IEEE Access
Multidisciplinary ⁝ Rapid Review ⁝ Open Access Journal

Received November 18, 2020, accepted December 13, 2020, date of publication January 4, 2021, date of current version January 14, 2021.

Digital Object Identifier 10.1109/ACCESS.2020.3049060

Towards Higher Sense of Presence: A 3D Virtual Environment Adaptable to Confusion and Engagement

TIAGO M. SILVA PEDRO[1,3] **AND JOSÉ LUÍS SILVA**[2]
[1]Instituto Universitário de Lisboa (ISCTE-IUL), ISTAR, 1649-026 Lisbon, Portugal
[2]ITI/LARSyS/Instituto Universitário de Lisboa (ISCTE-IUL), ISTAR, 1649-026 Lisbon, Portugal
[3]ADENE—Agência para a Energia, 1050-065 Lisbon, Portugal

Corresponding author: José Luís Silva (jose.luis.silva@iscte-iul.pt)

This work was supported by the Fundação para a Ciência e a Tecnologia (FCT), Portugal, under Project UIDB/50009/2020, Project UIDB/04466/2020, and Project UIDP/04466/2020.

ABSTRACT Virtual Reality scenarios where emitters convey information to receptors can be used as a tool for distance learning and to enable virtual visits to company physical headquarters. However, immersive Virtual Reality setups usually require visualization interfaces such as Head-mounted Displays, Powerwalls or CAVE systems, supported by interaction devices (Microsoft Kinect, Wii Motion, among others), that foster natural interaction but are often inaccessible to users. We propose a virtual presentation scenario, supported by a framework, that provides emotion-driven interaction through ubiquitous devices. An experiment with 3 conditions was designed involving: a control condition; a less confusing text script based on its lexical, syntactical, and bigram features; and a third condition where an adaptive lighting system dynamically acted based on the user's engagement. Results show that users exposed to the less confusing script reported higher sense of presence, albeit without statistical significance. Users from the last condition reported lower sense of presence, which rejects our hypothesis without statistical significance. We theorize that, as the presentation was given orally and the adaptive lighting system impacts the visual channel, this conflict may have overloaded the users' cognitive capacity and thus reduced available resources to address the presentation content.

INDEX TERMS Virtual reality, distance learning, cognitive informatics, human computer interaction.

I. INTRODUCTION

Virtual Reality (VR) is not a new idea [1], but the advent of Graphics Processing Units (GPUs) in the late 90s [2] (and the increase of the power of hardware and reduction of cost) has made it more feasible for consumer applications. This enabled the possibility of Collaborative Virtual Environments (CVE), which apply Virtual Environments (VE) to situations where multiple persons co-exist through their virtual representations. Examples of these are remote virtual presentations given by companies to actual or potential stakeholders, for example for teaching purposes. These presentations involve an exchange of information from one or few emitters, to one or several receptors.

This paper considers situations where virtual receptors are connected through simple hardware setups (laptop, mobile devices) and interaction devices (keyboard & mouse, touchscreens, microphone and webcam). We therefore assume a limited range of interaction modalities. Our aim is to propose a system that improves a sense of presence through ubiquitous devices, enabling this form of communication to the common user.

Most CVEs are based on text, speech, and mouse and keyboard input with limited or no access to other interaction modalities [3], especially when it comes to nonverbal communication. Systems of this kind are used as a basis for videogames, more specifically in the MMORPG market [4]. Players are usually represented by an avatar which graphically displays this representation to other players.

We believe that a drawback hindering the proliferation of CVEs is the lack of engagement these systems provide, which

The associate editor coordinating the review of this manuscript and approving it for publication was Lei Wei.

limit user immersion. One factor that contributes to this lack of engagement is the low level of natural interaction that generates nonverbal reciprocity (and consequently failure to convey emotion) between users, something that is critical to our daily human-to-human interactions [5].

One of the factors that can impact the efficiency of natural interaction is the emotional response of the receptors. The basic set of emotions [6] is widely studied, both in psychology, and as a basis for techniques of Automatic Emotion Recognition (AER). In a presentation scenario a typical set of emotions amongst participants consists of cognitive states of engagement, confusion, frustration, and boredom. Detecting when the user is in any of these states is valuable to enable action that will improve communication.

One way to reach an audience is through emotion and cognition. By being aware of the audience's emotional and cognitive context the system should be able to act accordingly to increase the engagement level.

Emotion is also impacted by the virtual environment, as it can make users experience certain emotions or psychological and physiological states (fear, positive/negative valence, arousal [7], [8]. The weather, daytime, lighting, and many other conditions have an impact on people's emotions. In a real-life environment these parameters are not controllable, but in a VE they can be manipulated to achieve an objective. Light is essential to the human body [9] and artificial lighting can be used to manage the participant's environment and sense of well-being.

The goals of this work are to:

- Develop a general 3D CVE that provides a platform for virtual business visits to a company's physical infrastructure. The experience generated by CVE aims to compensate for the shortcomings of a real-life visit;
- Build a ubiquitous interaction platform that allows the collection of emotional context data;
- Build a presentation scenario where the sense of presence is increased by written presentation scripts enhanced by Natural Language Processing techniques and automatic lighting adaptation.

Given the problems and challenges identified, we propose the following hypotheses:

H1 *A less confusing script on a virtual presentation increases the user's sense of presence.*

H2 *The automatic adaptation of the virtual environment's lighting condition, based on the user's head pose, increases his/her sense of presence.*

Here we take advantage of confusion and the user's head pose (perceived engagement) to enhance the sense of presence. Confusion is interesting as it is this state that is triggered by stimuli that leads to a cognitive disequilibrium [10], and as D'Mello *et al.* [11] state, the confusion state and its resolution can increment the learning gain. Therefore, detecting its source enables a better adaptation of the proposed system which will facilitate the overcoming of this state and increment learning gains. On the other hand, monitoring the user's head pose may give a clue about his/her engagement of the situation and let the system adapt its virtual environment to regain the user's engagement.

This paper is structured in the following chapters:

- Section II presents a review of the literature about confusion on learning virtual environments. We describe a model for learning-related emotional states, approach how confusion has been detected on learners and how it has been dealt with, as well as framing our contribution on this context. Still in this section, we describe how lighting has been studied across several disciplines and how it can impact human behavior.
- Section III describes a generic 3D CVE upon which specific scenarios can be built. This CVE provides an interface for emotion and speech recognition resorting to ubiquitous devices. On top of this we built an application for a presentation scenario where we tested our hypotheses.
- Section IV describes the participants of this study, the experimental protocol, the application and its relevant components to test each hypothesis.
- Section V presents the evaluation and discussion of the results of the experiment.
- Section VI briefly concludes and summarizes contributions and future work.

II. LITERATURE REVIEW

A. CONFUSION ON LEARNING ENVIRONMENTS

Much research has been concerned with affective detection in distance-learning scenarios, either to assess which affective states are most easily observed and relevant to this context, and how to automatically detect them. In contrast to the set of basic emotional states (anger, contempt, disgust, enjoyment, fear, sadness, surprise) that typically occur in emotion-driven situations, there is a set of more complex emotional states related to learning contexts. D'Mello, Graesser and colleagues have been conducting extensive research concerned with identifying and detecting learning-centered affective states and adapting their Intelligent Tutoring System, AutoTutor [12], to these states. When analyzing at a fine-grained level, it is suggested that the set of emotions experienced during learning is mainly comprised of boredom, confusion, engagement/flow, frustration, delight, neutral, and surprise [13]–[17].

Some studies have been trying to perform Automatic Emotion Recognition (AER) to detect some of these states through Action Unit (AU) detection [12], [18]–[20], physiological signals. [17], learner behavior [21], conversational cues [22], and gross body language [23]. However, there is strong evidence that a subset of emotions comprised of engagement/flow, confusion, frustration and boredom occur at a higher frequency than basic emotions [24], [25].

D'Mello and Graesser have conducted an experiment [25] that yielded a model that initially hypothesized affect transitions between engagement/flow → confusion, confusion → engagement/flow, confusion → frustration and frustration → boredom. In addition, surprise and delight were

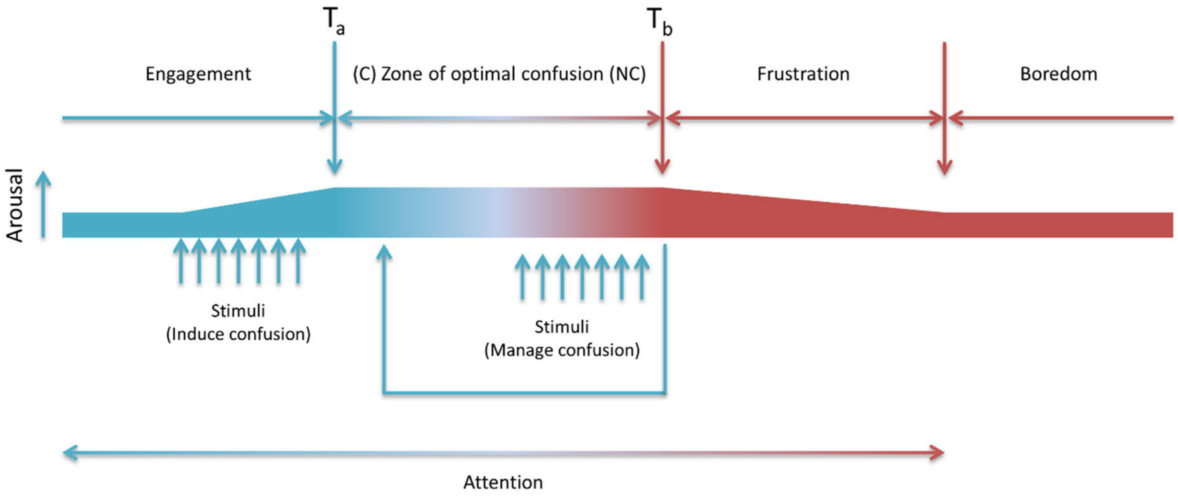


FIGURE 1. Adaptation, as proposed by Arguel [10], of the learning model of D'Mello and Graesser with the zone of optimal confusion.

occurring in the engagement/flow → confusion and confusion → engagement/flow transitions, respectively. Results confirmed most of these transitions with exception to frustration → boredom transition, which was only partially confirmed. The experiment was devised to validate the proposed model based on four hypotheses, from which the first three ones are the ones relevant for the current project:

1. The disequilibrium hypothesis states that certain stimuli lead the learner into a cognitive disequilibrium that highly relates to the engagement/flow → confusion transition;
2. The productive confusion hypothesis theorizes that the confusion → engagement/flow transition yields good learning gains as the learner can resolve the stimulus that drove him/her into the cognitive disequilibrium;
3. In opposition to the previous hypothesis, the hopeless confusion aims at explaining the confusion → frustration transition stating that in the same state of confusion the learner may not be able to resolve the stimulus that caused the disequilibrium;
4. The disengagement hypothesis states that if the learner stays in a frustration state for long, it will lead to a boredom state.

