FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

An approach to player's attention modelling in virtual reality environments

António David Casimiro

Mestrado Integrado em Engenharia Informática e Computação

Supervisor: Rosaldo José Fernandes Rossetti, PhD

Second Supervisor: João Tiago Pinheiro Neto Jacob, PhD

26th July, 2018

An approach to player's attention modelling in virtual reality environments

António David Casimiro

Mestrado Integrado em Engenharia Informática e Computação

26th July, 2018

Abstract

The human brain is constantly overwhelmed with stimuli from the surrounding environment, and its processing capacity is limited. Nevertheless, it is able to perceive the surrounding environment in an efficient manner, by filtering out non-relevant stimuli and concentrating on the relevant ones. The attention mechanisms influence such filtering and are executed in a seamless fashion.

Representing such attention processes in a computationally tractable way can be a daunting task. Indeed, no agreement seems to exist in neuropsychology and psychology literature as to how attention mechanisms are conducted, which may not have an exact computational counterpart. Current computational models of attention try to encapsulate those findings in several ways. However, no agreement seems to be found as well on the attention models.

The dissertation proposes a framework in order to capture the visual attention mechanisms. Such framework is divided into two components. The first is focused on the definition of a test bed which can be used as a proxy of the real environment. The definition of a controlled environment, which ensures the safety of the participants and allows the extraction of several parameters from the behaviour of participants, is paramount. The second component is related to the construction of computational models of visual attention, which are responsible for assigning a saliency value to objects. Such value indicates the degree of which the object "pops-out" when a person is observing the environment. The models are composed by two components: the bottom-up component of attention model is focused solely on the properties of the visual stimulus; the top-down factor, modelled by a Gaussian mixture model, is based on data collected from participants immersed in a virtual environment.

Participants were asked their opinion regarding aspects pertaining the immersiveness of the virtual environment. Collected results seem to suggest that the virtual environment is realistic and responsive to actions performed by users. As for the attention model, the obtained results seem to indicate that a poor performance of the model is obtained, with 17% of the conducted observations with a correct classification of the most salient object. Moreover, the contribution of the top-down component is non-existent; in fact, the same performance is obtained when the top-down component is considered, thus acting as a redundant component. Such redundancy suggests that the mixture model acts as a surrogate of the attention model when looking for the most salient object in the scene. On the other hand, the obtained results suggest either the adopted modelling strategy is inadequate, or the chosen performance metric is insufficient, or the validation procedure is not adequate. Further developments are thus required.

Keywords: Virtual Reality; Serious Games; Behavioural Modelling; Modelling and Simulation; Machine Learning; Attention Modelling.

Resumo

O cérebro humano é constantemente sobrecarregado por estímulos provenientes do mundo envolvente, e a capacidade de processamento do cérebro é limitada. Todavia, este é capaz de percepcionar o mundo envolvente de uma forma eficiente, filtrando os estímulos não relevantes. Os mecanismos de atenção intervêm nessa filtragem e são executados automaticamente.

A representação desses mecanismos de atenção através de um formalismo computacional pode constituir uma tarefa árdua. De facto, na literatura da neuropsicologia e psicologia, não existe concordância relativamente à forma como a atenção se processa, e a replicação desses mecanismos sob a forma de estruturas computacionais não é trivial. Os modelos computacionais de atenção existentes abordam, de várias formas, o problema exposto, recorrendo a diferentes metáforas. Contudo, nenhum modelo parece ser capaz de replicar os mecanismos de atenção de uma forma generalizada.

Nesta dissertação, é apresentada uma *framework* que propõe uma nova abordagem à modelação dos mecanismos de atenção. Essa *framework* é composta em duas componentes. Uma primeira componente prende-se com a definição de um ambiente virtual, onde podem ser executadas experiências de uma forma controlada, garantindo a segurança dos participantes e a recolha de vários dados dos participantes. A segunda componente foca-se na definição de modelos de atenção visual, capazes de ordenar os elementos observados de acordo com o seu grau de saliência. Tais modelos de atenção são compostos por uma componente *bottom-up*, que se foca unicamente nas propriedades do estímulo visual por forma a calcular o grau de saliência, e por uma componente *top-down* baseada em *Gaussian mixture models* responsável por modular a saliência proveniente da componente anterior, incorporando aspectos cognitivos. Os dados recolhidos pelos participantes serão usados para a construção de tais *mixture models*.

Quanto ao ambiente virtual, as respostas dadas pelos participantes parecem sugerir que este é realista e responsivo às acções dos utilizadores. Quando ao modelo de atenção, os resultados obtidos parecem sugerir uma fraca performance deste, com 17% das observações realizadas com uma correcta classificação do objecto mais saliente. Além disso, a mesma performance é obtida quando a componente *top-down* é considerada no modelo, actuando, assim, como um componente redundante; esta redundância pode sugerir que o *mixture models* pode funcionar como um substituto ao modelo de atenção. Por outro lado, os resultados podem sugerir que a abordagem adoptada para a componente *top-down* não é a mais adequada, que a métrica de performance adoptada pode não ser a mais apropriada para o problema em mãos, ou que o procedimento de validação pode não ser o mais acertado. Por conseguinte, é necessário proceder a desenvolvimentos futuros do trabalho.

Palavras-chave: Realidade Virtual; Jogos Sérios; Modelação Comportamental; Modelação e Simulação; *Machine Learning*; Modelação da Atenção

Acknowledgements

The work herein presented could not be possible without the help and support of some people.

First and foremost, I would like to thank all the participants that took part of the experiments. Specially, I would like to thank Xavier Fontes for his support on conducting the experiments. I would also to thank Professora Brígida Mónica Faria for her help and suggestions regarding the questionnaire used for the experiments. Acknowledgements are also due to project SIMUSAFE and, in particular, to ITCL (Technological Institute of Castilla y León); they provided a fantastic 3D urban scenario, which was a precious help for my work.

I would like to take this opportunity to thank my friends (in no particular order) Ana Amaral, Diogo Amaral, Pedro Câmara, Pedro Silva, Carlos Alves, Mauro Rodrigues, Ricardo Duarte, Vasco Pereira, and Rui Cardoso, as well as my fellow colleagues. If I have forgotten to put the name of someone, my sincere apologies; I will put your names in double in my PhD thesis.

My fellow labmates at LIACC must also be thanked – Tiago Neto, and João Neto – for all the help you provided me and for all the funny moments we had. I wish you the best of luck for your PhD. Also, I would like to thank (with repetitions) Tiago Neto, João Neto, Professor João Jacob, Professor Daniel Silva, Sara Paiva and Rui Andrade for introducing me to Dungeons & Dragons and for the Friday afternoon sessions; I've never had so much fun and I became a huge fan of the game.

I am grateful to Professor Rui Rodrigues, Professor Rui Nóbrega, Professor Henrique Cardoso, Professor Daniel Silva, Zafeiris Kokkinogenis and Thiago Rúbio for their support, comments and discussions regarding the developed work, which contributed to a richer outcome of the dissertation.

My sincere thanks to once again Professor João Jacob, which provided me a huge support throughout the dissertation. His help when technical problems arose was crucial for the conducted work. Moreover, when I was having many doubts regarding the focus of the dissertation, his support gave me courage to pursuit the selected research area and bring a closure to my dissertation. Of course, many thanks are due to my supervisor Professor Rosaldo Rossetti for giving me this dissertation proposal and supporting me throughout the dissertation. With his guidance, I was able to finish my dissertation, and therefore my course. Who would have thought he would be my supervisor when we first met in the first year of college... I am truly honoured that we have made acquaintance.

Finally, I would like to thank my family. In particular, I would like to thank my mother Ana Assucena, for supporting me throughout the dissertation, the difficulties that came up in college and throughout my life. You are a force of nature, and you are always showing me to never give up. I love you so much and thank you for everything, mom!

António David Casimiro

"Pedras no caminho?
Guardo todas, um dia vou construir um castelo."

Fernando Pessoa

Contents

List of Figures

List of Tables

Abbreviations & acronyms

2D	Two-dimensional
3D	Three-dimensional
API	Application Programming Interface
AR	Augmented Reality
CNN	Convolutional Neural Network
CPU	Central Processing Unit
DoG	Difference of Gaussians
FIT	Feature Integration Theory
FLOPS	Floating Point Operations per Second
FPS	Frames per Second
GPS	Global Position System
GPU	Graphics Processing Unit
HLSL	High Level Shading Language
HMD	Head Mounted Display
ID	Identification
IQR	Interquartile range
IVE	Immersive Virtual Environments
LoG	Laplacian of Gaussian
PDIS	Preparação para a Dissertação
R-CNN	Region-based Convolutional Neural Network
RGB	Red, Green, Blue
ROI	Region of Interest
UI	User Interface
UX	User Experience
VE	Virtual Environment
VR	Virtual Reality
WYSIWYG	What You See Is What You Get
XR	Crossed Reality

Chapter 1

Introduction

The human brain is constantly overwhelmed with stimuli that come from the surrounding environment, which represent a large amount of information to be processed by the brain. In terms of visual stimulus, it is estimated that 10^7 - 10^9 bits per second arrive at the optic nerve [IK01a, BI13]. Nevertheless, a large portion of the stimuli can be filtered out [DD95]. Such filtering occurs due to the fact that the brain has a limited capacity in processing stimuli, allowing less relevant information to be discarded. This, in turn, enables a more efficient response to environment changes and behavioural goal achievements [KC14]. Such information processing bottleneck is implemented by means of attention mechanisms [IK01a, KC14]. What is remarkable is that this mechanism is still effective even in the presence of noise in the environment (for instance, highly cluttered scenes) [GHV09]. The presence of attention mechanisms can be supported from an evolutionary perspective as a survival mechanism, as it allows to rapidly detect preys or threats in the environment [TIR05].

Attention has been an active research area [Wol00, BI13]. From a psychological and neurobiological perspective, such studies are important so as to have a better understanding of how humans perceive the surrounding environment. The ability to understand what and how objects capture the user's attention can reveal to be of utmost importance for marketing, for instance [WP08, WP06]. From a computational perspective, the role of attention as an information bottleneck is very appealing. Although millions of floating point operations per second (FLOPS) are feasible for today's computers, tasks such as image captioning [AHB+17, XBK+15, YJW+16] and realistic virtual human behaviour [HB05, CB01] can still represent a daunting endeavour. Moreover, the ability to reduce the complexity of information processing is relevant for real-time computation scenarios [KOS11, ZH07, BAA11, TCK+95].

The study of attention, in particular that of visual, usually requires the analysis of behaviour in terms of attention in a controlled way. User's position, eye gaze coordinates and observed objects are useful metrics to perform some sort of attention behaviour analysis. Equipment such as eye-trackers and GPS receptors are thus needed so as to obtain such metrics. However, in a real world

scenario, multiple problems may arise that may hinder the proper execution of the experiments. Logistics problems and costs are some examples. Additionally, liabilities for the participants represent a major factor to be taken into consideration. For example, if one wants to test an hypothesis which involves reaction times from car drivers when pedestrians are crossing the street, it is not viable to execute the experiment in a real scenario. Furthermore, the execution of experiments in the real environment may lead to a low degree of reproducibility. Maintaining weather conditions and position of the elements in the environment across different iterations of the experiment are some of the parameters that may reveal of being difficult to control. The usage of serious games can represent a possible alternative when conducting the experiments. Experiments can be executed in a controlled environment, where various parameters can be constrained, and data can be collected. The military and education are some examples of fields where serious games can be used [GA16]. Road safety is another example of such fields [GRP+14].

