, X. Liu, and A. K. Jain, "Demographic estimation from face images: Human vs. machine performance," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 37, no. 6, pp. 1148–1161, 2015.

[64] A. P. Dawid and A. M. Skene, "Maximum likelihood estimation of observer error-rates using the em algorithm," *Applied statistics*, pp. 20–28, 1979.

[65] Kaggle.com, https://www.kaggle.com/, [Last accessed: 11-November-2017].

[66] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks." CVPR, 2014.

[67] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Computer Vision and Pattern Recognition, 2016. CVPR 2016*. IEEE, 2016.

[68] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks." ICCV, 2015.

[69] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," arXiv, 2014.

[70] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. 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.

[71] E. Goeleven, R. De Raedt, L. Leyman, and B. Verschuere, "The karolinska directed emotional faces: a validation study," *Cognition and emotion*, vol. 22, no. 6, pp. 1094–1118, 2008.

[72] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 815–823.

[73] Tzutalin, https://github.com/tzutalin/labelImg, [Last accessed: 11-March-2018].