

# DAISEE: Dataset for Affective States in E-Learning Environments

Abhay Gupta<sup>1</sup>, Richik Jaiswal<sup>2</sup>, Sagar Adhikari<sup>2</sup>, Vineeth Balasubramanian<sup>2</sup>

<sup>1</sup> Microsoft India R&D Pvt. Ltd.

abhgup@microsoft.com

<sup>2</sup> Department of Computer Science, IIT Hyderabad

{cs12b1032, cs12b1034, vineethnb}@iiith.ac.in

**Abstract.** Extracting and understanding affective states of subjects through analysis of face videos is of high consequence to advance the levels of interaction in human-computer interfaces. This paper aims to highlight vision-related tasks focused on understanding “reactions” of subjects to presented content which has not been largely studied by the vision community in comparison to other emotions. To facilitate future study in this field, we present an effort in collecting DAiSEE, a free to use large-scale dataset using crowd annotation, that not only simulates a real world setting for e-learning environments, but also captures the interpretability issues of such affective states by human annotators. In addition to the dataset, we present benchmark results based on standard baseline methods and vote aggregation strategies, thus providing a springboard for further research.

## 1 Introduction

Inter-personal human communication includes not only spoken languages but also non-verbal cues such as hand gestures and facial expressions which are used to express feelings and give feedback. Affective states such as engagement, frustration, confusion, and boredom are very important not only to express our emotions but also to provide important suggestions during social interactions such as level of interest, our desire to take a speaking turn or to provide continuous feedback on the understanding of the information conveyed.

E-learning environments are one of the best examples for studying user affective states. With the accelerated growth of Massive Open Online Courses (MOOCs), there is a need to design more intelligent interfaces to simulate the interactions that occur between a teacher and students in a class. The main drawback of existing e-learning environments is that they do not provide real-time interactive feedback to students unless they use some related discussion forums or peer-engaged learning. Currently, MOOCs have a completion rate of 7-9% [24], with the completion rate for the first assignment being around 45%. An online survey [8] lists the top ten reasons for dropouts from such platforms; poor course design was one of the reasons, which included components such as lack



**Fig. 1:** Some examples of images from DAiSEE. The dataset captures real-world elements associated with e-learning environments.

of proper student 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 and make such platforms more interactive. Understanding affective states in these environments can help design more intuitive interfaces that further knowledge absorption by students and help decrease dropout rates, as well as personalize the learning experience.

This paper seeks to address the aforementioned issues, by working towards a system that can automatically recognize student affective states such as engagement, frustration, confusion and boredom frustration in e-learning environments. The proposed work can also be relevant to other application domains such as advertising, gaming and entertainment, where these affective states are important. Existing commercially available affective recognition systems have limited use in real-world environments (illustrated further in Section 2), necessitating further work on this problem. A major constraint, however, is that there is no image/video dataset available for affective states in e-learning environments, both in terms of the affective state as well as in terms of the typical real-world

environments used for e-learning. In this work, we develop a large vision dataset, DAISEE (**D**ataset for **A**ffective **S**tates in **E**-learning **E**nvironments), which will be made publicly available for further research (Figure 1). Considering that affective states such as engagement can be subtle, we crowdsource annotations for this dataset and test different vote aggregation methods on this dataset. We have attempted to make DAISEE rich in both data and annotations so as to facilitate further research in: (i) affective state recognition for real-world learning environments; and/or (ii) use of crowdsourced labels for classification problems with class labels that are not clearly defined. We further benchmark the performance of standard feature extractors and classifiers on this dataset to provide a baseline for further research.

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 and its salient features. In Section 4, we benchmark a baseline performance on this dataset using standard feature extraction and classification methods. Lastly, we summarize our analysis and suggest future directions with our dataset.

## 2 Related Work

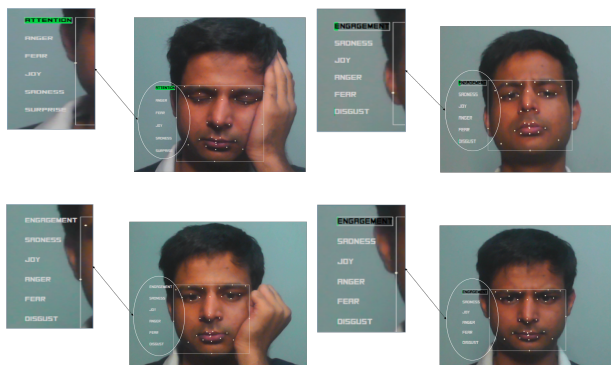
**MOOCs** Subsequent to the exponential growth of MOOCs over the last few years, e-learning has received significant attention from several research groups. Early endeavors concentrated on using machine learning methods to personalize educational modules, diversify assessment methods and make personalized recommendations based on learner preferences and browsing patterns, as in [3, 4, 7, 13, 22, 33]. Additionally, there have been endeavors in developing evolving e-learning frameworks that utilize visual information to give a further level of personalization; for instance, [31] utilizes eye-tracking for personalizing the interaction. Despite significant advancements in recent years, there have been very few efforts on understanding affective states of students in real-world e-learning environments, as discussed in the next section.

**Emotion Recognition and Engagement Detection** Determining the affective state of a user using computer vision and machine learning methods has been studied for over two decades [37, 45]. Until recently, most efforts focused on the six essential expressions (happiness, sadness, anger, disgust, fear, surprise) and the facial action units connected with them, as in [38, 44]. Despite recent efforts that attempt the recognition of subtle affective states [2], as well as model affective states in terms of dimensions such as valence and arousal [17, 20], very little work has been done in perceiving abstract affective states such as those relevant for e-learning (or advertising) settings, as described below.

Hernandez et al. [21] modeled the problem of determining engagement of a TV viewer as a binary classification task, using multiple geometric features extracted from the face, and SVMs for classification. Considering the lack of a publicly available dataset, a custom dataset - very small and labeled by a

single coder - was used in this work. On comparable lines as our work, Whitehill et al. [43] attempted to automatically understand engagement in learning environments. Once again, they used a custom dataset based on a few coders, which however is not available for further research. Besides, the dataset was captured under constrained settings and does not capture the nuances of real-world e-learning environments, thus motivating our work in this paper.

**Commercial Software** The relevance of affective state recognition to real-world applications can be gauged by the rising number of commercial applications that attempt to address this challenge. Applications such as Emotient [15], Emovu [16], and Sightcorp [36] provide an estimation of comparable affective states (called attentiveness, for instance, in SightCorp) in their frameworks. On one hand, all these applications are constrained only to attentiveness/engagement and do not consider other affective states relevant for e-learning. On the other, our studies with these applications show 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 [1] 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. Applications such as [36] use eye gaze of the subject as the sole determinant of the engagement level. Such correlations between eye gaze or facial keypoints and attention may not always hold, especially in e-learning environments.