The reported instances of boredom, engagement/flow, confusion frustration and neutral states were significantly higher than delight or surprise for both studies, which aligned with the described model, where delight and surprise are not essential nodes. Results show that hypotheses 1, 2 and 3 were confirmed and hypothesis 4 was only partially supported. The first three were centered on confusion, which stresses the important role of this affective state during information acquisition. There was also evidence of additional patterns of boredom → frustration and frustration → confusion, however, this falls outside the scope of this analysis due to its lack of robustness. Thus, confusion is at the core of the work developed to test H1.

With the central role and benefits of confusion for learning, some studies were carried to induce confusion in the subject and try to manage this level of confusion and keeping it at a level of productive confusion but avoiding the evolution to frustration (hopeless confusion) [26]–[28]. This regulation of confusion has been considered as the "zone of optimal confusion" and is displayed in Figure 1 as an adapted version of previous work [10].

D'Mello *et al.* [11] study results showing evidence that a moderate state of confusion can be beneficial for learning as long as it is overcome. Most studies focus on how to react to this confusion state [12] but they do not identify what was its source.

More recently, gamification methods have been used to motivate the learner and deal with the negative emotions of the spectrum [29]. The experiment evaluated students' performance when solving programming exercises. Results show that students under the condition that used emotion recognition and gamification yielded better performance metrics (mean time per exercise of 123.3 seconds with a standard deviation of 54.9) than those under the control condition (mean time per exercise of 168.6 seconds with a standard deviation of 85.4).

Building on previous work, Arguel presented strategies and features for interactive digital learning environments based on a review of the literature [30].

B. ADAPTIVE LIGHTING CONDITIONS

The effect of lighting is being studied as a variable that influences several traits in many fields of knowledge [29]–[33]. Another field of application of lighting management is in the workplace. Hawes *et al.* [34] study the effect color temperature yield on the emotional state of participants. Four workplace scenarios were set up with lights with different color temperatures in Kelvin degrees:

- 3345 K;
- 4175 K;
- 5448 K and;
- 6029 K.

The study was carried out with 24 participants. It involved a within-participants repeated-measure design where each participant visited the laboratories in five consecutive days, consisting of a first practice day and then a series of days exposed to each of the lighting conditions. For each test, the subject took the Profile of Mood States (POMS) [35], [37] to assess his/her emotional state after and before the test to assess the differences. Our study hypothesizes that the sense of presence changes as a function of the lighting conditions in a way that can be expressed in terms of emotional parameters such as valence or arousal. In this study, results showed that higher color temperatures were related to the more aroused states and lower depression rates. Their results support the hypothesis that "…lighting can alter environmental conditions enough to increase positive mood and decrease fatigue". Therefore, they can demonstrate that lower fatigue scores result in larger frames of higher aroused states.

Due to the nature of our 3D environment, we borrowed this experiment's color temperatures to serve as the levels of lighting of the adaptive system that we will use to evaluate H2. Previous studies have also been carried out on real-life setups; however, we believe that this may be a distinctive feature of a 3D VE as it can be dynamically adapted in real-time. This of course is not possible in real-life setups. There is not much research on how the manipulation of the VE's conditions can be used to its advantage. Even though much research on adaptive VEs based on emotion is centered on MOOCs, e-learning or training scenarios [36], [37], they do not take advantage of lighting conditions. Previous studies were carried on real-life setups; however, we believe this can be a distinctive feature of a 3D VE as it can be dynamically adapted on real-time, something that for now is not possible in real-life setups. There is not much research on how the manipulation of the VE's conditions can be used to its advantage. Even though much research on adaptive VEs based on emotion is centered on MOOCs, e-learning or training scenarios [38], [39], they do not take advantage of lighting conditions.

However, Yan *et al.* [40] used a Brain-Computer Interface (BCI) device to collect data about user engagement while attending to a virtual version of the opera Siegfried and the dance "The Tramps of Horses". The user is monitored with the BCI to detect disengagement and re-engagement and act accordingly. The way the system acts is by means of a set of pre-designed performing cues as defined in classic theatre performing theory. They have focused on the effects of scenic design (which includes lighting) including display blocks and stage effects such as smoke or fog that are known to be good engagement agents. It is also known that lighting enables the stage controllers to get the audience's attention to wherever it is desired.

In the reported experiment forty-eight users were exposed to three conditions. Participants were evenly distributed according to gender, with sixteen users were allocated to each condition: 1) without any performing cues; 2) single performing cues when a certain level of engagement was detected, and 3) multiple performing cues when a certain level of engagement was detected.

Results showed that the system could detect significant variations of engagement successfully for both performances. The adaptations could recover the users' engagement when triggered.

More recently, VEs have been used to study lighting preferences in offices [41], however, the adaptation of environment lighting remains unexplored.

III. DEVELOPED SYSTEM

A system was developed to accommodate the necessary features for this work. It is a host that provides enriched interaction upon which one can develop different scenarios. A 3D office environment provides a background and top-level scenario. It is built on Unity to take advantage of its Networking API.

A. SYSTEM ARCHITECTURE

The application is aimed at virtual visitors accessing the simulation using devices possessing a display (laptop, mobile devices) and across several platforms (Windows, OSX, iOS, Android). As there could be a wide range of device specifications, we designed a client-server architecture (Figure 2) with the heavy computation and plugins on the server side to ensure compatibility. This design choice also assures that the system would depend less on the device specifications, because the specifications would only have to meet the hardware requirements of Unity and the client application would not be slowed down by the "3rd-party processing" module on the server side.

The *Interaction Manager* is a core component of the player object as it manages the *Input Modalities* classes that collect all user input. The input that was implemented was video from webcam, audio from a microphone, mouse & keyboard interaction, and touch, depending on the platform. The *Interaction Manager* possesses a set of synchronized variables (tagged as "[SyncVar]" on Unity – SyncVars are variables of scripts that inherit from NetworkBehaviour, which are synchronized from the server to clients) that allows seamless bidirectional updates through Unity's Networking API. The *Multimodal Manager* possesses a reference to the *Interaction Manager* on the server side and is listening to these messages and triggers callback events upon receiving them. They are then channeled to the right *Third-party Software* (i.e. OpenFace [54] and Affectiva [55]) and analyzed. OpenFace is an open-source tool capable of facial landmark detection, head pose estimation, facial action unit recognition, and eye-gaze estimation. Affectiva is a real-time facial expression

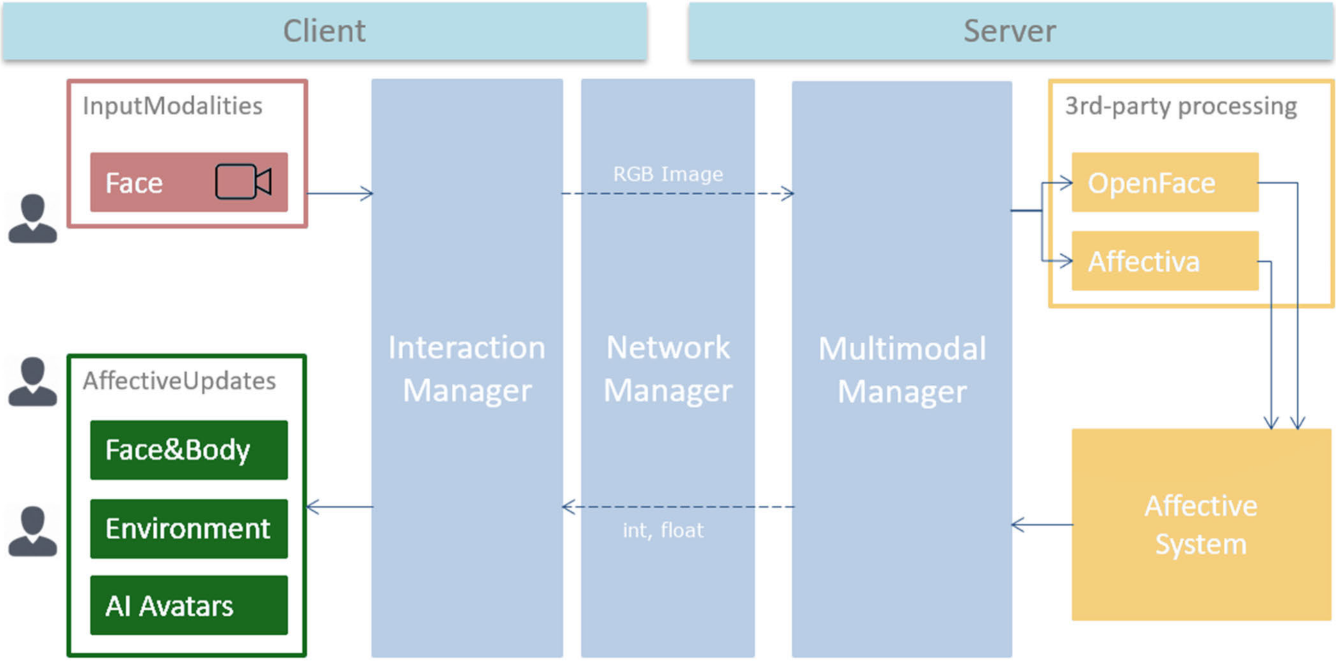


FIGURE 2. Server-client high-level architecture of the proposed system. The client provides audio and video input (RGB images) to the server which analyses it with 3rd-party software for emotion. The system acts based on this interpretation and broadcasts these changes to every user (updating float and int synchronized variables [SyncVar]. The user interface module is represented in red, in yellow the processing modules, and in green the output modules that are mapped on different components of the 3D VE.

recognition toolkit. After getting processed data from the *Third-party Software*, the *Multimodal Manager* calls the *Affective System* that interprets this data. These transformations are then updated by the *Multimodal Manager* and the *Interaction Manager* through the "[SyncVar]" variables. The *Network Manager* is a base component of Unity which provides core network functionality to the system, by synchronizing transformations, animations, and states.

