This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2021.3127692, IEEE Transactions on Affective Computing

Multimodal Engagement Analysis from Facial Videos in the Classroom

Ömer Sümer, Patricia Goldberg, Sidney D'Mello, Peter Gerjets, Ulrich Trautwein, and Enkelejda Kasneci

Abstract—Student engagement is a key component of learning and teaching, resulting in a plethora of automated methods to measure it. Whereas most of the literature explores student engagement analysis using computer-based learning often in the lab, we focus on using classroom instruction in authentic learning environments. We collected audiovisual recordings of secondary school classes over a one and a half month period, acquired continuous engagement labeling per student (N=15) in repeated sessions, and explored computer vision methods to classify engagement from facial videos. We learned deep embeddings for attentional and affective features by training Attention-Net for head pose estimation and Affect-Net for facial expression recognition using previously-collected large-scale datasets. We used these representations to train engagement classifiers on our data, in individual and multiple channel settings, considering temporal dependencies. The best performing engagement classifiers achieved student-independent AUCs of .620 and .720 for grades 8 and 12, respectively, with attention-based features outperforming affective features. Score-level fusion either improved the engagement classifiers or was on par with the best performing modality. We also investigated the effect of personalization and found that only 60 seconds of person-specific data, selected by margin uncertainty of the base classifier, yielded an average AUC improvement of .084.

Index Terms—Affective computing, computer vision, educational technology, nonverbal behaviour understanding.

1 Introduction

When are students engaged in learning during a class? What is the relationship between student engagement and the content and the quality of the learning material? And, how is student engagement related to learning outcomes and long-term learning goals? These research questions and more have drawn the interest of scientists from educational sciences, psychology, and similar fields to investigate student engagement during learning. We advance this research using computational methods.

To begin our investigation of student engagement, we must first define the term engagement and contextualize its implications in the classroom setting. Several dictionaries share a similar definition of the term engagement; being engaged means "to involve oneself or become occupied; to participate" while engagement can be defined as "[being] actively committed". As it relates to human behavior, engagement is highly connected to commitment and involvement. In the educational context, student engagement has been the subject of research for the past three decades. This includes different attempts to define the term [1].

The definition by Fredricks et al. [2] is one of the most accepted and frequently used in education research. They define engagement as a multidimensional construct composed of three dimensions: *behavioral*, *cognitive*, and *emotional*. Those dimensions do not reflect isolated processes, but rather dynamically interrelated factors within an individual student. In the context of classroom and learning activities, behavioral engagement focuses on the act of participation and can include behaviors such as displaying attention and concentration, or asking questions. Emotional engagement encompasses affective reactions such as a student's interest or boredom. Whereas aspects of behavioral and emotional engagement are typically externally observable, cognitive engagement incorporates less overt, internal cognitive processes such as psychological resource investments in learning and self-regulation [2]. Importantly, previous research has found positive correlations between aspects of student engagement and academic achievement, emphasizing its central role in classroom learning [3]. To put it differently, students' engagement during classroom instruction determines the extent to which students learn, how well they develop intellectual skills, and how long they will persist in school [4]. Given its importance, in the present study, we aim to use affective computing techniques to measure student engagement in authentic classrooms based on visible indicators.

Two methods are proposed in affective computing literature to acquire the engagement labels needed for supervised learning: 1) self-reports and 2) external behavior observations. Self-reports are practical, relatively cheap, and easy to administer to a large sample, making them valuable for the measurement of engagement and beyond [1]. Despite their value, self-reports have certain drawbacks, namely a dependence on participant compliance and diligence [5]. Furthermore, self-reports can be used in two ways: after the lesson or multiple times as experience sampling. The

- *Ö. Sümer, P. Goldberg, and U. Trautwein are with the Hector Research Institute of Education Sciences and Psychology, University of Tübingen, Tübingen, 72072, Germany, Ö. Sümer is also with the Department of Computer Science, University of Tübingen, Tübingen, 72076, Germany.*
- *Sidney K. D'Mello is with the Institute of Cognitive Science and the Department of Computer Science, University of Colorado Boulder, Boulder, CO 80309. E-mail: sidney.dmello@colorado.edu*
- *P. Gerjets is with the Leibniz-Institut für Wissensmedien, Tübingen, 72076, Germany. E-mail: p.gerjets@iwm-tuebingen.de.*
- *E. Kasneci is with the Department of Computer Science, University of Tübingen, Tübingen, 72076, Germany. E-mail: enkelejda.kasneci@uni-tuebingen.de.*

former can lead to biases in retrospective recall, the latter risks disrupting the natural flow of instruction. To cause as little disruption as possible, self-reports administered in experience sampling studies have to be short. There is a risk, however, that shorter self-reports may not adequately cover the construct under investigation (more details can be found in [6]).

External behavior observations are another useful assessment tool for student engagement and have a long tradition in education research. External behavior observations have been used to investigate determinants of classroom processes such as quality of instruction [7], teacher-student relationships (e.g., [8]), the number of learning opportunities (e.g., [9]), and a teacher's choice of practices (e.g., [10]). In general, observer ratings are systematic approaches that aim to identify and interpret certain behaviors [11]. Their deployment in large-scale studies is notably limited by the necessity of providing human raters with specialized training, the difficulty of acquiring reliable labeling, and the cost involved. Moreover, in contrast to many other computer vision applications, crowdsourcing is not a viable option to label student engagement collected in authentic classrooms due to privacy considerations and the specialized training that is required for raters.

Self-reports and external behavior observations pose a challenge for large samples of classrooms. A solution is to automatically estimate engagement using machine learning and computer vision. Automated methods have two main advantages: they are fast and they have the potential to increase the sample size of classroom studies. In the field of affective computing, initial studies aimed at estimating student engagement focused on computer-based learning [12], [13], [14] and intelligent tutor systems (ITS) [15]. From ITS log files, such as students' reaction times, errors, and performance [16], [17], [18], preferred modalities for engagement analyses shifted to video [12], [13], [19], audio, and physiological measures (i.e., galvanic skin response [20], [21], EEG [22], [23], heart rate [24]).

In computer-based learning settings, the availability of log data is an important asset [25]. Furthermore, vision-based features can be extracted reliably using webcams. In the classroom, however, log data is typically unavailable and using sensors for each student can render studies expensive, intrusive, and ultimately may affect student behaviors. Thus, a widely accepted practice in classrooms is to record the instruction with field cameras located in the corners of the room. One drawback of this approach, however, is that audio and visual data is noisy and may be occluded, a challenge that we address in the present work.

1.1 Contributions of the Study

Although automated engagement analysis is widely studied in computer-based settings such as intelligent tutors and educational games, this study is, to our knowledge, one of the first to perform video-based engagement classification in the classroom on a large scale. In this paper, we review, in detail, engagement measurement studies in the field of affective computing. We then discuss the large-scale school study we conducted by collecting audio-visual recordings of classes during a one and a half month period. Observer ratings of student engagement were acquired using an instrument previously validated in university-level seminars [26].

The current study's primary focus is to develop engagement classification from limited and unconstrained data where traditional face alignment and facial action unit estimation methods have largely failed. Following the definition by Fredricks et al. [2], we focus on behavioral and emotional aspects of student engagement because they have observable behavioral correlates [27]. Visual attention (subsequently referred to as attention) and affective expressions can thus provide useful insight into these two subcomponents of engagement. Accordingly, we propose learning attention and affect features from two convolutional neural networks that we trained on head pose estimation and facial expression recognition as pretasks. In contrast to previous work that utilized handcrafted features in engagement analysis, the deep learning-based representations we propose work without precise facial alignment. Our engagement classification is performed using these learned feature embeddings. We also applied feature and score level fusion on these features. Beyond reporting baseline results using person-independent classification, we investigated personalization to address intrapersonal variation in student (dis)engagement as well.

2 Related Work

In recent years, the use of automated methods in classroom behavior analysis and engagement estimation has been on the rise. The popularity of such methods is largely due to the availability of big data and the progress of artificial intelligence. Notably, developments in deep learning have yielded significant results in social signal processing problems [50], [51], [52], [53], including classroom and learning analytics [54], [55].

We can categorize the literature of automated engagement estimation based on the following criteria:

- learning situation (computer-enabled settings, traditional classroom instruction and group-work, etc.)
- nonverbal features (various behavioral cues related to learning activities)
- computational methodology (in both feature extraction and machine learning)
- final objectives (explanation [i.e., showing a statistical relation] vs. fully automated predictive system for psychologically valid measurements of engagement)

In addition to these points, another consideration is the use of sensors [56]. Whereas sensor-free measurements depend on educational systems' log files, sensor-based measurements use physical devices such as physiological sensors (i.e., EDA, EEG, heart rate sensors) and audiovisual recordings acquired from cameras and voice recorders. As our motivation is to measure engagement as seamlessly as possible without necessitating any expensive and intrusive sensors, we limit our scope to engagement analysis using only visual modalities. Table 1 summarizes the literature of automated engagement analysis across three domains: classroom, computer-based settings (including intelligent tutors and screen-based learning games), and human-human, human-robot interactions (HHI/HRI).

TABLE 1: Automated Engagement Analysis in Classroom, Computer-based Learning, Human-Human/Human-Robot Interaction (HHI/HRI) Settings

Reference	Setting	Behavioral Cues	Engagement Measurement	Predictive Models
[28]	classroom	head pose	observer reports	✓
[29], [30]	classroom	head pose, body motion	self-reports (in-class)	✗
[31], [32]	classroom	head pose, gaze, facial expressions, posture	observer reports	✓
[33]	classroom	gaze mapping (heads up/down)	–	✗
[34]	classroom	head pose, gaze, FACS action units	observer reports	✓
[35]	classroom	real-time monitoring system capable of extracting many behavioral features (i.e. smile detector, hand raising, head pose, speech analysis)	–	✗
[36]	classroom	monitoring system (head pose and gaze estimation)	–	✓
[37]	computer-based	FACS action units and ITS log features	observer ratings	✓
[38]	computer-based	FACS action units	self-reports (user engagement survey [39], NASA-TLX [40])	✓
[12]	computer-based	handcrafted features from faces	observer reports	✓
[19]	computer-based	FACS action units and appearance features	self-/observer reports (MW)	✓
[13]	computer-based	FACS action units and gross body movement	observer reports (BROMP [41])	✓
[42]	computer-based	Kinect Animation Units, facial appearance, heart rate estimated from face videos	self-reports (concurrent & retrospective)	✓
[43]	computer-based	facial appearance features	crowdsourcing	✓
[44]	computer-based	head pose and gaze direction	observer reports	✓
[14]	computer-based	facial expressions, head pose, learning management system log data	observer reports	✓
ELEA [45]	HHI	–	observer ratings	✗
RECOLA [46]	HHI	–	self-reports	✗
MHHRI [47]	HHI & HRI	audio, physiological, and first-person vision	self-reports	✓
[48], [49]	HRI	facial expressions, body pose, audio (in children's storrytelling and therapy with robots)	–	✓

2.1 Learning Analytics in the Classroom

Despite the popularity of computer-based learning technologies, Intelligent Tutor Systems (ITS), and Massive Online Open Courses (MOOC), traditional classroom-based learning is still the dominant setting for primary through tertiary education. The popularity of classroom-based learning can be linked to the importance of sociological factors and collaboration throughout the learning process [57], [58]. For this reason, analytical tools in the classroom that measure students' learning-related behaviors and affective and cognitive engagement can play an essential role in research aiming to investigate and improve the efficiency of classroom-based learning.

