Multimodal Student Engagement Recognition in Prosocial Games

Athanasios Psaltis, Konstantinos C. Apostolakis, Kosmas Dimitropoulos, *Member, IEEE*, and Petros Daras, *Senior Member, IEEE*

***Abstract*—In this paper we address the problem of recognizing student engagement in prosocial games by exploiting engagement cues from different input modalities. Since engagement is a multifaceted phenomenon with different dimensions, i.e., behavioral, cognitive and affective, we propose the modeling of student engagement using real-time data from both the students and the game. More specifically, we apply body motion and facial expression analysis to identify the affective state of students, while we extract features related to their cognitive and behavioral engagement based on the analysis of their interaction with the game. For the automatic recognition of engagement, we adopt a machine learning approach based on artificial neural networks, while for the annotation of the engagement data, we introduce a novel approach based on the use of games with different degrees of challenge in conjunction with a retrospective self-reporting method. To evaluate the proposed methodology, we conducted real-life experiments in four classes, in three primary schools, with 72 students and 144 gameplay recordings in total. Experimental results show the great potential of the proposed methodology, which improves the classification accuracy of the three distinct dimensions with a detection rate of 85%. A detailed analysis of the role of each component of the Game Engagement Questionnaire (GEQ), i.e., immersion, presence, flow and absorption, in the classification process is also presented in this paper.**

***Index Terms*—Emotion recognition, engagement recognition, human computer interaction, serious games, student engagement.**

I. Introduction

ENGAGEMENT is an important determinant for user’s interaction with technology. The measurement of user's engagement enables not only the better design of interactive applications, but also the development of intelligent, sophisticated and adaptive environments. This is mainly due to the fact that engagement plays a key role in better understanding general user's behavior and overall efficacy of goal or task-oriented behavior within computer-based environments [1], such as social networks, video-games, web-based applications or educational environments. In the field of education, student engagement is a topic of interest for several decades, since it is closely associated with academic, behavioral and social outcomes. There are numerous studies in the literature, which show that low student engagement leads to poor academic performance or even high drop-out rates [2]. To this end, the need of maintaining and enhancing student engagement during learning process has significantly increased the research interest in automatic engagement recognition methods.

Student engagement is a very complicated and multifaceted phenomenon with different dimensions and therefore, there are various definitions of this term in the literature. Newmann [3] first discussed the importance of engagement in the educational process defining student engagement as the student's psychological investment in and effort directed toward learning, understanding, or mastering the knowledge, skills, or crafts that academic work is intended to promote. Additional studies investigated the relationship between school engagement and dropping out [4][5], while other researchers identified the engagement in terms of autonomy [6][7]. In a more recent study, Zepke *et al.* [8] also suggested that engaged students usually succeed on their activities when they are intrinsically motivated and feel capable of working in an autonomous way. On the other hand, Parsons and Taylor [6] argue that only performing the activities is not enough to identify the engagement of students. Since activities should take place in a specific time, students show their commitment by managing time efficiently while working on deadlines [9][10]. Additionally, other studies revealed that engagement can also be evaluated from the perspective of collaboration and teamwork [11][12], while recent research works have focused mainly on the strong relation between fun, satisfaction and engagement [13][14][15].

In order to better study the role of engagement in learning, Finn’s [16] introduced the “Participation-Identification” model, which makes a clear distinction between behavioral and affective engagement [17]. More specifically, the behavioral dimension of the model is related to the degree of student’s participation in the learning process, while identification refers mainly to the student’s emotional attitude towards learning, i.e., the student’s affect within the classroom and the sense of belonging in school. Later, Fredericks *et al.* [5] proposed one of the most frequently cited models of student engagement that is based on three distinct components: behavioral, affective and cognitive engagement.

Manuscript received 28 April, revised 8 August, accepted 16 August 2017. This work was supported by EU Horizon 2020 Framework Programme under grant agreement no. 644204 (ProsocialLearn project)

A. Psaltis, K.C. Apostolakis, K. Dimitropoulos, and P. Daras are with the Information Technologies Institute, ITI - CERTH, 6th km Charilaou-Thermi Rd, Thessaloniki, 57001 GREECE (e-mail: at.psaltis@iti.gr, kapostol@iti.gr, dimitrop@iti.gr, daras@iti.gr).

The latter includes features such as problem solving, preference for challenging work, investment in learning, metacognitive strategies and persistence during difficult tasks [18]. More recently, other researchers [19] have further proposed that engagement comprises four subtypes, such as academic, behavioral, cognitive and psychological.

This complex nature of engagement has given rise to a number of different measurement methods, which are usually based on self-reports, external observers (e.g., teacher checklists and rating scales) or automated measurement systems. Although primitive, both self-reports and teacher checklists are still the most popular ways of measuring engagement. This mainly stems from the fact that they are cost-effective and easy to implement approaches for collecting data in learning environments. However, they also suffer from limitations, since both approaches are time-consuming, lack temporal resolution and may be biased under some conditions (e.g., students may not answer honestly or evaluation results may rely on teacher's subjective opinion). On the other hand, automatic measurement approaches collect engagement data in real-time in order to adapt the content or the learning environment, making personalized learning an attainable goal. Although automated measurements provide high temporal resolution compared to other two approaches, in most of the cases they fail to capture or simply ignore the different dimensions of engagement. Techniques based on interaction log-files estimate engagement using solely the timing and accuracy of student's responses [20], while methods based on different sensing techniques, such as physiological and neurological sensors, or computer vision focus mainly on the affective dimension of engagement [21][22].

Towards this end, in this paper we propose a novel multimodal student engagement recognition approach for serious games in education inspired by the theoretical grounds of educational psychology. The proposed approach aims to capture the different dimensions of engagement, i.e., behavioral, cognitive and affective, based on the combination of real-time engagement cues from different input modalities, such as facial expressions, body postures and various game events. More specifically, this paper makes the following contributions:

- We propose a new multimodal student engagement recognition method, which combines engagement cues from both the students and the game. Our study focuses on serious games in education, such as prosocial games for cooperation and trust, however, the proposed methodology is generic and can be easily applied to a variety of serious game applications.
- For the detection of user's affective engagement during gameplay scenarios, we present a multimodal emotion recognition methodology based on facial expression and body motion analysis. A deep neural network is used for the combination of different modalities, while for the measurement of the affective engagement dimension, the average variation of the player's affective state in the Valance-Arousal space is estimated.
- We introduce a novel methodology for the annotation of engagement data. Instead of relying only on self-reports or external annotations, as is commonly done, in this paper we have developed two new versions of a prosocial game, namely "Path of Trust" [23], with different degrees of challenge (we developed a "boring" and a more challenging version of the game) in order to trigger different levels of engagement to the players. Subsequently, we collect the engagement data using a self-report approach based on the well-known Game Engagement Questionnaire [24] and we train our machine learning model using data from both the sensors (in our experiments we used Microsoft Kinect sensors) and the games.
- Finally, we present the benefits of merging modalities for estimating student engagement through real-life experiments conducted in three primary schools. The results presented in the paper involve experiments with 72 kids and 144 gameplay recordings in total.

The remaining of this paper is organized as follows: In Section II, similar works on estimating engagement are discussed, while in Section III the two versions of the "Path of Trust" game are presented. Section IV outlines our multimodal engagement recognition methodology and Section V describes the experimental procedure followed. Finally, the experimental results of our study are given in Section VI, while conclusions are drawn in Section VII.

II. Related Work

A variety of techniques for measuring engagement exists in the literature. In its most common form, the quantification of engagement is achieved using specialized psychometric tests in the form of self-reports [19][25] or observational checklists [26]. The former are filled out by the participants, while the later by an expert observing the experiment. The standard procedure for such a test involves enrolling a group of people to observe a video, solve a puzzle or play a game. Participants or external observers, e.g., teachers, are asked to fill in a questionnaire and following an analysis of the answers, student engagement is measured. A distinction can be made about whether the questionnaire is filled out before, during or after the execution of the experiment. In a recent study, Monkaresi *et al.* [22] used self-reports for concurrent and retrospective affective annotations during and after a structured writing activity. For concurrent annotations, students were asked to verbally report their level of engagement (engaged or not) in response to an auditory probe produced every two minutes during an essay writing activity.