IV. EVALUATION

To test both hypotheses we designed a study with three conditions (C1, C2 and C3). C1 as the control condition and C2 as the test condition are designed to test H1. H2 is tested using C2 as the control condition and C3 as the test condition. Recalling both hypotheses:

H1 "A less confusing script on a virtual presentation increases the user's sense of presence.";

H2 "The automatic adaptation of the virtual environment's lighting condition on a virtual presentation, based on the user's head pose, increases his/her sense of presence."

A. USER DESCRIPTION

Fifteen users evaluated each condition with pre- and post-questionnaires to assess the differences between each condition. Fifty-four users participated in the experiment but 9 were discarded due to initial technical problems in the data collection process (6 in C2 and 3 in C3) that was promptly resolved. The experiments used 45 participants, 14 females (31.11%) and 31 males (68.89%). Figure 3 shows that the mean age across conditions is similar, with a variance of 2.15 between C1 (SD = 6.36) and C2 (SD = 5.36), and 2.12 between C2 and C3 (SD = 6.20). Assuming a Gaussian-like distribution, two other participants were considered outliers from the sample using a 95% confidence interval as they distanced more than 1.96 standard deviations from the mean.

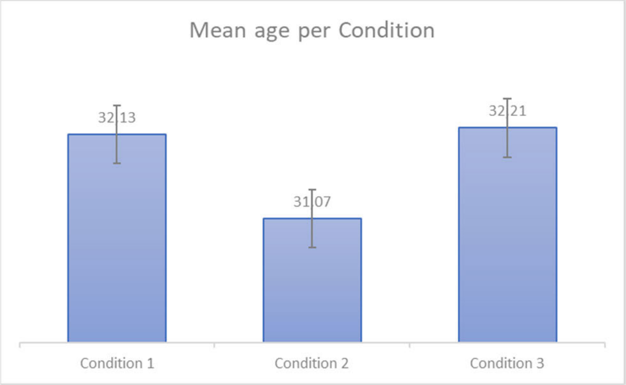


FIGURE 3. Mean age of the group of users per condition. The means are similar across all conditions with a variance of 2.15 between Condition I and II, and 2.12 between Condition II and III.

We asked every user about any hearing problems because that could affect head tracking. One pilot test showed that a user with hearing problems can unconsciously rotate the head with one ear towards the screen, as if trying to hear better. However, only one user reported hearing problems and during that test there were no reported problems arising from listening to the presentation.

B. EXPERIMENTAL SETTINGS AND PROCEDURE

Participants were allocated to slots of 30 minutes each. On average, the experiment took 25 minutes, depending on the participant. At the scheduled time, the participant was taken to an isolated room that was scheduled for these tests. The test was performed on an HP ProBook 640 G3 17" with a 1920 × 1080 resolution display with refresh rate of 60Hz and headphones.

The experiment involved three steps: first, the user was welcomed, thanked for participating in the usability test, and told that his/her help was vital to the success of the test; second, the researcher explained how the experiment would proceed, and third, the participant engaged in the experiment itself. During the second stage, the researcher explained that the experiment would be split up into three phases. First, the subject would complete a pre-questionnaire with demographics and profiling questions taken from the Immersive Tendencies Questionnaire (ITQ) [42]. Then the researcher explained that in the second phase the user would be watching a presentation, but that purpose of the presentation was to experience the 3D environment itself rather than to be concerned with its content. This was an attempt to make the participant at ease without being distracted. Our goal was that he/she would feel as comfortable and relaxed as possible, as if he/she was in a real presentation, where there is less pressure to maintain attention on the subject. The researcher also told the participant that while he/she was watching the presentation, the researcher would be present wearing headphones and facing away from the user to reduce any pressure. However, the researcher would still be present if there was any problem or if any intervention would be needed. The user was reminded that he just had to watch the presentation and did not have to perform any task. Afterwards, he/she would have to fill a final questionnaire, evaluating the scenario which had 11 questions taken from the Presence Questionnaire (PQ) [41] (Questions 1-11) [42] and others tailored specifically for this experiment (from Q12 to Q17). These questionnaires can be found in Appendix A and using a 7-point scale. Once these two stages were completed, the user was asked if he/she had any questions, doubts, or curiosities. It is important to stress that all information regarding privacy of data and freedom to leave the test at any point was also transmitted to the user immediately at the beginning of the virtual presentation.

C. VIRTUAL ENVIRONMENT

The scenario takes place in a 3D virtual meeting room and the user takes on the role of someone that is passively attending this presentation. The participant could not control the orientation of the camera. Even though this reduces immersion, it eliminates an uncontrollable variable. The presentation was composed of nine slides, each with a pre-written script, of around 11 minutes duration.

The avatar that was performing had two talking animations that were combined and synchronized with the script. Whenever there was a transition between slides, slide 'n' remained for two seconds before moving on to slide 'n+1'. Slide 'n+1' would also stay for two seconds before the avatar started speaking the respective written script. During these 4 (the 2 final seconds of slide 'n' plus the 2 final seconds of slide 'n+1') seconds of transition, the avatar would stay quiet and its hands animations were paused. This pause did not occur abruptly, the hands animation would start to fade to an idle position when it finished speaking. Hand animations were synchronized with the speech. The avatar's speech was synthesized using the IBM Watson Text-to-Speech service.[1]

The virtual environment was the same in C1 and C2, however, C3 introduced the automatic adaptation of the lighting condition. The default lighting is the one seen on Figure 4, however, in C3 it gets brighter by automatically responding to the level of the user's engagement. The scene replicates a real meeting room where a presentation is being given to visitors. There are three avatars in the scene: there is one facing the user that is giving the presentation, and two others facing away from the user that represent the visitors. These two avatars had minor animations that ran for one minute and were played on loop. The animations simulated small human gestures. The avatar that was presenting was running an animation of 3 minutes, mostly with hand gesture animation. The lighting was constant in C1 and C2, but in C3 it was adapting to the user's engagement. The script of the presentation can be found in Appendix B.

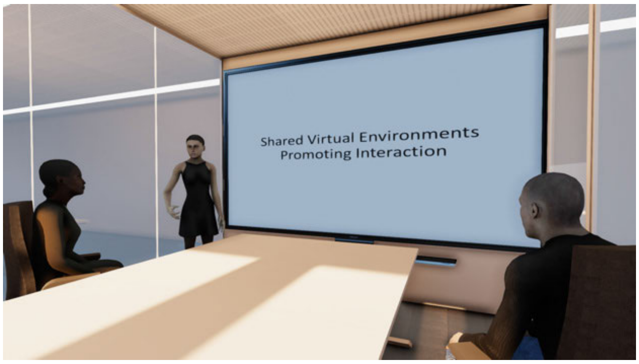
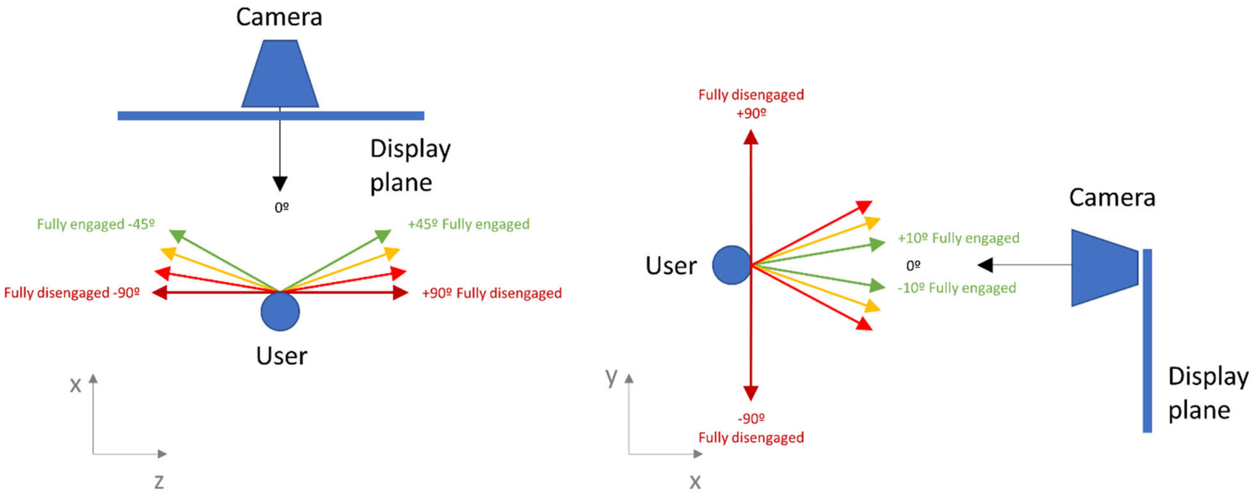


FIGURE 4. The 3D virtual environment used to test H1 and H2. There is a display where a slideshow was running, two passive avatars (back facing the user), and another presenting with synthesized speech. The lighting condition only varied on C3.

1) CONFUSION PREDICTION AND ENGAGEMENT DETECTION

To test the first hypothesis, we resorted to a confusion prediction model [43] that provided with a text excerpt of 50 words or more, generates the likelihood that excerpt has in inducing confusion on his/her listener. This likelihood can assume the discrete values of "low", "medium", or "high confusing".

[1] https://text-to-speech-demo.ng.bluemix.net/, accessed at 4th Feb 2019



(a)

(b)

FIGURE 5. a. As the angle between the user head orientation on the XZ plane (yaw rotation) and the laptop camera tends to 90°, the user engagement tends to 0. b. As the angle between the user head orientation on the XY plane (pitch rotation) and the laptop camera tends to 90°, the user engagement tends to 0.

The model was trained and tested with a set of 300 text excerpts and yielded an overall f-score of 0.57. This did not allow us to let this model identify "high" or "medium confusing" excerpts by itself, as it is only a tool meant to assist the human. Each text excerpt was annotated by five different annotators and ranked as only have "Slight agreement" between annotators with a Fleiss' Kappa coefficient of 0.16, which means this is not an easy task to gather consensus, and that establishes an upper cap for the trained model.