Learning analytics methods in the classroom can include video cameras in the corner of the room, direct recordings of students' faces and upper bodies, and external audio recorders. The quality of audio-visual feature extraction, in general, is not as fine-grained as in computer-based situations where a webcam, 1-2 meters away, captures a student's behaviors. However, classroom analytics can provide more insight into student-teacher, student-learning material, and student-student interactions than analysis focused on individual students.

Bidwell and Fuchs [28] presumably proposed the first classroom monitoring system capable of analyzing student engagement. Although their technical report did not incorporate any quantitative results, they defined a general workflow for classroom analytics by using several color and Kinect depth-sensing cameras during a lesson in a third-grade classroom. Three observers attended the lesson and coded each student's behavior using a mobile device during 20 second intervals according to the following categories: appropriate (engaged, attentive, and transition) and inappropriate (non-productive, inappropriate, attention-seeking, resistant, and aggressive). Due to the limitations of only recording a single lesson and collecting highly imbalanced data, Bidwell and Fuchs used a Hidden Markov Model (HMM) to classify three categories (engaged, attentive, and transition) from head pose based gaze-target mappings.

A more recent classroom monitoring system was proposed by Raca and Dillenbourg [29]. Their study proposed the use of student's motion information during class and behavioral synchronization between neighboring students' feature representation to estimate student attention. In [30], they handcrafted several features such as the row in which the student sat, the amount of still time (where head pose does not change significantly for a period), and head travel (normalized head pose change). As ground truth labels of attention, Raca and Dillenbourg used self-reports that students completed in approximately 10-minute intervals. These features, together with a Support Vector Machines (SVM) classifier, performed up to the accuracy of 61.86% (Cohen's $\kappa = 0.30$) to predict 3-scale attention (low, medium, and high). Their work showed that student attention can be automatically measured using visible behavioral cues. However, they used considerably long intervals (10 minutes) before self-reports were obtained and the person-independent evaluation was not clearly stated in their experiments. Moreover, they only employed attentional features (head pose and motion), and did not utilize any affective or behavioral nonverbal features.

Zalatelj and Kosir [31], [32] used a Kinect sensor and its commercial SDK to estimate body pose, facial expressions, and gaze. Subsequently, they computed behavioral cues (i.e., yawning, taking notes, etc.) from Kinect features and trained a bagged decision tree classifier to estimate observer-rated attention levels (low, medium, high). They also used manually-labeled behavioral features (i.e., writing, yawning, one's hand touching their head). [32] only analyzed a few minutes of video recording from datasets with 3 students, raising questions about generalizability. In [31], they increased the number of students to 18 (only 25-minute recordings per student) and reported an accuracy of 75.3% in three-level engagement classification. Even though they achieved moderate performance in the task, nearly one-fifth of their data had to be discarded due to the sensor's failure to accurately track face and body features. The range of Kinect and similar depth sensors is around 0.4 to 5.45 meters [59], and the ideal range for face alignment and body pose estimation is even less (i.e., Kinect One was used at 1.8 meters in [31]). This suggests that multiple sensors are required in a typical classroom with 20-30 students, potentially introducing additional cost and posing device synchronization issues.

Thomas and Jayagopi [34] collected video recordings of 10 students in three 12-minute segments while they were listening to motivational video clips on YouTube. Three observers labeled the engagement of each student in 10-second intervals based on whether a student was looking towards the screen (teacher area), talking to a neighbor, or gazing in another direction. Their approach was to use head pose, gaze direction, and facial action unit features using OpenFace [60] and classifiers such as SVM and logistic regression. They reported up to an AUC of .83 in two-class engagement classification. The main limitation of this study was the limited data size and concerns about the engagement labeling methodology. Specifically, students can still be attentive to the audio content while looking elsewhere or taking notes.

Goldberg et al. [26] is the first study that utilizes a psychologically valid and comprehensive engagement rating system. Their continuous observer-based rating system combines Chi & Wylie's ICAP (Interactive, Constructive, Active, Passive) framework [61] and on-task/off-task behavior analysis [62]. Using attentional (head pose and gaze direction) and affective (FACS action unit intensities) sets of features with support vector regression, they predicted continuous observer-ratings and reported correlations between estimated engagement levels and self-reports collected at the end of 40-minute teaching units (N=52). They also found that behavioral synchrony with immediate neighbors improved the estimation of engagement and reported regression results up to a Pearson correlation of .71 with the ground truth labels.

One of the main objectives of learning analytics in the classroom is the reporting of students' estimated attention and engagement to teachers. For instance, Fujii et al. [33] estimated head-down (i.e., taking notes or reading learning material) and head-up (gazing at whiteboard/teacher area) behaviors for each student and depicted a color-coded visualization for teachers with a synchronization rate of the classroom in terms of predicted classes, head-up and head-down. However, they tested the performance of the head-down/head-up detector on limited data (average accuracy of 89.8% in 30 minutes of video recordings and 5 students). Additionally, only reporting behavioral cues (looking at learning material or the teacher area) provides limited information on students engagement levels.

In a similar vein, two recent studies [35], [36] developed smart classroom monitoring systems. While Anh et al. [36] mapped gaze directions focused on three areas (board/teacher, table/notebook, and other directions) and visualized the distribution on a dashboard, Ahuja et al. [35] integrated various nonverbal features in their smart classroom, EduSense. These features included the state-of-the-art methods in face detection and alignment, body pose estimation, hand raise detection, and active speaker detection. [35] presented a technical analysis of real-time classroom monitoring systems, including the speed and latency of the system and their algorithms' performance. However, they did not report on student engagement. Even though nonverbal features are essential to understand engagement, they are not easy for a teacher to interpret on their own.

In summary, computer vision-based classroom analytics studies, though emerging, are still limited. The sample sizes are small and the majority do not estimate attention or engagement levels. For the studies that do estimate student attention/engagement, there remain concerns about the validity of the engagement labels.

2.2 Engagement Estimation in Computer-based Learning

Computer-based learning situations are more restricted than classroom situations because they only contain student-technology interactions. These studies generally capture video and audio from 1 to 2 meters away, resulting in better quality data for feature extraction methods. Furthermore, introducing an intervention during learning is more straightforward than in the classroom setting. For these reasons, automated engagement estimation is more prevalent in computer-based situations such as students playing an educational game, engaging in reading comprehension or writing tasks, or learning with ITSs (see [25], [56] for a review).

One study that predicted the level of engagement in computer-based settings (during which the participants perform a cognitive training task) was conducted by Whitehill et al. [12]. They used appearance-based facial features (Box filters, Gabor filters, CERT FACS features) and estimated levels of engagement using several classifiers such as GentleBoost, SVM, and multinomial logistic regression. They developed a manual rating system (4-scales) and annotated the video recordings at 60-sec or 10-sec intervals. The accuracy of their classifiers varied between 36-60%. Even though they conducted one of the initial studies, their setting was limited to interaction with an educational game. Their ratings entailed the assignment of single labels to the 3-minute videos and lacked continuous labels which can provide more precise information about student engagement levels.

In a similar computer-based setting, Monkarasi et al. [42] estimated engagement using Kinect face tracker ANimation Unit (ANU) features, LBP-TOP, and heart rate (estimated from videos of the face) features during a writing task. They used concurrent self-reports (every 2-minutes during the writing task) and retrospective self-reports after the participants finished the task. Both self-reports showed high correlation ($r = 0.82$, $p \leq 0.001$), and the engagement classification (low/high) achieved an AUC of .758 and .733 using concurrent and retrospective labels, respectively. Their study was based on Kinect sensors and handcrafted features such as local binary patterns (LBP) and statistical measures derived from facial landmarks. Thus, data capture was more sensitive to data loss because of head motion or face occlusion. Bosch et al. [13] used estimated FACS action units as features and predicted observer annotations [41] using different classifiers (Bayes Net, Updateable Naive Bayes, Logistic Regression, AdaBoost, Classification via Clustering, and LogitBoost). The six variables they predicted according to [41] are boredom, confusion, delight, engagement, frustration, and off-task. Engagement classification performed an AUC of .679 and 64% of accuracy.

Aslan et al. [14] developed a real-time engagement estimation system using facial features (facial landmarks, head pose) and learning management system log data (content duration, number of hints, difficulty level). They tested the usability of the system as an assistive tool for teachers. Their engagement estimation was in two dimensions: behavioral (on-task vs. off-task) [63] and emotional (satisfied, bored, confused) [64]. On-task/off-task engagement classification performed up to an F1 score of 80.09%, whereas their emotional engagement classifier had an F1 score of 48.12% to 89.30% in person-independent and person-specific settings, respectively. Including [14], previous studies in computer-based settings adopted low-dimensional geometric and appearance descriptors. In contrast, engagement analysis based on learned representations can be improved by learning better features.

Mind wandering (MW) is an important attentional component of engagement, defined as an attentional shift from task-related to task-unrelated thoughts [65]. It is consistently linked to negative performance in learning tasks [66]. The availability of automated methods to detect MW can reveal this covert aspect of engagement. The use of visual modalities, particularly facial videos, to detect MW is preferable to eye gaze [67] and physiological signals [68], which necessitate specialized sensors. Stewart et al. [69] is the first study that used visual modality, facial action units, and body motions to detect MW. They recorded facial videos while the participants watched a narrative film for 35 minutes. Each participant self-reported MW by pressing a key through the video screening. Facial action unit features and classifiers including logistic regression, naive Bayes, and support vector machines could spot MW in a person-independent setting with an F_1 score of .390. Later, [70] showed the generalizability of MW detection when trained and tested on different tasks (reading scientific text and watching a narrative film). Bosch et al. [71] showed the applicability of video-based MW detection in a classroom study (N=135) while learning from an intelligent tutor system. Even though mind wandering is related to behavioral disengagement and off-task behaviors, it involves internal thoughts, and there are fewer overt behavioral cues associated with mind wandering. In this study, we mostly limit our scope to visible engagement.

2.3 Human-Human and Human-Robot Interactions (HHI/HRI)

Another line of work is the attention analysis in human-human interactions (e.g., in group work) and in human-robot interactions. For example, Sanches-Cortes et al. [45] developed an audiovisual corpus of groups of four who engaged in a survival task and focused on estimating group performance, apparent personality, and perceived leadership and dominance. Similarly, Rinvegal et al. [46] used a survival task during remote collaboration using audio, video, and physiological signals as well as self-reported engagement. However, although survival tasks can be useful to measure group interactions, they do not represent typical learning situations which is the current focus.

Looking into more recent studies, Celiktutan et al. [47] collected an audiovisual dataset during human-human and human-robot interactions using first-person cameras. They acquired self-/acquaintance-assessed personality and self-reported engagement labels. However, limitations include the size of the dataset (18 participants, 6 hours, but composed of short clips and not in learning settings) and interactions wherein one participant or robot asks predefined questions. Another application in human-robot interactions is autism therapy for children [48], [72] and child-robot interactions (a dialogic storytelling task) [49], [73]. The measurement of engagement during children's storytelling or autism therapy may be more obvious. In these settings, it is comparably easier to differentiate between engaged and disengaged behaviors than in schools where most pupils learn to hide their disengagement. Furthermore, age affects attention levels [74]. Most pupils in higher grade levels learn to hide their disengagement. These factors have the potential to make visible cues between engaged and disengaged behaviors more difficult to spot. Despite the lack of expert-labeling criteria, these studies adopt a continuous engagement labeling approach and deep Q learning to actively sample training data and personalize models with limited data.

