



Department of Information Science and Technology

Virtual Environments Promoting Interaction

Tiago Miguel da Silva Pedro

A Dissertation presented in partial fulfillment of the Requirements for the Degree of
Master's in Computer Engineering

Supervisor:

José Luís Cardoso da Silva, PhD, Assistant Professor at
ISCTE-IUL

Co-supervisor:

Rúben Filipe Pereira, PhD, Assistant Professor at
ISCTE-IUL

September 2018

“Solitude matters, and for some people, it’s the air they breathe.”

“Everyone shines, given the right lighting.”

Susan Cain

Resumo

A Realidade Virtual (RV) tem sido alvo de investigação extensa na academia e tem vindo a entrar na indústria. Empresas comuns não têm acesso a esta tecnologia como uma ferramenta de colaboração porque estas soluções necessitam de dispositivos específicos que não estão disponíveis para o utilizador comum em escritório. Existem outras plataformas de colaboração baseadas em vídeo, voz e texto, mas a RV permite partilhar o mesmo espaço 3D. Neste espaço podem existir funcionalidades ou informação adicionais que no mundo real não seria possível, algo intrínseco à RV.

Esta dissertação produziu uma *framework* 3D que promove a comunicação não-verbal que tem um papel fundamental na interação humana e é principalmente baseada em emoção. Na academia é sabido que a confusão influencia os ganhos na aprendizagem quando gerida adequadamente. Desenhámos um estudo para avaliar como as características lexicais, sintáticas e n-gramas influenciam a confusão percebida. Construímos e testámos um modelo de aprendizagem automática que prevê o nível de confusão baseado nestas características, produzindo resultados não estatisticamente significativos que suportam esta hipótese. Este modelo foi usado para manipular o texto de uma apresentação e o *feedback* dos utilizadores demonstra uma tendência na diminuição do nível de confusão reportada no texto e aumento da sensação de presença. Outra contribuição vem das características intrínsecas de um ambiente 3D onde se podem executar ações que no mundo real não seriam possíveis. Desenhámos um sistema automático de iluminação adaptativa que reage ao *engagement* percebido do utilizador. Os resultados não suportam o que hipotetizámos mas não têm significância estatística, pelo que esta hipótese foi parcialmente rejeitada.

Três linhas de investigação podem provir desta dissertação. Primeiro, criar características mais complexas para treinar o modelo de aprendizagem, tais como árvores de sintaxe. Além disso, num *Intelligent Tutoring System* este modelo poderá ajustar o discurso do avatar em tempo real, alimentado por um detetor de confusão. As emoções básicas ajustam-se a um cenário social e podem enriquecê-lo. A emoção expressada facialmente pode estender este efeito ao corpo do avatar para alimentar o sincronismo social e aumentar a sensação de presença. Finalmente, baseámo-nos em dispositivos ubíquos, mas com a rápida evolução da tecnologia, podemos considerar que novos dispositivos irão estar presentes em escritórios. Isto abre possibilidades para novas modalidades.

Palavras-chave: Realidade virtual, 3D, confusão, comunicação não-verbal, sensação de presença

Abstract

Virtual reality (VR) has been widely researched in the academic environment and is now breaking into the industry. Regular companies do not have access to this technology as a collaboration tool because these solutions usually require specific devices that are not at hand of the common user in offices. There are other collaboration platforms based on video, speech and text, but VR allows users to share the same 3D space. In this 3D space there can be added functionalities or information that in a real-world environment would not be possible, something intrinsic to VR.

This dissertation has produced a 3D framework that promotes nonverbal communication. It plays a fundamental role on human interaction and is mostly based on emotion. In the academia, confusion is known to influence learning gains if it is properly managed. We designed a study to evaluate how lexical, syntactic and n-gram features influence perceived confusion and found results (not statistically significant) that point that it is possible to build a machine learning model that can predict the level of confusion based on these features. This model was used to manipulate the script of a given presentation, and user feedback shows a trend that by manipulating these features and theoretically lowering the level of confusion on text not only drops the reported confusion, as it also increases reported sense of presence. Another contribution of this dissertation comes from the intrinsic features of a 3D environment where one can carry actions that in a real world are not possible. We designed an automatic adaption lighting system that reacts to the perceived user's engagement. This hypothesis was partially refused as the results go against what we hypothesized but do not have statistical significance.

Three lines of research may stem from this dissertation. First, there can be more complex features to train the machine learning model such as syntax trees. Also, on an Intelligent Tutoring System this could adjust the avatar's speech in real-time if fed by a real-time confusion detector. When going for a social scenario, the set of basic emotions is well-adjusted and can enrich them. Facial emotion recognition can extend this effect to the avatar's body to fuel this synchronization and increase the sense of presence. Finally, we based this dissertation on the premise of using ubiquitous devices, but with the rapid evolution of technology we should consider that new devices will be present on offices. This opens new possibilities for other modalities.

Keywords: Virtual reality, 3D, confusion, nonverbal communication, sense of presence

Agradecimentos

Começo por agradecer às pessoas mais importantes na minha vida, a minha família e amigos. Nomeadamente aos meus pais, ao meu irmão, à minha avó, à minha namorada e à Dora por todo o apoio demonstrado durante todo este percurso, que não se resume apenas a este ano. A eles devo todo o meu percurso e sem eles seguramente não estaria onde estou. Obrigado pelo apoio em todas as mudanças de direção e em construir o meu caminho.

De seguida, quero expressar o meu profundo agradecimento ao meu orientador José Luís Silva pela orientação fantástica que me ofereceu, dando-me liberdade criativa, oferecendo sempre bons conselhos, por ter estendido a sua visão aos temas a que me propus abordar e pelos bons e divertidos momentos de discussão. Agradeço também ao meu coorientador Rúben Pereira pela total confiança que sempre demonstrou em mim, por esta oportunidade e pelo conhecimento transmitido. Endereço também um especial agradecimento ao Miguel Rangel pela oportunidade, pelo espírito e confiança transmitidos, e por uma experiência marcante.

Como esta dissertação foi o culminar de um longo percurso, que foi tudo menos linear, quero deixar os meus agradecimentos à Sara Eloy, pois foi com ela que tudo começou, e ao Miguel Dias pela confiança e pelos constantes desafios que me apresentou. Por fim, deixo um sentido agradecimento ao meu amigo Pedro Faria Lopes que, sem qualquer obrigação e por pura dedicação, vontade e alegria em ensinar, me guiou nestes últimos anos. Espero ter contribuído para a sua vontade em aproximar a Engenharia e a Arte.

Table of Contents

Resumo	i
Abstract	iii
Agradecimientos	v
List of Tables	ix
List of Figures	xi
List of Acronyms	xiii
Chapter I – Introduction.....	1
1.1 3D Virtual Environments	1
1.2 Motivation	3
1.3 Theoretical Framing	5
1.3.1. CVE scenarios.....	5
1.3.2. Emotion.....	7
1.4 Thesis Hypotheses.....	9
1.5 Objectives.....	10
1.6 Research Methodology.....	10
1.7 Dissertation Overview.....	12
Chapter II – Literature Review	13
2.1 Background	13
2.1.1. Nonverbal Behavior	13
2.1.2. Multimodal Emotion Recognition	14
2.2 Confusion Detection on Learning Environments.....	15
2.3 Adaptive Lighting Conditions.....	20
2.4 Summing Up	23
Chapter III – Technological Review.....	25
3.1 Automatic Facial & Emotion Recognition.....	25
3.2 Serious Game Engines Review	27
3.3 Summing Up	32
Chapter IV – 3D Virtual Environment.....	33
4.1 3D Modeling Pipeline	33
4.2 Distributed System Architecture	37
4.2.1. User Guide	38
4.2.2. Unity Networking	40
4.2.3. Component Description	41

4.2.4. AU Detection and Facial Emotion Recognition	43
4.3 Summing Up	45
Chapter V – Confusion Prediction	47
5.1 Methodology	47
5.2 Dataset description	49
5.2.1. Sample description.....	49
5.2.2. Text complexity features.....	51
5.3 Evaluation and Discussion	52
5.4 Summing Up	55
Chapter VI – Case Study.....	57
6.1 Methodology	57
6.1.1. User Description	58
6.1.2. Presentation Script	59
6.1.3. Virtual Environment	61
6.1.4. Experimental Settings and Procedure	63
6.2 Evaluation and Discussion	63
6.2.1. Results of H2.....	63
6.2.2. Results of H3.....	67
6.3 Summing Up	71
Chapter VII – Conclusion and Future Work.....	73
7.1 Academic and Industrial Contributions.....	73
7.2 Answers to Hypotheses and Discussion.....	73
7.3 Future Work	74
References.....	77
Appendix A - Questionnaires.....	85
1. Immersive Tendencies Questionnaire.....	85
2. Presence Questionnaire + Tailored questions	86
Appendix B – Presentation Scripts	89
1. Original Script.....	89
2. Rewritten Script	93
Appendix C – Emotional Body Posture Survey	94

List of Tables

Table 1. Different scenarios of CVEs	6
Table 2. Facial Action Code System (FACS).....	26
Table 3. Features for SGE comparative analysis framework.	28
Table 4. SGE used for serious games development.....	29
Table 5. Extra features considered for the project.	29
Table 6. Updated table of comparison for the project	30
Table 7. Pros and cons between UE4 and Unity.....	32
Table 8. Features selected by the recursive feature elimination algorithm.....	52
Table 9. Classification report of the original script	60
Table 10. Analysis from the three text excerpts that were chosen to be rewritten	61
Table 11. ITQ results for Condition I and Condition II.....	64
Table 12. Results from PQ and tailored questions for Condition I and Condition II	65
Table 13. ITQ results for Condition II and Condition III	67
Table 14. Results from PQ and tailored questions for Condition II and Condition III.....	69

List of Figures

Figure 1. a) The five most agreed emotions as to be the basic ones. b) Plutchik's wheel of emotions.	8
Figure 2. Multimodal AER framework.....	15
Figure 3. At the top is a real live performance of the opera Siegfried. The three lower displays are the 3 different conditions that users experimented.	22
Figure 4. Sample frame taken from the interaction interface between a learner and AutoTutor.	16
Figure 5. Model of learning-centered affective states	18
Figure 6. Adaptation of the learning model	20
Figure 7. Modeling pipeline.....	34
Figure 8. Revit's floor plan interface	34
Figure 9. Revit's 3D view	35
Figure 10. The mesh produced by Revit.....	36
Figure 11. Collision mesh computed by Unity	37
Figure 12. Server-client high-level architecture of the system	38
Figure 13. Connection menu that is presented to the user	39
Figure 15. Windows standalone interface on the left and Android interface on the right.....	40
Figure 16. Unity's High-level API broadcasting system	41
Figure 17. Class diagram	42
Figure 17. OpenFace and Affectiva's head pose estimation	44
Figure 18. Landmark detection and head orientation estimation.....	45
Figure 19. Interface to rate text excerpts	48
Figure 20. Chart representing the class distribution of the dataset.....	50
Figure 21. Each n-gram is represented on the x-axis with values for each class of confusion.....	53
Figure 22. Chart with train, validation and test f-scores for each estimator.....	54
Figure 23. Test stage results.....	54
Figure 24. Mean age of the group of users per condition	58
Figure 25. Level of education distribution for each condition.....	59
Figure 26. The 3D virtual environment used to test H2 and H3.....	62
Figure 27. Mean differences and respective p-values from ITQ from Condition I and Condition II	65

Figure 28. Mean differences and respective p-values from PQ and tailored questions from Condition I and Condition II..... 66

Figure 29. Mean differences and respective p-values from ITQ from Condition II and Condition III..... 68

Figure 30. Mean differences and respective p-values from PQ and tailored questions from Condition II and Condition III 70

List of Acronyms

- AER – Automatic Emotion Recognition
- AU – Action Unit
- CVE – Collaborative Virtual Environment
- CW – Collaborative Work
- DiD – Didactic
- FACS – Facial Action Code System
- FR – Facial Recognition
- ITS – Intelligent Tutoring System
- PrE – Presentation
- SoC – Socialization
- TrA – Training
- VE – Virtual Environment
- VR – Virtual Reality

Chapter I – Introduction

This research work comes in the context of applying its outcome to an industrial environment of a specific Company (whose name cannot be disclosed) but the theme was not completely defined. The Company requested a 3D distributed and collaborative environment of its offices where a virtual visitor could join and visit.

The theme presented lacked specifications, which allowed us to further research the state of the art to identify potential goals and hypotheses. In the academia this is a widely researched area, but its application on real-life industrial environments is not widely sought from an academic point of view. Thus, this context presented challenges that we wanted to embrace. Framing our work on the state of the art was bound to suffer from many deviations and setbacks as we assessed the scientific relevance and feasibility in due time of this dissertation as ideas came up. We will start out this chapter by relating the Company’s request into the Virtual Reality (VR) and Human-Computer Interaction (HCI) scientific fields. From there, the problem is dissected and analyzed, and the final hypotheses and goals of this dissertation are formalized.

As an early overview, we will briefly outline the contributions of this dissertation and the scenarios on which it was built upon. A multi-user shared 3D environment was identified as being the scenario the Company was requesting, divided into two more specific scenarios: 1) a multi-user distributed scenario where users socialize, and 2) a presentation scenario where users are given information through a live or offline presenter represented by an avatar. The distributed shared environment was developed as a “host” scenario to the two more specific scenarios. However, the scientific contributions of this dissertation lie within the presentation scenario as a strategic choice, but the distributed environment remains as a general 3D framework where future features can be developed in this industrial context.

1.1 3D Virtual Environments

As a large Company that employs thousands of professionals across several countries and cultures, work force diversification is one of the Company’s banners. Tied with this culture there is a need to enable people (employees or not) that are far away and unable to physically visit the headquarters, to get to know the offices. In this context, the Company intends to offer the visitor the ability

to communicate with the employees that are working *in-situ* at the office. In addition, the visitor should be able to be given presentations or demonstrations, taking advantage of techniques that would not be possible in a real-life setup. VR, based on a strong concept of HCI, is seen as the solution.

In an increasingly globalized world, we are witnessing the emergence of new inventive ways of digital communication on a daily basis. One of these is VR, which provides “super-powers” for those who use it. This concept allows people to live something that they otherwise could not, due to physical, time, or other constraints. Besides bridging this gap, these “super-powers” go even further by enabling people to experience new ways of living something, be it by flying (without an airplane) or being in the interstellar space, many light-years from Earth. The sense of presence in these environments is what makes them believable and its improvement is the main goal of our research. This main requirement of VR is, citing Schroeder, “(…) about “being there”: presence is therefore partly to do with the technology, and partly to do with the participants’ state of mind.” (R. Schroeder, 2002).

Even though VR exists for a long time (Bartle, 2010), the advent of Graphics Processing Units (GPUs) in the late 90s (Das & Deka, 2015) (and the increase of the general hardware power) levered its feasibility on consumer applications. In spite of this, there are still specific scenarios that can benefit from this approach (R. Schroeder, 2002). Some of these scenarios are contained on the field of Collaborative Virtual Environments (CVE) that applies VEs to situations where multiple persons co-exist through their virtual representations.

One of the main arguments against the validity of VEs is that they are not real, or they do not recreate real life. However, Jakobsson and Hudson-Smith (R. Schroeder, 2002) point out that this is not necessarily true because they show that users show commitment to the VE, which in turn socially forces them to behave in a way that their presence on the VE keeps being accepted by others. In fact, this sense of commitment and co-presence can be so strong that people can bond and experience emotions as shown by Slater and Steed (R. Schroeder, 2002).

At the core of a CVE is the user’s representation on the VE, the said avatar (Peña Pérez Negrón, Rangel Bernal, & Lara López, 2015). The avatar is especially important in the context of CVEs, in opposition to a single-user VE. An avatar lets its controller express behavior (in all its complexity) so that every other user is aware of each other’s context (Nguyen & Duval, 2015; Peña Pérez

Negrón et al., 2015; R. Schroeder, 2002). Even though the avatar is the user's representation, in the general context it does not necessarily need to have a human form. There are countless avatars that do not assume a human form, especially on the videogame industry, where the audience is more open to fantasy and detachment from reality due to its ludic status. Even in the context of serious CVEs it is possible to find avatars represented by abstract or non-humanoid models (Nguyen & Duval, 2015) which does not necessarily impact the fidelity of the system or the performance of the task at hand as some studies show (Schuemie, van der Straaten, Krijn, & van der Mast, 2001). However, the humanoid shape provides unique possibilities to bring the interaction level to new heights when it comes to reciprocity between humans.

A humanoid avatar can be customized to augment the sense of embodiment and is able to display information in a way the human naturally understands, like facial or body expressions and speech.

1.2 Motivation

In this digital world, companies often spread their installations across several countries or even continents. This leads to an increased unawareness by the employees of what their own company's services or products are. Virtual tours constituted of social and presentation scenarios aim at providing an engaging experience by fostering interpersonal communication and awareness. They both benefit from emotional context, but from different perspectives. In the social scenario the set of emotions can be comprised by the basic, universal ones (Ekman, 2016), but in the presentation scenario the set is more complex and include emotions such as engagement, confusion, frustration or boredom (Arguel, Lockyer, Lipp, Lodge, & Kennedy, 2017). Visitors and *in-situ* employees are typically connected through simple hardware setups (laptop, mobile devices) and interaction devices (keyboard & mouse, touchscreens, microphone and webcam), which considerably limit the range of interaction modalities. However, we take this barrier as a challenge to achieve an immersive virtual sense of presence through ubiquitous devices, striving to spread this communication media to the ordinary user.

There is already a set of successful CVEs however, most of them are based on text, speech and mouse & keyboard input with little to no other interaction modalities (R. Schroeder, 2002), especially when it comes to nonverbal communication. They are used as basis to videogames, more specifically in the MMORPG market (Tarng, Chen, & Huang, 2008). Players are usually

represented by an avatar which graphically displays this representation to other players. In terms of physical dynamic this avatar is usually limited to the pre-scripted animations of the game engine.

We believe that a drawback hindering the proliferation of CVEs is the lack of engagement these systems provide, easily breaking users out of immersion. We think that one factor that contributes to this lack of engagement is the low level of natural interaction that generates nonverbal reciprocity (and consequent fail of conveying emotion) between users, something that is critical to our daily human-to-human interactions (Peña Pérez Negrón et al., 2015).

One great and specific characteristic the abovementioned videogames have is the ambient of their virtual worlds. Its design (not necessarily in terms of graphics quality) (R. Schroeder, 2002) is often appealing to users who let themselves immerse in this fantasy. As they assume their environment as fantastist and otherworldly, they are aligning their users' expectations accordingly and often unconsciously. In turn, this alleviates the pressure users put on deviations (sometimes unwanted due to bugs) from a realistic environment because they are already in a place where real life rules do not necessarily apply, as long as the concept is kept congruent with the environment. This makes them more tolerable to the deviations taken from a real-life environment. In contrast, non-ludic CVEs usually try to replicate real-life environment which set higher expectations and users become more critical to similar deviations. We consider this a reason that hinders engagement because there is the risk of the user breaking out of immersion due to unmet expectations. However, one can take advantage of the power one has, to exploit environment conditions to achieve greater immersion and try to (re)engage the user when he starts do disengage. The environmental conditions that surround us play a role in the way we feel and accept that environment, as the architectural landscape of an environment can produce alterations on physiological measures (Dias, Eloy, Carreiro, Vilar, et al., 2014; Dias, Eloy, Carreiro, Proença, et al., 2014). Even though this does not represent a problem by itself, it presents as an opportunity to bring positive effects upon the user.

In a presentation scenario the set of emotions is comprised of the cognitive states of engagement, confusion, frustration, and boredom. Detecting when the user is in any of these states is valuable for acting accordingly. Confusion is particularly interesting as it is this state that is triggered by stimuli that leads to a cognitive disequilibrium (Arguel et al., 2017) and as D'Mello et al.

(D’Mello, Lehman, Pekrun, & Graesser, 2014) state, the confusion state and its resolution can increment the learning gain. Therefore, detecting its source enables a better adaptation of the system which will facilitate the overcoming of this state and increment learning gains.

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

1.3 Theoretical Framing

1.3.1. CVE scenarios

The collaboration and sharing of the same virtual environment by multiple users is being target of attention for many years now, both at the corporate and academic domains (Benford, Greenhalgh, Rodden, & Pycocock, 2001; R. Schroeder, 2002). These CVEs, along with social networks and other digital platforms of the likes, open a new paradigm that can be used in several fields (Nguyen & Duval, 2015).

We categorize CVEs based on some characteristics under 4 different scenario templates: Socialization (SoC), Collaborative Work (CW), Training (TrA) and Presentation (PrE) [Table 1]. These characteristics are: (1) the number of agents (flowing from emitter to receiver agents) – Roles, (2) whether the interaction can benefit from emotional context – Emotional environment, (3) whether the interaction is taken wrapped in a formal – Production environment, (4) whether the interaction needs coordination between tangible actions – Coordinated actions and (5) whether the template meets the Company’s scenario. These are seen only as templates that share high-level characteristics as they can fork into more specific scenarios that may differ at a lower-level (e.g. PrE can fork into a webinar, a class or simply a presentation at a company). These templates were designed disregarding the scenario identified by the Company (as described on “Chapter I - Introduction”) and the ‘Identified objectives’ field was only included later to avoid bias, rather than being an integrating part of the design process.

The SoC template is mainly characterized by the high demands of emotional cues needed to provide an engaging experience. The interaction on these scenarios typically happens between several actors whose roles switch frequently between an emitter (sending information) and a receiver (receiving information) state, but generally without the need for tangible actions. Due to the

inherent characteristics of these scenarios, emotional cues play an important role (Nguyen & Duval, 2015).

Table 1. Different scenarios of CVEs. Their different characteristics reflect on their requirements.

Characteristics Scenario	Roles	Emotional environment	Production environment	Coordinated actions	Identified objectives
Socialization	*..*	✓	✗	✗	✓
Collaborative Work	*..* 1..*	✗	✓	✓	✗
Training	1..*	✓	✓	✓	✗
Presentation	1..*	✓	✓	✗	✓

CW comprises the largest number of combinations between emitters and receivers. This case does not feature such a strong distinction between emitters and receivers, as all of them are being playing both roles all the time. The VE awareness is especially important here because it allows actors to play both roles at the same time in an efficient fashion. As a set of scenarios that are focused on productivity they are not so dependent on emotional cues for they are purely professional (Nguyen & Duval, 2015). One of the main researched topics in this field is how to efficiently coordinate tangible actions. It usually requires complex apparatus which is constrained/constrains the real physical environment of the user.

In opposition to the CW scenarios, the TrA template features a deeper interpersonal approach between one emitter and one/many receivers, but still not as dependent on emotional cues as SoC. These cues may help the emitter recognize if the receivers are engaged and following the training but are not essential to get the job done in a practical way, as it is on the SoC scenarios where emotional cues play a big part driving the interaction. As it is usually taken under the learning hat, it still focuses on productivity (or more of a learning rate, in this case) that is wanted to be kept on a high level, but not as rigid as in a CW scenario.

The PrE template is similar to TrA, differing only in the Coordinated Actions aspect. However, on the one hand Tra is a practical approach of learning, on the other, PrE is a theoretical one and here resides the difference. PrE does not need tangible actions, thus changing the interaction requirements.

Wrapping up and looking at the last column ('Identified objectives'), one can conclude that the SoC and PrE templates are the ones that best address the scenario identified by the company. The SoC template refers to a multi-user distributed environment where users socialize, whereas PrE

refers to a presentation scenario. First, the Soc allows the visitor to interact with the *in-situ* employees as well as letting them interact between each other; second, the PrE can be specified as the requirement of a visitor being given information. Neither includes Coordinated Actions, which meets our scenario of using only ubiquitous devices for interaction, at grasp of the ordinary user.

Finally, the identification of these templates and our positioning on them is what drives the logical stream of our State of the Art and contribution fields.

1.3.2. Emotion

It is not easy to define what an emotion is. Surely everyone knows when he/she is experiencing an emotion (and which one) however, nowadays is scientifically hard to define what exactly triggers it. Even so, there has been a significant body of research on the psychology of emotions since the 19th century. Darwin (Darwin, 1872) states that emotions are discrete and directly relate to specific parts of the body. This view has been followed by several researchers like Ekman and Friesen (Ekman & Friesen, 1969) or Izard (Izard, 1971). In opposition, Wundt (Wundt, 1896) suggests a dimensional view of emotions, varying along positive-negative valence and low-high intensity. Following Wundt's approach, Plutchik proposed his wheel of emotions in 1980 (Plutchik, 1980) which consists on a model that classifies emotions along the two said dimensions.

Research in both computer science and psychology has been highly focused in the discrete model but in recent years the multidimensional model advocated by Wundt has been getting more attention in both fields (Ekman, 2016; Gunes & Schuller, 2013). Regardless of the emotional model, Ekman's survey (Ekman, 2016) concluded that 88% of the surveyed population agreed about the universality of emotions. There was also agreement about five emotions, as depicted in Figure 1, to be the most basic:

- Anger – 91%,
- Fear – 90%,
- Disgust – 86%,
- Sadness – 80% and
- Happiness (Enjoyment) – 76%.

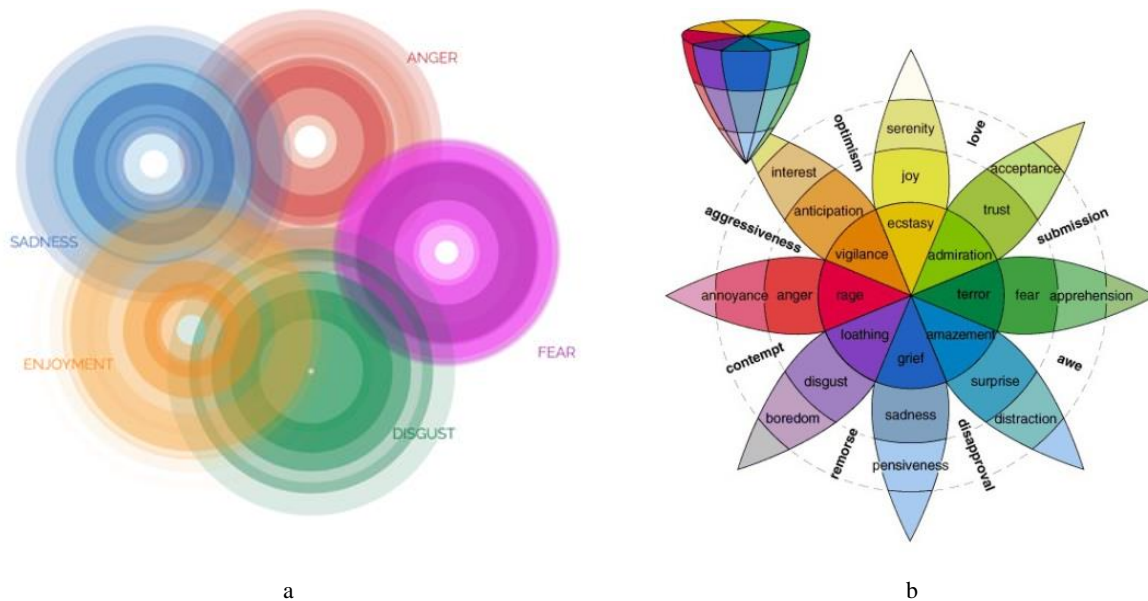


Figure 1. The five most agreed emotions as to be the basic ones and Plutchik's wheel of emotions.