The lighting adaptations used to test the second hypothesis were based on the engagement of the user, that is based on his/her head pose estimation. For this experiment we consider the user is more engaged as his/her head is more directly facing the laptop camera that is placed right above the laptop screen as Figure 5 depicts. Referring to Figure 5a and 5b, when considering the yaw of the user camera, if a vector that is cast with the user's head orientation (i.e., a vector parallel to the XZ-plane with positive direction along the X-axis) has the opposite direction of a vector cast with the laptop camera's orientation (i.e., a vector parallel to the XZ-plane with negative direction along the X-axis, Figure 5a), then the user is considered to be fully engaged. The same rationale is applied when considering the pitch of the user camera (Figure 5b). The engagement level is maximum while the user camera's vector is in between the green vectors but tends to a 0 as the user camera's vector approaches a 90° angle with the laptop camera's vector. Furthermore, eye closure was also being detected through the Emotion SDK from Affectiva and a weighted mean was being calculated in order to activate the lighting environment if prolonged eye closure was detected. The head pose estimation was also being evaluated by Affectiva.

2) PRESENTATION SCRIPT

A script was written for each presentation slide (can be found in Appendix B) to be spoken by the avatar that is presenting. The script for C1 was an original one, produced by the author of this paper. It was composed of nine different files, one for each slide. The version of the script for C2 and C3 was slightly different, after being ran through the machine learning model described on the previous section.

Each file (corresponding to the text excerpt of each slide) was split into smaller parts on the first ending punctuation occurrence after the first 50 words. This procedure generated 20 text excerpts that were fed into the classification model that classified them as having low, medium, or high predicted confusion. We also pulled the features from this analysis to further investigate and understand which parts should be rewritten.

Table 1 shows this data containing an "ID" field identifying the text excerpts produced from the nine script files, nine fields that correspond respectively [43] to:

- MLT – Mean Length of Text;
- VP/T – Verb Phrases per T-Unit;
- DC/T – Dependent Clauses per T-Unit;
- T/S – T-Units per Sentence;
- LS1 – Lexical Sophistication;
- VS1 – Verb Sophistication;
- NDWERZ – Number of different words (expected random 50);
- VV2 – Verb Variation;
- AdjV – Adjective Variation;
- ("conf") that corresponds to the classification of confusion (0 = low confusion, 1 = medium confusion, 2 = high confusion).

TABLE 1. Classification report of the original script. The nine parts of the script were split into 20 text excerpts and ran through the pipeline of the previous chapter. The "conf" field is the classification and the other fields are the lexical and syntactical features. N-gram features were not included in this table. 0 = low confusion, 1 = medium confusion, 2 = high confusion.

ID	MLT	VP/T	DC/T	T/S	LS1	VS1	NDWERZ	VV2	ADJV	conf
0	0.0413	0.1111	0.1500	0.6800	0.5185	0.1500	0.5972	0.7949	0.0909	0
1	0.1606	0.2222	0.2667	0.2000	0.4630	0.2000	0.8194	0.8718	0.3939	0
2	0.2202	0.2778	0.2000	0.2000	0.2778	0.0000	0.6319	0.4615	0.2727	1
3	0.4611	0.2222	0.1333	0.2000	0.6667	0.2833	0.9097	0.2821	0.4546	1
4	0.1615	0.0000	0.0000	0.2000	0.4815	0.0000	0.8958	0.3846	0.4546	0
5	0.1224	0.0741	0.1333	0.2000	0.2222	0.0000	0.9167	0.4872	0.4242	0
6	0.2202	0.2778	0.4000	0.4666	0.6667	0.2333	0.6597	0.5128	0.4849	0
7	0.5183	0.4444	0.6667	0.2000	0.5926	0.0000	0.6042	0.4103	0.4849	0
8	0.3561	0.4444	0.4000	0.2000	0.7037	0.2833	0.8333	0.7949	0.0909	0
9	0.5278	0.5185	0.4000	0.0000	0.6296	0.8333	0.4722	0.5897	0.3030	1
10	0.4038	0.8889	1.0677	0.2000	0.7407	0.2833	0.3750	0.9231	0.3333	2
11	0.0938	0.1111	0.1333	0.3600	0.8148	0.2000	0.6389	0.4615	0.5455	0
12	0.2774	0.1667	0.0000	0.2000	0.7963	0.3333	0.6250	0.4872	0.0606	1
13	0.2560	0.1667	0.3000	0.2000	0.7222	0.2333	0.5903	0.4103	0.4242	0
14	0.3466	0.1482	0.0000	0.2000	0.6482	0.4833	0.4931	0.4872	0.2727	0
15	0.2417	0.4444	0.4000	0.4666	0.4815	0.0000	0.3056	0.5385	0.1818	0
16	0.3418	0.2222	0.4000	0.2000	0.6111	0.2000	0.6597	0.4615	0.3636	1
17	0.2417	0.1111	0.2000	0.2000	0.6852	0.0000	0.5625	0.2564	0.4242	0
18	0.2302	0.2222	0.0800	0.2000	0.7222	0.5500	0.6181	0.6154	0.2424	0
19	0.0986	0.2963	0.3333	0.3600	0.4630	0.0000	0.2153	0.7436	0.2727	0

We analyzed this data to check which parts of the script should be rewritten and why they should be rewritten, according to the most relevant features for these parts. However, we did not follow this rigorously, since this tool is meant to assist humans in this task, rather than completely replacing them, in part due to its f-score value that does not allow it to work without human assistance. We used it to help us redesign the script and provide us with a guideline on what to rewrite. We wrote the script not trying to over-complicate it to get clearer results, but to stay faithful to what would be written in a normal context, even though we admit there may be bias in this method. Instead, we tried to write a natural script and thus we were not expecting the classification model to report many highly confusing parts.

D. METHODOLOGY

To test both hypotheses, we conducted an experiment under three different conditions. These conditions allowed us to test both hypotheses. These three conditions are:

- C1. The presentation with an original script;
- C2. The same presentation as in C1 but with a slightly different script, where this text was ran through a confusion predictive model [43] to aid in the rewriting of the script to report lower confusion levels;
- C3. The same presentation and script as in C2, but with the presence of automatic lighting adaptations based on the user's detected engagement.

The independent variables are the script that the avatar used to carry the presentation and the presence of automatic adaptation of the lighting conditions based on the user's engagement. The main dependent variable is the user's sense of presence. However, other variables were monitored, such as the reported confusion about the avatar's performance or the subject of the presentation. These variables were collected through self-report.

H1 was tested using the results of C1 and C2, where the variable was the script that was spoken by the avatar. The script for C1 was originally written without any analysis, whereas the script for C2 was a redesigned version of the original one. The original version was analyzed by a machine learning model [43] and had some parts identified as being low, medium or highly confusing. Based on its output, the parts that were reported as medium or highly confusing were rewritten to lower these levels. This redesigned script was used for C2 setting the variable to be tested. On C3 we kept this redesigned script but also introduced an automatic system that changes the lighting condition according to the user's detected engagement. The results from the pair of C2 and C3 allowed us to test H2.

V. RESULTS AND DISCUSSION

A. RESULTS OF H1

The ITQ lets us have insight about the level of immersion in an activity. The statistical results of ITQ of C1 and C2 are

TABLE 2. ITQ results for C1 and C2. Both groups concentrate well on enjoyable activities and lose track of time when doing so. Do not get immersed when watching sports, but one group is more leaned towards movies or TV dramas, whereas the other is more leaned to playing sports.

	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12	Q13
							Condition I						
M	4,67	4,47	4,67	3,80	3,50	4,47	2,60	3,00	3,93	5,14	5,67	4,13	5,20
SD	1,40	1,77	1,54	1,26	1,65	1,51	1,24	1,46	1,53	1,41	0,72	2,20	1,66
							Condition II						
M	5,14	4,36	3,86	4,43	4,15	4,17	3,00	3,57	4,21	4,21	5,86	3,50	5,64
SD	1,46	1,34	1,41	1,79	1,99	1,27	1,57	1,91	2,04	1,63	0,77	2,44	1,39

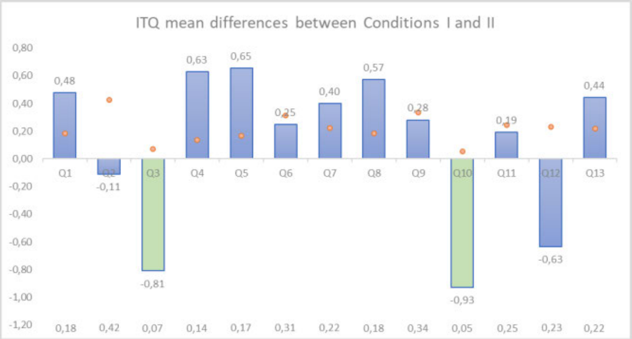


FIGURE 6. Mean differences and respective p-values from ITQ from C1 and C2. Users from C1 feel more involved in movies and when playing sports.

displayed on Table 2. The group of users from C1 reported results below or equal to 3 on Q7 and Q8 above or equal to 5 for Q10, Q11, and Q13. This user group is characterized by losing track of time when they are enjoying the activity they are performing, especially if that is playing sports. On the other hand, it does not seem like a group that gets involved when playing a passive role instead of actively participating when watching sports. Furthermore, they do little daydream, which may seem connected to the fact that they enjoy sports, something that leaves little room for daydreaming.

The user group from C2 reported scores below or equal to 3 only for Q7 and above or equal to 5 on Q1, Q11, and Q13. This user group has some touchpoints with the previous on Q7, Q11, and Q13. As the previous group, it also does not get much involved when acting passively when playing sports. It also reports the same trait of involving well on enjoyable activities and losing track of time when doing so. However, in this case, this may happen more when watching movies or TV dramas.

Figure 6 shows the mean difference between the ITQ results of users of C1 and C2. Blue and green bars represent this mean and the red dots are the p-values for each question (their respective values are at the bottom of the table). Q10 (M = -0.93, p = 0.05) shows a p-value of 0.05 and the highest difference between both groups, meaning that users from C1 become significantly more involved when playing sports. Q3 (M = -0.81, p = 0.07) may show C1 becomes more unaware of things happening around when watching movies. Positive average values mean that it is higher for C2, whereas negative average values means that it is higher on C1. This shows that users from C2 are more drawn to videogames and character development (although without statistical significance) when compared to users from C1, whereas the latter feel more involved when playing sports. However, there is a low overall mean difference of 1.41 on the ITQ scores between these groups.