A well-known example of a standardized questionnaire for engagement is the Game Engagement Questionnaire (GEQ) [24], which is used to quantify engagement of participants in games and includes a set of nineteen questions classified into four categories: absorption, flow, presence and immersion. On the other hand, O'Brien and Toms [27] introduced a conceptual model of engagement, namely User Engagement Scale (UES), for its general use in the context of human computer interaction, while later Wiebe *et al.* [28] investigated its use as a psychometric tool to measure engagement during video game-play. More recently, Phan *et al.* [29] proposed a new instrument for measuring video game satisfaction, called the Game User Experience Satisfaction Scale (GUESS), consisting of nine different subscales.

Although questionnaire fill-out procedures remain the gold-standard for inferring engagement, other semi-automated or fully automated methods for measuring engagement have emerged recently due to the advances in human-computer interaction and computer vision [30][31][32]. In a typical automated recognition procedure, sensors are used in order to collect various kinds of signals and then, following a process of data annotation, a predictor is built using machine learning. The predictor can then be applied in any real-time setting for inferring engagement. Most of the proposed methods are based on physiological sensors or computer vision techniques. The methods measuring physiological states attempt to monitor and analyze signals related to heart rate, blood pressure, galvanic skin response or brain operation [33][34][35]. However, since they require specialized hardware, such as Electrocardiogram (ECG), Electromyogram (EMG), Galvanic Skin Response (GSR), Respiration (RSP) or Electroencephalogram (EEG) sensors, they cannot be widely used in learning applications. On the other hand, computer vision techniques can analyze various cues from face, gaze, body or gestures [21][36][37] in order to automatically recognize user’s affective engagement. Due to the non-intrusive monitoring of user’s emotional state that they offer, computer vision-based approaches can be integrated easily into computerized learning environments and used in large scale studies.

In the field of video games, sensors have been employed by many researchers for estimating the engagement of users during gameplay. In particular, Chanel *et al.* [38] used physiological signals to infer user's emotion during a game and then adapted the difficulty of the game in order to retain engagement level of the player. On the other hand, Yannakakis and Hallam [39] developed a method for optimizing entertainment in games. They modeled player entertainment as an artificial neural network to provide the user's preference model. More recently, Shaker [40] presented an approach for predicting user’s level of engagement from nonverbal cues (visual and facial) within a game environment, such as Super Mario Bros, while in [41] a method based on the fusion of visual and behavioral cues for the modeling of user’s experience in games was proposed. Finally, other researchers focused on the study of visual attention in games through the use of head mounted eye trackers [42].

In education, automatic engagement recognition tasks historically found usage in the area of Intelligent Tutoring Systems (ITS). These systems intend to automate the learning process by adjusting teaching strategies based on the learner's engagement level [43]. Usually, the estimation of engagement in an ITS is based on the timing and accuracy of user’s responses to exercises, problems or test questions [44]. In recent years, engagement recognition has been the subject of increasing attention in the active research area of serious games. As several studies have shown, games are important in the development of knowledge and engagement of students [45]. The effectiveness of gamification for stirring up the engagement of students was further explored by da Rocha in [46], while Sabourin and Lester [47] presented an in-depth analysis of how affect and engagement interact with learning in game-based learning environments. Whitehill *et al.* [21] used face recognition to acquire student engagement related facial expressions during gameplay and then trained an SVM classifier in order to have an engagement estimator for face. Recently, Monkaresi *et al.* [22] proposed an automated engagement detection method using video-based estimation of facial expressions and heart rate. This experimental study focused on a structured writing activity rather than game based-learning. The study showed that the fusion of individual input signals can increase the accuracy of the engagement detection. Similarly, other researchers have proven that the accuracy in affect recognition tasks is increased by using multi-sensor approaches [48][49].

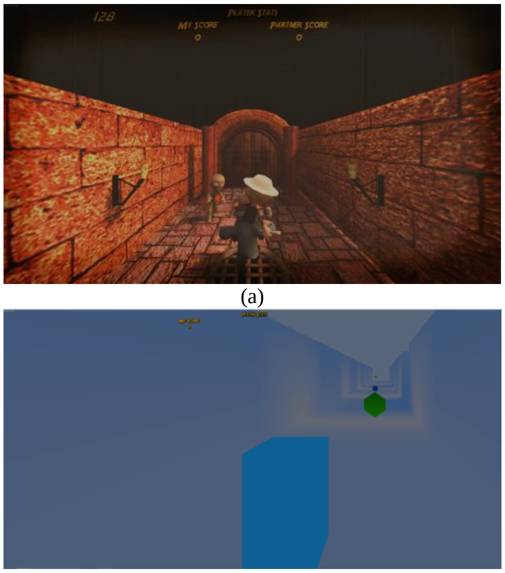
In the same manner that accuracy in affect recognition tasks is increased by fusing different input modalities, the computation of engagement can be enhanced by supplementing affective state estimation with contextual information. More specifically, a recent study on dialog systems showed that automatic engagement recognition combining affect with social-behavioral cues outperforms mono-modal solutions [50]. Towards this direction, in this paper we propose a novel multimodal machine learning approach for the automatic measurement of student engagement in serious games based on the merging of behavioral, cognitive and emotional engagement cues.

III. The Prosocial Game

Prosocial behavior refers to a type of social behavior that is intended to benefit other people (e.g., helping, sharing, cooperating, etc.), and/or society as a whole (e.g. donating, volunteering, etc.) [51]. From a psychological point of view, prosociality is composed of many core domains [52], of which *Trust* and *Cooperation* are key abilities. Experimental research has suggested that games in which the main characters (and therefore, players controlling them) model and carry out prosocial behaviors (e.g., prosocial games) may have a causal impact on actual player disposition for carrying out prosocial behaviors in real life [53].

In this paper, we developed two different versions of a prosocial game, namely “Path of Trust”, to collect various engagement cues. More specifically, the standard version of our game features both single and multiplayer modes, and is associated with developing desirable traits in the prosocial core domains of Trust and Cooperation [23]. Within “Path of Trust”, players take on the role of either of the two playable characters, who interact only through the suggestion (e.g., *the Guide*) and following of directions (e.g., *the Muscle*). A sense of trust must be built between the players (or the player controlling the Muscle character and NPC Guide in the single player game), in order for both characters to cooperate in collecting equal shares of the treasures scattered throughout the dungeon and guarded by monsters and instant game-over traps. To achieve its goals and increase the game’s appeal on the target audience, the game features narrative elements (colorful 3D characters and backstory, multiple endings etc.) and a Natural User Interface (NUI), which enables the players to navigate in the game's world using simple gestures.

For the experiments presented in this paper, we modified the original single version of our game [23]. More specifically, we placed two items in each corridor dungeon piece and we implemented a random item switching function. This function randomly determined whether one of the items in the corridor (*Treasure Piece*) would instantly be replaced with a *Mummy* hazard, slightly before players touched the item and collected the points to be gained from it. The rationale for including this game event lies within measuring the time it takes players to notice the swap (i.e., being attentive of the game), and subsequently react to it as they attempt to avoid the surprise hazard. A logging function was also implemented to keep track of the time responsiveness, accomplished outcome and player's affective state during these events.



(a)

(b)

Fig.1. a) The single version of "Path of Trust" and b) the stripped-down version of the game.

Additionally, in order to isolate the behavioral, cognitive and affective dimensions of student engagement using real-time facial, body and in-game engagement cues, we heavily modified the single player version of the game in a separate build, stripping down content and producing an intentional, boredom-inducing version of the game. In this version of the game, all colorful 3D graphics, textures, music and sound effects were replaced by simple geometric shapes. Instead of the two character avatars, in this version the player controls a simple cube totem, as seen in Fig.1. The game mechanics were slightly altered as well. Since all narrative elements were taken out of the original game, there was no exchange of information taking place between the player and the AI-controlled Guide. Instead, all the doors hindering the player from knowing what lies in the next room (thus contributing to spatial immersion in the game world) were omitted from this version, providing players with complete knowledge of the items lying in their path. As a result, this game version provided no challenge whatsoever, as the game’s slower pace and absence of closed doors allowed players plenty of time to move their totem out of harm’s way (red totems) and line it up to the treasure pieces (green and blue spheres).