1.1 Problem Statement

The problem statement is as follows:

> Is it possible to ubiquitously associate attention factors to objects, as they are perceived, in virtual environments? And, if so, is it possible to define a model in which such factors can be obtained in a computationally tractable way?

In other words, the dissertation will see to what extent it is possible to define a framework capable of sorting out the elements of the virtual environment in terms of attention. Such sorting should be as realistic as possible, capable of mimicking the biological processes that occur in the visual cortex [Mat, Tre03], and a computational representation of the model should be attained. Serious games will be used as a foundation for the experiments to be performed pertaining the developed model.

1.2 Objective

The objective of the dissertation is then to define a meta-model capable of outlining a function s

$$s\colon (p, o, Env, elem) \rightarrow \mathbb{R} \tag{1.1}$$

, where

s	: the attention function, representing the degree the element *elem* catches the user's attention
$p \in \mathbb{R}^3$	: the position of the person in the environment
$o \in \mathbb{R}^3$	: the person's orientation towards the surrounding environment (e.g. the heading of the field of view)
$Env \subseteq \mathbb{R}^3 \times \mathbb{R}^3$	: a function describing the placement and rotation of the elements in the virtual environment
$elem \in \mathbb{R}^3 \times \mathbb{R}^3$	: the element of interest

Note that

$$elem \notin Env \implies s(p,o,Env,elem) = 0$$

, meaning that objects not located in the environment do not have relevance in terms of attention. The model will take the collected data from the experiments conducted in a virtual environment in order to support the outputs. Note that no assumption is made as to the specific meaning of the value returned by s. Instead, this value should be used as a comparison factor among elements of the environment in terms of attention relevance to the user.

Therefore, the contribution of the dissertation is as follows:

- the proposal of a virtual environment in which users are immersed and experiments are conducted in a controlled way.
- the proposal of a visual attention model capable of labelling elements of the virtual environment according to their level of attention conspicuity.
- the definition of an experimental protocol to conduct the experiments in a systematic fashion.
- the proposal of a validation approach responsible for assessing the performance of the attention model, and to assess the user experience regarding the used virtual environment.

1.3 Motivation

The development of the project herein described is driven by multiple factors. The ability to replicate the attention mechanisms that occur in the visual cortex in a computational way represents an active topic of research. The literature review presented in Chapter 2 is the embodiment of such activity. From a medical perspective (such as psychology and neuropsychology fields of research), the understanding of the attentional influences and mechanisms represents a goal to be achieved. What catches the user attention, how a scene is perceived and processed by the human brain are some of the research topics in the field of attention [LG14].

From a computational perspective, real-time computation can benefit from the selective characteristic of the attention. Filtering out non-relevant inputs can translate into more responsive systems. The robotics, video/image processing and graphics fields can benefit from that [DOS17].

The research areas of behaviour and attention modelling can also take advantage of progresses in attention modelling research. Realistic actor modelling, as well as artificial entity generation, are examples of applications which can benefit from the outcome of the dissertation. The evaluation of how information is disseminated in a virtual environment and of the impact such information can have on virtual environments, with regard to their ability to retain user's attention is another example of application. Moreover, a social impact can be established. The creation of responsive and adaptive environments capable of reacting to the degree of attention of pedestrians is of utmost importance for road safety. The project SIMUSAFE [1] represents the outcome of the need to comprehend how the road environment can be defined so as to reduce the mortality resulted from reckless behaviour. The autonomous driving area is other of the examples where the prediction of the behaviour of the pedestrians and drivers is very important, and, therefore, a motivation for tackling the problem.

1.4 Dissertation Layout

The dissertation will be divided into the following chapters. In Chapter 2, a literature review is exposed. Since the development of a computational attention model is the focus of the dissertation, the subject matter of the chapter is to give an overview as to what has been developed in terms of proposed model architectures, taxonomies, as well as examples of application of such models. A preliminary exposition of introductory concepts is also given so as to allow the reader to be more familiar with the addressed concepts. A description of the proposed model architecture is given in Chapter 3. A section is dedicated to each of the two components that compose the attention model. Implementation details are also described in a proper section, and limitations of the adopted implementation are reported as well. A description of the conducted experiment, including the necessary equipment, is given in Chapter 4. The experimental protocol is outlined, and an explanation of the virtual environment used for the experiments is stated. The description of the obtained results and respective discussion is also provided. Finally, conclusions taken from the developed work are outlined in Chapter 5. A description of possible future work is also provided in the same chapter. So as to complement the dissertation, appendices related to the conducted experiments such as the used consent declaration, the questionnaire and respective responses are provided in Appendix A.

[1] project reference: 723386 - http://www.simusafe.eu

Chapter 2

Literature Review

In order to obtain a function of attention in a computationally tractable way, computational models of attention must be considered. As such, this chapter will describe the current research status in terms of computational attention models. However, before such listing, one must grasp the fundamental concepts of the area. Even such understanding falls outside the computer science field, it is necessary to grasp the concepts if one wants a realistic model. Therefore, the current chapter represents the effort conducted by the researcher to comprehend the concepts.

The current chapter is divided into three sections. The first one presents a series of fundamental concepts that will be introduced so as to allow the reader to better understand the concepts used throughout the literature review. The next section will focus on computational models of attention. A focus on the advancements in the area of serious games and data mining clustering analysis is presented in the third section. Lastly, conclusions will be drawn in the last section.

2.1 Visual Attention

An introduction to the human visual system is given in this chapter. An attempt at defining attention is given, and the structure of the human eye is described as to support the need of mechanisms of perception filtering. The theories of attention are described, and their embodiment as computational models is presented. Unless stated otherwise, [LG14], [BI13], [Sin14], and [IB15a] represent the main literature considered for this chapter. Additional literature is mentioned as needed.

2.1.1 A Definition of Attention

When trying to define the attention mechanisms that take place in the human brain, the following question imposes:

What is attention?

This has been a question several researchers have been eager to answer. However, no clear answer seems to arise [LG14].

Indeed, the definition of attention has changed as new findings are uncovered. Wilhelm Maximilian Wundt [Wun93] gives a definition of attention based on the various degrees of representation that concepts and ideas are held in consciousness. He proposed a distinction between *perception* (when the stimulus is processes by the attention processes) and *apperception*, a key concept of his theory, that occurs when stimuli enter higher regions of the processing chain. Gottfried Leibniz [LAG+89] notes that attention is required if one wants to become aware of some event. William James [Jam90] defines attention as the focus and concentration of consciousness, which implies throwing away "some things" in order to prioritise others. Luria [Lur73] and Solso [Sol88] emphasise the role that attention takes when filtering perceptions. In fact, the latter proposes that attention is the act of actively processing a small portion of the information that comes from the exterior.

No final and clear conclusion can be taken in terms of what attention is. A high-level, vague definition can be, however, established. Li *et al.* refers to attention as "allocation of cognitive resources on information" [LG14], and Sinai proposes attention as a "high level cognitive process that is cross-modal" [Sin14]. Since the dissertation's area of research is not one of psychology nor one of neuropsychology, such definitions are sufficient and represent the basis for the work of the dissertation.

With an established definition of attention, one may wonder the following:

What types of attention do exist?

Sohlberg and Mateer [SM89] categorise attention into 5 main types by considering how cognitive and behavioural resources are allocated:

- **Focused attention**: the ability to allocate resources to stimuli in a *discrete* fashion (concentrate on a part of the stimulus)
- **Sustained attention**: the ability to allocate resources to stimuli for a steady period of time
- **Selective attention**: the ability to filter out stimuli by concentrating on specific stimuli while ignoring distracting or competing stimuli
- **Alternating attention**: the ability to switch resource allocation between different contexts
- **Divided attention**: the ability to concurrently attend to different contexts

Although attention can be classified in the aforementioned types, studies in the field of computer vision tend to focus on selective *visual* attention. The developed architectures try to emulate the processes that occur at the human visual system on how the visual stimulus is perceived in a selective fashion, focusing on relevant elements of the scene while ignoring non-relevant ones. The next question that imposes is the following:

How can visual saliency affect the selectivity?

Imagine that you are looking at a painting. Neither the human brain nor the eye have the ability to attend to all details of the painting in the same way. Thus, a selective process takes place in order to focus on areas that are more conspicuous, and then shift the gaze so as to attend to the next most conspicuous part. In order to explain the different order of visit of the conspicuous areas, a possible explanation is that an importance value (*a saliency value*) is given to each part of the painting, thus establishing a ranking. This saliency value can be then used to guide the attention process in order to visit those areas. Therefore, one may conclude that visual saliency takes an important role in the attentional mechanism, by "allocating limited perceptual and cognitive resources on the most pertinent subsets of the sensory data" [LG14]. This optimal deployment of resources is very important for complex biological systems (such as humans) to allow a rapid detection of preys, predators or mates in a cluttered visual world. In other words, it allows the visual system to focus on relevant parts of the visual stimulus [IK01a, Itt07]. One may be wondering how these processes occur biologically. Neurobiological and psychological fields have no clear answer. Several tools such as electroencephalography (EEG), magnetoencephalography (MEG) and functional magnetic resonance imaging (fMRI) have been used in order to understand and define a function capable of encompassing the information processing mechanisms. Some applications of the aforementioned tools are given in [LCC13], [KTM+16], and [PMKA07], respectively. Also, a reference to these methods is given by Tsotsos *et al.* ([TIR05]).

2.1.2 Biological Mechanisms of Human Vision

The human eyes are the receptors of visual information. These are equipped with photosensitive cells located at the retina, that are stimulated by visual stimuli. Therefore, the retina is responsible for converting electromagnetic information into nerve impulses, to be later processed by the brain. In fact, the retina is composed by multiple layers of cells, and, at each layer, the visual stimulus is transformed by different processes:

1. **Photoreceptors and horizontal cells**: cones and rods are the photoreceptors responsible for converting light into electrical signals. Horizontal cells take part of the transformed light stimulus in order to obtain an average of the outputs.

2. **Bipolar cells and amacrine cells**: the bipolar cells are responsible for calculating highly contrasted signals. This is done by taking the difference between the output of the photoreceptors and the horizontal cells. A second local average is obtained by amacrine cells.

3. **Ganglion cells**: usually located near the inner surface of the retina, these cells take the output created by the bipolar and amacrine cells and further compress the signal to be sent through the optical nerve.

A graphical description of the layers is presented in Figure 2.1.