Table 3 shows the statistical results for the PQ for C1 and C2. The group of users from C1 reported results below or equal to 3 on Q12, Q13, and Q17, and above or equal to 5 for Q5, Q6, Q10, Q15, and Q16. Results from C1 show that from an overall view, the presentation subject and the avatar's performance were well accepted with values below 3. Furthermore, and as expected, users also reported a low value (with low standard deviation) when asked how noticeable the changes in the lighting condition were, and that is because this condition was static. High scores on the other questions show that this group was immersed on the virtual environment and that the lighting condition is something that relates with attention kept on the presentation.

The user group from C2 reported scores below or equal to 3 for Q12, Q13, and Q17, and above or equal to 5 on Q5, Q6, Q8, Q9, Q10, and Q15. The overall results from C2 follow those of C1 with low scores on how much confusion was induced through the avatar performance and the presentation subject, as well as low noticeable changes on lighting condition. This group shows higher scores on questions related to sense of presence when compared to users from C1.

Figure 7 shows the mean differences of the PQ between C1 and C2. There are no questions with a p-value below our significance threshold, however, there are some interesting differences. Q12 (M = -0.68, p = 0.11) and Q13 (M = -0.61, p = 0.13) relate to the confusion the avatar's presentation script or the presentation subject induce, and we can see that there are indications that on C2 users perceived them as less confusing, aligned with our goal of lowering confusion reports with the rewritten script. Q1 (M = 0.71, p = 0.10) and Q4 (M = 0.70, p = 0.09) show that users reported they were more involved with the visual aspects of the environment and

TABLE 3. Results from PQ and tailored questions for C1 and C2. PQ questions go from Q1 to Q11, the rest is tailored for this experiment. Users from both groups reported low scores on how confusing the presentation subject and the avatar's performance was.

	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12	Q13	Q14	Q15	Q16	Q17
									Condition I								
M	4,00	3,93	4,13	3,87	5,27	5,40	4,71	4,67	4,67	6,00	3,21	2,47	2,47	4,07	5,73	5,07	1,47
SD	1,25	1,67	1,60	1,46	1,71	1,59	1,71	1,88	1,35	1,07	1,52	1,60	1,77	1,49	1,16	1,03	0,74
									Condition II								
M	4,71	4,43	4,29	4,57	5,79	5,50	4,23	5,21	5,00	6,07	3,23	1,79	1,86	4,00	5,64	4,93	2,86
SD	1,68	1,65	1,90	1,34	1,19	1,34	1,42	1,05	1,80	1,49	1,19	1,31	1,03	1,75	1,01	1,38	1,70

TABLE 4. ITQ results for C2 and C3. Like other groups, Condition III user group also reports easiness on concentrating on enjoyable activities and losing track of time, but do not possess any standout traits like the other two.

	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12	Q13
							Condition II						
M	5,14	4,36	3,86	4,43	4,15	4,71	3,00	3,57	4,21	4,21	5,86	3,50	5,64
SD	1,46	1,34	1,41	1,79	1,99	1,27	1,57	1,91	2,04	1,63	0,77	2,44	1,39
							Condition III						
M	4,21	4,50	4,29	3,36	3,50	4,36	3,46	3,29	3,36	4,77	5,29	4,00	5,21
SD	0,80	1,40	1,20	1,28	2,10	1,60	1,85	1,54	1,39	1,64	0,91	1,92	0,89

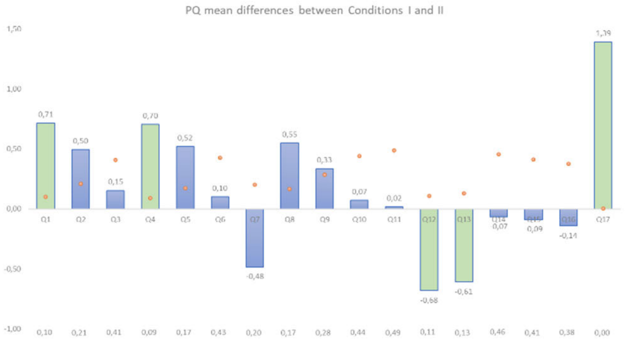


FIGURE 7. Mean differences and respective p-values from PQ and tailored questions from C1 and C2. The presentation subject (Q13) and avatar's performance (Q12) had lower scores on C2. Users also reported higher values of immersion on C2.

recognized it as more consistent with their real-world experiences. We were expecting slight differences on questions from the PQ because there were no differences on the visual aspects between both conditions, but we were expecting that a less confusing script would lead towards greater sense of presence. The group from C2 is characterized by being more drawn to videogames, which may have helped them be more involved in the experiment (there can also be the opposite perspective, as they are used to videogames, they may have higher expectations and therefore could feel less involved when compared to users from C1). The other reason may be related to the allocation of resources the users give to the visual and auditory channel. They reported less confusion towards the subject and the avatar's presentation script, which may have released them to be more attentive to the visual aspects of the environment. This same reason may explain in part the statistically significant difference on Q17 ($M = 1.39$, $p = 0.00$), which stands for how noticeable the changes were in the lighting condition of the environment. There were only slight movements in the lighting condition (shadows moving due to the movement of the avatars that are also watching the presentation), which may have been more noticed due to the abovementioned reason.

Even though there were no statistically significant results that support H1, there are some indicators that can fuel future research, with a larger sample, as Q12 and Q13 have moderate effect sizes of 0.47 and 0.42, respectively. Q1 and Q4 have moderately large effect sizes of 0.58 and 0.50 but it would be worth to have larger sample sizes to confirm if their p-values stay with values that accept the null hypothesis or if lean more towards the significance threshold. Q17 shows some effect that we did not expect and that would need further research to accurately explain its meaning.

B. RESULTS OF H2

The statistical results of ITQ of C2 are already analyzed on the previous section, please refer to it for a complete description. We include them on Table 4 for an easier comparison with values from C3. Unlike the users from other conditions, users from C3 report no values below or equal to 3 and only Q11 and Q13 above or equal to 5. Following the trend of the other groups, this one also reports it concentrates well on

TABLE 5. **Results from PQ and tailored questions for Condition II and Condition III. Condition III users report a high value of noticing changes on the lighting condition, which gives assurance when relating this variable to variations on dependent variables.**

	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12	Q13	Q14	Q15	Q16	Q17
Condition II																	
M	4,71	4,43	4,29	4,57	5,79	5,50	4,23	5,21	5,00	6,07	3,23	1,79	1,86	4,00	5,64	4,93	2,86
SD	1,68	1,65	1,90	1,34	1,19	1,34	1,42	1,05	1,80	1,49	1,19	1,31	1,03	1,75	1,01	1,38	1,70
Condition III																	
M	4,57	4,21	4,86	4,71	5,93	5,57	5,08	5,29	4,86	5,77	3,36	2,50	2,29	4,29	5,21	5,64	6,36
SD	1,02	1,19	1,35	1,27	0,92	1,16	1,77	1,27	1,10	0,89	1,55	1,29	1,33	1,27	1,37	0,93	0,84

enjoyable activities and loses all track of time in doing so. However, it does not possess any other traits that stand out.

Figure 8 displays the comparison between C2 and C3 ITQ results. Q1 ($M = -0.93$, $p = 0.02$) and Q4 ($M = -1.07$, $p = 0.03$) reveal that group from C3 feel significantly less involved in TV dramas or movies, significantly less identified with characters of plots, and have a harder time on concentrating on enjoyable activities as Q11 ($M = -0.57$, $p = 0.04$) reports. Despite this, its absolute value is still above 5, as reported on the previous paragraph. In addition, C3 dream less realistic dreams as Q9 ($M = -0.86$, $p = 0.10$) reveals. The group from C3 appears to have less tendency to be easily immersed on activities, so we could expect that to be reflected on the results. The overall mean difference between both groups is 3.07.

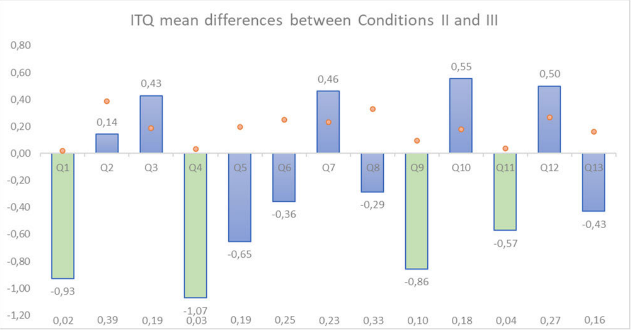


FIGURE 8. **Mean differences and respective p-values from ITQ from C2 and C3. C3 users significantly report they have a harder time concentrating on enjoyable activities and do not involve as much on TV dramas and identify less with characters on plots.**

Table 5 displays the PQ results from C2 and C3 that test H3. The analysis of the absolute results yielded by the group of C2 are already reported on the previous sections, please refer to it. C3 yielded results below or equal to 3 on Q12 and Q13, and results above or equal to 5 on Q5, Q6, Q7, Q8, Q10, Q15, Q16 and Q17. There are many results from the PQ that are above 5, which is a good indicator that the overall sense of presence of users is good. The high values reported of noticing changes in the lighting condition and its relevance to keep attention on the presentation also assures us that the lighting adaptations were not missed and any difference between these two conditions is due to this variable. Also, as expected, values related to confusion evaluation are still low.