This was intentionally designed to hinder the game from supporting the fundamental psychological need for competence, as described in self-determination theory (SDT) [54]. According to SDT, personal well-being is believed to be enhanced when one’s actions and interactions satisfy, among other things, a sense of efficacy. Maintaining interest and loyalty of players in video games is directly linked to competence need satisfaction described in the SDT [51]: too much of a challenge can lead to frustration, while too less of a challenge can ignite boredom. We therefore aimed for players to report no interest in re-visiting the stripped-down version of “Path of Trust”, as opposed to the original, which would allow us to trigger different levels of engagement during gameplay in a way that would make them detectable through the multi-modal, consumer-grade sensing devices used for mandatory game control (i.e., Microsoft Kinect).

IV. Automated Engagement Recognition

Due to the multifaceted nature of engagement, in this section we propose a multimodal approach, which combines engagement cues from both the player and the game. More specifically, for measuring the affective engagement, we extend our previous work on facial expression and body motion recognition [55][56] to identify the affective state of the player in the Valance-Arousal space using Kinect's data streams and, then, we estimate the average variation of the player’s affective state during gameplay. Subsequently, we combine this information with gameplay features, associated with the behavioral and cognitive engagement of the player, using a machine learning approach based on artificial neural networks.

A. *Facial Expression Analysis for Emotion Recognition*

Facial motion plays a major role in expressing emotions and conveying messages. The analysis of facial expressions for emotion recognition requires the extraction of appropriate facial features and consequent recognition of user’s emotional state that can be robust to facial expression variations among different users. Features extracted by applying facial expression analysis techniques can range from simply geo-locating and calculating actual anthropometric measurements, to summarizing an entire group of feature-group elements under a single emotional category, such as happiness or surprise.

For the experiments presented in this paper, we used Kinect SDK’s face tracking engine for the extraction of facial features. The engine is able to track facial muscle activities, i.e., Action Units (AUs), which can be seen as a form of mid-level representation of student's facial expressions. To classify expressions into the six basic emotion categories (anger, disgust, fear, happiness, sadness, and surprise), we concatenated the posteriors of all AUs in a unified vector representation [55] and we trained a neural network, as shown on the left part of Fig. 4.

B. *Body Motion Analysis for Emotion Recognition*

The majority of state of the art emotion recognition frameworks capitalize mainly on facial expression or voice analysis, however, research in the field of experimental and developmental psychology has shown that body movements, body postures, or the quantity or quality of movement behavior in general, can also be of help to differentiate between emotions [57]. To this end, we decided to extract a number of 3D body features, which are deeply inspired by the relevant psychological literature [56][58].

The 3D body movement features are extracted from joint-oriented skeleton tracking using the depth information provided by the Kinect sensor. More specifically, the extracted features are classified into the following broad categories: i) kinematic related features: kinetic energy, velocity and acceleration, ii) spatial extent related features: bounding box, density and index of contraction, iii) smoothness related features: curvature and smoothness index, iv) symmetry related features: wrists, elbows, knees and feet symmetry, v) leaning related features: forward and backward leaning of a torso and head as well as right and left leaning and vi) distance related features: distances between hands, distance between hand and head as well as hand and torso. An example of the kinetic energy measurement during the play of the "Path of Trust" prosocial game is demonstrated in Fig. 2.

For the combination of different set of features, we propose a two-layered network in which we have stacked seven ANNs (Artificial Neural Networks), six at the first layer and one at the second layer. Each layer is trained separately, starting from the base layer and moving up to the second, with no feedback from the higher layer to the lower one. Each ANN of the first layer receives as input the features of a different group of features. Then, the output probabilities of the first layer are fed as input to the second one and a separate ANN is trained. The output probabilities of the second layer constitute the classification result of the body motion analysis mono-modal classifier as shown in the right part of Fig.4.

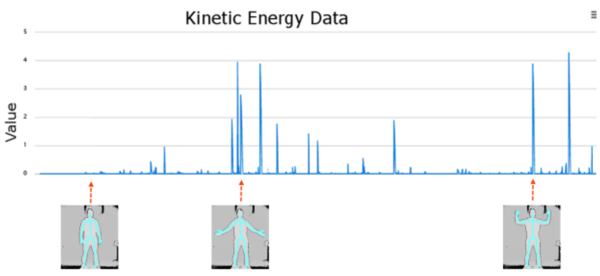


Fig.2. Kinetic energy data measurement using the Microsoft Kinect sensor.

C. *Multimodal Affective State Recognition*

The multimodal fusion process is responsible for measuring the affective state of students based on the combination of different modalities, i.e., facial expression and body motion analysis. Towards this end, in this paper we propose a multimodal fusion architecture for emotion recognition that uses stacked generalization on augmented noisy datasets and provides enhanced accuracy as well as robustness in the absence of one of the input modalities. Recent studies have shown that deep learning networks can be applied at feature level as well as at decision level, being trained directly on raw data or decisions accordingly. In this direction, we employ a late fusion scheme, where each intermediate classifier is trained to provide a local decision. In terms of affect, local classifiers return a confidence as a probability in the range of [0, 1] in a set of predefined classes, i.e., the six basic emotions. The local decisions are then combined into a single semantic representation, which is further analyzed to provide the final decision. The aforementioned scheme for late fusion using two Kinect's data streams is illustrated in Fig.3.

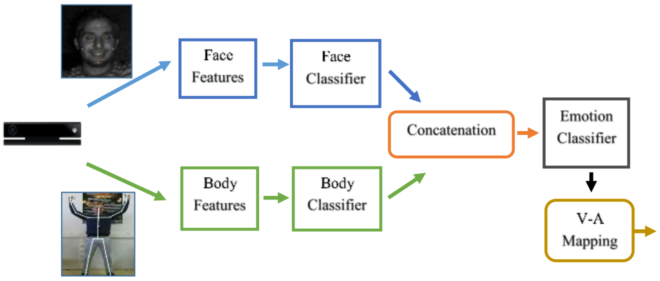


Fig.3. Bi-modal late fusion scheme.

In the proposed two-stream stacked generalization approach, the first stream models the appearance features of face, while the second stream models the topological characteristics of human body. Given a sequence of Kinect's data streams, feature vectors representing user's facial expressions and body gestures are extracted and fed into the unimodal ANN classifiers. Each classifier is composed of multiple nodes (neural units) that are connected each other by links associated with weights. The learning process of each ANN is based on a back propagation algorithm, which defines the parameters of the network. After the learning of each ANN model, the posteriors of the hidden variables can be used as a new representation of the data. By adopting such a unified representation, we can learn high-order correlations across different modalities.

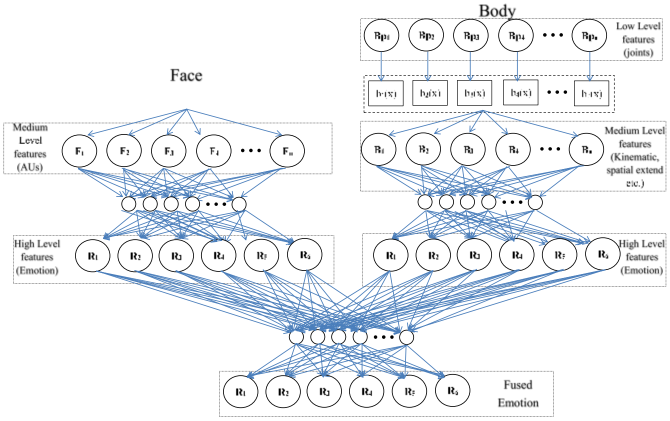

Body
Bp1
Bp2
Bp3
Bp4
• • •
Bpn
Low Level features (joints)
h1(x)
h2(x)
h3(x)
h4(x)
h n(x)
Face
Medium Level features (AUs)
F1
F2
F3
F4
Fn
B1
B2
B3
B4
Bn
Medium Level features (Kinematic, spatial extend etc.)
High Level features (Emotion)
R1
R2
R3
R4
R5
R6
High Level features (Emotion)
Fused Emotion

Fig.4. The proposed architecture of the multimodal fusion method.

