Automated Detection of Engagement using Video-Based Estimation of Facial Expressions and Heart Rate

Hamed Monkaresi, Nigel Bosch, Rafael A. Calvo, Sidney K. D'Mello

Abstract—We explored how computer vision techniques can be used to detect engagement while students (N = 22) completed a structured writing activity (draft-feedback-review) similar to activities encountered in educational settings. Students provided engagement annotations both concurrently during the writing activity and retrospectively from videos of their faces after the activity. We used computer vision techniques to extract three sets of features from videos, heart rate, Animation Units (from Microsoft Kinect Face Tracker), and local binary patterns in three orthogonal planes (LBP-TOP). These features were used in supervised learning for detection of concurrent and retrospective self-reported engagement. Area Under the ROC Curve (AUC) was used to evaluate classifier accuracy using leave-several-students-out cross validation. We achieved an AUC = .758 for concurrent annotations and AUC = .733 for retrospective annotations. The Kinect Face Tracker features produced the best results among the individual channels, but the overall best results were found using a fusion of channels.

Index Terms— Engagement detection, remote heart rate measurement, facial expression, writing task.

1 Introduction

It is widely acknowledged that the way users engage in an activity is an essential component of their experience with the activity. The way people engage with an activity has been studied from multiple perspectives in HCI [1] and psychology [2]. The term "engagement" itself is interpreted in somewhat different ways by different communities of researchers [1], but most definitions maintain that engagement involves attentional and emotional involvement with a task. Engagement is also not stable, but fluctuates throughout an interaction experience. In the area of HCI, Peters et al. [1] discuss four phases of engagement: the beginning (i.e. point) of engagement, sustained attention or engagement, disengagement (when attention fades) and re-engagement.

Our present emphasis is on engagement during learning (or educational activities). Many authors (c.f. [2]) agree on four types of engagement during learning. Behavioral engagement can be assessed by observing persistence and effort; emotional engagement can be assessed by detecting supportive emotions (e.g., interest, curiosity) and self-efficacy [3]. Cognitive engagement is demonstrated when the student shows a sophisticated approach to the activity, for example by using deep rather than superficial learning strategies. Agentic engagement [4] occurs when the student attempts to actively enrich the experience, instead of merely acting as a passive recipient.

Engagement and affect has been linked to increased productivity and learning [2], [5] and psychological well-being [6]. When it comes to engagement in educational activities, an important consideration is that engagement is malleable. Numerous studies have shown that educational interventions, learning designs, and feedback are some of the ways in which student engagement can be enhanced (c.f. [2]). The impact of these interventions has been evaluated using multiple methodologies, often with an analysis of students' motivation.

Measuring engagement has been important in educational research because it allows researchers to understand what decisions promote or hinder engagement. Studies that focus on students' engagement need a way of measuring it. This can be done with one of the two types of data identified by engagement theorists: internal to the individual (cognitive and affective) and external observable factors, such as perceptible facial features, postures, speech, and actions [2]. Methodologically, as is common in many affective computing applications (c.f. [7]), studying engagement requires bringing together observational data (e.g., facial expressions, speech) and subjective data (e.g., self-reports).

New sensing and affective computing techniques allow for novel methodological approaches to measuring engagement. Different modalities such as video [8], audio [9] and physiological measures [10] are being used for affect detection in learning contexts. Multimodal approaches have also been explored to improve the accuracy of affect detection in learning applications [11]. The emotional states of students can be inferred from these

- *H. Monkaresi is with the School of Electrical and Information Engineering, The University of Sydney. Darlington, Sydney, NSW 2006, Australia. E-mail: Hamed.Monkaresi@ Sydney.edu.au.*
- *N. Bosch is with the Department of Computer Science, The University of Notre Dame, 384 Fitzpatrick Hall, Notre Dame, IN 46556, USA. E-mail: pbosch1@nd.edu*
- *R. A. Calvo is with the School of Electrical and Information Engineering, The University of Sydney. Darlington, Sydney, NSW 2006, Australia. E-mail: Rafael.Calvo@ Sydney.edu.au.*
- *Sidney D'Mello is with the Departments of Computer Science and Psychology, The University of Notre Dame, 384 Fitzpatrick Hall, Notre Dame, IN 46556, USA. E-mail: sdmello@nd.edu*

measures via affective computing techniques that are increasingly being used in learning technologies (e.g., [10], [12]–[14]). For example, Whitehill et al. [12] utilized a video-based method to detect engagement while students played cognitive training games. In this paper, we attempt to detect engagement in a different educational task (writing) with a rather different methodological approach. Our focus on writing is motivated by the fact that writing is one of the most common activities in both work and educational contexts, so we aim to support writing tools that help students engage with and enjoy their writing activities.

1.1 Contributions and Novelty

This study contains several novel aspects: 1) We detect engagement during a writing task, which offers unique challenges including limited facial expressivity and frequent downward head poses; 2) We use remote video-based detection of heart rate as a channel for engagement detection; 3) We use self-reports, instead of relying on external annotations as is commonly done; 4) We compare both concurrent and retrospective self-reports for "ground truth" labels. These contributions are discussed in more detail below.

With respect to the first point, writing offers a particularly difficult context for engagement detection for two reasons. First, when students use a computer to type, they frequently focus their attention downwards towards the keyboard instead of at the screen. This causes the head to tilt down, which in turn makes facial feature detection less accurate due to non-frontal and inconsistent head pose. Second, writing itself is less likely to be associated with detectable expressions compared to some other educational activities because it is less interactive than, for example, interacting with a conversational tutor or playing an educational game. This leads to more subtle facial expressions.

Second, in addition to using facial features, we also use heart rate extracted using computer vision techniques. We used a video-based method for measuring heart rate to detect user's engagement. This method initially requires a physiological device in order to calibrate the remote HR monitor and improve HR estimation accuracy. In spite of this real-world limitation, we adopted this approach in order to explore the possibility of engagement detection with high-accuracy remote HR sensing.

With respect to the source (self-reports) and nature (concurrent and retrospective) of the engagement annotations, we focus on self-reports of engagement as opposed to external annotations (by researchers or via crowdsourcing techniques [15]–[17]) as is commonly done. Self-reports of engagement differ from external annotations in that they are derived entirely (in the case of concurrent reports) or at least in part (in the case of retrospective reports) from the internal state of students. In the case of concurrent self-reports, students report engagement based on their current state, while in cued-recall retrospective protocols, the student's reports are based on the memory of their internal state and on the video of their face (and sometimes computer screen) to aid recall of the situation. On the other hand, external annotations are not based on the student's internal state, and thus may be a less accurate (or at least different) representation of the student's internal state. This difference is particularly key for face-based engagement detection because external annotations (such as [16]) are often made based on videos of students' faces. However, it is difficult to separate instances of bonafide engagement from instances where students appeared to be engaged but may not have been. To the best of our knowledge this study is one of the first studies that focuses on video-based automatic detection of self-reported engagement.

2 Background and Related Work

Most previous work of affect detection has focused on detecting basic emotions [18], [19], but more recently some researchers have focused on the recognition of complex mental states, particularly attention and engagement [20], [21], [22]. Engagement can be measured from different behavioral expressions: eye-gaze movements, facial features, gestures, and so on. Nakano and Ishii [23] attempted to measure user's engagement during human-robot conversations based on the user's gaze patterns, and the robot asked questions when the user was disengaged. They showed that considering the user's engagement can have a positive effect on the user's verbal and nonverbal behavior during a conversation with the robot. Rich et al. [24] also proposed a framework to detect and maintain user engagement during human-robot interactions. Their approach relied on tracking eye-gaze, speech, and gesture. Eye gaze has been shown to be a useful indicator of attentional focus, including mind wandering or zoning out [25]. Unfortunately, eye tracking is affected by head movements and is not yet easily scalable in real-world contexts. The present emphasis is on physiology-based and facial-feature based engagement detection as these are the two methods we explore in this research.

2.1. Physiology-based detection

Central and peripheral physiological signals have been commonly used for detecting task engagement, alertness, and drowsiness. Most of the proposed methods for measuring physiological states attempt to record and analyze the electrical signals produced by heart, brain, muscles and skin. The main instruments used to monitor physiological signals include Electrocardiogram (ECG), Electromyogram (EMG), galvanic skin response (GSR), and Respiration (RSP). Electroencephalogram (EEG) is widely used to differentiate between alertness vs. drowsiness [26]. Various EEG-based engagement indices have been proposed [27]. Classification accuracies between 84% and 99% have been achieved for detecting driver drowsiness detection using EEG methods [28]. A few studies have used EEG indices for engagement detection during human computer interaction [29], [30]. Belle et al. [31] achieved an accuracy of 85.7 % for detecting users' engagement when they were watching video clips.

Cardiac activity has also been explored for automatic

affect and engagement/alertness detection. Heart rate (HR) and heart rate variability (HRV) are two important ECG measures which have been used widely for these purposes. Previous researches showed that HR is a good indicator for discriminating between different affective states [32], [33]. For example, the HR tends to be higher during fear, anger and sadness than during happiness, disgust and surprise [33]. HR and HRV have been shown to be indicators of alertness and drowsiness [34], [35]. Liang et al. [35] analyzed HRV, HR, blood pressure and palm temperature to detect driver fatigue. They showed that HRV features could be highly effective for detecting driver drowsiness. Patel et al. [34] proposed a system for fatigue detection based on HRV features and achieved an accuracy of 90%.

