Zaletelj and Košir *EURASIP Journal on Image and Video Processing* (2017) 2017:80
DOI 10.1186/s13640-017-0228-8

EURASIP Journal on Image and Video Processing

RESEARCH Open Access

Predicting students' attention in the classroom from Kinect facial and body features

Janez Zaletelj* and Andrej Košir

Abstract

This paper proposes a novel approach to automatic estimation of attention of students during lectures in the classroom. The approach uses 2D and 3D data obtained by the Kinect One sensor to build a feature set characterizing both facial and body properties of a student, including gaze point and body posture. Machine learning algorithms are used to train classifiers which estimate time-varying attention levels of individual students. Human observers' estimation of attention level is used as a reference. The comparison of attention prediction accuracy of seven classifiers is done on a data set comprising 18 subjects. Our best person-independent three-level attention classifier achieved moderate accuracy of 0.753, comparable to results of other studies in the field of student engagement. The results indicate that Kinect-based attention monitoring system is able to predict both students' attention over time as well as average attention levels and could be applied as a tool for non-intrusive automated analytics of the learning process.

Keywords: Attention estimation, Human behavior analysis, Learning analytics, Kinect sensor

1 Introduction

Automated learning analytics is becoming an important topic in the educational community, which needs effective systems to monitor learning process and provideş feedback to the teacher. Recent advances in visual sensors and computer vision methods enabled automated monitoring of behavior and affective states of learners at different levels from the university level [1] to the pre-school level [2]. Student affective states such as interested, tired, and confused are automatically determined from facial expressions [2–4], and attention state is computed from different visual cues such as face gaze, head motion, and body postures [1, 5].

The basic idea of our work is to utilize advanced capabilities of Kinect One sensor to unobtrusively collect behavioral data of multiple students during attending traditional lectures in the classroom. We propose a methodology to compute features from the Kinect data corresponding to visually observable behaviors and to apply machine learning methods to build models to predict attentive state of the individual students.

The first issue was defining student attention in the way to correspond to observations made by the teacher or other human observer. We analyze attention scores provided by human observers and match them with the observable behaviors, activities, gestures, etc. of the students (Section 3.4). Those results allow us to define a meaning of observable attention levels in terms of student behavior. The second issue was selection and derivation of meaningful features which can effectively discriminate attention levels. We selected features, provided by Kinect One feature detection system, which were correlated to observable behaviors and attention levels. The proposed set of features for attention estimation, derived from low-level Kinect features, is described in Section 4.3.

The final issue was selection of the appropriate machine learning method, which is able to learn a generalized attention model, applicable to any student or person in the classroom. We tackle this problem by preparing five combinations of input features and machine learning classifiers and data splitting strategies and analyze their accuracy on the test set of 18 persons (Section 5). We present detailed evaluation of results including comparison of performance of classifiers and discussion of method limitations.

*Correspondence: janez.zaletelj@fe.uni-lj.si
Faculty of Electrical Engineering, University of Ljubljana, Trzaska 25, 1000 Ljubljana, Slovenia

2 Related works

2.1 Student engagement and attention in the classroom

In the field of higher education, estimation of a long-term student engagement in the learning process is needed in order to evaluate courses and improve learning results [6, 7]. This evaluation is usually done through questionnaires, but with the proliferation of modern e-learning, it became possible to collect implicit usage data to estimate activity and engagement of students or children within learning activities [8]. A review of research on measurement of student engagement in technology-mediated learning [9] have provided a review of quantitative and qualitative observational measures (instruments) to measure behavioral, cognitive, and emotional indicators of student engagement. Attention was classified as one of the factors of cognitive engagement, while interest, anxiety, and boredom contributed to emotional engagement.

Attention is best described as the sustained focus of cognitive resources on information, while ignoring distractions. In the field of education, the terms of sustained attention or vigilance are used to describe the ability to maintain concentration over prolonged periods of time, such as during lectures in the classroom. Pedagogical research is often focused on maintaining student attention (concentration, vigilance) during lectures [10], because sustained attention is recognized as an important factor of the learning success [11]. However, tracking of individual students' attentive state in the classroom by using self-reports is difficult and interferes with the learning process, which is also the case for using psychophysical data sensors [12]. Visual observation is a non-intrusive method, and real-time video recording and encoding [13–16] can be used for manual attention coding; however, for long-term observations, automatic computer vision methods should be applied.

2.2 Automated measurement of affective parameters

Non-intrusive visual observation and estimation of affective parameters is commonly using recorded video (RGB) signal, for example, to estimate student engagement from facial expressions [3, 17], to estimate mood of children during one-to-one tutoring by using facial analysis [2], and to estimate driver's vigilance from his head pose [18]. A survey of automatic affect detection methods [4] identified various types of signals (video, EKG, EMG...) used in affect analysis. Video observation and face analysis usually require high-quality image and are applicable to single-person observation, which limits their usability or reduces accuracy and available complexity of image analysis [1] in the classroom setting. Eye tracking devices are very successful in measuring affective parameters such as concentration in the computerized learning environments, and Bixler et al. were using eye tracking data detect mind wandering during computerized reading [19]. Apart from visual signals, other types of measurements such as brainwaves (EEG) were utilized to assess attention level of students [12, 20].

Two state-of-the-art studies use machine learning methods to build models for automated estimation of student engagement from facial features. Monkaresi et al. [17] use combination of geometrical facial features (detected by Kinect sensor), texture description features (local binary patterns), and physiological features (heart rate) to estimate two-level engagement of students. Whitehill et al. [3] use computer vision methods to register faces and extract Box Filter features (Haar wavelets) and then train binary classifiers to estimate four states of engagement. Both works are studying the engagement of a single student during computerized learning which differs from our use case of attention during classroom teaching.

2.3 Kinect sensor and its usage

The introduction of low-cost depth sensors aimed at computer games such as Microsoft Kinect inspired a lot of research in various applications, especially those requiring detection of body skeletons. A review paper [21] provided an overview of usage of first-generation Kinect 360 sensor for human activity analysis, including body pose and activity recognition, and hand gesture analysis; however, they do not include any references on using head gaze information. Recent studies have utilized Kinect for gait assessment [22] and online human action recognition such as writing and cooking [23], and Won et al. [5] utilized two Kinect sensors to record body motions of teacher and the student during dyadic learning interactions to predict learning performance.

The review of affect estimation methods [4] states the importance of face gaze and facial expression as clues to assess cognitive engagement or inattention of students. The gaze direction have been detected from combined video and depth signals [24, 25] and utilized in the visual attention model to estimate human-to-human interaction. Human gaze has also been used for semantic mapping of human attention in the 3D environment [26]. Kinect One sensor provides advanced capabilities to detect face gaze and facial features, which have not yet been explored in available literature and are utilized in the proposed system.

3 Experimental methods

In this section, we present the experimental setup to acquire the test dataset, methods of data annotation, analysis of attention levels as observed by annotators, and their correspondence to student behavior.

3.1 Experimental setup

The goal of our experiments was to record student behavior in the classroom during lecture, thus obtaining video

and 3D data which would allow both human observation as well as automated analysis of their attention. The Kinect One sensor was set up to observe up to four students acting as test persons.