**Fig. 2:** Results of Affdex on videos from our dataset. The first entry on the text measures engagement. Top figure shows the user to be attentive (fully engaged); Middle figure reports zero engagement; and Bottom figure reports the user as minimally engaged.

**Datasets** Several datasets have been created to advance affective state recognition in recent years. The most popular datasets include DEAP [26], CK+ [28], AVEC [34] and Emotion Recognition in the Wild Challenge [11, 12]. However, none of these datasets focus on the subtle affective states that exist in e-learning environments (such as engagement, frustration, confusion and boredom), necessitating this work. To the best of our knowledge, this paper is the first systematic effort towards creating such a dataset and corresponding benchmark results.

Earlier related work that attempted a similar problem [21, 43], described earlier in this section, used custom datasets for this purpose. However, both these datasets are not accessible to the community for further research. Further, in the case of [21], depending on a single coder can lead to personal bias and impact generalizability. While [43] used a few trained coders for annotating their images, in this work, we have used crowdsourced labels to provide a commoner’s view to understanding engagement (and similar affective states), as opposed to a trained coder’s understanding which may have learned biases. Another disadvantage of both these datasets is that they disregard the Hawthorne effect (subject awareness of experiment objectives, described further in Section 3) in the creation of the dataset. Further, in both the efforts, the datasets were created with reasonably controlled settings, and do not capture the real-world issues of MOOC environments.

### 3 The DAiSEE Dataset

We now present the DAiSEE dataset that: (i) will be made public for further research; (ii) captures real-world settings in e-learning environments across all subjects; and (iii) provides labels for engagement, frustration, confusion and boredom levels that are crowdsourced. We first discuss the data collection procedure, followed by data annotation process, vote aggregation strategies, and finally describe the salient properties of the dataset.

#### 3.1 Data Collection

**Data Capture** We use a full HD web camera (1920x1080, 30 fps, focal length 3.6mm, 78° field of view) mounted on a computer focusing on subjects watching e-learning videos. 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 portray different learning environments. To model unconstrained settings, the subjects had the option to scroll through the videos. There are 95 subjects in the dataset belonging to the age group of 18-30, all of whom are currently enrolled students. In total, 30 hours of recordings were captured. The videos were captured in 5 different location settings: (1) lab setting with high background clutter; (2) dorm room; (3) lab setting with minimal background; (4) white background; and (5) miscellaneous locations (random locations where the subject was found). Each of these settings has different illumination conditions, which were then categorized manually as low or high.

**Data Pre-Processing** All recorded videos are divided into 10-second snippets (similar to [43]) and then individually processed for face detection using the standard Viola-Jones face detector [41]. The snippets in which faces are detected across all frames are retained. The resulting dataset has 7338 video snippets, each video snippet being 10 seconds long (300 frames) across 95 subjects, 5 different locations, and 2 different illumination conditions.

**Hawthorne Effect, Subject Privacy, and Anonymity** The Hawthorne effect [19], also referred to as the observer effect, is a type of reactivity in which individuals modify or improve 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 diminish the effects of such a circumstance, the subjects were recorded without their knowledge. This helped in limiting the Hawthorne effect. To account for the privacy interests of every subject, at the end of the data capture, they were informed of the recordings and their consent obtained to carry out research work. In the event that a subject declined consent, the captured videos were 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

As mentioned earlier, DAiSEE is created by tapping into the potential of crowd annotators. Crowd annotation brings in mass intelligence to interpret affective states that often can be subtle and prone to individual bias. Over the last few years, newer computer vision datasets [10, 27, 39] are increasingly relying on wisdom-of-the-crowd for annotations, due to the large amounts of data and easy availability of annotators on crowdsourcing platforms. Although the annotators can be non-experts, it has been shown that repeated labeling of examples by multiple annotators can produce high-quality labels [23, 35, 42]. Popular crowdsourcing platforms include Amazon’s Mechanical Turk (AMT), CrowdFlower, LiveOps, InnoCentive, and Samasource<sup>1</sup>. These frameworks also provide interfaces for fast and reliable crowd annotation. In this work, we used CrowdFlower for the annotations (reasons for our choice explained later in this section), similar to [5].

**Class Labels** Our dataset consists of labelings of four emotional states, viz., engagement, frustration, confusion and boredom. Recent work [6] has shown that the six essential expressions: anger, disgust, fear, joy, sadness and surprise [14] are not reliable in prolonged learning situations, such as classrooms and e-learning

---

<sup>1</sup> <http://www.mturk.com>; <http://www.crowdfunder.com>; <https://www.liveops.com>; <http://www.innocentive.com>; <http://www.samsource.com>

environments. Our choice of affective states such as engagement, frustration, confusion and boredom for this work is also supported by recent work in intelligent tutoring systems [32].

For each of the affective states, we provide four labels: (1) very low (2) low (3) high and (4) very high. This was obtained by conducting empirical studies with labels of other levels, including 5 and 3. In case of 5-scale, our labels included strongly positive, positive, neutral, negative and strongly negative; and the 3-scale study included positive, neutral and negative. Including neutral in the scales gave equivocal results as the annotators vote neutral when there is ambiguity. Hence, we chose a 4-scale/2-scale labeling strategy. A 4-scale is chosen over the 2-scale to increase the richness of the information of the labels offered and to also help learn the subtle differences in affective states of users in learning environments that extend beyond the dataset. (Besides, a 4-scale can easily be converted into a 2-scale result if required).

**Annotation Process** For DAiSEE, we have four different affective states, each having four classes, as described above. 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 underperforming annotators. For each video snippet, we get votes from 10 different annotators, similar to [18, 48].

**Why CrowdFlower?** The advantages of using CrowdFlower over other platforms is that it provides quality control mechanisms, advanced 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. Also, CrowdFlower provides worldwide access to the platform unlike certain other platforms like AMT (which is restricted to the United States).

### 3.3 Vote Aggregation Algorithms

After obtaining the votes for all video snippets, we use four different algorithms for aggregating annotator labels to obtain the single ground truth label for each video. The algorithms are Majority Voting [29], Dawid-Skene [9] and two variants of label aggregation using Maximal Conditional Entropy [47], namely, Categorical and Ordinal. The final ground truth label distributions after applying each of the aggregation algorithms for the four affective states is seen in Figure 3. We now briefly describe each of the vote aggregation methods.

**Majority Voting** The algorithm [29] determines the majority of votes cast in the annotations. In the event of a tie, one of the tied labels is randomly selected as the majority vote.



**Dawid-Skene** This is an unsupervised inference algorithm [9] that gives the maximum likelihood estimation of observer error rates using the EM algorithm. The algorithm has the following main steps:

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).