One of the main challenges associated with physiological-based affective computing applications is the *intrusiveness* of physiological sensors. Users must have access to a heart rate monitor, which typically must be physically attached to the skin. This issue can be addressed by using remote measurement techniques. Three different approaches have been investigated for remote, contactless measurement of vital signs such as heart rate. Microwave Doppler radar [36]–[38] was one of the earliest methods for sensing heart rate and respiration. Thermal imaging [39], [40] is another approach for heart rate detection using analysis of skin temperature modulation. More recent approaches include video-based imaging methods [41]–[43] that use photoplethysmography to detect HRV. Compared to other approaches, video-based measurement of vital signs is cheaper and easier to adopt [44]. Current studies showed that these methods can be used in HCI applications [45]. In this paper we explore video-based HR sensing for engagement detection.

2.2. Facial-feature based detection

The use of computer vision techniques in affective computing is gaining traction with recent advances in low-cost hardware sensors (e.g., cameras) that can be integrated into computerized learning environments, as evidenced by a large body of existing work (see review articles [19], [46], [47]). The Microsoft Kinect and other similar depth cameras will likely become standard in future computer hardware. Cameras provide a non-intrusive continuous way of capturing images of peoples' faces as they use cell phones, computers, and even automobiles. This facial information can be used to understand certain facets of the user's current state of mind, and many techniques have been developed to automate this measurement process [19], [47], [48].

Ekman & Friesen [49] proposed the Facial Action Coding System (FACS), a widely used method for describing facial muscle action units (AU) and corresponding expressions. Current facial expression recognition systems can recognize several AUs with reasonable accuracies [50]. For example, a new face-tracker module embedded in Microsoft Kinect SDK (v1.5) is able to track six action units.

Two main approaches have typically been used in the area of facial expression analysis: geometric-based and appearance-based approaches. Geometric features include shapes and positions of face components, and the location of fixed facial points [51] such as the corners of the eyes, eyebrows, etc. [52], [53]. Appearance-based methods recognize facial expressions by analyzing the changes of the face's surface in both static and dynamic space (e.g., dynamic texture-based techniques). Facial expression recognition systems that use appearance-based features have been reported in [54]. Several researchers have used different types of features: for example, Gabor wavelet coefficients [55], optical flow [56], and Active Appearance Models [57]. Bartlett et al. [54] investigated different methods, such as explicit feature measurement, Independent Component Analysis (ICA), and Gabor wavelets. In their studies, Gabor wavelets provided the best results [58].

There are strengths and weaknesses in both the geometric-based and appearance-based approaches. Geometric-based methods typically track the position of a number of facial points in time. With this approach, some features of facial appearances (e.g., shape of mouth, position of eyebrows) can be extracted, while features related to texture of the face (e.g., furrows and wrinkles) cannot be extracted. In contrast, appearance-based methods may be more sensitive to changes in illumination (e.g., brightness and shadows), head motions and differences between shapes of the faces [19]. Tian et al. [59] used a combination of geometric-based and appearance-based features (Gabor wavelets) for recognizing facial AUs. They claimed that the geometric features outperformed the appearance features, yet using a combination of both yielded the best results.

Affect and engagement detection from facial features has also been investigated in learning contexts. For example, Grafsgaard et al. [20] used the Computer Expression Recognition Toolbox (CERT) to track facial movements within a naturalistic video corpus of tutorial dialogue. The most frequent AUs including eyebrow raising (inner and outer), brow lowering, eyelid tightening, and mouth dimpling were selected to predict overall levels of engagement, frustration and learning gains using forward stepwise linear regression. Their findings suggested that upper face movements would be a reliable predictor of engagement, frustration, and learning. They achieved reasonable agreement between their predictions and manual annotations, albeit at a rather coarse grained level (i.e., across the entire learning session) [20].

Whitehill et al. [16] used three different computer vision techniques to detect engagement in students as they interacted with cognitive skills training software. Box Filter features (which measure differences in grayscale pixel values among different regions of the face), Gabor features, and CERT features were used independently to create machine learning models for engagement detection. Labels used in their study were obtained from retrospective annotation of videos by external annotators. Four levels of engagement were annotated ranging from complete disengagement (not even looking at the material) to strong engagement. Detection performance was quantified using 2AFC, a means of estimating the Area Under

the ROC Curve (AUC) for a classifier [60]. They were able to detect engagement in a user-independent fashion with AUC = .729 (averaged across all four levels of engagement). Gabor features with an SVM classifier proved to be the most effective method they tried. This represents state-of-the-art engagement detection, which we will compare to in the present study.

2.3 Present approach

Similar to some related work [16], we detected engagement with a video-based approach using computer vision techniques. Related work (see review [46]) has also shown improved detection performance from combining multiple channels of data, so we adopted multiple techniques designed to produce different types of features. We used a combination of geometric features (Kinect Face Tracker), appearance features (Local binary patterns in three orthogonal planes), and physiological features (heart rate) that were extracted using computer vision techniques. Machine learning classification models were trained in a person-independent manner to ensure generalization to new students. Ground-truth measurements of engagement were obtained concurrently and retrospectively, and models were built separately for both types of self-report. We compare state-of-the-art engagement detection results to our models built using a fusion of features extracted using different techniques.

3 Data Collection Methodology

3.1 Participants

The participants were 23 undergraduate/postgraduate engineering students from a public university in Australia. The students ranged in age from 20 to 60 years (M = 34 years, SD = 11) and there were 14 males and 9 females. One student did not complete the entire session, so data for this student was discarded. The study was approved by the University of Sydney's Human Ethics Research Committee prior to data collection. The students signed an informed consent prior to the study.

3.2 Procedure

The study took approximately one hour and was conducted indoors with a varying amount of ambient sunlight entering through windows in combination with normal artificial fluorescent light. Students were asked to sit in front of a computer and write an essay (using Google docs) about a place they had visited recently (a journalistic genre), an activity that requires some research (e.g., finding information about the location) but not much prior knowledge and is likely to trigger arousing emotional memories. After receiving the topic, the writing session was based on a 'draft-review-final' activity: 1) Students wrote a draft and submitted it after 30 minutes; 2) They waited for 10 minutes to receive feedback. They were asked to stay seated while feedback was being processed but were free either to work on other manuscripts or browse the Internet; 3) They received human and automated feedback on how to improve the quality of the writing; 4) They had an additional 20 minutes to revise their manuscript according to the feedback and submit the final version.

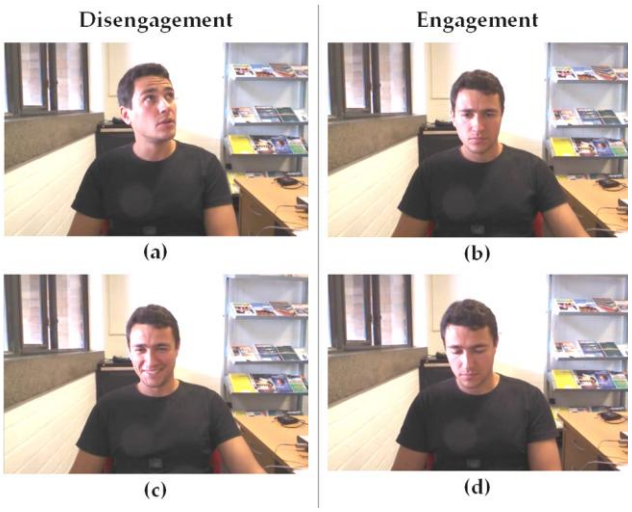

Fig. 1. Selected examples of disengagement (a and c) and engagement (b and d) annotations. (a) The student is distracted by noise. (b) The student is reviewing his writing. (c) The student disengaged from the writing task. (d) The student is typing and is engaged.

Videos of their faces and upper bodies were recorded with a Microsoft Kinect sensor in near mode. The sensor provides standard 2-dimensional color video, and 3-dimensional depth data. Color video was recorded in 24-bit RGB with 3 channels, 8 bits/channel, at 30 frames per second (fps) with a resolution of 640 × 480 pixels, and was saved in AVI format. The depth maps were also recorded at 30 fps with pixel resolution of 320 × 240. Sound was recorded using the Kinect microphone array. The student's heart rate was extracted from an ECG signal recorded by a BIOPAC MP150 system. Two ECG electrodes were placed on the student's wrists (left and right) and the ground electrode was placed on their ankle. Heart rate recordings were not used directly, but rather as ground-truth for video-based heart rate detection.

Concurrent and retrospective affective annotations were collected during and after the essay writing procedure, respectively.

3.2.1 Concurrent self-reports

For concurrent annotations, the system produced an auditory probe (beep) every two minutes *during* the essay writing activity. Students were instructed to verbally report their level of engagement (engaged in the writing task or not) in response to this probe. Their spoken responses were recorded with the Kinect microphone array. This method was used because it is a less-intrusive method for concurrent self-reporting compared to interrupting the writing session by asking students to fill out a questionnaire on their affect [61].

The impact of interrupting students to self-report their engagement level was not analyzed in this study. However, similar studies showed that this impact is not very high. For example, in a study involving a similar writing task, D'Mello and Mills [62] reported that the periodic (every 90 seconds) interruptions only had a negative im-

pact on 4.8% of the students even though they used a complex affect measurement instrument involving selecting among 11 affective states plus neutral. Nevertheless, the present study attempted to minimize the intrusiveness of the concurrent measure by: 1) simplifying the information to be self-reported (engaged vs. not-engaged) in order to decrease the mental effort required; 2) using a think-aloud method which is less disruptive compared to paper-based or computer based questionnaires; and 3) making the reporting task voluntary by allowing the students to ignore the prompt if they chose to do so.