The participants of the experiment were 22 undergraduate engineering students from a public university in Slovenia; there were 20 males and 2 females. Video and 3D data were recorded during 25-min lecturing sessions by Kinect One sensor. The four students were not reliably detected and tracked by Kinect body tracking engine, and they were not included in the dataset. The experimental dataset was obtained during four lecturing sessions in a classroom. The first two sessions were using a single Kinect sensor to frontally observe three students sitting behind a table from a distance of 1.8 m. The last two sessions involved two sensors, each one observing four students from the same distance. Students were asked to follow a lecture and take notes and answer questions prior and after the experiment. The relation between learning gain and observed attention was studied in our previous paper by Burnik et al. [27].

3.2 Data collection methods

The Kinect data was recorded by Matlab scripts using the methods provided by Kin2 Toolbox for Matlab, which encapsulates the Microsoft Kinect SDK 2.0. The real-time feature extraction phase captured the video and skeleton data during the experiment and recorded data on the disk drive. The offline processing and analysis of extracted data was performed by Matlab scripts.

The recording system extracted and stored several types od data provided by Kinect SDK's body and face tracking engine. Color frames with full HD resolution (1920 by 1080 pixels) were extracted at the frame rates up to 15 frames per second, encoded and stored as H.264 video file. Depth frames with resolution 512 by 424 pixels were recorded but were not used for attention prediction.

Kinect body skeleton is the most exploited 3D feature in literature [5, 22]. Skeleton is given as a set of 25 body joints, where each body joint is represented by a vector $p_j = [p_{j,x}, p_{j,y}, p_{j,z}]$ given in Kinect 3D coordinate system. As shown in Fig. 1, the origin of the coordinate system ($x = 0, y = 0, z = 0$) is located at the center of the Kinect IR sensor, the x coordinate grows to the sensors left, y grows up, and z grows out in the direction the sensor is facing. The unit is 1 m. At each time instance t, up to six skeletons can be detected and tracked. Skeleton indexing is however not consistent through time due to persons' occlusions or disappearance from the scene. The k-th skeleton at time t is given by 25 body joints, and we denote it as $S_k(t) = \{p_{1,k}(t), p_{2,k}(t), \cdots, p_{25,k}(t)\}$. Upper parts of the test persons' skeletons are visualized on the video frame in Fig. 5a.

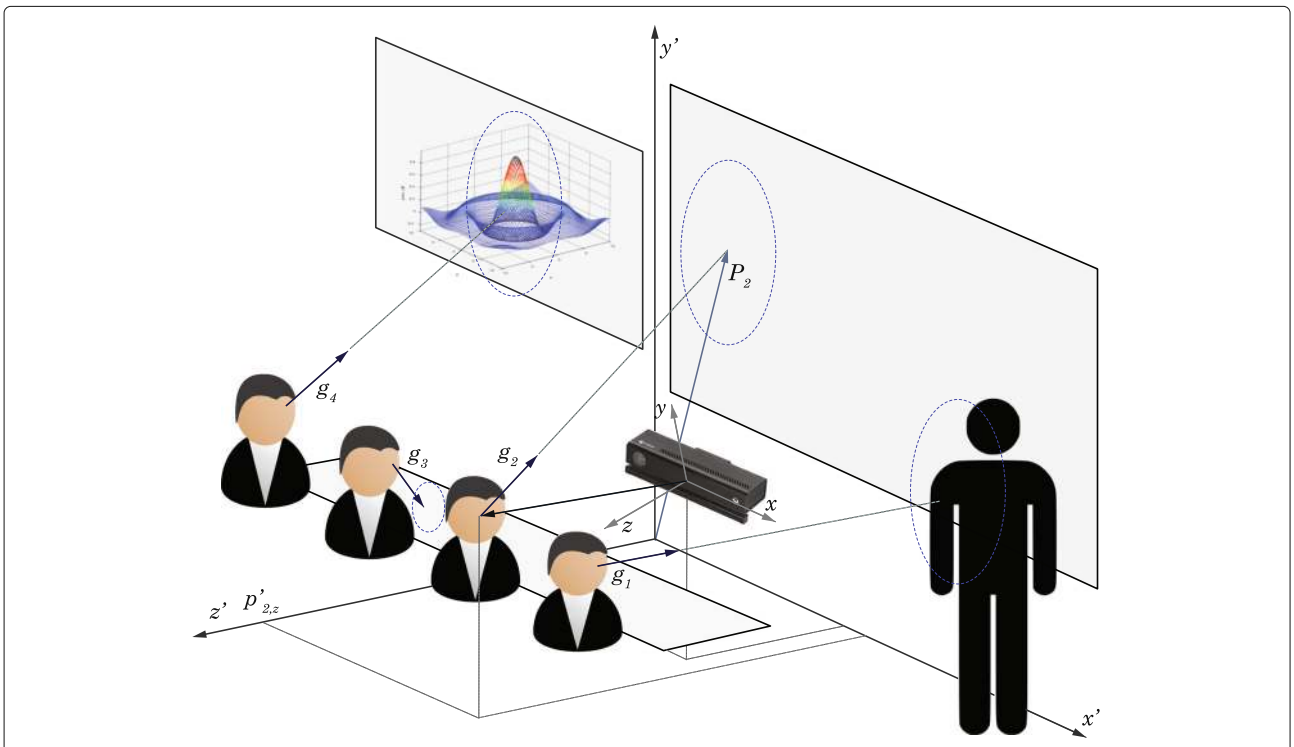

Fig. 1 Student observation in the classroom. The Kinect One sensor is recording behavior of multiple students to enable analysis of their gaze point and attention level during the lecture

The second set of 3D features is provided by Kinect SDK 2.0 face tracking engine. First, a detailed 3D face model composed of 1347 mesh vertices was recorded for each skeleton. Second, face orientation in 3D space was recorded, given by face yaw, roll and pitch angles, shown as lines in Fig. 2a. And third, 17 face Animation Units (defined within Kinect SDK) were recorded, expressing face deformations such as eye closed, lip corner depressed, jaw open, and eyebrow lowered in terms of numeric weights varying between 0 and 1.

We define a world coordinate system to estimate gaze point location of individual test persons located in the classroom (Fig. 1). The origin of a world coordinate system is set at the floor level and at the left corner of the classroom, with the x'-axis extending through the slide display area and the white-board. The y'-axis represent a height above the floor, and the z'-axis is extending towards the persons within the classroom. The 3D coordinates of a point within world (classroom) coordinate system are given by vector $p' = [p'_x, p'_y, p'_z]$.

3.3 Human estimation of attention level of students

The literature is lacking a consistent definition of the student's attention in the classroom. The review of video recordings of students during lecture has shown that students' attention towards the lecture is manifested by observable behavior such as gaze, writing, and mimics. As a starting point of the research, we asked human observers to estimate how attentive subjects appear to be during lecture by observing video recordings (Fig. 2a). The observed attention annotation procedure is described in detail in [27].