An important distinction must be made regarding the roles of the cones and rods as photoreceptors. The former are responsible for colour detection and visual acuity. However, cones require

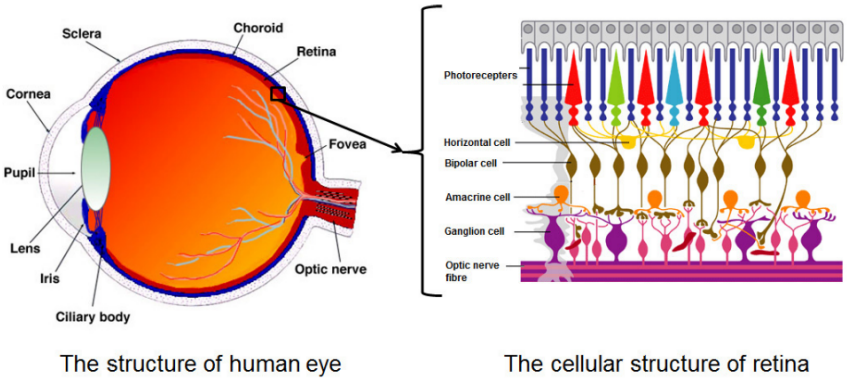


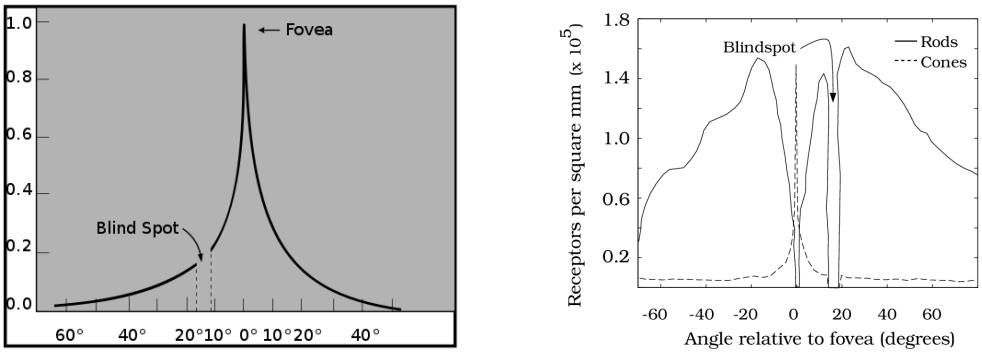
Figure 2.1: The structure of the human eye and respective cellular structure (Source: [LG14], which provides an adaptation from [WB04])

large amounts of photons in order to properly perceive the visual stimulus. On the other hand, rods are extremely sensitive, allowing them to perceive the visual stimulus even in low-light conditions, but colour perception is not made. Moreover, an analysis of the distribution of these cells throughout the retina shows a heterogeneous distribution, as show in Figure 2.2. Such distribution hints the fact that the visual stimulus is not perceived in a uniform way by the eye. In fact, high acuity values are located at the fovea, which is precisely where the largest concentration of cones is located. No photoreceptors are located at the "blind spot", which is the location where the optic nerve exits the eye.

It is worth noting that the retina can been seem as a compression filter. Indeed, 128 million photoreceptors [SSM08] are estimated to exist in the retina (120 million rods and 8 million cones). However, only one million optical nerves are available to transmit the nerve impulses to the brain. Therefore, the existence of compression mechanisms based on averaging and differencing operations is an assumption presented by Li *et al.* [LG14].

In addition, experimental studies have show that **centre-surround mechanisms** are implemented in the retina by establishing connections between bipolar and ganglion cells. Specifically, 2 antagonistic mechanisms can be outlined: *on-centre*, with an excitory central region and an inhibitory peripheral region; and *off-centre*, representing the reverse scenario. Such mechanisms are based on the connections between the cells. The properties given by the mechanisms allow to detect local irregularities in the visual stimulus (for example, edges).

With the visual stimulus converted into nerve impulses, the information is transmitted to the brain by the optical nerve. Throughout the various stages of processing at the brain, the information is refined. It is believed that several areas of the brain are focused on processing the visual stimulus in terms of various modalities such as colour and motion. While the visual information is being transmitted to higher brain regions, the stimuli compete with each other so as to gain relevance and become salient. Those who cannot compete in terms of saliency are filtered out. The



(a) Horizontal cross section of the acuity of the human eye (Source: Adaptation made by Vanessa Ezekowitz from an illustration of Hunziker [Hun06] - https://commons.wikimedia.org/wiki/File:AcuityHumanEye.svg.

(b) A schematic representation of the eye's retina cell distribution (Source: [Wan95]).

Figure 2.2: Physiology of the human eye.

saliency returned by this process is called **bottom-up saliency**, and its name is based on the fact that saliency is defined as stimuli goes from primary brain regions (bottom) to higher regions (top) of the brain. Since no cognitive processes occur at this stage, the process can also be referred to as a *data-driven* process.

However, cognitive processes take place while a person is perceiving the environment. In fact, neurons are capable of outputting different responses even when the properties of the visual stimulus is the same. Prior information such as task, goals, prior knowledge, or experience located at higher regions of the brain have an impact on how the stimulus is perceived. While the information goes from "bottom" to "up" regions, some biases are applied to the saliency competition. More specifically, the biases may tend to support semantic reasoning (object-based and cognitive biases), spatial locations (spatial biases) and feature dimensions (feature biases) [IB15a]. Reusing the previously introduced metaphor, such biases flow from "top" to "bottom" regions, thus naming the process as **top-down process**. In other words, the bottom-up processes are modulated by top-down signals, meaning that, for the same visual stimulus, different saliency responses can be obtained.

Two types of top-down factors can be established. *Volitional* factors are mainly focused on "acts of will". For instance, when a person is assigned a task to fulfil and asked to perceive the surrounding environment, these factors contribute for the final perception of the scene. On the other hand, *mandatory* factors cannot be voluntarily suppressed completely. It is believed that prior knowledge and experience are the main contributors. An experience conducted by Chun and Jiang [CJ98] showed that experience in perceiving scenes is transferred to new scenes with similar layout, showing the presence of top-down processes even when no volitional processes are present.

As for the modulation process of bottom-up responses, two approaches are outlined [LG14]:

- **additive baseline shift**: attended areas should have its saliency increased by a constant factor.
- **multiplicative gain modulation**: attended areas should have its saliency increased by a multiplicative factor.

These modulation processes are of utmost importance when defining a model of visual saliency that incorporates top-down processes. Indeed, a strategy must be used when one wishes to modulate bottom-up responses.

2.1.3 Psychological Theories of Visual Attention

So as to complement the aforementioned, a psychological approach to the visual attention process is presented. Phenomena such as pop-out, set-size, attention unit of measure, the registering of visual features, and contextual cuing effect are presented.

2.1.3.1 Pop-Out and Set-Size

In a visual search task, when a person is asked to search for an object with a particular set of feature, two phenomena take place. The *pop-out* effect occurs when the target object possesses a set of features that allows to distinguish it from the other distractor objects. The bottom-up processes "pop-out" the target object, making it more salient. Moreover, this process is independent from the number of distractors, thus being an efficient process to find salient objects.

On the other hand, when the target and the distractor objects share, at least, one feature, the pop-out effect is not enough, and the *set-size* effect takes place. In fact, conflicts in terms of saliency occur. Therefore, so as to find the target object, top-down processes are required, making the process dependent on the number of distractors. Moreover, the top-down modulation process is slower when compared with the time required by the bottom-up process, thus requiring a greater amount of time to process the stimulus. These effects can be exemplified in Figure 2.3.

Based on this fact, attention models usually consider the bottom-up module as their first one. If the target object "pops-out", it is almost immediately identified. A top-down component, if present, is located at a later phase.

2.1.3.2 Unit of attention

When perceiving a visual stimulus, what is the unit of work the attention mechanism takes into account? The ability to pinpoint such unit is of considerable interest for visual saliency computational model. Several studies have been conducted with the purpose of providing some enlightenment regarding this topic.

Three different units have been proposed: location, feature or object. When conducting the experiments, studies fix two of the aforementioned units while varying the third one. However, no consensus seems to be reached, as some studies propose that the attention is directed towards

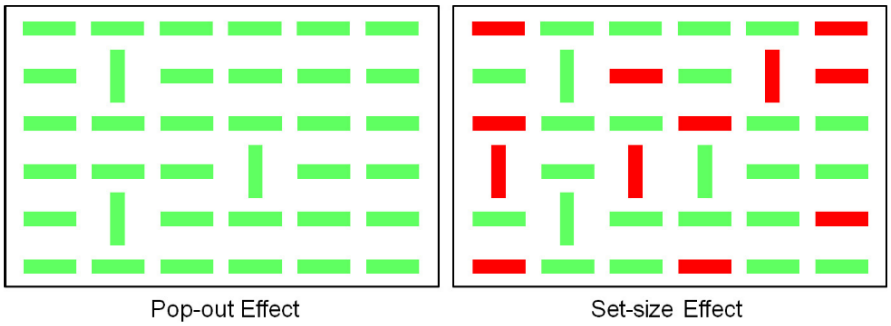


Figure 2.3: A demonstration of the pop-out and set-size effects (Source: [LG14])

salient locations [EE74, CPGV97], while others suggest that attention operates at the level of features or objects [ORK+97, AVLH98, ODK99].

The **Feature Integration Theory** (FIT) [TG80], proposed by Treisman and Gelade, is one of the renowned psychological models of human visual attention. It proposes that features of the visual stimulus are processed automatically in a parallel fashion in the early stages of perception to as to "pop-out" certain feature dimensions. If no feature is found to be distinct from the remaining ones, a serial search is conducted at a latter phase. At this phase, features are bound together and considered as a whole for object identification. Therefore, two stages can be identified: a *feature search*, which occurs in parallel and does not require the involvement of attention mechanisms; and a *conjunction search*, which occurs in a serial fashion and requires the allocation of resources to the attention mechanisms (thus, the attention works as a "glue" responsible for aggregating the individual features into perceptions [Gol09]). The stages can also be classified in terms of involvement of attention mechanisms: a *preattentive stage*, which corresponds to the feature search stage since no attention is required, and a *focused attention stage*, which corresponds to the conjunction search stage, as attention mechanisms are used.

The candidate visual features considered in the preattentive stage are called "preattentive features". However, no clear idea as to what features are considered as candidates is available. In computational visual saliency attention literature, colour, intensity and orientation are the most considered spatial features. For temporal saliency analysis, motion and flicker are the considered features [BI13].

2.1.3.3 Contextual Cueing Effect

As stated in previous sections, visual saliency is based on bottom-up factors modulated by top-down ones, in which the latter ones take into account the task to be performed, prior knowledge as well as experience. Experiments have shown that the top-down component of attention possesses some learning capabilities. Moreover, experiments have also shown that top-down factors take place even when no volitional processes are active. So, the following question can be formulated: why does top-down attention possesses some learning capabilities, and how is the acquired

knowledge transferred?

In fact, the visual world presents a highly structural format, which allows little space for randomness. This means that it is possible to encounter scenes that are somewhat similar, in which some properties (for instance, the scene elements' layout) are frequent among themselves. If that is the case, the visual system will try to minimise the amount of resources willing to spend [CJ98]. Features and spatial locations with lower probabilities of containing the target elements tend to be avoided. Hence, a quicker response can be achieved. In psychology, such phenomenon is called **contextual cuing effect**, and can be defined as "an attentional facilitation effect derived from past experiences of the visual world" [LG14].

When perceiving a new scene, the invariant stable properties of the scene (layout, for example) are obtained as a summary of the scene and used as a "lookup key" in the long-term memory for similar past experiences. If one is found, it is used to support the top-down biases, thus allowing an efficient processing of familiar scenes; otherwise, learning processes occur.

2.1.4 Computational Models of Attention

The objective of computational models of attention is to represent the visual attention of human beings in a computationally tractable way. Taking into consideration the available forms of attention, the models can be classified as **bottom-up attention models** and **top-down attention models**.

Laurent Itti and Ali Borji [IB15b, BI13] have proposed a taxonomy for bottom-up, saliency-based attention models, where such models can be segmented according to their definition of what represents a salient region and in what way that saliency can be obtained and represented. It is worth noting more than one category may be used to classify a given model.