**Multiclass Minimax Conditional Entropy** This algorithm [47] extends the Dawid-Skene algorithm by creating a two-dimensional confusion matrix for item difficulty based on annotator-error estimates and label errors. An overview of the algorithm is as below:

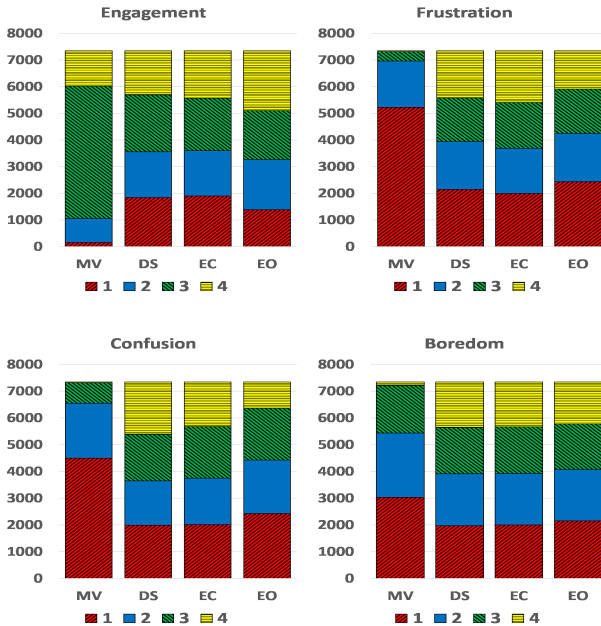
1. Using the labels given by multiple annotators, estimate the most likely “correct” label for each object.
2. Based on the estimated correct answer for each subject, compute the confusion matrices for both the annotator and item errors.
3. Taking into consideration the confusion matrices for both annotators and items, recompute the most likely “correct” label for each subject.
4. Repeat steps 2 and 3, until termination criterion is met (as before).

**Ordinal Minimax Conditional Entropy** This algorithm [47] is an extension of the multiclass minimax conditional entropy algorithm. It introduces a mechanism to map the ordinality of every label to the multiclass problem proposed above. The algorithm suggests comparing two ordinal labels by comparing these labels with respect to a reference label which varies for all possible values in a given set of ordinal labels.

### 3.4 Salient Features of DAiSEE

Every video snippet in DAiSEE (sample frames shown in Figure 1) is labeled with four attributes, viz. engagement, frustration, confusion and boredom. These attributes provide rich information about the learning experience of the subject, and can allow personalization of content as well as feedback for course, experience and environment design. The dataset takes into account the nature of different subjects across e-learning platforms at different locations. User affective states are captured at diverse locations which have significant background noise and clutter with people often walking around in the background and sometimes interacting with the subject. The distribution of videos across the different locations can be seen in Figure 4(a). An important aspect of typical e-learning environments is lighting conditions for different subjects, and we have captured



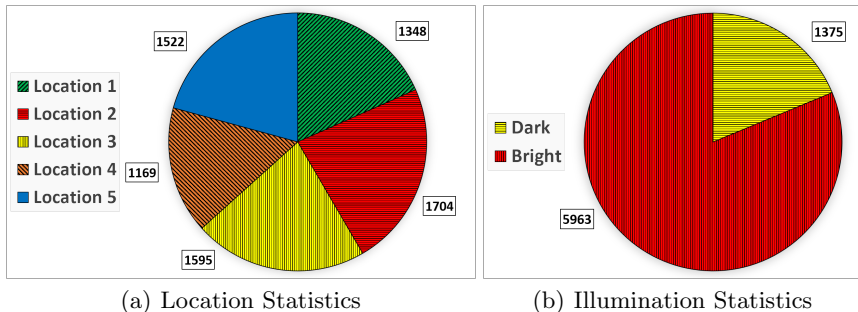


**Fig. 3:** Label distribution after applying the four aggregation methods for all affective states across all 7338 videos. Here MV represents Majority Voting, DS-Dawid-Skene, EC-Entropy Categorical and EO-Entropy Ordinal

videos both in bright and dark settings to reproduce this artifact in the dataset. The distribution of videos across low and high illumination conditions can be seen in Figure 4(b). All these features demonstrate the resemblance of DAISEE to videos captured in real-world e-learning environments. This dataset will be available free to use for the research community. Improvements in classification (or detection) accuracy by vision researchers will have a straightforward meaningful impact on designing more intelligent interfaces for e-learning (and other domains such as advertising and gaming).

## 4 Benchmark Results

As a dataset, DAISEE can be studied from perspectives of face/body/gaze detection, pose estimation, affective state recognition or vote aggregation in vision-based classification problems. Considering the original focus of this work, we benchmark results for understanding user affective states in e-learning environments. For this task, we use standard feature extraction methods and classifiers to obtain baseline results.



**Fig. 4:** *Left:* Distribution of videos across 5 location settings (described in Section 3.1). *Right:* Distribution across illumination settings: low and high.

## 4.1 Feature Extraction

We first perform face detection and alignment using the publicly available FaceAlign<sup>2</sup> tool. Once the faces have been aligned, we crop the faces in the frames of the videos. We note that future research on the dataset could use the entire human body (or even interactions in the scene) to discern subject attributes such as engagement (for e.g, if a subject is looking away to interact with a passerby, he/she is likely to be disengaged).

**3D HOG Descriptors:** We extract dense 3D HOG [25] from the cropped video and then process the features for a bag-of-words type representation. We apply  $k$ -means clustering on the descriptors with  $k = 256$  clusters [40]. This results in a 256-dimensional frequency histogram of facial features. We then normalize the histogram of the features.

**3D LBP Descriptors:** We extract the LBP-TOP [46] features for every video. The features for all the orientations are concatenated for the feature representation of a video.

**Deep Face Descriptors:** We use VGG-Face [30], a Convolutional Neural Network-based feature extraction method, for these descriptors. The architecture of VGG Face is shown in Figure 5. We extract the features from the  $fc7$  layer to get the most generic features as the  $fc8$  representations are tailored for the dataset proposed in [30]. The extracted features are processed for a bag-of-words representation. We apply  $k$ -means clustering on the descriptors with  $k = 50$  clusters. This results in a 50-dimensional frequency histogram. We then normalize the histogram of the features.

## 4.2 Classifiers

After all the features have been obtained, we use three classifiers, namely  $k$ -NN, SVM and Random Forests for classification. For SVM, we ran preliminary

<sup>2</sup> FaceAlign: <https://github.com/roblourens/facealign>



**Fig. 5:** Architecture of VGG Face

results with four different kernels: RBF, Linear, Poly-3 and Poly-4 and found that RBF kernels performed the best among them. For  $k$ -NN we varied  $k$  from 1 to 100 and found that at  $k = 49$  we obtained the best mean performance for the different feature extraction strategies. For random forests, we used a forest size of 150 with minimum split of 2 where we empirically observed the highest average accuracy.

