Smart Learning Environments

REVIEW Open Access

Engagement detection in online learning: a review

M. Ali Akber Dewan*, Mahbub Murshed and Fuhua Lin

* Correspondence: adewan@athabascau.ca
School of Computing and Information Systems, Faculty of Science and Technology, Athabasca University, Athabasca, Alberta, Canada

Abstract

Online learners participate in various educational activities including reading, writing, watching video tutorials, online exams, and online meetings. During the participation in these educational activities, they show various engagement levels, such as boredom, frustration, delight, neutral, confusion, and learning gain. To provide personalized pedagogical support through interventions to online learners, it is important for online educators to detect their online learners' engagement status precisely and efficiently. This paper presents a review of the state of the art in engagement detection in the context of online learning. We classify the existing methods into three main categories—*automatic, semi-automatic* and *manual*—considering the methods' dependencies on learners' participation. Methods in each category are then divided into subcategories based on the data types (e.g., audio, video, texts for learner log data etc.) they process for the engagement detection. In particular, the *computer vision based methods* in the *automatic* category that use facial expressions are examined in more details because they are found to be promising in the online learning environment. These methods are nonintrusive in nature, and the hardware and the software that these methods use to capture and analyze video data are cost-effective and easily achievable. Different techniques in the field of computer vision and machine learning are applied in these methods for the engagement detection. We then identify their challenges of engagement detection and explore available datasets and performance metrics for engagement detection, and provide recommendations for the future to advance the technology of engagement detection for online education.

Keywords: Engagement detection, Affect detection, Facial expression recognition, Action units, Emotion detection

Introduction

A key issue in online learning is to improve learners' engagement with their educational activities. Since the 1980s, learner engagement has been a key topic in the education literature (Whitehill et al. 2014). This interest may be driven by the concerns about high drop-out rates in online courses (Rothkrantz 2016). It is widely acknowledged that engagement and affect are linked to increased productivity and learning gain. Some research shows that engagement is malleable, and proper pedagogical interventions, learning designs and feedback can enhance learner engagement (Monkaresi et al. 2017). To provide personalized pedagogical support through intervention to online learners, detecting learners' engagement has become important to online

education. Fostering learners' engagement can benefit not only in online learning, but also in other learning settings such as traditional classrooms, educational games, and intelligent tutoring systems (Karumbaiah et al. 2017).

Several facets of learners' engagements have been discussed in the literature (Bosch 2016; Fredrick et al. 2004; Anderson et al. 2004). Bosch (2016) organizes engagement as three different forms: *affective, behavioral, and cognitive*. Fredrick et al. (2004) define engagement as *behavioral, cognitive, and emotional*, whereas Anderson et al. (2004) define as *academic*, *behavioral*, *cognitive*, and *psychological* in their research studies. *Affective* engagement refers to the emotional attitude, for example, being interested in a topic and enjoying learning about it (Bosch 2016)), whereas a*cademic* engagement refers to academic identification (e.g., getting along with teachers) and participation (e.g., time on tasks, not skipping classes) towards learning (Al-Hendawi 2012). *Behavioral* engagement draws on the idea of participation including participation in the classroom and extra-curricular activities, stay focused, submit assigned tasks, and follow the instructor's dictation (Christenson et al. 2012). *Cognitive* engagement refers to the thoughtfulness and willingness to exert the effort necessary to comprehend complex ideas and master difficult skills (e.g., focused attention, memory, and creative thinking (Anderson et al. 2004)). *Emotional* engagement encompasses positive and negative reactions to teachers, classmates, and academics (Fredrick et al. 2004). *Psychological* engagement refers to the sense of belonging and relationships with teachers and peers (Christenson and Anderson 2002).

Different types of engagements in the context of learning are useful to know for personalized intervention design to improve learners' experience. However, studies that focus on learner engagement need a way of measuring it (Harris 2008). This can be done with one of the two types of data identified by engagement theorists: internal to the individual (cognitive and affective) and external observable factors (perceptible facial features, postures, speech, and actions) (Bosch 2016). Some research studies also emphasized that measuring engagement requires bringing together observational data with the data internal to the individual (e.g., self-reports) (Whitehill et al. 2014).

This paper presents a review of the state of the art of engagement detection methods in the context of online learning, and then it identifies the challenges of detecting engagement in online learning. We classify the existing methods into three main categories—*automatic, semi-automatic* and *manual*—considering the methods' dependencies on learners' participation. And, then the methods in each category are divided into subcategories based on the types of data used, e.g., audio, video, learner log data etc. In particular, the *computer vision based methods* in the *automatic* category that use facial expressions are examined because they are promising in an online learning environment, nonintrusive in nature, and cost-effective when considering the hardware and the software needed for capturing and analyzing video data. Finally, we explore available datasets and performance metrics for engagement measurement, and provide recommendations for the future to advance the technology of engagement measurement for online education.

The remainder of the paper is organized as follows. In Section II, a taxonomy of engagement detection methods is proposed and related trends are discussed. Among the different methods, the *computer vision based methods* in the *automatic* category are found to be beneficial and further detailed in Section III. Benchmarking datasets,

performance metrics, and evaluation strategies along with some results are discussed in Section IV. Section V concludes the paper with some critical discussions and future recommendations.

Taxonomy of engagement detection methods

Several research studies on learners' engagement detection can be found in the literature. To review them, we propose a taxonomy which is shown in Fig. 1. First, we divide the existing methods for learners' engagement detection into three main categories — *automatic, semi-automatic,* and *manual* — based on the strategy and the type of users' involvement in the engagement detection process. The *manual* methods are further divided into *self-reporting* and *observational check-list* categories. The methods related to *engagement tracing* are categorized as *semi-automatic* in the taxonomy. The methods in the *automatic* category are divided into *computer vision based methods*, *sensor data analysis*, and *log-file analysis* depending on the information that these methods process for engagement detection. The *computer vision based methods* are further divided into three sub-categories — *facial expression*, *gestures and postures*, and *eye movement* — based on the modalities they use for the engagement detection. Although some research studies use the above modalities separately, some others find it promising to combine two or more of them to achieve a higher accuracy.

The *manual* category refers to the methods where learners' direct involvement is needed in the engagement detection process. In the *manual* category, *self-reporting* is a popular technique where a set of questionnaire is posted in which learners report their own level of attention, distraction, excitement, or boredom O'Brien and Toms 2010. All the survey questionnaires do not necessarily indicate the level of engagement of the learners directly, rather they imply engagement as a descriptive latent variable using factor analysis (Matthews et al. 2002; Wixon et al. 2016). *Self-reporting* is of great interest to many researchers because it is easy to administer and it provides some useful information regarding learner engagement. For example, it is useful to know that 25 and

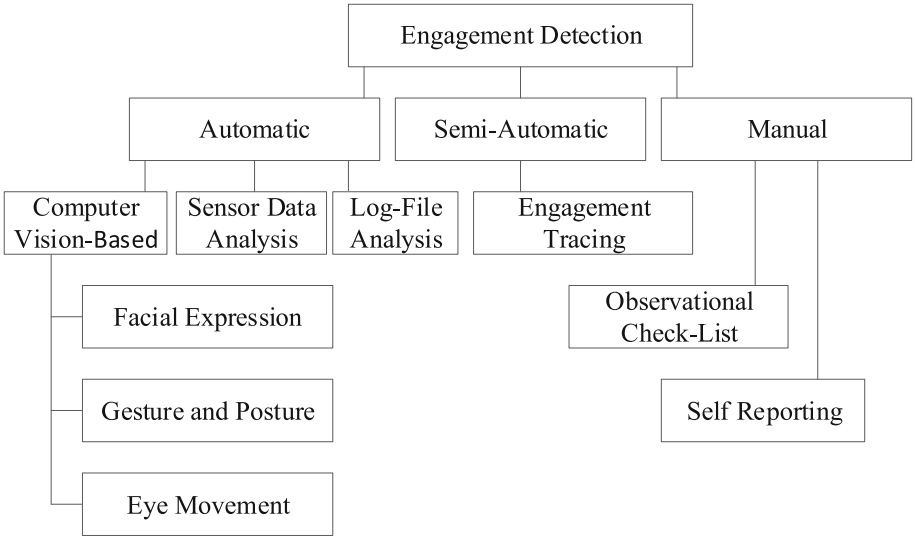

Fig. 1 Taxonomy of learners' engagement detection methods

60% of the learners report being bored and disengaged, respectively (Shernoff et al. 2000). However, the validity of the *self-reporting* results depends on a number of factors that are outside of the control of the researchers, such as learners' honesty, their willingness to report their emotion, and the accuracy of learners' perception about their emotions (D'Mello et al. 2014).

