Toward Active and Unobtrusive Engagement Assessment of Distance Learners

Brandon M. Booth*, Asem M. Ali†, Shrikanth S. Narayanan*‡, Ian Bennett§ and Aly A. Farag†
*Department of CS, †Department of ECE, ‡Department of EE, §TSN Inc.
*SAIL Lab, University of Southern California
†CVIP Lab, University of Louisville
Email: brandon.m.booth@gmail.com, asem.ali@louisville.edu, shri@sipi.usc.edu, ianmbennettc2y@gmail.com, aly.farag@louisville.edu

Abstract—Student behavior and lecturer oversight in the classroom is known to modulate study behaviors and impact performance and learning outcomes, but cannot at present be managed for distance learning students. Quantifying and automatically measuring student engagement during lectures in a scalable and accessible manner for these students is essential for improving academic success, but has not been studied widely in natural distance learning environments. We collect video recordings from a screen-mounted camera of students studying online lectures in a mostly unstructured setting and gather annotations from a panel of humans for assessing student engagement levels. We present results on the prediction of different representations of engagement, both with subject-independent and individual-specific models, and quantify the performance gap between the generalized and personalized models for engagement prediction. While the subject-independent performance is challenged by data sparsity, results show that the individual-specific models can predict engagement well even with very few labeled examples.

1. Introduction

Research in student engagement assessment is extensive and investigators generally agree that a holistic measurement of student engagement for predicting learning outcomes must encompass not only students’ performance in the class, but also their participation, interest, and development in learning outside the classroom [1]. Prior research suggests that the way students are monitored, behave, and learn in the classroom affects the manner in which they approach studying the content. Zhang [2] proposes that pedagogical practices emphasizing “deep” learning (comprehensive, not surface-level) and achievement (e.g. maximizing grades) leads to the best learning outcomes, both which require a substantial time and attention investment from students. The most successful pedagogical practices therefore should emphasize student engagement both in and out of the classroom.

Distance learning technologies have seen increased usage over the past several years [3] and will likely continue to expand. Measurements of engagement in distance education contexts have mainly focused on behavioral engagement [4], [5] and thus emphasized the use of learning management system interactions such as content views, forum views/posts, or language usage as predictor variables. Although these measurements have proven to help identify students on the path to educational success, they are scarce measures and may not distinguish well between inactive and at-risk students. A more active and observant engagement metric would be beneficial for monitoring student focus during lecture videos while also helping lecturers improve their presentation and flow.

Efforts to automate student engagement estimation are fairly recent. Researchers have explored using camera recordings of students in the classroom and shown that computer vision techniques that estimate head poses for each student can predict engagement with some success [6]. Motion-based features have also been used on classroom videos to assess engagement based on the observation that students are more attentive and more engaged when they are moving less [7]. Some efforts have explored using affect recognition and motion features for predicting engagement for individual students in front of computers [8], [9].

In this paper we focus on engagement as it pertains to students viewing course material via a massive open online course (MOOC) platform. In the classroom, teachers are able to judge the engagement levels of students in real time which provides passive feedback that enables these educators to adjust the structure of lectures and potentially identify at-risk students. This type of feedback is absent in an online setting where students may be multi-tasking and at greater risk of distraction [10].

We conduct a study to capture human behavior during online learning sessions for the purpose of understanding distance student engagement dynamics. We focus exclusively on video recordings of the students taken from a screen-mounted camera because this type of sensor is commoditized, unobtrusive, and likely to already exist on students’ devices. In order to capture behavior closer to what students may exhibit when distance learning in the wild, we allow them full freedom (to eat, drink, take notes, surf the web, etc.) during the learning sessions. Finally, human annotators rate the engagement levels of these students in real-time and on a continuous scale. Many annotation
schemes are possible, but we choose this approach because it affords more subtle variations in rating, it allows the full temporal context to be integrated into rating decisions, and because it is more natural and similar to what lecturers do while teaching.

We address three points in this work. First, we analyze the predictive power of state-of-the-art video features for engagement prediction during mostly unstructured learning sessions. Second, we test several machine learning models on these features with different ground truth representations of the engagement labels to assess the prediction performance end-to-end. Third, we compare results on personalized and cross-subject models quantifying the performance gap for our data set and reveal that good performance can be achieved per-subject with very few labels.