The multimodal Deep Neural Network in Fig.4 can be described as a composition of unimodal latent variables models (ANNs), where each ANN is trained separately in a completely supervised fashion. The output of these ANNs, i.e., the estimated posteriors, is subsequently fed into a new ANN (as shown in the lower part of Fig.4), which is responsible for the final decision, i.e., the student's affective state. The training of the Deep Neural Network is layer-wise starting at the base layer and moving up, without affecting or fine-tuning the weights of the pre-trained ANNs. In practice, it can be considered as a directed model, since there is no feedback from higher layers to the lower ones, as shown in Fig.4. This layer-wise architecture, motivated by deep learning networks, improves performance, while avoiding overfitting [59]. For single-modality analysis the training procedure lasted 150 epochs, while for the multi-modal analysis 100 epochs were shown to be sufficient. In all ANN configurations, each layer included 150 hidden units.

The multimodal affective state recognition is a frame-by-frame process. In order to have an indication of the player's affective engagement, i.e., a total measurement of the player's affective activity during gameplay, we initially map the dominant student's emotion (i.e., the emotion with the highest probability in each frame) to the Valance-Arousal (V-A) space and then we estimate the average variation of the player's affective state. For the mapping of the dominant emotion to the V-A space, the 2D position of the dominant emotion is estimated as follows:

$$\begin{cases} V_d = V_m + \lambda(1-p)\sigma_d \\ A_d = \mathrm{A}_m + \mu(1-p)\sigma_d \end{cases} \quad (1)$$

where V_m and A_m are the coordinates of the mean value of the dominant emotion in V-A space, σ_d indicates its standard deviation [61], p is the probability of the dominant emotion, while factors λ and μ are scalars with values equal to ±1 depending on V-A values of the second most dominant emotion (e.g., λ=1 if the valance value of the second most dominant emotion is greater than V_m). After mapping the emotions to V-A space, the average variation D_a of the student's affective state can be easily calculated as follows:

$$D_a = \frac{1}{F}\sum_{i=2}^{F}\left\|x_{i-1} - x_i\right\|^2 \quad (2)$$

where F is the total number of frames and x_i indicates the current affective state of the student in V-A space.

D. Game-play Features

Student's interactions with the game can provide valuable information for the two other engagement dimensions, i.e., behavioral and cognitive. Towards this end, in this section we aim to extract features based on the analysis of specific game-play events and their corresponding time-stamps in order to achieve a more targeted measurement of these two aspects of engagement. According to the literature, the dimension of behavioral engagement is defined as *"focused activity on a task"*, with a typical measurement being time on task [60][61] while playing the video-game. On the other hand, cognitive engagement is defined as "*mental activity associated with the presented content*" and is measured by successfully achieving the desired goal of the game or by pre and post testing of outcomes [61]. In this paper, the behavioral engagement of a student is measured by estimating his/her average time of responsiveness $R \in [0,1]$ in all challenges c_i of the game (e.g., in "Path of Trust", a challenge can be the collection of a diamond), with i=1,2…n:

$$R = \frac{1}{n}\sum_{c_i=1}^{n} R_{c_i} = \frac{1}{n}\sum_{c_i=1}^{n}\frac{t_r^{c_i}}{t_{total}^{c_i}} \quad (3)$$

where $t_r^{c_i}$ and $t_{total}^{c_i}$ indicate the student's time of responsiveness and the total available time in challenge c_i, respectively.

Similarly, for measuring the cognitive engagement, we estimate whether the desired goal has been achieved in each challenge c_i (e.g., whether the student has decided to cooperate with and trust the Guide or follow a different strategy that leads to a non-prosocial behavior, the number of selected diamonds out of the total available diamonds or the number of monsters/traps that the player has avoided out of the total number of monsters/traps in a challenge) and we estimate the total score S_{τ_j} in each task τ_j of the game, with j=1,2…m, (each task can contain one or more challenges). Finally, we estimate the average score $S \in [0,1]$ of the student in all tasks of the game:

$$S = \frac{1}{m}\sum_{\tau_j=1}^{m} S_{\tau_j} = \frac{1}{m}\sum_{\tau_j=1}^{m}\frac{g_a^{\tau_j}}{g_{total}^{\tau_j}} \quad (4)$$

where $g_a^{\tau_j}$ and $g_{total}^{\tau_j}$ indicate the number of successfully achieved goals and the total number of goals in a task τ_j, respectively. We have to note here that both metrics/features of (3) and (4) are normalized and are completely independent of the game, that is, a game developer can easily estimate the values of R and S $\in [0,1]$ in any serious game by simply defining the challenges, the tasks, the available time and the goals of his/her game.

E. Engagement Recognition

Having defined the parameters, i.e., the three dimensions, of our engagement model, we subsequently need to annotate our data in order to label them for the training of our classifier. Towards this end, we adopted a retrospective self-reports approach, based on GEQ questionnaire, for both games, as descripted in detail in the "Engagement Experiment" section. Fig. 5 illustrates the GEQ measurement approach [24] and shows how player's answers are mapped to the engagement scale. More specifically, symbols N, M, and Y displayed in Fig. 5 refer to "No", "Maybe", and "Yes", respectively to each question displayed on the right. Since each answer corresponds to a specific engagement value, we aggregated the values of all answers and we estimated the average engagement value for each gameplay.

Finally, the estimated engagement values were used as labels of the recorded data, i.e., engagement vectors E=[D_a, R, S], for the training of an ANN. In our experiments we used two classes, "Not Engaged" and "Engaged", however, one can add more classes, e.g., “Nominally Engaged” or “Very Engaged”.

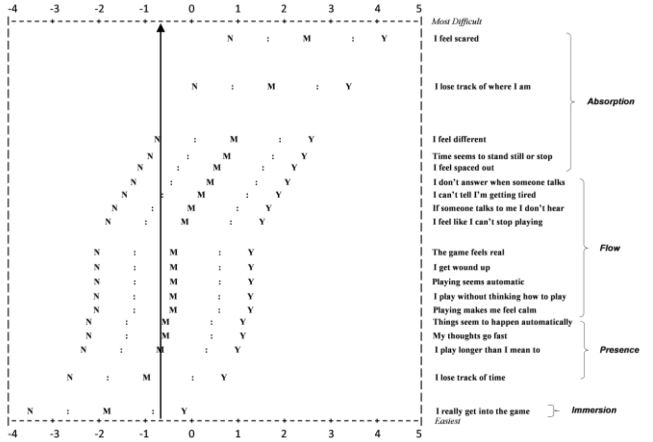


Fig.5. GEQ measurement approach [24].

V. The Engagement Experiment

A. Participants

Participants in our study were a total of N=72 children from three primary schools. The students ranged in age from 8 to 10 years old. In total, 38 boys and 34 girls completed the entire session. The study was approved by the Institute of Educational Policy (IEP), a private legal entity under the supervision of the Greek Ministry of Education, Research and Religious Affairs. Prior to the study, students’ parents signed consent forms allowing their children to participate in the study.

Fig.6. Experiment in a primary school: A student while playing the PoT game using the Microsoft Kinect sensor.

B. Procedure

The study took place in four different classes assigning around 16-20 children per classroom. Our user study coordination team assigned two persons per classroom to carry out the experiments, with one person hosting the game and the other assisting children with regards to the printed questionnaires queries. One desktop PC, equipped with a Microsoft Kinect sensor and with pre-installed versions of the single player *Path of Trust* games described in Section III, was situated within each classroom. The students were asked to play both versions of the game in succession, with a short time for filling out the GEQ in between sessions. A short description of the Path of Trust back-story was briefly delivered by the game host, along with instructions on the game gesture-driven interface with regards to moving the player character and touch to collect / avoid hazards mechanics. In order not to build-up any expectations about the games and receive a more genuine response to each game version’s appeal and motivational affordances, we allowed children to play the stripped-down version of the game first.

After playing, each student was asked to answer a GEQ to evaluate player’s engagement and overall game experience. The children were helped by the assigned experiment coordinator to clarify questions, which could confuse or were seen as somewhat difficult to assess by children on their own. The process was subsequently repeated with the original version of the game (see Fig. 6). Each game session had pure gameplay duration worth two and a half minutes (150 seconds) of time. In total, each child needed approximately 10-15 minutes to complete the study.