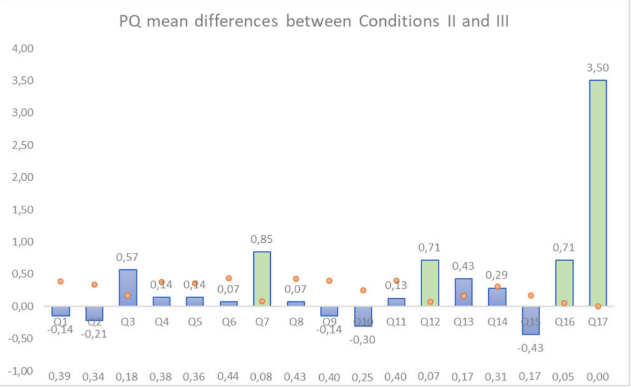


FIGURE 9. **Mean differences and respective p-values from PQ and tailored questions from C2 and C3. Users reported a higher score on Q17, which gives assurance that they did not miss the lighting adaptations. Q12, relating to confusion induced, was reported as higher, maybe due to a conflict between information channels.**

Figure 9 shows the mean difference between C2 and C3 PQs and the other questions tailored to this experiment. There is a value immediately draws our attention and that is the one from Q17 ($M = 3.50$, $p < 0.01$). It relates to the noticeability of the lighting condition adaptations, ensuring that users were aware of this system. In the PQ results there are no significant differences or indicators that users felt a higher sense of presence, except for Q7 ($M = 0.85$, $p = 0.08$) that relates to how well the user could localize sounds. This is unexpected, since the sound is the same across all conditions, except for the slight differences between the original script and the rewritten one, where the spoken text is different. However, this only varies between C2C1 and C3C2, and not between C1C2 and C2C3. Other unexpected result is from Q12 ($M = 0.71$, $p = 0.07$) where users report that the confusion induced by avatar's performance increased. We were expecting that this value would not reveal any differences or, if it did, it would be reported with lower values. However, this can be explained resorting to a previous explanation related with the visual and auditory channels. C3's visual environment was more dynamic when compared to C2's due to the lighting adaptations, which may require more mental resources from the user allocated to vision, and distracts the user from what is being transmitted through the auditory channel (in this case, the presentation script given by the avatar). Several authors provide evidence that attention is shared between auditory and visual channels [44]–[47]. In fact, inattentional deafness phenomenon is more frequent when i) visual task load increases [48], [49]; ii) the user is engaged in visual tasks of high perceptual load [50]; iii) the visual information conflict with the auditory channel [51]. Curiously, they also report in Q16 ($M = 0.71$, $p = 0.05$) that the lighting condition was more important to keep their attention on the presentation than users from C2, which aligns with H2. Q15 ($M = -0.53$, $p = 0.11$) shows that users felt more discomfort when exposed to the lighting adaptations, maybe because of this divergence between the auditory and visual channels. While they were trying to focus on something that was being said, they were being challenged by something they were being shown (the lighting changes). Finally, we were expecting some differences on PQ questions, and the only one standing out was Q7, which was already discussed. This lack of differences may be due to the C3 group of users having less tendency for immersion, which would explain this lack of differences. However, we also must consider the chance of this lighting adaptation system not contributing significantly for the user sense of presence.

Results suggest that a conflict between information channels is undesired and the channel that is being used to absorb information should be the one stimulated by an automatic environmental adaptation system. This does not support H2 but, nevertheless, it suggests that adapting environmental features are a valid mean to impact users and affect their cognitive performance. Q17 has a large effect size of 2.61. Q7, Q13, and Q15 have moderate effect sizes of 0.53, 0.36, and 0.36, respectively, whereas Q12 and Q16 have moderately large effect sizes of 0.55 and 0.61, respectively. As in the previous section, also in the testing of this hypothesis a larger sample may reveal clearer results.

Our work is novel in its approach to emotion-driven virtual environments through ubiquitous devices. To the best of our knowledge the presented Cabada *et al.* [29] and AutoTutor [12] works are the ones that most relate with our approach in what concerns the use of ubiquitous devices to detect and respond to users' emotional states. However, both works recognize emotions only based on facial images (Cabada *et al.* work) or only based on typed or spoken text analysis(AutoTutor work). In this aspect our approach is more complete by considering both text and video in different phases of the analysis. None of these works address 3D virtual environments nor the impact of changes in the user interface like we do (i.e. less confusing script and lighting condition changes). Regarding these latter aspects the most related work is the one of Yan *et al.* [40] that also deals with 3D virtual environments and the impact of changes in it using a Brain-Computer Interface (not considered a ubiquitous device). Nevertheless, in their work there was no consideration to confusion state or user's head pose while assessing users' engagement.

VI. CONCLUSION AND FUTURE WORK

This work produced a general 3D CVE framework that provides emotion and speech recognition, supported by a server-client architecture that concentrates heavy computation on the server, thus allowing remote users to take advantage of these features. A presentation scenario was built upon this CVE to carry out the study and test two hypotheses. There were three test conditions: a control condition, a second condition where we simplified the presentation text script with the help of a confusion prediction model, and a third condition where the virtual environment automatically adapted its lighting condition based on user engagement.

The two designed hypotheses approached the problem of distance learning and how emotion can be a catalyst for the sense of presence in Virtual Environments. It shows indications that supported our first hypothesis that a presentation script can be enhanced to yield a higher sense of presence by rewriting it based on its lexical and syntactic features. Furthermore, we also explored how lighting affects the sense of presence of users, and results yielded indications that do not support our second hypothesis that automatic lighting adaptation based on the user's head pose increases the sense of presence. However, this finding leaves interesting questions open to debate and for further future work. Information channel disparity may be at the root of these indications and we ought to carry out a study where automatic adaptations are conveyed through the auditory channel instead of the visual.

The work that was developed leaves some open avenues to be extended.

Emotion-driven natural interaction: Enriching interaction based on the user's emotions, in our opinion, has huge potential, since emotion is something that drives everything we do and every decision we make. One way emotion could fuel these scenarios through ubiquitous devices is through the mapping of emotion on the avatar's bodies. Spengler *et al.* [42] show evidence that high oxytocin levels are increased by synchronous social interactions which, in turn, play an important role on fostering prosocial behaviors. In this context, reciprocity is about providing and receiving nonverbal information and is at the core of a successful and engaging interaction [43]. By augmenting nonverbal synchronization between users with emotional body postures, the engagement level is increased as well as the acceptance of proposed the system.

Coordinated Actions: Depth sensors, eye trackers, or anything besides webcam, microphone, and mouse & keyboard is not available to the common user. However, we may see advances in the devices that are common on workplaces. Some laptop brands are already integrating them with embedded eye trackers, and we could expect that also depth sensors or biometrics start to be integrated in the future. This would open new opportunities for interaction and maybe, for instance, with depth sensors this framework could be expanded to support tangible actions.

APPENDIX A

A. DEMOGRAPHICS

- Age: __
- Gender: F_ M_
- Education:

High-school_ Bachelor_ Masters_ PhD_

B. IMMERSIVE TENDENCIES QUESTIONNAIRE

1. Do you easily become deeply involved in movies or TV dramas?

2. Do you ever become so involved in a TV program or book that people have problems getting your attention?

3. Do you ever become so involved in a movie that you are not aware of things happening around you?

4. How frequently do you find yourself closely identifying with the characters in a story line?

5. Do you ever become so involved in a video game that it is as if you are inside the game rather than moving a joystick and watching the screen?

6. How good are you at blocking out external distractions when you are involved in something?

7. When watching sports, do you ever become so involved in the game that you react as if you were one of the players?

8. Do you ever become so involved in a daydream that you are not aware of things happening around you

9. Do you ever have dreams that are so real that you feel disoriented when you awake?

10. When playing sports, do you become so involved in the game that you lose track of time?

11. How well do you concentrate on enjoyable activities?

12. How often do you play arcade or video games? (OFTEN should be taken to mean every day or every two days, on average.)

13. Do you ever become so involved in doing something that you lose all track of time?

C. PRESENCE QUESTIONNAIRE

1. How much did the visual aspects of the environment involve you?

2. How much did the auditory aspects of the environment involve you?

3. How compelling was your sense of objects moving through space?

4. How much did your experiences in the virtual environment seem consistent with your real-world experiences?

5. How completely were you able to actively survey or search the environment using vision?

6. How well could you identify sounds?

7. How well could you localize sounds?

8. How closely were you able to examine objects?

9. How involved were you in the virtual environment experience?

10. How quickly did you adjust to the virtual environment experience?

11. How much did the visual display quality interfere or distract you from performing assigned tasks or required activities?

1) TAILORED QUESTIONS

12. How confusing was the presentation given by the avatar?

13. How confusing was the subject presented by the avatar?

14. How engaging was the avatar that was presenting?

15. How comfortable was the lighting condition of the virtual environment?

16. How relevant was the lighting condition of the virtual environment to keep your attention on the presentation?

17. How noticeable were the changes in the lighting condition of the virtual environment?

APPENDIX B

A. ORIGINAL SCRIPT

1st slide: “Hello, my name is Amelia and I’ll be accompanying you during this usability test. Please note that this test isn’t about you, but rather about this system. It’s scheduled to take 20 minutes of your time, but you can leave anytime you want. It’s totally anonymous and only the answers you provide are kept. Furthermore, we will be collecting an approximation of your emotional context using video footage. However, we do not keep this footage. The frames of the video are analyzed in real-time and discarded automatically once emotions are recognized. By continuing, you acknowledge you agree with these terms.

The title of this presentation is ‘‘Shared Virtual Environments Promoting Interaction’’.

I’ll start by describing the scenario we adopted to conduct this research. Then I’ll give you some insight on basic and complex emotions and how they affect our daily social interactions and cognitive tasks.

I’ll proceed with the pipeline that was used to build the 3D virtual environment and talk about automatic emotion recognition. Finally, I’ll present you some examples of how emotion can be used to promote interaction on shared virtual environments.’’

2nd slide: ‘‘In such a global market, the awareness and marketing of a company is critical to success. People outside companies usually visit them to get to know the infrastructure but that’s not always possible. If someone is far away or cannot make the time to travel, companies will miss these people.

Virtual reality can answer this problem with virtual visits. But a visitor may want to talk with collaborators or have a brief presentation about the offices, technologies used or main areas of expertise. The problem is, people in companies usually work in environments similar to the ones depicted, leaving little room for interaction devices like depth sensors, head-mounted displays or haptic gloves. The only interaction devices at hand may be webcams, microphones, mouse, keyboard, and, eventually, an embedded eye tracker. This means that we can only consider these ubiquitous devices to enrich the interaction of the virtual environment.''

3rd slide: ''These are the 5 most scientifically agreed basic emotions. We experience them frequently in emotionally charged situations, often during our social interactions with others. The top image shows a set of basic emotions with facial expressions. These are called, arguably, the most universal emotions as identified by Paul Ekman. You may recall this author and these micro expressions from the famous TV series ''Lie to Me''. So, yes, that wasn't completely fictional and had this background. This researcher conducted experiments with several cultures across the world and identified these prototypical facial expressions. He coded micro expressions into what he calls ''Action Units''. But as the bottom image suggests, we don't show our emotions only through our faces. Our body expression also tells a lot. These two means of expression are considered part of nonverbal behavior.

And this is actually crucial to the information we convey to others. Body expression usually is captured with motion capture, which requires an apparatus that make it non-feasible for employees on offices. This limits the emotional reciprocity that we display in these environments and we think this is a major hindrance in the mainstream adoption of shared virtual environments on companies.''