Additionally, more recent approaches bring the concept of emotional system, rather than discrete or dimensional models. LeDoux (LeDoux, 1998) and Panksepp (Panksepp, 1998) state that similar neurological circuits and structures are activated for different conditions or situations. Like the previous described approaches, this approach also suggests that there are basic emotional systems that are “hardwired” on us. Under certain conditions the autonomous neural limbic system triggers the following emotional systems: Seeking, Fear, Rage, Lust, Care, Panic and Play (Panksepp, 2005).

The processing of emotions is said to start on the limbic system, which is a sub-cortical region and primitive region of the brain. Citing LeDoux, “Emotions ‘happen to us’ more than we ‘make them happen’”. Emotions are constructs that aim at protecting us from external threatening stimuli and rewarding the ones we perceive as being good for us. The thalamus is a component of the brain that is located in a sub-cortical region of the brain (the ‘visceral brain’) and is part of the limbic system which is at the core of the processing of emotions (MacLean, 1949). When it comes to fear processing, LeDoux (LeDoux, 1998) theorizes that there are two circuits for it. On the one hand, the short circuit connects the thalamus directly to the amygdala without the contribution of neo-cortical regions that are associated to cognitive reasoning. The hippocampus is linked to the storage of emotional memories and, coupled with the amygdala, codes the response to the stimulus (Richter-Levin, 2004) and activates the autonomic nervous system which is responsible for regulating muscular and internal organs activity. On the other hand, the long circuit gets contribution

from the cognitive system that helps modeling the final response, before the activation of the autonomic nervous system.

These events trigger responses at different levels. Physiological responses produce occurrences in the body at a low level, hard to perceive by others and in some cases, even by the subject itself. These alterations take place at the autonomic nervous system and consist on alterations of the pupil dilation, heartbeat rate, blood pressure, skin temperature or the variation of activation of certain brain regions, among others. Behavioral responses take place at a higher level as they are the responses modeled by the somatic nervous system, which is responsible by voluntary muscular movements and reflexes. These responses are easier to perceive as they are more self-aware when compared to physiological ones. Facial expressions, body postures, speech intonation and gaze are some of these responses.

1.4 Thesis Hypotheses

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

- H1 “It is possible to predict a sentence’s chance of generating confusion based on syntactic, lexical and n-gram features.”
- H2 “A less confusing sentence on a virtual presentation increases the user’s sense of presence.”
- H3 “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.”

With respect to the first hypothesis, predicting the confusion level of a sentence may help building better scripts when they are being designed, thus improving the acceptance and efficiency of the presentation scenario. There are systems that detect confusion (usually with a locally collected dataset with self-reported confusion which is then used to train a supervised model) but they do not deepen the cause of why this confusion was triggered. By identifying the syntactic, semantic, and n-gram features that trigger confusion, one can build better content and improve the self-adaptation techniques of these systems.

The outcome of the first hypothesis only addresses the ability of a model to accurately classify sentences on their confusion level, but this does not necessarily translate into a better sense of presence on VEs. However, the second hypothesis states that this has an impact on the sense of

presence as it is more likely that a user may stay more focused on a presentation, the more she/he is able to follow it.

As for the third, adapting the lighting landscape of the VE may increase the sense of presence of the user. In a real-life environment this condition is not controllable, but in a VE it can be manipulated according to an objective. There are other parameters that could benefit from this approach, but lighting seemed as the most appealing due to studies that were already carried in real-life environments concerning the subjective effect of lighting on working conditions.

1.5 Objectives

To demonstrate the validity of our hypotheses we aim at the following objectives:

1. Build and train a model that predicts the confusion level of a sentence to improve the presentation scripts.
2. Develop a 3D CVE that provides a platform for virtual visits and a general 3D framework to build SoC and PrE scenarios. This experience shall compensate some shortcomings verified in a real-life visit.
3. Allow interaction between virtual visitors and employees through their avatars resorting to speech, mapping of facial expressions, and emotion recognition that controls their lighting condition.
4. Build a ubiquitous interaction platform that allows the collection of emotional context data.
5. Create an affective system that models video input and feeds the virtual environment adaptation systems.
6. Build a PrE scenario supported on the general 3D framework.

1.6 Research Methodology

The deliverables of this project are three: one learning model that provides offline confusion prediction of sentences and that is bound to objective 1, a 3D CVE that encompasses objectives 2., 3., 4., and 5., and a presentation scenario integrated into the 3D CVE that meets objective 6. The 3D CVE that will be produced is a representation of the company environment in which employees are able to communicate between themselves, with external visitors, and give presentations, all of this wrapped in an emotion-driven context to augment the user's sense of presence.

Unfortunately, the learning trained model from 1. will not be able to be integrated in real-time into the 3D CVE during the lifetime of this thesis. To our knowledge, currently there are no public datasets labeled with confusion for Automatic Emotion Recognition (AER) and is unfeasible for the duration of this work to collect affective labeled data to integrate the outcome of this goal in the system. Such real-time classification would require one of two methods to train a model: professional, independent, and blind judges to label the extracted features with the desired affective states (a more reliable approach), or self-reported labeling, which would basically replicate what other studies have already done. Instead, we opted to discriminate the semantic, syntactic, and n-gram roles of a sentence in confusion induction. Furthermore, when such public dataset is released, and as no system is ever perfect, the learning model we propose can be continuously trained in real-time with the automatically labeled data and feed the adaption system in real-time.

This dissertation starts with a wide review of CVE, introducing the general theme and laying definitions of fundamental concepts. This serves to identify problems and motivations and define the hypotheses and objectives of a solution to these problems. Next, narrowing down to the contributions enunciated in the previous sections, a review of the literature was carried to position ourselves (refer to “Chapter II – Literature Review”). During this stage, a technological review was performed to choose which would be the best suited Serious Game Engine (SGE) and Facial Recognition (FR)/AER system for our objectives (refer to “Chapter III – Technological Review”).

Once the technological review was done, the development stage started as the literature review was still going on and we assumed an iterative process. This was done to integrate the first technological prototypes of SGE and FR/AER to detect any shortcomings or needs before the objectives were completely set on stone. At this point the development of the foundations of the CVE started, namely, the design of the 3D environment and prototypes of AER and other important features. With this approach we developed lo- to hi-fidelity prototypes to assess the feasibility of the features of the general 3D framework and have a realistic insight about the effort and time required to develop each of them. While this happens, we prioritized the contributions according to the effort/time required and the value of the outcome. The iterative approach extended to the depth of the literature review carried out to each contribution according to the priorities they were given.

The machine learning model produced for H1 was evaluated by analyzing its f-score to assess its quality and reliability when applied to real problems. Finally, for H2 and H3 a user evaluation

was carried to assess the user's sense of presence gains of the system with the objectives of these hypotheses integrated when compared to the system without them.

1.7 Dissertation Overview

In this first chapter we introduced the underlying fields that support this dissertation. On “Chapter II – Literature Review” we give an overview of nonverbal behavior, facial and emotion recognition and end it by reviewing the state of the art on confusion detection on learning environments and adaptive lighting systems.

“Chapter III – Technological Review” surveys automatic facial and emotion recognition tools to be integrated in the general 3D CVE framework. Besides these tools, it also surveys game engines that can be used to develop serious games.

The general 3D framework is described on “Chapter IV - 3D Virtual Environment” where the modelling pipeline of the 3D model specifically built for this dissertation as a use case is described. Then, we describe the system developed on Unity that provides the features described in “1.5 Objectives” section.

On “Chapter V – Confusion Prediction” the development of the machine learning model to test H1 is described, as well as an evaluation and discussion of this testing. “Chapter VI – Case Study” goes on with hypothesis testing and describes the experiment to test H2 and H3, along with the PrE scenario used for this effect.

Finally, “Chapter VII – Conclusion and Future Work” throws a reflection about the results, what could have been done better and outlines future work stemming from this dissertation.

Chapter II – Literature Review

In this chapter we provide an overview of the current state of the art on nonverbal communication on CVEs and review the literature fields where our hypotheses are narrowed down and backed up.

Section “2.1 Background” gives a general review of nonverbal behavior and Multimodal Emotion Recognition systems. “2.2 Confusion Detection on Learning Environments” surveys work that has been done regarding the adaptation of the VE based on stimuli collected from the user. Finally, section “2.3 Adaptive Lighting Conditions” gives an overview of the state of the art on the detection of emotions that were identified as being part of the learning flow: engagement, confusion, frustration and boredom and why this is relevant. From this overview, we go into further detail about the confusion state and how it is being handled.

2.1 Background

2.1.1. Nonverbal Behavior

Nonverbal behavior is something impregnated into our daily lives and plays a big role in our interactions (Peña Pérez Negrón et al., 2015). This behavior is typically conveyed through facial expressions, kinesics, proxemics (Hall et al., 1968), gaze, loudness and intonation of speech (Guye-Vuilieme, Capin, Pandzic, Thalmann, & Thalmann, 1999). On CVEs, people tend to replicate the same behavior they have in real life (Peña Pérez Negrón et al., 2015), therefore non-behavioral reciprocity in CVEs thus had to be based on these means. However, immersion is easily broken, which means that the way to do this cannot be intrusive or laborious (Nguyen & Duval, 2015), as our own real life communication is not.

Awareness and communication are critical for the sense of presence in any VE. There are different types of awareness and communication, especially in CVEs (Nguyen & Duval, 2015), where it gets more complex. Framing these into the scenario templates identified in section “1.3.1 CVE scenarios”, “Awareness of others”, and “Awareness of the Virtual Environment” are the ones that go with the requirements for those templates. “Awareness of others” relates to a user’s capacity in a CVE to be aware of other users, what they are doing or where they are looking, emotional or facial expressions, which is something important in a Socialization (SoC) scenario. The “Awareness of the Virtual Environment” is related to any way the user can be made more aware of the VE. In this dissertation this is attempted through adaptive lighting conditions.

In a similar fashion, “Audio Communication”, “Embodiment and Nonverbal Communication”, and “Visual Metaphors” are the conceptual means of communication that best suit their characteristics. “Audio Communication” is a part of the general 3D CVE framework that allows users to communicate via an audio chat, which enriches a SoC scenario. “Embodiment and Nonverbal Communication” refers to natural human-to-human nonverbal cues, like gestures, facial expressions or body postures. Action unit mapping for facial expressions or gaze are examples that could enrich a SoC scenario. “Visual Metaphors” refer to any visual cues to enhance communication, in this case the adaptive lighting, that is used on our PrE scenario and tested in H3. Awareness builds up the context of what is happening around us, and so it does on CVEs. Knowing the state of others and what surrounds us provides us with information on how to act. But being aware of the state of the world is not enough, for one must also be able to react (communicate) effectively.

2.1.2. Multimodal Emotion Recognition

Multimodal Emotion Recognition (also called Multimodal Affect Detection) (MER) generally follows the architecture displayed on Figure 2. There are other input modalities that are not included in the figure but are widely adopted and that can be categorized into behavioral or physiological responses (Arguel et al., 2017). Physiological responses are the ones that are triggered by the nervous system, like galvanic skin response, electrodermal activity, electrocardiography (Dias, Eloy, Carreiro, Proença, et al., 2014; Eloy et al., 2015; Jang, Park, Park, Kim, & Sohn, 2015), electroencephalography (Abdelrahman, Hassib, Marquez, Funk, & Schmidt, 2015; Yan et al., 2016). Multimodality has several advantages over unimodality, like robustness to noise (if one channel, i.e. facial, is noisy, the model can give more weight to other, i.e. speech), and has been proven that it can reach a higher accuracy (D’Mello & Kory, 2012, 2015).

The user is continuously monitored by these devices and the stream is analyzed in real-time with supervised learning algorithms that can act at feature- or decision-level. At feature-level, the fusion is performed at an earlier stage by joining features from all modalities in the same dataset upon which the trained model will act and perform the emotion classification. This approach allows the establishment of correlations between features coming from different modalities and the application of dimensionality reduction (i.e. through the Principal Component Analysis algorithm). However, the synchronization between the streams coming from different modalities may prove to be hard as they usually collect data with different rates and on different scales, which would require adding feature normalization to the process. On the other hand, at decision-level the

fusion of modalities is done after each algorithm provides its classification score. This eases the process when compared to feature-level, but it also does not share its advantages.

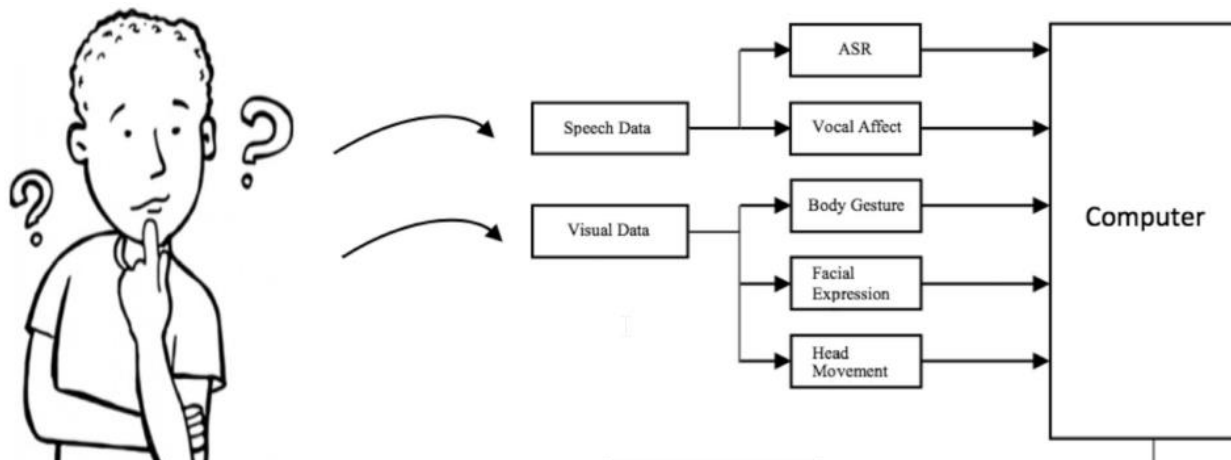


Figure 2. Multimodal AER framework. Speech and visual data are collected from the user and affective systems detect emotions from that data. That information is interpreted and produces output back to the user. ASR - Automatic Speech Recognition (Poria, Cambria, Bajpai, & Hussain, 2017).

At the moment of the writing of this dissertation and to the knowledge of the author, there were no commercially or academically available MER systems. Affectiva¹ offers facial emotion recognition and is working on speech emotion recognition, however, this service is not yet available, much less is it integrated in a MER.

2.2 Confusion Detection on Learning Environments

Distance-learning is a growing market, especially the so-called MOOCs (Massive Open Online Course) (Bersin, 2016). Most of these are based on video scripts where experts (either on the academia or the industry) teach the enrolled students about the matter of the course. However, Intelligent Tutoring Systems (ITS) are showing promising results when compared to human tutoring (Graesser, 2016). These systems often take place in VEs and have Animated Pedagogical Agents (APA) (see Johnson & Lester, 2016; N. L. Schroeder, Adesope, & Gilbert, 2013; Soliman & Guetl, 2010) that provide visual support and enhance the learner's engagement with the course. When compared to traditional tutoring scenarios (mostly classrooms or any kind of physical contact), this scenario carries advantages due to its ability of collecting data in a way that can be measured. This data (emotional state, click behavior, engagement level) is then used to adapt the ITS accordingly.

¹ <http://blog.affectiva.com/introducing-affectivas-emotion-recognition-through-speech>, accessed 24 September 2018

Much research has been conducted in affective detection on distance-learning scenarios, either to assess which affective states are most observed and relevant to this context, and how to automatically detect them. In opposition to the basic emotional states that typically occur in emotion-driven situations, in learning contexts there is a set of more complex, non-basic emotional states. Sidney D’Mello, Arthur Graesser and colleagues have been conducting extensive research on identifying and detecting learning-centered affective states and adapting their ITS, AutoTutor (D’Mello & Graesser, 2012a) [Figure 3], 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, surprise (Craig, Graesser, Sullins, & Gholson, 2004; D’Mello, Craig, Sullins, & Graesser, 2006; D’Mello, Lehman, & Person, 2010; D’Mello & Graesser, 2006; Hussain, AlZoubi, Calvo, & D’Mello, 2011).

Some studies have been trying to perform Automatic Emotion Recognition (AER) to detect some of these states through Action Unit (AU) detection (D’Mello & Graesser, 2012a; J F Grafsgaard, Wiggins, Boyer, Wiebe, & Lester, 2013; Joseph F. Grafsgaard et al., 2014; Joseph F Grafsgaard, Wiggins, Boyer, Wiebe, & Lester, 2013), physiological signals (Hussain et al., 2011), learner behavior (Bixler & D’Mello, 2013), conversational cues (D’Mello, Craig, Witherspoon, McDaniel, & Graesser, 2008), and gross body language (D’Mello & Graesser, 2009). 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 (D’Mello & Calvo, 2013; D’Mello & Graesser, 2012b).

A **AUTOTUTOR INTERFACE**

The screenshot shows the AutoTutor interface. On the left is a 3D avatar of a male conversational agent. The main area contains a 'Main Question' in red: 'How does information that you type in get passed from the keyboard to the hard disk?'. Below the question is an 'Auxiliary Diagram' showing a flow from 'Keyboard' to 'Central Processing Unit' to 'Hard Drive'. A 'Student Response' is shown at the bottom: 'through the CPU'. There are 'Submit' and 'Tutorial' buttons at the bottom left.

B **SAMPLE TUTORIAL DIALOGUE**

T: When you turn on the computer, how is the operating system first activated and loaded into RAM? *[main question]*

S: When the computer is turned on a file is automatically booted up.

T: Anything else? *[pump for info]*

S: This file contains the operating system.

T: What about storage? *[hint]*

S: The operating system is then read into RAM and activated.

T: Right. The CPU executes instructions in ROM, which helps the operating system to start. *[after positive feedback, there is an assertion]*

Figure 3. Sample frame taken from the interaction interface between a learner and AutoTutor.

D’Mello and Graesser have conducted an experiment (D’Mello & Graesser, 2012b) that yielded a model [Figure 4] that initially hypothesized affect transitions between engagement/flow → confusion, confusion → engagement/flow, confusion → frustration and frustration → boredom. In addition, surprise and delight were 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 3 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.

As confusion is the central subject of this section, an explanation of cognitive disequilibrium is due. Citing D’Mello and Graesser: “Cognitive disequilibrium is a state of uncertainty that occurs when an individual is confronted with obstacles to goals, interruptions of organized action sequences, impasses, contradictions, anomalous events, dissonance, incongruities, unexpected feedback, uncertainty, deviations from norms, and novelty.”. This means that this is the event that results from a stimulus applied to a learner when he/she is engaged on a learning process. Triggering this event is especially important because there is evidence that suggests that inducing confusion to lead the learner into a deep learning state produces higher learning gains (D’Mello et al., 2014).

The nodes in Figure 4 have the learning-centered affective states represented in parentheses and the associated events in bold. The solid lines represent the affect transitions that were hypothesized, whereas the middle transitions are expressed through dashed lines and can happen or not, but that will ultimately lead to the original affective states that are connected through solid lines.

This model assumes that the modeling starts when the learner is on one of the two leaf nodes (engagement/flow or boredom) and that he is actively focused on the content that is being delivered. The learner is usually in a long-engaged state with the content he/she is trying to master until a stimulus triggers a cognitive disequilibrium, leading him/her to a confusion state. But, on the one hand, if the stimulus is too disruptive, the middle state of surprise can be experienced before arriving at the disequilibrium. On the other hand, if the subject is already in a state of confusion and can resolve the impasse, this can generate a middle state of delight before engaging in the task again.

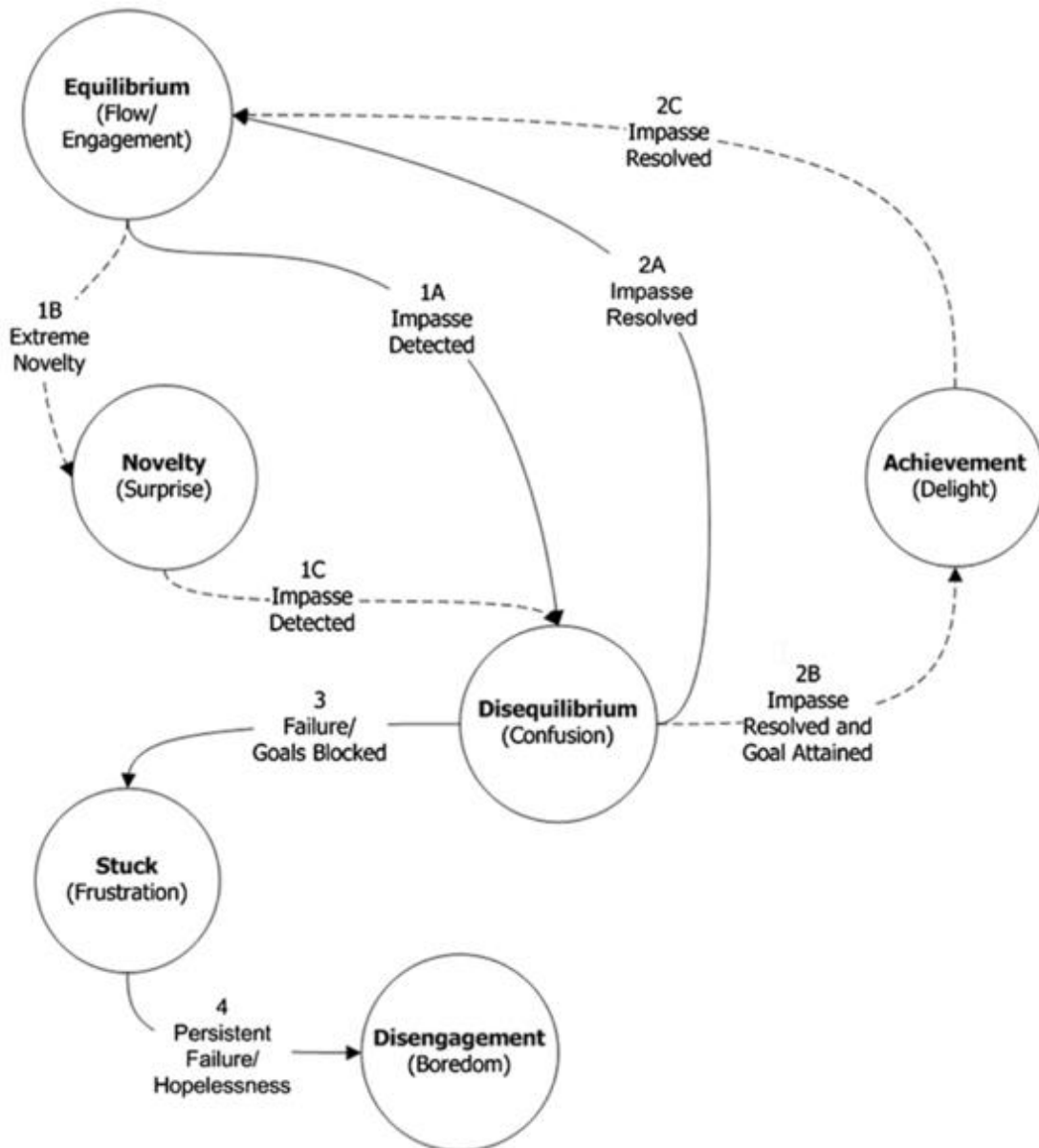


Figure 4. Model of learning-centered affective states as proposed by D'Mello and Graesser. It describes a set of relations between learning states.

Two studies were carried with emote-aloud feedback from participants ($N_1 = 28$, $N_2 = 30$) regarding the affective states they were experiencing during the test. The experimental design was similar for both studies, with the participants interacting with AutoTutor while their expressions were being recorded, as well as the whole log of the interaction and the screen activity. The test ran without interruptions and then they performed a retrospective evaluation of their affective states by watching the screen recording and their video. There were fixed and spontaneous judgments of affection during this evaluation. In the first study, the fixed judgments were given every 20 seconds of the video, with the spontaneous ones being given in between. On the second study, fixed judgments were given a few seconds after AutoTutor completed a tutor move, immediately before the participant answered the question and other randomly chosen points during the dialogue. The spontaneous judgments were given in between these fixed ones. After the experiment, participants were provided with two lists, one of them with affective states definitions and the other with affective states names and were asked to match them to establish a baseline of state understanding. They were able to keep these lists during the test.

The reported instances of boredom, engagement/flow, confusion frustration and neutral states were significantly higher than delight or surprise for both studies, which is line 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. All the first three were gravitating around confusion, which stresses out the important role of this affective state during information acquisition. There was also evidence of additional patterns of boredom \rightarrow frustration and frustration \rightarrow confusion, however, this falls out of the scope of this analysis due to its lack of robustness.

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 the level of productive confusion but avoiding the evolution to frustration (hopeless confusion) (Lehman, D'Mello, & Graesser, 2012; Lehman et al., 2011, 2013). This regulation of confusion has been considered as the “zone of optimal confusion” and is displayed in Figure 5 as an adapted version of previous work (Arguel et al., 2017).

D'Mello et al. (D'Mello et al., 2014) study results show evidence that a moderate state of confusion can be beneficial for learning, as long as it is overcome. Most of the ITS and APA focus on how to react to this confusion state (D'Mello & Graesser, 2012a) but they do not identify what was its source.