2. Data Acquisition

Data from 12 students was collected in the Experience Lab at the University of Southern California (USC). The lab is a multi-sensor bio- and psycho-physiologic data collection facility designed for capturing human behaviors during interactive computer tasks in a time-synchronous fashion. In this experiment, the lab was used to record front-facing video of students watching lecture videos. A Microsoft Kinect v2 sensor collected RGB frames at 30Hz of the face and upper body of each student seated at a desk in front of the computer screen. This arrangement is similar to typical camera setups in laptops or on personal computers at home. The recruitment and study procedures are described in the following sections.

2.1. Study Protocol

Students at USC enrolled in a graduate-level computer science course were enlisted for the study. Student recruitment occurred weeks prior to the final exam and all students in the class were encouraged to participate in the study by reviewing recorded lectures from the class in the Experience Lab. All interested volunteers from the class were accepted and the subject pool contained a mix of sexes and ethnicities of students in their mid twenties.

During the study, participants were provided a seat at an empty desk and privacy in a sound-resistant room. Drinks and snacks were provided and students were encouraged to eat, drink, and otherwise act naturally. Our aim was to capture behaviors similar to those that students might exhibit when distance learning in the wild. Participants were allowed to select the lecture they were most interested in reviewing and encouraged to watch as much or as little of the video as they desired after a minimum of 20 minutes. Observed behaviors included pausing the lecture, eating/drinking, taking notes, interacting with personal devices, and surfing the web for related information while learning.

3. Annotation Protocol

Ratings of engagement were provided by human annotators. USC undergraduate and graduate students were recruited to provide engagement ratings on a continuous [0, 1] scale in real time as they viewed the front-facing video of participants.

Annotation was performed using a mouse, a standard slider user interface widget, and a viewport showing the video from the front-facing camera. For each trial, annotators were instructed to move the slider in real time according to their perception of the student’s engagement. No clarifications or indications of how to interpret the term engagement were provided. Since many of the annotation tasks required at least an hour of time, annotators were allowed to pause and resume after taking a short break. Figure 1 shows the annotation interface.

Each session received approximately nine annotations. After finishing all annotation tasks, annotators completed an exit survey containing questions about their confidence in annotation accuracy and about any distractions that may have affected their labeling efforts. All annotators reported feeling confident both in their ability to assess engagement and in their actual labeling of it.

4. Engagement Prediction Experiments

We conduct several machine learning experiments to ascertain whether engagement can be predicted using state-of-the-art visual features extracted from the front-facing videos. Since the annotations themselves are produced using only the video, we hypothesize that a consistent mapping exists. We test different machine learning models against various ground truth engagement representations derived
from the annotations to try to find the best end-to-end system for student engagement prediction. Details for the extracted features, various ground truth methods employed, and prediction results are presented in the following subsections.

4.1. Feature Extraction

We extract the following features from students’ faces and upper bodies in each video frame:

- Facial landmarks
- Facial geometric features
- Facial action coding system (FACS) action units
- FACS eye movement codes
- Probability of emotional expressions
- Mean and median average optical flow velocities and average direction vector
- Head size and pose

First we extract a region of interest for the face in each frame using a local binary patterns (LBP) histogram cascade implemented in the OpenCV library [11]. Once the face is located and cropped within each frame, 68 facial landmarks are computed denoting the eyes, nose, mouth and face perimeter using the method described in [12]. Given the cropped face image and facial landmarks in each frame, a smaller region of interest (ROI) of the face is extracted, linearly skewed such that the eyes are horizontally aligned, and then resized to 32x32 pixels. A bank of 40 Gabor filters is applied to the square region producing a set of 40 textures corresponding to facial gradients at various orientations, spatial frequencies and scales.

From the facial landmarks, geometric features are extracted corresponding to pairwise 2D distances between landmarks per the method in [13]. FACS action units [14] are also extracted using an array of pre-trained binary linear support vector classifiers. Each classifier detects a unique action unit and is pre-trained on three FACS data sets from [15], [16], [17]. The average AUC across all 33 FACS classifiers is 89% in a leave-one-out cross validation test.

The eye movement codes are extracted from the Gabor-filtered ROIs corresponding to FACS action units 61-64 and also including centered eye gaze. A small 5x10 pixel region containing the eyes is pulled directly from the square ROIs and fed to a pre-trained RBF-kernel SVM. The SVM model is trained and tested on 1,888 images from the Columbia Gaze Database [18] and the RaDF database [19]. The prediction accuracy using this method is estimated to be 77%.