The five human observers, $j \in \{1, 2, 3, 4, 5\}$, were asked to estimate observed attention level of the students during lecture on the scale 1..5 with a time granularity of 1 s. We denote those estimates as *observed attention* $A^o_{u,j}(t) \in \{1, 2, 3, 4, 5\}$. To derive a *mean attention* score $A^m_u(t)$ of a student u at time t, we removed minimum and maximum estimate and calculated mean of the three visual attention scores (see Fig. 2b). Since observed attention estimates were not always consistent and in agreement, the mean attention score exhibits short-term fluctuations. To regularize estimated attention, we performed median filtering with a time window of 10 s and thresholding to three levels. The final *reference attention* of a student is denoted as $A^r_u(t) \in \{1, 3, 5\}$ and provides a human estimate of the current attention level of a student on a three-level scale; an example is shown in Fig. 2b.

3.4 Relation of attention to the behavioral cues

In addition to attention level, we labeled video clips for the presence of specific behavior such as writing for each of the test students. Starting and ending times were labeled, and binary signals representing those actions were calculated. The set of behavior reference signals includes the following features:

- Writing, $W_u(t) \in \{0, 1\}$, which was annotated when pencil was writing on paper and student was observing the notes

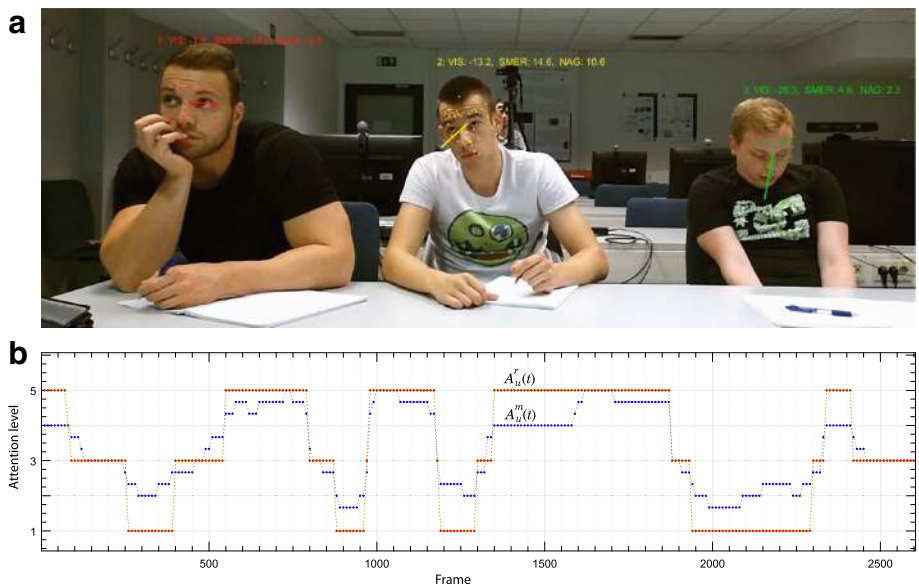

Fig. 2 Human estimation and annotation of person's attention level. By observing video footage (**a**), five human coders estimated attention level of each of the test persons. On graph (**b**), mean attention score $A^m_U(t)$ is shown as a blue line and final three-level reference attention $A^r_U(t)$ as a red line

- Yawning, $Y_u(t) \in \{0, 1\}$
- Supporting head, $S_u(t) \in \{0, 1\}$, where one hand is supporting or touching a face
- Leaning back, $L_u^b(t) \in \{0, 1\}$, describing upper body posture of the person
- Person's gaze, where we distinguish between four gaze directions (looking away, slides, white-board, notes). The gaze is represented by four binary features, $G_u^A(t), G_u^S(t), G_u^W(t), G_u^N(t) \in \{0, 1\}$

In order to associate visually observable behaviors of the test persons with the human-encoded attention level, we calculated mean values of annotated behavior signals for all levels of attention, over all test persons. In this way, we produced average score of selected behavior such as writing at each attention level, which gives an indication how observable behavior is related to observed attention. For example, for writing signal $W_u(t)$, we calculate writing-attention level correspondence $A_w(l)$, where $l \in \{1, 3, 5\}$ is an attention level:

$$A_w(l) = \frac{1}{N_u \cdot N_s} \sum_{u=1}^{N_u} \sum_{t=1}^{N_s} W_u(t)|_{A_u^r(t)=l} \quad (1)$$

Correspondences between the specific behavior and the attention level, calculated according to Eq. 1, are presented in Table 1. The values are interpreted as a rate of time the specific behavior was observed during all periods labeled with attention level l within our test set.

The results shown in Table 1 are graphically presented in Fig. 3. By observing video footage, we were able to identify body, facial, and other visual behavioral cues which correspond to levels of observed attention:

- High level of attention was associated with observing slides, writing notes (52% of time), and body leaning forward (88% of time).
- Medium level was associated with observing slides, body leaning forward (84%), and head supported by a hand (61%).
- Low level of attention was associated with gestures expressing tiredness or boredom, such as leaning back (41%), rubbing a neck, scratching head, yawning (26%), and looking away.

Table 1 Relation between observed attention levels and observable behaviors

	Attention level (l)		
Behavior	Low (1)	Medium (3)	High (5)
Writing, $A_w(l)$	0.0	0.05	0.52
Yawning, $A_y(l)$	0.26	0.06	0.001
Supporting head, $A_s(l)$	0.21	0.61	0.11
Leaning back, $A_b(l)$	0.41	0.16	0.12

The values represent rate of time the behavior was present during the periods with the specific attention level

Those observations allowed us to define a set of computed features for automated attention estimation in Section 4.3, which are closely related to the observed student behaviors as shown in Fig. 3.

4 Automated attention estimation from Kinect features

4.1 Kinect signal preprocessing

The frame-based Kinect data which were recorded during the observation session must be processed in order to extract specific features of interest and assign (map) them to correct test persons. During each experiment, we assigned an index u to each of the test persons.

The task of the mapping step at each time instance t is to assign each of the detected skeletons $k = 1..N_k$ to one of the actual test persons $u = 1..N_u$. This requires a predefined person setup which is given as an expected image position of the person's head. The mapping is done by finding a closest head position among all detected skeletons for each of the test persons' predefined positions. After the mapping step, we assign feature values of the k-th skeleton to the actual test person.

Kinect sensor provides skeleton and facial data at a rate of 30 fps. However, due to processing time to extract and store those features, the actual data rate is from 10 to 15 fps in our system. Data frames are lost due to limited processing speed and storage speed, causing dropped frames and non-uniformly sampled data. Kinect features also include certain level of noise, outliers are present in case of detection errors, or no signal is available when detection fails. Our attention estimation system deals with effects which last at least several seconds, so we chose 1 s as a sampling rate. Kinect signals are thus resampled at 1 frame per second by utilizing median filtering over a 1 s time window in order to preserve signal dynamics. The uniformly sampled Kinect feature signals are denoted $s_{j,u}(t)$.