Observational checklist is another popular method in the *manual* category for detecting learner engagement that relies on questionnaires completed by external observers instead of the learners. These questionnaires often consider teachers' personal opinion regarding the learners' engagement levels. They may also contain checklists for objective measures that are supposed to indicate engagement. Some example questions are "do the learners sit quietly?", "do they do their homework?", "are they on time?", "do they ask questions?" (Parsons and Taylor 2011). In some cases, external observers may rate learner engagement based on live or pre-recorded videos of educational activities (Kapoor and Picard 2005; 2012). Observers may also consider samples of the learner's work such as essays, projects, and class notes (Parsons and Taylor 2011). *Observational checklists* also have some limitations. Observational metrics may not always be related to engagement. For example, *sitting quietly*, *good behavior*, and *no tardy cards* appear to measure compliance and willingness to adhere to rules and regulations rather than engagement (Whitehill et al. 2014). Another major limitation for both the *self-reporting* and *observational checklist* is that they require a great deal of time and effort from both the learners and the observers to detect learners' engagement.

Methods in the *semi-automatic* category, learners' indirect involvements are needed in the engagement detection process. *Engagement tracing* is a popular method in the *semi-automatic* category that utilizes the timing and accuracy of learner responses to practice problems and test questions (Beck 2005). In order to evaluate the time and accuracy pattern, probabilistic inference is used and it remains consistent with an engaged or disengaged learner (Beck 2005; Johns and Woolf 2006). For example, very short response times on easy questions indicates that the learners are not engaged and are simply giving random answers without any effort. Although this method has been widely used in intelligent tutoring systems (ITS), not many applications of this method can be found in other educational settings, such as in online learning (Whitehill et al. 2014).

Methods in *automatic* category extract features from various traits captured by image sensors (e.g., eyes movement, facial expressions, and gestures and postures), physiological and neurological sensors (e.g., heart rate, EEG, blood pressure, or galvanic skin response) or by tracing learners' activities in their learning environments (e.g., total time spent on study, number of forum posts, average time to solve a problem, number of submissions correct etc.). These methods extract features automatically and do not interrupt learners in the engagement detection process. The methods in the *automatic* category are further divided into three groups: *log-file analysis*, *sensor data analysis*, and *computer vision based methods*.

In the *log-file analysis*, learners' actions preserved in log files are analyzed for the engagement detection. Especially, in an online learning environment, the learners' actions are stored in log files and this can provide valuable information for the engagement detection. Different data mining and machine learning approaches are used in the *log-file analysis*. Cocea and Weibelzahl (2009, 2011) analyzed log-files in a web-based

learning environment called *HTML-Tutor*. This research study analyzed 30 attributes of the online learners' from the log file including a number of pages accessed, average time spent on pages, number of tests attended, number of correctly answered tests, and number of incorrectly answered tests. Sundar and Kumar (2016) proposed an improvement over the above methods by combining the attributes of the log file with the user profile. In another research study, Aluja-Baneta et al. (2017) applied psychometric theory to 14 behavioral indicators to measure the learners' engagement in a virtual learning environment.

In the *sensor data analyses*, physiological and neurological sensor readings are used to measure engagement. In the neuroscience literature, engagement is typically equated with the level of arousal or alertness (Whitehill et al. 2014). Various physiological measures, such as EEG, blood pressure, heart rate, or galvanic skin response, are used to measure engagement and alertness (Chaouachi et al. 2010; Fairclough and Venables 2006; Goldberg et al. 2011). However, these measures require specialized sensors and are not convenient to use in real-life education settings.

The *computer vision based methods* offer a number of ways to measure learners' engagement by investigating the cues from the *gestures and postures*, *eye movement,* and *facial expressions* (D'Mello et al. 2009; D'Mello and Graesser 2010; Kapoor and Picard 2005; McDaniel et al. 2007). The main advantage of *computer vision based methods* is the unobtrusiveness of the assessment process and easy to use, similar to the classroom situation where a teacher observes whether a learner is motivated without interrupting his/her activities. Affective computing techniques and low cost of cameras and wide-spread of its availability in cell phones, tablets, computers, and even automobiles, are allowing to detect learners' engagement using computer vision (Monkaresi et al. 2017; Kamath et al. 2016). While vision-based methods for engagement detection have been pursued previously by the ITS community, much work remains to be done before making these automatic systems practical in a wide variety of education settings, such as in online learning.

Computer vision based methods for detecting engagement in online learning

Computer vision based methods are used to estimate learners' perceived engagement, i.e., engagement as judged by an external observer. Since teachers rely on perceived engagement to adapt their teaching behavior in a conventional classroom setting, the automation of perceived engagement detection is likely to be useful for online learning. The online courses can take advantage of this technological advancement for personalized intervention design, and reduce learners' frustration and dropout rates.

Generic framework

To facilitate the review, we present a generic framework (see Fig. 2) for learner's perceived engagement detection using the computer vision based methods. The framework is consisted with five different modules that include *detection*, *feature extraction*, *tracking*, *classification*, and *decision*.

In a computer vision based engagement detection system, video streams are captured using a webcam or a surveillance camera, where the camera provides a particular view of learners participating in a learning activity. The system seeks to detect the region of

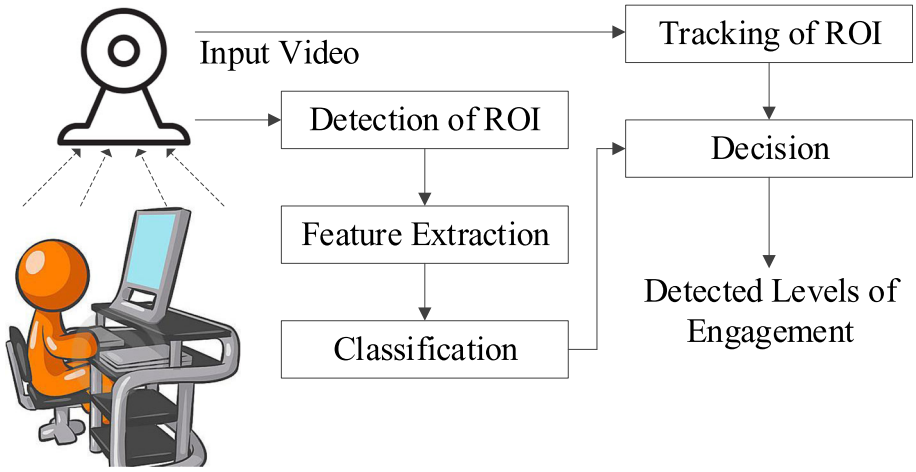

Fig. 2 A generic framework for computer vision based engagement detection system

interests (ROIs) (e.g., face, gestures, postures or eye) of the learners in the live video stream. Typically, engagement detection in such system is performed with a track-and-classify approach. The system first performs segmentation to isolate the ROIs using a *detection* module in each frame. For each ROI, features are then extracted in a *feature extraction* module and selected into patterns to initiate tracking and classification. A *classification* module is used to match input patterns against patterns extracted from training dataset and generates classification scores. A *tracking* module is designed for tracking the movement or changes in the ROIs in consecutive frames and generates tracking trajectories. Finally, a *decision* module combines classification scores over trajectories to output a list of engagement levels of the learners in the input video stream.

We found that the most commonly used modalities in *computer vision based methods* are *facial expressions*, *gestures and postures*, and *eye movement*. Thus, in the following subsections, we further review the engagement detection methods focusing on these three modalities.