3.2.2 Retrospective self-reports

Interval- and event- based segmentation are the two types of video segmentation used for retrospective video annotation. Each method has its strengths and weaknesses. When the behavior of a person is unpredictable, such as during writing activities, researchers [17], [63] suggest that event-based segmentation provides more useful annotations. Accordingly, an event-based video segmentation was used wherein each recorded video was divided into meaningful segments for annotation. More specifically, each event, such as a change in facial expression, or head or body posture movement, was segmented as an event by a researcher.

A week after their writing session, students returned to view their own extracted video segments and fill out a questionnaire after viewing each segment. They were free to play each segment as many times as needed to form an accurate judgment. The questionnaire simply asked them to report whether they were engaged or not ("*Were you engaged in the task or not?*"). Fig. 1 displays examples of disengaged and engaged instances as retrospectively annotated by the students.

3.3 Affect Reports and Engagement Levels

In concurrent self-reporting, the system asked students to verbally report their engagement every 2 minutes in each writing session (30 per session). Two students failed to report their level of engagement when the auditory probes sounded during the writing session, so no concurrent reports were available for those students. In all, 530 responses were obtained in response to 660 system probes. Students indicated that they were engaged for 425 cases (80%) compared to not being engaged (105 cases, or 20%).

For retrospective self-reports, 1,325 video segments were extracted from all videos. On average, 60.23 (SD = 8.25) video segments were extracted for each student. The average length of each video segment was 9.78 seconds (SD = 2.23). Students reported being engaged for a majority of the segments (996 segments or 75%), while they reported not being engaged for 315 segments (24%). 14 segments (1%) were labeled as "Not Applicable (N/A)."

There was a strong correlation between the proportion of engaged cases in concurrent and retrospective engagement reports (r = 0.82, p < 0.001). This provides evidence for the reliability of the two self-report measures.

Fig. 2 shows the average engagement level of students and their affective states during writing sessions as obtained from concurrent and retrospective self-reports. According to Fig. 2, the average of students' engagement was about 90% at the beginning of the task. Their engagement levels decreased gradually as they were approaching the middle of the session, when they had to submit their essays and wait 10 minutes to receive feedback (Break). Engagement reports corresponding to the breaks were not used in the affect detection analysis. Once they received the feedback they were engaged with writing for 90% of the reports. Again, their engagement waned as they neared the end of the session.

5 Engagement Detection Methodology

We followed three main steps for detecting students' engagement levels. In the first step, three types of features (Kinect face tracker, LBP-TOP, and Heart Rate - HR) were extracted from each video segment. The last 10 seconds of video before each annotation was considered for concurrent segments. The features were synchronized with corresponding concurrent and retrospective labels. Next, a feature selection technique was applied to the data in order to reduce the dimensionality of the feature space (as discussed below, feature selection was only applied to training data, not testing data). Finally, machine learning classification techniques were applied on selected features and validated with leave-several-students-out cross-validation for student-independent models. The summary of our methodology is illustrated in Fig. 3.

5.1 Feature Extraction

Three different methods were implemented for feature extraction from video recordings. These methods are explained in the following sub-sections.

5.1.1 Face Tracking Engine

The Kinect SDK's face tracking engine (v1.5) was used for facial feature extraction. This engine is able to track head position, ANimation Units (ANUs) and 100 facial points in real time. It should be noted that ANU codes are different from Action Units (AUs) proposed by Ekman [64]. For example ANU0 is equivalent to AU10, which measures upper lip raises.

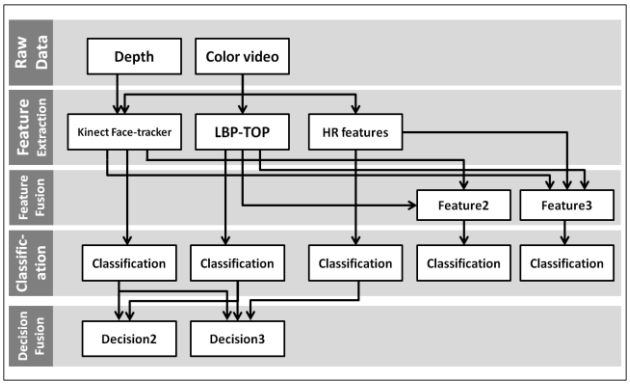

Fig. 3. An overview of our methodology for classifying engagement, (LBP-TOP: Local Binary Pattern in Three Orthogonal Plane)

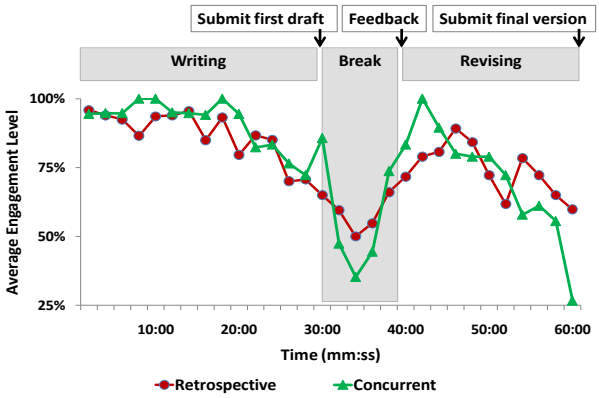

Fig. 2. Average engagement level of students (from retrospective and concurrent reports)

Six ANUs were tracked by the Kinect face tracking engine, which are a subset of those defined in the Candide3 model [65]. Four ANUs represent lip motions and two ANUs correspond to eyebrow motions. Each ANU was expressed as a numeric value ranging from -1 and +1. For example, ANU0 represents Upper Lip Raiser and a value of 0 indicates the teeth are fully covered by the upper lips; ANU0 equals +1 if teeth are fully visible; and it decreases towards -1 if the person pushes down the lip.

In addition to ANUs, the student's head position was captured via pitch, yaw and roll. Head translation was also measured in meters using the face tracking engine. In all, six ANUs were calculated for each detected face along with 3 values which specify head rotation in 3D space and three values which indicate the position of the head (X, Y and Z). Accordingly, twelve measures were calculated for each frame in the segment. Features were then created by aggregating the values of the 12 across the individual frames in the segment. Seven statistical measures (mean, median, standard deviation, max, min, range, difference between the values in the first and the last frame) were used to aggregate the 12 frame-level measures to obtain the final set of 84 segment-level FT features.

Lighting conditions, distance, and face occlusion are the main factors which affect the accuracy of the face tracking engine. The face tracking engine is also sensitive to the position of the head in front of the camera. According to the manual provided by Microsoft [66] the face tracking engine can track when the student's head pitch angle is less than 20 degrees, the roll angle is less than 90 degrees, and the yaw angle is less than 45 degrees. However, it works best when the head pitch, roll and yaw angels are less than 10, 45, and 30 degrees respectively. In some cases the face tracking engine could not detect the face and returned values which were outside of the abovementioned ranges. We detected and ignored these cases in preprocessing stages. Sometimes the face tracking engine mistakenly detected faces in the background. In order to avoid such false positives, we only considered the detected faces which were located in the normal range between 60 cm and 140 cm.

5.1.2 Local Binary Pattern in Three Orthogonal Planes

The Local Binary Pattern (LBP) proposed by Ojala et al. [67] is a powerful method for texture description. We used this method to describe the appearance and dynamics of facial objects. The LBP operator gets the color value of each central pixel and labels each neighborhood pixel by thresholding its color value with central pixel color value. Considering P neighboring pixels, 2^P different patterns could be considered. By calculating LBP for all pixels in an image and calculating the distribution of each pattern, a unique histogram could be extracted for each image. The histogram represents the number of occurrences of each specific local pattern. By defining different radii (R) and number of neighboring points (P), several types of LBP can be extracted. The best values for R and P depend on the application and general characteristics of the image sets.

Different variations of this method have been used for the problem of facial expression recognition [68]. Zhao and Pietikäinen [69] considered facial expressions as Dynamic Textures and used Volume Local Binary Patterns (VLBPs) for facial expression recognition. The approach showed promising results, although only the six prototypical emotions were recognized and no temporal segmentation was performed. In their approach, they normalized the face using the eye position in the first frame, but ignored any rigid head movement that may have occurred during the sequence. In addition, they used fixed overlapping blocks distributed evenly over the face instead of focusing on specific regions of the face, such as the mouth, eyes and eyebrows, which include valuable information about facial expressions. In this study, we used the Local Binary Patten in Three Orthogonal Planes (LBP-TOP) to recognize engagement. This method has been previously used to recognize three levels of valence and arousal and has achieved reasonable accuracy [70].

The LBP operator was originally designed for static images. Recently, Zhao and Pietikäinen [69] proposed an extended version of LBP to describe dynamic textures. Instead of considering a video sequence as a series of XY planes in axis T, it could be analyzed as a series of XT planes in axis Y and YT planes in axis X, respectively. Zhao and Pietikäinen [69] divided each video sequence into three orthogonal sets of 2-dimensional planes. The LBP could be computed for each set of planes separately. The LBP-TOP [69] descriptor for each video clip is calculated by concatenating three LBP histograms. Fig. 4, shows the LBP-TOP procedure. In such a representation, XT planes represent the changes in the appearance of the image during the time and the others (XT and YT) represent the motion patterns of each set (row/column) of pixels through Y and X axes separately.