To summarize, the literature in attention and engagement analysis is centered on computer-based learning settings as well as human-human and human-robot interactions. Collecting data for automated analysis in those domains is comparably more convenient than in the classroom. However, the impact of schools and classroom instruction exceeds the scope of these applications and, moreover, plays a crucial role in every student's life. Therefore, research analyzing attention and engagement in the classroom is highly important and could benefit from novel analytic approaches. Existing classroom-based studies are very limited in terms of data size. Most were conducted on university-level courses or on a small number of participants (mainly to test computer vision systems). While Raca and Dillenbourg [30] conducted the most comprehensive attention monitoring study in the classroom and showed the applicability of these technologies in a school setting, their study lacked expert-labeled attention/engagement measures and predictive learning models on a larger scale. We build off this existing work and extend it further in the current study.

3 DATA COLLECTION FOR AUTOMATED ENGAGEMENT ESTIMATION IN THE CLASSROOM

The study was conducted during regular lessons at a secondary school in Germany over a one and a half month period. The ethics committee from the Faculty of Economics and Social Sciences of the University of Tübingen approved our study procedures (Approval #A2.5.4-097_aa), and all teachers and parents provided written consent for their students to be videotaped. Students who did not consent to being videotaped attended a parallel session covering the same instructional content.

3.1 Participants

We collected audio-visual recordings of 47 classes from 5^{th} to 12^{th} grades, resulting in 128 participants overall. Each participant attended more than one class (3.84 on average).

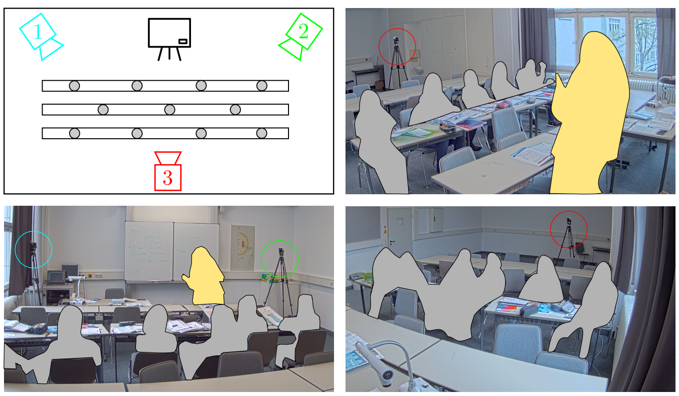

Fig. 1: Sample scene from the classroom. The synchronous cameras recorded the instruction simultaneously.

Therefore, the total number of samples across grades was over 360. The collection of labeled data for developing and benchmarking automated methods is time consuming. Thus, we identified a sub-sample of students based on their occurrence and visibility in multiple video recordings, resulting in 15 students from grade 8 (N=7) and grade 12 (N=8) in our analysis. Each participant appeared in five recordings on average (i.e., classes on different days and subjects). The total number of samples in our data was 75. Classes covered a wide range of subjects including Mathematics, Chemistry, Physics, IMP (**I**nformatics, **M**athematics, **P**hysics), History, Latin, French, German, and English.

3.2 Procedure

Before classes on the first day, students filled out a questionnaire covering demographic information (age, gender) and individual differences (BFI-2 XS, 15 items; [75]). After each session, students completed another questionnaire about their learning activities. Session recordings lasted between 30 and 90 minutes each. Video material during classes covered group work, individual work, and teacher-centered instruction. To best capture student attention on the instructor, we focused on teacher-centered components of the video (see Fig 1), extracting the main part of instruction time in intervals of 15 to 20 minutes from each recording. The intervals were manually annotated by human raters.

3.3 Self-Reported Learning Activities

After each session, we assessed students' involvement (four items, $\alpha = 0.73$; [76]), cognitive engagement (six items, $\alpha = 0.78$; [77]), and situational interest (six items, $\alpha = 0.92$; [78]) during the preceding instructional period.

3.4 Continuous Manual Annotation

To manually annotate students' observable behavior, we used a one-dimensional scale through the open software, CARMA [79], which enables continuous (1-second in our case) interpersonal behavior annotation via a joystick [80]. We combined the concept of on-task/off-task behavior [62], [81] with existing scales from the engagement literature. To define more fine-grained cues within the possible behavioral spectrum, **I**nteractive, **C**onstructive, **A**ctive, and **P**assive, we used the ICAP framework [61]. Thus, behaviors were annotated on a symmetric scale ranging from -2, indicating the disturbance of other students (i.e., interactive) or off-task behavior, to +2, indicating highly engaged, interactive, on-task behavior (see Fig 2). Values closer to 0 indicated rather unobtrusive, passive behavior. Two raters annotated the sub-set of students in all videos in random order, with inter-rater reliability ICC(2,2) for each student being 0.77 on average (absolute agreement). For subsequent analysis, the mean across the two raters was calculated for every learner in every second.

The existing observational instruments often use time samplings of 20-s intervals or longer (e.g., in [82], [83], [84], [85]). However, classroom interactions are rather dynamic and student behavior may change significantly within 20 seconds. To account for these changes and to capture the ground truth in a more fine-grained manner, we decided to acquire engagement labels per second. For more details about the manual annotation instrument, interested readers are referred to Goldberg et al. [26].

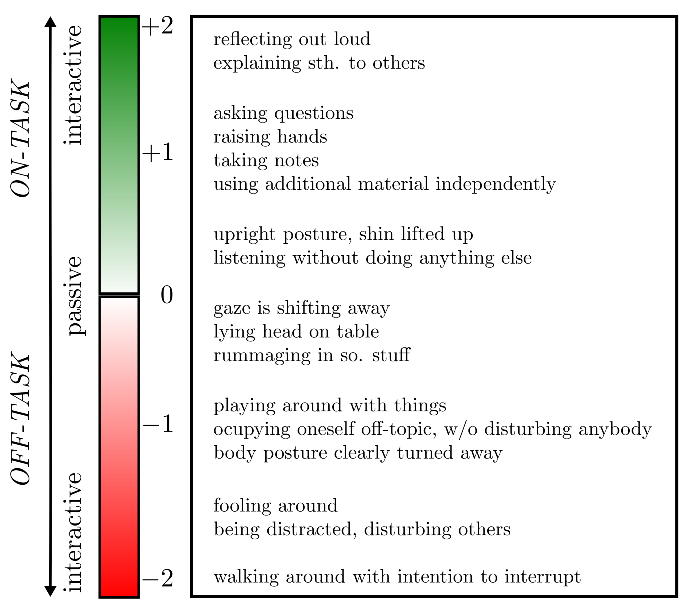

Fig. 2: Continuous scale of our manual rating instrument and visible behavioral indicators [26]

3.5 Preprocessing

For each video recording, we had three cameras as depicted in Figure 1. One camera was located in the rear part of the class covering the classroom and teacher. The other two cameras were placed on the left and right side of the teacher area (whiteboard) directed towards the class. We applied our computational pipeline to both the left and right camera and dynamically picked the stream where a particular student was more visible. Specifically, we used a single-stage face detector, RetinaFace [86], to detect all faces in the video streams. Subsequently, we picked several query face images that belonged to the students whose behaviors we intendeded to analyze. Instead of face tracking, we directly used those query images and extracted ArcFace embedding [87] for all face patches. By calculating the minimum cosine similarity between the query images and all faces, we created face tracklets for each student. Despite the challenges of occlusion and different camera angles, the face detection and recognition methods we employed could localize and recognize faces most of the time due to their training on large and unconstrained data sets. We used one-second (24 frames) continuous sequences where both face detection and recognition worked smoothly.

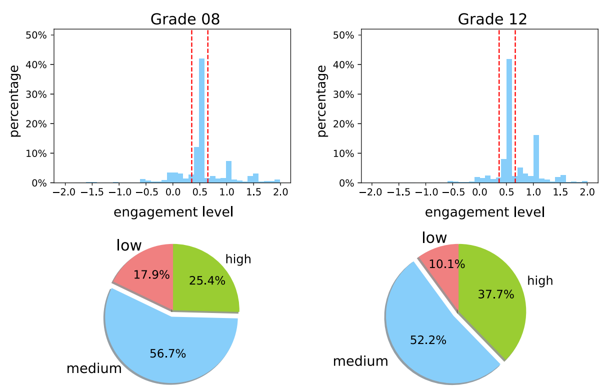

Fig. 3: The distribution of engagement labels in Grade 8 and 12. Pie charts show the percentage of quantized labels according to continuous labeling.

Table 2 shows the number of different day recordings per student and the total length of the data where preprocessing was successful. The total data length is 25,450 and 32,755 seconds for Grades 8 and 12, respectively.

In total, we collected over 15 hours of recording in 30 sessions. Compared to other classroom-based studies, the line of work by Raca & Dillenbourg [30] used four classes in 9 sessions. Even though their study was on large-scale data, their attention analysis was based on 10-minute intervals and self-reports. Similarly, sample sizes of other engagement studies in the classroom are limited: three videos of 12-minute recordings in [34], 25 minutes of video recordings in [31], 4 minutes in [32].

In the continuous labeling scale, values denoting disengagement were rarely observed and the labels were often imbalanced. Thus, we followed the previous works [72], [73] that discretized the continuous scale into three groups: low [-2, 0.35], medium (0.35, 0.65], and high engagement (0.65, 2.0]. Figure 3 depicts the continuous and discrete distribution of labels in grades 8 and 12.

4 Methodology

4.1 Problem Statement

To classify student engagement level, we used video recordings of classes. Formally, we employed sequences $\mathcal{S} = \{I_1, I_2, \cdots I_N\}$ where $n = 1, \cdots, N$ denotes the time intervals of a second (24-frames). Using any of the channels, attention or affect, we extracted feature vectors from each sequence $\mathbf{x} = \{x_1, x_2, \cdots x_N\}$ with $\mathbf{x} \in \mathcal{R}^{\mathcal{T} \times M \times D_m}$. The 24-frame feature sequences are associated with engagement label $y = \{0, 1, 2\}$. When training engagement classifiers (except for LSTM models), we used the middle frame as a training sample. To predict the engagement labels, we used either the majority voting of all 24-frame predictions or a single prediction in a temporal learning model.

TABLE 2: The number of classes and the total duration of recording where face detection was successful (in seconds) per each student.

Grade 8								
student	**S4**	**S7**	**S8**	**S11**	**S13**	**S14**	**S16**	
#class	2	7	7	3	6	4	6	
seconds	836	5450	5309	2269	4404	2674	4508	
Grade 12								
student	S1	S2	S3	S4	S5	S6	S7	S8
#class	9	8	3	3	4	3	6	4
seconds	6363	6695	2662	2708	4219	2605	3844	3659

4.2 Feature Representation

In most of the classes, students were listening to the teacher instead of speaking. Due to occlusion of the students' upper bodies in many of the recordings, nonverbal features such as speech and body pose were not always available. However, faces were usually visible and computationally faster and more reliable to detect. Consequently, our analysis depends on preprocessed faces as described in 3.5.