Facial expressions

It has been hypothesized that a good deal of information used by humans to make engagement judgment is based on human faces, and facial expressions are directly linked to the perceived engagement (Whitehill et al. 2008, 2014; Ekman et al. 2002; Littlewort et al. 2011). Using cameras provide a continuous and non-intrusive way of capturing face images as a learner uses a mobile device or a personal computer for his or her learning activities. The captured facial information is used to understand certain facets of the learner's current state of mind. Many different methods have been proposed to automate this detection process by analyzing the face images (Booth et al. 2017; Bosch et al. 2014). Based on how the information from a face appearance is used, these methods are divided into two groups: *part-based* and *appearance-based*. Both the *part-based* methods and the *appearance-based* methods use *geometric* and *holistic* features in their engagement detection process (Dewan et al. 2018).

Part-based methods

Part-based methods refer to the techniques that analyze different parts of a face (e.g., eyes, mouth, nose, forehead, chin and so on) for the engagement detection. A comprehensive way to analyze the parts of a face is the Facial Action Coding System (FACS). Ekman and Friesen (1978) is the pioneer in developing the FACS system for the analysis of facial expressions. The FACS uses facial muscle movements also known as action units (AUs) in order to design the theoretical measure of specific discrete emotions (Ekman and Friesen 1978; Ekman et al. 2002). FACS has been extensively used by psychologists and neuroscientists on various aspects of facial expression analysis. The AUs can occur either singly or in combination. Ekman et al. (2002) acknowledged that although the number of AUs is relatively small, more than 7000 AU combinations are observed in our everyday life, and certain AUs or certain combinations of AUs are more frequent than the others. For example, happiness is sometimes viewed as a combination of AU12 and AU6. Using FACS, every possible facial expression can be empirically described as a mixture of AUs.

Measuring AUs is a descriptive analysis of behavior, whereas measuring facial expressions, such as anger or happiness, is an inferential process (Ekman and Friesen 1978). Any observational system (e.g., engagement detection) requires inferences about that which is being measured (Whitehill et al. 2014; Grafsgaard et al. 2013a). Although the FACS has been widely used in facial expression recognition in the last several years, it has just got attention for the engagement detection in learning context (Grafsgaard et al. 2013b). In the literature, the mapping of AUs to those each of the expressions are relatively well defined (Martinez et al. 2017). However, the mapping of AUs to learning-centered affective states are still at its early stage. Some mapping has been done in recent years and these are listed in Table 1. McDaniel et al. (2007) mapped the AUs into 5 different engagement levels — boredom, confusion, delight, flow, frustration, and surprise — when the learners interact with a web-based educational tool called *AutoTutor*. The learners' facial expressions are coded using FACS by the human experts. The set of AUs that accompany the above affective states are listed in Table 1. The authors acknowledged that the above estimated affective states were the most

Table 1 List of AUs involved in engagement detection using facial expressions in online learning

Engagement Levels	Action Units	Methods
Boredom	AU4, AU7, AU12	McDaniel et al. (2007)
Confusion	AU1, AU4, AU7, AU12	
Delight	AU4, AU7, AU12, AU25, AU26	
Frustration	AU12	
Neutral	AU4, AU7, AU12, AU25, AU26	
Difficulties in viewing speed	AU1, AU2, AU4, AU5, AU9, AU10, AU12, AU14, AU15, AU17, AU20, AU45	Whitehill et al. (2008, 2014)
Confusion	AU4, AU7	Grafsgaard et al. (2013a, 2013b)
Frustration	AU1, AU2, AU4	
Learning Gain	AU2, AU14	
Confusion	AU45, AU1, AU4	Bosch (2016); Bosch et al. (2014, 2015, 2016)
Frustration	AU45, head pose feature	
Learning Gain	AU2, AU4, AU5, AU12, AU15, AU23	Vail et al. (2016a); Vail et al. (2016b)

prominent uses in online learning. They also acknowledged that some of these affective states were correlated with learning gains, e.g., boredom was negatively correlated with learning, whereas confusion and flow were positively correlated.

Instead of discriminating the engagement in different levels, Booth et al. (2017) measured learners' engagement in a scalable and accessible manner. In this study, videos were collected from a screen-mounted camera of learners studying online lectures. AUs (Al-Hendawi 2012; Aluja-Baneta et al. 2017; Aslan et al. 2014; Bartlett et al. 2006; Bosch 2016; Chen et al. 2013; Christenson et al. 2012; Cocea and Weibelzahl 2009; Dewan et al. 2018; D'Mello et al. 2014; D'Mello et al. 2009; Fredrick et al. 2004; Grafsgaard et al. 2013b; Grafsgaard et al. 2013c) were used with facial landmarks, eye gaze, emotion probabilities, average optical flow magnitude and direction, and head pose and size. A decision on engagement detection was done using the K-nearest neighbor (KNN) classifier. Both subject-independent and individual-specific models were analyzed, where the individual-specific models outperformed the others. Whitehill et al. (2008) estimated the difficulty level of an online lecture by analyzing 12 AUs which were automatically recognized by support vector machines (SVM) and Gabor energy filters (Bartlett et al. 2006). The output of the 12 AUs detectors (see Table 1) were analyzed to make inference on the difficulty levels the learners feel about an online lecture. By using this measurement, the speed of the instruction was adjusted automatically to avoid any frustration, confusion, and boredom that could potentially happen to the online learners. In another study, Whitehill et al. (2014) investigated the face and facial landmark (i.e., eyes, nose, and mouth) with four binary classifiers, one for each engagement category: not engaged, nominally engaged, normally engaged, and very engaged. The classifiers were formed by combining GentleBoost with Box Filter features (Boost (BF)), SVM with Gabor features (SVM (Gabor)), Multinomial logistic regression (MLR) with expression outputs from the Computer Expression Recognition Toolbox CERT (MLR (CERT)) (Littlewort et al. 2011).

CERT has also been used in several other research studies for engagement detection in the context of learning. CERT gives intensity values for facial AUs from a wide range of FACS, thus empowering fine-tuned analyses for inferring affective states of learners' using facial expression analysis. Grafsgaard et al. (2013b) analyzed facial movements consisting of brow raising, brow lowering, eyelid tightening, and mouth dimpling that occurred during computer-mediated tutoring using the CERT. In this study, upper face movements were found to be predictive of engagement, frustration, and learning. Mouth dimpling was found to be a positive predictor of learning and self-reported performance. The authors also acknowledged that both the intensity and frequency of facial expressions could be used to predict tutoring outcomes. In another study, Grafsgaard et al. (2013a) acknowledged that AU2 was negatively correlated with learning gain, whereas AU4 was positively correlated with frustration. AU14 was positively correlated with both frustration and learning gain.

Bosch et al. (2014) used CERT to track FACS facial features, and the features were used to build classification models to detect five engagment levels—confusion, frustration, boredom, neutral, and engaged. In this study, the engagement levels—confusion and frustration—were detected with a higher chance than the levels—boredom, neutral, and engaged. AU45 was found to be a predictive feature to identify confusion and frustration, where AU1 or a combination of AU1 and AU4 was found as a separator for the confusion from the frustration. The authors also conducted experiments in different

real-world settings in a school environment with some unique challenges and achieved similar success (Bosch 2016; Bosch et al. 2015, 2016). For example, Bosch et al. (2016) used FACET (commercial version of CERT) to estimate the presence of 19 AUs along with head poses, head positions, and interaction patterns of learners' with learning environments. Features extracted from these patterns were used with C4.5 trees and Bayesian classifiers. The most common affective states observed by the authors were engagement, followed by frustration, boredom, delight, and confusion. Vail et al. (2016a) examined the responses received from CERT with the skin conductance responses, postures, and gestures. Authors argued that among the modalities they used, facial expressions and skin conductance responses were found to be highly predictive of learning gain. CERT was used to analyze different AUs of learners, and the AU4, AU5, AU15, and AU23 were found to be highly predictive to learners' learning. In another study, Vail et al. (2016b) acknowledged that the intense expression of AU12 or AU5 represented higher engagement. AU12 was likely to be related to higher engagement and AU5 to signifying paying attention to and concentrating on the task.