The **cognitive models** take into consideration psychological and neuropsychological findings in order to build a master saliency map of attention, which encodes the areas of the scene that, when comparing with theirs surrounding region, are more conspicuous, and thus more susceptible to grab the focus of attention. Most models framed in this category are strongly influenced by the aforementioned Feature Integration Theory [BI13]. The development of such saliency maps are based on computer vision techniques, which are used to decompose the visual input into elementary visual features (such as colour, intensity, or orientation) at multiple spacial scales. Then, those features, using centre-surround differences and normalisation operators, are combined so as to produce a map capable of showing the most salient regions. These models represent the first attempts to model the visual attention, and they are based on findings stated in the Feature Integration Theory. The architecture outlined by Koch and Ullman [KU87] represents one of the first attempts of modelling bottom-up attention, and it is depicted in Figure 2.4. In their work, a set of elementary features of the visual stimulus (such as colour, orientation, luminosity, motion, among others) is obtained at different scales – forming a pyramid representation – using linear filtering (such as Gabor filters, which is widely used in computer vision) and is represented into different topographical maps, to which they call "the early representation". At a latter stage, such features are combined into a single map capable of exhibiting saliency properties. In order to

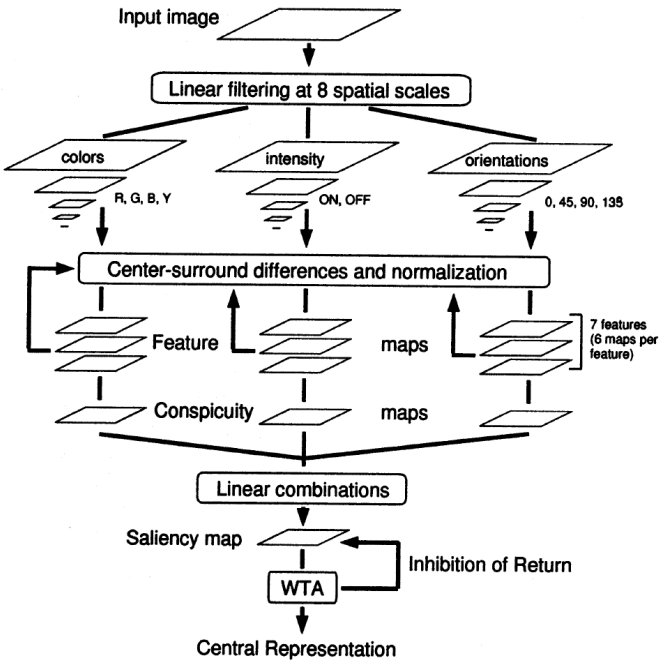


Figure 2.4: Diagram depicting the proposed architecture by Koch (Source: [IK00])

iterate throughout the salient regions of the stimulus, a *winner-take-all* network (WTA) – which selects the most salient region – along with an inhibition of return mechanism – to iterate throughout the salient regions in a descending order of saliency [AL10]. The work of Itti *et al.* [IKN98] represents one of the first attempts of implementation of the proposed architecture by Kock *et al.*.

In **information-theoretic models**, salient areas represent the most informative regions of the stimulus when comparing to neighbour regions. In other words, these models allocate a higher saliency to regions with rare features (that is, with low likelihood). In terms of information theory, the rarity of such features can be seen as the degree of surprise. The Shannon's self-information measure [Sha48], given by

$$I(F) = -log(p(F)) \tag{2.1}$$

, where F represents the feature at hands, is thus used in this context in order to measure such rarity. A fitting of a probability distribution $p(F)$ is used, which can then be used to extract the most salient regions (with low probability) of the stimulus by performing $p(F)^{-1}$ at every position on the image. The model proposed by Bruce *et al.* [BT09], the Attention based on Information Maximisation model, is one of the models framed in this category. In the proposed model, so as to reduce the dimensionality of the problem ($M \times N \times \#(RGB)$), Independent Component Analysis (ICA) [Com92] is used and probability density functions of the features are calculated. The self-information measures are obtained from the joint likelihood of the probabilities of the density functions of the coefficients obtained by ICA, which are then used to obtain the most informative areas of the image.

Other types of models resort to probabilistic frameworks in order to model the attention mechanisms. Such models are called **graphical models**, and they can resort to approaches such as (Dynamic) Bayesian Networks, Hidden Markov Models, and Conditional Random Fields so as to model saliency, as Borji *et al.* suggest in their survey [BI13]. Avraham *et al.* proposed a stochastic model [AL10], in which the saliency of a given area is defined as a probability. As stated in their work, the resulting model should not try to give a plausible explanation for the human attention, but instead it should give an idea, in a probabilistic sense, of how relevant is a given region of the stimulus. In a mathematical sense, the goal is to find an estimation of the probability $p(l_i = 1)$, which gives the probability of the region i being relevant (the value 1 represents the label of relevant areas).

The **decision-theoretic models** state that "atention is driven optimally with respect to the task"[IB15b] at hands. From the unifying perspective of Gao & Vasconcelos [GV09], which tries to combine bottom-up and top-down attention components, saliency can be described as a *discriminant process*, that is, features that best tell apart the area of interest of the stimulus from the surrounding environment are considered as salient features. Therefore, saliency represents the measure of discriminability (as a score) of the visual stimulus with respect to the previous classification [RSFL15].

The aforementioned models perform the detection of salient areas in the spatial domain. However, **spectral analysis models** perform such detection in the frequency domain of the image. Jou & Zhang [HZ07] proposed a model in which the Fourier transform is used so as to detect the "novelty parts" – $H(Innovation)$ – of the image, that is, the salient areas. According to them, and following a information theory perspective, the information of the image $H(Image)$ can be decomposed as

$$H(Image) = H(Innovation) + H(Prior\ Knowledge) \tag{2.2}$$

, where $H(Prior\ Knowledge)$ represents the invariant parts of the signal, which represents the "global" information of the image. In natural image statistics, scale invariance – also known as $1/f$ law, is one of the invariant factors that can be found in images. Based on this fact, the amplitude of the averaged Fourier spectrum can be used as the invariant part of the image. The image is processed in the frequency domain in order to obtain the log amplitude spectrum, and, then, the invariant part of the image is removed from the log spectrum so as to obtain the spectral residual (that is, the unexpected part of the image). The saliency map can be obtained in a subsequent phase by performing the inverse Fourier transform on the spectral residual [WCJ11, LG14].

Lastly, the **pattern classification models** resort to machine learning techniques so as to retrieve a mapping between the stimulus and the saliency of such stimulus. Peters & Itti [PI07] proposed a framework that fits in this category. Their framework is composed by two modules encompassing the bottom-up and top-down aspects of attention. The bottom-up component is based on the work proposed by Koch & Itti [IKN98], already mentioned at the beginning of the section, which captures the relevance of visual areas from low-level features (colour, contrast, among others). The innovative part resides on the top-down component, where a broad perspective of

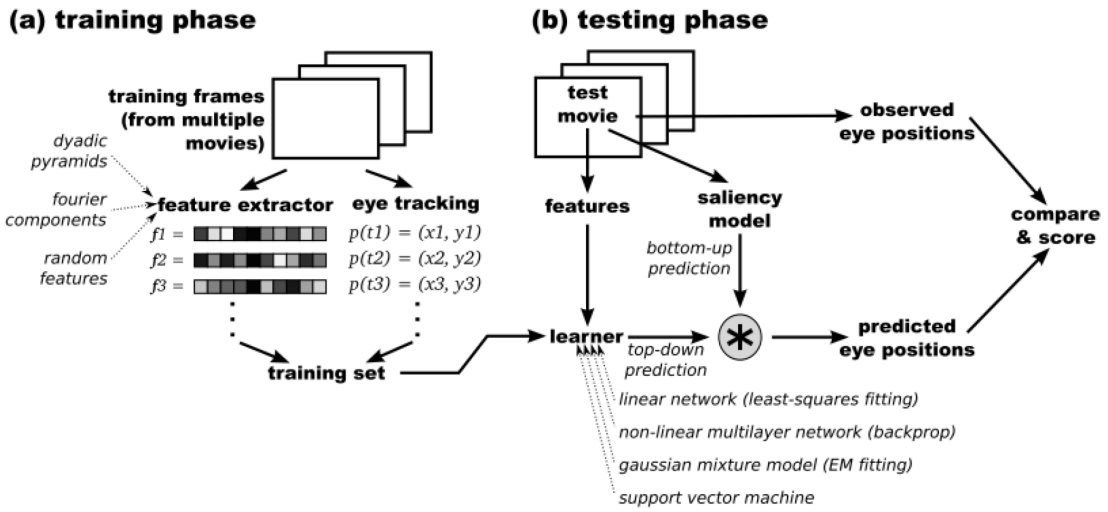


Figure 2.5: Workflow of the proposed task-dependent model, in which top-down attention elements are considered [PI07]

the visual stimulus is considered and a goal-oriented task is imposed. This *broad perspective* is also known as *gist*, which is the general perception and meaning (the basic-level category of the scene, the spacial layout of the elements of the scene, among other factors) of the scene a person have in a fraction of a second, which can be useful when users have to perform a task with a goal in mind, since top-down attention will focus the attention to areas of relevance to the task at hands [Oli05, Ren00]. Figure 2.5 gives an overall perspective of the adopted work flow in order to build a model capable of predicting eye positions. As such, the "gist" of the scene and the eye positions are considered so as to build a map capable of predicting the eyes' position of the user. It is worth noting that gaze is often considered as a surrogate for attention [BI13, HB05]. In addition, Yarbus [Yar67] has shown that gaze patterns take into account the task being performed. Therefore, this model can represent a good predictor for visual saliency.

Regarding computational models capable of modelling *top-down* aspects of attention, some investigation have taken place in the last years, although with less vigour. Indeed, when comparing with bottom-up attention models, modelling cognitive aspects of attention can become a daunting task, due to the complexity of modelling and representing such processes, as well as due to the nonexistence of consensus in the psychology field pertaining to the processes that occur at the level of attention. In this models, a representation of the concepts *goal* and *task* is required, so as to emulate the cognitive processes that take place in the human brain. In this models, the saliency-based mechanisms take into consideration the biases of top-down attention, which allows the saliency calculus to be more task-aware. The usage of explicit cognitive reasoning mechanisms (which takes into consideration the task at hands and updates the relevance of the selected areas) or the usage of fuzzy logic (which incrementally learns the features most relevant for the task at hands, and generates a "top-down signal" that allows the dynamic adjustment) are examples of techniques that consider cognitive concepts.

2.1.4.1 Related work

Some domains of application can be outlined where the use of attention models is predominant. Visual attention models have been used for the purposes of *image captioning and visual question answering*. In particular, Anderson *et al.* [AHB+17], proposes a model based on region-based convolutional neural network (R-CNN), in which R-CNNs are used to create a bottom-up response modulated by a long short term memory (LSTM) network layer responsible for defining top-down biases. These models can then be used to query images regarding their content.

Attention models can also be used for image compression. This can be an interesting research field, since the amount of available information is increasing daily [MB12, Joh14]. Itti [Itt04] applies a saliency-based attention model to video feeds, in which salient regions are kept with full detail, whereas non-salient regions suffer a compression factor. Approximately, a compression by half of the original size can be achieved.

The development of realistic avatars is also another area of application of attention models. Such realism is of utmost importance when high-fidelity is required (for instance, serious games – Section 2.2 gives a brief description on what serious games are, including some references about domains of application). Peters *et al.* [PO01] proposed a model for autonomous virtual environment capable of perceiving the virtual environment by using human-like attention mechanisms. This allows the processing module to be more focused on relevant elements of the environment while discarding non-relevant ones. Artificial gaze and motion generation research is proposed as well [CB01, CB01, BCC99]. The objective is to endow artificial entities with behaviours similar to those of humans. Simulation and animations may take benefits from these contributions as these areas usually require the replication of natural behaviours with high level of fidelity.