Emotional expression probabilities are derived using another classifier. The distance of the Gabor ROIs to each FACS action unit classifier’s separating hyperplane is interpreted as a measure of predictive certainty. These certainties are used to train a RBF-kernel SVM to predict one of the eight basic emotional expressions (neutral, anger, contempt, disgust, fear, happiness, sadness, surprise) using four labeled data sets [15], [19], [20], [21] including both elicited and genuine expressions. In total, 3,365 images are used with a leave-one-subject-out cross-validation paradigm resulting in an average accuracy of 80%.

Optical flow is generated using OpenCV’s [11] dense optical flow library based on [22]. Since the subjects in our study are confined to a room, no motion other than the subjects’ occurs. We compute the mean and median flow velocities by averaging over all pixels in the dense flow image. The average flow direction is generated by aggregating the 2D flow directions over all pixels and normalizing the result.

Finally, the head size and pose are extracted using the fR Face Recognition SDK [23]. This toolbox is pre-trained on millions of faces in the wild and reliably extracts head pose and a binary mask for pixels belonging to the head. We estimate the head size as the sum of head pixels and also use the three Euler angle pose estimates provided as features for our machine learning experiments.

Functionals of all of these features within 10-second sliding time windows (overlapping) are used as supplementary features for machine learning. We choose windows of this size based on the results from [8] suggesting that average engagement can be summarized well over 10-second intervals. The standard collection of functionals is used including: mean, variance, quartiles, min, max, and range. Our final set includes 476 features in total, and with the functionals applied, the count increases to 3,808.

4.2. Ground Truth Representation

Several methods for establishing a ground truth set of engagement labels are tested in order to help find the best end-to-end prediction scheme. Ultimately we employ a novel approach to obtain the best ground truth. All tested ground truth labels are listed in the Labels column in Table 1. We first fuse the separate annotations pertaining to each student into a single continuous time series using the Optimal τ method described in [24] and name this fusion the continuous set of labels. All other ground truth labels are computed from this lag-corrected and fused signal. We measure inter-rater reliability by computing the intraclass correlation (ICC(3,k)) separately for each subject and averaging. The ICC value is above 0.6 yielding a “good” agreement among the annotations according to [25].

The following descriptions refer to label names in Table 1. The binary labels are generated directly from the continuous labels using a 0.5 threshold. This interpretation is designed to test whether annotators label engaged versus disengaged behaviors with respect to the mid point of the annotation scale. The trinary and quintary labels are produced from the continuous labels by binning the data across subjects so that each bin contains the same number of samples. Though less interpretable, this scheme avoids having bins with too few samples which can hinder machine learning. The delta binary signal is created from the continuous signal by computing forward differences \( f : \{0, 1\} \times \{0, 1\} \to \{-1, 0, 1\} \) and corresponds to momentary engagement trends.

We also generate a set of ground truth labels via a novel approach where the continuous signal is binned using thresholds obtained by clustering the extrema and plateaus.
This method leverages a recurring observation [26], [27], [28] that human annotators more successfully capture trends and less accurately represent exact ratings. Thus, though the ratings of individual peaks and valleys may not be accurate or consistent, we hypothesize clusters of these points better represent the intended distribution of the data on average.

To detect these extrema and plateaus, we employ total variation (TV) denoising. TV denoising has been successfully used to remove salt and pepper noise from images while simultaneously preserving signal edges [29]. We use the TFOCS Matlab library [30] to find a new sequence $y_t$ that approximates the continuous annotation sequence $x_t$ and minimizes:

$$
\min_{y_t} \left[ \sum_t \|x_t - y_t\|^2_2 + \lambda \sum_t \|y_{t+1} - y_t\|_1 \right]
$$

The parameter $\lambda$ controls the influence of the temporal variation term and degree to which $y_t$ is approximately piecewise-constant. This parameter needs tuning to produce a desirable sequence and for this study we hand-tune $\lambda$ to a value of 0.05. This produces an approximately piecewise constant version of the continuous engagement signal.

Nearly-constant regions of the TV-denoised signal are extracted and correspond to the peaks, valleys, and plateaus in the original continuous engagement signal. For each constant region, we record the engagement value and then run k-means clustering with $k = 3$ on this list of scalars to generate three clusters corresponding to categories of high engagement, engagement, and disengagement. Finally, the continuous signal is converted to k-means trinary labels by thresholding on the midpoints between cluster means.