### 4.3 Performance Metrics

5-fold cross-validation is used to measure the generalization capabilities of the baseline methods. Figure 3 shows that under certain vote aggregation methods such as majority voting, the final dataset is imbalanced (data from some labels are far higher in volume than others). To ensure a fair comparison, we use the average classwise accuracy as the performance metric in this work (accuracy is measured for each class individually, and their average is presented), where the average is computed across the folds of cross-validation.

### 4.4 Benchmark Results

DAiSEE is rich in various parameters, having 3 feature extraction strategies, 3 classifiers and 4 vote aggregation strategies for each of the 4 affective states. To help distinguish each of the affective states, we plot the results individually for all the affective states in Figure 6. Some example results are shown in Figure 9. Figure 7 provides all the results for the engagement recognition problem in a single illustration, to showcase the impact of feature extraction methods, classification methods, vote aggregation methods, as well as location/illumination settings for the data capture. These results are discussed further in the next section. Similar results for other affective states are presented in the supplementary material, owing to space constraints.

### 4.5 Analysis and Discussion

We analyze the performance of our benchmark results on DaiSEE, including the dataset's behavior to changes in illumination and location settings, below.

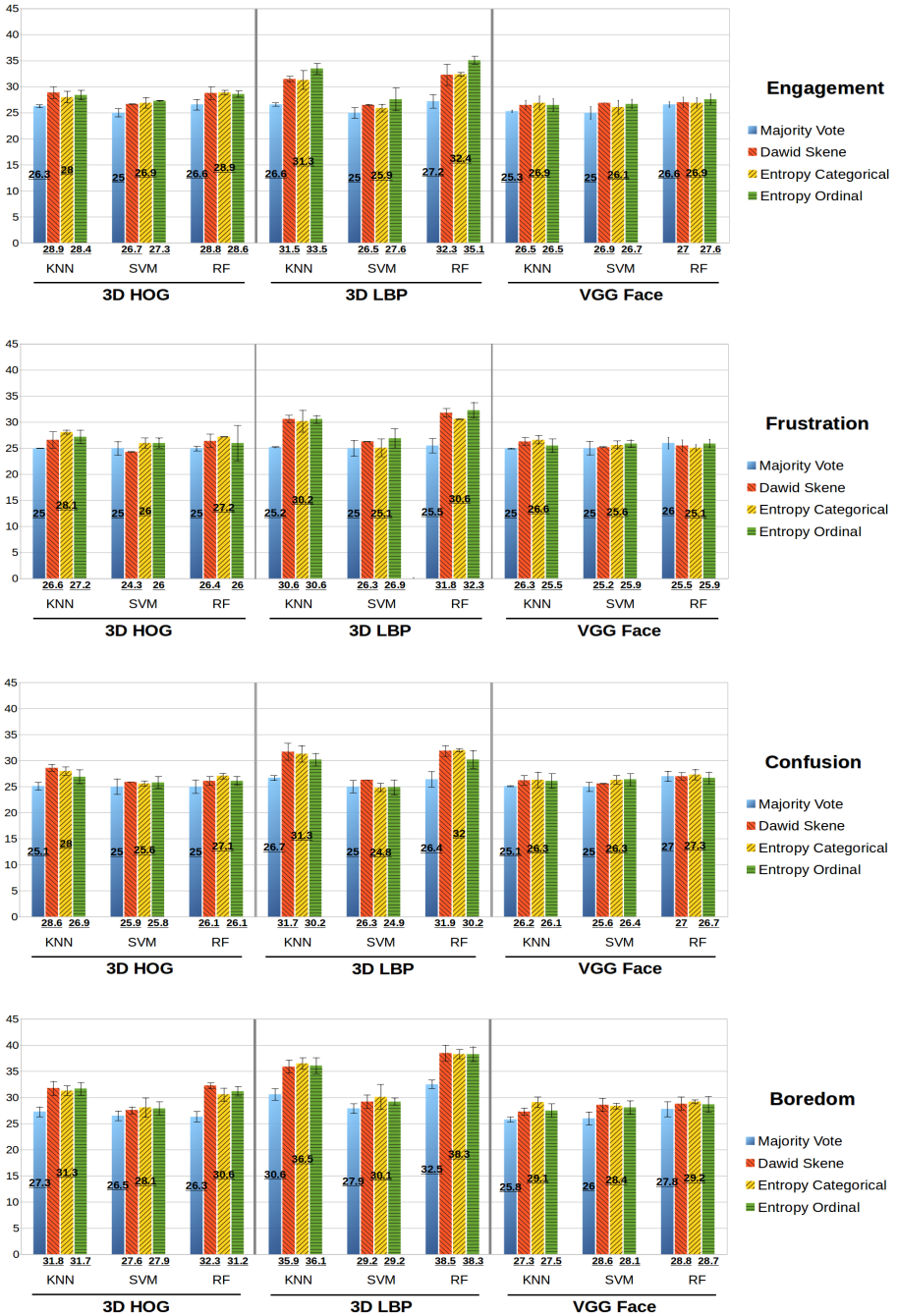
The average classwise accuracy results, seen in Figure 6, are, in general, consistent with the findings in [47] (Entropy-based vote aggregation performs better

than Dawid-Skene, which performs better than Majority Voting). On further inspection, it is observed that Dawid-Skene and Ordinal Entropy outperform Categorical Entropy, in general (barring a few exceptions). This implies that the natural ordering in the labels in the dataset (scale of 1 to 4) favors Ordinal Entropy over Categorical Entropy. Also, following the results in [47], we see that when there are very subtle differences among the classes, it is difficult for the annotators to distinguish between them and this introduces more confusion in the labeling of a data for a given category, which results in poorer results for Categorical Entropy as compared to Dawid-Skene or Ordinal Entropy. Thus, in general, it can be said that Dawid-Skene and Entropy Ordinal are good representations of the crowd’s opinion, when labels are ordinal. While we have used standard vote aggregation methods, it is possible that newer vote aggregation methods that incorporate annotator statistics may provide a more consistent ground truth, and we leave this for further research. Towards this, we provide annotator statistics from CrowdFlower along with the DAiSEE dataset.

Figure 7 shows, expectedly, that the performance of the methods under high illumination is better than under low illumination. Among the location settings, we observe that the performance is best in location setting 1, which is a lab setting with high background clutter. On deeper analysis, we found that although this setting had high clutter, it was very well-illuminated. Considering that we only considered video snippets with good face detections, the background issue was mitigated. In general, one may infer from this that good face detection with good illumination provides the best performance for recognizing the kind of affective states this work intends to study.