Motivated by the fact that engagement is a multidimensional construct, we can extract two different sets of information from face images: attentional and emotional features. There are several studies in the literature that used available face processing tools such as OpenFace [60] for engagement estimation [44], [48].

The main drawback of this approach, however, is that it depends on very accurate face alignment. In a classroom setting, however, the distance of the camera to the students varies between 2-10 meters, which reduces image quality and eventually leads to poor facial keypoint localization. Hence when we processed the classroom data using OpenFace [60], it could only process approximately 30-40% of a student face with high confidence. Furthermore, even though facial action unit-based approaches provide valuable information on affect, they almost always anticipate nearly frontal images. Considering these issues, we extracted affect and attention features from a network that we trained on

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2021.3127692, IEEE Transactions on Affective Computing

discrete facial expression recognition and head pose estimation, respectively. Both networks were trained without precise face alignment using 68-point facial landmarks.

Figure 4 shows the feature learning method for affect and attention. Feature learning does not involve any educational data. In the affect branch (Affect-Net), we used one of the most unconstrained and large-scale affect datasets, AffectNet [88]. We also trained a ResNet-50 network using softmax cross-entropy loss to predict categorical models of affect (seven discrete facial expressions): neutral, happy, sad, surprise, fear, disgust, and anger. The training set of AffectNet was composed of 23,901 images whereas the validation set had 3,500 images. We aligned all facial images using five facial keypoints estimated by the face detector [86] and aligned by similarity transform to the size of 224×224. The training was done using an SGD solver with an initial learning rate of 0.1 (decayed ten times in every 30 epochs) for 100 epochs. The best accuracy on the validation set reached 58.37%. This performance is comparable to the [88]'s benchmark 58%. State-of-the-art methods that employed pyramid super resolution, label smoothing [89], and knowledge distillation [90] improved up to 60.68%, and 60.60% by using additional training data. In subsequent engagement classification experiments, we discarded the prediction of both networks and used the feature activations of the layer before the last fully connected layer of the AffectNet model for affect embedding.

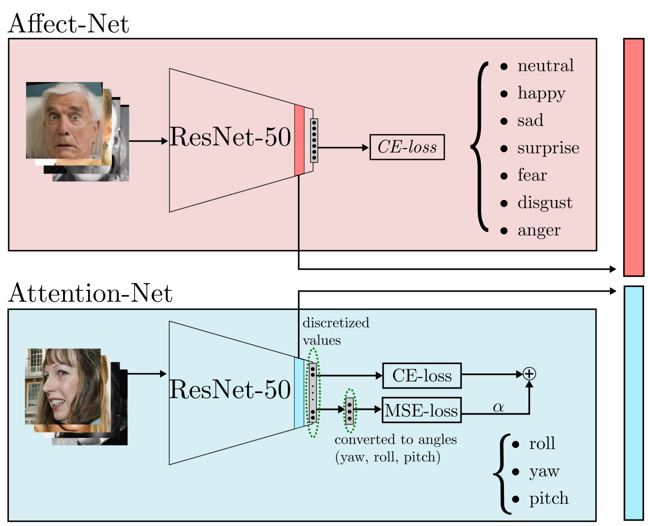

Fig. 4: Feature learning for affect and attention. Two ResNet-50 backbones are separately trained for facial expression recognition and head pose estimation. The learned features will be used subsequently for engagement estimation on classroom data.

In the attention branch (Attention-Net), we used the 300W-LP [91] dataset to train another ResNet-50 network to estimate head pose. By adopting the approach in [92], we optimized two losses jointly: softmax classification on discretized values and mean squared loss on continuous values of head pose angles (yaw, pitch, and roll). The features of the final Attention-Net model that we used performed a mean angular error of 7.36 on the AFLW2000 dataset [91]. The details of our Attention-Net training procedure are provided in the supplementary material.

The CNN-based approach is advantageous for head pose estimation because it is more robust than Perspective n-Point (PnP)-based methods that find correspondence between estimated facial keypoints on an image and their corresponding 3D locations in an anthropological face model [92]. In challenging cases where those methods fail (for instance, partial occlusion or varying angles of camera view), CNN-based methods can return satisfactory predictions and, more importantly, map the inputs in a continuous low-dimensional embedding according to poses. We do not use estimated head pose or facial expressions. We are only interested in the learning embedding feature representation to use as attentional or affective features.

AffectNet and 300W-LP datasets face many challenging situations such as various camera angles and image quality, illumination, background, and makeup. Thus, Affect-Net and Attention-Net branches trained on these datasets can learn robust feature representation to avoid these difficulties. Compared to the handcrafted appearance features such as Local Binary Patterns or Gabor filters, deep embeddings can be extracted without precise alignment and are extendable by training with new DNN architectures on more data. Training Attention-Net and Affect-Net representations on head pose estimation (300W-LP) and facial expression (AffectNet) datasets, instead of training them directly for engagement classification on classroom data, helps to avoid overfitting caused by the limited number of subjects represented in the classroom data, a major advantage of this method.

4.3 Engagement Classification

We compared different classifiers on affective and attention features. Frame-based classifiers were trained on the middle frame of each 1-second sequence to avoid redundant training samples when all frames were used. In the test phase, we retrieved predictions for all 1-second (24 frame) sequences and applied majority voting. The shallow classifiers that we used included Support Vector Machine (SVM) and Random Forests (RF). All model training and dimensionality reduction was conducted in a person-independent manner where an individual participant's data could either be in the training or test fold, but not in both. Considering the behavioral differences between grades, we built separate models for grades 8 and 12.

For SVMs, we tested both linear and radial basis function (rbf) kernels. When the data is linearly inseparable (as in engagement classification), the input feature space is projected into a high-dimensional space through an inner product transformation called a kernel function, $k(x_n, x_m) = \phi(x_n)^{\phi}(x_m)$. Subsequently, an SVM classifier is applied. Instead of explicitly calculating ϕ, we calculated all pairwise similarities in the training set. For instance, linear and rbf kernels were defined as $x_n^\intercal x_m$ and $\exp(-\gamma||x_n - x_m||^2)$. Due to the high number of samples, training SVM-based models with many instances and features (i.e., 2048-dimensional features and 20-25K training samples) can lead to storage and computational costs for kernel methods. Thus, before training SVM models, we applied Principal Component Analysis (PCA) and used the top 48 components that explain 99% of the variance in the training set. Principal

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2021.3127692, IEEE Transactions on Affective Computing

components were calculated in training sets separately. We applied the same transformation to the test set in order to preserve person-independence. For RF, we used the feature embeddings directly without dimension reduction.

For DNN's, instead of retraining the entire representation up to the first layers of the ResNet-50 architecture, we trained a Multi-Layer Perceptron (MLP). Even though the data subset that we acquired for manual annotation and used in our analysis was over 15 hours, we still faced the problem of a limited number of participants and the restricted context of in-class learning. Thus, training a network from scratch would result in overfitting and failure to recall previously learned features from a larger set of participants, a useful action for modeling engagement. We used two-layer MLPs (one for AffetcNet and another for AttentionNet features) each with an input layer of 2048 neurons and a hidden layer of 128 neurons. Training was conducted in mini-batches of 256 using soft-max cross-entropy loss and an SGD solver with a learning rate of 0.001 and a maximum 200 of iterations. In each trial, we kept a random 10% of the training data as a validation set for early stopping and preserved the person-independent setting. As with the SVM models, we applied majority voting on individual frames to acquire the prediction for 1-second sequences.

In addition to those approaches, we used a recurrent neural network model, long short-term memory (LSTM) [93], to learn temporal patterns in the data.

We provided 2048-length Attention-Net or Affect-Net embeddings as input to a two-layer LSTM network with a hidden layer size of 128. The output of the LSTM network on the last time step was fed to a fully connected layer of 64 neurons, and the entire model was trained using softmax cross-entropy loss and Adam solver [94] with a learning rate of 0.001. All LSTM models were trained for 5 epochs.

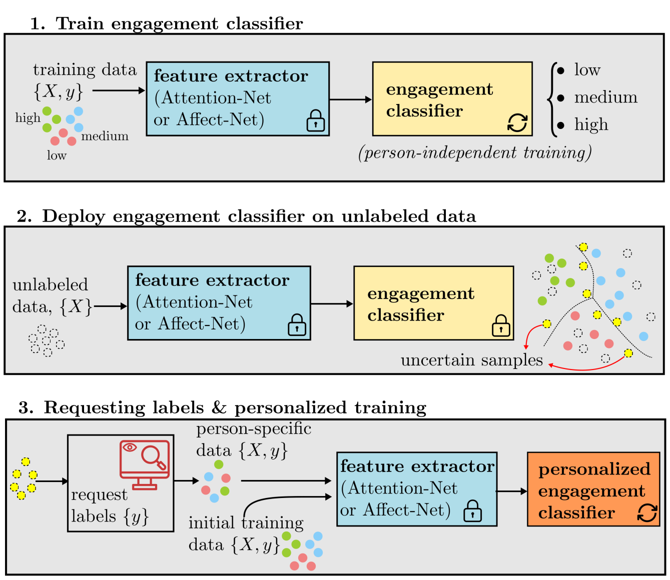

Fig. 5: Active Learning for Personalized Engagement Classification (The parameters of the parts with lock sign are kept frozen).

4.4 Personalization of Engagement Classifiers

Since engagement and disengagement during instruction can differ significantly from one student to another, engagement classifiers often benefit from personalization. Traditional active learning algorithms typically propose a single instance to label at a time. This may result in a longer waiting period for the expert labeler during the personalization phase of the engagement classifier. In contrast, we assume the labeler starts from an engagement classifier trained in a person-independent manner and labels a set of instances. For SVM-based classifiers, engagement probability can be calculated via Platt scaling whereas the mean predicted class probabilities of the trees can be used in Random Forests. MLP and LSTM classifiers provide an engagement probability output because they were trained with softmax cross-entropy loss. We used these probabilities as an uncertainty score for unlabeled instances.

In order to investigate the effect of personalization with a small amount of data, we utilized the margin uncertainty principle [95]. This principle considers the samples with the smallest margin between first and second to be the most likely class probabilities. After training a person-independent engagement classifier, we ran this classifier on person-specific data and requested labels of instances where the class probabilities between the first and second most likely predictions were the smallest. In this way, we only used the labels of instances around the separation hyperplane of low, medium, and high engagement. The margin uncertainty rule can be written as follows:

$$x^*_{marg} = \arg\min_{x} \left[P_{M_{init}}(\hat{y_1} \mid x) - P_{M_{init}}(\hat{y_2} \mid x)\right] \quad (1)$$

where $\hat{y} = P_{M_{init}}(\hat{y} \mid x)$ is the prediction with highest posterior probability, $\hat{y_1}$ and $\hat{y_2}$ are first and second most likely predictions.

Figure 5 depicts our personalization framework using batch-mode active learning. The deep feature embedding component (i.e., Attention-Net and Affect-Net) was not retrained in the personalization of engagement classifiers. Thus, these steps only updated the engagement classifier and did not require a longer training time. The first step shows the initial training of the engagement classifier. Subsequently, deep feature extractors and engagement classifiers were deployed on person-specific data. A small number of samples with high uncertainty were selected and added to the training set.