Each component of the face has a specific texture and applying the LBP-TOP operator on the entire facial image might not provide useful information. Hence, we divided each facial video sequence into three blocks of images: left-eye, right-eye and mouth regions. These regions were detected and extracted automatically using an extended boosted cascade classifier [71]. Then, the deformation of eyes and mouth was monitored during each video segment. In order to have the same size blocks in each image, the detected objects (eyes and mouth) were resized to fixed sizes. The radius for extracting LBP-TOP patterns was set to 3. These setting (P = 8, R = 3) for LBP-TOP method provided the best result in our previous study on valence and arousal detection [72]. So, for each video sequence, 2,304 (3 regions × 3 orthogonal planes × 2^8 patterns) features were extracted.

5.1.3 Extracting HR features

We estimated HR based on the method introduced by Poh et al. [43] and then improved by Monkaresi et al. [45]. The first step was to detect and track the face in the recorded video using an extended boosted cascade classifier implemented in OpenCV (v. 2.2). The algorithm focuses on the regions which are more likely containing photoplethysmogram signal. The face area detected by the OpenCV library sometimes contains parts of the background region, which should be omitted before further analysis. Poh et al. [43] suggest considering the center 60% of the width of the detected face and the full height as the Region of Interest (ROI) in order ensure that there are no unwanted background regions. Next, the ROI was divided into the RGB channels and the average of each color (RGB) amplitude values was calculated across all pixels in the ROI. These three raw signals were inputs for an Independent Component Analysis (ICA). Before applying ICA, the raw signals were detrended [73] and normalized to improve the quality of the signals.

ICA [74] is a special case of Blind Source Separation (BSS) techniques which attempt to separate a multivariate signal into statistically independent subcomponents by assuming that the subcomponents are non-Gaussian signals. ICA finds the statistically independent components while the level of independence is maximum. Here, we adopted a linear ICA based on the Joint Approximate Diagonalization of Eigenmatrices (JADE) algorithm [75]. In the linear ICA, it is assumed that the observed signals contain linear mixtures of source signals. Typically, the number of basic source signals cannot be identified by ICA but the number of recoverable sources is less than or equal to the number of observations. In our case, the outputs of ICA were three independent components.

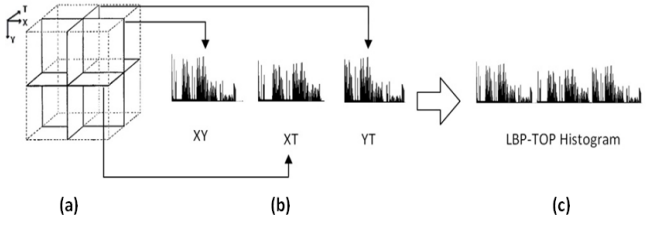

Fig. 4. LBP-TOP procedure: (a) A video sequence is divided into three sets of orthogonal planes, (b) LBPs are extracted and corresponding histograms are created for each set of planes, (c) Three histograms are concatenated to create the LBP-TOP histogram.

In order to identify the component that contains the HR signal, further analysis was needed. Poh et al. in [43] selected the second component manually as they argued that the HR signal could be observed clearly from that component. Monkaresi et al. [45] proposed a machine learning method to automatically estimate HR from the three components. Their proposed method improved the root mean squared error (RMSE) for HR prediction from 43.76 beats per minute (BPM) to 3.64 BPM, thus demonstrating the accuracy of vision-based HR measurement.

According to Monkaresi et al.'s method, nine features were extracted from the three power spectral density (PSD) curves of the independent components. From each independent component three features were extracted: The first feature is the frequency of highest peak in the PSD curves before applying the noise reduction method. The frequency of highest peak might represent the HR frequency. The noise reduction method tried to find and ignore noises based on previous correct estimations. If the difference between new estimation and the previous HR prediction was greater than 12 BPM, then new estimation was ignored and the algorithm examined the next highest peak. The noise reduction method repeated this process to find the first peak in the PSD curves which had the distance less than 12 BPM from previous estimation. The frequency of that peak was considered as the second feature. The depth of searches in the noise reduction method was considered as the third feature. Then, a nearest neighbor model (used for regression by averaging the values of the nearest neighbors) was trained using the nine features as input vectors and actual HR extracted from the ECG signal recorded by the BIOPAC system. The training and testing process was done based on a k-fold cross-validation approach (k = 10) for each student. The estimated HR values were used to extract HR features for each video segment. A RMSE of 8.49 BPM for HR prediction was achieved using this algorithm. It is clear that this level of error is not acceptable in clinical applications. However we assumed that this estimation can be used for affect and engagement detection if this study could demonstrate that it can support the idea of replacing HR sensors with cameras in certain types of applications in affective computing research. Seven statistical features (mean, median, standard deviation, max, min, range, difference) were extracted from the heart rate estimations for each video segment.

5.2 Feature Selection and Supervised Classification

We built engagement-detection models using each of the individual channels (HR, FT, LBP-TOP) as well as two models using a fusion of equal numbers of features from each of these channels (feature-level fusion) and two models using a fusion of the classifications from each of these channels (decision-level fusion). For the feature-level fusion we built models using features from all three channels (*Feature3*). We used seven features from each of

the channels in the Feature3 model because the HR channel has only seven features. Additional features could not be used without effectively weighting the influence of one channel more heavily than another. Second, we built models using features from two channels (*Feature2*), namely FT and LBP-TOP. We were able to selectively choose the most predictive features from each channel to build Feature2 because FT and LBP-TOP have many more features than HR. Finally, we also built decision-level fusion models using the classification output of all three channels (*Decision3*) and the classification output of FT and LBP-TOP (*Decision2*) to correspond to the feature-level fusion models.

Decision-level fusion classifiers used individual channel base classifiers to make a classification by using the decision (engaged or not engaged) of whichever base classifier had the highest decision probability. One of the key advantages of using a decision-level fusion classifier is that the base classifiers could be trained on instances regardless of whether or not those instances were available in the other channels. For example, Feature2 could only be trained on instances where features could be detected in both FT and LBP-TOP, but Decision2 used individual base classifiers so the FT model was trained on all the valid FT instances (within the training dataset) even when those instances could not be detected using the LBP-TOP method (and vice versa for the LBP-TOP base classifier). Classifiers were tested using instances that were available in all channels, so that a decision could be obtained from each base classifier.

5.2.1 Feature Selection

In order to reduce the dimensionality of extracted features, RELIEF-F feature selection was used (on training data only) [76]. This technique assigns weight to features according to the Euclidean distance of instances from one class to other instances in the same class, compared to the distance to instances in another class. Features were then ranked by weight and a certain proportion of the top-ranked features were used. We explored six different proportions of features used in FT channel (.10, .20, .30, .40, .50, .75), five in the LBP-TOP channel (.003, .005, .008, .012, .016), and three in the HR channel (.25, .50, .75). The HR data had only seven features, so we did not use as many different proportions as in the FT channel. The LBP-TOP channel had such a large number of features that it was not feasible to try a large proportion of them as was done in the FT channel (with .75) so we tried only five small proportions.

5.2.2 SMOTE

Synthetic Minority Oversampling Technique (SMOTE [77]) was used in some of the resulting models as a means to handle data imbalance in the training data (but not testing data). SMOTE creates synthetic instances by projecting new data points in the feature space at random positions that fall between a point in the minority class and the nearest within-class neighbors of that point. We compared the effect of using SMOTE to no resampling.

5.2.3 Outlier Handling

Outliers were considered to be any value in an instance that was more than three standard deviations from the mean of that feature. We experimented with two outlier treatments: leaving them unchanged and Winsorization (replacing outliers with the value three standard deviations above or below the mean).

5.2.4 Classifiers

We used WEKA [78], a machine learning tool, for classification. We experimented with different classifiers for each model that was built.

Updateable Naïve Bayes. This classifier finds the distribution (mean and standard deviation) for each feature within each class. The class label for a test instance is then predicted by applying Bayes' theorem to determine the probability of the instance being a member of a particular class given the distribution for that class, and choosing the class with the highest such probability.

Bayes Net. This classifier is a graphical model that also utilizes Bayes' theorem to compactly represent conditional probabilities of a set of features. A simple network structure was used in which individual features were assumed to be independent of each other but related to the class label.

Logistic Regression. This classifier works by regressing the class label on the features using the logistic function. The logistic function is bounded on the [0, 1] interval and can thus be interpreted as indicating one class or the other depending on which side of a cutoff (e.g., .5) the prediction lies for a test instance.

Classification via Clustering. K-means clustering (with K = 2 to match the number of classes) is applied to the training data to produce two clusters. Each cluster is then assigned a

TABLE 1
Summary of models built with concurrent self-reports

Channel	AUC	Classifier	SMOTE	Winsorization	No. Features	No. Instances	Prop. Engaged
HR	0.544	Nearest Neighbor	No	Yes	4	408	0.841
FT	0.635	Logistic	Yes	Yes	26	243	0.810
LBP-TOP	0.645	Bayes Net	Yes	No	37	222	0.795
Feature3	0.751	Updateable Naïve Bayes	No	Yes	21	142	0.788
Feature2	**0.758**	**Updateable Naïve Bayes**	**No**	**Yes**	**26**	**146**	**0.763**
Decision3	0.635	Decision-Level Voting	Mixed	Mixed	-	-	0.820
Decision2	0.690	Decision-Level Voting	Yes	Mixed	-	-	0.800

label depending on which class it corresponds to best. Test instances are then assigned labels according to which cluster they correspond to.