We observe from Figure 6 that  $k$ -NN and random forests consistently seem to outperform SVMs. A more comprehensive evaluation of different kernels for SVMs can possibly provide better performance, which is a direction for further research. Also, improved strategies of evaluation such as out-of-bag error for random forests can be considered as a future line of work. We also note that in terms of feature extraction methods, 3D LBP outperforms 3D HOG and VGG-Face, although marginally. We believe that this is because we use a bag-of-words representation for 3D HOG and VGG-Face, which may not necessarily capture the characteristics of the video. Learning features in an unsupervised manner using deep learning architectures that are tailored to this specific problem is an important direction of work that could result in significant improvements, considering the recent successes of deep learning.

In summary, we observe that the classwise accuracy performance is only marginally better than random (considering there are 4 classes), thereby showing the difficulty of working with this dataset. We note that the dataset captures the nature of real-world e-learning environments in an organic manner, with varying user poses, positions and background noise typically encountered in such settings. While we use cropped face yvideos in this work, we will also make available the original videos to encourage research that considers the complete scene too. Methods that use geometric features (such as facial fiducials), facial action units, body pose and eye gaze may need to be explored to improve the



**Fig. 6:** Average classwise accuracy for each affective state for all classifiers,  $k$ -NN, SVM and Random Forests(RF), 3 feature extraction methods and 4 vote aggregation strategies

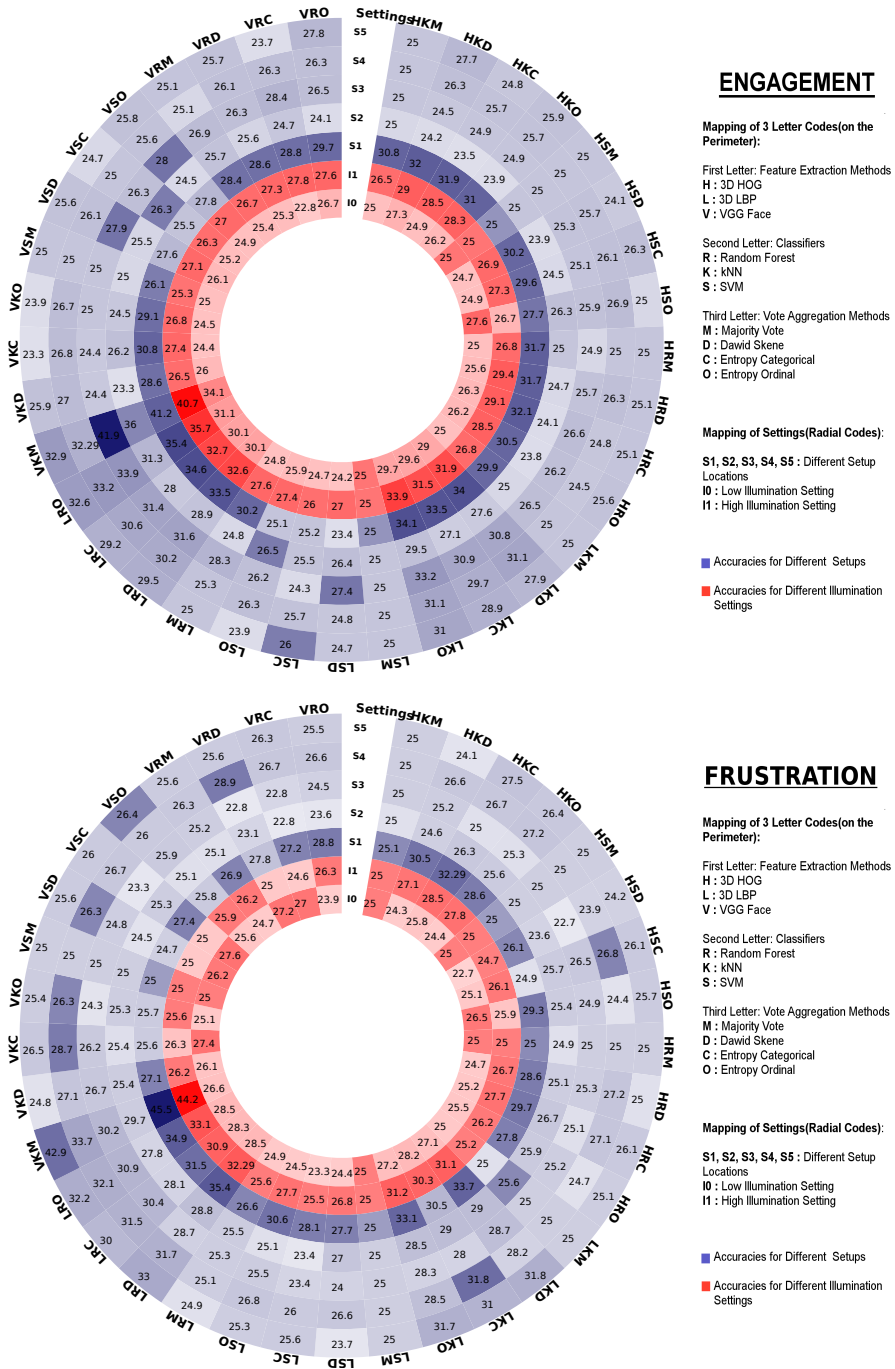
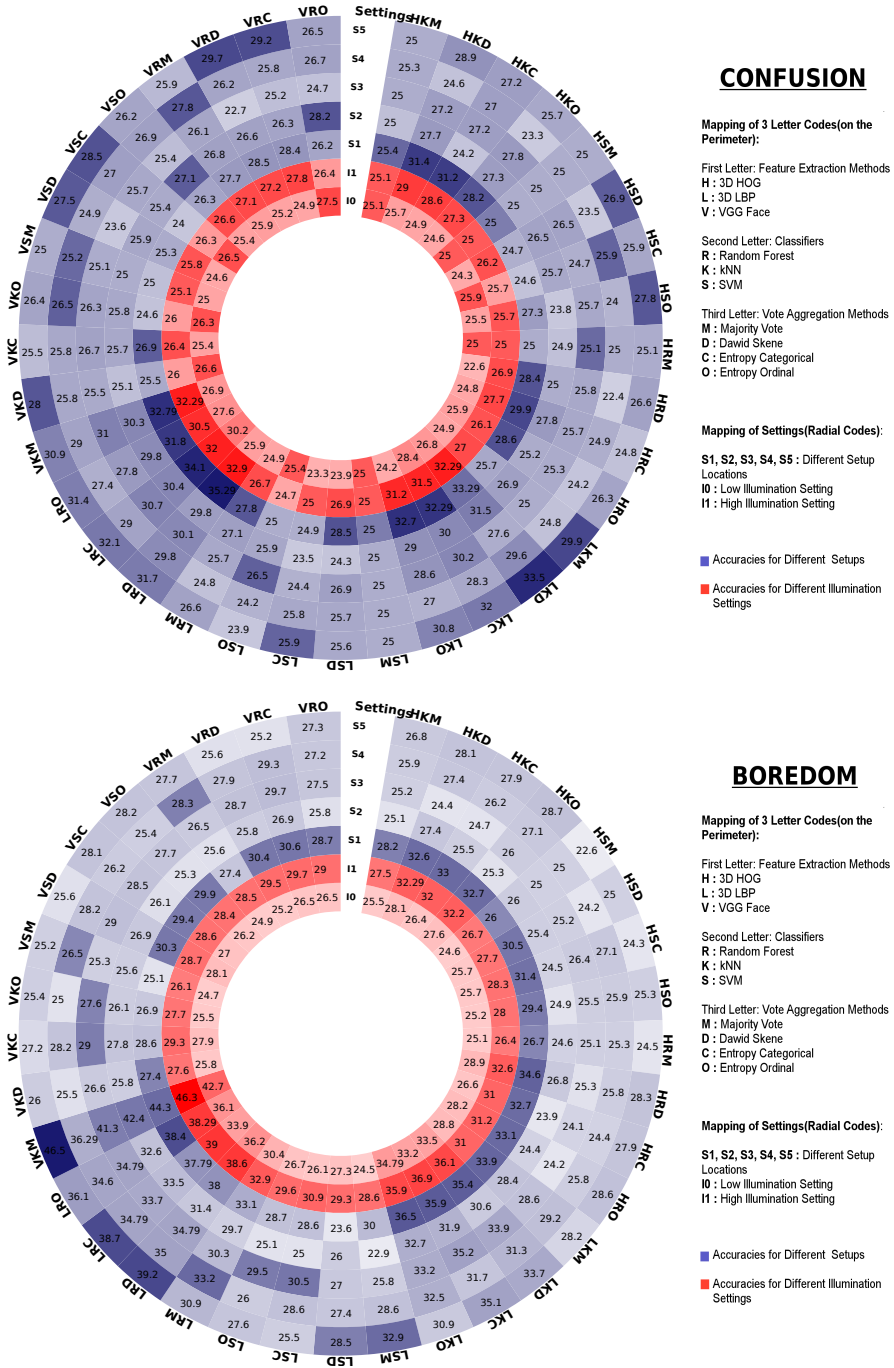