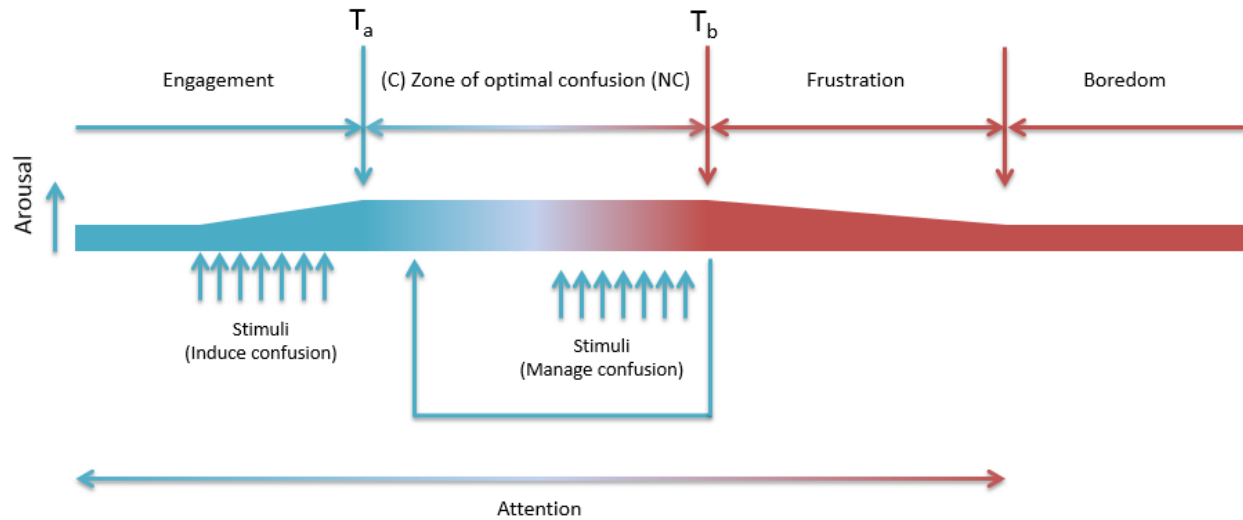


Figure 5. Adaptation, as proposed by Arguel, of the learning model of D'Mello and Graesser with the zone of optimal confusion.

2.3 Adaptive Lighting Conditions

We believe that the familiarity of the virtual environment itself is crucial to the sense of presence. The pure architectural space, stripped from objects, is unconsciously recognizable (“I know this place, but not quite sure where from”) and can make users experience certain emotion (fear, positive/negative valence, arousal) (Dias, Eloy, Carreiro, Proença, et al., 2014; Dias, Eloy, Carreiro, Vilar, et al., 2014), but it is on salient details that reside the anchors for familiarity and orientation in space (Eloy et al., 2015). Beyond these salient details, small scale entities like trees, people or a pencil give sense of scale to the environment through cognition. In addition, the weather, daytime 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 according to an objective. These manipulations can be made according to the context of the users, but the VE itself is a part of this context, creating this symbiotic and cyclical relationship.

As essential as it is to the human body (Wurtman, 1975), light is a constant throughout history. Since artificial lighting came to be, people also manage this condition according to their needs. The effect of lighting is being studied as a variable that influences several traits in many fields of knowledge (Knez & Kers, 2000; Kuijsters, Redi, De Ruyter, & Heynderickx, 2015; Mott, Robinson, Walden, Burnette, & Rutherford, 2012; Park & Farr, 2007; Quartier, Vanrie, & Van Cleempoel, 2014). Another field of application of lighting management is on the workplace. Hawes et al. (Hawes, Brunyé, Mahoney, Sullivan, & Aall, 2012) propose a work where they study

the effect color temperature as on the emotional state of the subjects. Four workplace scenarios where set up with lights with different color temperatures in Kelvin degrees:

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

The study was carried with 24 participants with a within-participants repeated-measure design where each participant visited the laboratories in five consecutive days to take a first practice day and then be exposed to each of the lighting conditions. For each test, the subject took the Profile of Mood States (POMS) (McNair, Lorr, & Droppleman, 1971) to assess his/her emotional state after and before the test to assess the differences. Objectively, in our study we hypothesize that the sense of presence changes as a function of the lighting condition by means of emotional parameters such as valence or arousal. In this study, results showed that higher color temperatures were related to more aroused states and lower depression rates. Moreover, their results support the theory that “(...) lighting can alter environmental conditions enough to increase positive mood and decrease fatigue”. This is directly related to their findings that the 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 that our adaptive system that will use to evaluate H3. All of the 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. Most of the research on adaptive VEs based on emotion is centred on MOOCs, e-learning or training scenarios (Scott, Soria, & Campo, 2016; Vaughan, Gabrys, & Dubey, 2016) and they do not take advantage of lighting conditions.

However, Yan et al. (Yan et al., 2016) used a Brain-Computer Interface (BCI) device to collect data about users engagement while attending to a virtual version of the opera *Siegfried* and the dance *The Tramps of Horses*. The user is monitored with this BCI that detects disengagement and re-engagement and acts according to this input. The way the system acts is through a set of pre-designed performing cues based on the classic theatre performing theory. They have focused on scenic design (which includes lighting) like adapting display blocks, simulate stage effects, like

smoke or fog, which is known to be a good engagement agent, and lighting lets the stage controllers get the audience attention to wherever is desired.

Forty-eight users were exposed to three conditions with evenly distributed gender, with sixteen users per 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. Figure 6 shows a live performance of the *Siegfried* performance on the top, along with the three experimental conditions.

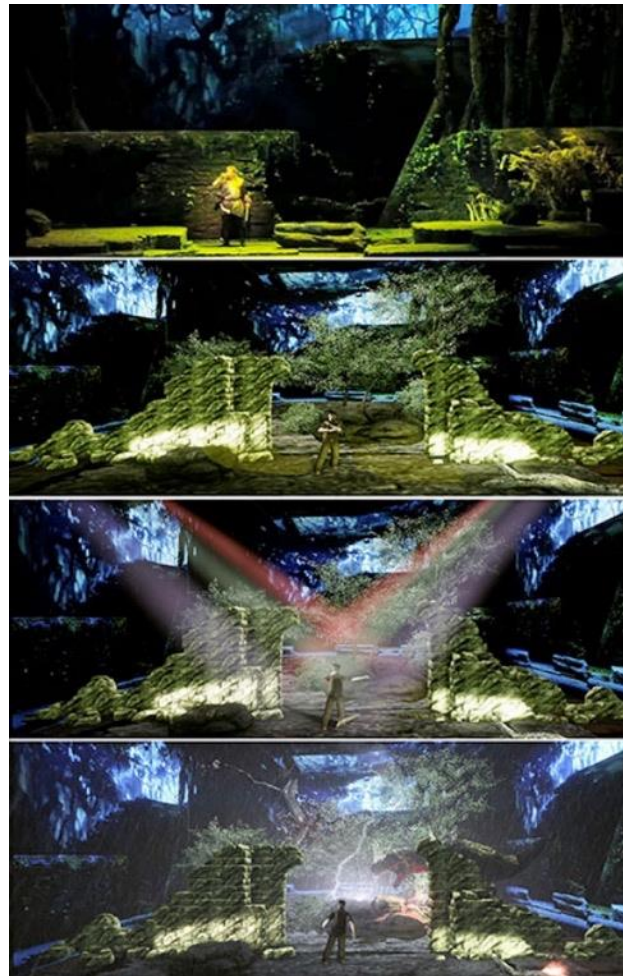


Figure 6. At the top is a real live performance of the opera *Siegfried*. The three lower displays are the 3 different conditions that users experimented.

The test took place in a 30-square meter laboratory with a 135° circular projection plane for the 3D visualization. Results show that the system could detect significant variations of engagement successfully for both performances and the adaptations could recover the users' engagement when triggered.

2.4 Summing Up

MER appears to be still under development both on the academia and on the industry, without a public solution to be used to carry research with such a tool. A natural interaction based on non-verbal behavior is an ambitious goal that can be achieved with a contribution of affective sensing. However, there are other harder constraints, mainly physical ones. The strength a person applies when shaking a hand is something that would require a way to apply this same effect on the other person, however, currently there is no way to do this without resorting, for example, to force feedback devices, and such devices are not present in common offices. In this dissertation we draw a first effort to meet the different kinds of awareness and natural communication to increase VEs acceptance as a mainstream tool, but there is still much work to be done.

In this chapter we also discussed two examples of how a set of non-basic emotions can be explored in VEs. Lighting is regarded as one of the most basic needs for the human body and, as such, has been studied across several fields of knowledge. One of the presented studies is particularly interesting as adaptations of the lighting condition are performed during the test and based on the user's brain activity. Its results show that lighting can, in fact, be an important condition to manage user's engagement. Despite focusing on the lighting condition, there are other environmental conditions that may be used to affect the user's emotional context. This a complex theme as there are several variables that affect lighting (like temperature, or position on the visible spectrum), let alone a set of environmental conditions.

It is useful to keep a user focused on a task, but in the context of learning it is also important to induce constructive confusion and manage it. There is extensive research on how confusion affects learning, and tests to detect and act upon it. The main premise is that confusion can be beneficial for learning if properly managed. Research has been carried in the learning scenario to detect confusion using several modalities and to deal with this confusion, but it does not try to identify what triggers this confusion in first place. We consider this of high relevance so that a system that identifies confusion can also pinpoint what part in the speech was likely to induce this confusion. There can be a handful of sources for confusion depending on how information is conveyed. It can be an image hard to interpret, complex text, a different tone of voice, or other countless options.

Chapter III – Technological Review

This chapter divides into two reviews of critical technology for this work. First, of Automatic Emotion Recognition (AER) (also named Affective Recognition or Detection) and Facial Recognition (FR) SDKs (Software Development Kit) and APIs (Application Programming Interface) and secondly, of serious game engines (SGE).

3.1 Automatic Facial & Emotion Recognition

AER enables an emotional context awareness of the user that can be used to take diverse actions. Therefore, this dissertation is supported on this technology. However, we stress that creating a system that outperforms existing ones in terms of recognition is not a focus of this dissertation, nor it is to compare them. For a more detailed review of the state of the art on affective recognition please refer to Section “2.1.2 Multimodal Emotion Recognition”.

There are three major modalities that capture signals from the human senses and are used for affective sensing in computation: video, audio and text. AER solutions based on video that focus on facial features to compute emotion, like Affectiva², Kairos³, Emotion API⁴ from the Microsoft Cognitive Services, eyeris⁵, FaceReader⁶ or Sightcorp⁷, are based on the Facial Action Code System (FACS) (P. Ekman and W. V. Friesen, 1977) that provides a series of metrics called “Action Units” (AU). These AUs are detected based on FR which is performed using facial landmarks. They code the movement of meaningful facial muscles as depicted in Table 2. The discrete emotional model followed by Ekman (Ekman & Friesen, 1969) can use combinations of these AUs to prototype emotions. OpenFace (Baltrusaitis, Robinson, & Morency, 2016) is an open-source toolkit for academia that is able to compute these AUs and estimate other parameters like head pose and eye gaze. Even though OpenFace provides eye gaze estimation and more robust measures of head tracking and AUs, it lacks the trained emotional model that the other solutions have. Therefore, it only provides a FR SDK without AER.

In spite of the approaches listed in the section “1.2.3 Emotion”, in the field of AER performed

² www.affectiva.com, accessed 24th September 2018

³ kairos.com/, accessed 24th September 2018

⁴ azure.microsoft.com/en-gb/services/cognitive-services/emotion/, accessed 24th September 2018



















⁵ <http://www.eyeris.ai/>, accessed 24th September 2018

⁶ www.noldus.com/human-behavior-research/products/facereader, accessed 24th September 2018

⁷ sightcorp.com/, accessed 24th September 2018

with FR the most adopted model (Sariyanidi, Gunes, & Cavallaro, 2015) is the one proposed by Ekman and Friesen (Ekman & Friesen, 1971). With this work, Ekman follows Darwin’s approach, but FACS itself does not code emotions. However, a compound of specific AUs is able to code an emotion (Matsumoto, Keltner, & Shiota, 2016).

Table 2. Facial Action Code System (FACS) (P. Ekman and W. V. Friesen, 1977). Each Action Unit codes a different facial movement.

AU01	AU02	AU04	AU05	AU06	AU07
					
Inner Brow Raiser	Outer Brow Raiser	Brow Lowerer	Upper Lid Raiser	Cheek Raiser	Lid Tightener
AU09	AU10	AU12	AU14	AU15	AU17
					
Nose Wrinkler	Upper Lip Raiser	Lip Corner Puller	Dimpler	Lip Corner Depressor	Chin Raiser
AU20	AU23	AU25	AU26	AU28	AU45
					
Lip Stretcher	Lip Tightener	Lips Part	Jaw Drop	Lip Suck	Blink

With the advent of Machine Learning and Natural Language Processing there are also several AER solutions based on text analysis like IBM Watson’s Natural Language Understanding⁸ (former Alchemy API) or IBM Watson’s Tone Analyzer⁹, Text Analytics API¹⁰ from the Microsoft Cognitive Services, Repustate API¹¹, among others. Some of these only provide sentiment analysis (measure positivity-negativity or valence), whereas others provide emotional classification.

Finally, AER solutions based on audio are not so well spread as the other modalities, but it has been getting some attention. The range of solutions is scarcer but there are solutions like VokatURI¹², DeepAffects’ Emotion Recognition API¹³, sensAI¹⁴ from audeERING and upcoming Emotion API for Speech¹⁵ from Affectiva. Besides these commercial solutions, there are also solutions developed in the academic community like D.A.V.I.D.¹⁶ or EmoVoice¹⁷.

⁸ www.ibm.com/watson/services/natural-language-understanding/, accessed 24th September 2018

⁹ www.ibm.com/watson/services/tone-analyzer/, accessed 24th September 2018

¹⁰ azure.microsoft.com/en-us/services/cognitive-services/text-analytics/, accessed 24th September 2018

¹¹ www.repustate.com/sentiment-analysis/, accessed 24th September 2018

¹² vokatURI.com/, accessed 24th September 2018

¹³ www.deepaffects.com/emotion-recognition-api/, accessed 24th September 2018

¹⁴ audeering.com/technology/sensai/, accessed 24th September 2018

¹⁵ blog.affectiva.com/introducing-affectivas-emotion-recognition-through-speech/, accessed 24th September 2018

¹⁶ cream.ircam.fr/?p=44, accessed 24th September 2018

¹⁷ www.informatik.uni-augsburg.de/lehrstuehle/hcm/projects/tools/emovoice/, accessed 24th September 2018

In their study, Paulmann and Pell (Paulmann & Pell, 2011) show evidence that emotion recognized from multimodal communication from congruent sources is able to achieve a higher accuracy than through uni-modal means. However, in the uni-modal domain, facial expressions and semantics from text show significantly higher accuracy when compared to auditory prosody. Facial expressions seem to be the most promising modality even when compared to text since text requires cognitive processing which can turn it into a more unnatural channel. In addition, text signal is a discrete input, whereas facial expressions remain as continuous input even when it conveys a neutral emotion.

There is a lack of commercial multimodal AER solutions and, even though there are some academic studies that properly fuse different modalities for AER, there is no academic solution released that can be used for this effect. Therefore, we focus on AER through FR and explored not only OpenFace, but also Affectiva for different goals. Both SDKs provide facial landmark detection, which in turn leads to AU detection, gaze and head pose estimation. Besides these features, Affectiva also provides access to emotion classification based on AU detection and provides an API for Unity.

3.2 Serious Game Engines Review

Ali and Usman have built a framework for choosing a game engine for serious games and review work that compare serious game engines (SGE) (Ali & Usman, 2016). They rate each publication according to four criteria: 1) the presence and application of a filter (defined per paper) to choose which SGEs to evaluate, 2) the features used to evaluate each SGE, 3) validation with a case study and 4) the range (number) of games engines compared. From the 10 reviewed publications, the one from Vasudevamurt and Uskov (Vasudevamurt & Uskov, 2015) is considered because it analyses a big collection of SGEs ($N = 23$), while it also includes the features [Table 3] needed for this project. In addition, the following specific features were added to the feature list:

- Supports VR,
- Supports AR,
- Third-party integration for:
 - Kinect (or other motion device),
 - Eye-tracking,
 - Speech recognition,

- Emotion recognition.

Table 3. Features for SGE comparative analysis framework.

Feature	Requirements to SGE features
<i>Graphics</i>	SGE must support both 2D and 3D graphics (these days only a few game engines exist that support both 2D and 3D graphics)
Work flow editor	Work flow editor is an important part of SGE – it provides the game developer with no coding experience with an opportunity to create complex actions in the game; it must be built on a concept of object-oriented graphical editor
<i>World (level) editor</i>	World (level) editor, as a crucial SGE component, must help the SG developer to edit or create a map/world using available objects
Character model editor	SGE’s model editor should enable the SG developer to create or edit game’s 2D or 3D assets/characters
Texture editor	Textures are images which are superimposed on 3D model to give it a realistic feel
Cinematic support	SGE’s cinematic support enables the SG developer to create in game movies, where they want a specific incident to happen; thus, game developer becomes a movie director
<i>Physics</i>	Physics – a vital part of SGE; it adds the realistic feel to game play, including lighting, collision detection, solidity of an object, water and cloth physics, weather, etc
<i>Artificial Intelligence (AI) support</i>	SGE must provide SG developers with complex behavior algorithms (the existing examples are limited to Unreal Engine 3 and CryEngine 3 SGE).
<i>Networking</i>	SGE should provide comprehensive networking support for SG developers to develop advanced and sophisticated distributed, real-time multiplayer games; the key topics in networking include latency, reliability, bandwidth, and security issues.
<i>Creation of online game</i>	SGE must be focused on Web-based games that are cross platform systems (these days only a few 3D game engines support this features)
Programming experience required	SGE must support SG developers with various programming skills; particularly, SG developers with good programming skills should have opportunities to control almost all the aspects of the game engine by themselves. On the other hand, SG developers with low level of programming skills should be supported by SGE by concentrating mostly on content creating and using the work flow editor do add logic in SG.
<i>Scripting</i>	SGE should provide SG developer with complete scripting support to write his/her own complex code to optimize SG efficiency in terms of performance and memory usage
Platforms supported	SGE must support easy cross platform deployment of developed SG; however, the SG experienced developers may need to look out for some performance issues of SG on the targeted technical platform

Firstly, Vasudevamurt and Uskov present Table 4 of serious games used in three specific fields of application and thus it can be concluded that, excluding the column “Educational SG identified”, Unity SIM, Unreal Engine 3 and Unigine SIM are the most used.

Table 4. SGE used for serious games development in selected areas (Vasudevamurt & Uskov, 2015).

Game engine	Educational SG identified	Simulation SG identified	Virtual Reality SG identified
Unity SIM	12	10	4
Torque	13	2	5
Unreal Engine 3	1	4	5
Unigine SIM	0	6	4
Neoaxis	3	2	0
CryEngine 3	0	2	0

On Table 3 Vasudevamurt and Uskov provide a description for each feature that was used to evaluate the different SGEs. The new extra features added to this specific project can be found on Table 5. Features marked in italic on both tables are especially relevant for this project and will determine the recommended SGE.

Table 5. Extra features considered for the project.

Feature	Requirements to SGE features
<i>Supports VR</i>	The SGE must be able to build the SG for VR, either natively or through the integration of a third-party application
<i>Supports AR</i>	The SGE must be able to build the SG for AR, either natively or through the integration of a third-party application
<i>Motion control device</i>	The SGE must support the integration of motion control devices such as Microsoft Kinect or PlayStation Move for gesture or skeleton/joint recognition – either natively or through third-party applications
<i>Eye-tracking</i>	The SGE must support the integration of eye-tracking technology to take advantage of the user’s gaze – either natively or through third-party applications
<i>Speech recognition</i>	The SGE must support the integration of speech recognition – Speech-to-text, Text-to-speech, Speaker identification – either natively or through third-party applications
<i>Emotion recognition</i>	The SGE must support the integration of emotion recognition (based on video, speech tone or audio signal), either natively or through third-party applications

According to the marked features, six SGEs were selected and are presented in Table 6, where they are compared based on the joint features of Table 3 and Table 5. What is presented is a new table with a binary scale (“have” vs. “does not have”) that compare those platforms. A binary scale was adopted because only the presence/non-presence of features is being evaluated, rather than their quality or ease of use, which would require further investigation and evaluation.

Table 6. Updated table of comparison for the project.

Feature	Engine					
	Unreal Engine 4 ¹⁸	CryEngine 5.3 ¹⁹	Unity3D 5.6.1 ²⁰	NeoAxis 3D 3.5 ²¹	Torque3D 3.6.1 ²²	Unigine 2 ²³
3D Graphics	✓	✓	✓	✓	✓	✓
World (level) editor	✓	✓	✓	✓	✓	✓
Physics	✓	✓	✓	✓	✓	✓
Artificial Intelligence (AI) support	✓	✓	✓	✓	✓	✓
Networking	✓	✓	✓	✓	✓	✓
Creation of online game	✓	✗	✓	✗	✓	✗
Scripting	✓	✓	✓	✓	✓	✓
Supports VR	✓	✓	✓	✗	✓	✓
Supports AR	✓	✗	✓	✗	✗	✓
Motion control	✓	✗	✓	✓	✓	✓
Eye-tracking	✓	✗	✓	✗	✗	✗
Speech recognition	✓	✗	✓	✗	✗	✗
Emotion recognition	✗	✗	✓	✗	✗	✗
Platforms	Windows, PS4, Xbox One, Mac OSX, iOS, Android, VR (several), Linux/SteamOS, HTML5	CMake	iOS, Android, Windows Phone, Tizen, Windows Store Apps, Mac OSX, Linux/Steam Os, WebGL, PS4, PS Vita, Xbox One, Wii U, Nintendo 3DS, VR (several), AndroidTV, Samsung SMART TV, tvOS, Nintendo Switch, Fire OS, Facebook Gameroom	Windows, Mac OSX	Windows, Mac OSX	Windows, Linux, Mac OSX

¹⁸ www.unrealengine.com, accessed 24th September 2018

¹⁹ www.cryengine.com, accessed 24th September 2018

²⁰ unity3d.com, accessed 24th September 2018

²¹ www.neoaxis.com, accessed 24th September 2018

²² torque3d.org, accessed 24th September 2018

²³ unigine.com, accessed 24th September 2018

As Table 6 shows, every SGE that was analyzed can produce the basic 3D experience with higher or lower visual fidelity in exchange for low or free subscription plan (exception made for Unigine 2). The differences lie on the available target platforms, support for VR, AR, and natural interfaces and devices.

The deployment for mobile and web-based platforms is a must-have to take advantage of mobility scenarios that are comprised on ubiquitousness. On the other hand, the support of natural interfaces and devices are mandatory for more complex and immersive experiences. VR, AR, motion control, eye-tracking, speech and emotion recognition can all provide far more immersive experiences, without relying on simple mouse and keyboard, but using speech, gaze and emotion. Based on the support for natural interfaces and devices, and in the capacity to deploy to mobile or web-based platforms, Unreal Engine 4 (UE4) and Unity 5.6.1 stand out from the others, followed by Torque3D 3.6.1 that has the advantage of being the only open-source SGE, but only supporting VR and motion control, failing to deliver on others. AER/FR, already identified as a major feature of this work, puts a major toll on the choice of the adopted SGE.

In conclusion, UE4 and Unity differ only in the lack of support for AER/FR, in the licensing and target platforms for deployment. Thus, Unity represents the best choice due to the critical role of AER/FR and multiplatform deployment.

To the best of our knowledge, there is still no scientific deep comparison between Unity and UE4 for the development of serious games. There are several comparisons on the indie game development community however, there is no clear answer as to which SGE is better. These comparisons and tests have no background scientific methodology and they are not aimed at serious games but rather at generic games. Determining which SGE is best for the development of serious games is not one of the goals of this project, so the theoretical pros and cons of each SGE are weighted in Table 7 to determine which would be chosen.

Ultimately, Unity was chosen as the SGE to use for the practical development of this dissertation. Its small learning curve and easy prototyping were decisive and some of its cons would not really apply in this context. This work does not require AAA graphics quality and will not result in such a big project as a commercial game that would lead to a point where the assets would be hard to manage. As a lone project, the big community it possesses is also important when problems may arise.

Table 7. Pros and cons between UE4 and Unity.

	Pros	Cons
Unreal Engine 4	<ul style="list-style-type: none"> • AAA graphics out-of-the-box • Open-source engine code • Supports Oculus Rift • Blueprints (visual scripting) for rapid prototyping • Designed for large projects (more mature) 	<ul style="list-style-type: none"> • Steep learning curve • Users report major bugs with Android, HTML5 and iOS • Few plugins in the Marketplace
Unity 5	<ul style="list-style-type: none"> • Small learning curve • Big community • Great UX • Lots of assets that can be used for rapid prototyping • Emotion recognition support • Supports a lot of platforms (including HoloLens) • More matured for mobile <ul style="list-style-type: none"> ○ Mobile VR devices have low HW requirements ○ Vuforia AR extension free 	<ul style="list-style-type: none"> • Some features (like AI) come from third-party assets and are not natively built which can create undesired dependencies • AAA graphics require Pro license • Poor version control • Managing assets for large projects is hard • Some features may require Unity Pro

3.3 Summing Up

As we are dealing with emotional states to inform the system, AER is a core theme of this dissertation. There are several uni-modal tools for facial and vocal recognition of emotions, but the development of multimodal ones is underway. This evolution is critical for a strong integration of emotional feedback on CVEs. However, mapping AUs on the avatar’s face can already prove useful when trying to achieve reciprocity between users.

As important as the AER tools is the SGE chosen to build the practical side. There is a lot of offer in this field with some SGEs being more complete than others. Some offer an SDK, whereas others offer a complete suite, like Unity or Unreal. In the end, the decision came to these two SGEs and both are well-known. Unity is widely adopted in the indie game development community, but has also produced AAA games, whereas Unreal Engine has the Unreal trademark supporting it and offers several features.

Chapter IV – 3D Virtual Environment

This chapter describes the conception and ambient of the 3D virtual environment that contains the stage where the Presentation (PrE) scenario took place. The environment replicates a portion of the Company office and was developed to accommodate the social and presentation scenarios. With a unified environment that comprised both scenarios, the virtual visitor could experience the flow of entering the virtual office, interact with other remote users or watch a presentation. However, the identified hypotheses are all tested in this last scenario and takes advantage of the distributed system architecture. This system can be considered a host that provides enriched interaction upon which one can develop different scenarios with this 3D office environment as background and top-level scenario.

The pipeline followed to build this 3D environment is described in the next section. Following that pipeline, we will describe the system architecture built on top of the Unity high-level networking API and its components. Finally, we present Affectiva and OpenFace in the context of this dissertation, tests carried to compare similar features, and how they were integrated with the system.

4.1 3D Modeling Pipeline

The conception of the 3D model started out with a set of floor plans that were imported to Autodesk Revit 2017²⁴ (from now on, only “Revit”). A diagram of the pipeline can be seen on Figure 7, summarizing the flow required to produce the 3D model. Briefly, Revit was used to build and texturize the 3D geometry of the physical, inanimate space. This model was exported to Autodesk FBX²⁵ (from now on, only “FBX”), a well-known and standardized file format that supports 3D geometry embedded with materials with textures, lights, cameras, animations, among other data. The FBX file was imported into Autodesk 3DS Max 2017²⁶ (from now on, only “3DS Max”) to correct some transformations and object features so that it could be well consumed by Unity3D. The game mechanics were implemented in this last software as well as the environment graphics concerning lighting and some texturing.