Rotation Forest. In the Rotation Forest classifier, features are randomly split into groups of three and then principle components analysis (PCA) is applied to each group. A separate classifier (C4.5 decision tree) is then applied to the PCA components to build a base classifier on a randomly chosen 50% of instances. This process is repeated ten times to produce an ensemble of classifiers, and then test instances are classified by assigning them to the class with the highest average confidence.

Dagging. This classifier creates an ensemble of ten base classifiers (decision stump classifiers) by randomly splitting training data into ten folds and training a separate classifier on each. Test instances are then classified by majority vote of the base classifiers.

5.2.5 Validation

All models were trained and tested with student-level nested cross validation. The training set consisted of all data from 66% of students (chosen randomly), while the testing set consisted of the data from the remaining students. Feature selection was performed with nested cross-validation within the training set only. That is, feature selection was performed repeatedly on the training set using data from 66% of the students in the training set (44% of total students) and averaging results across 10 rounds of feature selection. The entire classification process was repeated 100 times for each model with different randomly chosen students in the training and testing sets.

We used estimated Area Under the ROC Curve (AUC) as our primary classification performance metric because it accounts for chance level accuracy (AUC = .5), closely approximates the A′ statistic, and is more robust to class imbalances than common measures of classification performance such as kappa and accuracy [79]. An AUC of 0 represents completely incorrect classification while an AUC of 1 represents perfect classification

6 Classification Results

In this section the performance of the classifiers for discriminating two levels of engagement (Engaged vs. Not Engaged) using different channels are reported. According to the methods explained in the Section 5, for each video segment, 84 features were extracted for the FT channel, 2,304 features were extracted using the LBP-TOP method, and 7 features were extracted for the HR channel. In addition to these three channels, results of combinations of channels are also presented. The first fusion model (Feature3) combined all three channels using feature-level fusion, while in the second fusion model (Feature2) only the best-performing two individual channels were combined. Similarly, decision-level fusion models on all three channels (Decision3) and the best two channels (Decision2) were built. Results are reported separately for retrospective and concurrent labels.

6.1 Concurrent Self-Report Classification Results

The best performing models that were built with data from the concurrent self-reports are presented in Table 1. Before delving into the results, it is important to discuss issues pertaining to the varying number of instances and features across the models.

TABLE 2
Confusion matrices for the best individual channel and best channel fusion models built with concurrent self-report labels

	Actual	Classified		Priors
LBP-TOP		*Engaged*	*Not Engaged*	
	Engaged	**.723 (hit)**	.277 (miss)	0.795
	Not Engaged	.544 (false alarm)	.456 (correct rejection)	0.205
Feature2		*Engaged*	*Not Engaged*	
	Engaged	**.841 (hit)**	.159 (miss)	0.763
	Not Engaged	.528 (false alarm)	.472 (correct rejection)	0.237

The number of instances varied between channels, as can be seen in Table 1. The HR channel was least susceptible to missing instances, as it relied on less specific facial features than FT or LBP-TOP and could handle situations with more extreme head pose and occlusion. Fusing multiple channels at the feature level required data from each channel, so Feature2 was missing instances that were missing in either the FT or LBP-TOP channel, and Feature3 was missing instances in one or more of any of the three channels. Loss of instances did not adversely affect engagement base rates as noted in Table 1.

The decision-level fusion models used the same number of instances for testing as the feature-level fusion models (142 for Decision3 and 146 for Decision2), as discussed in Section 5.2. The number of training instances was larger for the decision-level fusion models than for the feature-level fusion models, however, because 408 instances could be used during training of the HR base classifier, 243 for FT, and 222 for LBP-TOP. Note that instances were still chosen only from the training data in each iteration, however, so that proper training/testing independence was preserved.

Variance in the number of features across models was a factor of number of available features and feature selection. For the Feature2 model, the number of features was chosen by considering the number of features selected in the individual channel models (63 in FT, 28 in LBP-TOP) and taking the minimum of those two. Thus 28 features were used from the FT and LBP-TOP channels, 14 from each. This was done to ensure that a channel with more features (e.g., LBP-TOP) did not dominate a channel with fewer features (e.g., FT and HR). Feature3 used seven features from each modality because the HR channel only had seven features.

A number of key conclusions can be drawn from Table 1. First, performance of individual channels could be ordered as: HR < FT < LBP-TOP, though the difference between FT and LBP-TOP is quite muted. Second, fusing channels resulted in noticeable performance improvements over the best individual channel (FT). Third, fusing the individual FT and LBP-TOP models across both fusion schemes yielded more accurate results over fusing all

TABLE 3
Summary of models built with retrospective self-reports

Channel	AUC	Classifier	SMOTE	Winsorization	No. Features	No. Instances	Prop. Engaged
HR	0.590	Updateable Naïve Bayes	Yes	Yes	4	817	0.792
FT	0.666	Updateable Naïve Bayes	No	No	42	470	0.784
LBP-TOP	0.644	Updateable Naïve Bayes	No	No	37	650	0.761
Feature3	0.697	Logistic	Yes	No	21	323	0.774
Feature2	**0.733**	**Updateable Naïve Bayes**	**No**	**Yes**	**36**	**342**	**0.773**
Decision3	0.683	Decision-Level Voting	Mixed	Mixed	-	-	0.777
Decision2	**0.730**	**Decision-Level Voting**	**No**	**No**	**-**	**-**	**0.763**

three channels (presumably due to the lower performance of the HR model). Fourth, feature-level fusion outperformed decision-level fusion, and the most accurate model was the feature-level Feature2 model (i.e. FT + LBP-TOP).

To examine the advantages of channel fusion over individual channels in greater detail, we assessed the confusion matrices (see Table 2) for the best individual channel (LBP-TOP) and the best fusion model (Feature2). We note that most of the improvement in detection performance achieved by channel fusion can be attributed to more accurate detection of true positives (hits); identification of true negatives (i.e., correct rejection) was consistent across the two models.

6.2 Retrospective Self-Report Classification Results

Table 3 shows an overview of performance for different channels when using retrospective labels. Similar to the concurrent labels, instances for the retrospective models also differed across channels due to challenges caused by head pose and other occlusions. However, the number of instances for each channel and fusion was higher for retrospective labels than for concurrent labels. As evident in Table 3, instance removal did not drastically alter the base rate of the retrospective engagement distribution.

As with concurrent fusion models, we created a fusion model with the best two individual channels (FT and LBP-TOP), called Feature2. The model for the FT channel had 42 features while the LBP-TOP model had 37, so we used 36 features in our Feature2 model (18 from each channel) since 37 could not be evenly divided between the two channels. The Feature3 model had 21 features, 7 from each modality, due to the limitation of 7 features in the HR channel.

The results indicated that, similar to the concurrent models, HR resulted in the lowest performance. FT was the best individual channel, though its accuracy was similar to LBP-TOP. The fusion models provided a notable increase in performance over individual channels for both feature- and decision- level fusion. Once again, the Feature2 and Decision2 models outperformed the Feature3 and Decision3 models. However, unlike the concurrent models, performance of the Feature2 and Decision2, models was mostly equivalent.

Confusion matrices for the best individual model (FT) and best fusion model (Feature2) are shown in Table 4. In the case of retrospective labels, classification performance was largely improved in terms of correct rejection rates. Specifically, both FT and Feature2 models had nearly identical hit rates, but the correct rejection rates were much higher for the fusion model.

It should be noted that although Feature2 model showed a small advantage over the Decision2 model (AUC of .733 vs. .730), the Feature2 model had a disadvantage in terms of instances used for training and testing. Decision2 had nearly equal performance to Feature2, but with the added advantage of additional training data (470 instances for FT, 650 for LBP-TOP) that may produce better performance in cases where the additional training data could be especially advantageous.

6.3 Comparison of Concurrent and Retrospective Results

Fig. 5 shows a comparison of the best concurrent and retrospective models. For the individual channels, the retrospective models outperformed the concurrent models for HR and FT but not for LBP-TOP. Fusion improved results for both concurrent and retrospective models, albeit in different ways. Specifically, feature-level fusion was more successful for the concurrent models and vice versa for the retrospective models. Overall, the *best* concurrent model had a higher AUC (.758) than the best retrospective model (.733), but *on average,* performance for both label types was equivalent (mean AUC = .665 for concurrent and .678 for retrospective).