VI. Experimental Results

In this section, we present the results from testing the efficacy of the proposed method on estimating student engagement. More specifically, we evaluate the performance of the propose multimodal affective state recognition algorithm and then present experimental results from real-life experiments conducted in primary schools. In the first case, our main goal is to compare the proposed multimodal affective state classifier against mono-modal as well as early-fusion approaches, while in the second one we aim to evaluate our student engagement methodology in discriminating between the two levels of engagement. Finally, we attempt to study the role of each component of the GEQ questionnaire to the classification process.

A. Multimodal Affective State Recognition

In order to evaluate the performance of our multimodal affective state recognition algorithm, we created a new dataset with Microsoft Kinect recordings of subjects performing 5 basic emotions (Anger, Fear, Happiness, Sadness, Surprise), which are commonly encountered in a typical game (see Fig.7). We defined a “neutral” category to classify all frames where there is no motion indicating any of the distinct remaining emotion classes. Body movements and facial expressions were selected based on the literature, as described in Section IV. The dataset contains 750 videos from 15 subjects [62]. Each video begins with a close to neutral expression and proceeds to a peak expression. The total duration of each video is 3 seconds. Subjects were shown a short video with the aforementioned movements and afterwards they were asked to perform 5 times each movement according to their personal style. The emotion label refers to what expression was requested rather than what may actually have been performed.

Inspired by the work in [63] we propose the training of the fusion model using an augmented noisy dataset with additional samples that have only a single-modality as input. In practice, we added samples that have neutral state for one of the input modalities (e.g., face) and original values for the other input modality (e.g., body). Thus, a third of the training data contains only facial expressions, while another third contains only body gestures, and the last one has both face and body information. The figure below showcases a selection of these movements.

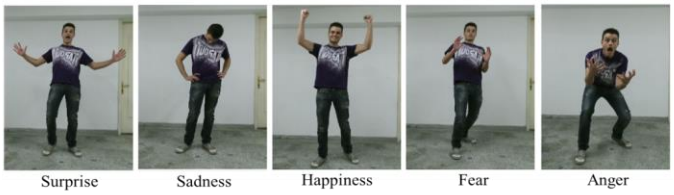


Fig.7. Dataset Movements expressing five emotions

For the evaluation of the proposed multimodal affective state recognition method, we examined whether the proposed fusion algorithm performs better than the intermediate mono-modal classifiers, i.e., the classifiers that are based only on facial expression analysis or body motion analysis, as well as a number of different multimodal approaches (e.g., Linear Weighted, Non Linear SVM and Shallow NN). As shown in Fig.8, the proposed model outperforms all other methods, both mono-modal and multimodal (early and late fusion approaches), with a recognition rate of 98.3%. As we can see, the two mono-modal classifiers provide high recognition rates similar to those of early fusion algorithms, i.e., non Linear SVM and Shallow NN, while the proposed fusion method outperforms the linear weighted-based late fusion approach, with an improvement of 6.7%. We have to note here that in order to have a fair comparison between the different classification methods, we applied the same validation approach to all experiments. More specifically, for the experimental results of Fig.8, we applied cross validation using the same folds for all methods and then we averaged all classification rates.

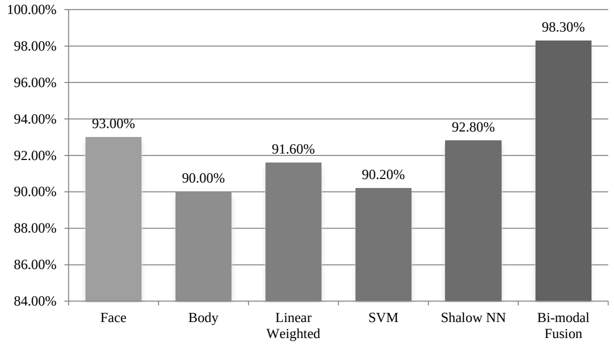


Fig.8. Comparison of the proposed fusion algorithm against face and body mono-modal classifiers, two early fusion approaches (non-linear SVM and Shallow NN) and a late fusion multimodal method (Linear Weighted).

B. *Engagement Recognition*

In this section, we aim to evaluate the performance of the proposed automatic engagement recognition method and at the same time to present a thorough analysis related to the three dimensions of engagement, i.e., behavioral, cognitive and affective. For the experimental evaluation of our method, we considered two classes, “Not Engaged” and “Engaged”, corresponding to engagement value intervals of [-2, 0] and (0, 2], respectively. More specifically, the average engagement value of students playing the standard version of the PoT game was 0.728, with a standard deviation of 0.48 (see Table I), while the mean engagement value of students with the stripped-down version of PoT was -0.483, with a standard deviation of 0.396, as shown in Table II. As we can also see in Fig.9 the majority of participants showed a clear preference for the original version of the game. These results also verify our hypothesis that the use of games with different degrees of challenge can trigger different levels of engagement to the users and provide an alternative, reliable approach for annotating engagement data.

TABLE I
DESCRIPTIVE STATISTICS FOR POT GAME

	N	Min	Max	Mean	Std. Dev.
Total	72	-0.878	1.564	0.728	0.48

TABLE II
DESCRIPTIVE STATISTICS FOR THE STRIPPED-DOWN VERSION OF POT GAME

	N	Min	Max	Mean	Std. Dev.
Total	72	-1.651	0.359	-0.483	0.396

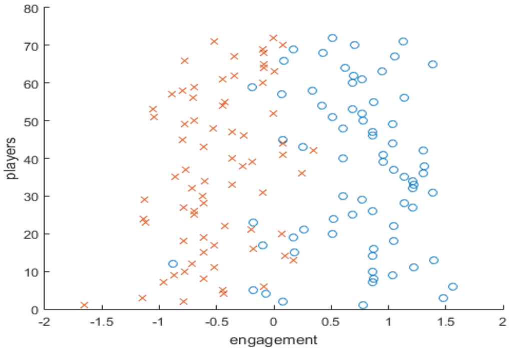


Fig.9. Students' individual engagement scores, based on GEQ analysis. Red ‘x’ marks indicate the engagement level of players corresponding to the stripped-down version of the game, while blue circles correspond to the PoT.

Experimental results in all classes showed that the PoT game increased the engagement level of students, i.e., the mean engagement of students in all classes is positive, while the engagement level of students with the stripped down version of the game is negative (see Table III and Table IV). Results also indicate that the difference in engagement level of students between the two versions of the game was statistically significant in all classes, with z>1.96 or z<-1,96 and p<0.05. In other words, the results support the alternative hypothesis H_1, i.e., games with different degrees of challenge can trigger different levels of engagement, and, therefore, we can easily reject the null hypothesis H_0, i.e., the engagement level of students remains the same. In addition, by analyzing the engagement values of participants with regards to their scores in PoT game, we can clearly see (Table V) that students with high scores (N_{HS}=32) tend to have higher engagement values, i.e., mean value m=0.8 and standard deviation std=0.484, than those with low scores (N_{LS}=40), i.e., mean value m=0.671 and standard deviation std=0.475. Of course the populations of these two categories belong to the ‘Engaged’ class, i.e., the mean engagement of students is positive. As we can see in Table V, there is a statistically significant difference in the engagement level of these two groups (students with high and low scores) from the engagement level of students played the stripped down version. Hence, we can easily reject the null hypothesis H_0 even in the case of students with low scores in PoT game.