The uncertain samples with the highest margin between the first and the second most likely predictions should lie around the separating hyperplane. As an example, let's use an unlabeled sample with probability outputs of 45%, 40%, and 15% for high, medium, and low engagement categories, respectively. Even a few of these samples can help increase the separation between high and medium engagement. They can also be more effective in personalization than samples more congruous to the labeled samples and further from the separating hyperplane. Instead of a single update, we sampled a small batch of unlabeled images to label and retrained the initial model iteratively. The use of only a small number of person-specific samples should not create significant time and memory overhead for retraining engagement classifiers. With iterative updates,

TABLE 3: Performance Comparison of Engagement Classifiers on Classroom Data using Attention-Net and Affect-Net Features and Different Classifiers.

Classifier	AUROC			
	Grade-8		Grade-12	
	Attention-Net	*Affect-Net*	*Attention-Net*	*Affect-Net*
SVM (linear)	.560 ± .05	.570 ± .06	.656 ± .09	.563 ± .06
SVM (rbf)	.603 ± .05	.604 ± .03	.697 ± .07	.595 ± .08
RF	.620 ± .04	.608 ± .03	.708 ± .05	.600 ± .09
MLP	.615 ± .05	.597 ± .03	.701 ± .06	.622 ± .05
LSTM	.603 ± .05	.610 ± .04	.719 ± .05	.612 ± .09

each personalization step was applied on a day or a week of recording. The existing person-independent engagement classifier was personalized and adapted to a specific subject.

5 RESULTS

As indicated above, we performed engagement classification experiments separately in grades 8 and 12 because visual engagement across grades can vary. With the exception of the test subject, every student was used for training and the same experiment was repeated for each test student and modality (affect vs. attention) in both grades. Table 3 shows the performance of various classifiers using Attention-Net and Affect-Net features. We used weighted Area Under the ROC Curve as a performance measure in the three-level engagement classification task since it measures the performance of a classifier at different thresholds. Furthermore, it is more attune to class imbalances than to metrics such as accuracy.

Engagement classification. AUCs in general ranged from .56 to .719 with a mean of .623, which reflect an improvement from chance AUC of 0.5. The best performing unimodal classifiers used attention features and a RF in grade 8 (AUC of 0.62) and an LSTM for Grade 12 (AUC of .72).

When visual indicators were compared, Attention-Net features yielded .01 to .03 better AUC than Affect-Net for Grade 8. On the other hand, the margin between the average AUCs of Grade 12 students is more notable. Attention-net features performed .08-.11 better than Affect-Net features in Grade 12. This may be related to the easy distraction, movement, and increased gaze drifts characteristic of students in both grades. As a result, attention features were more effective than affect features in engagement classification.

Another comparison is the type of classifier used to examine engagement. In our experiments, linear SVM classifiers had the lowest performance (.03 to .06 lower AUCs). However, there were no explicit performance differences among SVM with rbf kernel, RF, and MLP classifiers across both grades and feature sets. Typically, we expect deep learning-based methods, MLP for instance, to better model engagement than shallow classifiers, but performance in this case was comparable to RF and SVM-rbf. This may be due to the limited sample size of the data, the multifaceted aspect of learning problems, and imbalances in feature and label distribution. We can intuitively argue that a better performance in initial tasks (Attention-Net in head pose estimation and Affect-Net in facial expression recognition) can eventually lead to better engagement classifiers. We investigated the relationship between initial tasks and engagement classification performance and validated it through additional experiments presented in the supplementary material.

Looking into DNN-based classifiers, the use of temporal information by LSTM classifiers negligably improved the performance of MLP only in the settings of Affect-Net/Grade 8 (+.013 in AUC) and Attention-Net/Grade 12 (+.018 in AUC). The limited improvement of LSTMs may be due to the short time window (24-frame) over one second. We adopted this approach to match the continuous engagement labeling method which generated engagement labels per second, and consequently deemed it suitable to provide real-time feedback for applications deployed in a school setting.

TABLE 4: Performance Comparison of Different Fusion Strategies using Random Forest Classifiers.

Grade	Feature Set	Avg. AUROC
8	Attention-Net	.620
8	Affect-Net	.608
8	Feature-level Fusion	**.633**
8	Score-level Fusion	.632
12	Attention-Net	**.708**
12	Affect-Net	.600
12	Feature-level Fusion	.616
12	Score-level Fusion	.694

We tested different fusion strategies using RF engagement classifiers due to their higher performance and speed and to make them comparable to the personalized models that we use as RF engagement classifiers. For feature level fusion, different feature embeddings were concatenated to train a single engagement classifier. Score level fusion averaged the probability outputs of two separate classifiers trained on Affect-Net and Attention-Net representations. Table 4 shows the performance of feature-level and score-level fusion for grade 8 and 12. For grade 8, both fusion strategies yielded comparable improvement, +.012-.013 over the best modality (Attention-Net). On the other hand, score level fusion in grade 12 was on par with the unimodal attention classifier, whereas feature-level fusion was much lower.

After reviewing the overall results and considering the difficulty of interpreting a student's level of engagement using only facial videos, we find these results to be moderate. This is notable given that the criteria for the manual annotation of engagement (as depicted in Figure 2) is not directly related to gaze direction or facial expression.

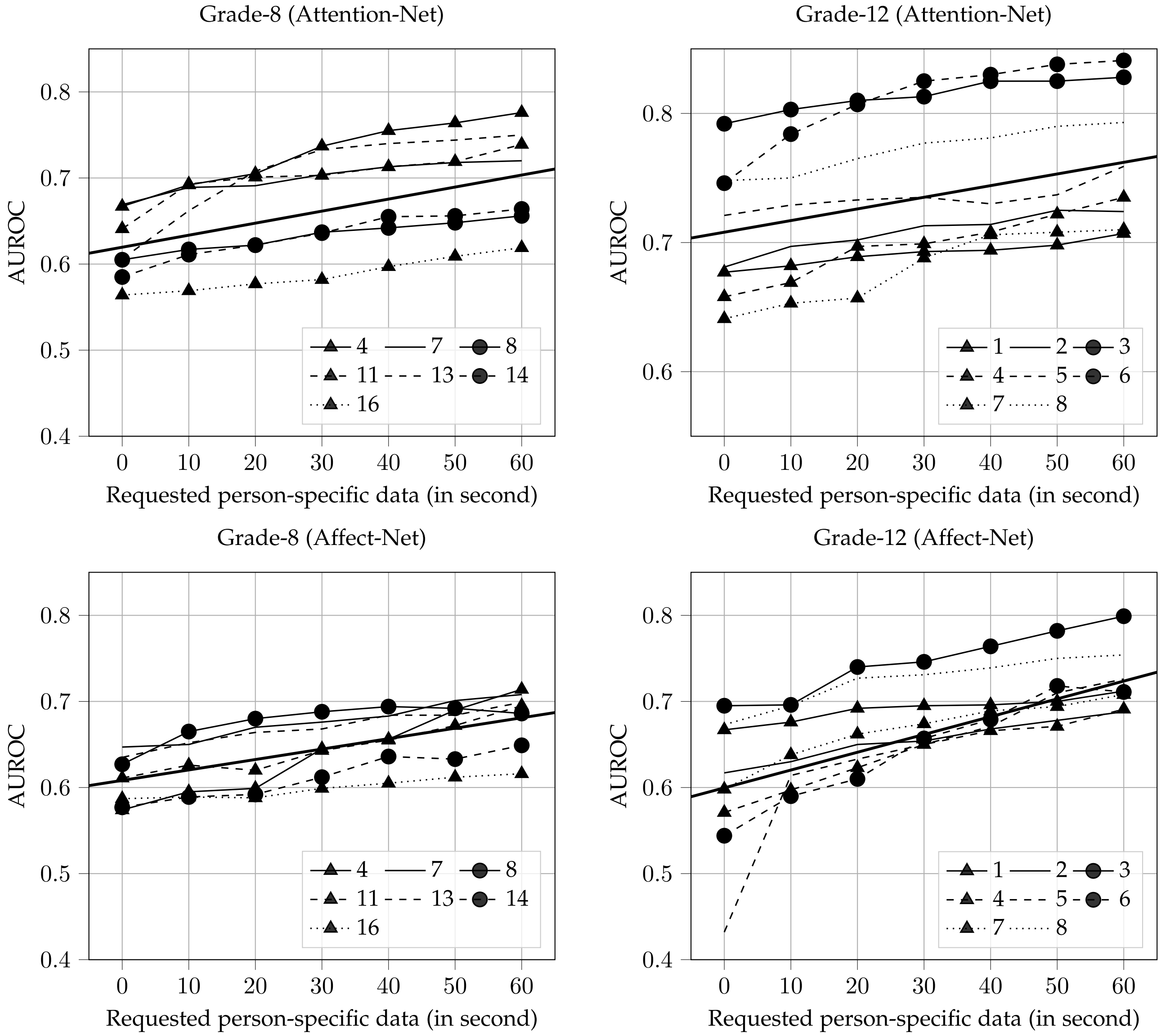

Fig. 6: The Effect of Personalization on Different Engagement Classifiers (All classifiers are based on RF. The legends show the corresponding AUC performance per student, and each thick line represents the overall trend of personalization.)

Personalized models. We selected RF classifiers for the personalized models because of their successful performance in the above person-independent experiments and their speed in training as compared to other classifiers. Readers may recall that instead of directly training and testing on person-specific data, we started with person-independent models and adapted them based on small amounts of person-specific data in a simulated active learning setting. The number of samples from each student varied (as depicted in Table 2). Thus, we limited person-specific labels requested by the oracle to 60 seconds for each student. Specifically, starting with the model person-independent model, we sampled 60 samples using different sampling strategies, and compared ROC performance to initial performance. The 60 samples were acquired after 6 steps by selecting only 10 samples per step, adapting the classifier with the new samples iteratively.

TABLE 5: Confusion Matrices for the Best Person-Independent and Personalized Models.

Method	**Actual**	**Classified**			**Priors**
(Grade 12)		*low*	*medium*	*high*	
Attention-Net, RF	*low*	.099	.442	.458	.101
	medium	.053	.735	.345	.522
	high	.075	.400	.525	.377
Attention-Net, RF (personalized)	*low*	.185	.387	.429	.101
	medium	.027	.768	.205	.522
	high	.032	.360	.608	.377

The effect of personalization based on RF engagement classifiers using Attention-Net and Affect-Net features in grades 8 and 12 is depicted in Figure 6 for each student. As the amount of data per student varies, 60 samples correspond to different proportions of each person's data. Thus, we also report the requested (%) percent of samples.

With the exception of one student (S4 in Grade 8), the

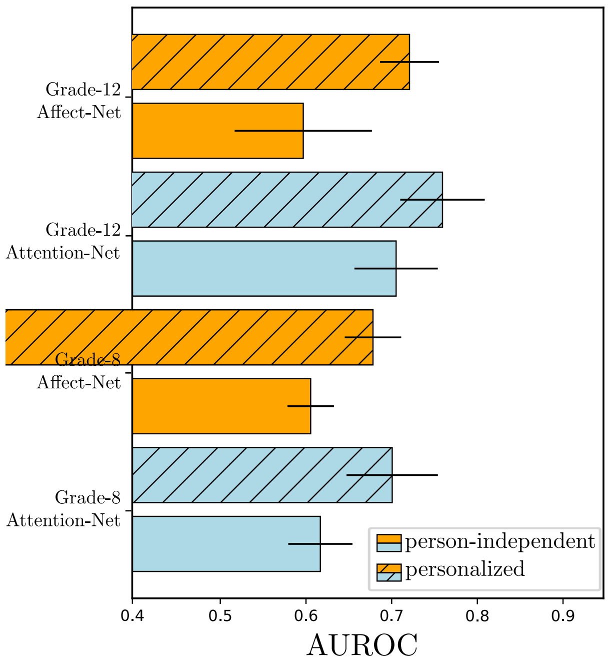

Fig. 7: The overall improvement of personalization in AUROC using Attention-Net and Affect-Net features in Grades 8 and 12.

amount of data was large enough. The requested data (60 samples) corresponded to only 2-3% of all the data. The effect of personalization varied from .03 to .29 in terms of AUC improvements. Affect features had greater improvements after personalization for both grades. Overall, there was a 6.89% and 9.83% AUROC improvement for attention and affect features, respectively.