6.4 Comparison to State-of-the-Art Results

The work by Whitehill et al. [16] perhaps reflects the state of the art in vision-based engagement detection in learning contexts. They used human coders to *retrospectively* judge the level of student engagement based on videos of the students' faces recorded during a cognitive skills training activity. The four levels of engagement ranged from complete disengagement (level 1, i.e. not looking at the screen or closed eyes) to very engaged (level 4). Their best person-independent model, an SVM classifier with Gabor features, yielded an average AUC of 0.729 for detecting each level of engagement from all the other levels (i.e., level 1 vs. levels 2, 3, and 4, etc). In contrast, our best person-independent concurrent model achieved an AUC of .758, which reflects a measurable improvement over

TABLE 4
Confusion matrices for the best individual channel and best channel fusion models built with retrospective self-report labels

	Actual	Classified		Priors
		Engaged	*Not Engaged*	
FT	*Engaged*	.790 (hit)	.210 (miss)	0.784
	Not Engaged	.546 (false alarm)	**.454 (correct rejection)**	0.216
		Engaged	*Not Engaged*	
Feature2	*Engaged*	.789 (hit)	.211 (miss)	0.773
	Not Engaged	.465 (false alarm)	**.535 (correct rejection)**	0.227

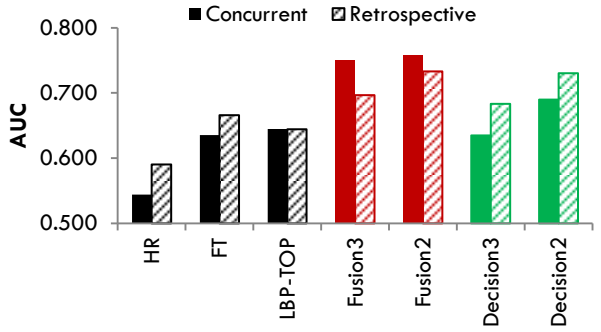

Fig. 5. Concurrent and retrospective results

the previous state-of-the-art for person-independent engagement detection. However, detection associated with level 1 was very high (AUC = .914), so their approach may be more successful for complete disengagement (e.g., not even looking at screen). For the three remaining levels involving some degree of engagement (levels 2-4), their average AUC was .667, which may indicate that our approach is more appropriate when there is some amount of engagement.

It is important to note, however, that a direct comparison of techniques between studies is not without its difficulties, since the datasets differ in terms of both videos and annotation methods. For example, the engagement annotations in Whitehill et al. were obtained by researchers after the learning session, while the labels in our data were provided by student self-reports both during and after the session. However, despite the difficulty of comparison we believe that our results have the potential for establishing a new state of the art standard in face-based engagement detection.

7 Discussion and Conclusions

The pervasiveness of cameras, be it simple webcams or more sophisticated technologies such as the Kinect, opens new opportunities in the design of software interfaces. Though there is considerable research on video-based detection of affect, much (but not all) of this research focuses on the basic emotions (anger, fear, surprise, disgust, sadness, and happiness). Taking a different approach, the present paper focused on automatic detection of engagement, which is a complex state with affective, cognitive, behavioral, and agentic components [4]. In the remainder of this section, we discuss our main contributions and findings and consider limitations and avenues for future work.

7.1. Contributions and Findings

The main goal was to introduce automated video-based methods to detect engagement in a learning context. This is an important step towards developing software tools and interventions that promote engagement. Our results showed that engagement can be detected in realistic scenarios with moderate accuracy. We compared the performance of the engagement detectors to state-of-the-art engagement detection results that also use video data (section 6.4). Our results showed improved engagement detection performance when compared to current results that discriminated across multiple levels of engagement.

The most successful models overall were created with a fusion of features and with concurrent engagement labels. In particular, the model using a fusion of the best two channels (Feature2) with concurrent labels provided the best overall result, with AUC = .758. In comparison, the best retrospective model had lower (but respectable) accuracy (AUC = .733 using the Feature2 model). Importantly, the models were validated using student-level nested-cross validation, so we have some confidence that the results will generalize to new students with similar demographics.

One important contribution of the present study involved the fact that it directly compared both concurrent and retrospective engagement annotations. We found a tradeoff between hit rate and false alarm rate that may affect which model is most suitable for use in an application. If concerned primarily with high-precision engagement detection, the choice for which model to use is clear: the concurrent Feature2 model has the highest hit rate. However, the best retrospective model had higher correct rejection rate than the best concurrent model (i.e., better detection of disengagement). The cost to students that may arise from incorrectly classifying a student as not engaged when they are actually engaged may be quite different from the cost of missing an occurrence of disengagement, and is arguably more important for learning environments that need to intervene when students become disengaged. The set of retrospective labels was also larger than the concurrent labels, so in a situation where it might be most advantageous to have a large amount of training data, the retrospective decision-level model Decision2 might be the best choice.

Furthermore, the advantages of decision-level fusion are worth considering while choosing a "best" detector for deployment. One benefit of decision-level fusion is that the individual base classifiers could be trained with more data than the feature-level fusion models. A decision-level fusion model can potentially operate in more situations than any one individual detector. For example, it could use a decision from the FT channel detector when the LBP-TOP channel is unavailable, or vice-versa.

Another important contribution of this work was the use of video-based methods for remote monitoring of heart rate. Although physiological-based engagement detection has shown promise (as reviewed in Section 2), video-based detections is more scalable due to the widespread availability, low cost, and lower intrusiveness of cameras compared to physiological sensors. Thus, we attempted to capitalize on the merits of both approaches by considering video-based remote sensing of a physiological signal (HR). The results indicated that performance of the HR channel was lower than the facial expression based channels (FT and LBP-TOP). One reason for this result is that HR only provides one index into cardiac activity. Other features such as HRV are more informative and have been known to be more strongly correlated with mental states [35], but could not be extracted with the proposed approach. Future advances in remote physiological sensing might be needed before their potential can be fully understood.

7.2 Limitations and Future Work

It is important to point out several limitations with this research. First, gathering and analyzing behavioral data in naturalistic scenarios is one of the most challenging issues. The current study suffered limitations in this regard. Due to limits in analyzing head motions, and frequent face occlusions, our methods were not able to extract features from some video segments, thereby leading to data loss. For example, some students were too close to the screen during writing, and their face (or part of it) was not in the range of video recording. We had to ignore these video segments for classification. The Kinect Face Tracker was more sensitive to these conditions compared to the LBP-TOP method. Furthermore, our approach relied on a manual segmentation of the videos, which might need to be replaced with random segmentation.

Another limitation of this study is associated with the method we used for HR estimation. Current video-based methods are far from perfect when detecting HR in naturalistic and practical scenarios. Several studies showed that photoplethysmography signals could not be detected when the subject was moving [43], [80]. Machine learning approaches [45], [80] have been proposed to improve the accuracy of HR detection when the users behave normally in front of camera. We used a method that provided acceptable accuracy, but needed to be trained with HR signals which were recorded by a BIOPAC system. In a real world-context, this defeats the purpose of remote-sensing since one needs a sensor-based approach to parameterize the vision-based approach. It could also be argued that using the actual HR signals might improve the results reported in this paper. We acknowledge this limitation given that our goal was to experiment with an accurate video-based HR tracking system. Improving the accuracy of remote physiological measurement techniques, ostensibly without requiring a learning phase, for practical applications is an important item for future work.

7.3. Concluding Remarks

It is our hope that improved automatic detection of engagement in computerized education environments will lead to more effective learning and a more engaging experience for students. To that end we presented our methods and results for detecting student engagement in the context of a writing task. Our methods showed that combining facial texture- and appearance- based features resulted in the most accurate student-independent engagement detectors. We also considered the possibility of engagement detection from remote-sensing of heart rate, however, this did not result in improved performance, ostensibly due to the limited amount of information that could be sensed. Future improvements to remote heart rate detection will provide more opportunities to explore the combination of channels and further increase the effectiveness of engagement detection in computerized learning environments.

ACKNOWLEDGMENTS

Sidney D'Mello and Nigel Bosch were supported by the National Science Foundation (NSF) (DRL 1235958) and the Bill & Melinda Gates Foundation. Rafael Calvo was supported by the Young and Well Cooperative Research Centre. Rafael Calvo and Hamed Monkaresi were supported by the LEADS project of the SSHRC, Canada. Any opinions, findings and conclusions, or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of the NSF.

REFERENCES

[1] C. Peters, G. Castellano, and S. de Freitas, "An exploration of user engagement in HCI," in *Proceedings of the International Workshop on Affective-Aware Virtual Agents and Social Robots - AFFINE '09*, 2009, pp. 1–3.

[2] S. L. Christenson, A. L. Reschly, and C. Wylie, Eds., *Handbook of Research in Student Engagement*. New York: Springer, 2012.

[3] E. A. Linnenbrink and P. R. Pintrich, "The Role of Self-Efficacy Beliefs Instudent Engagement and Learning Intheclassroom," *Read. Writ. Q. Overcoming Learn. Difficulties*, vol. 19, no. 2, pp. 119–137, Apr. 2003.

[4] J. Reeve and C.-M. Tseng, "Agency as a fourth aspect of students' engagement during learning activities," *Contemp. Educ. Psychol.*, vol. 36, no. 4, pp. 257–267, Oct. 2011.

[5] W. A. Kahn, "Psychological Conditions of Personal Engagement and Disengagement at Work," *Acad. Manag. J.*, vol. 33, no. 4, pp. 692–724, 1990.

[6] R. A. Calvo and D. Peters, *Positive Computing: Technology for Wellbeing and Human Potential*. Cambridge, MA: MIT Press, 2014.

[7] R. A. Calvo, S. K. D'Mello, A. Kappas, and J. Gratch, Eds., *The Oxford Handbook of Affective Computing*. New York: Oxford University Press, 2014.

[8] J. F. Grafsgaard, J. B. Wiggins, K. E. Boyer, E. N. Wiebe, J. C. Lester, and N. Carolina, "Embodied Affect in Tutorial Dialogue : Student Gesture and Posture," in *Proceedings of the 16th international conference on Artificial intelligence in education*, 2013, pp. 1–10.