Other applications may take advantage of the physiology of the human eye. As described in Section 2.1.2, the visual acuity is very high at the fovea, meaning that the peripheral areas are not responsible for high detailed vision. Image renderers can take advantage of this fact by adapting the fidelity of the rendered objects according to the user's attention. Those elements that are more salient have a higher resource allocation from the renderers. This efficient management of resources can be very important for interactive scenarios (VR, for instance). A framework of attention is proposed by Lee *et al.* [SKS09], in which adaptive rendering is stated as an example of application of the model.

Other research areas such as object recognition, image matching, image segmentation, object tracking, active vision, human-robot interaction, and robot navigation and localisation [BAA11] can be presented. However, the purpose of this section was to give a non-exhaustive list of areas where attention models have a relevant role.

2.2 Serious Games

At its core, "serious games" are not just for entertaining purposes, but also for learning and research purposes. Indeed, from a broader perspective, the purpose of games is to entertain users and,

possibly, allow interactions between them. Little to none research impact is expected from them. Platforms such as PlayStation®, Xbox® and computers in general are used to as to support the execution of games. However, serious games go beyond games *per se*. A simple definition of serious games can be defined as "a serious game is a game in which education (in its various forms) is the primary goal, rather than entertaining" [MC06]. A specific purpose, whether that is educational or a research one, that goes beyond entertaining, is what drives them. Note, however, that serious games does not imply that they can also have an entertaining component [GA16]. The development of serious games is usually a multidisciplinary exercise, as various areas of expertise are required for a proper development [GA16]. Game engines such as Unity3D®, a cross-platform game engine developed by Unity Technologies [1], which allows the development of 3-dimensional virtual scenarios, can support the development of serious games. An example of application is given by Mykoniatis *et al.* [MAPK14].

Several domains of application for serious games can be outlined. Serious games have been used as a training tool in military contexts. In fact, the US Army is on the lead in terms of development of serious games for military purposes [LJ13]. Flight [SORJM15] and battle[2] simulators are some examples. The objective of such simulators is to provide military training and rehearsal. An interesting component of such simulators is their ability to provide multiple adjustable parameters, so as to provide high fidelity levels. As for medical domain, serious games are also used as an educational and training tool. A systematic literature review was conducted by Wang *et al.* [WDGK16] in the usage of serious games for medical purposes, where a comprehensive list of games is available. In addition, efforts have been put towards the use of serious games in educational scenarios. For example, the University of California has developed a mixed-reality virtual environment [HSD+13] where teachers can practice pedagogy and management in a virtual classroom.

The scope of serious games goes beyond entertaining. Their focus is to provide educational, research and training tools. Several platforms are available and they can be used as a foundation for the development of serious games. Several areas of application can be outlined, such as military, medical and education domains.

2.3 Data Mining Clustering

In Section 3.2.1, data mining clustering algorithms are applied to the eye-tracker data collected throughout the conducted experiments. This is done so as to obtain different clusters of observed objects in the virtual scene. Therefore, a literature review pertaining data mining clustering was conducted. A brief enumeration of the different clustering algorithms, as well as their respective description, is provided in this section. Note that not all algorithms are listed in this section; those that, according to the researcher, are seem to be more adequate for the purposes of this project are

[1] https://unity3d.com

[2] "VBS3: The future battlespace", from Bohemia Interactive Simulations - https://bisimulations.com/products/virtual-battlespace (Accessed: 20th June 2018)

enumerated. The surveys proposed by Xu *et al.* [XW05], Berkhin [Ber06], and Wong [Won15] represent the main literature for this section. Additional references are provided as necessary.

The purpose of clustering algorithms is to group data according to properties/features of the data instances. In order to compare instances, a measure of similarity must be defined, in which the features of the instances are taken into consideration. Clustering algorithm should then place together in the same cluster instances that are similar, while instances belonging to different cluster should have highly dissimilar features. Since no information other than the properties of the instances is used, clustering is an *unsupervised learning* method.

k-means [Har75, HW79] is one of the renowned clustering algorithms. Being a *partitional and prototype-based* algorithm, each instance is assigned to a single cluster (thus leading to non-overlapping clusters), and clusters are represented by a vector of properties whose value may not correspond to any instance of the dataset, respectively. Such vector is called *centroid* and its values can be seen as a profile of the elements belonging to the cluster. The set of vectors that characterise each of the `k` clusters is the output of the algorithm. However, some drawbacks can be outlined. The algorithm requires the number of clusters (`k`) to be previously defined, thus representing one of the hyper-parameters of the algorithm. In addition, k-means is capable of producing cluster of convex shape. If the data contains patterns that cannot be represented in a convex shape, an ideal clustering model may not be obtained. An initial random seeding step for the initial centroids is conducted, which, if not chosen properly, can lead to suboptimal results. Several variations have been proposed: the *k-medoids* [KR87] forces the usage of an instance of the cluster to be its profile; *x-means* [PM00] uses information metrics such as Bayesian Information Criterion and Akaike Information Criterion to serve as heuristics when choosing the ideal number of clusters; and k-means++ [AV07] proposes a new seeding technique for the initial step so as to reduce the randomness by leaning towards points located far away from the already chosen ones.

DBSCAN[3] [EKSX96] is a *partitional and density-based* clustering algorithm. Clusters are formed based on how close instances are among themselves. If they are close, then such instances must be similar and thus they must belong to the same cluster. This approach allows the algorithm to be more protected against outliers, since, in order to be classified as such, the instances must be isolated from the remaining instances. However, the lack of interpretability is a drawback of this method, since no model is outputted. The only piece of information that is obtained is what cluster the instance belongs to. Some extensions have been proposed (such as *GDBSCAN* [SEKX98] and *OPTICS* [ABKS99]).

Expectation-maximisation clustering [XW05] represents a probabilistic approach to the clustering problem. The model assumes that instances are generated by multiple probability distributions (mixture components), collectively defined by a mixture model. Different density functions from different families and parameters can be considered in the model, and each component is associated with a weight (mixture weight) representing its contribution to the final model. The mixture components' parameters are then fitted to the data by using Likelihood Maximisation

[3] Density-based spatial clustering of applications with noise

(LM) estimation, which tries to find the parameters of the distributions that maximises the probability of the instances being generated by those distributions. Since an analytic approach to such estimation is not tractable, an iterative estimation using the expectation-maximisation algorithm is performed. The final output of the model is the mixture model, in which the mixture weights and the distribution parameters are described. The concept of belonging to a particular cluster is replaced by a probabilistic membership. Using Bayes' Theorem, one can estimate the degree of the membership to a particular cluster. The expectation-maximisation algorithm can be extended. Variational inference [BKM17] is one of such extensions, which adds regularisation factors by taking into account information from prior distributions (for instance, Dirichlet processes). This extension allows a suitable number of components to be taken into account into the final mixture model.

In **Agglomerative Hierarchical clustering** [MR05], clusters are obtained in a progressive fashion, leading towards a hierarchy of clusters. Two approaches can be outlined: a *top-down / divisive*, in which all instances belong to the same clusters and a progressive division of the clusters is made; and a *bottom-up / agglomerative* approach, in which each instance has its own cluster and these are joined iteratively. A tree representing the hierarchical structure is obtained, and the connections can be cut according to a given criteria in order to form clusters. Usually, a dendrogram is used as a graphical support. Several criterias can be outlined as to how distances between clusters are defined as a function of observations, such as: *single linkage* (distance between the closest instances, one from each set), *complete linkage* (distance between the most distant instances, one from each set), *average linkage* (average distance of every pair of instances, in which each instance of a pair comes from each set), and *Ward's linkage* (takes into account the increase of the total within-cluster variance after merging).

Finally, **Affinity Propagation** [FD07] takes a graphical approach (not probabilistic, bu based on graphs) at the clustering problem, by taking as input a measure of similarity between pairs of instances. Real-value messages are exchanged between the instances. After executing various iterations, the clusters and respective profiles (which were labelled by the algorithm as "representative") will emerge. The number of clusters is thus obtained automatically by the algorithm.

It is worth noting that only a brief number of clustering algorithms was presented here, as the focus of the dissertation is in the area of visual modelling. Nevertheless, since clustering is going to be presented in the methodological approach, a introductory research was conducted.

2.4 Summary

In this chapter, the literature review conducted by the researcher was presented. Since specific terminology is present in the research domain of attention, an introduction of those concepts was presented. An attempt at defining what attention is was given; however, it was not possible to reach a solid conclusion. Taking into account the fact that the focus of the current dissertation is not one of psychology nor one of neuropsychology, a high level description provided by Li *et al.* and Sinai were considered as sufficient to be used as basis for the work of the dissertation.

The main 5 types of attention proposed by Sohlberg and Mateer (focused, sustained, selective, alternating, and divided attention) were briefly described. Also, a brief description regarding the concept of *saliency* was provided.

Before listing the computational models that try to emulate the biological processes of visual attention, a description of the biological mechanisms of human vision was required. The cellular structure of the eye's retina was presented, and the centre-surround mechanisms were referred. The bottom-up and top-down attention processes were explained, followed by an outline of the approaches concerning how bottom-up responses are modulated by top-down biases. Additionally, a psychological point of view regarding visual attention was presented. The pop-out, set-size and contextual cueing effects were briefly reported, and the type of units it is believed that attention operates were listed. The renowned Feature Integration Theory, proposed by Treisman and Gelade, was referred.

The computational models of attention were presented, based on the taxonomy suggested by Itti *et al.*. Their work considered the following classification: cognitive, information-theoretic, graphical, decision-theoretic, spectral analysis, and pattern classification models. As for models that take into account top-down processes, no explicit taxonomy was suggested; however, a grouping can be suggested based on which level (space, feature, or object) these biases operate. A brief description of related work concluded the presentation of the models. To finalise the chapter, an overview regarding serious games and data mining clustering techniques was provided.

Chapter 3

Methodological Approach

In the previous chapters, a description of the problem was given, as well as some contextual framing of the dissertation. In addition, background knowledge that seem necessary from the researcher's point of view for a minimal understanding of the developed work was given, and a literature review regarding the fundamentals of attention and computational attention models was presented. In this chapter, a description of the implementation proposed by the researcher is reported. Each type of attention component will be given a section, and each one will refer to the corresponding component.

As stated in Equation 1.1, the purpose of the dissertation is to outline a framework capable of ranking the elements of an environment according to their degree of importance in terms of visual attention. In other words, the dissertation proposes the definition of a function s in which, given an attention profile, the location of the person in the world, the head orientation towards the surrounding environment, and an entity (other than the person himself) located in the same environment, the visual saliency of the entity from the person's perspective is clearly stated.