Table 5 shows the confusion matrices of the RF classifier using Attention-Net features in Grade 12 before and after personalization. Without personalization, high engagement was misclassified as medium (.400 and .360) engagement, and low engagement was misclassified as medium and high. This may be due to class imbalances since the majority are medium engagement (52.2%). Personalization improves engagement classification across the board.

In summary, our experiments in personalization yielded average AUC improvements of .084 using only 60 seconds of personal data. The largest improvement, as depicted in Figure 7, was +.124 of AUC in Affect-Net features and RF classifier in Grade 12. Labeling 60 one-second samples selected from different parts of a video is more manageable than labeling the entire recording and takes only a few minutes for an expert annotator. This procedure enables a practical use of our proposed workflow. In return for this effort, the performance gain was substantial for both feature sets and grades.

6 Discussion

We aimed to develop video-based models to detect engagement during learning in authentic classroom environments. We collected a large-scale classroom observation dataset along with observer ratings of student engagement for grades 8 and 12 (N=15). In contrast to previous works that mainly used handcrafted local (i.e., local binary patterns, Gabor filters) and precomputed features such as head pose or estimated facial action units, we used deep learning methods for feature extraction and a combination of shallow and deep classifiers. We discuss our main findings, limitations/future work, applications, and ethical implications below.

6.1 Main Findings

Our main findings are that: (1) attention-based features were more effective at predictive engagement than affect-related features; (2) fusion of affective- and attention- features led to small boosts in accuracy; (3) there were limited benefits to deep learning methods over shallow classifiers; (4) overall engagement could be classified with moderate accuracy in a person-independent setting from individual streams; (5) engagement classification was higher for grade 12 vs. grade 8; and (6) even a small amount of person-specific data could considerably enhance classification accuracy.

6.2 Limitations and Future Work

From the technical perspective, the limited sample size is related to both the technical constraints of infrastructure in such field studies (e.g. the preparation of such a recording requires 20 minutes) and the manual effort associated with data annotations. Additionally, the presence of cameras can put pressure on students and cause their behavior to change when they know instruction is being recorded. Collecting a significant amount of audiovisual recordings from the same classes over the course of a school year as a longitudinal study could overcome some of these effects and allow researchers to investigate engagement in time.

Another limitation of this study is its focus on only the visible dimension of engagement. The detection of mind wandering through observation of students' facial expressions is a relevant emerging research topic [71]. Combining automated methods to detect mind wandering with engagement analysis may yield a better understanding of students' affective and cognitive engagement.

Even though our study presents a step towards measuring facial representations in the classroom, it was not possible to learn them on the engagement data due to the limited sample size. The use of self-supervision and representation learning on unlabelled classroom data may result in better representations for engagement analysis in future work. Additionally, our models failed to detect low engagement, likely due to data skew. The distribution of continuous labeling was also highly imbalanced. To solve these issues, we propose collecting more data in uncontrolled environments or, in order to obtain additional low engagement samples, employing interventions to manipulate engagement.

Our study focused on facial videos in particular. However, speech features can also provide valuable information for engagement classification and complement visual modalities. We observed that noisy capture from a distance makes audio signal separation more challenging. Thus, recording audio using separate voice recorders per desk synchronized with cameras could be the topic of future work. This could include the use affective acoustic signal processing and even speech recognition and language processing.

6.3 Applications and Ethical Considerations

The reported results in this study suggest that engagement classifiers could be applied to automate data processing within the scope of classroom instruction research and could be personalized using a small amount of data. Engagement models utilizing the same feature extractor used here can be applied to real-time students. Instead of recording videos, such a system would only record behavioral data and help increase the sample size for studies in classroom research. Data collection, storage, and privacy concerns are some of the significant issues that need to be addressed before large scale classroom studies can be conducted.

The use of these models outside of research settings is more limited. If used, it must be done in a privacy-preserving and ethical manner. In particular, our approach envisions, beyond anonymization [96], the immediate deletion of raw video recordings as part of the responsible use of the data. Instead, only aggregated information from the student group may be stored and individual-level scores discarded. As a result, the explicit mapping of a student's individual engagement scores and features can be avoided. Further improvement in the performance of engagement classification and a transition from student engagement to classroom-level analysis has the potential to make engagement analysis a more useful tool.

We are well aware that a potential application for engagement classifiers is a real-time classroom observation system such as [35]. Although such affective and cognitive interfaces summarizing engagement analytics as a teaching aid are growing in popularity, we strongly oppose any use of such solutions for real-world classroom monitoring. This is for both ethical reasons and due to the lack of empirical data on possible negative side effects with regard to student motivation and learning in such arrangements. As such, we explicitly do not advocate using these methods for any evaluative assessment of student motivations/learning and instruction quality. This is for many reasons including, but not limited to: the moderate accuracy of the models, the focus on only sub-components of engagement, the fact that engagement is only annotated by raters and not cross-referenced with other data sources, engagement levels may be impacted by external causes such as difficult family circumstances, and student engagement levels are not fully within the control of the instructors. To be blunt, using such tools for student and teacher evaluation and for any form of accountability would likely constitute a major ethical misuse of the technology. Importantly, advances in AI are necessary to address the fairness, accountability, transparency, and bias of the algorithms before they are deployed in any application. In this context, only a continuous, reflective dialog with social stakeholders can lead to sustainable solutions.

ACKNOWLEDGMENTS

Ömer Sümer and Patricia Goldberg are associate members of the LEAD Graduate School & Research Network, funded by the Ministry of Science, Research and the Arts of the state of Baden Württemberg within the framework of the sustainability funding for the projects of the Excellence Initiative II. This work is also supported by Leibniz-WissenschaftsCampus Tübingen "Cognitive Interfaces". Sidney D'Mello was supported by the NSF National AI Institute for Student-AI Teaming (iSAT) (DRL 2019805) and NSF IIS 1523091/ 1748739. The opinions expressed are those of the authors and do not represent views of the funding agencies.

REFERENCES

[1] S. L. Christenson, A. L. Reschly, and C. Wylie, *Handbook of research on student engagement*. Springer Science & Business Media, 2012.

[2] J. A. Fredricks, P. C. Blumenfeld, and A. H. Paris, "School engagement: Potential of the concept, state of the evidence," *Review of Educational Research*, vol. 74, no. 1, pp. 59–109, 2004. [Online]. Available: https://doi.org/10.3102/00346543074001059

[3] H. Lei, Y. Cui, and W. Zhou, "Relationships between student engagement and academic achievement: A meta-analysis," *Social Behavior and Personality: an international journal*, vol. 46, no. 3, pp. 517–528, 2018.

[4] M. Janosz, *Part IV Commentary: Outcomes of Engagement and Engagement as an Outcome: Some Consensus, Divergences, and Unanswered Questions*. Boston, MA: Springer US, 2012, pp. 695–703. [Online]. Available: https://doi.org/10.1007/978-1-4614-2018-7_33

[5] G. Eisele, H. Vachon, G. Lafit, P. Kuppens, M. Houben, I. Myin-Germeys, and W. Viechtbauer, "The effects of sampling frequency and questionnaire length on perceived burden, compliance, and careless responding in experience sampling data in a student population," *Assessment*, vol. 0, no. 0, p. 1073191120957102, 2020, pMID: 32909448. [Online]. Available: https://doi.org/10.1177/1073191120957102

[6] J. Fredricks and W. Mccolskey, *The Measurement of Student Engagement: A Comparative Analysis of Various Methods and Student Self-report Instruments*. Springer Science & Business Media, 01 2012, pp. 763–782.

[7] W. J. Van de Grift, S. Chun, R. Maulana, O. Lee, and M. Helms-Lorenz, "Measuring teaching quality and student engagement in south korea and the netherlands," *School Effectiveness and School Improvement*, vol. 28, no. 3, pp. 337–349, 2017.

[8] R. C. Pianta, K. M. La Paro, and B. K. Hamre, *Classroom Assessment Scoring System™: Manual K-3.* Paul H Brookes Publishing, 2008.

[9] N. Karweit and R. E. Slavin, "Measurement and modeling choices in studies of time and learning," *American Educational Research Journal*, vol. 18, no. 2, pp. 157–171, 1981.

[10] S. D. Beyda, S. S. Zentall, and D. J. Ferko, "The relationship between teacher practices and the task-appropriate and social behavior of students with behavioral disorders," *Behavioral disorders*, pp. 236–255, 2002.

[11] J. M. Girard and J. F. Cohn, "A primer on observational measurement," *Assessment*, vol. 23, no. 4, pp. 404–413, 2016, pMID: 26933139. [Online]. Available: https://doi.org/10.1177/1073191116635807

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

[13] N. Bosch, S. K. D'mello, J. Ocumpaugh, R. S. Baker, and V. Shute, "Using video to automatically detect learner affect in computer-enabled classrooms," *ACM Trans. Interact. Intell. Syst.*, vol. 6, no. 2, Jul. 2016. [Online]. Available: https://doi.org/10.1145/2946837

[14] S. Aslan, N. Alyuz, C. Tanriover, S. E. Mete, E. Okur, S. K. D'Mello, and A. Arslan Esme, "Investigating the impact of a real-time, multimodal student engagement analytics technology in authentic classrooms," in *Proceedings of the 2019 chi conference on human factors in computing systems*, 2019, pp. 1–12.

[15] S. D'Mello, R. W. Picard, and A. Graesser, "Toward an affect-sensitive autotutor," *IEEE Intelligent Systems*, vol. 22, no. 4, pp. 53–61, 2007.

[16] R. Cerezo, M. Sánchez-Santillán, M. P. Paule-Ruiz, and J. C. Núñez, "Students' lms interaction patterns and their relationship with achievement: A case study in higher education," *Computers & Education*, vol. 96, pp. 42–54, 2016.

[17] F. Okubo, T. Yamashita, A. Shimada, and H. Ogata, "A neural network approach for students' performance prediction," in *Proceedings of the seventh international learning analytics & knowledge conference*, 2017, pp. 598–599.

[18] J. W. You, "Identifying significant indicators using lms data to predict course achievement in online learning," *The Internet and Higher Education*, vol. 29, pp. 23–30, 2016.

[19] N. Bosch, "Detecting student engagement: Human versus machine," in *Proceedings of the 2016 Conference on User Modeling Adaptation and Personalization*, ser. UMAP '16. New York, NY, USA: ACM, 2016, pp. 317–320. [Online]. Available: http://doi.acm.org/10.1145/2930238.2930371

[20] E. Di Lascio, S. Gashi, and S. Santini, "Unobtrusive assessment of students' emotional engagement during lectures using electrodermal activity sensors," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 2, no. 3, pp. 1–21, 2018.