[9] K. Forbes-riley and D. Litman, "When Does Disengagement Correlate with Learning in Spoken Dialog Computer Tutoring ?," in *Proceedings of the 15th international conference on Artificial intelligence in education*, 2011, pp. 81–89.

[10] M. S. Hussain, O. Alzoubi, R. A. Calvo, and S. D. Mello, "Affect Detection from Multichannel Physiology during Learning Sessions with AutoTutor," in *Artificial Intelligence in Education*, G. Biswas, S. Bull, J. Kay, and A. Mitrovic, Eds. Springer, 2011, pp. 131–138.

[11] A. Kapoor and R. W. Picard, "Multimodal affect recognition in learning environments," in *Proceedings of the 13th annual ACM international conference on Multimedia*, 2005, pp. 677–682.
[12] J. Whitehill, Z. Serpell, A. Foster, Y.-C. Lin, B. Pearson, M. Bartlett, and J. Movellan, "Towards an Optimal Affect-Sensitive Instructional System of cognitive skills," in *IEEE Conference on Computer Vision and Pattern Recognition: Workshop on Human-Communicative Behavior*, 2011, pp. 20–25.
[13] A. Graesser, A. Witherspoon, B. McDaniel, S. D'Mello, P. Chipman, and B. Gholson, "Detection of Emotions during Learning with AutoTutor," in *Proceedings of the 28th Annual Meetings of the Cognitive Science Society*, 2006, pp. 285–290.
[14] R. A. Calvo and S. D'Mello, Eds., *New Perspectives on Affect and Learning Technologies*, vol. 3. New York: Springer, 2011.
[15] M. Soleymani and M. Larson, "Crowdsourcing for affective annotation of video: Development of a viewer-reported boredom corpus," in *Proceedings of the ACM SIGIR 2010 workshop on crowdsourcing for search evaluation (CSE 2010)*, 2010, pp. 4–8.
[16] J. Whitehill, Z. Serpell, Y.-C. Lin, A. Foster, and J. R. Movellan, "The Faces of Engagement: Automatic Recognition of Student Engagement from Facial Expressions," *IEEE Trans. Affect. Comput.*, vol. 5, no. 1, pp. 86–98, Jan. 2014.
[17] S. Afzal and P. Robinson, "Natural Affect Data - Collection & Annotation in a Learning Context," in *3rd International conference on Affective Computing and Intelligent Interaction, ACII 2009*, 2009, pp. 1–7.
[18] S. D'Mello and R. A. Calvo, "Beyond the Basic Emotions : What Should Affective Computing Compute ?," in *CHI 2013 - Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2013.
[19] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A survey of affect recognition methods: audio, visual, and spontaneous expressions.," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 1, pp. 39–58, Jan. 2009.
[20] J. F. Grafsgaard, J. B. Wiggins, K. E. Boyer, E. N. Wiebe, and J. C. Lester, "Automatically Recognizing Facial Expression: Predicting Engagement and Frustration," in *Educational Data Mining*, 2013.
[21] D. Bohus and E. Horvitz, "Models for Multiparty Engagement in Open-World Dialog Models for Multiparty Engagement," in *Proceedings of SIGDIAL 2009*, 2009, no. September, pp. 225–234.
[22] B. McDaniel, S. D'Mello, B. King, P. Chipman, K. Tapp, and A. Graesser, "Facial Features for Affective State Detection in Learning Environments," in *29th Annual meeting of the cognitive science society*, 2007, pp. 467–472.
[23] Y. I. Nakano and R. Ishii, "Estimating user's engagement from eye-gaze behaviors in human-agent conversations," in *Proceedings of the 15th international conference on Intelligent user interfaces - IUI '10*, 2010, pp. 139–148.
[24] C. Rich, B. Ponsler, A. Holroyd, and C. L. Sidner, "Recognizing engagement in human-robot interaction," *2010 5th ACM/IEEE Int. Conf. Human-Robot Interact.*, pp. 375–382, Mar. 2010.
[25] S. K. D'Mello, J. Cobian, and M. Hunter, "Automatic Gaze-Based Detection of Mind Wandering during Reading," in *Proceedings of the 6th International Conference on Educational Data Mining (EDM 2013)*, 2013, pp. 364–365.
[26] M. V. M. Yeo, X. Li, K. Shen, and E. P. V. Wilder-Smith, "Can SVM be used for automatic EEG detection of drowsiness during car driving?," *Saf. Sci.*, vol. 47, no. 1, pp. 115–124, Jan. 2009.
[27] F. G. Freeman, P. J. Mikulka, L. J. Prinzel, and M. W. Scerbo, "Evaluation of an adaptive automation system using three EEG indices with a visual tracking task," *Biol. Psychol.*, vol. 50, no. 1, pp. 61–76, May 1999.
[28] A. Sahayadhas, K. Sundaraj, and M. Murugappan, "Detecting driver drowsiness based on sensors: a review," *Sensors*, vol. 12, no. 12, pp. 16937–53, Jan. 2012.
[29] R. Stevens, T. Galloway, and C. Berka, "EEG-related changes in cognitive workload, engagement and distraction as students acquire problem solving skills," in *11th International Conference on IUser Modeling*, C. Conati, K. McCoy, and G. Paliouras, Eds. Corfu, Greece: Springer, 2007, pp. 187–196.
[30] M. Chaouachi, P. Chalfoun, I. Jraidi, and C. Frasson, "Affect and Mental Engagement: Towards Adaptability for Intelligent Systems," in *Proceedings of the Twenty-Third International Florida Artificial Intelligence Research Society Conference (FLAIRS 2010)*, 2010, no. Flairs, pp. 355–360.
[31] A. Belle, R. H. Hargraves, and K. Najarian, "An automated optimal engagement and attention detection system using electrocardiogram.," *Comput. Math. Methods Med.*, vol. 2012, pp. 1–12, Jan. 2012.
[32] S. D. Kreibig, "Autonomic Nervous System Activity in Emotion: A Review," *Biol. Psychol.*, vol. 84, pp. 394–421, 2010.
[33] R. W. Levenson, P. Ekman, and W. V. Friesen, "Voluntary Facial Action Generates Emotion-Specific Autonomic Nervous System Activity," *Psychophysiology*, vol. 27, pp. 363–384, 1990.
[34] M. Patel, S. K. L. Lal, D. Kavanagh, and P. Rossiter, "Applying neural network analysis on heart rate variability data to assess driver fatigue," *Expert Syst. Appl.*, vol. 38, no. 6, pp. 7235–7242, Jun. 2011.
[35] W. C. Liang, J. Yuan, D. C. Sun, and M. H. Lin, "Changes in physiological parameters induced by indoor simulated driving: effect of lower body exercise at mid-term break.," *Sensors (Basel).*, vol. 9, no. 9, pp. 6913–33, Jan. 2009.
[36] C. Li, J. Cummings, and J. Lam, "Radar remote monitoring of vital signs," *Microw. Mag.*, vol. 10, no. February, pp. 47–56, 2009.
[37] E. F. Greneker, "Radar sensing of heartbeat and respiration at a distance with applications of the technology," in *RADAR*, 1997, no. 449, pp. 150 – 154.
[38] K. M. Chen, D. Misra, H. Wang, H. R. Chuang, and E. Postow, "An X-band microwave life-detection system," *IEEE Trans. Biomed. Eng.*, no. 7, pp. 697–701, 1986.
[39] M. Garbey, N. Sun, A. Merla, and I. Pavlidis, "Contact-free measurement of cardiac pulse based on the analysis of thermal imagery," *IEEE Trans. Biomed. Eng.*, vol. 57, no. 4, pp. 1418–1426, 2004.
[40] J. Fei and I. Pavlidis, "Thermistor at a distance: unobtrusive measurement of breathing.," *IEEE Trans. Biomed. Eng.*, vol. 57, no. 4, pp. 988–98, Apr. 2010.
[41] W. Verkruysse, L. O. Svaasand, and J. S. Nelson, "Remote plethysmographic imaging using ambient light," *Opt. Express*, vol. 16, no. 26, pp. 21434–45, Dec. 2008.
[42] C. Takano and Y. Ohta, "Heart rate measurement based on a time-lapse image.," *Med. Eng. Phys.*, vol. 29, no. 8, pp. 853–7, Oct. 2007.
[43] M. Z. Poh, D. J. McDuff, and R. W. Picard, "Non-contact, automated cardiac pulse measurements using video imaging and blind source separation.," *Opt. Express*, vol. 18, no. 10, pp. 10762–74, May 2010.
[44] M. Z. Poh, D. J. McDuff, and R. W. Picard, "Advancements in Noncontact, Multiparameter Physiological Measurements Using a Webcam," *IEEE Trans. Biomed. Eng.*, vol. 58, no. 1, pp. 7–11, Jan. 2011.
[45] H. Monkaresi, R. A. Calvo, and H. Yan, "A Machine Learning Approach to Improve Contactless Heart Rate Monitoring Using a Webcam," *IEEE J. Biomed. Heal. Informatics*, vol. 18, no. 4, pp. 1153–1160, 2014.
[46] S. D'Mello and J. Kory, "Consistent but modest: a meta-analysis on unimodal and multimodal affect detection accuracies from 30 studies," in *Proceedings of the 14th ACM international conference on Multimodal interaction*, 2012, pp. 31–38.
[47] R. A. Calvo and S. D'Mello, "Affect Detection : An Interdisciplinary Review of Models , Methods , and Their Applications," *IEEE Trans. Affect. Comput.*, vol. 1, no. 1, pp. 18–37, 2010.
[48] M. F. Valstar, M. Mehu, B. Jiang, M. Pantic, and K. Scherer, "Meta-Analysis of the First Facial Expression Recognition Challenge.," *IEEE Trans. Syst. Man. Cybern. B. Cybern.*, vol. 42, no. 4, pp. 966–979, Jun. 2012.
[49] P. Ekman and W. V. Friesen, *Facial Action Coding System: A Technique for the Measurement of Facial Movement*. Palo Alto, CA: Consulting Psychologists Press, 1978.
[50] M. F. Valstar, M. Mehu, B. Jiang, M. Pantic, and K. Scherer, "Meta-Analysis of the First Facial Expression Recognition Challenge.," *IEEE Trans. Syst. man, Cybern. Part B, Cybern.*, vol. 42, no. 4, pp. 966–979, Jun. 2012.
[51] M. Pantic and L. J. M. Rothkrantz, "Facial action recognition for facial expression analysis from static face images," *Syst. Man, Cybern.*, vol. 34, no. 3, pp. 1449–1461, 2004.
[52] M. Pantic and I. Patras, "Dynamics of Facial Expression: Recognition of Facial Actions and Their Temporal Segments," *IEEE Trans. Syst. Man. Cybern.*, vol. 36, no. 2, pp. 433–449, 2006.
[53] M. F. Valstar and M. Pantic, "Combined Support Vector Machines and Hidden Markov Models for Modeling Facial Action Temporal Dynamics," in *Lecture Notes on Computer Science*, 2007, vol. 4796, pp. 118– 127.