From the conducted literature review, the attention process is divided into two interconnected components. A first one, *bottom-up*, is focused entirely on the properties of the stimulus. The second component, *top-down*, modulates the output of the bottom-up component in order to take into account cognitive phenomena. Therefore, the aforementioned function should reflect this division. Equation 3.1 breaks down the defined function into the two components.

$$s(p,o,Env,elem) \triangleq bu(p,o,Env,elem) \odot td(p,o,Env,elem) \tag{3.1}$$

, where

s	: the attention function, whose value represents the degree that the element $elem$ catches the user's attention
$p \in \mathbb{R}^3$	: the position of the person in the environment
$o \in \mathbb{R}^3$	: the person's orientation towards the environment (e.g. the heading of the field of view)
$Env \subseteq \mathbb{R}^3 \times \mathbb{R}^3$	: a function describing the placement and rotation of the elements in the virtual environment
$elem \in \mathbb{R}^3 \times \mathbb{R}^3$	: the element of interest
bu	: the bottom-up component
td	: the top-down component
$\odot$	: the operator responsible for combining the attention factors of both components

Equation 3.1 allows the reader to have a high-level representation of the developed model. The meta-model implementation will use the model proposed by Itti *et al.* [IKN98] in order to model the bottom-up component of attention. The top-down component based on mixture models will modulate the bottom-up output by considering the obtained experimental results with the called subjects. Each of the following sections will address the implementation details of each attention component. Note that, throughout the dissertation, the term "user" refers to an abstract entity that is perceiving the environment and interact with it accordingly. As for "participant" and "subject", these terms are going to be used interchangeably to refer to all people that took part of the conducted experiments.

For both attention components, a virtual environment is used. The bottom-up element makes use of it so as to capture the visible elements of the scene as the eyes of a human being would do in a real scene. As for the top-down, the purpose of the virtual scene is to capture data that reflects the actions and cognitive processes of the user while he is immersed in the scene. For this project, the virtual environment was developed in Unity [1], a WYSIWYG component-based game engine. Each element of the game is described by several components with adjustable parameters. In addition, scripts based on general programming languages such as C# can be used to add custom behaviour to such elements. This modularity-driven programming allows programmers to promote reusability and a better separation of concerns. The engine supports the development of programs framed in a cross reality (XR) paradigm. In particular, it gives support for Augmented Reality (AR) and Virtual Reality (VR) applications. For the dissertation, a focus was given to the latter, where peripherals such as HTC Vive (Figure 3.1 shows the equipment used throughout the experiments) were used. Note that the equipment was provided by LIACC [2].

[1] version 2017.3.1f1 was the used version

[2] Laboratório de Inteligência Artificial e Ciência de Computadores

(a) Wands (b) Head Mounted Display (HMD)

Figure 3.1: *HTC Vive*: the equipment used for this project, so as to allow a participant to be immersed in the virtual environment and to allow interactions with the latter

3.1 Bottom-up Visual Attention

As stated in Chapter 2, bottom-up visual attention takes into account the properties of the visual stimulus, and thus no volitional processes take place at this phase. Several models have been proposed in order to explain, in a computationally tractable way, the processes that take place when a person is observing the surrounding environment in terms of attention processes. As such, different definitions of a salient region are given by those models depending on their rationale for explaining the attention processes. One of such models is the one proposed by Itti *et al.*, which implements the architecture proposed by Koch *et al.*. It defines a single topographical saliency map from the combination of a set of conspicuity maps, one for each of the features of the visual stimulus (colour difference, intensity, and local edge orientation, among others) [Itt00, IKN98, KU87]. The saliency map encodes the conspicuity at every location in the visual field by a scalar quantity. Such map can then support the selection of visual salient objects based on the spatial distribution of saliency. Several other models have taken this model as their basis, and it has become a standard benchmark for comparison purposes [AL10]. Indeed, throughout the conducted literature review, several articles would reference Itti's model. In addition, this model is supported by biological evidences, namely by the Feature Integration Theory (FIT) [TG80], which dictates that when a stimulus is being perceived, features are perceived in a parallel fashion, while objects are identified in a serial fashion at a later phase. Given this high profile, this model was chosen for implementation in this work. It is also worth noting that a C++ implementation of this model is available at iLab (`http://ilab.usc.edu`). A transformation of the code to C# was conducted so as to adapt it to the needs of the project. In addition, some optimisations were performed, namely the use of data and task parallelism. In order to allow the reader to become more familiar with the model, this section will present a deep analysis of the model and the respective implementation.

The output of the model is a topographical map capable of encoding how much a given region is salient, taking into account properties from the input image. For that, three conspicuity

maps are combined, where each of them focus on specific dimensions of the input image, such as colour, intensity, and local orientation information. Each of these conspicuity maps represent the combination of several images from dyadic pyramids, and differences between the pyramid's levels simulate centre-surround opponency that is present in the cell's receptive field of the visual system. The reader is referred to Chapter 2 for an introductory approach of these concepts. An operationalisation of these operations will be explained in greater detail throughout this chapter.

3.1.1 From Image to Saliency Map

The model starts by decomposing the input image into 3 maps, according to several properties of the image, namely the intensity and colour differences (red-green and blue-yellow). This chromatic opponency is taken into account by the human visual system [EZW97] in the visual cortex when analysing the stimulus for regions of interest, thus being considered by the model. The intensity map, an arithmetic average of the colour components, is obtained for each pixel, which can be represented in a mathematical notation by

$$I_i = \frac{R_i + G_i + B_i}{3} \tag{3.2}$$

, where R, G and B represent the red, green and blue components of the pixel, respectively. Note that, aside from intensity information, this map is also latter used to obtain local orientation information. In Unity, an image is represented as an array of `Color` objects, and each of them carries the values of the pixel components. The colour differences maps are then obtained according to the formulas stated in the definition of the model. A normalisation of the colour components is performed, so as to ensure that the difference is not affected by the intensity of the colour. The normalisation factor is given by

$$n_i = \frac{255}{R_i + G_i + B_i} \tag{3.3}$$

, being i the index of the pixel.If the sum of the colour components results in a null value (a black pixel), it is assumed that the normalisation factor is 0, which results in a null difference. The normalised red, green, and blue colours can be obtained by Equations 3.4, 3.5, and 3.6, respectively.

$$R_i^n = R_i \times n_i \tag{3.4}$$

$$G_i^n = G_i \times n_i \tag{3.5}$$

$$B_i^n = B_i \times n_i \tag{3.6}$$

After that, four colour channels are obtained by the following equations.

$$R_i^{'} = R_i^n - 0.5 \times (G_i^n + B_i^n) \tag{3.7}$$

$$G_i^{'} = G_i^n - 0.5 \times (R_i^n + B_i^n) \tag{3.8}$$

$$B_i^{'} = B_i^n - 0.5 \times (R_i^n + G_i^n) \tag{3.9}$$

$$Y_i^{'} = R_i^n + G_i^n - 2 \times [|(R_i^n - G_i^n)| + B_i^n] \tag{3.10}$$

Note that negative values of $R_i^{'}$, $G_i^{'}$, $B_i^{'}$, and $Y_i^{'}$ are clamped to zero. Finally, Equations 3.11 and 3.12 give the red-green and blue-yellow colour difference channels, respectively.

$$RG_i = R_i^{'} - G_i^{'} \tag{3.11}$$

$$BY_i = B_i^{'} - Y_i^{'} \tag{3.12}$$

If the luminance of the pixel if less than a given threshold (for this project, it was considered a threshold of 25.5, as proposed by the implementation available at iLab), it is assumed that the pixel has an insufficient response (it is too dark, and therefore no hue difference is perceived), and thus the colour difference is considered to be zero. As stated by Itti, the proposed expressions to obtain and normalise the colour and intensity channels are crude and represent and approximation of the true colours detected by the photosensitive cells of the eye. Such simplification is explained in terms of simplicity and efficiency.

From the obtained maps, dyadic pyramids are obtained. In the field of computer vision and image processing fields, a pyramid represents a multi-scale representation of the input signal (for example, an image), in which the input signal is represented at multiple scales. A level of the pyramid is the result of a convolution and a decimation operations on the previous level of the pyramid. Usually, a smoothing operation is performed before the decimation step in order to avoid aliasing in the resulting image, and this can be achieved by executing a convolution with a low-pass filter such as a Gaussian one. If, after the decimation step, the resulting image is half of the dimensions when compared with the original image, then the pyramid is said to be dyadic. This type of structures are used by the attention model to support the centre-surround difference operations and perform within-feature saliency competition.

From the intensity, red-green, and blue-yellow differences, dyadic Gaussian pyramids are built for each of these features. A Gaussian kernel of size 5×5 is used in the convolution operation. 9 levels are obtained (the original image is at level 0), and each level represents a scaled version of the original image. This means that, at level 8 (the lowest one), the ratio between the original and the scaled down version is $1:256$. Let $I(\sigma)$, $RG(\sigma)$, and $BY(\sigma)$ represent the pyramids, respectively. The pyramids are represented by a parameter σ, which represents a specific level (scale) of the image. Note that $\sigma \in \mathbb{N} \wedge \sigma \in [0,8]$.

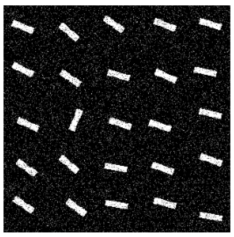


Figure 3.2: Example of a search task image where elements only differ in terms of edge orientation (Source: [IKN98])

The bottom-up attention model also takes into consideration information regarding local orientation for the detection of salient regions, that is, information regarding the orientation of objects. This type of information is crucial to detect pertinent regions when no differences in terms of colour and intensity properties can be found. When considering Figure 3.2, the elements only differ in terms of orientation, where the element at coordinates $(1,2)$[3] "pops out". The model tries to incorporate this information into the saliency map by using oriented Gabor pyramids. Before applying Gabor filters, edge information must be obtained first. For that, one can use the Laplacian of Gaussian (LoG) function as the kernel for the convolution operation. However, a Difference of Gaussians (DoG) is used as an approximation of the former convolution, which corresponds to the difference between an image and its blurred version. In the frequency domain, the gaussian filter can be seen as a lowpass filter (since it attenuates high frequency signals from the image), and the difference operation as a bandpass filter. To obtain the blurred version, a 9×9 Gaussian kernel was used. A dyadic laplacian pyramid is the result of performing DoG over several iterations on the same image, in which each iteration results on an image half of the size of the previous one.

This laplacian pyramid is then subject to a convolution operation with Gabor filters, which are the result of the scalar product between a complex sinusoid, known as the *carrier*, and a 2-D Gaussian shaped function[4], called the *envelope*. The carrier takes a spatial frequency parameter, which dictates how the kernel reacts to the input signal in a given direction. In the particular case of the attention model, the kernel reacts differently depending on the edge orientation of the objects. Note that edges with an orientation orthogonal to the filter's orientation get the highest output response. Four orientations (0°, 45°, 90°, 135°) of the filter are considered, which results in four oriented pyramids. The consideration of more orientations did not lead to significant performance changes [Itt00]. Let $O_{0^\circ}(\sigma)$, $O_{45^\circ}(\sigma)$, $O_{90^\circ}(\sigma)$, and $O_{135^\circ}(\sigma)$ represent the oriented pyramids at those orientations, respectively. As previously, σ is the scale.

In order to compute the feature maps, centre-surround operations are taken. The objective of such operations is to emulate the receptive field of the retinal ganglion cells, which are located near the inner surface of the retina of the eye. These cells are excited when a difference between the centre and the surround cells is found, thus they encode for contrast in the received visual stimulus. In terms of implementation, this can be emulated as the difference between fine and

[3] assuming the top-left element as the origin of the coordinate system

[4] in this model, an isotropic Gaussian-shaped function was used

coarse scales, in which the former is an image containing a substantial amount of details when compared with the image representing the coarse scale, where only a gist of the input image is present. Note that this difference is performed for each considered feature (colour, intensity, and orientation). With these scales, an across-scale difference between these maps can be performed. This is achieved by performing interpolation to the finer scale and pixelwise subtraction between the selected scales, an operation represented by $\ominus$. The model proposed by Itti *et al.* performs these subtractions using different scales for fine and coarse, thus achieving a multiscale feature extraction. This is useful since salient areas can be of multiple sizes.