²⁴ <https://www.autodesk.com/products/revit/overview>, accessed 24th September 2018

²⁵ <https://www.autodesk.com/products/fbx/overview>, accessed 24th September 2018

²⁶ <https://www.autodesk.eu/products/3ds-max/overview>, accessed 24th September 2018



Figure 7. Modelling pipeline. Revit was used to model the architectural space, whereas 3DS Max is used to fix transformations and other details. The game logic is implemented in Unity.

Revit is a software produced and maintained by Autodesk that was developed to serve the Architecture, Engineering and Construction (AEC) community. As such, it provides a series of tools that facilitate the production of 3D construction models across all fields involved in AEC (e.g. heating, ventilation, and air conditioning (HVAC), structure, infrastructure or architecture). These tools differ depending on the field one is focused on. For this design we used the set of architectural tools that provide the placement of inanimate objects like walls, windows, pillars, floors/slabs, ceilings or doors based on the floor plans. Figure 8 shows the interface of Revit for the “Floor Plans” section (check the lower left corner, on “Project Browser”) with “Level 2” of an example project selected. As the model is being built from 2D drawings, the 3D model can be checked on the “3D Views” [Figure 9] section of the Project Browser. Due to compliance, the floor plans and 3D model built for this project cannot be entirely published in this document, thus images of a template project are provided to explain these tools.

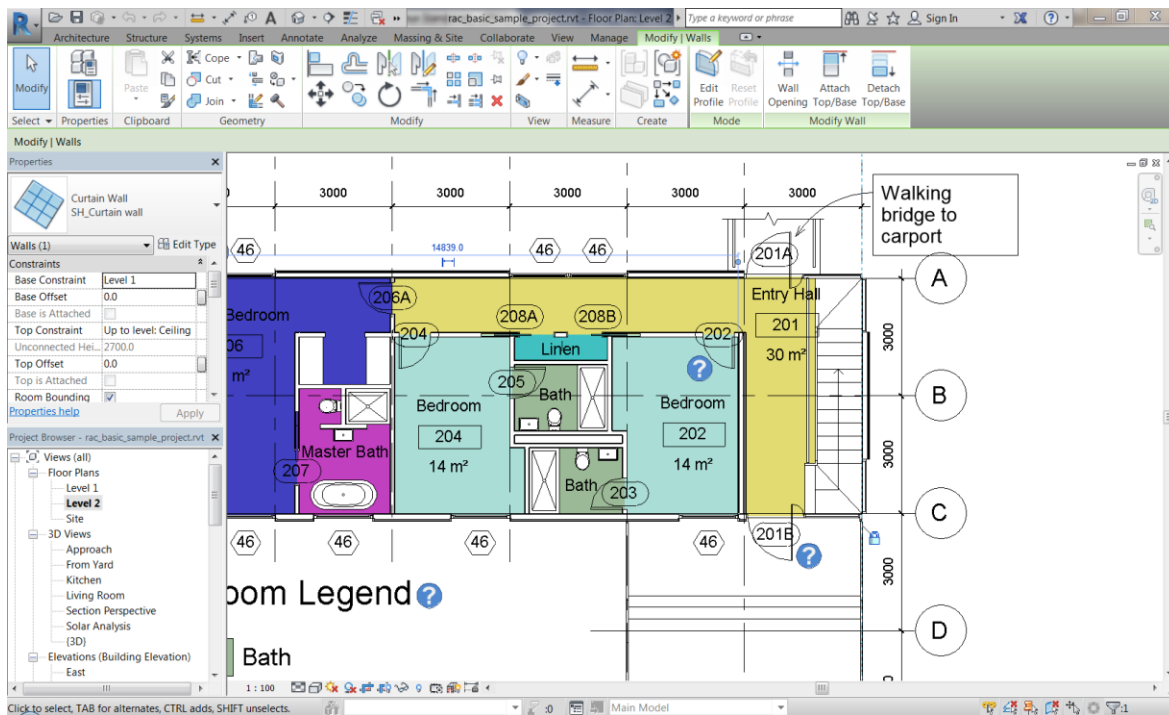


Figure 8. Revit’s floor plan interface. It provides architectural modeling tools for lifting walls, placing windows, doors, and other housing elements.

Revit is a level of abstraction above software like 3DS Max or Blender²⁷. 3DS Max is a pure 3D modelling software where the editing is at the level of the vertex, edge or polygon, providing tools like extrusion, slicing polygons to create new vertices, or UV mapping. The primitives of Revit are compounds of the primitives of 3DS Max. Its primitives are walls, windows, slabs and any other kind of object that is part of the body of knowledge of AEC. These objects are still made of vertices, edges and polygons, however, on Revit they cannot be edited at that level.

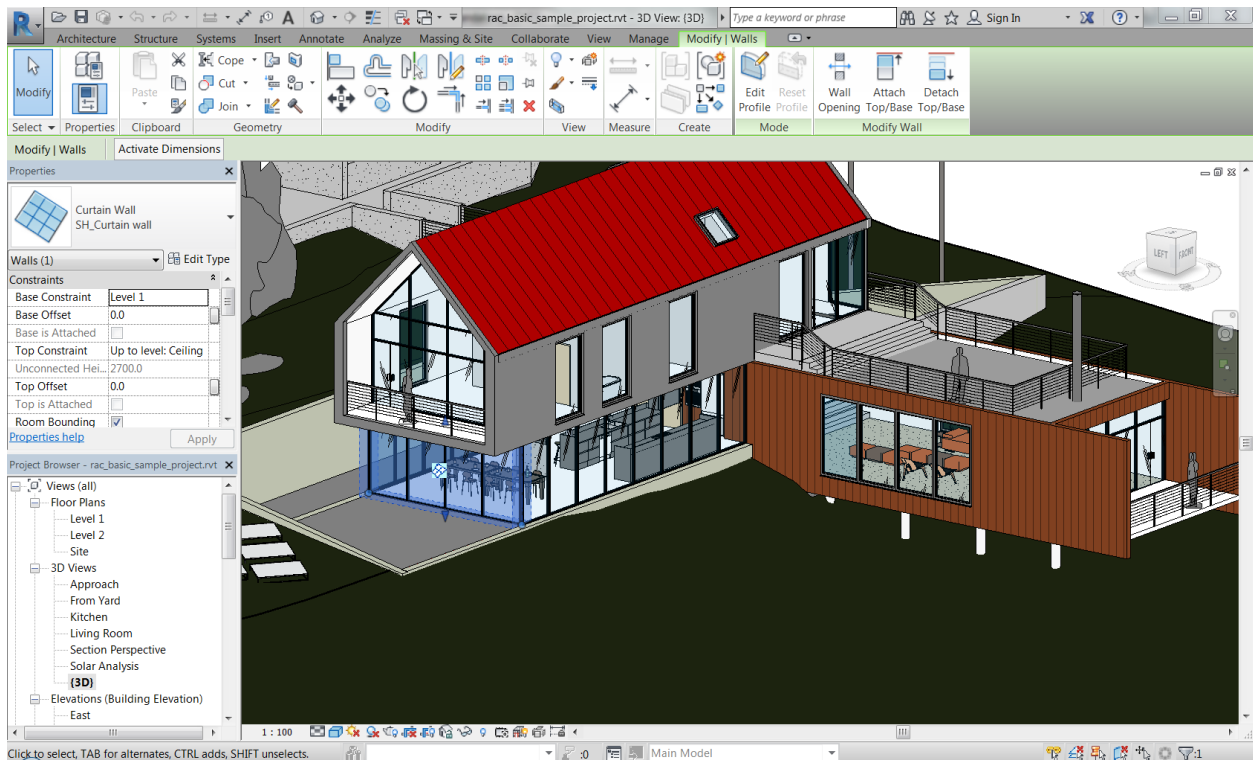


Figure 9. Revit's 3D view. Whatever changes are done on floor or section planes are reflected on the 3D view. This mesh can be exported into a compliant format and consumed by 3DS Max.

This could be dangerous due to the lack of control one has when building these models however, an analysis on 3DS Max of a model produced in Revit shows correctly oriented normal vectors, no gaps, but messy geometry in planes that have openings (like windows or doors) with big groups of edges connecting to the same vertices [Figure 10]. However, even though that geometry is not perfect, it is acceptable and consumable by Unity with no other actions needed. Furthermore, Unity collision geometry [Figure 11] is well computed by Unity by not presenting a topology as messy as the original geometry produced in Revit. This geometry is of the utmost importance due to the computation related to collisions, where reducing the number of primitives may result in a boost

²⁷ <https://www.blender.org/>

of performance.

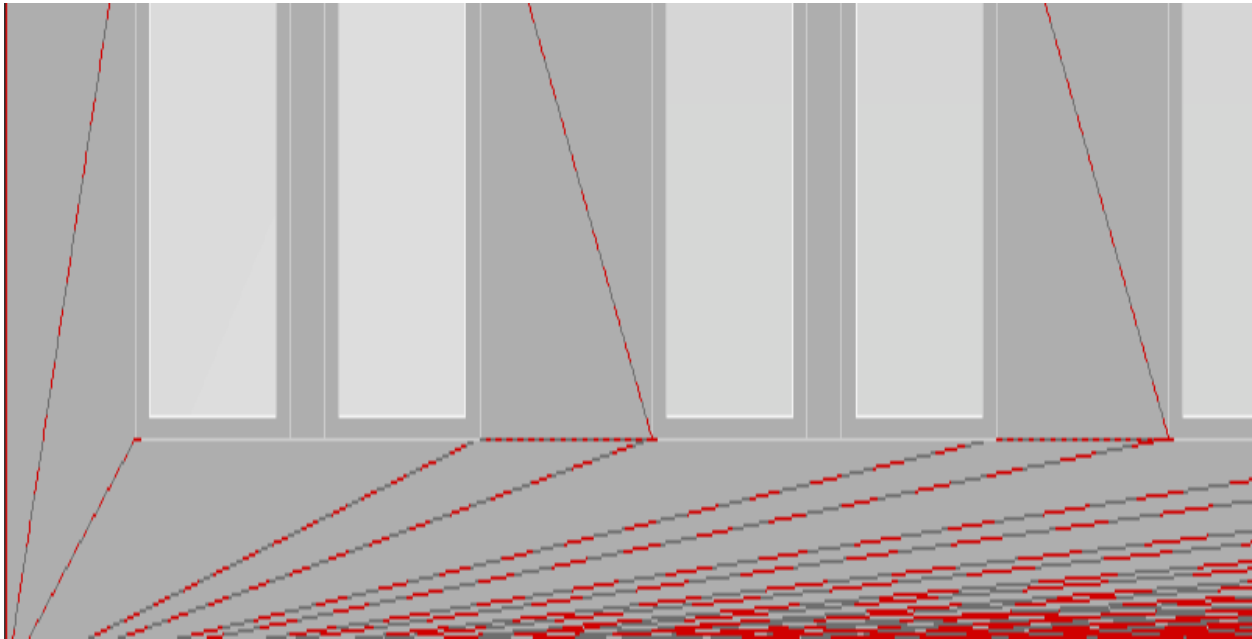


Figure 10. The mesh produced by Revit, seen on 3DS Max. It is not the ideal mesh as many edges share the same vertex.

The main advantage of using Revit over a native 3D modeling software is the swiftness and easiness it offers when producing 3D models of buildings. It requires considerably less time to build this same model on Revit which makes the trade-off between time and quality of the model's topology well worth it. Nevertheless, geometry produced on Revit can be easily exported into 3DS Max or any other native 3D modeling software to have it corrected if need be.

The model was produced with a metric scale in which 1 Revit unit corresponded to 1 meter. The transition from Revit to 3DS Max is not seamless as the scale used in Revit does not correspond to the one used in 3DS Max. When exporting the FBX file from Revit, one needs to specify the unit conversion to meters. Once this FBX file is imported to 3DS we need to convert the units to meters and scale the model to a 30.48 factor and reset the transformations on each object so that we keep the size of the objects but get a scale factor of 1.0 instead of 30.48. Without this step the importation of the model to Unity would become messy. Apart from this unit scaling we also need to select all objects and center the pivot of each object so that we can manipulate them easier. Once done, the model is ready to be exported again to an FBX file that can be properly consumed by Unity where the entire gameplay and graphical environment was developed.

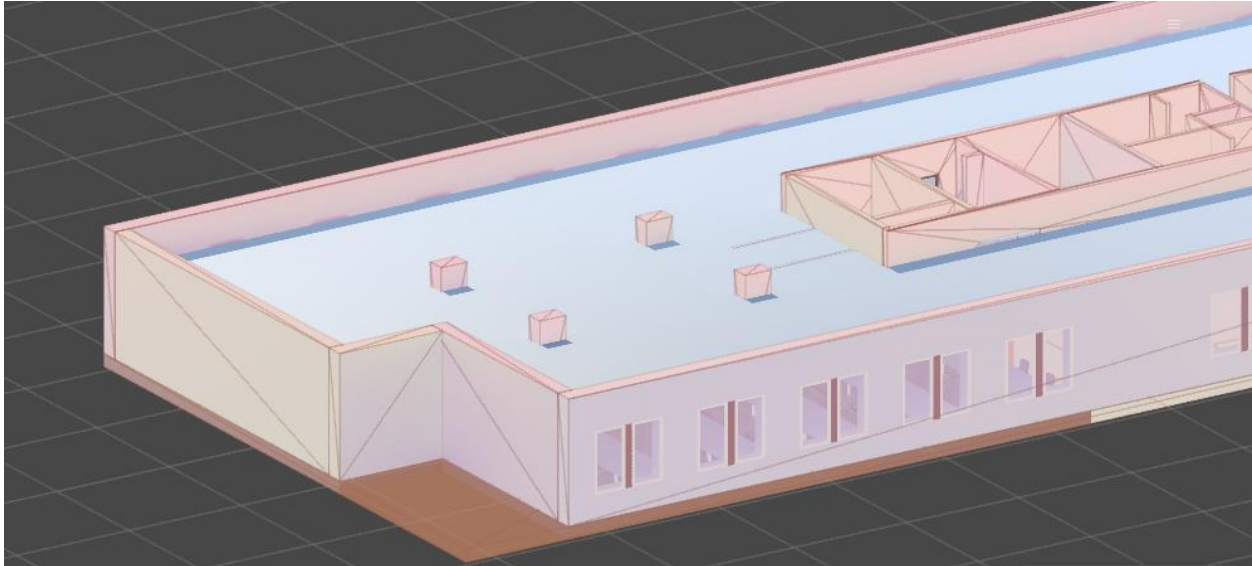


Figure 11. Collision mesh computed by Unity. A bounding box is computed by Unity depending on the type of collision type. The not-so-optimized mesh created by Revit is not reflected on the collision mesh.

Section “3.2 **Serious Game Engines Review**” displays high-level features of Unity based on the work of Vasudevamurt and Uskov (Vasudevamurt & Uskov, 2015), which dates back to Unity 5.6.1. The development of this dissertation was carried using version 2017 for the most part. There are expected differences between these versions, however, these high-level features are still valid.

Networked multiplayer is one of the main features and will be described in the next section. Briefly, the system is designed with a client-server architecture implemented at the expense of the Unity High-level API (HLAPI) (described in the next section). The goal of this architecture is two-fold. First, to allow any user to connect to this environment with a server always up and running. But this could be achieved with a peer-to-peer architecture, however, this could not serve the second purpose, which is to centralize the heavy computing on the server, allowing the clients to run smooth but still taking advantage of computing-expensive features.

4.2 Distributed System Architecture

The application is aimed at virtual visitors accessing the simulation from 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 12] with the heavy computation and plugins on the server side to ensure compatibility and avoid having to compile these toolkits for every different platform (i.e. to build the application for Windows, each plugin would have to be compiled into a Dynamic-link Library (DLL) file, whereas

for Android these same plugins would have to be compiled into an Shared Object (SO) file), saving time and effort on deployment and maintenance. This design choice also assures that the system would depend less on the device specifications, i.e. the specifications would only have to meet the hardware requirements of Unity and the client application would not be slowed down by the processing done on the “3rd-party processing” module on the server side.

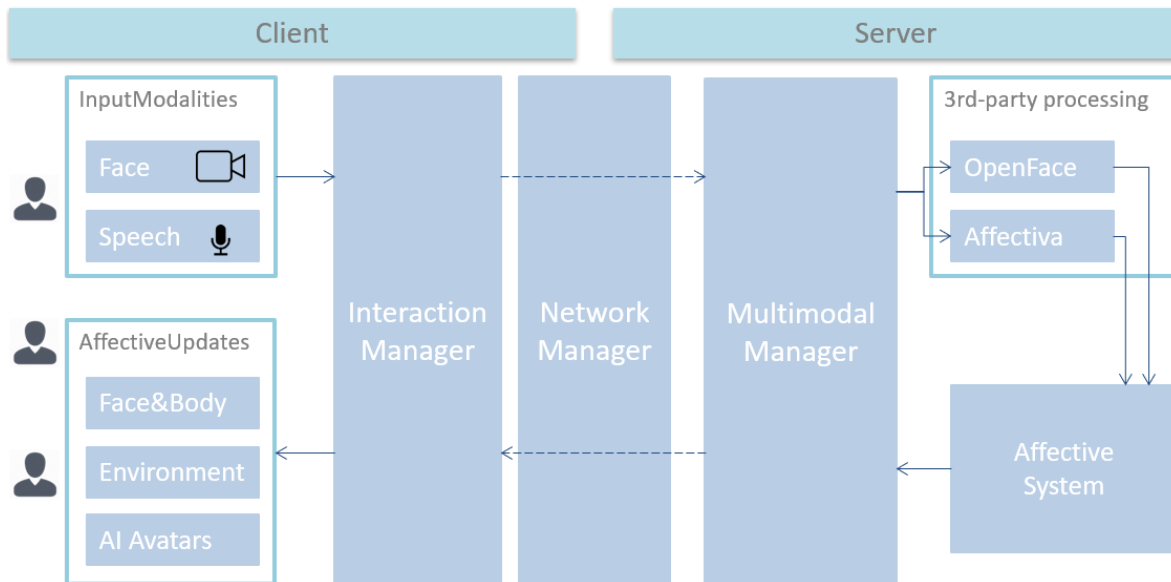


Figure 12. Server-client high-level architecture of the system. The client provides audio and video input to the server which analyses it with 3rd-party software for emotion. The system acts based on this interpretation and broadcasts these changes to every user.

4.2.1. User Guide

The user is first presented with a “Connection Menu” that is composed by two main panels that display options to set the type of connection and Facial Recognition (FR) software to be used. Figure 13.b shows a version of the user interface and portions **a** and **c** show other possible settings of both panels.

The left panel lets the user configure the connection by choosing an Internet or Local Area Network (LAN) connection. The Internet connection (“Internet” checkbox checked) is the default behaviour and the left panel assumes that of Figure 13.a. It uses the Matchmaking interface integrated with Unity’s Multiplayer service with the concept of “Rooms”. A room is an instance of the application that is running somewhere and can be joined by whoever runs the same application and chooses to list rooms available for this application. This synchronization is done through Unity’s Multiplayer Service and once a user asks to list rooms available, a list will appear, and the user chooses which one he/she wants to join. However, a server must be running so that users can

join. A password is required to unlock the creation of rooms and this should be provided only by the application administrator. On the other hand, a user can configure a LAN connection by unchecking the “Internet” checkbox. The left panel becomes displayed as it is on Figure 13.b and new parameters show up to configure this connection. The “IP Address” is the local IP address of the machine that is running the target server application, the “Port” field is the port number where the client will be connecting on the server and the “Peer Type” is the type of user. This “Peer Type” can be a “Client”, “Host” or “Server” and this option is defaulted to “Client”. To access the “Host” or “Server” rules the application administrator must provide a password. If starting as a “Server” peer, the application will be running only with this capability and without any gameplay; if ran as a “Host”, in addition to the server instance, a client instance will also start on the same machine and be automatically registered on the server instance. A “Client” peer only has to fill the “IP Address” and “Port” fields.

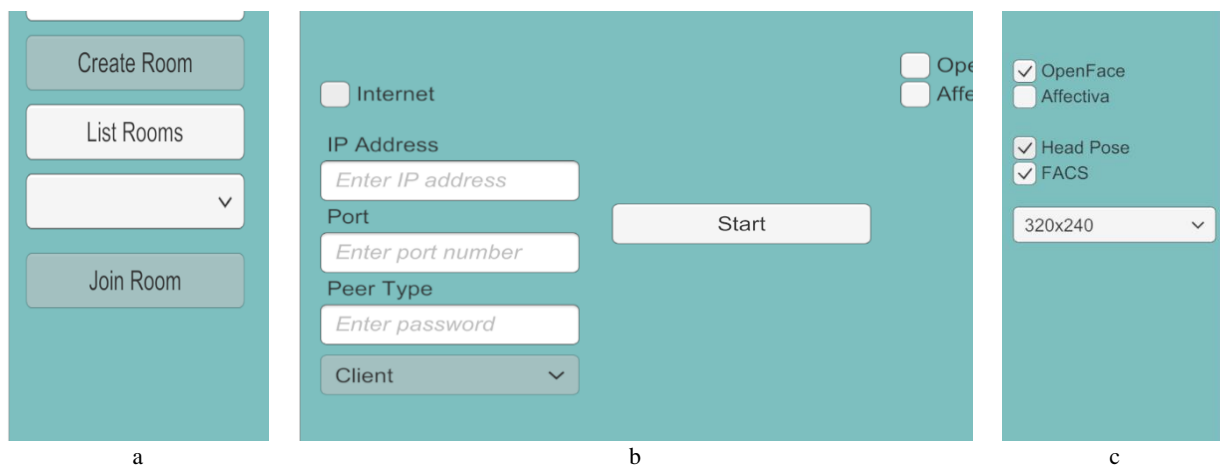


Figure 13. Connection menu that is presented to the user. The left panel offers Internet and LAN connections, displaying a set of different parameters, according to the type of connection. The right panel displays the options for Facial Recognition software and respective features.

On the right-side panel the user can choose the FR software that wants to run (Figure 13.c and right panel on Figure 13.b). Once any of the options are selected (either “OpenFace” or “Affectiva”), new parameters appear. The “Head Pose” checkbox parameter sets if the user’s real-life head orientation will be mapped to his/her avatar’s head and neck orientation. The “FACS” (standing for “Facial Action Coding System”) parameter allows the user to define if action units will be detected and mapped on the avatar’s face. Furthermore, Affectiva performs Automatic Emotion Recognition and any component on the “AffectiveUpdates” module that takes advantage of emotion recognition will be enabled. Once “FACS” is enabled, a dropdown menu also shows up to let the user choose which frame resolution should be captured by the camera input. Higher resolutions

yield better results but also require more processing power.

When the user connects to a server (either on a LAN or an Internet connection) an avatar is spawned and assigned to the connecting client. The client application was tested and developed to be deployed for Android (however, the FR features were not tested thoroughly) and as a Windows standalone application. The user interface [Figure 14] has slight differences between both platforms that accommodate both ways of interaction. The Windows application can be controlled with mouse and keyboard with WASD keys controlling translation and mouse controlling orientation. If “Head Pose” was checked on the connection menu, the mouse controls the avatar’s body orientation (and any translation transformation is applied on the avatar’s body local reference frame) and the head rotation of the user is mapped on the head and camera of the avatar. On an Android device, there are two joysticks on each bottom corner of the GUI that mimic the mouse and keyboard controls of navigation. If “Head Pose” is checked on player settings, the same mechanics apply on Android devices as well.

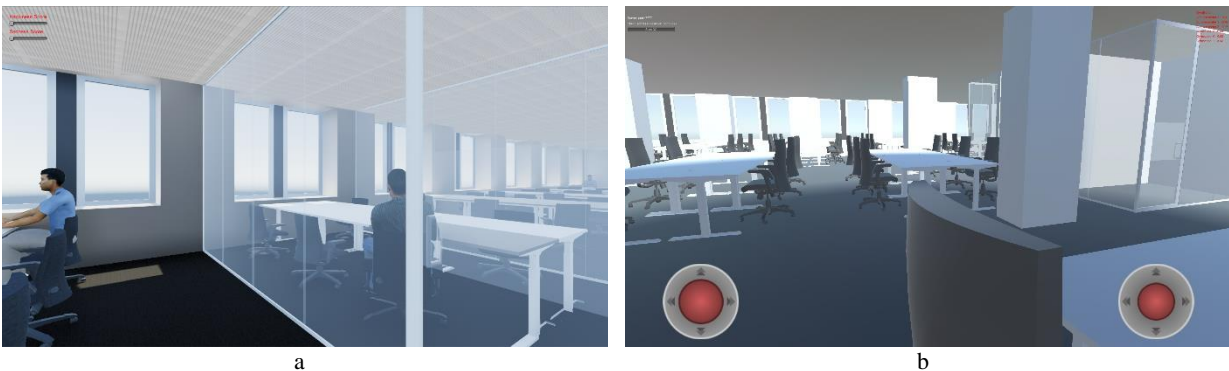


Figure 14. Windows standalone interface on the left and Android interface on the right. The Android app is controlled with joysticks. The red text on the left image represent emotion scores, whereas the red text on the right represents the device’s gyro and accelerometer values.

4.2.2. Unity Networking

Unity contains its own high- and low-level networking APIs. To fully understand the system architecture, an explanation of the High-Level API (HLAPI) concept is due, even though this documentation can be found online²⁸.

This API provides important features that ease the management and implementation of a distributed application. The server is responsible for spawning networked objects and maintain their state synchronization across all client instances. Every time a client tries to establish a connection with the server, the server spawns a default player object that represents that specific client. This

²⁸ <https://docs.unity3d.com/Manual/UNet.html>

player object is replicated across the server and all client instances like Figure 15 shows. Furthermore, each client will have authority over its own player object, which is represented by the thick outlined player icons on each client on the figure, whereas the server has authority over the overall 3D scene. An instance (be it a server or a client instance) having authority over an object means that only that instance can manage that object. These transformations are communicated to the server that synchronizes them with the other clients. Position and rotation are examples of what is synchronized, as well as animation states. Also, individual variables can be tagged for synchronization, which will prove useful to our goal.

However, the 3D scene is not only composed of player objects. Objects that do not change over the gameplay do not need to be synchronized and are simply packed on each client executable bundle. One such example of static objects are walls, the floor or any other inanimate objects that are unlikely to move or change.

Summing up, the HLAPI facilitates the management and optimization of the synchronization of objects between clients. Each networked object has an instance on the server and on each client with a unique identifier that ensures synchronization.