4.2 Inter-person normalization and smoothing

Analysis of distribution of feature values revealed that they contain significant inter-personal differences. Signals differ either due to person-specific absolute values such as head position in 3D space, or due to varying amplitude of the signals (such as body lean back angle) between persons. In order to build a general behavior model, we need to align signals and normalize their range of values (Fig. 4).

We compute mean $\bar{s}_{j,u}$ and standard deviation $\sigma_{j,u}$ over all samples and compute normalized feature signal as $\tilde{f}_{j,u}(t) = (s_{j,u}(t) - \bar{s}_{j,u})/\sigma_{j,u}$. Normalized features provide better separation of attention classes. In order to improve time consistency of results, we computed smoothed

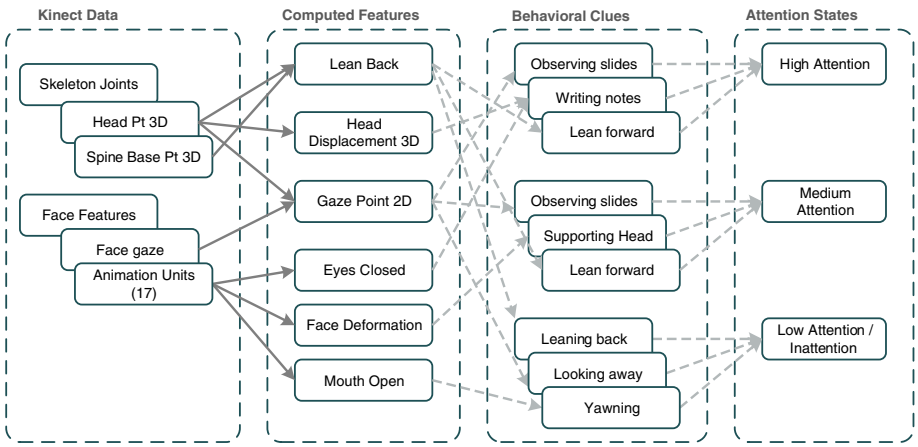

Fig. 3 Relation between computed features based on Kinect signals and the observable behavioral cues. High correlation between computed feature and the behavior is shown as dotted line

versions of feature signals $\hat{f}_{j,u}(t)$ by using Gaussian filter of width 11 s. Normalized and smoothed feature signals are shown in Fig. 5.

4.3 Computed feature set for attention estimation

The final feature set A consists of seven features computed by combining different Kinect signals, which are normalized over all persons. The feature set B consists of the same signals which are temporally filtered using weighted smoothing filter of the size of 11 s, see Fig. 5b.

4.3.1 Upper body posture

Upper body posture was found to be highly correlated with student activities such as observing slides and writ-

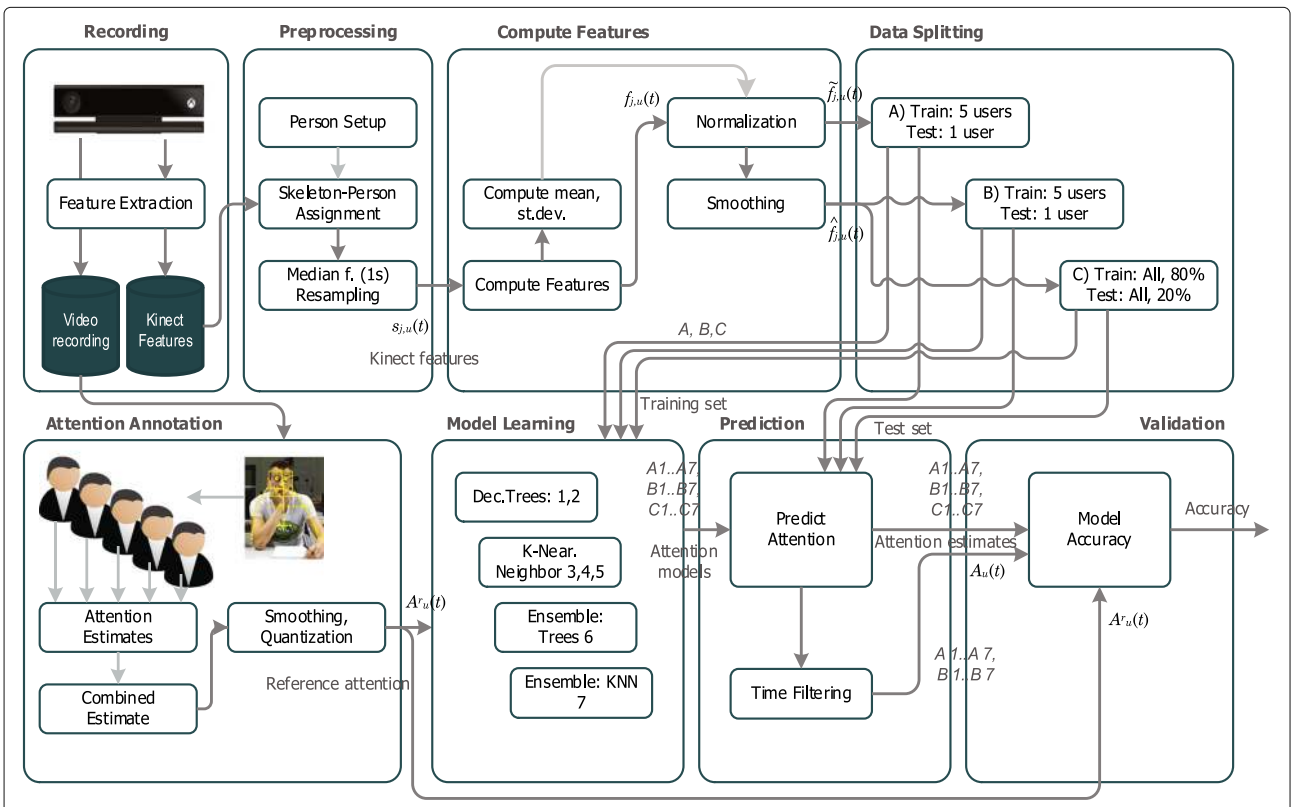

Fig. 4 Data processing diagram. Feature set for machine learning is computed from pre-processed Kinect features, and reference attention is provided by five human observers. Several classifiers are learned from the feature samples and the reference attention data, and model accuracy is computed on the test data set