Remember that σ represents the scale and $\sigma \in \mathbb{N} \wedge \sigma \in [0,8]$. Let $c \in \{2,3,4\}$ be the pyramid scale that is considered as the centre of the receptive field, and $s = c + \delta$ the surround scale, with $\delta \in \{3,4\}$. Seven feature map types are considered: red-green difference, blue-yellow difference, intensity and one for each orientation. The feature maps are obtained by computing the centre-surround differences from the obtained pyramids, as shown by Equations 3.13, 3.14, 3.15, and 3.16. In total, 42 feature maps are created, in which 6 are for intensity, 12 for colour, and 24 for orientation.

$$I_f(c,s) = |I(c) \ominus I(s)| \tag{3.13}$$

$$RG_f(c,s) = |RG(c) \ominus RG(s)| \tag{3.14}$$

$$BY_f(c,s) = |BY(c) \ominus BY(s)| \tag{3.15}$$

$$O_f(c,s,\theta) = |O_\theta(c) \ominus O_\theta(s)| \tag{3.16}$$

The combination of the respective feature maps is the next step. The result is 3 conspicuity maps, one for each feature type. As the final step of the model, these maps are combined in order to have a single saliency map. However, those conspicuity maps encode saliency for different non-comparable visual modalities, where different extraction mechanisms are used and unrelated saliency ranges are considered. Thus, the values are not comparable among themselves, meaning that a simple summation may not be sufficient nor accurate. Itti and Koch propose several feature combination strategies [IK01b], namely:

- **Naïve Summation** - all maps are normalised to a predefined range (e.g., $[0,1]$) and summed up.
- **Supervised Learning Linear Combinations** - each map has a weighting factor, which represents its contribution to the final map. Such weight is obtained from applying supervised learning techniques into manually outlined target regions image.
- **Contents-based Global Non-Linear Amplification** - promotes the maps that have few "peaks of activity" (only a few regions of the map have high saliency values). The maps that

are homogeneous in terms of response are suppressed, since no relevant information can be obtained from these.

- **Iterative Localised Interactions** - iteratively applies a two-dimensional difference of Gaussians kernel convolution, allowing robust results even in the presence of noise. The objective is to simulate local competition between neighbouring salient regions.

In terms of performance, the last two strategies led to better results, when compared to the first one [IK01b]. The **Contents-based Global Non-Linear Amplification** was the adopted strategy, due to its simplicity in terms of computation and performance in terms of results, as reported by Itti *et al.* [IK01b]. Let $\mathcal{N}$ represent the application of this strategy. The operationalisation of $\mathcal{N}$ can be described as follows: firstly, all maps are normalised into a given range; then, for each map, the global maximum M is found, as well as the average of all the other local maxima $\overline{m}$; finally, the map is globally multiplied by $(M-\overline{m})^2$, which promotes maps in which the difference between the peak and the rest of the map is greater.

The next step is focused on obtaining the conspicuity maps for intensity I_c, colour C_c and orientation O_c. These maps are obtained by combining the previously computed feature maps. This combination is done by scaling each feature map to scale 4 (the scale considered by the model for the final saliency map, which, in dyadic pyramids, this represents $\frac{1}{2^4} = \frac{1}{16}$ of the size of the input image) and then performing a pixelwise addition. Let $\bigoplus$ represent this combination.

$$I_c = \bigoplus_{c=2}^{4} \bigoplus_{s=c+3}^{c+4} \mathcal{N}(I_f(c,s)) \tag{3.17}$$

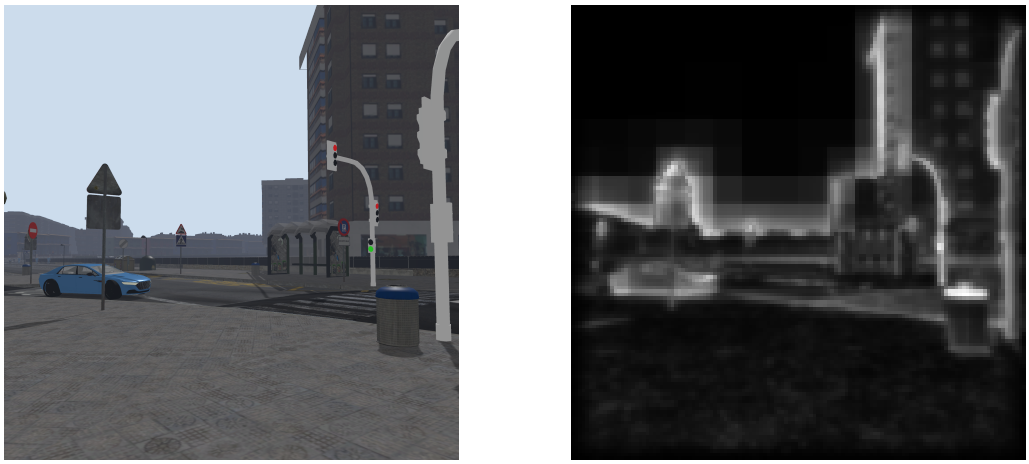
$$C_c = \bigoplus_{c=2}^{4} \bigoplus_{s=c+3}^{c+4} [\mathcal{N}(RG_f(c,s)) + \mathcal{N}(BY_f(c,s))] \tag{3.18}$$

$$O_c = \sum_{\theta \in \{0°,45°,90°,135°\}} \mathcal{N}(\bigoplus_{c=2}^{4} \bigoplus_{s=c+3}^{c+4} \mathcal{N}(O_f(c,s,\theta))) \tag{3.19}$$

The final step is to obtain the saliency map. A final summation of the normalised conspicuity maps is performed, as shown by Equation 3.20. Figure 3.3 shows an example of the application of the model. Notice that, in Figure 3.3b, salient regions are more highlighted. Also note that, so as to obtain an image of the saliency map, a value rescaling is performed using a range of $[0,255]$, for a 8-bit colour channel encoding, in which the same value is used for every channel.

$$S = \frac{\mathcal{N}(I_c) + \mathcal{N}(C_c) + \mathcal{N}(O_c)}{3} \tag{3.20}$$

Having the saliency map, it is required to perform the association between the saliency of a given region and the object located at that region, so as to know how salient is an object to the user. In order to do this, it is required to establish a relationship between image-space (the saliency map) and object-space (the virtual environment). In this project, this was achieved by assigning a unique identification for each type of object (including its state, if relevant) and performing a false colour



(a) The original image. (b) The obtained saliency map.

Figure 3.3: An example of the saliency map obtained by the Itti's model.

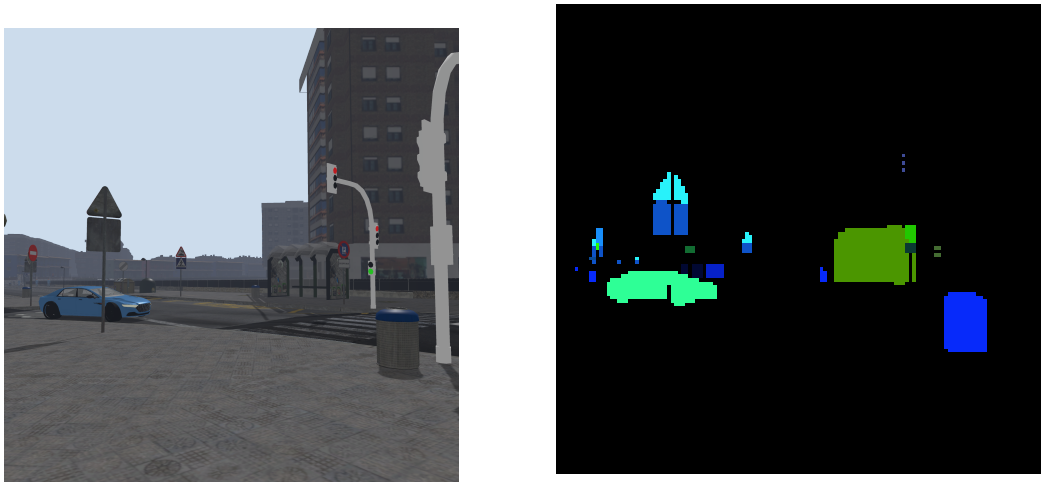
off-screen rendering of an object identification map (or object ID map, for short) of the scene. Each unique ID is assigned a unique false colour, thus establishing an injective mapping. This means that an inverse mapping can be obtained, which is important as the model needs to obtain the original ID in order to know what object it is. Having the input image, the object ID map is obtained with the same size and resolution as the saliency map. Figure 3.4 shows an example of such map.

With the saliency map and the object identification map, both of the same width and height, one can assign a saliency value for the object. Both maps are searched in a lockstep fashion, and, for each identified object, the saliency value is updated. With the assumption that the saliency is the same throughout the object's surface, the incremental average process is used, as shown by Algorithm 1. The average object saliency value for each observable object represents the output of the algorithm.

A relation between the saliency value and the object is obtained with the execution of the presented model. This allows to know how salient a given object of the scene is from the subject's perspective by taking into consideration information related to the properties of the visual stimulus. At this phase, the *bu* function in Equation 3.1 is defined. However, higher level processes also take place, biasing attention in a manner that bottom-up factors alone cannot explain. The next section will focus on this component of attention.

3.1.2 Implementation Details

Up until now, a detailed description of the bottom-up attention model proposed by Itti *et al.* was presented. In addition, an overall explanation of the object ID map is provided. The purpose of such descriptions was to provide the reader some enlightenment as to how an association between saliency and scene objects was obtained. However, some implementation challenges were dealt



(a) The original image.

(b) The object identification map. Note that each object type is assigned a different colour.

Figure 3.4: An example of the object identification map. Such map is used to detect what objects are present in a given image.

Algorithm 1: Algorithm for extracting the saliency values for each object of the scene.

```
1  function PROCESS-SALIENCY-MAP (SaliencyMap, ObjectIDMap);
   Input  : SaliencyMap - the saliency values for each position of the viewport
            ObjectIDMap - the object ID for each position of the viewport
   Output: ObjectSaliency(objectID) - a function that gives a correspondence between an
            object and its average saliency value
2  Assert SaliencyMap.dimensions = ObjectIDMap.dimensions
3  foreach pixel p in ObjectIDMap do
4      objectID ←GET-OBJECT-ID(p)
5      saliency ← SaliencyMap(p.coord)
6      if objectID! = ∅ then
7          csv ← ObjectSaliency(objectID).value
8          nt ← ObjectSaliency(objectID).numberOfTerms + 1
9          nsv ← csv + (saliency − csv)/nt
10         ObjectSaliency(objectID).value ← nsv
11         ObjectSaliency(objectID).numberOfTerms ← nt
12     end
13 end
14 return ObjectSaliency
```

with throughout the development of the dissertation, and they are briefly described throughout this section.

As a side note, regarding the code implementation of the aforementioned operations, the facilities offered by an object-oriented programming language such as C# were used. The aforementioned operations are offered by `TextureImage`, a class which takes a `Color` array representing the image to be analysed, and the width and height of the image. In order to facilitate the manipulation of images throughout the implementation, the class `MyImage` was created, whose instances contain an array of floats (whose values focus on a given property of the image, such as the pixel intensity) of the same size as the number of pixels of the image. In addition, such instances have the width and height values of the corresponding images they represent so as to perform several validation operations (for instance, when two images are added, both images must have the same dimensions). Various operations such as pixelwise addition and subtraction of images are available, including the multiplication operation of all pixel values by a scalar. The latter operation will reveal to be useful when normalisation operations are applied. The two colour difference maps, as well as the intensity maps, are abstracted by `MyImage` instances.