TABLE III
ENGAGEMENT STATISTICS FOR POT GAME

	Class1	*Class2*	*Class3*	*Class4*
Mean	0.628	0.930	0.763	0.588
STD	0.643	0.417	0.308	0.423
SE	0.088	0.093	0.093	0.098
z-Score	**12.630**	**15.242**	**13.441**	**10.885**
p-Value	**0.0**	**0.0**	**0.0**	**0.0**

TABLE IV
ENGAGEMENT STATISTICS FOR THE STRIPPED-DOWN VERSION OF POT GAME

	Class1	*Class2*	*Class3*	*Class4*
Mean	-0.57	-0.593	-0.421	-0.321
STD	0.444	0.376	0.391	0.320
SE	0.107	0.112	0.112	0.119
z-Score	**-12.180**	**-11.754**	**-10.232**	**-8.800**
p-Value	**0.0**	**0.0**	**0.0**	**0.0**

TABLE V
ENGAGEMENT STATISTICS FOR STUDENTS WITH HIGH AND LOW SCORES USING POT GAME

	High Scores	*Low Scores*
Mean	0.8	0.671
STD	0.484	0.475
SE	0.084	0.075
z-Score	**18.334**	**18.427**
p-Value	**0.000**	**0.000**

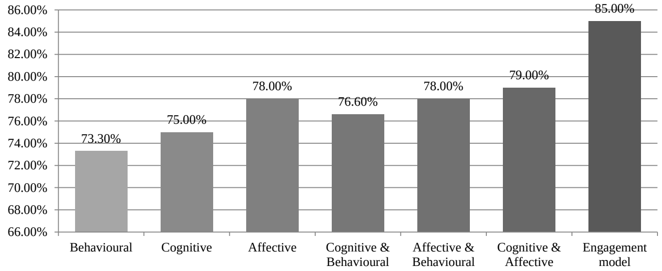


Fig.10. Comparison of the proposed methodology against behavioral, cognitive, affective engagement and their combinations in pairs.

For the evaluation of our student engagement recognition method, we applied a cross validation approach with four folds. The proposed method is compared against behavioral, cognitive, affective engagement and their combinations in pairs. More specifically, as shown in Fig.10, our student engagement recognition method outperforms all other approaches with a classification rate of 85%. As we can also see, the role of the affective dimension of engagement is crucial in the classification process, with a detection rate of 78% against 73.3% and 75% for behavioral and cognitive engagement, respectively. However, the two other dimensions (behavioral and cognitive) provide also valuable engagement information (with a classification rate of 76.6%), contributing significantly to the classification process, especially when there is a lack of any emotional expression.

C. GEQ Analysis

In the last stage of our analysis, we aim to study the role of the four components of the Game Engagement Questionnaire, i.e., immersion, presence, flow and absorption, in the classification process of the proposed student engagement recognition algorithm. Table VI presents in detail the engagement statistics of each component for the two versions of the game (left side: PoT game, right side: stripped-down version of PoT). The statistical analysis shows that all components of GEQ questionnaire reject hypothesis H_0, with z>1.96 or z<-1,96 and p-value<0.05. As we can see from Table VI, students of class 4 were more absorbed than the students of the other classes when playing the stripped-down version of the PoT game, with a mean engagement value m=0.527 (z=-6.226, p=0.0). However, this engagement value is much lower than the average engagement value (m=1.644) of absorption component for all classes using the standard version of PoT game and lower than the engagement value produced by the same group of students playing the PoT game (m=1.611).

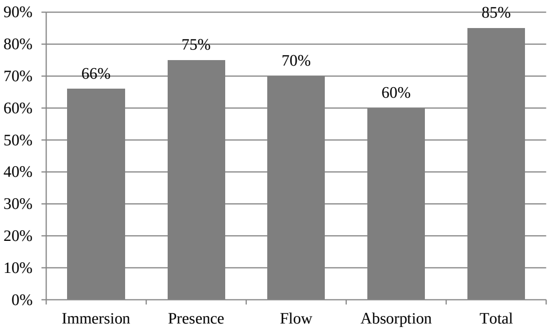


Fig. 11. Comparison of the total classification rate of the method against the classifcation rate of each component, i.e., immersion, presence, flow and absorption.

In Fig.11, we present the classification rates of each component and we compare them with the classification rate of the proposed method in order to validate their contribution to the total classification rate. To have comparable results, each time we trained our classifier with the same set of features and we considered again two classes "Not Engaged" and "Engaged". The engagement value intervals corresponding to each class are defined separately for each component based on the mean values of Table VI, i.e., we found the mean values of each component for the PoT game and its stripped down version, and then we estimated the average in order to define the value intervals corresponding to each class. As we can easily see, all components contribute significantly to the classification process, with presence and flow playing the most crucial role, i.e., detection rates 75% and 70% respectively. On the other hand, the absorption component seems to contribute less than the other three components of GEQ, with a detection rate of 60%. This is mainly due to the fact that questions belonging to absorption component lead generally to high engagement

TABLE VI

ENGAGEMENT STATISTICS FOR THE FOUR GEQ COMPONENTS USING POT GAME (LEFT SIDE) AND ITS STRIPPED-DOWN VERSION (RIGHT SIDE)

Immersion	*Class1*	*Class2*	*Class3*	*Class4*
Mean	-0.92	-0.17	-0.17	-0.274
STD	1.265	0.00	0.00	0.418
SE	*0.26*	*0.275*	*0.275*	*0.291*
z-Score	**4.765**	**7.252**	**7.252**	**6.479**
p-Value	**0.0**	**0.0**	**0.0**	**0.0**

Presence	*Class1*	*Class2*	*Class3*	*Class4*
Mean	-0.186	0.294	-0.169	-0.078
STD	0.808	0.709	0.616	0.714
SE	*0.157*	*0.166*	*0.166*	*0.176*
z-Score	**6.268**	**8.843**	**6.046**	**6.217**
p-Value	**0.0**	**0.0**	**0.0**	**0.0**

Flow	*Class1*	*Class2*	*Class3*	*Class4*
Mean	0.639	0.746	0.718	0.369
STD	0.672	0.407	0.417	0.51
SE	*0.092*	*0.097*	*0.097*	*0.103*
z-Score	**10.796**	**11.348**	**11.053**	**7.029**
p-Value	**0.0**	**0.0**	**0.0**	**0.0**

Absorption	*Class1*	*Class2*	*Class3*	*Class4*
Mean	1.569	1.773	1.625	1.611
STD	0.859	0.628	0.629	0.779
SE	*0.114*	*0.12*	*0.12*	*0.127*
z-Score	**12.199**	**13.264**	**12.037**	**11.241**
p-Value	**0.0**	**0.0**	**0.0**	**0.0**

Immersion	*Class1*	*Class2*	*Class3*	*Class4*
Mean	-2.169	-2.114	-2.116	-2.254
STD	1.489	1.308	0.855	0.960
SE	0.168	0.177	0.177	0.188
z-Score	**-10.531**	**-9.682**	**-9.692**	**-9.874**
p-Value	**0.0**	**0.0**	**0.0**	**0.0**

Presence	*Class1*	*Class2*	*Class3*	*Class4*
Mean	-1.220	-1.206	-1.115	-1.137
STD	0.801	0.581	0.866	0.568
SE	0.162	0.171	0.171	0.181
z-Score	**-7.301**	**-6.838**	**-6.308**	**-6.069**
p-Value	**0.0**	**0.0**	**0.0**	**0.0**

Flow	*Class1*	*Class2*	*Class3*	*Class4*
Mean	-0.485	-0.476	-0.195	-0.234
STD	0.492	0.463	0.307	0.266
SE	0.117	0.123	0.123	0.131
z-Score	**-9.490**	**-8.924**	**-6.650**	**-6.566**
p-Value	**0.0**	**0.0**	**0.0**	**0.0**

Absorption	*Class1*	*Class2*	*Class3*	*Class4*
Mean	0.117	-0.006	0.104	0.527
STD	0.555	0.376	0.487	0.501
SE	0.160	0.169	0.169	0.179
z-Score	**-9.522**	**-9.765**	**-9.109**	**-6.226**
p-Value	**0.0**	**0.0**	**0.0**	**0.0**

values for both games, as shown in Table VI (despite the fact that PoT produces higher engagement levels than the stripped down version of the game). In any case, the analysis of Fig.11 makes evident that each of the four components contributes to the classification process producing in total a classification rate of 85%.

VII. CONCLUSIONS AND DISCUSSION

In this paper we presented a novel methodology for the automatic recognition of student engagement in prosocial games that combines engagement cues from both the user and the game. Inspired by the theoretical grounds of educational psychology, the proposed method aims to capture the different dimensions of engagement, i.e., behavioral, cognitive and affective, by exploiting real-time engagement cues from different input modalities. More specifically, we apply body motion and facial expression analysis to identify the affective state of students, while we extract features related to their cognitive and behavioral engagement based on the analysis of their interaction with the game.