2D and 3D information from different sensors were combinely used with AUs for engagement detection. Saneiro et al. (2014) analyzed 2D points of a face, 3D head poses, and animation and shape units, where the animation and shape units include jaw lowered, lip stretcher, brow lowered, lip corner depressor, and outer brow raiser from a Kinect camera. This method applied machine learning technique to infer five engagement levels—excited, relaxed, resolution, interested and concentrated—when dealing with cognitive tasks. Psaltis et al. (2017) combined AUs to body motions from Kinect sensors and the gameplay events to detect affective states—*engaged* and *not-engaged*—with the intensity values in ranges $[-2, 0]$ and $[0, 2]$, respectively. Artificial Neural Network (ANN) was used for engagement classification. Sathik and Jonathan (2013) examined different non-verbal communications and AUs to interpret the comprehension level of learners in a virtual classroom.

Appearance-based methods

In *appearance-based* methods, features extracted from whole-face regions are used to generate patterns for engagement classification. Among different feature extraction techniques, Local Binary Patterns (LBP) and Histogram of Oriented Gradients (HOG) are found to be popular for engagement detection. Monkaresi et al. (2017) used LBP in three orthogonal planes (LBP-TOP) of face appearance with Kinect face tracker and heart-rate for learners' engagement detection in educational activities. Both decision-level and feature-level fusion were used with Updateable Naive Bayes, Bayes Net, K-means clustering, Rotation Forest, and Dagging classifiers. Two level engagement detection (i.e., engaged and not-engaged) was done, where the feature-level fusion was found to be more successful than the decision-level fusion when small size dataset was available to train the classifiers. The accuracy of the facial expression-based channels (LBP-TOP and face tracker) was found to be higher than the heart-rate channel. Kamath et al. (2016) presented an instance-weighted multiple kernel learning SVM model that considers vote distributions from crowdsourcing platforms for learner's engagement detection during e-learning sessions.

Deep learning approaches have been used in engagement detection. Kaur et al. (2018) used LBP-TOP and Deep Multi-Instance Learning (DMIL) for engagement detection. Since the labeling of the engagements at frequent intervals in user videos is expensive and noisy, in this research study, the prediction and localization of learner engagement were formulated as a Multi-Instance Learning (MIL) problem and derived baseline scores based on DMIL. The dataset was annotated using crowdsourcing, where the labelers were instructed to label the videos on the basis of their engagement intensity (from facial expressions) ranging from disengaged, barely engaged, engaged, and highly engaged. Gupta et al. (2018) investigated different models of CNN with face appearance features for detecting four engagement levels—engagement, boredom, confusion, and frustration. Each of the engagement levels were further ranked from low to high scales.

Features extracted from face appearance are combined with different visual cues to enhance learners' engagement detection. Happy et al. (2013) automatically identified learners' cognitive state using non-intrusive visual cues—facial expression, ocular parameters, gestures, and postures—captured by webcam. The LBP features extracted from the face appearance were used to determine the basic emotions such as happiness, surprise, anger, fear, and sadness. SVM was used to identify the eye state as open or closed to observe the interest of learner. Body movements from vision cue was used to understand attention and interest. The postures of hand and head gestures were used to detect boredom and frustration. By combining information from the above cues, the system inferred the learner's state of alertness to generate appropriate feedback. Hwang and Yang (2008) proposed a fuzzy-based engagement detection method using face appearance. The moves of facial features, distances between facial features and facial edges, and records of mouse and keyboard operations in computers were used to evaluate learners' engagement. Drowsiness, turning head to talk, and leaving seats were identified as a low engagement attitude in the context of online learning in this research study.

Face appearance information along with 2D and 3D information collected from different sensors are used for engagement detection. Frank et al. (2016) proposed a framework for engagement detection that includes facial appearance, voice, body postures and motion using 2D and 3D sensors. A SVM classifier is used to classify in one of the six individual engagement levels — disengagement, relaxed engagement, involved engagement, intention to act, action, and involved action. This method is applied to detect engagement levels in a group meeting. Khelfallah et al. (2015) proposed a web-based intelligent tutoring system called Remote Laboratory that allowed learners from anywhere to use the Internet and perform computational experiments in real laboratory equipment, where the learners' levels of engagement were examined in terms of frustration and serenity by using 70 small classifiers.

Gestures and postures

Gesture and postures are two important forms of non-verbal communication through our body language. These are important components of embodied affect with ties to cognitive-affective states that may help or hinder learning. Grafsgaard et al. (2013c) analyzed gestures and postures in a computer-mediated tutorial dialogue, where relationships between learner

postures, gestures, dialogue, and tutor were investigated. Hand-to-face and hand-over-face gestures were found to be promising for informing the runtime behavior of tutoring. Gestures and postures combined key mechanisms of holistic methods of nonverbal behavioral communication and included affects while learning. In this study, learner data was collected from database logs, webcam video, skin conductance, and Kinect depth video to infer the knowledge about learners' engagement. Hand gestures were also used by Tofighi et al. (2016) to identify disengagement, attention, intention, and actions (DAIA). In DAIA, several binary classifiers were designed to detect various hand movements, e.g., raise hands above the waist, different levels of hand speeds, and so on. These classifiers detect user intention for performing an action. A Finite State Transducer (FST) of engagement detection was finally used to flow among different emotional states by analyzing the decisions by the classifiers.

Some research studies focused on revealing learner-tutor interactions by analyzing gestural-activities in a spatial and temporal domain (Sathayanarayana et al. 2014). In these systems, gestures are captured by an overhead camera and then manually labeled as deictic, beat, iconic, and writing for the training of a classifier. Sathayanarayana et al. (2014) employed visual deictic gestures to understand learner-tutor interactions. A graph-based visual saliency (GBVS) (Harel et al. 2006) was used to detect potential deictic gestural regions in the input image. Then an HOG and SVM based classifier was used to determine actual deictic gesture points from candidate regions. Learner-tutor interactions were finally inferred from incorporating gestural-activity information in a spatial and temporal domain. Among different visual cues from gestures and posters, the deictic gestures found to be as a key component to reveal learner-tutor interactions in this research study.

Eye movement

Users' gazes and regions of interests from eye trackers have been used to understand the moods of learners while engaging in any educational activity in online learning. Aslan et al. (2014) used an eye tracker to detect the users' gazes and combined this information with statistical facial features and depth information. Nine pilot sessions on five machine learning algorithms—decision trees, random forest, naive Bayes, logistic regression, and multilayer perceptron—were tested for engagement detection. The authors acknowledged that the use of touch-free 2D and 3D cameras to collect the above information enabled the system to get more accurate facial landmarks and achieved better results for engagement detection. Krithika and Lakshmi (2016) employed the moving patterns of eyes with head motions to infer information on concentration levels in an e-learning environments.

Raina et al. (2016) presented an eye-tracking-based model to reduce content skipping, thus enhancing engagement in online learning. Two learning modules — one with a large amount of content on a single screen (linear) and the other with the same content broken into smaller chunks — were tested. Learners required to go through the content and answer a set of questions. Authors examined learners' content skipping behavior between linear and segmented modules based on reading scores (Buscher et al. 2008) and reading depths by eye tracking.

Although the methods based on eye-tracking are effective, the main challenge of these methods is proper eye-calibration. To receive accurate data precision, these methods require several calibration rounds for each participant. Participants wearing eye-glasses or having eye disorders have difficulty in calibration and often need to be excluded from the studies (Raina et al. 2016). Another major challenge is restricting participants' movement to stay within an eye-tracker range which is not feasible in a real-life educational environment.

Dataset and evaluation techniques

Information about open and online datasets and knowing about metrics and evaluation techniques might be of great interest for researchers in any research field. A review on available datasets, metrics for evaluation and evaluation techniques suitable for users' engagement detection in the context of learning, and some research results are discussed in this section.