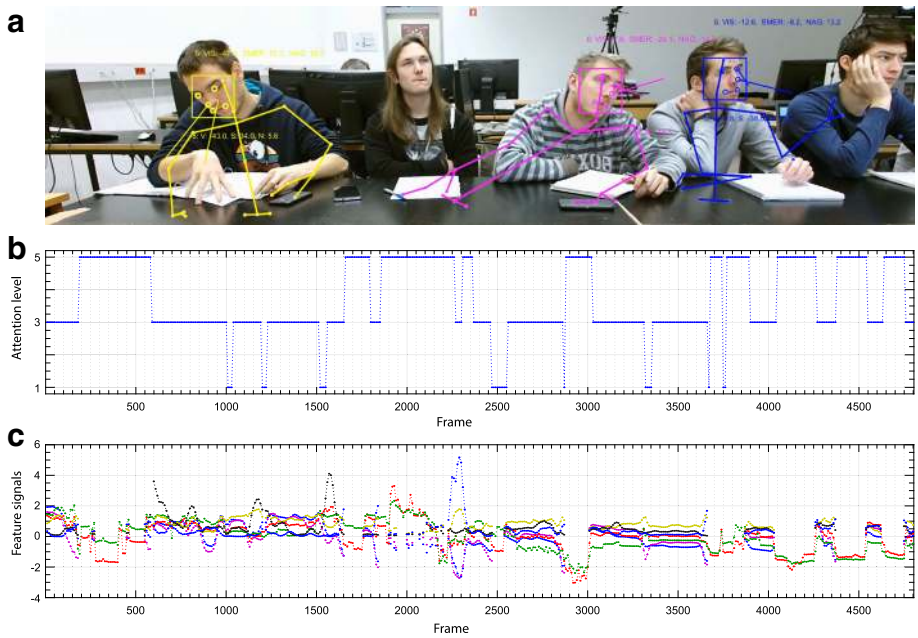

Fig. 5 Examples of final feature set for test person 6. Image (**a**) represents a single frame of Kinect video with visualization of Kinect data including face rectangles, face points, gaze vectors, and body skeletons. Graph (**b**) shows reference attention level, and graph (**c**) shows seven normalized and smoothed feature signals

ing. Features were computed from body joints given in 3D camera space, $p_{j,u}(t)$. Leaning forward when observing slides or writing resulted in changed head position in the 3D camera space.

Head displacement. In order to characterize changes in the head position during activities such as writing, we computed displacement vector of the current person's head position $p_{4,u}(t)$ from the mean position over the experiment $\bar{p}_{4,u}$, resulting in 3D displacement vector $D_u(t) = p_{4,u}(t) - \bar{p}_{4,u}$. Only head displacement in the vertical direction (y-axis) was found to be significant, and the feature 1 is normalized head displacement, $\tilde{f}_{1,u}(t) = (D_{u,y}(t) - \bar{D}_{u,y})/\sigma_{D_{u,y}}$

Body lean indicator. In order to characterize the overall upper body posture, we calculated the vector from head to lower spine, $d_u(t) = p_{4,u}(t) - p_{1,u}(t)$. We then calculated body lean indicator as an angle to the vertical coordinate axis y, $L_u(t) = \arctan \frac{d_{u,z}(t)}{d_{u,y}(t)}$. Feature 3 $\tilde{f}_{3,u}(t)$ represents normalized body angle.

Head angle indicator. Head angle in the z direction was found to be correlated with writing and observing slides. We calculated vector from head to upper spine point, $h_u(t)$, and compute head angle to the vertical coordinate axis y, $H_i(t) = \arctan \frac{h_{u,z}(t)}{h_{u,y}(t)}$. Feature 4 $\tilde{f}_{4,u}(t)$ represents normalized head angle.

4.3.2 Face gaze point

Kinect SDK provides an estimation of the relative head gaze, given as a vector of angles $g_u(t) = [\gamma_{x,u}(t), \gamma_{y,u}(t), \gamma_{z,u}(t)]$, where $\gamma_{x,u}(t)$ corresponds to head yaw and $\gamma_{z,u}(t)$ corresponds to pitch. Using the head position in the 3D camera space $\bar{p}_{4,u}$ and the Kinect sensor position in the world space K', we calculate projection of the head gaze onto the $x-y$ plane in the world coordinates, resulting in the 2D world gaze point coordinates $P_u(t) = [p'_{u,x}(t), p'_{u,y}(t)]$ (see Fig. 1). We use normalized y coordinate of the gaze point within our feature set as feature 2, $\tilde{f}_{2,u}(t) = (p'_{u,y}(t) - \bar{p}'_{u,y}(t))/\sigma_{p'_{u,y}}$.

4.3.3 Facial features

Facial features are derived from the 17 animation unit values which represent deformations of the detailed 3D face model.

- *Closed eyes* feature is computed as a maximum value of the two animation units, Right Eye Closed and Left Eye Closed, and correlates to writing and observing note activities.
- *Mouth open* feature is computed from the Jaw Open animation unit, and corresponds to yawning.
- *Face deformation* is computed from the Left Cheek Puff and Right Cheek Puff animation units and corresponds to supporting head with the hand, which causes 3D head model to become deformed.

The final set of seven normalized features $\tilde{f}_{1,u}(t) \ldots \tilde{f}_{7,u}(t)$ for the person 6 ($u = 6$) is shown in Fig. 5c.

4.4 Classifiers for attention estimation

The goal of the study was to build attention classifiers to automatically estimate three-level attention from recorded Kinect features. The first issue was to select optimal classifiers and their parameters in order to build a general, person-independent attention model, which is not over-fitted to the training data. The second issue was to select proper preprocessing of the features to achieve best prediction accuracy. We thus tested normalization and smoothing of the computed features.

We have included seven classifiers in the comparison, ranging in flexibility from simpler models such as decision trees to more complex models. The training process was using five-fold cross-validation to prevent model overfitting. The tested classifiers and their parameters are the following:

1. Decision tree (simple). Split criterion is Gini's diversity index; maximum number of splits is 4.
2. Decision tree (medium). Maximum number of splits is 20.
3. K-nearest neighbors (coarse). Distance metric is Euclidean, distance weight is equal, and number of neighbors is 100.
4. K-nearest neighbors (medium). Number of neighbors is 10.
5. K-nearest neighbors (weighted). Number of neighbors is 10, distance metric is Euclidean, and distance weight is squared inverse.
6. Bag of decision trees. Ensemble method is bag; learner type is decision tree. Number of learners is 30; max. number of splits is 20.
7. Subspace K-NN. Ensemble method is subspace; learner type is nearest neighbor. Number of learners is 30; subspace dimension is 4.

4.5 Training datasets and data splitting

We trained seven simple to complex classifiers and compared their overall accuracy in predicting attention on the testing data. All models were trained on the three different training datasets denoted as A, B, and C. The datasets are composed of two parts, set of seven features computed from Kinect data for attention prediction, and reference observed attention data. The general properties of the datasets are as follows:

- Datasets A and B contain samples of six test subjects recorded during two lecturing sessions. Each subject was assigned observed attention levels during 260 s of the lecture, leading to 260 data samples per subject and a total of 1560 samples in the datasets A/B. The difference between the two sets is in the preprocessing of the samples. The samples of the dataset A were normalized, while the dataset B contains features which were normalized and temporally smoothed (see Section 4.2).
- Dataset C contains samples of another 12 test subjects recorded during two lecturing sessions. The data of each subject were collected during 480 s, so the total number of samples in dataset C is 5760. The feature samples in the dataset were normalized and temporally smoothed.

The datasets were split into training set used during learning and the test set used during evaluation. For datasets A and B, we used the following strategy to split the data: training of the models was done on five persons (1300 data samples) and the remaining person (260 samples) was used for evaluation of the accuracy of predictions. The training/evaluation was repeated six times, and the results of accuracy of methods were averaged.