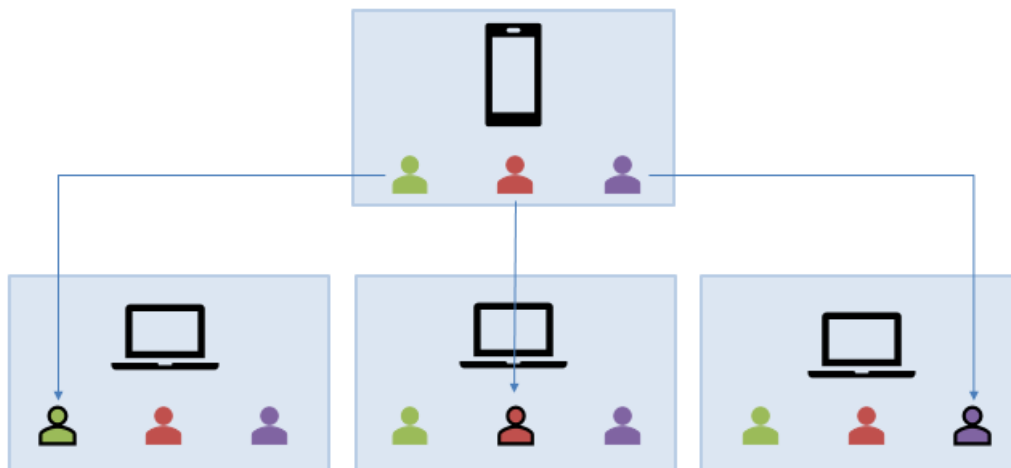


Figure 15. Unity's High-level API broadcasting system. Network-tagged components automatically synchronize with server and other clients. The server and each client have a local instance of every other user.

4.2.3. Component Description

The client modules (“Input”, “Output”, “Navigation & Animation”) can be found on the left side of the scheme and the server modules on the right (“Third-party Software”, “SMultimodal-Manager”, “SGameManager”, “AffectiveSystem”). There are three main components: the “Interaction Manager”, the “Network Manager”, and the “Multimodal Manager”. Figure 16 displays a class diagram that is the implementation of the architecture presented on Figure 12.

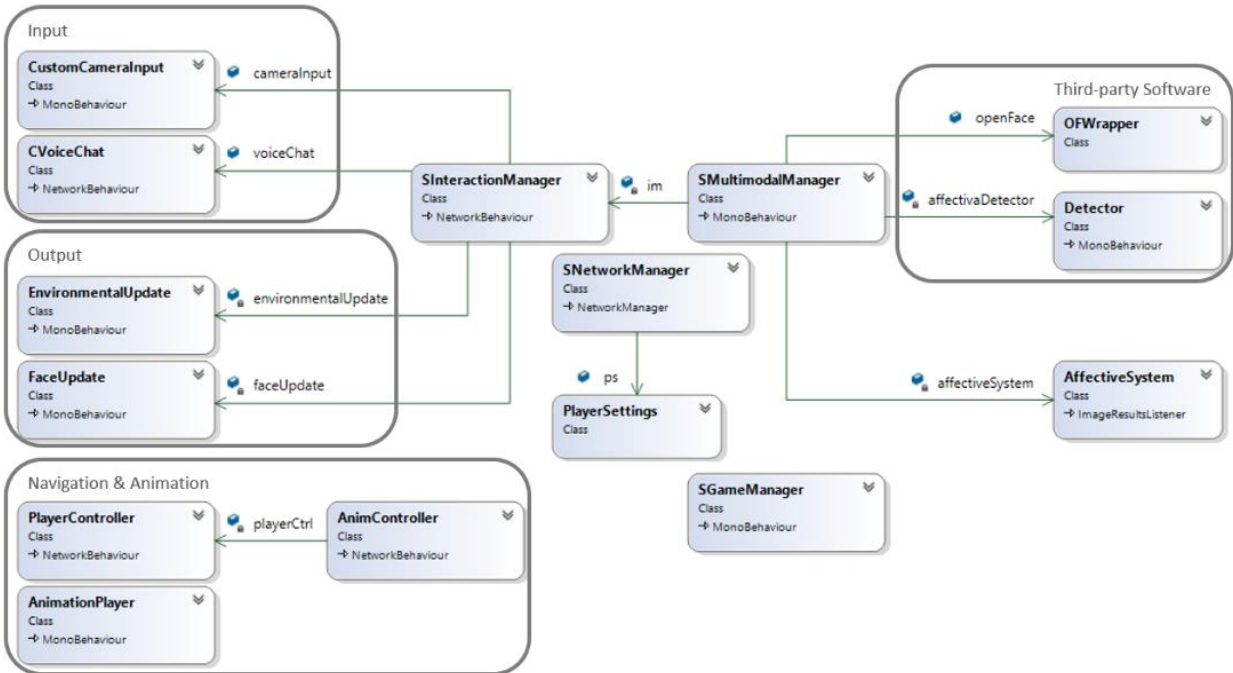


Figure 16. Class diagram. The “SInteractionManager” contains references for instances of the “Input” module classes and channels to the “SMultimodalManager” for processing and interpretation. This data is synchronized back to the “SInteractionManager” and sent to the “Output” classes.

The “Interaction Manager” is the 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 the microphone, mouse & keyboard interaction, and touch, depending on the platform. When the user finishes setting the player settings on the “Connection Menu”, the “SNetworkManager” of the client instance sends a “PlayerSettingsMessage” to its counterpart on the server instance. There is an “SGameManager” on the server that is responsible for instantiating and managing all players in the scene. It is also this component that adds an “SMultimodalManager” to each player that, in turn, possesses a reference to the “SInteractionManager” on the server side. The “SInteractionManager” possesses a set of synchronized variables (tagged as “[SyncVar]” on Unity) that allow seamless bidirectional updates through Unity’s Networking API. The “SInteractionManager” collects data from the “CustomCameraInput” and “CVoiceChat” and sends it to the server by sending messages that extend Unity’s “MessageBase”²⁹ class. The “SMultimodalManager” is listening to these messages and triggers callback events upon receiving them. They are then channeled to the right “Third-party Software” and analyzed. The “OFWrapper” is a C# OpenFace wrapper to expose some methods and allow it to be integrated into Unity and the

²⁹ <https://docs.unity3d.com/ScriptReference/Networking.MessageBase.html>

“Detector” is a class that implements Affectiva’s interface to make API calls. After getting processed data from the “Third-party Software”, the “SMultimodalManager” calls the “AffectiveSystem” that interprets this data to inform the “Output” classes on the client-side. This is done through the “SMultimodalManager” that contains a reference to the “SInteractionManager”. The synchronized variables are updated on the server instance which is automatically synchronized on the client side. The “EnvironmentalUpdate” and “FaceUpdate” are constantly consuming the synchronized variables, so any changes that occur on the server instance will be instantly reflected on the client.

The “SNetworkManager” is the core component of the distribution of the system. As explained in the previous section, it is responsible for synchronizing all networked objects. Besides the synchronized variables, the “PlayerController” and the “AnimController” also extend the “NetworkBehavior” class. Unity has a special way of dealing with synchronizing player transformations and animations. The player object contains three important special components that deal with this synchronization: the “Network Identity”³⁰, the “Network Transform”, and the “Network Animator”. The “Network Identity” provides the game object with a unique identifier across the entire network to eliminate ambiguity between similar objects. The “Network Transform” uses this identifier to synchronize the player’s rotation and translation from its owner’s client instance with the server and broadcast it back to all other clients. This component is responsible for the player’s overall transformations and the “Network Animator” applies the same procedure over animation states that produce local transformations on the avatar’s skeleton.

The third-party processing module contains off-the-shelf plugins that can contribute to the “AffectiveSystem”. For this project, only FR was integrated with two tools, the Emotion SDK from Affectiva and OpenFace. These tools are described in section “3.1 Automatic Emotion Recognition”. The idea behind this architecture is that the Multimodal Manager channels other kinds of input (like speech audio, body posture with pressure chairs, or any other modalities one would like to integrate) to third-party tools that return recognized emotions and expressions. This data can then feed the Affective Model, building more robust outputs.

4.2.4. AU Detection and Facial Emotion Recognition

Emotion and FR are important themes to this dissertation. Its contributions revolve around the

³⁰ <https://docs.unity3d.com/Manual/class-NetworkIdentity.html>

detection or foresight of emotion, facial expressions or head orientation. We've tested Affectiva and OpenFace for this effect and both provide the detection of facial landmarks, Action Units (AU), and head pose estimation. In addition, Affectiva also provides AER based on facial features and OpenFace provides gaze estimation. The detection of facial landmarks is foundational for the other features, but we won't use this specific feature for our contributions, so we will rather focus on head pose estimation, AER, and AU detection.

One of the main requirements to test H3 is head pose estimation, a feature provided by both parties. As such, a test was carried to evaluate which provides the best estimation, in terms of robustness to noise and amplitude of face detection. We ran tests with both solutions under the same conditions that yielded the results presented on Figure 17.

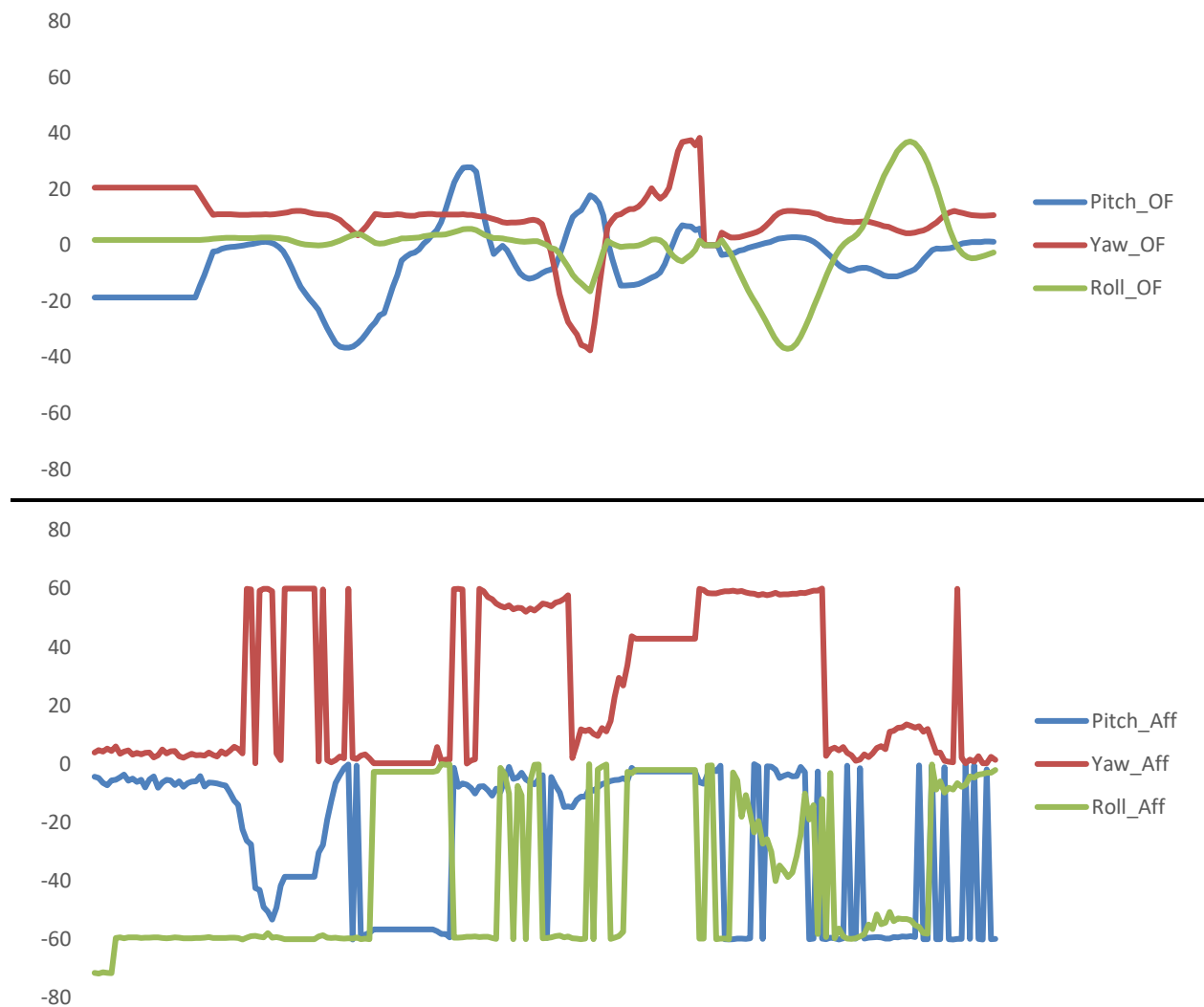


Figure 17. OpenFace and Affectiva's head pose estimation. OpenFace's head tracking is more robust as it is harder to lose head track than Affectiva's. The Dirac's on Affectiva's chart represent track loss.

The charts represent the values of pitch, yaw, and roll collected over time under the same conditions for OpenFace and Affectiva. OpenFace shows more robust results on head tracking robustness to loss of tracking and noise, as it yields softer curves and does not display the Dirac's seen on Affectiva's. The Dirac's occur when the software loses head tracking. Likewise, when Affectiva's able to track the head, it produces more noise than OpenFace. Figure 18 shows the display of OpenFace and the respective avatar with inner eye brow AU and head orientation estimation matched. We resort to this head orientation estimation to estimate the user's engagement to test H3 and the AU mapping is part of the general CVE of this dissertation.

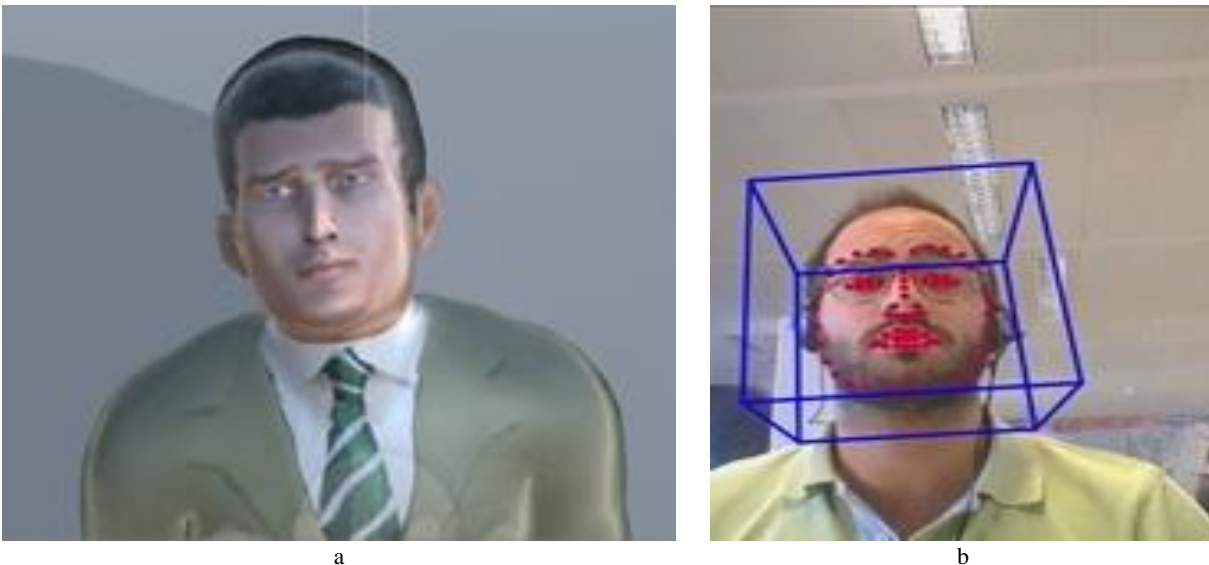


Figure 18. Landmark detection and head orientation estimation. On the right is OpenFace's display of head pose estimation with a blue bounding box and facial landmarks in red. At the left the avatar's head rotation and inner upper eyebrows are matched with those of the user.

4.3 Summing Up

This chapter approached the planning and development of the general CVE that establishes a framework for the contributions of this dissertation. It provides a distributed and emotion-driven framework upon which specific 3D applications can be built. The architecture follows a server-client design where most of the workload is done on the server to free up the client to be ran on mobile or desktop across multiple devices and platforms. However, this can also require a lot of centralized processing power if the number of users scale up quickly. As a future endeavour, this architecture could be refactored to allocate the emotion processing on the client, once mobile devices specifications are advanced enough to support this workload. The distribution is built upon Unity's High-level API that concentrates emotion processing on the server.

Autodesk Revit is a well-known software to the Architecture, Engineering and Construction fields. It accelerates the 3D modeling of architectural models, freeing up time for the implementation of other features. The game logic is provided by Unity which proved to be a good decision, confirming the reasons that made us choose it, enunciated in the previous chapter. Not every feature was able to be tested in every target platform, but instead the development and testing were focused on the Windows standalone application, since the final usability tests were going to be carried in this platform.

Chapter V – Confusion Prediction

This chapter is replicated from our publication (Silva Pedro, Luís Silva, & Pereira, 2018) that resulted from the work of this dissertation.

To our knowledge there are no publicly available datasets that serve our purpose, so we set out to build one to test H1. Our objective was to have a dataset composed of training examples labelled with a confusion level and a set of features that could syntactically and lexically describe these examples, as well as produce n-grams. To achieve this goal, three tasks were required: collect and process the corpus of the dataset, extract features for each training example and label them.

5.1 Methodology

The collection was carried by means of manual web scraping where 39 presentation transcripts from various fields of knowledge were collected. These transcripts contain full presentations and we wanted our classifier to have a granularity of excerpts with at least 50 words, but the closest possible to this number. With the NLTK package for Python³¹ we split these transcripts on text excerpts with at least 50 words, while keeping the full sentences (that is, splitting in the next ending punctuation after the first 50 words, resulting in text excerpts with varying lengths but as close to 50 words as possible). The splitting of these 39 transcripts resulted in a pool of 600 text excerpts from which we pulled 300.

The next task consisted in extracting features from these text excerpts. We resorted on the Lexical Complexity Analyzer (LCA) (Lu, 2012) and L2 Syntactical Complexity Analyzer (L2SCA) (Lu, 2010) available at <http://aihaiyang.com/software/> at the date this dissertation was written. These are web-based tools that require only text strings as inputs and provide several lexical and syntactic measures. We opted to use this tool because it provided the wanted features without having to produce any code or delve into other toolkits. The output of these strings is comma-separated values (CSV) files with the selected features for each string.

Finally, the last task was to collect classifications of confusion for each excerpt. We asked 51 annotators to classify 30 excerpts each, which gave a total of 1500 valid annotations (the annotators could skip some text excerpts if they wanted to). This sample was composed of 41 (80.39%) male and 10 (19.61%) female annotators. An application was provided for each annotator to classify

³¹ www.nltk.org/, accessed 26th September 2018

these excerpts. A first slide introducing the context and goal of the task was presented to the annotator, stating the task was anonymous and only his/her answers would be recorded. Next, a more detailed description of the task was presented, stressing the task was not about the annotators reading or comprehension skills, but rather about the complexity of the test itself. One challenge with which we were faced was to clarify that the confusion evaluation was only about the syntactic and lexical shape of the text, and not about the content of the text itself. In this description slide it was also stressed that some of the content may not be familiar to the annotator, but to try to rate anyway, disregarding the content and focusing on how well he could read and understand the text in terms of the complexity of its sentences and words.

Next, a slide provided the annotator with a set of instructions for the task. He/she had a slider ranging from 0 (not confusing) to 6 (very confusing) to rate the excerpt, a button to confirm it, another to press in case it was not possible to detach the confusion state from the underlying content of the excerpt, and a skip button to skip any excerpt. A voice was reading every excerpt and the button mentioned above are unlocked after it finishes reading. The annotator could exit the task any time he wanted to or pause it while the voice was reading the excerpt, in that case the reading stopped, and the excerpt was hidden. After the instructions were given, a screenshot was shown so the annotator could get himself accustomed with the interface, shown on Figure 19.

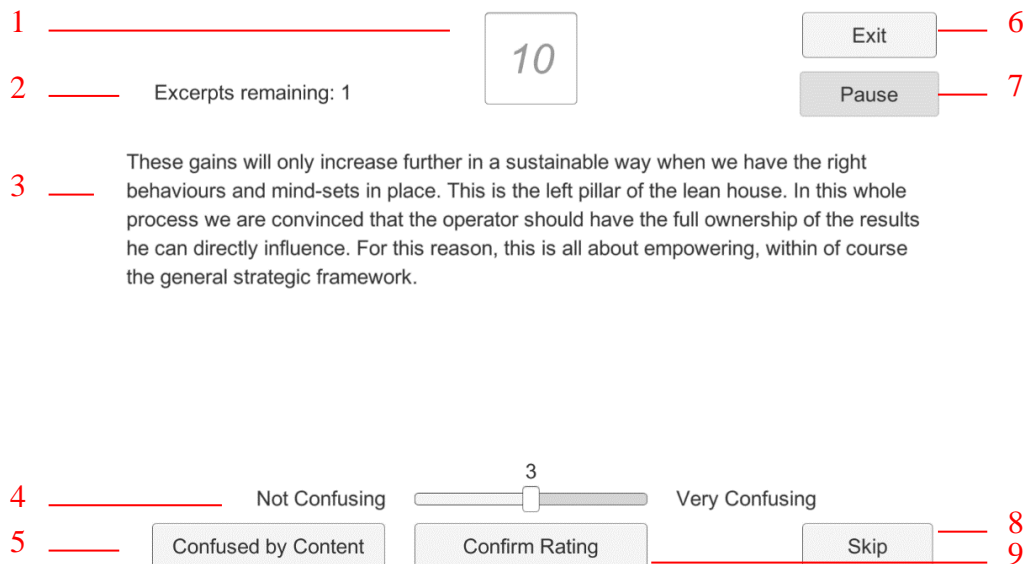


Figure 19. Interface to rate text excerpts. 1 – 10-sec timer to rate the excerpt; 2 – Count of number of excerpts remaining to rate; 3 – Text excerpt body; 4 – Slider to set the confusion level and labels to establish the scale; 5 – Button to confirm that it was not possible to decide if the confusion felt was due to content or shape; 6 – Button to exit the application; 7 – Button to pause the text-to-speech; 8 – Button to skip an excerpt; 9 – Button to confirm the rating set on the slider.

Before the task started, the annotators were asked if they understood the separation between syntactic and lexical text complexity and the content. From the whole sample, 46 (90.20%) stated they understood, 5 (9.80%) said they somewhat understood, and none reported they did not understand. The ones who reported they only somewhat understood were clarified until comfortable with this definition. It is noteworthy to say that all the annotators had a technological background. This detail allowed them to be easily accustomed with the application due to the intrinsic contact with digital applications, or due to the typical English language skill that is required to work in this field. However, this characteristic can introduce bias in the sample. During the task, a synthesized voice (either male or female) was reading the excerpt. When it finished the reading, the 10-sec timer was unlocked, and the annotator could classify.

After performing the task, the annotator reported if he/she found it boring. Most of the annotators, 39 (76.47%), reported they did not find the task boring, 10 (19.61%) found it somewhat boring and only 2 (3.92%) said it was boring. This can assure us that there was no significant bias on classifications due to the repetitiveness of the task. The confusion levels ranged from 0 to 6, where 0 corresponded to no confusion reported, and 6 to maximum confusion. Besides these levels, there was also the “Confused by content” button where, in case of confusion, the annotator could not decide if the source of confusion was the content or the lexical and syntactical complexity.

5.2 Dataset description

5.2.1. Sample description

The task described in the previous section resulted in a dataset composed of 300 English text excerpts classified over 8 different categories: from 0 to 6 confusion or “confused by content” (CC). The ratio of skipped excerpts is around 1.9%, which is a good sign that annotators engaged well with the task and confirms their feedback about the boredom of the task. Also, the vast majority of them are triggered by exceeding the time to answer (only one skipped excerpt was purposely skipped), which may be due to indecision or the learning curve of the application mechanics.

Due to the subjectivity of what a level of confusion is and the size of the dataset, we decided to condense the answers into four categories:

- “Low confusion” is considered when there is a majority (three or more annotations) of 0 or 1 ratings,
- “Medium confusion” is considered when there is a majority of 2 or 3 ratings,

- "High confusion" is considered when there is a majority of 4 to 6 ratings,
- "Confused by content" is considered when there is a majority of 'CC' ratings.

Furthermore, when there was no agreement between at least three annotators, the excerpt was classified as having no agreement. In fact, there is agreement for only 67.7% of the dataset. Our ideal goal is to achieve near-human accuracy when classifying text excerpts. For such a subjective task we would not have the means to assess if the model was performing well, motivating us to exclude the excerpts that did not meet this criterion of having agreement. In addition, for now we cannot hope that a model would be able to surpass the human skill in this case.

As can be stated from the bullet points above, we decided to give a higher weight to “High confusion”, and the motivation for this is two-fold. First, due to the imbalance of the dataset where, even with this weighting, the classes described above are represented in the dataset approximately by the percentages presented on Figure 20, after dropping the excerpts that did not meet the agreement criterion. Secondly, because we consider it is more important to solve highly confusing excerpts than oversimplifying them.

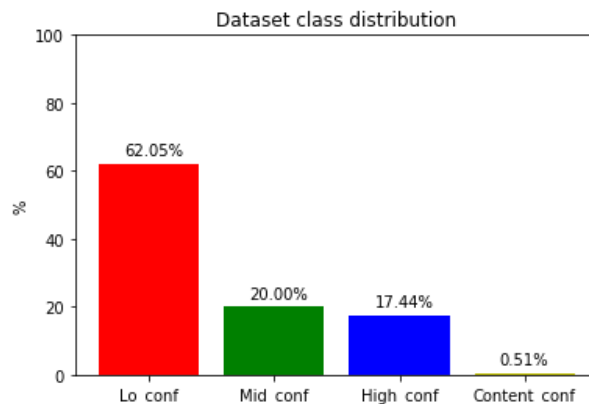


Figure 20. Chart representing the class distribution of the dataset. With 62.05% the “Low confusion” class is the most represented and unbalances the dataset. “Medium confusion” and “High confusion” are equally balanced. “Confused by content” classifications are negligible.

From an optimistic perspective these values can tell us that the annotators understood well what the task was about. However, we cannot discard the possibility of its misinterpretation and giving high values of confusion due to content and not text complexity. Due to the negligibility of the “Confused by content” presence, we decided to remove those excerpts from the dataset, for their presence would only increase the complexity of the problem. The Fleiss’ kappa coefficient (Fleiss, 1971) obtained from this sample ranks in the “Slight agreement” segment with only 0.16 (Landis & Koch, 1977). This is not necessarily bad, but rather translates to a hard problem since even

among excerpts that collected a majority (3 or more) of same categorical annotations there is only shy inter-annotator agreement.

Due to its size, 80% of the dataset was taken for purposes of training with a 10-fold cross-validation and the remnant made the test set. The two sets were similarly balanced class-wise according to the distribution given in Figure 20. Feature scaling was performed to place the values between 0 and 1 and not biasing the learning towards features with higher scales.

5.2.2. Text complexity features

The LCA and L2SCA provided us with syntactic and lexical features. Coh-Metrix (Graesser, McNamara, Louwerse, & Cai, 2004) is a similar syntactic complexity analyzer that was developed for English native speakers. We have chosen L2SCA over Coh-Metrix because in our final use case, the producers of written text are not English native speakers and L2SCA was developed towards this audience. In addition to these, we vectorized each excerpt into unigrams, bigrams and trigrams, and kept those that occurred more than certain lower and upper thresholds in the entire corpus. We tested our model with unigrams up to trigrams because higher dimensional n-grams are known to become so sparse that it renders themselves unusable (Allison, Guthrie, & Guthrie, 2006), especially in small datasets like ours.