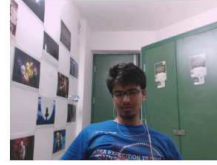


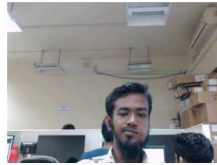


Fig. 7: Average classwise accuracies for engagement and frustration recognition for the different location and illumination settings with respect to different vote aggregation methods, classifiers and feature extraction methods (Best viewed in color. Darker the shade, better the performance.)



**Fig. 8:** Average classwise accuracies for confusion and boredom recognition for the different location and illumination settings with respect to different vote aggregation methods, classifiers and feature extraction methods (Best viewed in color. Darker the shade, better the performance.)



Engagement	Frustration	Confusion	Boredom
			
Ground Truth - 4	Ground Truth - 3	Ground Truth - 2	Ground Truth - 4
Predicted Label - 4	Predicted Label - 3	Predicted Label - 2	Predicted Label - 4

Engagement	Frustration	Confusion	Boredom
			
Ground Truth - 1	Ground Truth - 4	Ground Truth - 4	Ground Truth - 4
Predicted Label - 4	Predicted Label - 1	Predicted Label - 1	Predicted Label - 1

**Fig. 9:** Example predictions from DAiSEE. All the results are generated using 3D-LBP using random forests classifier and ordinal entropy vote aggregation. The top image shows correct classifications and the bottom image shows misclassification results.

performance w.r.t. the baselines shared in this work. It is also possible that our ground truth labels are noisy due to their crowdsourced nature. Considering annotator statistics during vote aggregation can help in improving ground truth labeling itself too.

#### 4.6 Confusion Matrix

In order to better understand the results, we present the confusion matrix for one of the result configurations. Figure 10 provides us with the confusion matrix for engagement with 3DLBP being the feature extraction method, random forests as the classifier and Entropy Ordinal as the vote aggregation strategy. This setting was chosen because it provided the best results. Expectedly, this shows that the *Very High* state (label 4) provides the best results, considering it is easiest to notice.

#### 4.7 Results for Soft Accuracies

We also studied the benchmark results with a modified performance metric based on *soft accuracies*. We compute soft accuracy in the following manner. If the predicted label is 2, but the true label is 1, then we consider this instance to

Confusion matrix		Obtained labels			
		1	2	3	4
Actual labels	1	217	453	306	414
	2	205	688	397	581
	3	123	452	513	756
	4	104	350	435	1344

**Fig. 10:** Confusion matrix for the following configuration: Affective state - Engagement; Feature extraction method - 3DLBP; Classifier - Random forests; Vote aggregation strategy - Entropy Ordinal