[21] K. S. McNeal, M. Zhong, N. A. Soltis, L. Doukopoulos, E. T. Johnson, S. Courtney, A. Alwan, and M. Porch, "Biosensors show promise as a measure of student engagement in a large introductory biology course," *CBE—Life Sciences Education*, vol. 19, no. 4, p. ar50, 2020.

[22] A. T. Poulsen, S. Kamronn, J. Dmochowski, L. C. Parra, and L. K. Hansen, "Eeg in the classroom: Synchronised neural recordings during video presentation," *Scientific reports*, vol. 7, no. 1, pp. 1–9, 2017.

[23] D. Bevilacqua, I. Davidesco, L. Wan, K. Chaloner, J. Rowland, M. Ding, D. Poeppel, and S. Dikker, "Brain-to-Brain Synchrony and Learning Outcomes Vary by Student–Teacher Dynamics: Evidence from a Real-world Classroom Electroencephalography Study," *Journal of Cognitive Neuroscience*, vol. 31, no. 3, pp. 401–411, 03 2019. [Online]. Available: https://doi.org/10.1162/jocn_a_01274

[24] D. K. Darnell and P. A. Krieg, "Student engagement, assessed using heart rate, shows no reset following active learning sessions in lectures," *PLOS ONE*, vol. 14, no. 12, pp. 1–13, 12 2019. [Online]. Available: https://doi.org/10.1371/journal.pone.0225709

[25] R. S. Baker and J. Ocumpaugh, *Interaction-Based Affect Detection in Educational Software*. Oxford Library of Psychology, 2015, pp. 233–245.

[26] P. Goldberg, Ö. Sümer, K. Stürmer, W. Wagner, R. Göllner, P. Gerjets, E. Kasneci, and U. Trautwein, "Attentive or not?: Toward a machine learning approach to assessing students' visible engagement in classroom instruction," *Educational Psychology Review*, 2019. [Online]. Available: https://doi.org/10.1007/s10648-019-09514-z

[27] B. McDaniel, S. D'Mello, B. King, P. Chipman, K. Tapp, and A. Graesser, "Facial features for affective state detection in learning environments," in *Proceedings of the Annual Meeting of the Cognitive Science Society*, vol. 29, no. 29, 2007.

[28] J. Bidwell and H. Fuchs, "Classroom analytics: Measuring student engagement with automated gaze tracking," University of North Carolina at Chapel Hill, Tech. Rep., 11 2011.

[29] M. Raca and P. Dillenbourg, "System for assessing classroom attention," *Proceedings of 3rd International Learning Analytics & Knowledge Conference*, 2013. [Online]. Available: http://infoscience.epfl.ch/record/185814

[30] M. Raca, "Camera-based estimation of student's attention in class," Ph.D. dissertation, EPFL, Lausanne, 2015.

[31] J. Zaletelj and A. Kosir, "Predicting students' attention in the classroom from kinect facial and body features," *EURASIP Journal on Image and Video Processing*, vol. 2017, pp. 1–12, 2017.

[32] J. Zaletelj, "Estimation of students' attention in the classroom from kinect features," in *Proceedings of the 10th International Symposium on Image and Signal Processing and Analysis*, Sept 2017, pp. 220–224.

[33] K. Fujii, P. Marian, D. Clark, Y. Okamoto, and J. Rekimoto, "Sync class: Visualization system for in-class student synchronization," in *Proceedings of the 9th Augmented Human International Conference*, ser. AH '18. New York, NY, USA: Association for Computing Machinery, 2018. [Online]. Available: https://doi.org/10.1145/3174910.3174927

[34] C. Thomas and D. B. Jayagopi, "Predicting student engagement in classrooms using facial behavioral cues," in *Proceedings of the 1st ACM SIGCHI International Workshop on Multimodal Interaction for Education*, ser. MIE 2017. New York, NY, USA: Association for Computing Machinery, 2017, p. 33–40. [Online]. Available: https://doi.org/10.1145/3139513.3139514

[35] K. Ahuja, D. Kim, F. Xhakaj, V. Varga, A. Xie, S. Zhang, J. E. Townsend, C. Harrison, A. Ogan, and Y. Agarwal, "Edusense: Practical classroom sensing at scale," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 3, no. 3, pp. 71:1–71:26, Sep. 2019. [Online]. Available: http://doi.acm.org/10.1145/3351229

[36] B. Ngoc Anh, N. Tung Son, P. Truong Lam, L. Phuong Chi, N. Huu Tuan, N. Cong Dat, N. Huu Trung, M. Umar Aftab, and T. Van Dinh, "A computer-vision based application for student behavior monitoring in classroom," *Applied Sciences*, vol. 9, no. 22, 2019. [Online]. Available: https://www.mdpi.com/2076-3417/9/22/4729

[37] S. K. D'Mello, S. D. Craig, and A. C. Graesser, "Multimethod assessment of affective experience and expression during deep learning," *Int. J. Learn. Technol.*, vol. 4, no. 3/4, p. 165–187, Oct. 2009. [Online]. Available: https://doi.org/10.1504/IJLT.2009.028805

[38] J. F. Grafsgaard, R. M. Fulton, K. E. Boyer, E. N. Wiebe, and J. C. Lester, "Multimodal analysis of the implicit affective channel in computer-mediated textual communication," in *Proceedings of the 14th ACM International Conference on Multimodal Interaction*, ser. ICMI '12. New York, NY, USA: Association for Computing Machinery, 2012, p. 145–152. [Online]. Available: https://doi.org/10.1145/2388676.2388708

[39] H. L. O'Brien and E. G. Toms, "The development and evaluation of a survey to measure user engagement," *Journal of the American Society for Information Science and Technology*, vol. 61, no. 1, pp. 50–69, 2010. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/asi.21229

[40] S. G. Hart and L. E. Staveland, "Development of nasa-tlx (task load index): Results of empirical and theoretical research," *Human mental workload*, vol. 1, no. 3, pp. 139–183, 1988.

[41] J. Ocumpaugh, R. Baker, M. A. Mercedes, and R. T., "Baker rodrigo ocumpaugh monitoring protocol (bromp) 2.0 technical and training manual," Columbia University, Tech. Rep., 2015.

[42] H. Monkaresi, N. Bosch, R. A. Calvo, and S. K. D'Mello, "Automated detection of engagement using video-based estimation of facial expressions and heart rate," *IEEE Transactions on Affective Computing*, vol. 8, no. 1, pp. 15–28, Jan 2017.

[43] A. Kamath, A. Biswas, and V. Balasubramanian, "A crowdsourced approach to student engagement recognition in e-learning environments," in *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, March 2016, pp. 1–9.

[44] A. Kaur, A. Mustafa, L. Mehta, and A. Dhall, "Prediction and localization of student engagement in the wild," in *2018 Digital Image Computing: Techniques and Applications (DICTA)*, Dec 2018, pp. 1–8.

[45] D. Sanchez-Cortes, O. Aran, and D. Gatica-Perez, "An audiovisual corpus for emergent leader analysis," in *Proceedings of the 10th International Symposium on Image and Signal Processing and Analysis*, 2011, p. 1–6.

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

[47] O. Celiktutan, E. Skordos, and H. Gunes, "Multimodal human-human-robot interactions (mhhri) dataset for studying personality and engagement," *IEEE Transactions on Affective Computing*, vol. 10, no. 4, pp. 484–497, Oct 2019.

[48] O. Rudovic, J. Lee, M. Dai, B. Schuller, and R. W. Picard, "Personalized machine learning for robot perception of affect and engagement in autism therapy," *Science Robotics*, vol. 3, no. 19, 2018. [Online]. Available: https://robotics.sciencemag.org/content/3/19/eaao6760

[49] H. W. Park, I. Grover, S. Spaulding, L. Gomez, and C. Breazeal, "A model-free affective reinforcement learning approach to personalization of an autonomous social robot companion for early literacy education," in *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence,*

EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019. AAAI Press, 2019, pp. 687–694. [Online]. Available: https://doi.org/10.1609/aaai.v33i01.3301687

[50] B. Schuller, *Deep Learning Our Everyday Emotions*. Cham: Springer International Publishing, 2015, pp. 339–346. [Online]. Available: https://doi.org/10.1007/978-3-319-18164-6_33

[51] B. Kratzwald, S. Ilić, M. Kraus, S. Feuerriegel, and H. Prendinger, "Deep learning for affective computing: Text-based emotion recognition in decision support," *Decision Support Systems*, vol. 115, pp. 24–35, 2018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167923618301519

[52] S. Zhao, S. Wang, M. Soleymani, D. Joshi, and Q. Ji, "Affective computing for large-scale heterogeneous multimedia data: A survey," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 15, no. 3s, Dec. 2019. [Online]. Available: https://doi.org/10.1145/3363560

[53] P. V. Rouast, M. T. P. Adam, and R. Chiong, "Deep learning for human affect recognition: Insights and new developments," *IEEE Transactions on Affective Computing*, vol. 12, no. 2, pp. 524–543, 2021.

[54] X. Chen, H. Xie, D. Zou, and G.-J. Hwang, "Application and theory gaps during the rise of artificial intelligence in education," *Computers and Education: Artificial Intelligence*, vol. 1, p. 100002, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2666920X20300023

[55] F. Ouyang and P. Jiao, "Artificial intelligence in education: The three paradigms," *Computers and Education: Artificial Intelligence*, vol. 2, p. 100020, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2666920X2100014X

[56] S. D'Mello, E. Dieterle, and A. Duckworth, "Advanced, analytic, automated (aaa) measurement of engagement during learning," *Educational psychologist*, vol. 52, no. 2, pp. 104–123, 2017. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/29038607

[57] R. E. Slavin, "When does cooperative learning increase student achievement?" *Psychological bulletin*, vol. 94, no. 3, p. 429, 1983.

[58] A. M. O'Donnell, "The role of peers and group learning." *Handbook of educational psychology*, pp. 781–802, 2006.

[59] M. Tölgyessy, M. Dekan, L. Chovanec, and P. Hubinskỳ, "Evaluation of the azure kinect and its comparison to kinect v1 and kinect v2," *Sensors*, vol. 21, no. 2, p. 413, 2021.

[60] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L. Morency, "Openface 2.0: Facial behavior analysis toolkit," in *2018 13th IEEE International Conference on Automatic Face Gesture Recognition (FG 2018)*, 2018, pp. 59–66.

[61] M. T. H. Chi and R. Wylie, "The icap framework: Linking cognitive engagement to active learning outcomes," *Educational Psychologist*, vol. 49, no. 4, pp. 219–243, 2014. [Online]. Available: https://doi.org/10.1080/00461520.2014.965823

[62] A. Helmke, "Das münchner aufmerksamkeitsinventar (mai). manual für die beobachtung des aufmerksamkeitsverhaltens von grundschülern während des unterrichtes," Max-Planck-Institut für psychologische Forschung, Tech. Rep. 6, 1988.

[63] E. Okur, N. Alyuz, S. Aslan, U. Genc, C. Tanriover, and A. A. Esme, "Behavioral engagement detection of students in the wild," in *International Conference on Artificial Intelligence in Education*. Springer, 2017, pp. 250–261.