From a total of 56 features we considered 40. The ones we left out are simpler features that compose other more complex ones (i.e. two of the left-out features are the number of words (W) and the number of sentences(S), and a more complex feature that we included is the mean number of words per sentence, which stands for “Mean Length Sentence”, MLS) which would render them redundant. We performed automatic recursive feature elimination (RFE) with cross-validation using the *scikit-learn* machine learning toolkit for Python³² which left us with the set of features that is described on Table 8.

In addition to these features, several bigrams were also considered by the RFE algorithm. The choice of using bigrams rather than uni- or trigrams was made by performing RFE with the lexical and syntactic features plus each set of n-grams separately with bigrams yielding the best results. These results will be presented in the next section, as well as the final results from model selection and final test scores.

³² <http://scikit-learn.org/>, accessed 26th September 2018

Table 8. Features selected by the recursive feature elimination algorithm. For further explanation of these features refer to (Lu, 2010, 2012).

Feature	Description	Alias
Mean Length of T-unit	# of words / # of T-units	MLT
Verb phrases per T-unit	# of verb phrases / # of T-units	VP/T
Dependent Clauses ratio	# of dependent clauses / # of T-units	DC/T
Sentence coordination ratio	# of T-units / # of sentences	T/S
Lexical Sophistication I	# of sophisticated lexical words / # of lexical words	LS1
Verb Sophistication I	# of different types of sophisticated verbs / # of verbs	VS1
Number of Different Words (expected random 50)	Mean T of 10 random 50-word samples	NDWERZ
Noun Variation	# of different nouns / # of lexical words	VV2
Adjective Variation	# of different adjectives / # of lexical words	AdjV

5.3 Evaluation and Discussion

Recursive feature elimination (RFE) is a technique in which a selected estimator is tested recursively against various subsets of a feature set, eliminating features in each recursion until reaching a subset that optimizes the selected error metric (i.e. maximizing accuracy or f-score on classification, or minimizing mean squared error on regression). We chose f-score (Chinchor, 1992) as the error metric for the same two reasons we have given higher relevance to the “High confusion” class: the imbalance of the dataset and valuing the detection of highly confusing excerpts over the less confusing.

Figure 21 presents f-score values for RFE with cross-validation for each feature space (lexical and syntactic features with unigrams, bigrams or trigrams). Bigrams hold the highest average f-score across all classes when compared with the other n-grams. It also holds the highest individual f-score for the “High confusion” class based on the same comparison. These two facts led us to include bigrams in the feature space rather than uni- or trigrams. Even so, class imbalance takes its toll on the f-score as this score is higher for “Low confusion” than for any other class across all

n-grams, even with weights favoring both “High” and “Medium confusion” classes. As we would want f-score values to be as close as possible for each n-gram, trigrams are clearly more spread which means the weights did not work as well in this case. Trigrams are widely used but it did not fit as well in our dataset, probably due to its size. With small samples, trigrams tend to be sparser and not perform well. However, we can hypothesize that its performance may increase if more data is available since it can capture more syntax. The algorithm we used to perform the RFE was a linear support vector classifier (SVC) with balanced class weights due to the dataset imbalance, regularization term of 1.0, gamma inverse to the number of features used to fit the algorithm, tolerance for the stopping criterion of 0.001 with a ‘one-vs-all’ decision function.

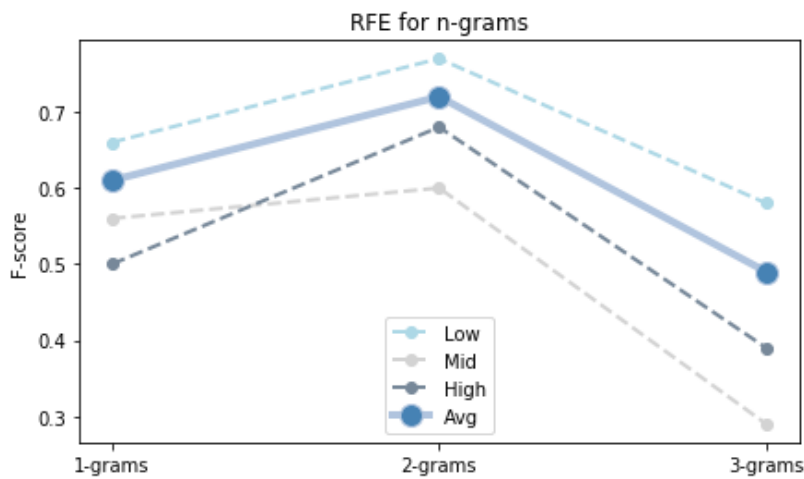


Figure 21. Each n-gram is represented on the x-axis with values for each class of confusion, as well as their average. The average is higher for bigrams (2-grams) and the individual f-score for “High confusion” is also significantly higher for bigrams, in this case surpassing even “Medium confusion”.

Figure 22 shows the results of grid search with cross-validation over several algorithms. Multilayer Perceptron (MLP) and other instance of SVC (different from the one used to perform RFE) yielded the best results. Both algorithms seem to be overfitting, especially the MLP with its training f-score of 1.0 and significant difference between training and validation f-scores reinforcing that. However, when ran over the test set, it performed better as this f-score was closer to the validation f-score than those with SVC. SVC seemed less overfitted as it only has about 0.90 f-score for the training set and 0.72 on the validation set, but it performs worse in the test set.

The MLP was trained with an identity activation function, alpha equal to 0.001, 4 hidden layers with 20 nodes, constant learning rate initialized as 0.01, and 500 max iterations until convergence. The linear SVC was trained with a regularization term of 0.1, ‘one-vs-one’ decision function shape, and gamma of 10.0.

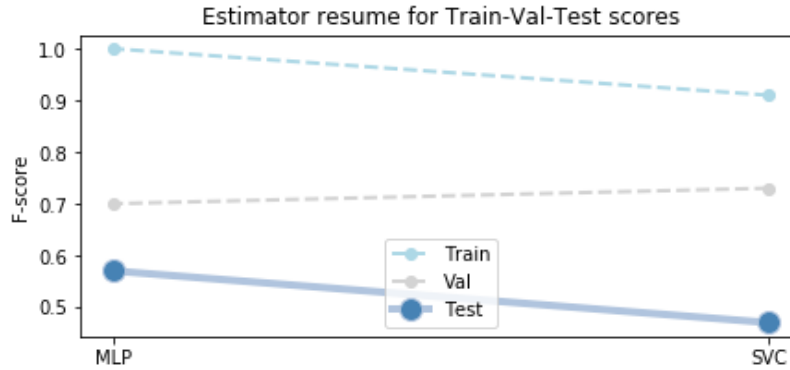


Figure 22. Chart with train, validation and test f-scores for each estimator. The test f-score of MLP was higher than SVC, however, one could expect the opposite as the train and validation f-scores of the SVC seemed more robust and less overfitted.

A finer look to the test results [Figure 23] shows that “High” and “Medium confusion” also have higher individual values of f-score, thus aligning with our objective of maximizing these values. With an average f-score of 0.47, SVC resulted in 0.57, 0.38 and 0.17 f-scores for the “Low”, “Medium” and “High confusion” classes, respectively. On the other hand, MLP yielded an average f-score of 0.57 with individual f-scores of 0.70, 0.38 and 0.33 for “Low”, “Medium” and “High confusion” classes, respectively. When compared to SVC, MLP yields significantly higher f-score for the “High confusion” class which aligns with our goal. Still, “Low confusion” f-score is significantly separated from the rest.



Figure 23. Test stage results. “High” and “Medium” confusion show higher values of f-score, aligned with the goal of maximizing them.

This study represented a first effort on trying to predict confusion from transcripts produced from spoken text with lexical, syntactic, and n-gram features. The results are not near the performance of other text-related problems however, the calculated Fleiss’ kappa coefficient reminds us that we are not facing an easy problem.

5.4 Summing Up

In this chapter we present the methodology and results of the testing of H1. Fifty-one annotators were asked to rate 300 text excerpts which resulted in 1500 valid annotations. A machine learning model was trained with this data after exploring and preprocessing the data with RFE. A grid search revealed that an MLP and an SVC were the models that yielded the best results. We extracted lexical, syntactic, and n-grams features from the text excerpts to feed this model and performed RFE over them to select the best-performing features.

This was an important stage of this dissertation that allowed the test of H2 based on the validation of H1 described in this chapter. Even though the test score was not near perfection, this is meant to be an assisting tool to the human and not a fully autonomous one. Typically, these models require larger datasets than ours and this would be an obvious aspect to improve the model.

Chapter VI – Case Study

A Presentation (PrE) scenario was built on top of the general CVE developed in the context of this dissertation. This chapter describes this scenario that served the experiment that tested H2 and H3. The case study simulates a scenario where the user is a virtual visitor coming into the Company's office and watching a presentation about a Company's project. The presentation consists of an avatar presenting a part of this dissertation, however, we made sure there were no hints about the purpose of this user evaluation.

6.1 Methodology

To test both hypotheses, we conducted an experiment under three different conditions. These conditions allowed us to test both hypotheses by comparing the differences of the results between conditions. These three conditions are:

- Condition I. The presentation with an original script,
- Condition II. The same presentation as in Condition I but with a slightly different script, where this text was ran through the model described on Chapter V and rewritten to report lower confusion levels, and
- Condition III. The same presentation and script as in Condition II, 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.

H2 was tested from the results of Condition I and Condition II, where the variable was the script that was spoken by the avatar. The script for Condition I was originally written without any analysis, whereas the script for Condition II was a redesigned version of the original one. The original version was analysed by the machine learning model the previous chapter 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 Condition II setting the variable to be tested. On Condition III we kept this redesigned

script but also introduced an automatic system that changes the lighting condition according to the user's detected engagement. The full source of engagement is highly debatable. For this experiment we consider the user is more engaged as his/her head is more directly facing the screen. In other words, if a vector can be cast with the user's head orientation (i.e., a vector aligned with the XZ-plane with positive direction) and has the opposite direction (i.e., a vector aligned with the XZ-plane with negative direction), then the user is considered to be fully engaged. On the other hand, if both vectors have the same direction and aligned with the same plane, the user is considered to be disengaged. The results from the pair of Condition II and Condition III allowed us to test H3.

6.1.1. 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, which totalizes 45 valid users, 14 female (31.11%) and 31 male (68.89%) users. Figure 24 shows that the mean age across conditions is similar, with a variance of 2.15 between Condition I (SD = 6.36) and Condition II (SD = 5.39), and between Condition II and Condition III (SD = 6.34).

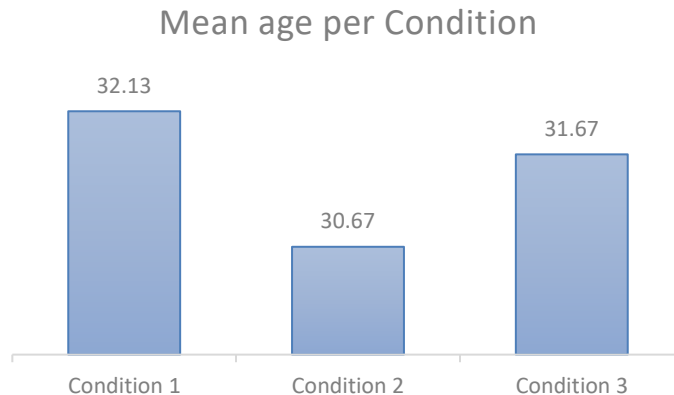


Figure 24. Mean age of the group of users per condition. The means are similar across all conditions with a variance of 2.15 between Condition I and II, and Condition II and III.

Figure 25 shows the level of education of the users, with a predominance of users with a master's degree, especially in the last condition. However, all users work on the Information Technology field, so we believe the level of education did not have a significant impact on results. We also asked every user about any hearing problems because that could affect the head tracking. One pilot test showed that a user with hearing problems can unconsciously rotate the head with the one ear towards the screen, as if trying to hear better. However, only one user reported hearing problems

and during the test there were no reported problems of listening to the presentation.

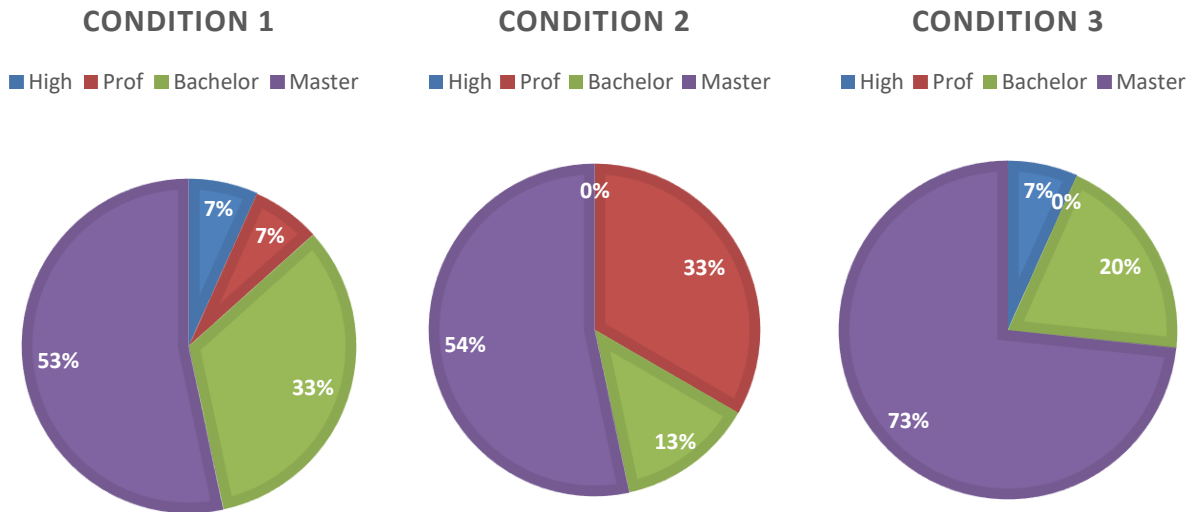


Figure 25. Level of education distribution for each condition. There is an abundance of users holding a master's degree. Users holding a bachelor are always present and there is a strong presence of users with professional education on Condition II.

6.1.2. Presentation Script

A script was written for each presentation slide (can be found on Appendix A) to be spoken by the avatar that is presenting. The script for Condition I was an original one, produced by the author of this dissertation. It was composed of nine different files, one for each slide. The version of the script for Condition II and Condition III was slightly different, after being ran through the machine learning model described on the previous chapter.

Each file (corresponding to the text excerpt of each slide) followed the same procedure as the excerpts from the previous chapter and 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 confusion interpretation. We also pulled the features from this analysis to further investigate and understand which parts should be rewritten.

Table 9 shows this data containing an "ID" field identifying the text excerpts produced from the nine script files, nine fields that correspond to the ones displayed on Table 8 on the previous chapter, and a final one ("conf") that corresponds to the classification of confusion (0 = low confusion, 1 = medium confusion, 2 = high confusion). We analyzed this data to check which parts of the script had should be rewritten and why they should be rewritten, according to the most relevant features for these parts. However, we did not follow this religiously, 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 give help us redesigning the script and provide us with a guideline on what to rewrite. We wrote the script trying to not over-complicate it to get clearer results and stay faithful to what would be script written in a normal context. Instead we tried to write a natural script and thus we were not expecting the classification model to report many highly confusing parts.

Table 9. 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	0.041334	0.111111	0.15	0.68	0.518519	0.15	0.597222	0.794872	0.090909	0
1	0.160572	0.222222	0.26668	0.2	0.462963	0.2	0.819444	0.871795	0.393939	0
2	0.22019	0.277778	0.2	0.2	0.277778	0	0.631944	0.461538	0.272727	1
3	0.46105	0.222222	0.13332	0.2	0.666667	0.283333	0.909722	0.282051	0.454545	1
4	0.161525	0	0	0.2	0.481481	0	0.895833	0.384615	0.454545	0
5	0.122414	0.074067	0.13332	0.2	0.222222	0	0.916667	0.487179	0.424242	0
6	0.22019	0.277778	0.4	0.46664	0.666667	0.233333	0.659722	0.512821	0.484848	0
7	0.518283	0.444444	0.66668	0.2	0.592593	0	0.604167	0.410256	0.484848	0
8	0.35612	0.444444	0.4	0.2	0.703704	0.283333	0.833333	0.794872	0.090909	0
9	0.527821	0.518511	0.4	0	0.62963	0.833333	0.472222	0.589744	0.30303	1
10	0.403816	0.888889	1.06668	0.2	0.740741	0.283333	0.375	0.923077	0.333333	2
11	0.093798	0.111111	0.13332	0.36	0.814815	0.2	0.638889	0.461538	0.545455	0
12	0.277424	0.166667	0	0.2	0.796296	0.333333	0.625	0.487179	0.060606	1
13	0.255961	0.166667	0.3	0.2	0.722222	0.233333	0.590278	0.410256	0.424242	0
14	0.346582	0.148156	0	0.2	0.648148	0.483333	0.493056	0.487179	0.272727	0
15	0.241652	0.444444	0.4	0.46664	0.481481	0	0.305556	0.538462	0.181818	0
16	0.341812	0.222222	0.4	0.2	0.611111	0.2	0.659722	0.461538	0.363636	1
17	0.241652	0.111111	0.2	0.2	0.685185	0	0.5625	0.25641	0.424242	0
18	0.230206	0.222222	0.08	0.2	0.722222	0.55	0.618056	0.615385	0.242424	0
19	0.098568	0.296289	0.33332	0.36	0.462963	0	0.215278	0.74359	0.272727	0

One excerpt was classified as being highly confusing and another five as medium confusing. We looked at these and analysed the weight of each feature. Excerpt 10 was automatically accepted as the only high confusing and because it had some of the highest scores in each feature. Then, we rejected excerpts 16 and 2 because they did not possess high scores on any feature and we interpreted this as misclassification. Then we rejected excerpt 12 as being mid-confusing because the only high score it has is “LS1”, which stands for “Lexical Sophistication”. This means that it had lexically complex words and, after reading the respective excerpt, we realized that, as this presentation sometimes is technical, it is sometimes impossible to avoid this lexical complexity, which means that in the case of excerpts with many technical words, we did not take “LS1” into account since it is unavoidable. Excerpt 3 is also rated as mid-confusing, however, that is reportedly due to long t-units (“MLT”) and number of different words (“NDWERZ”). Again, after analysing this excerpt, it is an excerpt that lists devices and types of inputs, which rises “NDWERZ”, but it is

something that is needed and unavoidable. Lastly, we also accepted excerpt 9 because it reported high mean length of t-units and verb sophistication (“VS1”), as well as above average score of verb phrases per t-unit (“VP/T”). Furthermore, we also analysed the excerpts reported as low-confusion. From all, we decided to include excerpt 7 because it also contained one of the highest “MLT” and significant scores on other features. After identifying these three excerpts, we rewrote them keeping in mind which features were raising the reported confusion level. We ran them again through the classifying model and obtained the data reported on Table 10.

We mainly focused on reducing the length of each sentence as, from our empirical perspective, this is one of the features that mostly affects the readability of a sentence. With this process we were able to reduce “MLT” and the number of dependent clauses per t-unit (“DC/T”) on excerpt 7. However, this was achieved at the cost of raising “LS1” a bit. Excerpt 9 kept being reported as mid-confusing, however, we significantly reduced “MLT” and “VP/T”, slightly reduced “DC/T” and noun variation (“VV2”), at the expense of slightly raising “LS1”, number of t-units per sentence (“T/S”) and significantly increase the adjective variation (“ADJV”). On excerpt 10 we were able to reduce from high-confusion to low-confusion by significantly reducing “MLT”, “VP/T”, and “DC/T”. In turn, “NDWERZ” and “ADJV” slightly increased.

Table 10. Analysis from the three text excerpts that were chosen to be rewritten.

ID	MLT	VP/T	DC/T	T/S	LS1	VS1	NDWERZ	VV2	ADJV	conf
7	0.22019	0.277778	0.4	0.2	0.685185	0	0.5625	0.435897	0.30303	0
9	0.161525	0.177778	0.24	0.2	0.685185	0.833333	0.395833	0.358974	0.787879	1
10	0.055643	0.222222	0.2	0.2	0.740741	0.283333	0.506944	0.871795	0.424242	0

6.1.3. Virtual Environment

This is the 3D scenario used to run the experiment. It is integrated into the distributed environment, is part of the experience and concretizes the Presentation scenario described on “1.3.1 CVE scenarios”. Figure 26 shows this environment with one avatar presenting and two others watching.

The scenario takes place in a meeting room from the Company offices’ general 3D model and the user takes the role of someone that’s passively attending this presentation. The user cannot control the orientation of the camera as if turning the avatar’s head. Even though this reduces immersion, it eliminates an uncontrollable variable because every user should have the same experience. The presentation was composed of nine slides, each with a pre-written script, with around 11 minutes of duration.

The two passive avatars display a slight movement only to try and increase immersion without

investing too much resources on animation. It is important to stress that we did not want to spend too much resources on building a photo-realistic environment because our main goal is to interpret differences between Condition I and Condition II, and between Condition II and Condition III, rather than evaluate the absolute responses of users. The avatar that was performing had two talking animations that melded and were synchronized with the script. Whenever there was a transition between slides, the current slide would stay for two seconds before being replaced by the next slide. This next slide would also stay for two seconds before the avatar started speaking the respective written script. During these 4 seconds of transition, the avatar would stay quiet and its hands animations were paused. This pause did not occur abruptly, but rather the hands animation would start to fade to an idle position when it finished speaking. Then again, the hands animations would only start when it started speaking. The avatar's speech was synthesized resorting to IBM Watson Text-to-Speech service, with the online tool³³.



Figure 26. The 3D virtual environment used to test H2 and H3. There is a display where a slideshow was running, two passive avatars (back facing the user), and another presenting with text-to-speech generated speech. The lighting condition only varied on Condition III.

The virtual environment was the same in Condition I and Condition II, however, Condition III introduced the automatic adaptation of the lighting condition. The default lighting is the one seen on Figure 26, however, in Condition III it automatically gets brighter as the user's engagement is

³³ <https://text-to-speech-demo.ng.bluemix.net/>

lower.

6.1.4. Experimental Settings and Procedure

There was an active recruitment to gather all 45 valid users that were allocated into slots of 30 minutes each. In average, the experiment took 25 minutes, depending on the user. At the scheduled time, the user was taken to an isolated room that was scheduled for these tests.

The procedure had four steps: first, the user was welcomed, thanked for participating in the usability test, and told that his/her help is vital for the success of this dissertation; second, the researcher explained how the experiment would proceed, and third, 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) (Witmer & Singer, 1998). Then the researcher explained that in the second phase the user would be watching a presentation but that the subject of the presentation is not what is really important but the 3D environment itself. This was an attempt to make the user at ease with being distracted. Our goal was that he/she would feel as comfortable and relaxed as possible, as if he/she were in a real presentation, where there is less pressure to keep the attention on the subject. The researcher also told the user that while he/she was watching the presentation, the researcher would be with headphones and facing away from the user to lower the pressure. However, the researcher would still be present if there is 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. After, he/she would have to fill a final questionnaire evaluating the scenario which had 11 questions (from Q1 to Q11) taken from the Presence Questionnaire (PQ) (Witmer & Singer, 1998) and others tailored specifically for this experiment (from Q12 to Q17). These questionnaires can be found in Appendix A. Finally, after these stages, 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.

6.2 Evaluation and Discussion

6.2.1. Results of H2

The ITQ let us have insight about the tendency each group must become immersed on an activity. The statistical results of ITQ of Condition I and Condition II are displayed on Table 11. The group

of users from Condition I reported results below or equal to 3 on Q7 (M = 2.60, SD = 1.24) and Q8 (M = 3.00, SD = 1.46) above or equal to 5 for Q10 (M = 5.14, SD = 1.41), Q11 (M = 5.67, SD = 0.72), and Q13 (M = 5.20, SD = 1.66). 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 Condition II reported scores below or equal to 3 only for Q7 (M = 2.93, SD = 1.53) and above or equal to 5 on Q1 (M = 5.13, SD = 1.41), Q11 (M = 5.93, SD = 0.80), and Q13 (M = 5.73, SD = 1.39). 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.

Table 11. ITQ results for Condition I and Condition II. Both groups concentrate well on enjoyable activities and lose track of time when doing so. Do not get immersed when acting passively when watching sports, but one group is more leaned towards playing videogames, 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	5.13	4.47	4.07	4.53	4.36	4.87	2.93	3.53	4.40	4.33	5.93	3.73	5.73
SD	1.41	1.36	1.58	1.77	2.06	1.36	1.53	1.85	2.10	1.63	0.80	2.52	1.39

Figure 27 shows the mean difference between the ITQ results of users of Condition I and Condition II. 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). There are no questions with p-value below our significance threshold of 0.05 however, there are some that are close to it, which may show trends. Q4 (M = 0.73, $p = 0.10$), Q5 (M = 0.86, $p = 0.11$) and Q10 (M = -0.81, $p = 0.08$) (green bars) stand out. Positive average values mean that it is higher for Condition II, whereas negative average values means that it is higher on Condition I. This shows that users from Condition II are naturally more drawn to videogames and character development when compared to users from Condition I, whereas the latter feel more involved when playing sports.

ITQ mean differences between Conditions I and II



Figure 27. Mean differences and respective p-values from ITQ from Condition I and Condition II. Users from Condition II are more drawn to videogames, whereas the ones from Condition I feel more involved when playing sports.

Table 12 shows the statistical results for the PQ for Condition I and Condition II. The group of users from Condition I reported results below or equal to 3 on Q12 (M = 2.47, SD = 1.60), Q13 (M = 2.47, SD = 1.77), and Q17 (M = 1.47, SD = 0.74), and above or equal to 5 for Q5 (M = 5.27, SD = 1.71), Q6 (M = 5.40, SD = 1.59), Q10 (M = 6.00, SD = 1.07), Q15 (M = 5.73, SD = 1.16), and Q16 (M = 5.07, SD = 1.03). Results from Condition I 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 directly relates with attention kept on the presentation.