Dataset

The need for large, labeled, publicly available datasets for training, evaluating, and benchmarking has been widely acknowledged, and a number of efforts to address this need have been made in the last few years. In user engagement detection, while many research studies use their in-house datasets, very few of them are made publicly available online. A summary of publicly available and annotated datasets are listed in Table 2.

Gupta et al. (2018) created the DAiSEE dataset with an intent to capture learners' engagement in online courses. This dataset includes 112 individuals, where 80 male and 32 female. The videos in the dataset were collected in unconstrained environments, such as at dorm rooms, crowded lab spaces and libraries, with three different illumination settings—light, dark, and neutral. The videos were captured with a webcam mounted on a computer focusing on learners' watching some video tutorial. The annotation of the video frames were done in four different levels—engaged, bored, confused, and frustrated, by relying on the "wisdom-of-the crowd". The annotations were further rated from 0 to 3 based on the intensity. The advantage of the above annotation is that it can be changed to any other n-levels of engagement as required. The dataset HBCU (Whitehill et al. 2014) were taken from 34 individuals from two different pools, where 9 male and 35 female. In both pools, individuals participated in

Table 2 Datasets for learners' engagement detection

Datasets	Videos	Number of subjects	Emotions	Modality
DAiSEE (Gupta et al. 2018)	9068	112 (80 male and 32 female)	engaged, frustration, boredom, confusion	facial expressions
HBCU (Whitehill et al. 2014)	120	34 (9 male and 25 female)	not engaged, nominally engaged, engaged, very engaged	facial expressions
in-the-wild (Kaur et al. 2018)	195	78 (25 female and 53 male)	disengaged, barely engaged, normally engaged, highly engaged	facial expressions
SDMATH (Sathayanarayana et al. 2014)	20	20 (10 male and 10 female)	deictic gestures	speech, gestures, eye gaze, facial expressions

Cognitive Skills Training study arranged by the Historically Black College/University (HBCU) and the University in California (UC). Annotation of the dataset was done manually by human experts.

Kaur et al. (2018) introduced "in-the-wild" dataset, where videos were captured from 78 individuals — 25 females and 53 males. The dataset was collected in an unconstrained environment, such as at a computer laboratory, hostel rooms, and an open ground, via Skype. The annotation of the data was done to one of the four possible engagement levels—disengaged, barely engaged, normally engaged, and highly engaged based on crowdsourcing. Sathayanarayana et al. (2014) introduced the dataset SDMATH, where the videos were captured for one-to-one mathematics tutoring sessions. This dataset offers a set of richly labeled data with both video and audio modalities.

Although each of the datasets mentioned above has their own characteristics and advantages, they do have some limitations. In these datasets, the videos were recorded with limited participants from a particular race. For example, all the participants in HBCU datasets are African-American, whereas in DAiSEE and "in-the-wild" are Asian. Also, the male-female ratio in the datasets are high. The above issues may cause generalization problem during training and testing with the classifiers. Another limitation of these datasets is the ambiguity in labeling the frames with appropriate engagement levels. The frames in the DAiSEE and "in-the-wild" are labeled based on crowdsourcing, where as in the HBCU by human experts. In both cases, ambiguity in labeling frequently occur due to not having a clear guideline for mapping facial indicators to different affective states or engagement levels of the online learners. To alleviate this issue, frames with ambiguous labeling are often removed during the experiments, which eventually reduces sizes and removes the diversity of information in the datasets. Visual cues along with users' activity in the learning environment, self-evaluation and transfer learning could further be investigated to solve the above problems.

Evaluation techniques and metrics

Different metrics and techniques are used to evaluate the performance of engagement detection systems. A widely used technique for evaluating the performance of engagement detection is the investigation of correlation between *human* and *automatic perceptions* of engagements. Whitehill et al. (2014) compared the *automatic perceptions* of engagement and the learner *pre-* and *post-test performance* to evaluate the performance of an engagement detection system. Grafsgaard et al. (2013a, 2013b) evaluated the performance by comparing the results of automated engagement detection with manual annotations. Cohen's Kappa, R^2, multinomial logistic regression, Pearson's correlations and Krippendorff's alpha are often used for correlation-measurement (Kaur et al. 2018; D'Mello et al. 2009; Vail et al. 2016b).

For the *automatic perceptions* of engagement, different classifiers are used, where the accuracy of classification is measured in a *Receiver Operational Characteristic* (ROC) space. In the ROC space, the area under the curve (AUC) is estimated that provides a global measure of the system performance (Monkaresi et al. 2017; Bosch et al. 2015; Bosch et al. 2016). In practice, an empirical ROC curve is obtained by connecting the observed *tpr* and *fpr* for a classifier at each threshold. The AUC assesses ranking in

terms of class separation – the fraction of positive–negative pairs that are ranked correctly. For instance, with an AUC = 1, all positives are ranked higher than negatives indicating a perfect discrimination between classes. A random classifier has an AUC = 0.5, and both classes are ranked at random. The partial AUC, pAUC (5%), is measured by taking the AUC at $0 < fpr \leq 5\%$ from the ROC curve.

Class priors for positive samples and negative samples may vary over time in a real scenarios. Traditional ROC analysis cannot distinguish between two classifiers for specific class miss-classification costs. ROC curves and the AUC allow for a performance evaluation that is independent of costs and priors by integrating performance over a range of decision thresholds. However, it is important to observe performance as the proportion of the correctly predicted positive samples out of the total number of input samples predicted to belong to an affective state. Otherwise, when processing highly imbalanced data, and the minority positive samples are of interest, a system may outperform others by predicting a very large number of samples as minority, resulting in an increased *tpr* at the expense of an increased *fpr* (Sathayanarayana et al. 2014; Cocea and Weibelzahl 2009, 2011). Given the imbalance between a target and non-target captures, performance is assessed in the Precision–Recall (PR) space (Bosch et al. 2015), where the area under the PR curve (AUPR) provides another global scalar measure.

The 2-alternative forced choice (2AFC) is another measure which expresses the probability of correctly discriminating a positive example from a negative example in a 2-alternative forced choice classification task (Whitehill et al. 2008; Fei and Pavlidis 2010; Mason and Weigel 2009). The 2AFC is an unbiased estimate of the area under the ROC curve, which is commonly used in the facial expression recognition literature. A 2AFC value of 1 indicates perfect discrimination, whereas 0:5 indicates that the classifier is "random by the chance".

When evaluating machine learning models, the validation step helps to find the best parameters for the classification model while also preventing it from becoming over-fitted (Booth et al. 2017; Cocea and Weibelzahl 2009). Two of the most popular strategies to perform the validation step are the *hold-out strategy* and the *k-fold strategy*. In the case of *hold-out strategy*, it uses fully independent data and only needs to be run once so has lower computational costs. However, this strategy is subject to higher variance given the smaller size of the data. In *k-fold* cross-validation strategy, the set of all labeled frames are partitioned into k folds such that no engagement label appear in more than one-fold; hence, the cross-validation gives an estimate of how well the classifier would perform on the various engagement levels on which the classifier was trained (Monkaresi et al. 2017). The advantage of *k-fold* is that it is prone to less variation because it uses the entire training set. The limitation of this strategy is that it requires higher computational costs as the model needs to be trained k times at the validation step plus one more at the test step.

Results and discussions

Although the main focus of this study is to review existing methods, benchmarking datasets and metrics used for performance evaluation of systems for learners' engagement detection, we also did some discussions on the results from the existing literature. Whitehill et al. (2014) used HBCU dataset for the automatic detection of learners' engagement from facial expressions. As mentioned earlier, this research study compared

three machine learning techniques — Boost (BF), SVM (Gabor), and MLR (CERT) (Whitehill et al. 2014). Four fold subject-independent cross-validation with the 2AFC was done to measure the accuracy for engagement detection. The average accuracies achieved by the MLR (CERT), Boost (BF), and the SVM (Gabor) were 0.714, 0.728 and 0.729, respectively. The performances of the classifiers are also validated by comparing with manual annotations by using Cohen's k, where the above classifiers achieved the correlations 0.275, 0.329, and 0.306, respectively.