For the dataset C, we used time-based splitting (folding) of the data into training set containing 80% of samples of all users and test set containing 20% of samples of all users. In each of the five training/evaluation phases, attention level of all subjects was predicted from the testing set data. This five-fold cross-validation strategy was used to compute average accuracy of the classifiers.

5 Results

In this section, the performance of the proposed automatic attention classifiers is reported on two data sets involving 6 and 12 subjects and compared to the state-of-the art methods.

5.1 Training, testing data sets, and result sets

The original data set collected from six subjects was used to evaluate the effects of preprocessing of samples and postprocessing of results on the accuracy of predictions. Learning datasets A and B were created by normalization (dataset A) and additional temporal smoothing of normalized values (dataset B). The training/evaluation was done in six iterations, and in each iteration, a different subject u was used for evaluation of predictions. We denote predicted attention levels by classifier m for the test subject u as $A^{(R)}_{u,m}(t)$, where R denotes a specific result set.

Two additional sets of predictions were created from the results of datasets A and B. In order to increase temporal consistency of predictions, we employed post-processing of predicted values by temporal median filtering with a window of 11 s. We denote those results as A' and B'. The filtered attention levels are computed as $A^{(A')}_{u,m}(t) = \text{median}(A^{(A)}_{u,m}(t))$.

The dataset C was split into training and evaluation set by time, and five iterations were performed. The predicted attention of a subject u by a classifier m is denoted as $A^{(C)}_{u,m}(t)$, and the samples are collected from five iterations of training and testing.

5.2 Evaluation of attention estimation accuracy

Accuracy of the proposed system was evaluated by comparing the predicted attention levels $A_{u,m}(t)$ to the reference attention levels provided by human annotation $A^r_u(t)$. We compute the accuracy of predictions of a classifier m for the subject u on a dataset R, denoted as $Acc^{(R)}_{u,m}$, as a rate of correct predictions over time,

$$C_{u,m}(t) = \begin{cases} 1;\ A_{u,m}(t) = A^r_u(t) \\ 0;\ A_{u,m}(t) \neq A^r_u(t). \end{cases} \quad (2)$$

$$Acc^{(R)}_{u,m} = \frac{1}{N_s} \sum_{t=1}^{N_s} C_{u,m}(t) \quad (3)$$

The accuracy of predictions of tested classifiers for each of the test persons is shown in Fig. 6a–e. Each line connects results of a single classifier, allowing us to observe inter-subject variations of accuracy of a selected classifier, as well as range of prediction accuracies for a selected person.

The overall accuracy of a classifier m on the selected dataset $Acc^{(R)}_m$ was computed as an average over test subjects. The total number of predicted attention levels was 1560 for datasets A and B and 5760 for dataset C. Table 2 presents a comparison of attention estimation accuracies of seven evaluated classifiers for five data sets and allows us to estimate the influence of several factors on the accuracy of predictions.

Temporal smoothing of feature values (dataset B versus dataset A) improved the accuracy of four out of seven methods, and the average gain was +3.1%. Temporal smoothing was also used in the dataset C which achieved best overall accuracy. *Temporal filtering of predictions* in result sets A' and B' improved accuracy on average for 7.8%. Both results indicate that the successive observed attention values are highly correlated, and the actual attention level is sustained over longer periods of time.

Model flexibility. In the dataset A/B, (simple) decision tree classifier consistently provided best accuracy of predictions for the unknown person; thus, it achieved best inter-subject generalization. It is interesting that low flexibility model was superior over complex models, although the training algorithm reported differently during learning. This indicates probable over-fitting of complex models to the training data due lower number (1300) of training samples. Different results on classifier accuracy were observed on the dataset C, which is larger and includes 4560 training samples. The least flexible classifier (simple decision tree) produced lowest average accuracy, while highly flexible model (bagged trees) produced best accuracy of 0.753. It is thus important to adjust model flexibility to the number of available training samples in order to achieve inter-subject generalization.

Consistency and reliability of results. The results on the larger dataset C proved to be more consistent among the classifiers, as the range of results was much smaller (from 0.708 to 0.753) and the overall accuracy was higher. The larger dataset thus provided more reliable and consistent results.

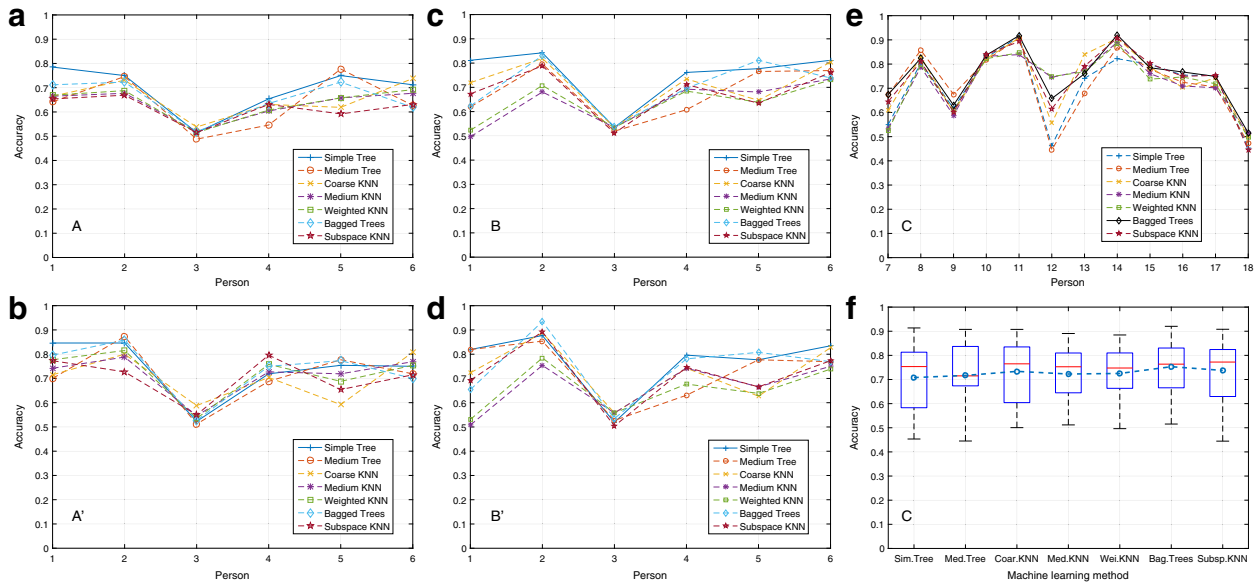

Fig. 6 Results on accuracy of attention prediction. Attention estimation accuracies of seven classifiers are shown for each test subject on datasets: **a** A (normalized features), **b** A′ (normalized with post-filtering), **c** B (normalized and smoothed features), **d** B′ (with post-filtering), and **e** C (normalized and smoothed features). Graph **f** represents distributions (shown as boxes) and average accuracies of each classifier over all test subjects of dataset C

Table 2 Comparison of attention prediction accuracies of tested machine learning methods

	Dataset				
Method *m*	A	A′	B	B′	C
1 (Simple tree)	*0.694*	*0.740*	*0.756*	*0.771*	0.708
2 (Medium tree)	0.638	0.712	0.636	0.674	0.717
3 (Coarse KNN)	0.654	0,701	0.717	0.731	0.734
4 (Medium KNN)	0.633	0.715	0.601	0.640	0.723
5 (Weight. KNN)	0.638	0.721	0.613	0.641	0.725
6 (Bagged trees)	0.655	0.734	0.706	0.747	*0.753*
7 (Subspace KNN)	0.615	0.703	0.680	0.712	0.738

5.3 Inter-personal differences

As we aimed to build a generalized attention prediction model applicable to various persons, we have encountered observable differences in the behavior of the test persons. The frequency of specific behaviors varied among test subjects, for example, yawning was only present at three persons out of six (in dataset A). There were notable differences in the dynamics and amplitude of head and body motions. And not least important, the success of Kinect feature detection (3D head model fitting, detection of open eyes and mouth, etc.) varied for each test person.