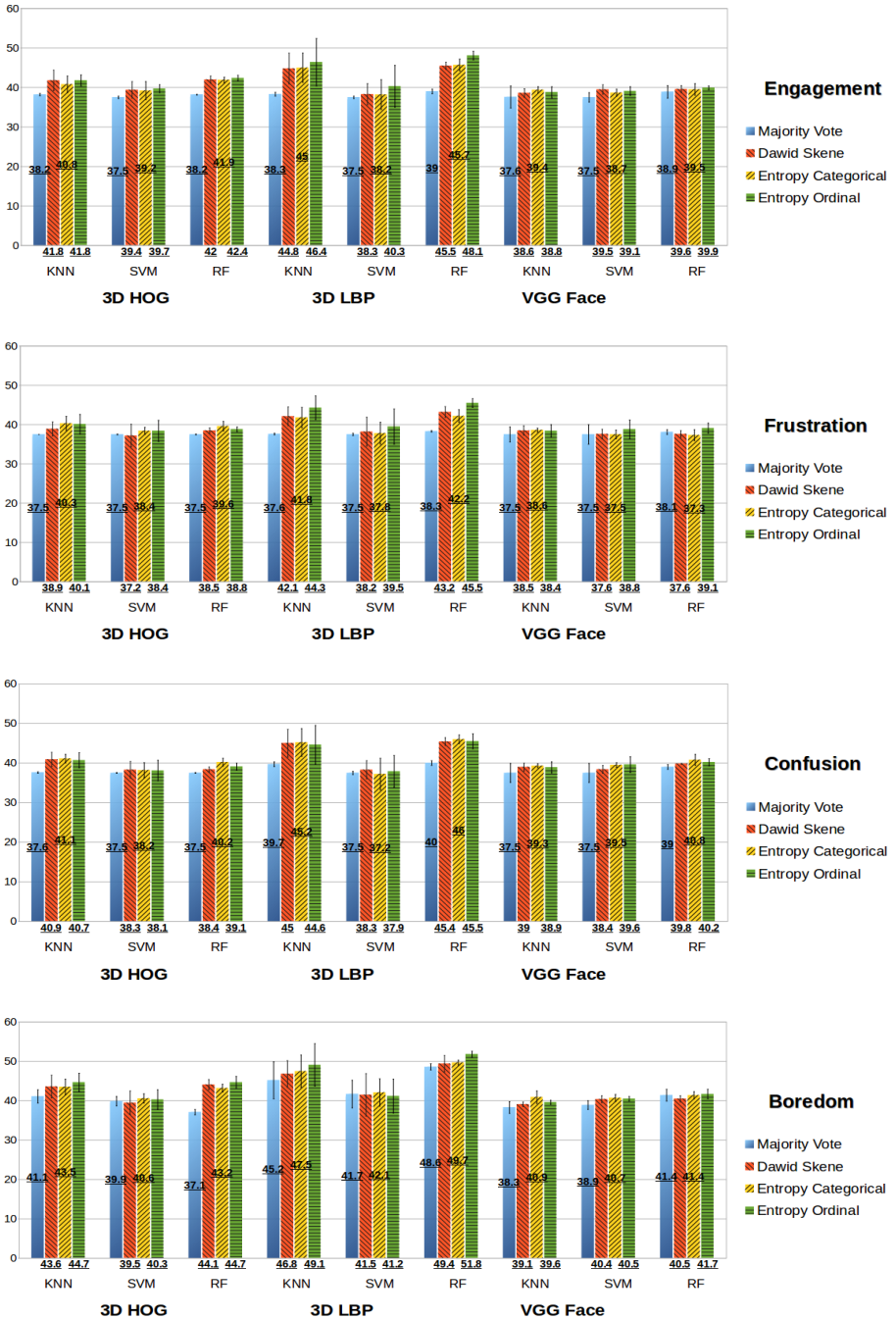
be half-correct while computing accuracy. However, if the true label is 3 or 4, we do not consider this instance to be correctly classified. Similarly, if the predicted label is 4, but the true label is 3, we consider this instance to be half-correct while computing accuracy. This performance metric considers the ordinal nature of the annotations. The soft accuracies for the results, seen in Figure 11 shows the benchmark results based on the average classwise soft accuracy. As expected, the results show improvement under this performance metric. However, the inferences from the results remain unchanged (as discussed in Section 4.5) even under these results.

#### 4.8 Applicability to Real-world Affective State Understanding

To test the usefulness of DAiSEE to real-world affective state recognition, we ran models trained while obtaining the benchmark results on newer data captured using a webcam stream. Figure 12 shows the results. In this experiment, an Intel Xeon E5 with two 8-core 1.2GHz CPUs with 64GB RAM is used and we are able to process up to 100 frames/min in real-time video streams. The feature extraction and model validation are performed using a single core while the face detection and alignment use both the cores. While we do not have ground truth for these frames, the results show promise from a subjective evaluation.

## 5 Conclusions and Future Work

This paper introduces DAiSEE, a crowdsourced dataset focused on modeling affective states in e-learning environments. The proposed dataset has rich information including 4 different affective states, each having 4 different ground truth labels across various locations, illumination conditions and videos presented during the capture. DAiSEE is very useful in gauging the affective states of users in e-learning environments (and potentially other application domains such as advertising, where user engagement is very critical) and further motivates the development of applications that make use of such affective states. We analyze



**Fig. 11:** Average classwise soft accuracy for each affective state for all classifiers,  $k$ -NN, SVM and Random Forests(RF), 3 feature extraction methods and 4 vote aggregation strategies (where soft accuracy is computed as described in Supplementary Section 4.7).



E	3	1	2	4
F	2	2	2	1
C	2	1	2	1
B	2	4	3	1

**Fig. 12:** Labels predicted by using 3D-LBP with Random Forests classifier using Ordinal Entropy Vote Aggregation Method. E=Engagement; F=Frustration; C=Confusion; B=Boredom

various feature extraction methods and classifiers against 4 different vote aggregation methods and learn that Dawid-Skene and Ordinal Entropy are good models of the crowd’s opinion for this context. We provide benchmark results for these feature aggregation methods and classifiers, along with the dataset (and annotator statistics) for further research. Few pointers to improve the performance are also shared in this work.

Advancements in understanding affective states of students in e-learning environments can help boost completion rates in MOOCs by personalizing the learning experience, as well as providing feedback to the instructor for course design. We expect that the proposed dataset will provide a significant boost to focused development of better affective state recognition methods in e-learning (or advertising/gaming) environments. Our immediate future work will include extensions with newer face detectors that may provide better detections and more video snippets in the dataset. In such a case, newer baselines will be shared on the dataset webpage in the near future.

## References

1. Afdex: <http://www.affectiva.com/solutions/afdex/>, [Online; accessed 13-March-2016]
2. Baron-Cohen, S.: Mind reading [: the interactive guide to emotions. Jessica Kingsley Publishers (2003)
3. Baylari, A., Montazer, G.A.: Design a personalized e-learning system based on item response theory and artificial neural network approach. *Expert Systems with Applications* 36(4), 8013–8021 (2009)
4. Brusilovsky, P.: Knowledgetree: A distributed architecture for adaptive e-learning. In: *Proceedings of the 13th international World Wide Web conference on Alternate track papers & posters*. pp. 104–113. ACM (2004)
5. Burke, A.: Crowdsourcing scientific progress: how crowdflower’s hordes help harvard researchers study tb. *Forbes*. October 16 (2011)