[54] M. S. Bartlett, G. Littlewort, M. Frank, C. Lainscsek, I. Fasel, and J. Movellan, "Fully Automatic Facial Action Recognition in Spontaneous Behavior," in *7th IEEE International Conference on Automatic Face and Gesture Recognition (FGR06)*, 2006, pp. 223–230.
[55] G. Guo and C. R. Dyer, "Learning from examples in the small sample case: face expression recognition.," *IEEE Trans. Syst. man, Cybern. Part B*, vol. 35, no. 3, pp. 477–88, Jun. 2005.
[56] K. Anderson and P. W. McOwan, "A real-time automated system for the recognition of human facial expressions.," *IEEE Trans. Syst. man, Cybern. Part B*, vol. 36, no. 1, pp. 96–105, Feb. 2006.
[57] S. Lucey, A. B. Ashraf, and J. F. Cohn, "Investigating Spontaneous Facial Action Recognition through AAM Representations of the Face," in *Face Recognition*, no. June, K. Delac and M. Grgic, Eds. I-Tech Education and Publishing, 2007, pp. 275–286.
[58] G. Littlewort, M. S. Bartlett, I. Fasel, J. Susskind, and J. Movellan, "Dynamics of facial expression extracted automatically from video," *Image Vis. Comput.*, vol. 24, no. 6, pp. 615–625, Jun. 2006.
[59] Y. Tian, T. Kanade, and J. F. Cohn, "Evaluation of Gabor-Wavelet-Based Facial Action Unit Recognition in Image Sequences of Increasing Complexity," in *Fifth IEEE International Conference on Automatic Face and Gesture Recognition*, 2002, pp. 229 – 234.
[60] S. J. Mason and A. P. Weigel, "A Generic Forecast Verification Framework for Administrative Purposes," *Mon. Weather Rev.*, vol. 137, no. 1, pp. 331–349, Jan. 2009.
[61] S. D'Mello and C. Mills, "Emotions while writing about emotional and non-emotional topics," *Motiv. Emot.*, no. In press, Apr. 2013.
[62] S. D'Mello and C. Mills, "Emotions while writing about emotional and non-emotional topics," *Motiv. Emot.*, vol. 38, no. 1, pp. 140–156, Apr. 2013.
[63] M. Mahmoud, T. Baltrusaitis, P. Robinson, and L. D. Riek, "3D Corpus of Spontaneous Complex Mental States," in *Affective Computing and Intelligent Interaction*, S. D'Mello, A. Graesser, B. Schuller, and J.-C. Martin, Eds. Springer, 2011, pp. 205–214.
[64] P. Ekman, W. V. Friesen, and J. C. Hager, *Facial Action Coding System*. A Human Face, 2002.
[65] J. Ahlberg, "CANDIDE-3 -- an updated parameterized face," Sweden, 2001.
[66] "Kinect For Windows SDK - Face Tracking." [Online]. Available: http://msdn.microsoft.com/en-us/library/jj130970.aspx. [Accessed: 30-Jul-2013].
[67] T. Ojala, M. Pietikäinen, and D. Harwood, "A comparative study of texture measures with classification based on featured distributions," *Pattern Recognit.*, vol. 29, no. 1, pp. 51–59, 1996.
[68] C. Shan, S. Gong, and P. W. McOwan, "Facial expression recognition based on Local Binary Patterns: A comprehensive study," *Image Vis. Comput.*, vol. 27, no. 6, pp. 803–816, May 2009.
[69] G. Zhao and M. Pietikäinen, "Dynamic texture recognition using local binary patterns with an application to facial expressions," *IEEE Trans. Pattern Anal. Mach.*, vol. 29, no. 6, pp. 915–928, 2007.
[70] H. Monkaresi, M. S. Hussain, and R. A. Calvo, "A Dynamic Approach for Detecting Naturalistic Affective States from Facial Videos during HCI," in *AI 2012: Advances in Artificial Intelligence*, M. Thielscher and D. Zhang, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 170–181.
[71] M. Castrillón, O. Déniz, C. Guerra, and M. Hernández, "ENCARA2: Real-time detection of multiple faces at different resolutions in video streams," *J. Vis. Commun. Image Represent.*, vol. 18, no. 2, pp. 130–140, Apr. 2007.
[72] H. Monkaresi, R. A. Calvo, and M. S. Hussain, "Automatic natural expression recognition using head movement and skin color features," in *Proceedings of the International Working Conference on Advanced Visual Interfaces (AVI '12)*, 2012, pp. 657–660.
[73] M. P. Tarvainen, P. O. Ranta-aho, and P. A. Karjalainen, "An advanced detrending method with application to HRV analysis," *IEEE Trans. Biomed. Eng.*, vol. 49, no. 2, pp. 172–175, 2002.
[74] P. Comon, "Independent component analysis, a new concept?," *Signal Processing*, vol. 36, no. 3, pp. 287–314, 1994.
[75] J. F. Cardoso, "High-order contrasts for independent component analysis.," *Neural Comput.*, vol. 11, no. 1, pp. 157–92, Jan. 1999.
[76] I. Kononenko, "Estimating attributes: Analysis and extensions of RELIEF," in *Machine Learning: ECML-94*, F. Bergadano and L. De Raedt, Eds. Springer, Berlin Heidelberg, 1994, pp. 171–182.
[77] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, Jun. 2011.
[78] G. Holmes, A. Donkin, and I. H. Witten, "WEKA: a machine learning workbench," in *Proceedings of the 1994 Second Australian and New Zealand Conference on Intelligent Information Systems,1994*, 1994, pp. 357 –361.
[79] L. . Jeni, J. F. Cohn, and F. de la Torre, "Facing Imbalanced Data–Recommendations for the Use of Performance Metrics," in *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction (ACII)*, 2013, pp. 245–251.
[80] Y. Hsu, Y. Lin, and W. Hsu, "Learning-based heart rate detection from remote photoplethysmography features," in *IEEE International Conference on Acoustic, Speech and Signal Processing*, 2014, pp. 4433–4437.

Hamed Monkaresi is a researcher at the Positive Computing lab at the University of Sydney. He completed his PhD in 2014 at the University of Sydney. He also received his MSc in Information Technology from Amirkabir University of Technology (Tehran Polytechnic). His research interests are in affective computing and computer vision. More specific interests include Facial Expression Recognition, and using machine learning techniques for understanding complex mental states.

Nigel Bosch is a PhD candidate in the department of Computer Science at the University of Notre Dame, where he has been in the Emotive Computing research group since June 2012. His research interests involve affect detection from video-based signals in learning contexts and studying the affective component of learning in computer programming students.

Rafael Calvo is a Professor in the School of Electrical Engineering and Information Engineering, at the University of Sydney. He has a PhD in Artificial Intelligence applied to automatic document classification (e.g., web site classification). He has taught at several Universities, high schools and professional training institutions. He worked at the Language Technology Institute at Carnegie Mellon University, at Universidad Nacional de Rosario (Argentina) and on sabbaticals at The University of Cambridge and University of Memphis. Rafael also has worked as an Internet consultant for projects in Australia, Brazil, the US and Argentina. Rafael is the recipient of 5 teaching awards for his work on learning technologies, and the author of two books and many publications in the fields of learning technologies, affective computing and computational intelligence. Rafael is Associate Editor of the IEEE Transactions on Learning Technologies and of IEEE Transactions on Affective Computing and Senior Member of IEEE.

Sidney D'Mello is an Assistant Professor in the departments of Computer Science and Psychology at the University of Notre Dame. His interests include affective computing, artificial intelligence, human-computer interaction, speech recognition, and natural language understanding. He has published over 180 journal papers, book chapters, and conference proceedings in these areas. He is an associate editor for IEEE Transactions on Affective Computing and IEEE Transactions on Learning Technologies. D'Mello received his PhD. in Computer Science from the University of Memphis in 2009.