4th slide: ''In addition to that set of basic emotions, there are others that are more complex and linked to other situations that are not as emotionally charged. States like confusion, engagement, frustration or boredom are seen during cognitive tasks like learning something new, or attending to a presentation…that is, when someone is passing information on to you.

Confusion may be the most interesting of this set as it is located on the boundary of engagement and disengagement.''

5th slide: ''This spectrum shows how we can evolve from an engaged state to a situation of boredom during a task that requires cognitive processing. Assuming you start out engaged with the activity, if a stimuli is applied, a cognitive disruption occurs. This event triggers a gain on arousal and eventually hits the first threshold, coded as t a on the scheme. Past t a you're in a state of confusion. As you are kept longer and longer in this state, you start to lean towards the second threshold, coded as t b on the scheme. Past this threshold you evolve into a frustrated state and, if not solved, to boredom, where you disengage from the activity and lose attention.

Confusion is particularly interesting because, as would've been thought, it's not always bad to be confused about something. When you are confused you are being cognitively challenged, which makes you try to overcome this confusion by truly understanding the subject that caused it, which is constructive. However, there's a fine line between constructive and non-constructive confusion. If you're not able to overcome it, and as time keeps on going, you're led into frustration, which is non-constructive and should be avoided.''

6th slide: ''So, for now we're done with the theoretical introduction about emotion. Let's proceed to how we built the 3D virtual environment. This simple scheme on the bottom left shows the software that was used. There was a collaborative effort between Revit, and 3DS max to do the heavy lifting of 3D modelling. On the top right we can see the interface of Revit. It provides tools to quickly build a 3D architectural model based on floor plans. 3DS max is used to correct some details and prepare the model to be exported to Unity with an FBX file.

Unity carries on with the game play mechanics, from avatar movement to the lighting condition and distributed logic. Within its interface, it provides access to lighting parameters, placement of 3D models and some texturing. It also provides a scripting API in C sharp that we used to create the mechanics of movement, the client-server distributed logic, and the automatic emotion recognition. The bottom right render shows the virtual environment on Unity.''

7th slide: ''Automatic emotion recognition is a key feature of proposed the system, so that we can take advantage of emotional context. We experimented with two tools that can estimate the head, and gaze orientation and facial expressions. On the top we have a screenshot of a sample application using OpenFace. This is an open-source academic tool that performs automatic facial landmark detection and derives action units from it. Each action unit codes a muscular movement on the face. For instance, on the top right we can see the values for inner, and outer brow raising, nose wrinkling, among others. The green/blue box centred on the head shows the estimation of the head pose orientation and the vectors coming from out of the eyes is the gaze estimation.

We can use these action units directly to map them on the avatar's face or we can use them to understand if the user is feeling any emotion. However, to produce emotional body posture animation we really have to understand if the user is experiencing any emotion. Coupled with head orientation estimation, we hope we can produce body expressions coherent with facial expressions. However, OpenFace has one shortcoming. It does not have a model that maps the action units to emotions. In opposition, the other tool we used, Affectiva, not only gives us the action units, but also detects emotions.''

8th slide: ''How do we apply these emotion concepts within virtual environments? Here we only present two examples of how this could be achieved. Once we have the emotional context of the user, we can adapt the behavior of AI avatars or even extend these emotions to your avatar's body expression. For instance, if you're sad or confused, the AI avatar may understand this and adopt a more cheerful posture or even ask you if everything's alright. Within this same environment there can be another people represented by their respective avatars, as you are with yours. In this case, if each user's avatar can display its user's emotion through facial and body expressions, the environment will be richer. Hopefully, this leads to higher reciprocity between users and even between users and AI avatars. Ultimately, this will raise user engagement.

The left scheme shows a scientifically published architecture for this affective system. The perception module contains the input devices that capture sound, video, and other modalities. These feed the cognitive module that interpret and model this data. This model is then used to trigger actions on the motor module that correspond to the examples given previously. One such example is on the image on the right where an avatar has body expressions based on its controller's emotions. The bottom example, the AutoTutor, is a special type of AI avatar. It's called an Intelligent Tutoring System and it's like a virtual teacher with intelligence. It understands wrong and correct answers from the student and leads the interaction with the goal of instructing him on a specific matter."

9th slide: "And it's like this that I finish this presentation. Now you will be asked to answer a questionnaire regarding this experience.

After you've finished the questionnaire, Tiago will be happy to answer any questions you may have.

I hope you enjoyed my presence and you can be totally honest in the next questionnaire. Remember that what is being evaluated is the system, not you!

Thank you!"

ACKNOWLEDGMENT

The authors would like to thank Rúben Pereira and Michael Harrison for the given support.

REFERENCES

[1] R. Bartle, *International Handbook of Internet Research*. Dordrecht, The Netherlands: Springer, Jun. 2010.

[2] P. K. Das and G. C. Deka, "History and evolution of GPU architecture," in *Emerging Research Surrounding Power Consumption and Performance Issues in Utility Computing*, G. C. Deka, Ed. Hershey, PA, USA: IGI-Global, 2015, pp. 109–135.

[3] R. Schroeder, *The Social Life of Avatars*, no. 1. London, U.K.: Springer, 2002.

[4] P.-Y. Tarng, K.-T. Chen, and P. Huang, "An analysis of WoW players' game hours," in *Proc. 7th ACM SIGCOMM Workshop Netw. Syst. Support Games (NetGames)*, 2008, p. 47.

[5] A. P. P. Negrón, N. E. R. Bernal, and G. L. López, "Nonverbal interaction contextualized in collaborative virtual environments," *J. Multimodal User Interfaces*, vol. 9, no. 3, pp. 253–260, Sep. 2015.

[6] P. Ekman, "What scientists who study emotion agree about," *Perspect. Psychol. Sci.*, vol. 11, no. 1, pp. 31–34, Jan. 2016.

[7] M. S. Dias, S. Eloy, M. Carreiro, and P. Proença, "Designing better spaces for people," in *Proc. 19th Int. Conf. Comput. Archit. Design Res. Asia (CAADRIA)*, 2014, pp. 739–748.

[8] M. S. Dias, S. Eloy, M. Carreiro, E. Vilar, S. Marques, A. Moural, P. Proênça, J. Cruz, J. d'Alpuim, N. Carvalho, A. S. Azevedo, and T. Pedro, "Space perception in virtual environments: On how biometric sensing in virtual environments may give architects," in *Proc. 32nd eCAADe Conf.*, vol. 2, 2014, pp. 271–280.

[9] R. J. Wurtman, "The effects of light on the human body," *Sci. Amer.*, vol. 233, no. 1, pp. 69–77, 1975.

[10] A. Arguel, L. Lockyer, O. V. Lipp, J. M. Lodge, and G. Kennedy, "Inside out," *J. Educ. Comput. Res.*, vol. 55, no. 4, pp. 526–551, 2017.

[11] S. D'Mello, B. Lehman, R. Pekrun, and A. Graesser, "Confusion can be beneficial for learning," *Learn. Instruct.*, vol. 29, pp. 153–170, Feb. 2014.

[12] S. D'mello and A. Graesser, "AutoTutor and affective autotutor," *ACM Trans. Interact. Intell. Syst.*, vol. 2, no. 4, pp. 1–39, Dec. 2012.

[13] S. K. D'Mello and A. Graesser, "Affect detection from human-computer dialogue with an intelligent tutoring system," in *Proc. Int. Workshop Intell. Virtual Agents*, 2006, pp. 54–67.

[14] S. Craig, A. Graesser, J. Sullins, and B. Gholson, "Affect and learning: An exploratory look into the role of affect in learning with AutoTutor," *J. Educ. Media*, vol. 29, no. 3, pp. 241–250, Oct. 2004.

[15] S. K. D'Mello, S. D. Craig, J. Sullins, and A. C. Graesser, "Predicting affective states through an emote-aloud procedure from AutoTutor's mixed-initiative dialogue," *Int. J. Artif. Intell. Educ.*, vol. 16, no. 1, pp. 3–28, 2006.

[16] S. K. D'Mello, B. Lehman, and N. Person, "Monitoring affect states during effortful problem solving activities," *Int. J. Artif. Intell. Educ.*, vol. 20, no. 4, pp. 361–389, 2010.

[17] M. S. Hussain, O. AlZoubi, R. A. Calvo, and S. K. D'Mello, "Affect detection from multichannel physiology during learning sessions with AutoTutor," in *Artificial Intelligence in Education* (Lecture Notes in Computer Science), G. Biswas, S. Bull, J. Kay, and A. Mitrovic, Eds. Berlin, Germany: Springer, 2011, pp. 131–138.

[18] J. F. Grafsgaard, J. B. Wiggins, K. E. Boyer, E. N. Wiebe, and J. C. Lester, "Automatically recognizing facial expression: Predicting engagement and frustration," in *Proc. 6th Int. Conf. Educ. Data Mining (EDM)*, 2013, pp. 43–50.

[19] J. F. Grafsgaard, J. B. Wiggins, K. E. Boyer, E. N. Wiebe, and J. C. Lester, "Automatically recognizing facial indicators of frustration: A learning-centric analysis," in *Proc. Humaine Assoc. Conf. Affect. Comput. Intell. Interact.*, Sep. 2013, pp. 159–165.

[20] J. F. Grafsgaard, J. B. Wiggins, A. K. Vail, K. E. Boyer, E. N. Wiebe, and J. C. Lester, "The additive value of multimodal features for predicting engagement, frustration, and learning during tutoring," in *Proc. 16th Int. Conf. Multimodal Interact. (ICMI)*, Nov. 2014, pp. 42–49.

[21] R. Bixler and S. D'Mello, "Detecting boredom and engagement during writing with keystroke analysis, task appraisals, and stable traits," in *Proc. Int. Conf. Intell. User Interfaces (IUI)*, 2013, p. 225.

[22] S. K. D'Mello, S. D. Craig, A. Witherspoon, B. McDaniel, and A. Graesser, "Automatic detection of learner's affect from conversational cues," *User Model. User-Adapt. Interact.*, vol. 18, nos. 1–2, pp. 45–80, Feb. 2008.

[23] S. D'Mello and A. Graesser, "Automatic detection of learner's affect from gross body language," *Appl. Artif. Intell.*, vol. 23, no. 2, pp. 123–150, Feb. 2009.

[24] S. D'Mello and R. A. Calvo, "Beyond the basic emotions," in *Proc. CHI Extended Abstr. Hum. Factors Comput. Syst. (CHI EA)*, 2013, p. 2287.