### 4.3. Engagement Prediction via Simple Classifiers

We test various machine learning algorithms on different ground truth representations for cross-subject engagement prediction and present the results in Table 1. Just prior to training, missing features (where no face is detected) are imputed using the mean feature values. In all cases, feature selection is employed prior to training where features with a Pearson correlation less than 0.4 with the engagement labels are discarded. This selection method decreases training time and also improves the performance on this data set. The correlation threshold is chosen using a grid search over the range $[0.1, 0.9]$ with 0.05 spacing. Among the selected features, one of any pair of features with an absolute mutual correlation greater than 0.95 is removed, and finally all features are Z-normalized. When using feature functionals, the final feature count is reduced by over 50 percent on average and the retained features include: eye gaze, facial landmarks, FACS action units (1,2,4,5,8,14,16,18,20,22,23,29,32,33), emotion probabilities for neutral and fear, average optical flow magnitude and direction, geometric facial features, and head size and pose. In other words, all categories of features are at least partially conserved.

Sklearn’s Python-based machine learning library [31] is utilized for all models and the hyperparameters are optimized using a grid search. For brevity, we report the optimal configuration only for the best model whose results are in bold in Table 1. This K-nearest neighbor model uses 201 neighboring points for classification.

For the continuous labels, all algorithms use a mean squared error loss function, and for all discrete label cases a weighted zero-one loss is employed where the weights are assigned per subject proportional to each session’s percentage of total samples. Leave-one-subject-out cross-validation is used to test each model and Table 1 shows weighted average scores across all subjects.

### 4.4. Engagement Prediction via Ensemble

Separately from our continuous annotations of engagement, we obtain discrete frame-wise annotations for subsets of the videos and use them to augment engagement prediction accuracy. Representative segments from some of the videos are selected and labeled frame-by-frame via majority voting by a panel of human annotators. These supplementary labels correspond to four states: no face visible (invalid frame), disengaged, engaged (attentive), and highly engaged (focused). We construct three binary classifiers using these labels, train them on the same video feature set, and show their performance in Table 2. The no face
classifier is trained to predict whether a face is not visible or obstructed. The attention classifier is trained only on frames for which a face is visible and predicts whether the student is visually engaged or disengaged. The focus classifier is trained only on frames with visible faces where the student is not disengaged and predicts whether the student is simply engaged or highly engaged.

We form an ensemble classifier using these binary predictors, the best KNN model for engagement prediction, and the k-means trinary engagement labels. The late fusion logic is displayed in Figure 2. We achieve a F-score of 0.369, an improvement over the 0.278 from the simple KNN classifier.

```
procedure LATEFUSION(data, engage_model, attention_model, focus_model)
  no_face_labels ← PREDICT(engage_model, data)
  no_face_labels ← PREDICT(no_face_model, data)
  attention_labels ← PREDICT(attention_model, data)
  focus_labels ← PREDICT(focus_model, data)
  for t = 1 to T do
    engage_label ← engage_labels(t)
    no_face_label ← no_face_labels(t)
    attention_label ← attention_labels(t)
    focus_label ← focus_labels(t)
    if no_face_label is True then
      engagement(t) ← 0
    else if attention_label is False then
      engagement(t) ← max(0, engage_label - 1)
    else if focus_label is False then
      engagement(t) ← engage_label
    else
      engagement(t) ← 2
    end if
  end if
end procedure
```

Figure 2. Late fusion ensemble method

4.5. Personalized Engagement Prediction

In a final set of experiments we test the utility of several learning models with our best ground truth labels (k-means trinary) on individual subjects. Table 3 shows average F1 scores when using five-fold cross validation on data per subject and averaging the performance across all subjects. To help mitigate learning deficiencies due to label imbalances, each fold retains a proportional fraction of each unique label value. This strategy avoids validating held-out sets containing unique or lacking certain labels.

5. Discussion

The performance in our cross-subject experiments are too low to demonstrate suitable engagement prediction. These results differ from those obtained in other related engagement experiments in the wild, such as the studies from Whitehill et al. [8] and Bosch et al. [9]. Given that the data collection methodology in this study more closely resembles distance learning settings, employs natural human-based engagement assessment, and focuses on end-to-end analysis, the performance differences are not surprising.