Experimental results showed that the role of the affective engagement is crucial in the classification process, providing higher detection rate than the two other engagement dimensions. However, by considering cognitive and behavioral engagement, we can further improve the classification accuracy. Moreover, these two dimensions can provide valuable engagement information, especially, when no sensor technology is used for the recognition of player's emotion, e.g., in the case of mobile games, or when there is a lack of any emotional expression (e.g., players suffering from autism, depression etc.). Here we have to note that the training of our classifier was performed with real engagement data, i.e., the students participated in our experiments were free to express or not their emotions. Hence, in some cases the detected emotional state was the neutral class or the average variation of the player's affective state in the Valance-Arousal space was rather small. However, due to the combination of the three engagement components, the proposed method showed promising results with a classification accuracy of 85%.

To improve the recognition of the affective engagement, in this paper we propose a multimodal affective recognition algorithm, which improves the classification accuracy of mono-modal classifiers, e.g., based on facial expression analysis. Currently, we are working on the combination of sound analysis and heart rate analysis (using bands) with body and facial expression analysis in order to further improve the robustness of our algorithm. Other sensor technologies, such as ECG or EEG sensors, could also be used in the future, but only for research purposes, since these solutions are not considered suitable for learning applications.

For the collection and annotation of the engagement data, we introduced a novel approach, which is based on the use of games with different degrees of challenge in conjunction with a retrospective self-reporting method, i.e., GEQ questionnaire. A detailed analysis of the experimental results in terms of the contribution of GEQ components in the classification process was also elaborated in this paper, showing that flow and presence components play the most crucial role in the recognition of student engagement.

Since the engagement data collected using a retrospective approach may be biased, in our experiment we used an adequate number of students (72 students from 4 different classes) and two games with a quite different degree of challenge. The experimental results and the statistical analysis for each class verify our hypothesis that the use of games with different degrees of challenge triggers different levels of engagement. More specifically, the engagement of students in all classes was positive for the PoT game and negative for the stripped-down version of the game. As we can see in Fig. 9, there are only few outliers, which in practice do not significantly affect the classification process of our neural network. During the training of our classifier, we normalized our features to be independent of the time (affective and cognitive engagement) and the total number of goals (behavioral engagement). For this reason, the proposed engagement recognition algorithm can be used by other game developers to collect engagement data in real-time (i.e., during the game play) for any time interval in their game.

In the future, we aim to use the proposed methodology of engagement recognition in combination with an on-line adaptation algorithm in order to develop sophisticated games that will provide personalized learning through dynamic gameplay adaptation.

Acknowledgment

The research leading to this work has received funding from the EU Horizon 2020 Framework Programme under grant agreement no. 644204 (ProsocialLearn project).

References

[1] E. N. Wiebea, A. Lamba, M. Hardya, D. Sharekb, “Measuring engagement in video game-based environments: Investigation of the user engagement scale”, *Computers in Human Behavior*, vol. 32, March 2014, pp. 123–132.

[2] R. Larson, and M. Richards, “Boredom in the middle school years: blaming schools versus blaming students”, *American Journal of Education*, vol. 99, no.4, Aug 1991, pp. 418-43.

[3] F. M. Newmann, *Student Engagement and Achievement in American Secondary Schools*, Teachers College Press, 1992.

[4] P. Sullivan, A. Mornane , V. Prain, C. Campbell, C. Deed, S. Drane, M. Faulkner, A. Mcdonough, C. Smith, “Junior secondary students' perceptions of influences on their engagement with schooling”, *Australian. Journal of Education*, vol. 53, no. 2, 2009, pp. 176-191.

[5] J. A. Fredricks, P.C. Blumenfeld, A.H., Paris, “School engagement: potential of the concept, state of the evidence”, *Review of Educational Research*, vol.74, no.1, 2004, pp. 59-109.

[6] J. Parsons, L. Taylor, “Improving student engagement”, *Current Issues in Education*, vol.14, no.1, 2011, pp. 1-32.

[7] J. Reeve, H. Jang, D. Carrell, S. Jeon, J. Barch, “Enhancing students' engagement by increasing teachers' autonomy support”, *Motivation and Emotion*, vol.28, no.2, 2004, pp. 147-169.

[8] N. Zepke, L. Leach, P. Butler, “Student engagement: What is it and what influences it”, *Teaching and Learning Research Initiative*, 2010, pp. 1-14.

[9] T. M. Akey, *School context, student attitudes and behavior, and academic achievement: An exploratory analysis*”, MDRC publication, Jan. 2006.

[10] S. Saeed, D. Zyngier, “How motivation influences student engagement: a qualitative case study”, *Journal of Education and Learning*, vol. 1, no.2, 2012, pp. 252-267.

[11] M. E. Bulger, R. E. Mayer, K. C. Almeroth, S. D. Blau, “Measuring learner engagement in computer-equipped college classrooms”, *Journal of Educational Multimedia and Hypermedia*, vol. 17, no. 2, 2008, pp. 129-143.

[12] N. Zepke, “Improving student engagement: Ten proposals for action”, *Active Learning in Higher Education*, vol. 11, 2010, pp. 167–77.

[13] M. Prensky, “The motivation of gameplay: the real twenty-first century learning revolution”, *On the horizon*, vol. 10, no. 1, 2002, pp. 5-11.

[14] D. J. Shernoff, M. Csikszentmihalyi, B. Shneider, E.S. Shernoff, “Student engagement in high school classrooms from the perspective of flow theory”, *School Psychology Quarterly*, vol. 18, no. 2, 2003, pp. 158-176.

[15] P. Chatterjee, “Entertainment, engagement and education in e-learning”, *Training & Management Development Methods*, vol. 24, no. 2, 2010, pp. 601-621.

[16] J. D. Finn, G. M. Pannozzo, K. E. Voelkl, “Disruptive and inattentive withdrawn behavior and achievement among fourth graders”, *The Elementary School Journal*, 1995, pp. 421-434.

[17] J. Culver, Relationship Quality and Student Engagement, Wayne State University, Jan. 2015.

[18] A. R. Anderson, S. L., Christenson, M. F. Sinclair, C. A. Lehr, “The importance of relationships for promoting engagement with school”, *Journal of School Psychology*, vol. 42, no. 2, 2004, pp. 95–113.

[19] J. J. Appleton, S. L. Christenson, D. Kim, A. L. Reschly, “Measuring cognitive and psychological engagement: Validation of the student engagement instrument”, *Journal of School Psychology*, vol. 44, 2006, pp. 427-445.

[20] J. Beck “Engagement tracing: Using response times to model student disengagement”, *Artificial Intelligence in Education*, 2005, pp. 88-95.

[21] J. Whitehill, Z. Serpell, Y-C. Lin, A. Foster, J. Movellan, “The Faces of Engagement: Automatic recognition of student engagement from facial expressions” *IEEE Transactions on Affective Computing*, vol. 5, no. 1, Jan.-Mar. 2014, pp. 86-98

[22] H. Monkaresi; N. Bosch, Nigel; RA Calvo; S. D’Mello, “Automated Detection of Engagement using Video-Based Estimation of Facial Expressions and Heart Rate”, *IEEE Transactions on Affective Computing*, vol. 8, no. 1, Jan.-March 2017, pp. 15-28.

[23] K. Apostolakis, A. Psaltis, K. Stefanidis, K. Kaza, S. Thermos, K. Dimitropoulos, E. Dimaraki, P. Daras, "Exploring the Prosociality Domains of Trust and Cooperation, through Single and Cooperative Digital Gameplay in Path of Trust", *International Journal of Serious Games*, vol: 3, no: 3, 2016, pp. 39-57

[24] J. H. Brockmyer, C. M. Fox, K. A. Curtiss, E. McBroom, K. M. Burkhart, J. N. Pidruzny, “The development of a Game Engagement Questionnaire: A measure of engagement in video game-playing”, *Journal of Experimental Social Psychology*, vol. 45, no. 4, July 2009, pp. 624-634.

[25] M. Lalmas, H. O’Brien, and E. Yom-Tov “Measuring User Engagement”, *Synthesis Lectures on Information Concepts Retrieval and Services*, vol. 6, no. 4, Nov. 2014.