[25] S. D'Mello and A. Graesser, "Dynamics of affective states during complex learning," *Learn. Instruct.*, vol. 22, no. 2, pp. 145–157, Apr. 2012.

[26] B. Lehman, S. K. D'Mello, A. C. Strain, M. Gross, A. Dobbins, P. Wallace, K. Millis, and A. C. Graesser, "Inducing and tracking confusion with contradictions during critical thinking and scientific reasoning," in *Proc. Int. Conf. Artif. Intell. Educ.*, 2011, pp. 171–178.

[27] B. Lehman, S. D'Mello, A. Strain, C. Mills, M. Gross, A. Dobbins, P. Wallace, K. Millis, and A. Graesser, "Inducing and tracking confusion with contradictions during complex learning," *Int. J. Artif. Intell. Educ.*, vol. 22, nos. 1–2, pp. 85–105, 2013.

[28] B. Lehman, S. D'Mello, and A. Graesser, "Interventions to regulate confusion during learning," in *Intelligent Tutoring Systems*, vol. 7315, S. A. Cerri, W. J. Clancey, G. Papadourakis, and K. Panourgia, Eds. Berlin, Germany: Springer, 2012, pp. 576–578.

[29] R. Z. Cabada, M. L. B. Estrada, J. M. R. Félix, and G. A. Hernández, "A virtual environment for learning computer coding using gamification and emotion recognition," *Interact. Learn. Environ.*, vol. 28, pp. 1–16, Dec. 2018.

[30] A. Arguel, L. Lockyer, G. Kennedy, J. M. Lodge, and M. Pachman, "Seeking optimal confusion: A review on epistemic emotion management in interactive digital learning environments," *Interact. Learn. Environ.*, vol. 27, no. 2, pp. 200–210, Feb. 2019.

[31] N.-K. Park and C. A. Farr, "The effects of lighting on Consumers' emotions and behavioral intentions in a retail environment: A cross-cultural comparison," *J. Interior Des.*, vol. 33, no. 1, pp. 17–32, Sep. 2007.

[32] K. Quartier, J. Vanrie, and K. Van Cleempoel, "As real as it gets: What role does lighting have on consumer's perception of atmosphere, emotions and behaviour?" *J. Environ. Psychol.*, vol. 39, pp. 32–39, Sep. 2014.

[33] A. Kuijsters, J. Redi, B. De Ruyter, and I. Heynderickx, "Lighting to make you feel better: Improving the mood of elderly people with affective ambiences," *PLoS ONE*, vol. 10, no. 7, pp. 1–22, 2015.

[34] M. S. Mott, D. H. Robinson, A. Walden, J. Burnette, and A. S. Rutherford, "Illuminating the effects of dynamic lighting on student learning," *SAGE Open*, vol. 2, no. 2, pp. 1–9, 2012.

[35] I. Knez and C. Kers, "Effects of indoor lighting, gender, and age on mood and cognitive performance," *Environ. Behav.*, vol. 32, no. 6, pp. 817–831, Nov. 2000.

[36] B. K. Hawes, T. T. Brunyé, C. R. Mahoney, J. M. Sullivan, and C. D. Aall, "Effects of four workplace lighting technologies on perception, cognition and affective state," *Int. J. Ind. Ergonom.*, vol. 42, no. 1, pp. 122–128, Jan. 2012.

[37] D. M. McNair, M. Lorr, and L. F. Droppleman, "Manual for the profile of mood states," Educ. Ind. Test. Service, San Diego, CA, USA, Tech. Rep., 1971.

[38] E. Scott, A. Soria, and M. Campo, "Adaptive 3D virtual learning environments—A review of the literature," *IEEE Trans. Learn. Technol.*, vol. 10, no. 3, pp. 262–276, Sep. 2017.

[39] N. Vaughan, B. Gabrys, and V. N. Dubey, "An overview of self-adaptive technologies within virtual reality training," *Comput. Sci. Rev.*, vol. 22, pp. 65–87, Nov. 2016.

[40] S. Yan, G. Ding, H. Li, N. Sun, Y. Wu, Z. Guan, L. Zhang, and T. Huang, "Enhancing audience engagement in performing arts through an adaptive virtual environment with a brain-computer interface," in *Proc. 21st Int. Conf. Intell. User Interface (IUI)*, Mar. 2016, pp. 306–316.

[41] A. Heydarian, E. Pantazis, A. Wang, D. Gerber, and B. Becerik-Gerber, "Towards user centered building design: Identifying end-user lighting preferences via immersive virtual environments," *Autom. Construct.*, vol. 81, pp. 56–66, Sep. 2017.

[42] B. G. Witmer and M. J. Singer, "Measuring presence in virtual environments: A presence questionnaire," *Presence, Teleoperators Virtual Environ.*, vol. 7, no. 3, pp. 225–240, Jun. 1998.

[43] T. S. Pedro, J. L. Silva, and R. Pereira, "Predicting the confusion level of text excerpts with syntactic, lexical and N-gram features," in *Proc. 10th Int. Conf. Educ. New Learn. Technol.*, 2018, pp. 8417–8426.

[44] S. P. Banbury, W. J. Macken, S. Tremblay, and D. M. Jones, "Auditory distraction and short-term memory: Phenomena and practical implications," *Hum. Factors, J. Hum. Factors Ergonom. Soc.*, vol. 43, no. 1, pp. 12–29, Mar. 2001.

[45] M. Brand-D'Abrescia and N. Lavie, "Task coordination between and within sensory modalities: Effects on distraction," *Perception Psychophysics*, vol. 70, no. 3, pp. 508–515, Apr. 2008.

[46] V. Santangelo, M. O. Belardinelli, and C. Spence, "The suppression of reflexive visual and auditory orienting when attention is otherwise engaged," *J. Exp. Psychol., Hum. Perception Perform.*, vol. 33, no. 1, pp. 137–148, 2007.

[47] S. Sinnett, A. Costa, and S. Soto-Faraco, "Manipulating inattentional blindness within and across sensory modalities," *Quart. J. Exp. Psychol.*, vol. 59, no. 8, pp. 1425–1442, Aug. 2006.

[48] L. Giraudet, M.-E. Saint-Louis, and M. Causse, "Electrophysiological correlates of inattentional deafness: No hearing without listening," in *Proc. HFES Eur. Chapter Conf.*, Toulouse, France, Oct. 2012, pp. 89–99.

[49] A. F. Kramer, L. J. Trejo, and D. Humphrey, "Assessment of mental workload with task-irrelevant auditory probes," *Biol. Psychol.*, vol. 40, nos. 1–2, pp. 83–100, May 1995.

[50] J. S. P. Macdonald and N. Lavie, "Visual perceptual load induces inattentional deafness," *Attention, Perception, Psychophysics*, vol. 73, no. 6, pp. 1780–1789, Aug. 2011.

[51] S. Scannella, M. Causse, N. Chauveau, J. Pastor, and F. Dehais, "Effects of the audiovisual conflict on auditory early processes," *Int. J. Psychophysiol.*, vol. 89, no. 1, pp. 115–122, Jul. 2013.

[52] F. B. Spengler, D. Scheele, N. Marsh, C. Kofferath, A. Flach, S. Schwarz, B. Stoffel-Wagner, W. Maier, and R. Hurlemann, "Oxytocin facilitates reciprocity in social communication," *Social Cognit. Affect. Neurosci.*, vol. 12, no. 8, pp. 1325–1333, Aug. 2017.

[53] M. Büscher, J. O'Brien, T. Rodden, and J. Trevor, "'He's behind you': The experience of presence in shared virtual environments," in *Collaborative Virtual Environments*. London, U.K.: Springer, 2001, pp. 77–98.

[54] T. Baltrusaitis, P. Robinson, and L. P. Morency, "OpenFace: An open source facial behavior analysis toolkit," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, 2016.

[55] D. McDuff, A. Mahmoud, M. Mavadati, M. Amr, J. Turcot, and R. El Kaliouby, "AFFDEX SDK: A cross-platform real-time multi-face expression recognition toolkit," in *Proc. CHI Conf. Extended Abstr. Hum. Factors Comput. Syst. (CHI EA)*, 2016, pp. 3723–3726.

TIAGO M. SILVA PEDRO was born in Santarém, Portugal, in 1990. He received the B.Arch. degree and the M.S. degree in computer engineering from the Instituto Universitário de Lisboa (ISCTE-IUL), Lisbon, Portugal, in 2014 and 2018, respectively.

While pursuing his B.Arch. degree, he worked as a Research Intern with ISTAR-IUL, from 2014 to 2017, where he coauthored eight international conference papers and one book chapter exploring how architectural design impacts human behavior through virtual reality experiments with physiological and subjective evaluation. From 2017 to 2018, he worked as a trainee at Siemens. He is currently working as an IT Consultant with Celfocus, where he is developing chatbot solutions with speech interaction. He is currently applying to a Ph.D. scholarship in the fields of entrepreneurship, human–computer interaction, and immersive environments. His current research interests include human–computer interaction, distance learning, virtual environments, and applied cognitive neuroscience.

JOSÉ LUÍS SILVA received the M.S. degree in informatics and systems engineering from the University of Minho, Braga, Portugal, in 2007, and the Ph.D. degree in computer science from the Portuguese MAP-i Consortium (University of Minho, University of Aveiro and University of Porto), in 2012.

From 2012 to 2013, he held a postdoctoral position at the University of Toulouse, France, in collaboration with Airbus. From 2013 to 2016, he was an Invited Assistant Professor with the Exact Sciences and Engineering Department, Madeira University, Portugal. Since 2016, he has been Assistant Professor with the Department of Information Science and Technology, Instituto Universitário de Lisboa (ISCTE-IUL), Lisbon, Portugal. His work has been published at venues, such as IEEE Access, IJHCS, ENTCS, ACM EICS, and INTERACT. His main research interests include software engineering, human–computer interaction, ubiquitous computing, and virtual environments. He is member of the LARSyS Laboratory and the ISTAR-IUL Research Center.

Dr. Silva is member of the IFIP TC 13—Working Group 13.2. His awards and honors include the FCT Doctoral Degree Grant (Portuguese Government), the ISCTE-IUL Scientific Award, the Best Iberian Ph.D. Thesis in Systems and Information Technologies from AISTI, and the PhD Award from Fraunhofer Portugal Challenge.