All those factors influenced efficiency of attention prediction for different subjects. Graphs a–e in Fig. 6 allow us to estimate inter-subject differences in accuracy of the predictions. In the dataset A/B, subject no. 3 was the most difficult for attention estimation with an average accuracy of 0.55. Visual observations confirmed that the behavior of this person was most dissimilar to other persons, as he exhibited longer periods of inattention and performed tiredness gestures (yawning, looking away) which otherwise appeared rarely in the training set (among other persons). Person 3 was also writing while he was leaning back on the chair, which was unusual in our dataset. On the other hand, person 5 caused most dissimilar scores among different classifiers, which indicates difficulties in predicting attention due to very small amplitude of head and body motion. He performed many subtle finger gestures which influenced observers' estimates but were not captured within our feature set.

Similar observations can be made for larger dataset C. While the consistency of results of tested classifiers was higher, there was still a notable difference of 0.31 between the highest and lowest accuracy among 12 test subjects. The predictions were more accurate for the persons who appeared to be more attentive to observers and were regularly taking notes. Upright body position, consistent behavior over time, and large amplitudes of head motion when changing gaze point all increased the success of prediction. Four test subjects with lowest accuracy of predictions also appeared to have low average attention to observers, as shown in Fig. 7. Figure 6f presents distributions and average accuracies of seven classifiers over 12 test subjects of dataset C. While the average accuracy of the best method is 0.753, the range of scores over test persons is from 0.52 to 0.93.

5.4 Estimation of average attention of subjects

The proposed automatic attention estimation system is able to predict time-varying attention $A_u(t)$ of a person u by observing his body and facial features. In the context of learning, it is however also important to provide a more general average attention of a person during a specific time period, for example, during a single lecture topic. We thus calculate average attention prediction $\bar{A}_u(t_1, t_2)$ of students during the observation time of 480 s and compare those estimates with the reference averages $\bar{A}^r_u(t_1, t_2)$ computed from reference attention.

The reference average attention scores for users 7 to 18 (shown in Fig. 7 as a dashed line) were in range from 2.1 to 4.1, while predicted attention averages (solid line) were in range from 3.0 to 4.0. The error in predicting average attention was in range from -0.33 to 0.86, and average error was 0.20. The system was less successful

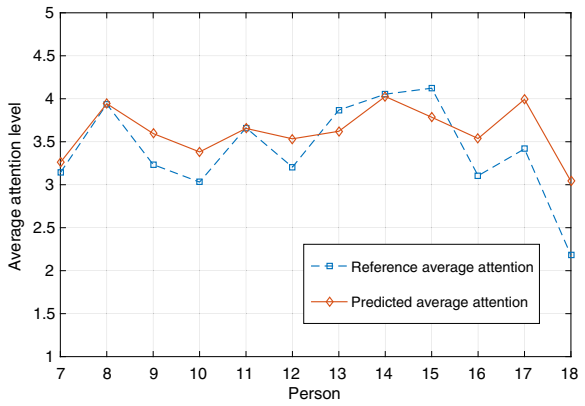

Fig. 7 Average attention of subjects. The blue boxes represent average attention of the test subjects calculated from human observations, while red diamonds represent estimated average attention of subjects by the best attention classifier (bagged trees)

in predicting lower attention levels, so the students with longer duration of low attention periods were predicted attention level which was too high.

5.5 Comparison to state-of-the-art

Two recent studies by Whitehill et al. [3] and Monkaresi et al. [17] propose systems for automatic estimation of student engagement from face and other features in the context of computerized learning. The experiments involve cameras which observe subjects during playing cognitive skills training game on iPad [3] and during writing an essay on a computer. The first study [3] utilized human labelers to estimate how engaged the subject appear to be on a four-level scale from either 60- or 10-s video clips, while our time resolution of annotations and predictions was 1 s. Their best model, SVM classifier with Gabor features, achieved subject-independent engagement recognition accuracy of 0.729 for detecting each level from all the other levels. The second study [17] uses concurrent self-reports on a two-level scale every 2 min, and retrospective self-reports after observing 10-s clips as a reference data. They achieved an accuracy of 0.758 for a person-independent two-level engagement detection. Although it is difficult to directly compare the results due to differences in tasks, datasets, and annotation methods, our results are comparable in terms of accuracy.

5.6 Discussion

In the following section, we discuss some of the limiting factors which affect the automated estimation of attention level in the classroom.

First, the ground truth data on attention computed from human observer estimates is not entirely reliable. In the absence of prior hard definition of how to annotate attention from observable behavior human coders have often disagreed on the level of attention at each time instance. Our definition of three attention levels was derived from the mean scores and their correspondence to the observable behaviors. This however does not mean that the actual annotated levels follow exactly this definition, which affects the achievable accuracy of the proposed system.

Second, the size of the training dataset is rather limited, and its total length for 18 persons is 122 min. The variations of human behavior which are present within the data set is limited and is not covering all possible student behaviors. Furthermore, inter-personal differences in behavior during lecture were clearly visible and influenced the accuracy of a person-independent classifier.

We noticed the issue of reliability and accuracy of Kinect data. Kinect sensor employs computer vision algorithms to detect facial features, such as eyes, nose, and mouth, and to fit a detailed 3D face model. The detection sometimes fail and in other cases produce erroneous results. The reliability of gaze detection depends on the orientation of the face (frontal or not) and presence of obstructing objects such as hands. The similar issue is with person's skeleton which is not accurate due to obstructed view of the person sitting behind a table. We could not include hand 3D coordinates due to low reliability.

And finally, the exploited set of seven features computed from low-level Kinect data was not comprehensive enough to be able to describe all observed behavioral differences of the test persons. More comprehensive feature set would clearly be able to more reliably detect important behavior such as writing, hand and finger gestures, and facial expressions.

6 Conclusion

In this paper, we proposed a novel approach to estimate the attention level of students in the classroom using a set of features computed from the data obtained by the Kinect One sensor. On the basis of visual observation of behavioral cues and their correlation with the attention level estimated by human observers, we derived a set of body, gaze, and facial features related to observed students' behavior. This computed feature set was utilized within seven machine learning algorithms to predict a three-level attention score with a time resolution of 1 s. We have evaluated several options for the preprocessing of signals and post-processing of the results and compared the efficiency of seven classifiers with different levels of flexibility. The success of building a person-independent attention prediction model was validated by testing attention prediction on the dataset of 18 persons with moderate accuracy of up to 0.753, comparable to the state-of-the-art studies in the field.