Table 12. Results from PQ and tailored questions for Condition I and Condition II. PQ questions go from Q1 to Q11, the rest are 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	4.80	4.47	4.33	4.73	5.80	5.47	4.36	5.27	5.13	6.13	3.50	1.73	1.80	4.07	5.73	4.87	2.73
SD	1.66	1.60	1.84	1.44	1.15	1.30	1.44	1.03	1.81	1.46	1.50	1.28	1.01	1.71	1.03	1.36	1.71

The user group from Condition II reported scores below or equal to 3 for Q12 (M = 1.73, SD = 1.28), Q13 (M = 1.80, SD = 1.01), and Q17 (M = 2.73, SD = 1.71), and above or equal to 5 on Q5 (M = 5.80, SD = 1.15), Q6 (M = 5.47, SD = 1.30), Q8 (M = 5.27, 1.03), Q9 (M = 5.13, SD = 1.81), Q10 (M = 6.13, SD = 1.46), and Q15 (M = 5.73, SD = 1.03). The overall results from Condition II follow those of Condition I 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 even more higher scores on questions related to sense of presence when compared to users from Condition I.

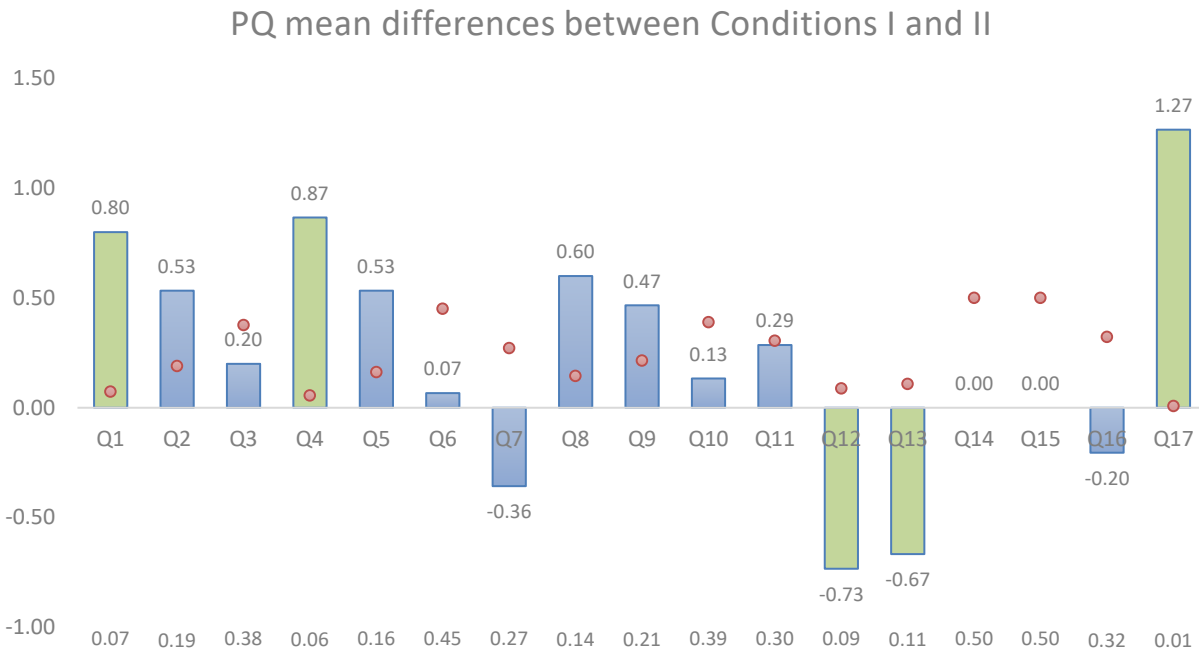


Figure 28. Mean differences and respective p-values from PQ and tailored questions from Condition I and Condition II. The presentation subject (Q13) and avatar's performance (Q12) had lower scores on Condition II. Users also reported higher values of immersion on Condition II.

Figure 28 shows the mean differences of the PQ between Condition I and Condition II. There are no questions with a p-value below our significance threshold, however, there are some interesting differences. Q12 (M = -0.73, $p = 0.09$) and Q13 (M = -0.67, $p = 0.11$) relate to the confusion the avatar performance or the presentation subject induce, and we can see that there are indications that on Condition II users perceived them as less confusing, aligned with our goal of lowering confusion reports with the rewritten script. Q1 (M = 0.80, $p = 0.07$) and Q4 (M = 0.87, $p = 0.06$) show that users reported they were more involved with the visual aspects of the environment and 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. However, the reasons for this may be twofold. The group from Condition II 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 Condition I). 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, 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 ($M = 1.27, p = 0.01$), which stands for how noticeable were the changes 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 H2, there are some indicators that can fuel future research, with a larger sample, as the effect sizes for Q12 and Q13 have moderate effect sizes of 0.51 and 0.46, respectively. Q1 and Q4 have moderately large effect sizes of 0.54 and 0.60 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 research to accurately explain its meaning.

6.2.2. Results of H3

Table 13. ITQ results for Condition II and Condition III. 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 stand out traits like the other two.

	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12	Q13
Condition II													
M	5.13	4.47	4.07	4.53	4.36	4.87	2.93	3.53	4.40	4.33	5.93	3.73	5.73
SD	1.41	1.36	1.58	1.77	2.06	1.36	1.53	1.85	2.10	1.63	0.80	2.52	1.39
Condition III													
M	4.07	4.33	4.07	3.27	3.33	4.33	3.50	3.27	3.40	4.50	5.07	4.00	5.13
SD	0.96	1.50	1.44	1.28	2.13	1.54	1.79	1.49	1.35	1.87	1.22	1.85	0.92

The statistical results of ITQ of Condition II are already analysed on the previous section, please refer to it for a complete description. We include them on Table 13 for an easier comparison with values from Condition III. Unlike the users from other conditions, users from Condition III report

no values below or equal to 3 and only Q11 ($M = 5.07$, $SD = 1.22$) and Q13 ($M = 5.13$, $SD = 0.92$) above or equal to 5. Following the trend of the other groups, this one also reports it concentrates well on enjoyable activities and loses all track of time in doing so. However, it does not possess any other traits that stand out.

Figure 29 displays the comparison between Condition II and Condition III ITQ results. There are several differences between these groups. Q1 ($M = -1.07$, $p = 0.01$) and Q4 ($M = -1.27$, $p = 0.02$) reveal that group from Condition III 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 ($M = -0.87$, $p = 0.01$) reports. In spite of this, its absolute value is still above 5, as reported on the previous paragraph. In addition, Q5 ($M = -1.02$, $p = 0.10$) shows that they feel less involved when playing videogames and dream less realistic dreams as Q9 ($M = -1.00$, $p = 0.07$) reveals. On other groups Q11 seems to be related with Q13, something that also follows that trend on this group with Q13 ($M = -0.60$, $p = 0.09$), reporting lower values of losing track of time when doing something. The group from Condition III appears to have less tendency to be easily immersed on activities, so we could expect that to be reflected on the results.

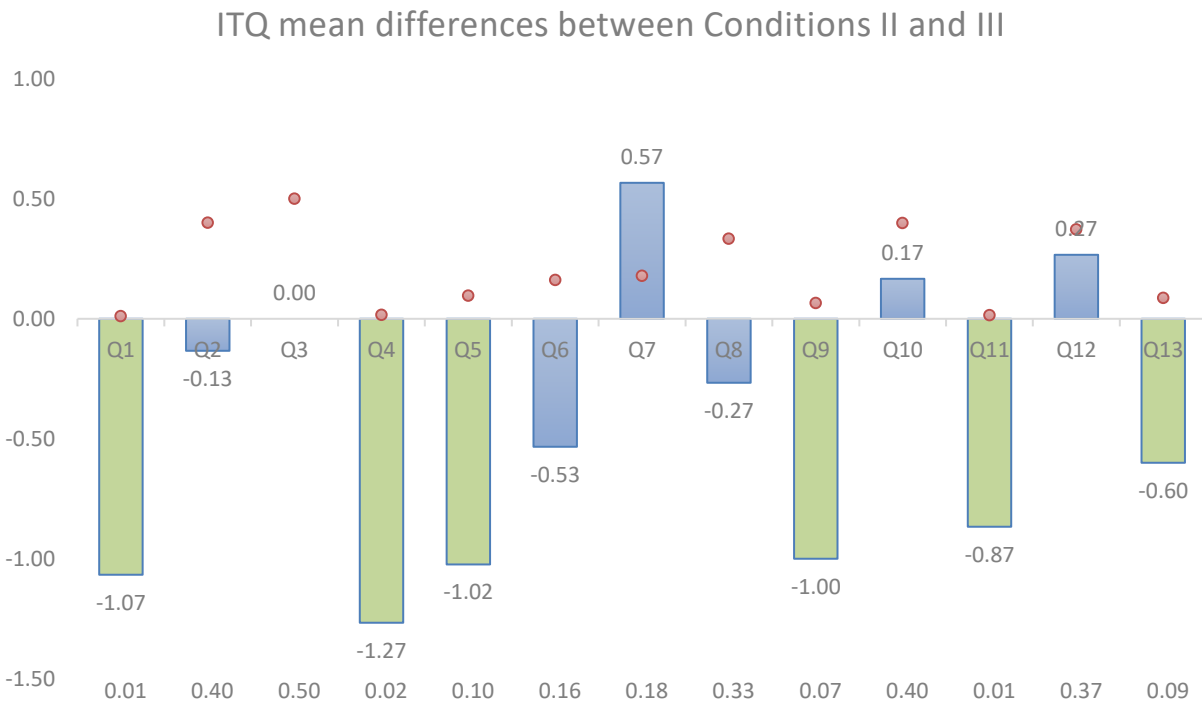


Figure 29. Mean differences and respective p-values from ITQ from Condition II and Condition III. Condition III users significantly report they have a harder time concentrating on enjoyable activities and do not involve as much on TV dramas and identify less with characters on plots.

Table 14 displays the results from Condition II and Condition III that test H3. The analysis of the absolute results yielded by the group of Condition II are already reported on the previous sections, please refer to it. Condition III yielded results below or equal to 3 on Q12 (M = 2.52, SD = 1.25) and Q13 (M = 2.33, SD = 1.29), and results above or equal to 5 on Q5 (M = 5.87, SD = 0.92), Q6 (M = 5.60, SD = 1.12), Q7 (M = 5.14, SD = 1.73), Q8 (M = 5.27, SD = 1.22), Q10 (M = 5.79, SD = 0.86), Q15 (M = 5.20, SD = 1.32), Q16 (M = 5.53, SD = 0.99, and Q17 (M = 6.20, SD = 1.01). 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.

Table 14. 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	4.80	4.47	4.33	4.73	5.80	5.47	4.36	5.27	5.13	6.13	3.50	1.73	1.80	4.07	5.73	4.87	2.73
SD	1.66	1.60	1.84	1.44	1.15	1.30	1.44	1.03	1.81	1.46	1.50	1.28	1.01	1.71	1.03	1.36	1.71
Condition III																	
M	4.53	4.20	4.73	4.73	5.87	5.60	5.14	5.27	4.80	5.79	3.40	2.53	2.33	4.33	5.20	5.53	6.20
SD	0.99	1.15	1.39	1.22	0.92	1.12	1.73	1.22	1.08	0.86	1.50	1.25	1.29	1.23	1.32	0.99	1.01

Figure 30 shows the mean difference between Condition II and Condition III 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 = 3.47, $p < 0.01$). It relates to the noticeability of the lighting condition adaptations and with this result we are assured that users were well 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 = 0.79, $p = 0.09$) 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 Condition I and Condition II, and not between Condition II and Condition III. Other unexpected results are from Q12 (M = 0.80, $p = 0.05$) and Q13 (M = 0.53, $p = 0.11$) where users report that the confusion induced by avatar's performance increased with statistical significance, and the subject of the presentation also seems to have been reported as more confusing. We were expecting that these values would not reveal any differences or, if they did, they would be reported with lower values. However, this can be explained with resource to a

previous explanation related with the visual and auditory channels. Condition III's visual environment was much more dynamic due to the lighting adaptations, which may require more mental resources from the user allocated to vision and unfocuses the user from what is being transmitted through the auditory channel (in this case, the presentation script given by the avatar). Curiously, they also report in Q16 ($M = 0.67, p = 0.07$) that the lighting condition was more important to keep their attention on the presentation than users from Condition II, which aligns with H3. 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 Condition III group of users having significantly less tendency for immersion, which would explain this lack of differences. However, we also have to consider the chance of this lighting adaptation system not contributing significantly for the user sense of presence.

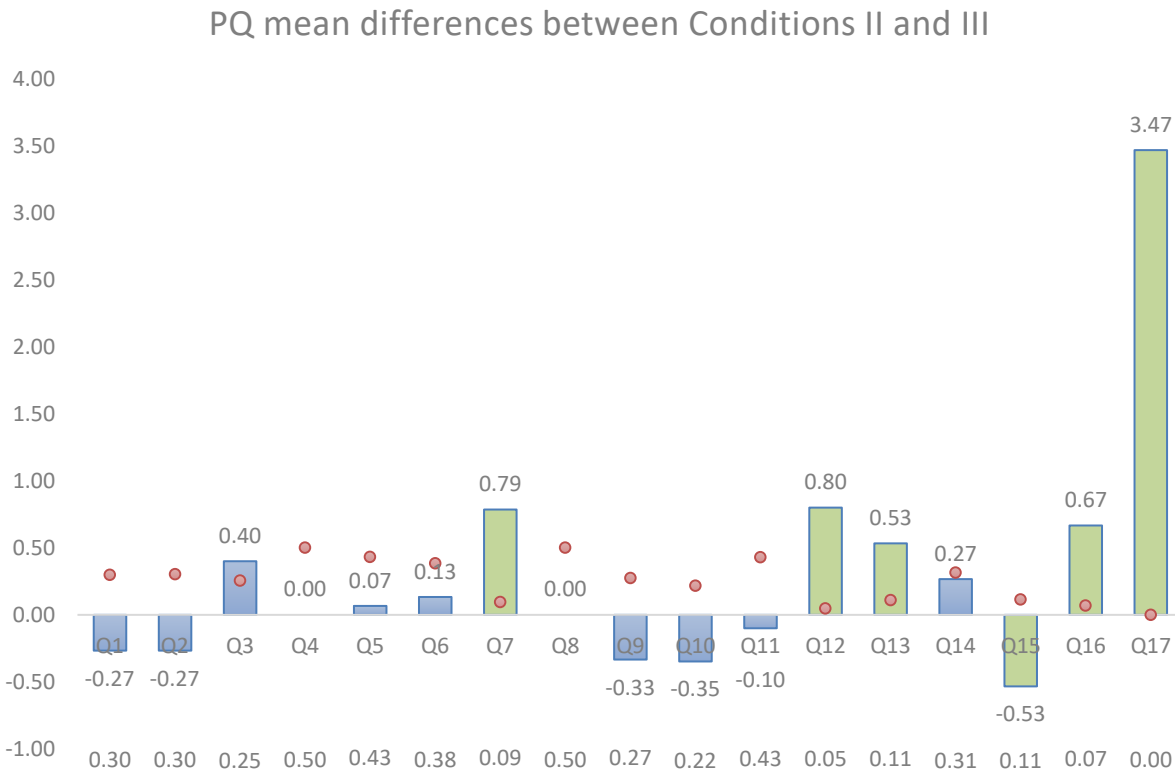


Figure 30. Mean differences and respective p-values from PQ and tailored questions from Condition II and Condition III. Users reported a higher score on Q17, which gives assurance that they did not miss the lighting adaptations. Q12 and Q13, relating to confusion induced, were reported as higher, maybe due to a conflict on information channels.

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 H3 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 2.47. Q7, Q13, and Q15 have moderate effect sizes of 0.49, 0.46, and 0.45, respectively, whereas Q12 and Q16 have moderately large effect sizes of 0.63 and 0.56, respectively. As in the previous section, also in the testing of this hypothesis a larger sample may reveal clearer results.

6.3 Summing Up

In this chapter we described the experiment we used to test H2 and H3. We had three conditions to test these two hypotheses, by forming two pairs. Condition I's results were compared to Conditions II's, and Condition II's were compared with Conditions III's. All of these are based on the same 3D virtual environment, with two independent variables to evaluate the variation of sense of presence between conditions. We varied the script the avatar was speaking to the user from Condition I to Condition II, to evaluate if a less confusing script would lead to less reported confusion and had an impact on the sense of presence. From Condition II to Condition III the variable was the presence of automatic lighting adaptations.

The user pool was composed of subjects that work in the Information Technology sector, which brings a strong bias towards this evaluation. However, all users being from the same sector reduces the variability of background and strong deviations from user experience with digital interfaces and virtual environments. Even so, although user groups from Condition I and Condition II were similar on the ITQ, Condition III user group had strong differences on the ITQ, proving to have significantly less tendency for immersion, which may have introduced bias on the results.

However, this experiment yielded interesting results partially confirming H2 and partially rejecting H3, but still providing useful insight on how this line of research can be carried further. Results showed indicators that a script can be optimized to lower confusion induction resorting to a learning model and this may augment the sense of presence. We theorize that since the user has to spend less cognitive resources on understanding the script, he/she might get more immersed on the virtual environment. Users from Condition III reported higher confusion from the presentation subject and avatar's performance, maybe due to overburdening of the visual channel, which did not let them allocate as much cognitive resources on what was being conveyed by the auditory

channel. Hsia (Hsia, 1973) states that “Between-channel redundancy refers to the similarity between two channels (...) Conceivably, between-channel redundancy is unity when both visual and auditory channels transmit identical information; conversely, it is zero when the visual and auditory channels emit completely different information.”. We think that, in this case, this overburdening is harmful due to the lack of similarity between both channels. The user group from Condition III has less tendency for immersion, which may have lowered the PQ results that, otherwise, could have been higher.

Chapter VII – Conclusion and Future Work

In this dissertation we approached the problem of Companies that want to spread awareness about their offices and provide virtual tours to people who cannot visit them, due to time, distance, or other constraints. Two scenarios were identified as part of this virtual tour: a Socialization and a Presentation scenarios.

7.1 Academic and Industrial Contributions

A general 3D CVE framework was built to accommodate applications that meet these scenarios. It was developed on an industrial environment and brings a more natural interaction based on devices that are ubiquitous across the industry. This framework provides a server-client architecture and interaction channels based on ubiquitous devices, such as voice chat, mapping of facial expressions or automatic lighting adaptation based on facial and emotion recognition. The learning model developed represents both an academic and industrial contribution, since, to the best of our knowledge, there was no study or tool to assess the confusion of text based on their lexical, syntactic, and n-gram features.

The last contribution this work brings is the automatic adaptive lighting system based on user engagement. It represents an academic contribution since only few studies take automatic adaptation of lighting as a variable to augment the user's sense of presence and re-engagement. However, this system cannot yet be taken as an industrial contribution. It represents a first effort of this kind of system based on video input, which produces a lot of noise and deals with sensible data.

7.2 Answers to Hypotheses and Discussion

Our hypotheses are the following:

H1 “It is possible to predict a sentence's chance of generating confusion based on syntactic, lexical and n-gram features.”

H2 “A less confusing sentence on a virtual presentation increases the user's sense of presence.”

H3 “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 Multilayer Perceptron was trained on syntactic, lexical and n-gram features, and achieve an overall f-score of 0.57, with a scores of 0.70, 0.38, and 0.33 for “low”, “mid”, and “high”

confusion, respectively. The Fleiss' kappa coefficient for this dataset falls on the "Slight agreement" with a value of 0.16, which reminds us that there was little agreement between annotators, which, in part, explains the low absolute value of the chosen error metric. This allows us to partially confirm H1, because it is possible to use a prediction model to assist a human in determining the confusion level of a text excerpt. This tool was used later to assist us in evaluating H2. With this in mind, we consider these as promising results that we hope will improve with more data. There is obvious room for improvement over the size of the dataset, but also over the design of the data collection itself. The content of the text excerpts was varied and, even if the annotators could detach from it, there is possible bias coming from their personal preferences which may make them more committed to rate some excerpts than others. One possible solution for this may be collecting excerpts over a single mainstream theme.

The case study of this dissertation is a Presentation scenario where we tested H2 and H3 with 45 users under three experimental conditions. H2 is partially supported on Presence Questionnaire questions that evaluate the involvement of the visual aspects of the environment, as well as how consistent the environment was with the user's real-life experiences. In addition, the confusion reported was lower. However, none of these questions achieved statistical significance. An unexpected report of users noticing more lighting changes when exposed to a less confusing script requires further research, but we theorize that this could be related to the user having more cognitive resources to visually explore the environment, since a less confusing script may require less resources.

The second part of this experiment partially rejected H3 since users did not report any differences on the Presence Questionnaire items and reported that the avatar's performance was significantly more confusing. This opens an interesting question, because the script did not vary on H3 conditions. Once again, we theorize that this may have to do with information channels (Hsia, 1973). On H3 we were overburdening the visual channel, while the information was being conveyed by the avatar's speech.

7.3 Future Work

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

Confusion detection: Collecting a larger dataset is an obvious step towards improving the error metric of the model. At the moment this dataset comprises a broad set of themes and this could be

a source of bias because different people react differently towards their favorite themes. However, with a large enough dataset we expect these individual differences are dissolved. For this, an automatic web scrapper should be developed, rather than manually collecting presentations, as was done in this first effort. Another possible way is to resort to CoreNLP (Manning et al., 2014)³⁴ to explore other syntactic and lexical features, such as syntactic trees, and to eliminate the constraint of only processing excerpts with more than 50 words. Moreover, the ultimate goal is to turn this offline model into a real-time one, detecting the user's confusion level in real-time, relating it to the current text's syntactic, lexical and n-gram features, and automatically adapting the next part of the script according to this.

Another possible field of application of this technology is on Massive Open Online Courses (MOOC) integrated with an Intelligent Tutoring System (ITS) like AutoTutor. Even though its authors already approach the theme of confusion, a system that understands why a speech is being confusing could tailor it in real-time in order to get the best results.

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. (Spengler et al., 2017) 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 (Büscher, O'Brien, Rodden, & Trevor, 2001). By augmenting nonverbal synchronization between users with emotional body postures, the engagement level is increased as well as the acceptance of the system. For more information, refer to the survey about emotional body posture in Appendix C.

Coordinated Actions: This dissertation is supported on the premise of basing its interaction only on ubiquitous devices that are available for the common employee. At the moment of this writing, depth sensors, eye trackers, or anything besides webcam, microphone, and mouse & keyboard is not common among their workplaces. However, we may see advances in the devices that

³⁴ <https://stanfordnlp.github.io/CoreNLP/>, accessed 25th September 2018

are common on workplaces. Actually, 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 a bunch of options for interaction and maybe, for instance, with depth sensors this framework could be expanded to support Collaborative Work and Training scenarios.

References

- Abdelrahman, Y., Hassib, M., Marquez, M. G., Funk, M., & Schmidt, A. (2015). Implicit Engagement Detection for Interactive Museums Using Brain-Computer Interfaces. *Proceedings of the 17th International Conference on Human-Computer Interaction with Mobile Devices and Services Adjunct - MobileHCI '15*, 838–845. <https://doi.org/10.1145/2786567.2793709>
- Ali, Z., & Usman, M. (2016). A framework for game engine selection for gamification and serious games. In *2016 Future Technologies Conference (FTC)* (pp. 1199–1207). IEEE. <https://doi.org/10.1109/FTC.2016.7821753>
- Allison, B., Guthrie, D., & Guthrie, L. (2006). Another look at the data sparsity problem. *Text, Speech and Dialogue*, 327–334. <https://doi.org/10.1007/11846406>
- Arguel, A., Lockyer, L., Lipp, O. V., Lodge, J. M., & Kennedy, G. (2017). Inside Out. *Journal of Educational Computing Research*, 55(4), 526–551. <https://doi.org/10.1177/0735633116674732>
- Baltrusaitis, T., Robinson, P., & Morency, L. P. (2016). OpenFace: An open source facial behavior analysis toolkit. *2016 IEEE Winter Conference on Applications of Computer Vision, WACV 2016*. <https://doi.org/10.1109/WACV.2016.7477553>
- Bartle, R. (2010). International Handbook of Internet Research, (June 2010). https://doi.org/https://doi.org/10.1007/978-1-4020-9789-8_2
- Benford, S., Greenhalgh, C., Rodden, T., & Pycocock, J. (2001). Collaborative virtual environments. *Communications of the ACM*, 44(7), 79–85. <https://doi.org/10.1145/379300.379322>
- Bersin, J. (2016). Use Of MOOCs And Online Education Is Exploding: Here's Why. Retrieved January 22, 2018, from www.forbes.com/sites/joshbersin/2016/01/05/use-of-moocs-and-online-education-is-exploding-heres-why
- Bixler, R., & D'Mello, S. K. (2013). Detecting boredom and engagement during writing with keystroke analysis, task appraisals, and stable traits. In *Proceedings of the 2013 international conference on Intelligent user interfaces - IUI '13* (p. 225). New York, New York, USA: ACM Press. <https://doi.org/10.1145/2449396.2449426>
- Büscher, M., O'Brien, J., Rodden, T., & Trevor, J. (2001). "He's Behind You": The Experience of Presence in Shared Virtual Environments. In *Collaborative virtual environments* (pp. 77–98). https://doi.org/10.1007/978-1-4471-0685-2_5
- Chinchor, N. (1992). The statistical significance of the MUC-4 results. In *Proceedings of the 4th conference on Message understanding - MUC4 '92* (p. 30). Morristown, NJ, USA: Association for Computational Linguistics. <https://doi.org/10.3115/1072064.1072068>
- Craig, S. D., Graesser, A. C., Sullins, J., & Gholson, B. (2004). Affect and learning: An exploratory look into the role of affect in learning. *Journal of Educational Media*, 29, 241–250.
- D'Mello, S. K., & Calvo, R. A. (2013). Beyond the basic emotions. In *CHI '13 Extended*