[26] J. Ocumpaugh, R. S. Baker, and M. M. T. Rodrigo, “Baker-Rodrigo observation method protocol 1.0 training manual”, Ateneo Laboratory for the Learning Sciences, EdLab, Manila, Philippines, Tech. Rep. 1, 2012.

[27] H. L. O’Brien, E. G. Toms, “The development and evaluation of a survey to measure user engagement”, *Journal of the American Society for Information Science and Technology*, vol. 61, 2010, pp. 50–69.

[28] E. N. Wiebe, A. Lamb, M. Hardy, D. Sharek, “Measuring engagement in video game-based environments: Investigation of the User Engagement Scale”, *Computers in Human Behavior*, vol. 32, 2014, pp. 123–132.

[29] M. H. Phan, J. R. Keebler, B. S. Chaparro, “The Development and Validation of the Game User Experience Satisfaction Scale (GUESS)”, *Human Factors and Ergonomics Society*, vol. 58, no. 8, Dec 2016, pp.1217-1247.

[30] S. D’Mello and A. C. Graesser, “Automatic detection of learners’ emotions from gross body language”, *Applied Artificial Intelligence*, vol. 23, no. 2, 2009, pp. 123–150.

[31] S. D’Mello, A. C. Graesser, “Affect detection from human-computer dialogue with an intelligent tutoring system”, Lecture Notes in Computer Science: Intelligent Virtual Agents: 6th International Conference, Berlin, Heidelberg, Germany, Springer, 2006, pp. 54-67.

[32] S. Asteriadis, P. K. Tzouveli, K. Karpouzis, S. D. Kollias, “Estimation of behavioral user state based on eye gaze and head pose - application in an e-learning environment”, *Multimedia Tools Appl.*, vol. 41, no. 3, 2009, pp. 469-493.

[33] A. Sahayadhas, K. Sundaraj, M. Murugappan, “Detecting driver drowsiness based on sensors: a review”, *Sensors*, vol. 12, no. 12, Jan. 2012, pp. 16937 – 53.

[34] A. Belle, R. H. Hargraves, K. Najarian, “An automated optimal engagement and attention detection system using electrocardiogram”, *Comput. Math. Methods Med.*, vol. 2012, Jan. 2012, pp. 1–12.

[35] S. D. Kreibig, “Autonomic Nervous System Activity in Emotion: A Review”, *Biol. Psychol*., vol. 84, pp. 394–421, 2010.

[36] S. D'Mello, S. Craig, A. Graesser, "Multimethod assessment of affective experience and expression during deep learning", *Int. J. Learn. Technol.*, vol. 4, no. 3, 2009, pp. 165–187.

[37] S. D'Mello and A. Graesser, "Multimodal semi-automated affect detection from conversational cues, gross body language, and facial features", *User Model. User-Adapted Interaction*, vol. 20, no. 2, 2010, pp. 147–187.

[38] G. Chanel, C. Rebetez, M. Betrancourt, T. Pun, "Emotion assessment from physiological signals for adaptation of game difficulty", *IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans*, vol. 41, no. 6, Nov. 2011, pp. 1052-1063.

[39] G. N. Yannakakis, and J. Hallam, "Towards Optimizing Entertainment in Computer Games", Applied Artificial Intelligence, vol. 21, 2007, pp.933-971.

[40] N. Shaker, M. Shaker, "Towards Understanding the Nonverbal Signatures of Engagement in Super Mario Bros", *User Modeling, Adaptation, and Personalization*, July 2014, pp.423-434.

[41] N. Shaker, S. Asteriadis, G. N. Yannakakis, K. Karpouzis, "Fusing visual and behavioral cues for modeling user experience in games", *IEEE Transactions on Cybernetics*, vol. 43, no. 6, 2013, pp. 1519–1531.

[42] M. S. El-Nasr and S. Yan, "Visual attention in 3-D video games," ACM SIGCHI International Conference on Advances in computer entertainment technology, 2006.

[43] B. Woolf, W. Burleson, I. Arroyo, T. Dragon, D. Cooper, R. Picard, "Affect-aware tutors: recognizing and responding to student affect", *International Journal of Learning Technology*, vol. 4, no. 3, 2005, pp. 129-164.

[44] A. Arnold, R. Scheines, J. E. Beck, B. Jerome, "Time and Attention: Students, Sessions, and Tasks", Workshop on Educational Data Mining, Pittsburgh, 2005, pp. 62-66.

[45] J. Ellis, S. Heppell, J. Kirriemuir, A. Krotoski, A.McFarlane, *Unlimited learning. Computer and video games in the learning landscape*, ELSPA - Entertainment and Leisure Software Publishers Association, 2006.

[46] Da Rocha Seixas, A. S. Gomes, J. de Melo Filho, "Effectiveness of Gamification in the Engagement of Students", *Comput. Hum. Behav.*, vol.58, 2016, pp.48-63.

[47] J. L. Sabourin, J. C. Lester, "Affect and engagement in game-based learning environments", *IEEE Transactions on Affective Computing*, vol. 5, no. 1, 2014, pp. 45–56.

[48] R. A. Calvo and S. D'Mello, "Affect detection: An interdisciplinary review of models, methods, and their applications", *IEEE Transactions on Affective Computing*, vol. 57, no. 4, 2010, pp. 18–37.

[49] E. Kruijff, A. Marquardt, C. Trepkowski, J. Schild, A. Hinkenjann, "Enhancing User Engagement in Immersive Games through Multisensory Cues", VS-GAMES 2015, pp. 1-8.

[50] Z. Yu, A. Papangelis, A. Rudnicky, "TickTock: A Non-Goal-Oriented Multimodal Dialog System with Engagement Awareness". AAAI Spring Symposium Series, 2015

[51] N. Weinstein, R. M. Ryan, "When helping helps: autonomous motivation for prosocial behavior and its influence on well-being for the helper and recipient", *Journal of personality and social psychology*, vol. 98, no. 2, 2010, pp. 222-244.

[52] L. Vuillier, C. Cook, , *ProsocialLearn D2.1 User requirements*, Tech. Report, 2015

[53] T. Greitemeyer, D. O. Mügge, "Video games do affect social outcomes a meta-analytic review of the effects of violent and prosocial video game play", Personality and Social Psychology Bulletin, vol. 40, no.5, 2014, pp. 578-589.

[54] E. I. Deci, R. M. Ryan, "The" what" and" why" of goal pursuits: Human needs and the self-determination of behavior", *Psychological inquiry*, vol. 11, no. 4, 2000, pp. 227-268.

[55] A. Psaltis, K. Kaza, K. Stefanidis, S. Thermos, K. Apostolakis, K. Dimitropoulos, P. Daras, "Multimodal affective state recognition in serious games applications", IEEE International Conference on Imaging Systems and Techniques, October 4-6, 2016.

[56] K. Kaza , A. Psaltis , K. Stefanidis , K. Apostolakis , S. Thermos , K. Dimitropoulos, P. Daras, "Body motion analysis for emotion recognition in serious games", HCI International, Toronto, Canada, July 2016.

[57] H. G. Wallbott, "Bodily expression of emotion", *European Journal of Social Psychology*, vol. 28, 1998, pp.879–896.

[58] S. Piana, A. Stagliano, A. Camurri, and F. Odone, "A set of full-body movement features for emotion recognition to help children affected by autism spectrum condition", In IDGEI International Workshop, 2013.

[59] G. E. Hinton, S. Osindero, and Y. The, "A fast learning algorithm for deep belief nets", Neural computation, vol.18, no.7, pp. 1527-1554, 2006.

[60] G. Hookham, K. Nesbitt, F. Kay-Lambkin, "Comparing usability and engagement between a serious game and a traditional online program", Australasian Computer Science Week Multiconference, 2016.

[61] L. A. Annetta, L. A., M. T. Cheng, S. Holmes, "Assessing twenty-first century skills through a teacher created video game for high school biology students", *Research in Science and Technological Education*, vol. 28, no. 2, 2010, pp. 101-114.

[62] Multimodal Affective State Recognition dataset [Online]. Available: https://vcl.iti.gr/masr-dataset.

[63] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng., "Multimodal deep learning", 28th International conference on machine learning, 2011, pp. 689-696..

The original publication can be found in:
http://ieeexplore.ieee.org/xpl/tocresult.jsp?isnumber=4804729