Sathayanarayana et al. (2014) used the SDMATH dataset to detect the deictic tip for the hand gesture recognition. This method employed graph based visual saliency (GBVS) and SVM to detect the deictic tip, and achieves the TPR of 85% while considering within 12 pixels from the ground truth annotations. Although this study intended to understand the learner-tutor interaction using visual deictic gestures, the annotations of the dataset and the discussion of the results were limited to deictic tip detection considering this as a preliminary work towards an automatic understanding of the learner-tutor interaction.

Kaur et al. (2018) used the "in-the-wild" dataset for their performance evaluation. This method employed three-fold cross validation with multiple kernel learning (MKL) SVM and the average accuracy and the maximum accuracy obtained 43.98% and 50.77%, respectively. The authors acknowledged the improvement of performance up to 75.77% when the frames with ambiguous labels were removed from the dataset. Gupta et al. (2018) used DAiSEE dataset for engagement detection through using three different models of Convolution Neural Networks (CNNs) — InceptionNet, C3D, and Long-Term Recurrent Convolutional Network (LRCN). The models were applied to detect boredom, engagement, confusion, and frustration, where InceptionNet achieved the accuracies of these engagement levels 36.5%, 47.1%, 70.3%, and 78.3%, respectively. The C3D and LRCN achieved the accuracies of these engagement levels of 45.2%, 56.1%, 66.3%, 79.1%, and 53.7%, 61.3%, 72.3%, 73.5%, respectively. Like the previous study (Kaur et al. 2018), the authors of this study also acknowledged an improvement in accuracies by removing the ambiguous annotations from the dataset.

We have some observations about the results we discussed above. Annotations of the datasets are clearly playing an important role in improving the accuracies of engagement detection. Since this is still not clear how many engagement levels should be considered during learners' learning activities and what facial expressions are linked with what engagement levels, ambiguous labeling of the frames frequently occurs. Many of the research studies reported that removing the images on which there is a large label disagreement help to improve the engagement detection results. However, this could bias the results to be too optimistic because the "harder" images might be ones on which labelers tend to disagree. Also, it is suggested that the larger number of images available for training can compensate for the noisier labels. Another observation is that although many papers do performance evaluation of their systems on publicly available datasets, this does not necessarily lead to a true comparison among the systems. The way in which systems are trained and evaluated can differ significantly, leading to incomparable results (Martinez et al. 2017). Moreover, the issue of unbalanced data makes comparisons harder even further (Jeni et al. 2013). Building personalized models using online and transfer learning methodologies (Chen et al. 2013; Chu et al. 2017) is the way forward in our opinion. This is due to several reasons, as the lack of training

data, large subject differences, and the dependency of the displayed expressions on a large number of factors such as environments, tasks or moods, which would be hard to cover exhaustively even if much larger amount of training data was available.

Conclusions and future works

This paper has presented a review of engagement detection methods in learning context. In the review, a taxonomy of state-of-the-art methods is proposed with two levels. Although the *computer vision based methods* are found to be promising in engagement detection, they do have some limitations. Automatic gathering and analyzing the behavioral data in naturalistic scenarios is still challenging for the *computer vision based methods*. For example, the existing algorithms face challenges to analyze head motion and facial occlusions. In such a situation, these algorithms are not able to extract features from some video segments, thereby leading to data loss. Another challenge is to extract robust features from the region of interests due to segmentation error. Although a lot of attention has been given towards deploying facial expression analysis, the challenges encountered in these endeavors are not only in terms of technical issues.

Until now very few datasets are available online that can be used for engagement detection in the context of online learning. However, the importance of this kind of datasets has been recognized. Researchers are paying more and more attention for creating this kind of datasets and making the datasets publicly available. In creating datasets for engagement detection from facial expressions, three major challenges are encountered by researchers. Many research studies stated that the link between specific facial expressions to a particular type of learning activities (e.g., reading, writing, participating in the online meeting and watching online video tutorials) is hard to define. It is also not clear enough that how many affective states or engagement levels (or types) are effective to recognize an online learner when fine discrimination of his/her engagement is needed. Another potential pitfall is the frequency with which affective states should be reported in an input video.

This is not clear enough how frequently the decision on engagement detection should be made – frame by frame, a short fragment of a video or an entire video clip? In case of a short fragment, what the length of a video clip is suitable to assign a single level? During labeling training data, it is unclear what exactly should be the standard for deciding what emotions a learner is truly having. Should it be the learner or the trained judges? Although the highest interrater reliability was obtained between the trained judges, it might nothing more than an artefact brought on by the training. This is also not clear what environmental constraints are needed to be considered while capturing videos for engagement detection in the context of online learning. Many research studies also emphasized the importance of doing further research to know what direct correlation the engagement detection results have to the actual task performance.

Addressing the above challenges can contribute to advance the research of automatic engagement detection in a computerized educational environment and lead to more effective learning and a more engaging experience for learners. Along with these, we make the following recommendations for further improvement in this research field.

As also many other researchers suggested in their research studies (Psaltis et al. 2017; Grafsgaard et al. 2013c; Aslan et al. 2014) that combining different modalities can help to improve the accuracy of engagement detection. For example, facial expression, eye

tracing, body parts motion, ocular parameters, gestures, postures, voice, and gaze are needed to experiment with biometric information (e.g., galvanic skin response, heart rate, electromyography of the jaw, respiration rate, respiration amplitude) collected from learners' smartwatch and brainwave-sensing eyeglasses (Koydemir and Ozcan 2018). Features extracted from the *engagement tracing*, *self-reporting* and *observational check-list* can also be experimented with the above automatically extracted features to improve in engagement detection results.

Future research should focus on setting annotation criteria to label the benchmarking datasets. It is acknowledged that the affective states relevant to learning such as frustration, boredom, learning gain or the levels of engagement are more difficult to define than the commonly studied domains of emotion recognition, i.e., happy, sad, angry, disgusted, fearful, surprised, or neutral or facial AUs classification (Whitehill et al. 2014). Hence, arriving at a sufficiently clear definition and devising an appropriate labeling procedure, including the timescale at which labeling should be taken place, is important for ensuring both the reliability and validity of the training labels (Porayska-Pomsta et al. 2013). While determining useful features for detecting learner's engagement remains a challenging question, a detailed understanding of mapping between the features and the affective states will also be needed (Fairclough and Venables 2006). Intuitively there remain subjective interpretations, however, the quality of labeling is highly dependent on the expertise of the human annotator (Cocea and Weibelzahl 2009). Annotators form different backgrounds may label engagements differently. To maintain consistency, annotations done by experts and crowdsourced labels need to be combined and evaluated during the ground truth validity. Additionally, feasible collaborative sensory data can also need to be combined for ground truth validity (Fairclough and Venables 2006).

Future research in engagement detection with automatic AU detection would be a great benefit. AU detection based engagement detection methods use pre-defined learned models. However, these models require manual AU labeling in the context of learning which is challenging. It is also difficult to share common labeled datasets and tools with the other researchers. Thus, further initiative is needed to develop tools and benchmark standards to create and share common datasets and labeling instead of accessing them privately. Another problem in AU based research is that the co-occurrences of AUs are much harder to model since not all AUs are independent and treating their combinations as a new label is impractical Vail et al. (2016b). Therefore, Wang et al. (2013) suggest to consider modeling the "semantics" of behavior learning (i.e., temporal co-occurrences of AUs) to maintain the consistency. Combining this knowledge can benefit existing research further.

Recent advances in machine learning tools, such as CNNs and DBNs, require more data volumes than currently available. Collecting and analyzing behavioral data in naturalistic scenarios is itself a challenging issue. Learner engagement detection systems cannot be useful unless we can address the issues related to environmental constraints. Some of the challenges include illumination variation, occlusions, head poses, errors in speech detection, objects appearing too far or close, and so on. So far, very limited attempts have been taken to resolve this problem. For example, D'Mello and Graesser (2010) utilize a general body language to alleviate the difficulty in detecting boredom from facial expression and dialogue. Valstar et al. (2017) limit some of the challenges

using a specialized environment with a head-pose monitoring mechanism. However, more efforts are needed to address these open challenges related to the real-world learning environment constraints.