6. Calvo, R.A., D'Mello, S.: Affect detection: An interdisciplinary review of models, methods, and their applications. *Affective Computing, IEEE Transactions on* 1(1), 18–37 (2010)
7. Castro, F., Vellido, A., Nebot, À., Mugica, F.: Applying data mining techniques to e-learning problems. In: *Evolution of teaching and learning paradigms in intelligent environment*, pp. 183–221. Springer (2007)
8. Culture, O.: Moocs interrupted, [http://www.openculture.com/2013/04/10\\_reasons\\_you\\_didnt\\_complete\\_a\\_mooc.html](http://www.openculture.com/2013/04/10_reasons_you_didnt_complete_a_mooc.html), [Online; accessed 11-March-2016]
9. Dawid, A.P., Skene, A.M.: Maximum likelihood estimation of observer error-rates using the em algorithm. *Applied statistics* pp. 20–28 (1979)
10. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. pp. 248–255. IEEE (2009)
11. Dhall, A., Goecke, R., Lucey, S., Gedeon, T.: Static facial expression analysis in tough conditions: Data, evaluation protocol and benchmark. In: *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on*. pp. 2106–2112. IEEE (2011)
12. Dhall, A., et al.: Collecting large, richly annotated facial-expression databases from movies (2012)
13. Dolog, P., Henze, N., Nejdli, W., Sintek, M.: Personalization in distributed e-learning environments. In: *Proceedings of the 13th international World Wide Web conference on Alternate track papers & posters*. pp. 170–179. ACM (2004)
14. Eckman, P.: Universal and cultural differences in facial expression of emotion. In: *Nebraska symposium on motivation*. vol. 19, pp. 207–284. University of Nebraska Press Lincoln (1972)
15. Emotient: <http://www.emotient.com/>, [Online; accessed 13-March-2016]
16. Emovu: <http://www.emovu.com/e/>, [Online; accessed 13-March-2016]
17. Gunes, H., Schuller, B., Pantic, M., Cowie, R.: Emotion representation, analysis and synthesis in continuous space: A survey. In: *Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*. pp. 827–834. IEEE (2011)
18. Han, H., Otto, C., Liu, X., Jain, A.K.: Demographic estimation from face images: Human vs. machine performance. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 37(6), 1148–1161 (2015)
19. Hawthorne Effect, W.: [https://en.wikipedia.org/wiki/hawthorne\\_effect](https://en.wikipedia.org/wiki/hawthorne_effect) (2016), [https://en.wikipedia.org/wiki/Hawthorne\\_effect](https://en.wikipedia.org/wiki/Hawthorne_effect)
20. He, L., Jiang, D., Yang, L., Pei, E., Wu, P., Sahli, H.: Multimodal affective dimension prediction using deep bidirectional long short-term memory recurrent neural networks. In: *Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge*. pp. 73–80. ACM (2015)
21. Hernandez, J., Liu, Z., Hulten, G., DeBarr, D., Krum, K., Zhang, Z.: Measuring the engagement level of tv viewers. In: *Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on*. pp. 1–7. IEEE (2013)
22. Huang, M.J., Huang, H.S., Chen, M.Y.: Constructing a personalized e-learning system based on genetic algorithm and case-based reasoning approach. *Expert Systems with Applications* 33(3), 551–564 (2007)
23. Ipeirotis, P.G., Provost, F., Sheng, V.S., Wang, J.: Repeated labeling using multiple noisy labelers. *Data Mining and Knowledge Discovery* 28(2), 402–441 (2014)

24. Jordan, K.: Initial trends in enrolment and completion of massive open online courses. *The International Review of Research in Open and Distributed Learning* 15(1) (2014)
25. Klaser, A., Marszałek, M., Schmid, C.: A spatio-temporal descriptor based on 3d-gradients. In: *BMVC 2008-19th British Machine Vision Conference*. pp. 275–1. British Machine Vision Association (2008)
26. Koelstra, S., Mühl, C., Soleymani, M., Lee, J.S., Yazdani, A., Ebrahimi, T., Pun, T., Nijholt, A., Patras, I.: Deap: A database for emotion analysis; using physiological signals. *Affective Computing, IEEE Transactions on* 3(1), 18–31 (2012)
27. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: *Computer Vision–ECCV 2014*, pp. 740–755. Springer (2014)
28. Lucey, P., Cohn, J.F., Kanade, T., Saragih, J., Ambadar, Z., Matthews, I.: 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*. pp. 94–101. IEEE (2010)
29. Moore, J.S.: A fast majority vote algorithm. In: *Automated Reasoning: Essays in Honor of Woody Bledsoe* (1981)
30. Parkhi, O.M., Vedaldi, A., Zisserman, A.: Deep face recognition. *Proceedings of the British Machine Vision* 1(3), 6 (2015)
31. Pivec, M., Trummer, C., Pripfl, J.: Eye-tracking adaptable e-learning and content authoring support. *Informatica* 30(1) (2006)
32. Rajendran, R.: *Enriching the Student Model in an Intelligent Tutoring System*. Ph.D. thesis, The IITB-Monash Research Academy (2014)
33. Romero, C., Ventura, S., Press, W.: *Data mining in e-learning*. Wit Press Southampton (2006)
34. Schuller, B., Valster, M., Eyben, F., Cowie, R., Pantic, M.: Avec 2012: the continuous audio/visual emotion challenge. In: *Proceedings of the 14th ACM international conference on Multimodal interaction*. pp. 449–456. ACM (2012)
35. Sheng, V.S., Provost, F., Ipeirotis, P.G.: 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*. pp. 614–622. ACM (2008)
36. SightCorp: <http://www.sightcorp.com/>, [Online; accessed 13-March-2016]
37. Tao, J., Tan, T.: Affective computing: A review. In: *Affective computing and intelligent interaction*, pp. 981–995. Springer (2005)
38. Tian, Y.L., Kanade, T., Cohn, J.F.: Recognizing action units for facial expression analysis. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 23(2), 97–115 (2001)
39. Van Horn, G., Branson, S., Farrell, R., Haber, S., Barry, J., Ipeirotis, P., Perona, P., Belongie, S.: 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*. pp. 595–604 (2015)
40. Vedaldi, A., Fulkerson, B.: Vlfeat: An open and portable library of computer vision algorithms. In: *Proceedings of the 18th ACM international conference on Multimedia*. pp. 1469–1472. ACM (2010)
41. Viola, P., Jones, M.: Rapid object detection using a boosted cascade of simple features. In: *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*. vol. 1, pp. I–511. IEEE (2001)

42. Welinder, P., Perona, P.: Online crowdsourcing: rating annotators and obtaining cost-effective labels (2010)
43. Whitehill, J., Serpell, Z., Lin, Y.C., Foster, A., Movellan, J.R.: The faces of engagement: Automatic recognition of student engagement from facial expressions. *Affective Computing, IEEE Transactions on* 5(1), 86–98 (2014)
44. Yang, P., Liu, Q., Metaxas, D.N.: Boosting coded dynamic features for facial action units and facial expression recognition. In: *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*. pp. 1–6. IEEE (2007)
45. Zeng, Z., Pantic, M., Roisman, G.I., Huang, T.S.: A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 31(1), 39–58 (2009)
46. Zhao, G., Pietikainen, M.: Dynamic texture recognition using local binary patterns with an application to facial expressions. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 29(6), 915–928 (2007)
47. Zhou, D., Liu, Q., Platt, J., Meek, C.: Aggregating ordinal labels from crowds by minimax conditional entropy. In: *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*. pp. 262–270 (2014)
48. Zhou, D., Basu, S., Mao, Y., Platt, J.C.: Learning from the wisdom of crowds by minimax entropy. In: *Advances in Neural Information Processing Systems*. pp. 2195–2203 (2012)