The proposed automatic attention estimation system has a clear potential usage as a tool for automated analytics of the learning process, providing a mechanism for large-scale analytics of student behavior in the classroom by using affordable but very capable hardware. This opens a possibility for teachers to evaluate their lectures and observe fine-grained effect on the students and possibly adapt them to increase participation and attention of students and thus improve results of the learning process, as well as their teaching methods.

Acknowledgements

The authors would like to thank Dr. U.Burnik for preparing a lecture and the participants who took part in the experiment.

Funding

This work was partially funded by the Slovenian Research Agency (Javna agencija za raziskovalno dejavnost RS), grant no. P2-0246 (B), Algorithms and optimization methods in telecommunications.

Availability of data and materials

The data is available upon request by e-mail.

Authors' contributions
Both authors contributed equally to this work. Both authors read and approved the final manuscript.

Ethics approval and consent to participate
The study does not include medical research involving patients and does not involve ethical issues related to such research. The participants provided an informed consent prior to the study.

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

Publisher's Note

Received: 7 June 2017 Accepted: 12 November 2017
Published online: 01 December 2017

References

1. D Dinesh, A Narayanan, K Bijlani, in *2016 International Conference on Information Science (ICIS), Kochi, India*. Student analytics for productive teaching/learning (Institute of Electrical and Electronics Engineers (IEEE), Piscataway, 2016), pp. 97–102
2. NJ Butko, G Theocharous, M Philipose, JR Movellan, in *Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference On*. Automated facial affect analysis for one-on-one tutoring applications (Institute of Electrical and Electronics Engineers (IEEE), Piscataway, 2011), pp. 382–287
3. J Whitehill, Z Serpell, Y-C Lin, A Foster, JR Movellan, The faces of engagement: Automatic recognition of student engagement from facial expressions. IEEE Trans. Affect. Comput. **5**(1), 86–98 (2014)
4. RA Calvo, S D'Mello, Affect detection: An interdisciplinary review of models, methods, and their applications. IEEE Trans. Affect. Comput. **1**(1), 18–37 (2010)
5. AS Won, JN Bailenson, JH Janssen, Automatic detection of nonverbal behavior predicts learning in dyadic interactions. IEEE Trans. Affect. Comput. **5**(2), 112–25 (2014)
6. J Fredricks, W McColskey, J Meli, B Montrosse, J Mordica, K Mooney, *Measuring student engagement in upper elementary through high school: A description of 21 instruments. (issues & answers report, rel 2011–no. 098)*, (2011). Technical report, U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Southeast
7. JA Fredricks, PC Blumenfeld, AH Paris, School engagement: Potential of the concept and state of the evidence. Rev. Educ. Res. Spring. **74**(1), 59–109 (2004)
8. R Martinez-Maldonado, A Clayphan, K Yacef, J Kay, Mtfeedback: Providing notifications to enhance teacher awareness of small group work in the classroom. IEEE Trans. Learn. Technol. **8**(2), 187–200 (2015)
9. CR Henrie, LR Halverson, CR Graham, Measuring student engagement in technology-mediated learning: A review. Comput. Educ. **90**, 36–53 (2015)
10. MS Young, S Robinson, P Alberts, Students pay attention!: Combating the vigilance decrement to improve learning during lectures. Act. Learn. High. Educ. **10**(1), 41–55 (2009)
11. EF Risko, N Anderson, A Sarwal, M Engelhardt, A Kingstone, Everyday attention: Variation in mind wandering and memory in a lecture. Appl. Cogn. Psychol. **26**(2), 234–42 (2012)
12. C-M Chen, J-Y Wang, C-M Yu, Assessing the attention levels of students by using a novel attention aware system based on brainwave signals. Br. J. Educ. Technol. **48**(2), 348–469 (2015)
13. C Yan, Y Zhang, J Xu, F Dai, L Li, Q Dai, F Wu, A highly parallel framework for HEVC coding unit partitioning tree decision on many-core processors. IEEE Signal Proc. Lett. **21**(5), 573–6 (2014)
14. C Yan, Y Zhang, J Xu, F Dai, J Zhang, Q Dai, F Wu, Efficient parallel framework for hevc motion estimation on many-core processors. IEEE Trans. Circ. Syst. Video Technol. **24**(12), 2077–89 (2014)
15. C Yan, Y Zhang, F Dai, X Wang, L Li, Q Dai, Parallel deblocking filter for HEVC on many-core processor. Electron. Lett. **50**(5), 367–8 (2014)
16. C Yan, Y Zhang, F Dai, J Zhang, L Li, Q Dai, Efficient parallel hevc intra-prediction on many-core processor. Electron. Lett. **50**(11), 805–6 (2014)
17. H Monkaresi, N Bosch, RA Calvo, SK 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)
18. N Alioua, A Amine, A Rogozan, A Bensrhair, M Rziza, Driver head pose estimation using efficient descriptor fusion. EURASIP J. Image Video Process. **2016**(1), 1–14 (2016)
19. R Bixler, S D'Mello, Automatic gaze-based user-independent detection of mind wandering during computerized reading. User Model. User-Adap. Inter. **26**(1), 33–68 (2016)
20. N-H Liu, C-Y Chiang, H-C Chu, Recognizing the degree of human attention using EEG signals from mobile sensors. Sensors. **13**(8), 10273 (2013)
21. J Han, L Shao, D Xu, J Shotton, Enhanced computer vision with microsoft kinect sensor: A review. IEEE Trans. Cybern. **43**(5), 1318–34 (2013)
22. S Springer, GY Seligmann, Validity of the kinect for gait assessment: A focused review. Sensors. **16**(2), 194 (2016)
23. G Zhu, L Zhang, P Shen, J Song, An online continuous human action recognition algorithm based on the kinect sensor. Sensors. **16**(2), 161 (2016)
24. SS Mukherjee, NM Robertson, Deep head pose: Gaze-direction estimation in multimodal video. IEEE Trans. Multimed. **17**(11), 2094–2107 (2015)
25. A Saeed, A Al-Hamadi, A Ghoneim, Head pose estimation on top of haar-like face detection: A study using the kinect sensor. Sensors. **15**(9), 20945–66 (2015)
26. L Paletta, K Santner, G Fritz, A Hofmann, G Lodron, G Thallinger, H Mayer, in *ICVS'13 Proceedings of the 9th International Conference on Computer Vision System. Lecture Notes In Computer Science*. Facts—a computer vision system for 3D recovery and semantic mapping of human factors (Springer-Verlag, Berlin, 2013), pp. 62–72
27. U Burnik, J Zaletelj, A Košir, Video-based learners' observed attention estimates for lecture learning gain evaluation. Multimedia Tools and Applications (2017). https://doi.org/10.1007/s11042-017-5259-8