Learners' engagement detection is biased toward demographic variables, learners' age, and geographic locations (Sundar and Kumar 2016). Further research is needed to test detectors on a larger dataset with more demographic variability. Future study should also investigate *what, how, when* and *why* learners' get disengaged and how to re-engage them effectively. Future research should go more detail into the temporal domain and investigate at what frequency an engagement expression appears and how quickly it goes away. Further effort should also be given to examine how engaged/disengaged behaviors are associated with learning outcomes.

Acknowledgements
Authors would like to thank Athabasca University, Canada, to support this research through the Academic Research Fund (ARF) and Research Initiative Grant (RIG).

Funding
Academic Research Fund (ARF) and Research Incentive Grant (RIG), Athabasca University, Canada.

Availability of data and materials
Not applicable.

Authors' contributions
All authors contributed equally. All authors read and approved the final manuscript.

Competing interests
The authors declare that they have no competing interests.

Publisher's Note

Received: 13 July 2018 Accepted: 7 December 2018
Published online: 03 January 2019

References

M. Al-Hendawi, Academic engagement of students with emotional and behavioral disorders: Existing research, issues, and future directions. Emotional and Behavioural Difficulties **17**(2), 125–141 (2012)
T. Aluja-Baneta, M.-R. Sanchob, I. Vukic, Measuring motivation from the virtual learning environment in secondary education. Journal of Computational Science, 1–7 (2017)
A.R. Anderson, S.L. Christenson, M.F. Sinclair, C.A. Lehr, Check and connect: The importance of relationships for promoting engagement with school. J. Sch. Psychol. **42**(2), 95–113 (2004)
S. Aslan, Z. Cataltepe, I. Diner, O. Dundar, A.A. Esme, R. Ferens, M. Yener, *Learner Engagement Measurement and Classification in 1:1 Learning* (International Conference on Machine Learning and Applications, Detroit, 2014)
M. Bartlett, G. Littlewort, M. Frank, C. Lainscsek, I. Fasel, J. Movellan, Automatic recognition of facial actions in spontaneous expressions. J. Multimed. **1**(6), 22–35 (2006)
J.E. Beck, *Engagement Tracing: Using Response Times Tomodel Student Disengagement* (Conference on Artificial Intelligence in Education, Amsterdam, 2005)
B.M. Booth, A.M. Ali, S.S. Narayanan, I. Bennett, A.A. Farag, *Toward Active and Unobtrusive Engagement Assessment of Distance Learners* (International Conference on Affective Computing and Intelligent Interaction, San Antonio, 2017)
N. Bosch, *Detecting Student Engagement: Human Versus Machine* (Conference on User Modeling Adaptation and Personalization, Halifax, 2016)
N. Bosch, Y. Chen, S. D'Mello, *It's written on your face: Detecting affective states from facial expressions while learning computer programming* (Intelligent Tutoring Systems, Honolulu, 2014)
N. Bosch, S.K. D'Mello, R.S. Baker, J. Ocumpaugh, V. Shute, M. Ventura, W. Zhao, *Automatic Detection of Learning-Centered Affective States in the Wild* (International Conference on Intelligent User Interfaces, Atlanta, 2015)
N. Bosch, S.K. D'Mello, R.S. Baker, J. Ocumpaugh, V. Shute, M. Ventura, W. Zhao, *Detecting Student Emotions in Computer-Enabled Classrooms* (International Joint Conference on Artificial Intelligence, New York, 2016)
G. Buscher, A. Dengel, L.v. Elst, *Eye Movements as Implicit Relevance Feedback* (Conference on Human Factors in Computing Systems, New York, 2008)
M. Chaouachi, P. Chalfoun, I. Jraidi, C. Frasson, *Affect and Mental Engagement: Towards Adaptability for Intelligent Systems* (Florida Artificial Intelligence Research Society Conference, Florida, 2010)
J. Chen, X. Liu, P. Tu, A. Aragones, Learning person-specific models for facial expression and action unit recognition. Pattern Recogn. Lett., 1964–1970 (2013)
S. Christenson, A. Reschly, C. Wylie, *Handbook of Research on Student Engagement* (Springer, New York, 2012)

S.L. Christenson, A.R. Anderson, The centrality of the learning context for students' academic enabler skills. School Psychological Review **31**(3), 378–393 (2002)
W.-S. Chu, F.D. Torre, J.F. Cohn, Selective transfer machine for personalized facial expression analysis. IEEE Trans. Pattern Anal. Mach. Intell., 529–545 (2017)
M. Cocea, S. Weibelzahl, Log file analysis for disengagement detection in e-learning environment. User Model. User-Adap. Inter. **19**, 341–385 (2009)
M. Cocea, S. Weibelzahl, Disengagement detection in online learning: Validation studies and perspectives. IEEE Trans. Learn. Technol. **4**(2), 114–124 (2011)
S. D'Mello, B. Lehman, R. Pekrun, A. Graesser, Confusion can be beneficial for learning. Learn. Instr. **29**, 153–170 (2014)
S.K. D'Mello, S.D. Craig, A.C. Graesser, Multimethod assessment of affective experience and expression during deep learning. Int. J. Learn. Technol. **4**(3), 165–187 (2009)
S.K. D'Mello, A. Graesser, Multimodal semi-automated affect detection from conversational cues, gross body language, and facial features. User Model. User-Adap. Inter. **20**(2), 147–187 (2010)
A. Dewan, F. Lin, D. Wen, M. Murshed, Z. Uddin, *A Deep Learning Approach to Detecting Engagement of Online Learners* (IEEE International Conference on Internet of People, Guangzhou, 2018)
P. Ekman, W. Friesen, *Facial Action Coding System: A Technique for the Measurement of Facial Movement* (Consulting Psychologists Press, 1978)
P. Ekman, W. Friesen, J. Hager, *Facial Action Coding System* (A Human Face, 2002)
S. Fairclough, L. Venables, Prediction of subjective states from psychophysiology: A multivariate approach. Biol. Psychol. **71**(1), 100–110 (2006)
J. Fei, I. Pavlidis, Thermistor at a distance: Unobtrusive measurement of breathing. IEEE Trans. Biomed. Eng. **5**(4), 988–998 (2010)
M. Frank, G. Tofighi, H. Gu, R. Fruchter, *Engagement Detection in Meetings* (International Conference on Computing in Civil and Building Engineering, Osaka, 2016)
J.A. Fredrick, P.C. Blumenfeld, A.H. Paris, School engagement: Potential of the concept, state of the evidence. Rev. Educ. Res. **74**(1), 59–109 (2004)
B. Goldberg, R. Sottilare, K. Brawner, H. Holden, *Predicting Learner Engagement during Well-Defined and Ill-Defined Computer-Based Intercultural Interactions* (International Conference on Affective Computing and Intelligent Interaction, Memphis, 2011)
J.F. Grafsgaard, J.B. Wiggins, K.E. Boyer, E.N. Wiebe, J.C. Lester, *Automatically Recognizing Facial Indicators of Frustration: A Learning-Centric Analysis* (International Conference on Affective Computing & Intelligent Interaction, Geneva, 2013a)
J.F. Grafsgaard, J.B. Wiggins, K.E. Boyer, E.N. Wiebe, J.C. Lester, *Automatically Recognizing Facial Expression: Predicting Engagement and Frustration* (International Conference on Educational Data Mining, Memphis, 2013b)
J.F. Grafsgaard, J.B. Wiggins, K.E. Boyer, E.N. Wiebe, J.C. Lester, *Embodied affect in tutorial dialogue: Student gesture and posture* (Artificial Intelligence in Education, Memphis, 2013c)
A. Gupta, A. D'Cunha, K. Awasthi, V. Balasubramanian, DAiSEE: Towards user engagement recognition in the wild. Computer Vision and Image Processing. arXiv preprint arXiv **1609**, 01885 (2018)
S.L. Happy, A. Dasgupta, P. Patnaik, A. Routray, *Automated Alertness and Emotion Detection for Empathic Feedback during e-Learning* (IEEE 5th International Conference on Technology for Education, Kharagpur, 2013)
J. Harel, C. Koch, P. Perona, *Graph-Based Visual Saliency* (International Conference on Neural Information Processing Systems, Vancouver, 2006)
L.R. Harris, A phenomenographic investigation of teacher conceptions of student engagement in learning. Aust. Educ. Res. **35**(1), 57–79 (2008)
K. Hwang, C. Yang, *Fuzzy Fusion for Affective State Assessment in Distance Learning Based on Image Detection* (International Conference on Audio, Language and Image Processing, Shanghai, 2008)
A.L. Jeni, F.J. Cohn, D.F. Torre, *Facing Imbalanced Data Data-Recommendadtions for the Use of Performance Metrics* (Humaine Association Conference on Affective Computing and Intelligent Interaction, Geneva, 2013)
J. Johns, B. Woolf, *A Dynamic Mixture Model to Detect Student Motivation and Proficiency* (Proceedings of the Twenty-First National Conference on Artificial Intelligence, Boston, 2006)
A. Kamath, A. Biswas, V. Balasubramanian, *A Crowdsourced Approach to Student Engagement Recognition in e-Learning Environments* (IEEE Conference on Applications of Computer Vision, New York, 2016)
A. Kapoor, R.W. Picard, *Multimodal Affect Recognition in Learning Environments* (ACM international conference on Multimedia, New York, 2005)
S. Karumbaiah, B. Woolf, R. Lizarralde, I. Arroyo, D. Allessio, N. Wixon, *Addressing Student Behavior and Affect with Empathy and Growth Mindset* (International Conference on Educational Data Mining, Wuhan, 2017)
A. Kaur, A. Mustafa, L. Mehta, A. Dhall, Prediction and localization of student engagement in the wild. Computer Vision and Image Processing **arXiv**, 1804.00858 (2018)
J. Khelfallah, Slama, & Hadj, J. B, Facial expression recognition for intelligent tutoring systems in remote laboratories platform. Procedia Computer Science **73**, 274–281 (2015)
H.C. Koydemir, A. Ozcan, Wearable and implantable sensors for biomedical applications. Annu. Rev. Anal. Chem. **11**, 127–146 (2018)
L.B. Krithika, P.G.G. Lakshmi, Student emotion recognition system (SERS) for e-learning improvement based on learner concentration metric. Procedia Computer Science **85**, 767–776 (2016)
G. Littlewort, J. Whitehill, T. Wu, I. Fasel, M. Frank, J. Movellan, M. Bartlett, *The computer expression recognition toolbox (CERT)* (IEEE International Conference on Automatic Face & Gesture Recognition and Workshops, Santa Barbara, 2011)
B. Martinez, M.F. Valstar, B. Jiang, Automatic analysis of facial actions: A survey. IEEE Trans. Affect. Comput., 1–22 (2017)
S.J. Mason, A.P. Weigel, A generic forecast verification framework for administrative purposes. Mon. Weather Rev. **137**(1), 331–349 (2009)
G. Matthews, S. Campbell, S. Falconer, L. Joyner, J. Huggins, K. Gilliland, J. Warm, Fundamental dimensions of subjective state in performance settings: Task engagement, distress, and worry. Emotion **2**(4), 315–340 (2002)