- Abstracts on Human Factors in Computing Systems on - CHI EA '13* (p. 2287). New York, New York, USA: ACM Press. <https://doi.org/10.1145/2468356.2468751>
- D'Mello, S. K., Craig, S. D., Sullins, J., & Graesser, A. C. (2006). Predicting affective states through an emote-aloud procedure from AutoTutor's mixed-initiative dialogue. *International Journal of Artificial Intelligence in Education, 16*, 3–28.
- D'Mello, S. K., Craig, S. D., Witherspoon, A., McDaniel, B., & Graesser, A. (2008). Automatic detection of learner's affect from conversational cues. *User Modeling and User-Adapted Interaction, 18*(1–2), 45–80. <https://doi.org/10.1007/s11257-007-9037-6>
- D'Mello, S. K., & Graesser, A. (2006). Affect Detection from Human-Computer Dialogue with an Intelligent Tutoring System (pp. 54–67). https://doi.org/10.1007/11821830_5
- D'Mello, S. K., & Graesser, A. (2009). Automatic Detection of Learner's Affect From Gross Body Language. *Applied Artificial Intelligence, 23*(2), 123–150. <https://doi.org/10.1080/08839510802631745>
- D'Mello, S. K., & Graesser, A. (2012a). AutoTutor and affective autotutor. *ACM Transactions on Interactive Intelligent Systems, 2*(4), 1–39. <https://doi.org/10.1145/2395123.2395128>
- D'Mello, S. K., & Graesser, A. (2012b). Dynamics of affective states during complex learning. *Learning and Instruction, 22*(2), 145–157. <https://doi.org/10.1016/j.learninstruc.2011.10.001>
- D'Mello, S. K., & Kory, J. (2012). Consistent but modest: a meta-analysis on unimodal and multimodal affect detection accuracies from 30 studies. In *Proceedings of the 14th ACM international conference on Multimodal interaction - ICMI '12* (p. 31). New York, New York, USA: ACM Press. <https://doi.org/10.1145/2388676.2388686>
- D'Mello, S. K., & Kory, J. (2015). A Review and Meta-Analysis of Multimodal Affect Detection Systems. *ACM Computing Surveys, 47*(3), 1–36. <https://doi.org/10.1145/2682899>
- D'Mello, S. K., Lehman, B., Pekrun, R., & Graesser, A. (2014). Confusion can be beneficial for learning. *Learning and Instruction, 29*, 153–170. <https://doi.org/10.1016/j.learninstruc.2012.05.003>
- D'Mello, S. K., Lehman, B., & Person, N. (2010). Monitoring affect states during effortful problem solving activities. *International Journal of Artificial Intelligence in Education, 20*(4), 361–389.
- Darwin, C. (1872). The expression of the emotions in man and animals. *London, UK: John Murray*. <https://doi.org/10.1037/h0076058>
- Das, P. K., & Deka, G. C. (2015). History and Evolution of GPU Architecture. In G. C. Deka (Ed.), *Emerging Research Surrounding Power Consumption and Performance Issues in Utility Computing* (pp. 109–135). IGI-Global. <https://doi.org/10.4018/978-1-4666-8853-7.ch006>
- Dias, M. S., Eloy, S., Carreiro, M., Proença, P., Moural, A., Pedro, T., ... Azevedo, A. S. (2014). Designing better spaces for people. *Rethinking Comprehensive Design: Speculative Counterculture - Proceedings of the 19th International Conference on Computer-Aided Architectural Design Research in Asia, CAADRIA 2014*, 739–748. Retrieved from <https://www.scopus.com/inward/record.uri?eid=2-s2.0->

84904706822&partnerID=40&md5=f75b87716a4bca0835c334fdd439e882

- Dias, M. S., Eloy, S., Carreiro, M., Vilar, E., Marques, S., Moural, A., ... Pedro, T. (2014). Space Perception in Virtual Environments: on how biometric sensing in virtual environments may give architects. *Fusion - Proceedings of the 32nd ECAADe Conference*, 2, 271–280.
- Ekman, P. (2016). What Scientists Who Study Emotion Agree About. *Perspectives on Psychological Science*, 11(1), 31–34. <https://doi.org/10.1177/1745691615596992>
- Ekman, P., & Friesen, W. V. (1969). The Repertoire of Nonverbal Behavior: Categories, Origins, Usage, and Coding. *Semiotica*, 1(1). <https://doi.org/10.1515/semi.1969.1.1.49>
- Ekman, P., & Friesen, W. V. (1971). Constants across cultures in the face and emotion. *Journal of Personality and Social Psychology*, 17(2), 124–129. <https://doi.org/10.1037/h0030377>
- Eloy, S., Ourique, L., Pedro, T., Resende, R., Dias, M. S., & Freitas, J. (2015). Analysing People's Movement in the Built Environment via Space Syntax, Objective Tracking and Gaze Data. *ECAADe 2015 - 33rd Annual Conference*, 1, 341–350. Retrieved from <http://info.tuwien.ac.at/ecaade2015/index.html>
- Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5), 378–382. <https://doi.org/10.1037/h0031619>
- Graesser, A. C. (2016). Conversations with AutoTutor Help Students Learn. *International Journal of Artificial Intelligence in Education*, 26(1), 124–132. <https://doi.org/10.1007/s40593-015-0086-4>
- Graesser, A. C., McNamara, D. S., Louwerse, M. M., & Cai, Z. (2004). Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, & Computers*, 36(2), 193–202. <https://doi.org/10.3758/BF03195564>
- Grafsgaard, J. F., Wiggins, J. B., Boyer, K. E., Wiebe, E. N., & Lester, J. C. (2013). Automatically recognizing facial expression: Predicting engagement and frustration. *The 6th International Conference on Educational Data Mining EDM 2013*, 43–50.
- Grafsgaard, J. F., Wiggins, J. B., Boyer, K. E., Wiebe, E. N., & Lester, J. C. (2013). Automatically Recognizing Facial Indicators of Frustration: A Learning-centric Analysis. In *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction* (pp. 159–165). <https://doi.org/10.1109/ACII.2013.33>
- Grafsgaard, J. F., Wiggins, J. B., Vail, A. K., Boyer, K. E., Wiebe, E. N., & Lester, J. C. (2014). The Additive Value of Multimodal Features for Predicting Engagement, Frustration, and Learning during Tutoring. In *Proceedings of the 16th International Conference on Multimodal Interaction - ICMI '14* (pp. 42–49). New York, New York, USA: ACM Press. <https://doi.org/10.1145/2663204.2663264>
- Gunes, H., & Schuller, B. (2013). Categorical and dimensional affect analysis in continuous input: Current trends and future directions. *Image and Vision Computing*, 31(2), 120–136. <https://doi.org/10.1016/j.imavis.2012.06.016>
- Guye-Vuilieme, A., Capin, T. K., Pandzic, I. S., Thalmann, N. M., & Thalmann, D. (1999). Nonverbal Communication Interface for Collaborative Virtual Environments. *Virtual Reality*, 4, 49–59. <https://doi.org/10.1007/BF01434994>

- Hall, E. T., Birdwhistell, R. L., Bock, B., Bohannon, P., Diebold, A. R., Durbin, M., ... Vayda, A. P. (1968). Proxemics [and Comments and Replies]. *Current Anthropology*, 9(2/3), 83–108. <https://doi.org/10.1086/200975>
- Hawes, B. K., Brunyé, T. T., Mahoney, C. R., Sullivan, J. M., & Aall, C. D. (2012). Effects of four workplace lighting technologies on perception, cognition and affective state. *International Journal of Industrial Ergonomics*, 42(1), 122–128. <https://doi.org/10.1016/j.ergon.2011.09.004>
- Hsia, N. J. (1973). On Redundancy. *Association for Editorial Journalism Convention*.
- Hussain, M. S., AlZoubi, O., Calvo, R. A., & D’Mello, S. K. (2011). Affect Detection from Multichannel Physiology during Learning Sessions with AutoTutor. In G. Biswas, S. Bull, J. Kay, & A. Mitrovic (Eds.), *Lecture Notes in Computer Science* (pp. 131–138). Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-21869-9_19
- Izard, C. E. (1971). *The face of emotion*. New York: Appleton-Century-Crofts. Retrieved from <https://books.google.co.uk/books?id=7DQNAQAAMAAJ>
- Jang, E.-H., Park, B.-J., Park, M.-S., Kim, S.-H., & Sohn, J.-H. (2015). Analysis of physiological signals for recognition of boredom, pain, and surprise emotions. *Journal of Physiological Anthropology*, 34(1), 25. <https://doi.org/10.1186/s40101-015-0063-5>
- Johnson, W. L., & Lester, J. C. (2016). Face-to-Face Interaction with Pedagogical Agents, Twenty Years Later. *International Journal of Artificial Intelligence in Education*, 26(1), 25–36. <https://doi.org/10.1007/s40593-015-0065-9>
- Karg, M., Samadani, A. A., Gorbet, R., Kuhnlenz, K., Hoey, J., & Kulic, D. (2013). Body movements for affective expression: A survey of automatic recognition and generation. *IEEE Transactions on Affective Computing*, 4(4), 341–359. <https://doi.org/10.1109/T-AFFC.2013.29>
- Kleinsmith, A., & Bianchi-Berthouze, N. (2013). Affective Body Expression Perception and Recognition: A Survey. *IEEE Transactions on Affective Computing*, 4(1), 15–33. <https://doi.org/10.1109/T-AFFC.2012.16>
- Knez, I., & Kers, C. (2000). Effects of Indoor Lighting, Gender, and Age on Mood and Cognitive Performance. *Environment and Behavior*, 32(6), 817–831. <https://doi.org/10.1177/0013916500326005>
- Kuijsters, A., Redi, J., De Ruyter, B., & Heynderickx, I. (2015). Lighting to make you feel better: Improving the mood of elderly people with affective ambiances. *PLoS ONE*, 10(7), 1–22. <https://doi.org/10.1371/journal.pone.0132732>
- Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33(1), 159–174. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/843571>
- LeDoux, J. (1998). *The Emotional Brain: The Mysterious Underpinnings of Emotional Life*. Simon & Schuster.
- Lehman, B., D’Mello, S., & Graesser, A. (2012). Interventions to Regulate Confusion during Learning. In S. A. Cerri, W. J. Clancey, G. Papadourakis, & K. Panourgia (Eds.) (7315th ed., pp. 576–578). Springer, Berlin, Heidelberg. <https://doi.org/10.1007/978-3-642-30950->

- Lehman, B., D’Mello, S. K., Strain, A. C., Gross, M., Dobbins, A., Wallace, P., ... Graesser, A. C. (2011). Inducing and Tracking Confusion with Contradictions during Critical Thinking and Scientific Reasoning. In G. Biswas, S. Bull, J. Kay, & A. Mitrovic (Eds.) (pp. 171–178). Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-21869-9_24
- Lehman, B., D’Mello, S., Strain, A., Mills, C., Gross, M., Dobbins, A., ... Graesser, A. (2013). Inducing and Tracking Confusion with Contradictions during Complex Learning. *International Journal of Artificial Intelligence in Education*, 22(1–2), 85–105. <https://doi.org/10.3233/JAI-130025>
- Lu, X. (2010). Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics*, 15(4), 474–496. <https://doi.org/10.1075/ijcl.15.4.02lu>
- Lu, X. (2012). The Relationship of Lexical Richness to the Quality of ESL Learners’ Oral Narratives. *Modern Language Journal*, 96(2), 190–208. https://doi.org/10.1111/j.1540-4781.2011.01232_1.x
- MacLean, P. D. (1949). Psychosomatic disease and the visceral brain; recent developments bearing on the Papez theory of emotion. *Psychosomatic Medicine*, 11(6), 338–353. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/15410445>
- Manning, C., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S., & McClosky, D. (2014). The Stanford CoreNLP Natural Language Processing Toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations* (pp. 55–60). Stroudsburg, PA, USA: Association for Computational Linguistics. <https://doi.org/10.3115/v1/P14-5010>
- Matsumoto, D., Keltner, D., & Shiota, M. (2016). Facial Expressions of emotion. In *Handbook of Emotions* (4th ed.). Guilford Press.
- McNair, D. M., Lorr, M., & Droppleman, L. F. (1971). *Manual for the Profile of Mood States*. (E. and I. T. Services, Ed.). San Diego, CA.
- Mott, M. S., Robinson, D. H., Walden, A., Burnette, J., & Rutherford, A. S. (2012). Illuminating the effects of dynamic lighting on student learning. *SAGE Open*, 2(2), 1–9. <https://doi.org/10.1177/2158244012445585>
- Nguyen, T. T. H., & Duval, T. (2015). A survey of communication and awareness in collaborative virtual environments. *2014 International Workshop on Collaborative Virtual Environments, 3DCVE 2014*, 1–8. <https://doi.org/10.1109/3DCVE.2014.7160928>
- P. Ekman and W. V. Friesen. (1977). *Manual for the Facial Action Code System*. Palo Alto: Consulting Psychologists Press.
- Panksepp, J. (1998). *Affective Neuroscience: The Foundations of Human and Animal Emotions*. New York: Oxford University Press.
- Panksepp, J. (2005). Affective consciousness: Core emotional feelings in animals and humans. *Consciousness and Cognition*, 14(1), 30–80. <https://doi.org/10.1016/j.concog.2004.10.004>
- Park, N.-K., & Farr, C. A. (2007). The Effects of Lighting on Consumers’ Emotions and

- Behavioral Intentions in a Retail Environment: A Cross-Cultural Comparison. *Journal of Interior Design*, 33(1), 17–32. <https://doi.org/10.1111/j.1939-1668.2007.tb00419.x>
- Paulmann, S., & Pell, M. D. (2011). Is there an advantage for recognizing multi-modal emotional stimuli? *Motivation and Emotion*, 35(2), 192–201. <https://doi.org/10.1007/s11031-011-9206-0>
- Pedica, C., & Högni Vilhjálmsson, H. (2010). Spontaneous avatar behavior for human territoriality. *Applied Artificial Intelligence*, 24(6), 575–593. <https://doi.org/10.1080/08839514.2010.492165>
- Peña Pérez Negrón, A., Rangel Bernal, N. E., & Lara López, G. (2015). Nonverbal interaction contextualized in collaborative virtual environments. *Journal on Multimodal User Interfaces*, 9(3), 253–260. <https://doi.org/10.1007/s12193-015-0193-4>
- Plutchik, R. (1980). *Emotions: A Psychoevolutionary Synthesis*. New York: Harper & Row.
- Quartier, K., Vanrie, J., & Van Cleempoel, K. (2014). As real as it gets: What role does lighting have on consumer's perception of atmosphere, emotions and behaviour? *Journal of Environmental Psychology*, 39, 32–39. <https://doi.org/10.1016/j.jenvp.2014.04.005>
- Richter-Levin, G. (2004). The amygdala, the hippocampus, and emotional modulation of memory. *The Neuroscientist: A Review Journal Bringing Neurobiology, Neurology and Psychiatry*, 10(1), 31–39. <https://doi.org/10.1177/1073858403259955>
- Sariyanidi, E., Gunes, H., & Cavallaro, A. (2015). Automatic Analysis of Facial Affect: A Survey of Registration, Representation, and Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(6), 1113–1133. <https://doi.org/10.1109/TPAMI.2014.2366127>
- Schroeder, N. L., Adesope, O. O., & Gilbert, R. B. (2013). How Effective are Pedagogical Agents for Learning? A Meta-Analytic Review. *Journal of Educational Computing Research*, 49(1), 1–39. <https://doi.org/10.2190/EC.49.1.a>
- Schroeder, R. (2002). *The Social Life of Avatars*. (R. Schroeder, Ed.). London: Springer London. <https://doi.org/10.1007/978-1-4471-0277-9>
- Schuemie, M. J., van der Straaten, P., Krijn, M., & van der Mast, C. A. P. G. (2001). Research on Presence in Virtual Reality: A Survey. *CyberPsychology & Behavior*, 4(2), 183–201. <https://doi.org/10.1089/109493101300117884>
- Scott, E., Soria, A., & Campo, M. (2016). Adaptive 3D Virtual Learning Environments – A Review of the Literature, 1382(c), 1–15. <https://doi.org/10.1109/TLT.2016.2609910>
- Silva Pedro, T., Luís Silva, J., & Pereira, R. (2018). Predicting the Confusion Level of Text Excerpts with Syntactic, Lexical and N-gram Features (pp. 8417–8426). <https://doi.org/10.21125/edulearn.2018.1959>
- Soliman, M., & Guetl, C. (2010). Intelligent Pedagogical Agents in immersive virtual learning environments: A review. In *The 33rd International Convention MIPRO* (pp. 827–832).
- Spengler, F. B., Scheele, D., Marsh, N., Kofferath, C., Flach, A., Schwarz, S., ... Hurlmann, R. (2017). Oxytocin facilitates reciprocity in social communication. *Social Cognitive and Affective Neuroscience*, 12(8), 1325–1333. <https://doi.org/10.1093/scan/nsx061>

- Stephens-Fripp, B., Naghdy, F., Stirling, D., & Naghdy, G. (2017). Automatic Affect Perception Based on Body Gait and Posture: A Survey. *International Journal of Social Robotics*, 9(5), 617–641. <https://doi.org/10.1007/s12369-017-0427-6>
- Tarng, P.-Y., Chen, K.-T., & Huang, P. (2008). An analysis of WoW players' game hours. In *Proceedings of the 7th ACM SIGCOMM Workshop on Network and System Support for Games - NetGames '08* (p. 47). New York, New York, USA: ACM Press. <https://doi.org/10.1145/1517494.1517504>
- Vasudevamurt, V. B., & Uskov, A. (2015). Serious game engines: Analysis and applications. In *2015 IEEE International Conference on Electro/Information Technology (EIT)* (pp. 440–445). IEEE. <https://doi.org/10.1109/EIT.2015.7293381>
- Vaughan, N., Gabrys, B., & Dubey, V. N. (2016). An overview of self-adaptive technologies within virtual reality training. *Computer Science Review*, 22, 65–87. <https://doi.org/10.1016/j.cosrev.2016.09.001>
- Vinayagamorthy, V., Gillies, M., Steed, A., Tanguy, E., Pan, X., Loscos, C., & Slater, M. (2006). Building Expression into Virtual Characters. *The Eurographics Association*, 1–42. <https://doi.org/citeulike-article-id:7494884>
- Witmer, B. G., & Singer, M. J. (1998). Measuring Presence in Virtual Environments: A Presence Questionnaire. *Presence: Teleoperators and Virtual Environments*, 7(3), 225–240. <https://doi.org/10.1162/105474698565686>
- Wundt, W. (1896). Emotions. In *Grundriss der Psychologie*.
- Wurtman, R. J. (1975). The Effects of Light on the Human Body. *Scientific American*, 233(1), 69–77.
- Yan, S., Ding, G., Li, H., Sun, N., Wu, Y., Guan, Z., ... Huang, T. (2016). Enhancing Audience Engagement in Performing Arts Through an Adaptive Virtual Environment with a Brain-Computer Interface. *Iui '16*, 306–316. <https://doi.org/10.1145/2856767.2856768>
- Zacharatos, H., Gatzoulis, C., & Chrysanthou, Y. L. (2014). Automatic Emotion Recognition Based on Body Movement Analysis: A Survey. *IEEE Computer Graphics and Applications*, 34(6), 35–45. <https://doi.org/10.1109/MCG.2014.106>
- Zhang, L., Gillies, M., Dhaliwal, K., Gower, A., Robertson, D., & Crabtree, B. (2009). E-drama: Facilitating online role-play using an ai actor and emotionally expressive characters. *International Journal of Artificial Intelligence in Education*, 19(1), 5–38. Retrieved from http://www.scopus.com/record/display.url?eid=2-s2.0-73449137566&origin=inward&txGid=_GDv8OcXZxR9jLqcfWT6OtG%3A2

Appendix A - Questionnaires

1. Immersive Tendencies Questionnaire

Demographics

- Age: __

- Gender: F_ M_

- Education:

High-school_ Bachelor_ Masters_ PhD_

Immersive Tendencies Questionnaire

1. Do you easily become deeply involved in movies or TV dramas?
Not easily Very easily Don't know
2. Do you ever become so involved in a TV program or book that people have problems getting your attention?
Not involved Very involved Don't know
3. Do you ever become so involved in a movie that you are not aware of things happening around you?
Not involved Very involved Don't know
4. How frequently do you find yourself closely identifying with the characters in a story line?
Not frequently Very frequently Don't know
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?
Not involved Very involved Don't know
6. How good are you at blocking out external distractions when you are involved in something?
Not good Very good Don't know
7. When watching sports, do you ever become so involved in the game that you react as if you were one of the players?
Not involved Very involved Don't know
8. Do you ever become so involved in a daydream that you are not aware of things happening around you?
Not involved Very involved Don't know

9. Do you ever have dreams that are so real that you feel disoriented when you awake?
 Not disoriented Very disoriented Don't know
10. When playing sports, do you become so involved in the game that you lose track of time?
 Not involved Very involved Don't know
11. How well do you concentrate on enjoyable activities?
 Not concentrated Very concentrated Don't know
12. How often do you play arcade or video games? (OFTEN should be taken to mean every day or every two days, on average.)
 Don't play Often Don't know
13. Do you ever become so involved in doing something that you lose all track of time?
 Not involved Very involved Don't know

2. Presence Questionnaire + Tailored questions

Presence Questionnaire

1. How much did the visual aspects of the environment involve you?
 Not involving Very involving Don't know
2. How much did the auditory aspects of the environment involve you?
 Not involving Very involving Don't know
3. How compelling was your sense of objects moving through space?
 Not compelling Very compelling Don't know
4. How much did your experiences in the virtual environment seem consistent with your real-world experiences?
 Not consistent Very consistent Don't know
5. How completely were you able to actively survey or search the environment using vision?
 Not completely Very completely Don't know
6. How well could you identify sounds?
 Not well Very well Don't know
7. How well could you localize sounds?
 Not well Very well Don't know
8. How closely were you able to examine objects?
 Not closely Very closely Don't know

9. How involved were you in the virtual environment experience?
 Not involved Very involved Don't know
10. How quickly did you adjust to the virtual environment experience?
 Not quickly Very quickly Don't know
11. How much did the visual display quality interfere or distract you from performing assigned tasks or required activities?
 Not interfering Very interfering Don't know

Tailored questions

12. How confusing was the presentation given by the avatar?
 Not confusing Very confusing Don't know
13. How confusing was the subject presented by the avatar?
 Not confusing Very confusing Don't know
14. How engaging was the avatar that was presenting?
 Not engaging Very engaging Don't know
15. How comfortable was the lighting condition of the virtual environment?
 Not comfortable Very comfortable Don't know
16. How relevant was the lighting condition of the virtual environment to keep your attention on the presentation?
 Not relevant Very relevant Don't know
17. How noticeable were the changes in the lighting condition of the virtual environment?
 Not noticeable Very noticeable Don't know

Appendix B – Presentation Scripts

1. 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 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!”

2. Rewritten Script

All slides are equal to the original script, except for the fourth and fifth slides, which are listed below.

4th slide: “In addition, there are others that are more complex and linked to situations that are not as emotionally charged. States like confusion, engagement, frustration or boredom occur during cognitive tasks like learning or attending to a presentation. That is, when someone is passing information. Confusion may be the most interesting 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. The longer you are in this state, the closer you lean towards the second threshold, coded as t_b . Past this threshold you get frustrated and eventually bored. In this case, you disengage from the activity and lose attention. Contrary to common sense, it's not always bad to be confused. Thus, confusion is particularly interesting. Cognitive challenges induce confusion. Trying to understand the subject that caused this confusion makes you overcome it. This is a constructive confusion. However, there's a fine line between constructive and non-constructive. Non-constructive confusion occurs when you stay confused for too long and can't overcome it. This will lead you to frustration, which should be avoided.”

Appendix C – Emotional Body Posture Survey

Most of the work related to emotion through body cues has been done from the perspective of how emotion can be automatically recognized from the body, much as what has been done regarding facial expressions or sentiment analysis from text or speech. In fact, it has been on these last modalities that the research has mainly focused on (Kleinsmith & Bianchi-Berthouze, 2013). Nevertheless, recognizing emotion from body gesture has been shown as a promising field with multidisciplinary applications and there are now several surveys about the body to express emotion.

These surveys (Kleinsmith & Bianchi-Berthouze, 2013; Stephens-Fripp, Naghdy, Stirling, & Naghdy, 2017; Zacharatos, Gatzoulis, & Chrysanthou, 2014; Vinayagamoorthy et al., 2006) focus on emotion recognition from body posture whereas Karg and colleagues (Karg et al., 2013) go a bit further and also survey generation systems for body affective expression. They stress out how this field is focused on techniques of AER from body cues and generation of emotional body animation applied to avatars of non-person characters (NPC). Only a few studies explore how these generation systems can be used to enhance nonverbal communication of user's avatars.

Most of the surveyed studies resort to professional optical motion sensors to extract kinematic features and infer emotional states. This apparatus is not viable in a company's workspace and this highlights the relevance of finding ways to provide congruent emotional body animation with ubiquitous devices. The two next papers we present are some of the few studies that were carried when trying to enrich social scenarios with body animation without resorting to motion sensors.

(Pedica & Högni Vilhjálmsón, 2010) study the human territoriality and how can this be applied social VEs. In games like World of Warcraft and Second Life, players can naturally control the movement of their avatars and is usual to engage in social interactions. In these situations, the theory of proxemics (Hall et al., 1968) requires too much micro-management from users to be applied properly. Therefore, the authors developed a model that constantly updates the avatars position and orientation based on the territorial field and the social interaction they are in. Four studies were carried to evaluate the validity of this approach:

- “A person joining a conversation”,
- “A participant moving around within the conversation”,
- “A person trying to avoid a conversation”,

- “A person passing by a conversation”.

For each study there was a control condition where this approach was disabled (there was no automatic orientation and positioning) and another condition where this model was enabled. Overall results show that this approach improved believability and enriched the social scenario. Even though it was not framed in this scenario, this approach seems promising for the ubiquitous scenario we described because it lessens the workload required from the user and facilitates blending VEs in the daily routine of employees. However, there is one caveat to consider. This approach was designed and evaluated considering the user is on a third-person perspective, which is considerably different from the first-person perspective our system employs. Tests would be needed to evaluate if this automatic micro-management was not too intrusive since the virtual camera of the user would be constantly adjusting without any direct input from him. Input from the emotional context of the user could also prove useful to the effectiveness of this approach. A negative or positive emotional state impacts a person’s behavior, which could also reflect on his position and orientation towards others.

(Zhang et al., 2009) present EMMA, a virtual agent that can be incorporated by users and reacts to their emotional context. It is designed for emotionally charged scenarios where the user becomes an actor and must perform according to the role he is given. There are three scenarios: “Big Night Out”, “Homophobic Bullying” and “Crohn’s Disease” (please refer to the original paper for a more detailed description) with different participants that have specific roles. Its AER is performed through text input which has already been proven to be one of the most reliable modalities when using uni-modal recognition (Paulmann & Pell, 2011). Since the scenario is based on an artificial drama we believe that the discontinuity of the text input signal is not a hindrance to the flow of the interaction.

The evaluation was designed under three conditions: 2D with no animation or affect detection, 3D with AI characters and limited animation (no affective animation) and 3D with AI characters and full animation (with affective animation). Results show that transitioning from 2D to 3D with no affective animation just slightly improves their subjective evaluation of the avatars, enjoyment and sense of presence, but when affective animation is introduced their subjective evaluation, enjoyment and sense of presence increases significantly. This suggests that it is not enough to present realistic avatars, but also enrich them with realistic behavior. However, the authors expected that social interaction was greatly improved when moving from 2D to 3D with affective animation, but

what was rather observed was that the significant improvement was verified when moving from 2D to 3D with no affective animation.

This study provides useful insight over how users perceive other avatars' emotional body animation through AER from text. However, it was evaluated with scripted AI characters, which does not fully represent the dynamics of human users interacting in this scenario. An interesting evaluation would be to record an actual performance of the scenarios with human users and let the AER only animate the users' own avatar bodies. This is what we suggest in our first hypothesis. Therefore, in relation to this study, we position ourselves in evaluating the self-perceived congruency of the user's avatar own body and facial expressions according to what emotion he is experiencing, rather than only evaluating how AER can animate AI avatars bodies.