[64] N. Alyuz, E. Okur, E. Oktay, U. Genc, S. Aslan, S. E. Mete, B. Arnrich, and A. A. Esme, "Semi-supervised model personalization for improved detection of learner's emotional engagement," in *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, ser. ICMI '16. New York, NY, USA: Association for Computing Machinery, 2016, p. 100–107. [Online]. Available: https://doi.org/10.1145/2993148.2993166

[65] J. Smallwood and J. W. Schooler, "The restless mind." *Psychological bulletin*, vol. 132, no. 6, p. 946, 2006.

[66] S. D'Mello, "What do we think about when we learn?" in *Deep comprehension*, K. Millis, D. L. Long, J. P. Magliano, and K. Wiemer, Eds. New York, NY, USA: Routledge, 2018, pp. 52–67.

[67] S. Hutt, K. Krasich, C. Mills, N. Bosch, S. White, J. R. Brockmole, and S. K. D'Mello, "Automated gaze-based mind wandering detection during computerized learning in classrooms," *User Modeling and User-Adapted Interaction*, vol. 29, no. 4, pp. 821–867, 2019.

[68] N. Blanchard, R. Bixler, T. Joyce, and S. D'Mello, "Automated physiological-based detection of mind wandering during learning," in *Intelligent Tutoring Systems*, S. Trausan-Matu, K. E. Boyer, M. Crosby, and K. Panourgia, Eds. Cham: Springer International Publishing, 2014, pp. 55–60.

[69] A. Stewart, N. Bosch, H. Chen, P. Donnelly, and S. D'Mello, "Face forward: Detecting mind wandering from video during narrative film comprehension," in *Artificial Intelligence in Education*, E. André, R. Baker, X. Hu, M. M. T. Rodrigo, and B. du Boulay, Eds. Cham: Springer International Publishing, 2017, pp. 359–370.

[70] A. Stewart, N. Bosch, and S. K. D'Mello, "Generalizability of face-based mind wandering detection across task contexts." *International Educational Data Mining Society*, 2017.

[71] N. Bosch and S. D'Mello, "Automatic detection of mind wandering from video in the lab and in the classroom," *IEEE Transactions on Affective Computing*, pp. 1–1, 2019.

[72] O. Rudovic, M. Zhang, B. Schuller, and R. Picard, "Multi-modal active learning from human data: A deep reinforcement learning approach," in *2019 International Conference on Multimodal Interaction*, ser. ICMI '19. New York, NY, USA: Association for Computing Machinery, 2019, p. 6–15. [Online]. Available: https://doi.org/10.1145/3340555.3353742

[73] O. Rudovic, H. W. Park, J. Busche, B. Schuller, C. Breazeal, and R. W. Picard, "Personalized estimation of engagement from videos using active learning with deep reinforcement learning," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2019, pp. 217–226.

[74] D. Lufi and I. Haimov, "Effects of age on attention level: changes in performance between the ages of 12 and 90," *Aging, Neuropsychology, and Cognition*, vol. 26, no. 6, pp. 904–919, 2019, pMID: 30427264. [Online]. Available: https://doi.org/10.1080/13825585.2018.1546820

[75] C. J. Soto and O. P. John, "Short and extra-short forms of the big five inventory–2: The bfi-2-s and bfi-2-xs," *Journal of Research in Personality*, vol. 68, pp. 69 – 81, 2017. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0092656616301325

[76] B. Frank, *Presence messen in laborbasierter Forschung mit Mikrowelten Entwicklung und erste Validierung eines Fragebogens zur Messung von Presence*. Springer-Verlag, 2015.

[77] S. E. Rimm-Kaufman, A. E. Baroody, R. A. A. Larsen, T. W. Curby, and T. Abry, "To what extent do teacher-student interaction quality and student gender contribute to fifth graders' engagement in mathematics learning?" *Journal of Educational Psychology*, vol. 107, no. 1, pp. 170–185, 2015.

[78] M. Knogler, J. M. Harackiewicz, A. Gegenfurtner, and D. Lewalter, "How situational is situational interest?: Investigating the longitudinal structure of situational interest." *Contemporary Educational Psychology*, vol. 43, pp. 39–50, 2015.

[79] J. M. Girard, "CARMA: Software for continuous affect rating and media annotation," *Journal of Open Research Software*, vol. 2, no. 1, p. e5, 2014.

[80] I. Lizdek, P. Sadler, E. Woody, N. Ethier, and G. Malet, "Capturing the stream of behavior: A computer-joystick method for coding interpersonal behavior continuously over time," *Social Science Computer Review*, vol. 30, no. 4, pp. 513–521, 2012. [Online]. Available: https://doi.org/10.1177/0894439312436487

[81] M. Hommel, "Aufmerksamkeitstief in reflexionsphasen – eine videoanalyse von planspielunterricht," *Wirtschaft und Erziehung*, pp. 12–18, 01 2012.

[82] M. K. Alimoglu, D. B. Sarac, D. Alparslan, A. A. Karakas, and L. Altintas, "An observation tool for instructor and student behaviors to measure in-class learner engagement: a validation study," *Medical education online*, vol. 19, no. 1, p. 24037, 2014.

[83] R. S. Baker, A. T. Corbett, K. R. Koedinger, and A. Z. Wagner, "Off-task behavior in the cognitive tutor classroom: when students" game the system"," in *Proceedings of the SIGCHI conference on Human factors in computing systems*, 2004, pp. 383–390.

[84] N. Karweit and R. E. Slavin, "Measurement and modeling choices in studies of time and learning," *American Educational Research Journal*, vol. 18, no. 2, pp. 157–171, 1981.

[85] J. J. Carta, C. R. Greenwood, D. Schulte, C. Arreaga-Mayer, and B. Terry, "Code for instructional structure and student academic response: Mainstream version (ms-cissar)," Bureau of Child Research, University of Kansas, Tech. Rep., 1988.

[86] J. Deng, J. Guo, Z. Yuxiang, J. Yu, I. Kotsia, and S. Zafeiriou, "Retinaface: Single-stage dense face localisation in the wild," in *arxiv*, 2019.

[87] J. Deng, J. Guo, X. Niannan, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *CVPR*, 2019.

[88] A. Mollahosseini, B. Hasani, and M. H. Mahoor, "Affectnet: A database for facial expression, valence, and arousal computing in

the wild," *IEEE Transactions on Affective Computing*, vol. 10, no. 1, pp. 18–31, 2019.

[89] T.-H. Vo, G.-S. Lee, H.-J. Yang, and S.-H. Kim, "Pyramid with super resolution for in-the-wild facial expression recognition," *IEEE Access*, vol. 8, pp. 131 988–132 001, 2020.

[90] L. Schoneveld, A. Othmani, and H. Abdelkawy, "Leveraging recent advances in deep learning for audio-visual emotion recognition," *Pattern Recognition Letters*, vol. 146, pp. 1–7, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167865521000878

[91] X. Zhu, Z. Lei, X. Liu, H. Shi, and S. Z. Li, "Face alignment across large poses: A 3d solution," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 146–155.

[92] N. Ruiz, E. Chong, and J. M. Rehg, "Fine-grained head pose estimation without keypoints," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2018, pp. 2155–215 509.

[93] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997. [Online]. Available: https://doi.org/10.1162/neco.1997.9.8.1735

[94] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Y. Bengio and Y. LeCun, Eds., 2015. [Online]. Available: http://arxiv.org/abs/1412.6980

[95] D. D. Lewis and J. Catlett, "Heterogeneous uncertainty sampling for supervised learning," in *Machine learning proceedings 1994*. Elsevier, 1994, pp. 148–156.

[96] Ö. Sümer, P. Gerjets, U. Trautwein, and E. Kasneci, "Automated anonymisation of visual and audio data in classroom studies," *Workshops of the Thirty-Fourth Association for the Advancement of Artificial Intelligence (AAAI) Conference*, 2020.

Ömer Sümer received the BSc degree in electronics engineering from Naval Academy in Istanbul, Turkey, and the MSc degree in electronics engineering from Istanbul Technical University in Istanbul, Turkey. As of February 2021, he received his Ph.D. degree in computer science at the University of Tübingen. His research interests involve machine learning, computer vision, and their application in social and affective computing, focusing on learning analytics.

Patricia Goldberg received her BSc degree in cognitive science from the University of Osnabrück and her MA degree in education science from the University of Freiburg, Germany. She did her PhD in Psychology at the Hector Research Institute of Education Sciences and Psychology at the University of Tübingen. In her research, she is focusing on attentional processes in teacher-learner interactions to improve teacher training and classroom instruction.

Sidney D'Mello Sidney D'Mello (PhD in Computer Science) is an Associate Professor in the Institute of Cognitive Science and Department of Computer Science at the University of Colorado Boulder. He is interested in the dynamic interplay between cognition and emotion while individuals and groups engage in complex real-world tasks. He applies insights gleaned from this basic research program to develop intelligent technologies that help people achieve to their fullest potential by coordinating what they think and feel with what they know and do. D'Mello has co-edited seven books and published almost 300 journal papers, book chapters, and conference proceedings. His work has been funded by numerous grants and he currently serves(d) as associate editor for Discourse Processes and PloS ONE. D'Mello is the Principal Investigator for the NSF National Institute for Student-Agent Teaming.

Peter Gerjets received the diploma in psychology from the University of Göttingen, in 1991. From 1991 to 1995 he was a research associate with the University of Göttingen where he received the PhD degree in 1994. Afterwards he has been working as assistant professor with the Saarland University in Saarbrucken where he finished his habilitation in 2002 before taking over his current position at the University of Tübingen. Since 2002 he has been working as principal research scientist at the Knowledge Media Research Center and beside as full professor for research on learning and instruction at the University of Tübingen. He was honoured with the Young Scientist Award of the German Cognitive Science Society in 1999 and served in the editorial boards of the Journal of Educational Psychology, the Educational Research Review, the Computers in Human Behavior, and the Educational Technology, Research, and Development. His current research focuses on multimodal and embodied interaction with digital media as well as on learning from multimedia, hypermedia, and the Web. He is a member of DGPs, APS, and EARLI and served as coordinator of the EARLI Special Interest Group 6:Instructional Design.

Ulrich Trautwein is a Professor of Education Sciences and Executive Director of the Hector Research Institute of Education Sciences and Psychology at the University of Tübingen. He is also Co-Director of the LEAD Research Network. Prior to this, Professor Trautwein was a senior researcher at the Max Planck Institute for Human Development in Berlin. His area of research is empirical educational research, broadly defined.In particular, his research interests are directed primarily to the development of self-referent cognitions in the school context, school management and the influence of homework on school achievement. Trautwein has published a large number of articles on a number of topics, including the development of conscientiousness, expectancy-value beliefs, and academic effort, the effectiveness of homework assignments and completion, and the results of several randomized controlled field trials in school settings.

Enkelejda Kasneci is a Professor of Computer Science at the University of Tübingen, Germany, where she leads the Human-Computer Interaction Lab. As a BOSCH scholar, she received her M.Sc. degree in Computer Science from the University of Stuttgart in 2007. In 2013, she received her PhD in Computer Science from the University of Tübingen. For her PhD research, she was awarded the research prize of the Federation Südwestmetall in 2014. From 2013 to 2015, she was a postdoctoral researcher and a Margarete-von-Wrangell Fellow at the University of Tübingen. Her research evolves around the application of machine learning for intelligent and perceptual human-computer interaction. She serves as academic editor for PlosOne and as a TPC member and reviewer for several major conferences and journals.