B. McDaniel, S. D'Mello, B. King, P. Chipman, K. Tapp, A. Graesser, *Facial Features for Affective State Detection in Learning Environments* (Proceedings of the Annual Meeting of the Cognitive Science Society, California, 2007)
H. Monkaresi, N. Bosch, R. Calvo, S. D'Mello, Automated detection of engagement using video-based estimation of facial expressions and heart rate. IEEE Trans. Affect. Comput. **8**(1), 15–28 (2017)
H.L. O'Brien, E.G. Toms, The development and evaluation of a survey to measure user engagement. J. Am. Soc. Inf. Sci. Technol. **61**(1), 50–69 (2010)
J. Parsons, L. Taylor, *Student engagement: what do we know and what should we do?* (University of Alberta, Technical Report, Edmonton, 2011)
K. Porayska-Pomsta, M. Mavrikis, S. D'Mello, C. Conati, R. Baker, Knowledge elicitation methods for affect modelling in education. Int. J. Artif. Intell. Educ. **22**(3), 107–140 (2013)
A. Psaltis, K.C. Apostolakis, K. Dimitropoulos, P. Daras, Multimodal student engagement recognition in prosocial games. IEEE Transactions on Computational Intelligence and AI in Games (2017)
S. Raina, L. Bernard, B. Taylor, S. Kaza, *Using Eye-Tracking to Investigate Content Skipping: A Study on Learning Modules in Cybersecurity* (IEEE Conference on Intelligence and Security Informatics, Tucson, 2016)
L. Rothkrantz, *Dropout Rates of Regular Courses and MOOCs* (International Conference on Computer Supported Education, Rome, 2016)
M. Saneiro, O.C. Santos, S. Salmeron-Majadas, J.G. Boticario, Towards emotion detection in educational scenarios from facial expressions and body movements through multimodal approaches. Sci. World J., 1–14 (2014)
S. Sathayanarayana, R.K. Satzoda, A. Carini, M. Lee, L. Salamanca, J. Reilly, G. Littlewort, *Towards Automated Understanding of Student-Tutor Interactions Using Visual Deictic Gestures* (IEEE Conference on Computer Vision and Pattern Recognition Workshops, Columbus, 2014)
M. Sathik, S.G. Jonathan, Effect of facial expressions on student's comprehension recognition in virtual educational environments. SpringerPlus **2**(455), 1–9 (2013)
D.J. Shernoff, M. Csikszentmihalyi, B. Schneider, E.S. Shernoff, Student engagement in high school classrooms from the perspective of flow theory. Sociol. Educ. **73**, 247–269 (2000)
P. Sundar, S. Kumar, Disengagement detection in online learning using log file analysis. International journal of computer technology and applications **9**(27), 195–301 (2016)
G. Tofighi, H. Gu, K. Raahemifar, *Vision-based engagement detection in virtual reality* (The Digital Media Industry and Academic Forum, Santorini, 2016)
A.K. Vail, J.F. Grafsgaard, K.E. Boyer, E.N. Wiebe, J.C. Lester, *Predicting Learning from Student Affective Response to Tutor Questions* (International Conference on Intelligent Tutoring Systems, Zagreb, 2016a)
A.K. Vail, J.B. Wiggins, J.F. Grafsgaard, K.E. Boyer, E.N. Wiebe, J.C. Lester, *The Affective Impact of Tutor Questions: Predicting Frustration and Engagement* (International Conference on Educational Data Mining, Raleigh, 2016b)
M.F. Valstar, E. Sanchez-Lozano, J.F. Cohn, A. Jeni, L. A, J.M. Girard, Z. Zhang, M. Pantic, Addressing head pose in the third facial expression recognition and analysis challenge. Computer Vision and Pattern Recognition (2017)
Z. Wang, Y. Li, S. Wang, Q. Li, *Capturing Global Semantic Relationships for Facial Action Unit Recognition* (IEEE International Conference on Computer Vision, Sydney, 2013)
J. Whitehill, M. Bartlett, J. Movellan, *Automatic Faacial Expression Recognition for Intelligent Tutoring Systems* (IEEE Conference on Computer Vision and Pattern Recognition, Anchorage, 2008)
J. Whitehill, Z. Serpell, Y.-C. Lin, A. Foster, J. Movel, The faces of engagement: Automatic recognition of student engagement from facial expressions. IEEE Transactions on Affective Computing ***5***(1), 86–98 (2014)
N. Wixon, D. Allessio, S. Schultz , W. Burleson, I. Arroyo, K. Muldner, B.Woolf (2016). . Conference on User Modeling Adaptation and Personalization. Halifax