Figure 3 shows confusion matrices for the best-performing versions of each of the three approaches: cross-subject classifier, cross-subject ensemble, and an intra-subject classifier. The cross-subject ensemble conditioned on binary behavior features outperforms any other cross-subject model, but still not well enough to encourage further exploration using this approach. Even the best end-to-end method trained across subjects fails to identify visual feature combinations that generalize across individuals and are predictive of engagement. Our best personalized intra-subject model, however, performs significantly better on average and with little variation between subjects. The large performance gap between the cross-subject and intra-subject models and good performance of the personalized models suggest that the video features used in this study do not generalize across subjects.

If we consider scaling this approach, annotations of engagement would be necessary for each individual. Remarkably, if we train the best intra-subject model (random forest) on only a tiny random fraction of the data instead of using five-fold cross validation but still ensuring that all engagement labels are represented, then we can achieve nearly identical F1 scores. This result is consistent with an observation from Whitehill et al. [8] indicating that there is a lot of similarity across frames and thus a smaller number of annotations are necessary. In our initial tests, an average of only 25 annotated frames are necessary to achieve a F1 score within a few percent of the original when training individualized models. We intend to explore this in depth in future work.

Lastly, the novel k-means trinary method employed yields ground truth labels that give only modest performance benefits but still offer the best prediction performance overall. This demonstrates that some gains may be obtained by extracting coarse rankings from the less precise ratings and suggests that other rank-based methods for ground truth generation should be explored.

6. Conclusion

We assess the predictive power of state-of-the-art video features for estimating student engagement as perceived by human annotators in our distance learning corpus containing unstructured learning sessions. Several machine learning experiments are conducted using a variety of models and different representations of the ground truth labels to find the most promising end-to-end engagement prediction method. Our highest performance is achieved using a novel ground truth representation where constant regions from the fused continuous engagement annotations are clustered and used to quantize the signal. Prediction results in cross-subject experiments are low suggesting that the video features tested...
TABLE 2. ENGAGEMENT PREDICTION PERFORMANCE USING THE CONDITIONAL BINARY ENSEMBLE WITH DIFFERENT LEARNING ALGORITHMS

<table>
<thead>
<tr>
<th>Labels</th>
<th>Video Features</th>
<th>Functions of Video Features</th>
<th>Performance Metric</th>
</tr>
</thead>
<tbody>
<tr>
<td>Majority Class</td>
<td>SVC</td>
<td>KNN</td>
<td>RF</td>
</tr>
<tr>
<td>No Face</td>
<td>0.586</td>
<td>0.397</td>
<td>0.611</td>
</tr>
<tr>
<td>Attention</td>
<td>0.642</td>
<td>0.419</td>
<td>0.591</td>
</tr>
<tr>
<td>Focus</td>
<td>0.791</td>
<td>0.492</td>
<td>0.652</td>
</tr>
</tbody>
</table>

TABLE 3. AVERAGE INTRA-SUBJECT ENGAGEMENT PREDICTION RESULTS

<table>
<thead>
<tr>
<th>Labels</th>
<th>Video Features</th>
<th>Functions of Video Features</th>
<th>Performance Metric</th>
</tr>
</thead>
<tbody>
<tr>
<td>K-means trinary</td>
<td>Majority Class</td>
<td>SVC</td>
<td>KNN</td>
</tr>
<tr>
<td></td>
<td>0.254</td>
<td>0.357</td>
<td>0.321</td>
</tr>
</tbody>
</table>

Figure 3. Confusion matrices for the best-performing learning model in different experiments. Rows correspond to true labels, columns to predicted labels, and element values are normalized across the rows. The k-means trinary labels are used in each case.

are not agnostic to individual differences. Models trained and cross-validated on individual subjects, however, perform very well even when only a modest fraction of the annotated video frames are used.

Acknowledgments

We thank Dr. Michael Zyda for providing space and hardware for the Experience Lab. Funding for this research was provided by the National Science Foundation under NSF Award Number 1431024. Any opinions, findings, conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect those of the National Science Foundation.

References


[16] Y. Tong, W. Liao, and Q. Ji, “Facial action unit recognition by exploit-


[25] D. V. Cicchetti, “Guidelines, criteria, and rules of thumb for evaluat-