The current section can be divided into 3 subsections. The first subsection focuses on the mechanisms used to obtain the object identification mapping. Another subsection emphasises the impact that separate kernels have in terms of computational performance for the model. A last subsection proposes an overall description regarding the use of task and data parallelism as means for performance gains.

3.1.2.1 Object Identification Mapping

As previously stated, an object map is used so as to deduce what object is located at a given pixel location. From a computer graphics perspective, a relationship between image-space (the saliency map) and object-space (the virtual environment) is obtained. Several strategies can be used to solve this mapping problem, with various degrees of performance efficiency. In this project, a strategy that, in the opinion of the researcher, entails minimal computation costs is used. Two operation stages involving the object map can be outlined. A first stage concerns the rendering of such map from the identification assigned to the object to a colour; then, another stage can depict the inverted process, from the colour to the object ID, and, inherently, to the object. Both stages are described in terms of implementation in the current subsection.

Unity offers an A.P.I. to programmers allowing them to obtain the object located at a specific pixel of the image. The game engine casts a ray from a given origin (usually, the position of the observer) and passing through a given pixel of the viewport. If such ray collides with an object of the virtual environment, then a mapping between the pixel and the object is obtained. However, for a given pixel, the raycast operation bears a considerable computation cost (for instance, geometry calculus, and collision detection operations), and performing the operation for each pixel can entail considerable performance costs. A more efficient approach can be used based on shaders. A shader represents a computer program that can be executed by the graphics card, and, since these components are specifically designed to deal with graphic computations (and thus being proficient

at those), one can take advantage of this aspect in order to solve the object identification problem. *High Level Shading Language* (HLSL) is the adopted language by Unity to develop shaders, and, therefore, it was used to develop a shader to solve the aforementioned problem.

Each object is linked to a unique ID (an integer), and such ID can also incorporate semantic information. For instance, for the traffic lights, the ID can the object and its current state. If one considers id_{tl} to be the ID of the traffic light, id_{tl} can also indicate that the red light is turned on. Moreover, $id_{tl} + c,\ c \neq 0$ can represent the same object, but in a new state (with the green light activated, for instance).

However, as aforementioned, shaders are specifically designed to perform graphic processing, and simply feeding an integer to them does not solve the problem at hands. A transformation of such ID is thus required. A RGB colour representation is available for shaders, with 8 bits per channel, meaning that a range of 256 different values is available for each channel. So as to compensate for possible errors, a base-six numeral system is used. Equation 3.21 dictates how a pixel colour is obtained.

$$(r,g,b) = (\left[\left\lfloor \frac{id}{36} \right\rfloor \ \%\ 6\right] * 50, \left[\left\lfloor \frac{id}{6} \right\rfloor \%\ 6\right] * 50, [id\ \%\ 6] * 50) \tag{3.21}$$

, where

$r,g,b \in [0,255] \cap \mathbb{N}$: the red, green, and blue components of the pixel, respectively, encoded for a 8-bit channel

$id \in \mathbb{N}$: the object's ID

Having a false colour rendering of the observed scene, the next step concerns with obtaining the original object ID from the colour components, presented by Equation 3.22.

$$id = \frac{r}{50} * 36 + \frac{g}{50} * 6 + \frac{b}{50} \tag{3.22}$$

, where

$id \in \mathbb{N}$: the object's ID

$r,g,b \in [0,255] \cap \mathbb{N}$: the red, green, and blue components of the pixel, respectively, encoded for a 8-bit channel

With the definition of the aforementioned operations, one can deduce what object is located at a specific pixel coordinate. This, in turn, allows one to detect how salient an object is on the scene, by combining the information provided by the saliency map and the object ID map.

3.1.2.2 Separate Kernels

The application of kernels (such as a Gaussian one) in images can be done by applying a convolution operation, so as to extract the input signal's response to the kernel. In image processing, the

convolution can be given by Equation 3.23, where $f(x,y)$ represents the pixel's value at coordinates (x,y), h is the filter (or kernel) function, and $g(x,y)$ gives the image's response to the kernel at coordinates (x,y) [Sze10].

$$g(x,y) = \sum_i \sum_j f(x-i, y-j) \times h(i,j) \tag{3.23}$$

Algorithm 2 shows a high level description of how the convolution operation can be implemented. Assuming an image of size $M \times N$ and a kernel of size $U \times V$, for each $M \times N$ pixels, $U \times V$ additions and $U \times V$ multiplications are performed. Using Big-O notation to represent the temporal complexity, the convolution operation has a complexity of $O(MN(2UV)) = O(MNUV)$, which represents a considerable impact in terms of performance. In other words, for each pixel of the image, all filter coefficients are visited, which translates in $U \times V$ add-multiply operations for each pixel.

Algorithm 2: Pseudo code for the convolution operation

```
1  function CONVOLUTION (image, kernel);
   Input  : image - the input image
            kernel - the filter to be used in the convolution operation
   Output: response - the image's response to the kernel
2  foreach image row in image do
3      foreach pixel in image row do
4          accum ← 0
5          foreach kernel row in kernel do
6              foreach coefficient in kernel row do
7                  pixel value ←
                     GET-PIXEL-VALUE(image, pixel.coords, coefficient.coords)
8                  accum ← accum + coefficient × pixel value
9              end
10         end
11         response(pixel.coords) = accum
12     end
13 end
14 return response
```

However, some filters can be written as a product of two or more simpler filters, which brings some efficiency to the convolution operation. Such filters are called *separate filters*. If the filter is separate, the convolution of a 2D-image with such filter can be obtained by performing a one-dimensional horizontal convolution, followed by a one-dimensional vertical convolution, thus being called a *separate convolution*. The temporal complexity of a separate convolution operation is $O(MN(U+V))$, which represents a reduction when compared to the previous version. In order to check whether a given kernel can be separable, one may perform an empirical inspection or by analysing the analytic expression of the kernel. However, if one considers the 2-D kernel as a 2-D matrix, a *singular value decomposition* is a simpler verification method. Some kernels, such as

the Gaussian, are known for their separability property, and such property is taken advantage of in computer vision research. In particular, the considered bottom-up attention model uses separate convolution.

3.1.2.3 Task and Data Parallelism

A first approach for the implementation of the adopted bottom-up attention model followed a sequential fashion, meaning that each intermediate result of the model is produced one after another. Since the operations performed by the model are computationally demanding, the resulting saliency computation would require a considerable amount of time. Unity executes user interface-related tasks in a single thread, meaning that UI updates and script execution are run in the same thread. If an operation requires more CPU time and it is executed in a blocking fashion (that is, sequential-based programming), the user interface will stall. Therefore, it is required to execute the attention model in a separate, worker thread so as to minimise the impact on the main thread. In addition, in this worker thread, the execution of the model must take the least amount of time so as to keep up with the player's interaction and update the object's saliency values accordingly. The usage of parallel programming was the adopted solution for this problem.

If one takes a closer look into the data dependencies between the multiple maps that are defined by the bottom-up attention model, some operations can be performed independently from others, since no dependencies are found between then. It is, then, possible to define *units of work* that can be executed concurrently, a form of parallelisation called *task parallelism*. By taking into account the parallel nature of the model, one can take advantage of the resources provided by the system, namely multi-core CPUs, thus yielding a truly parallel execution. This allows a model execution with a greater performance when compared to the sequential approach.

C# language offers high-level abstractions allowing a user to build task-based parallel programs. A `Task` instance represents an asynchronous operation, and the class of those instances offers some static methods to simplify the implementation of parallel programs. Examples of such methods are `Task.Factory.StartNew`, a factory method that allows the creation of new tasks, and `Task.Factory.ContinueWhenAll`, another factory method allowing to add synchronisation points throughout the code. The latter method is of the utmost importance when one wants to wait for more than one task to be completed.

The implementation of task parallelism requires a global understanding of the global architecture of the system before coding it. It is necessary to identify the units of work in the code that cannot be performed in parallel, so as to maximise the benefits taken from the parallelisation features. Outlining the data dependencies between tasks can reveal to be very useful in identifying such units. An outline done by the researcher is presented in Figure 3.5. Note that each colour represents a unit of work that must be run in a sequential fashion, thus representing a task to be created.

Data parallelism was also considered. With this type of parallelism, data is distributed across different working entities (e.g. threads) and each of these entities execute the same instructions over the assigned data portion. Therefore, this allows for a load distribution among the available

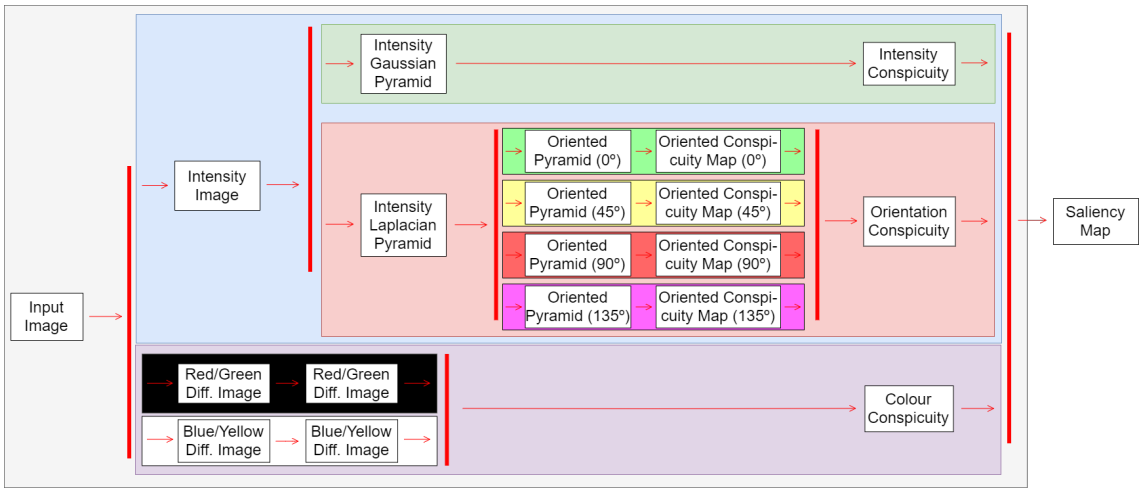


Figure 3.5: Data dependency graph for the bottom-up attention model

working entities. `for` loops represent a good place where data parallelism can be applied. In the model, the cycles were parallelised using the static method `Parallel.For` available in the `.NET` framework.

In conclusion, task and data parallelism were the facilities used to parallelise the execution of the bottom-up model. This was done so as to speedup its execution, allowing it to keep up with the user's interaction with the system.

3.1.3 Limitations

Up until now, an overall description of the bottom-up component of the attention model was provided. However, it is worth pointing out some of the limitations and assumptions this solution presents so as to allow the reader to have an in-depth insight of the proposed solution. These limitations are focused on two areas. The first is related with the time performance of the model, which can hinder the application of the model in real-time solutions; the second area is related to the processes responsible for linking the saliency to the respective objects.

In order to detect how salient an object is from the user's point of view, incremental average is used. As the name suggests, the average of the saliency value is iteratively computed. However, obtaining the average entails the assumption that the saliency value is the same throughout the object's surface. One of the problems that may arise with this assumption is related to the existence of different saliency values for different sections of the object, which, in turn, results in an erroneous value for the saliency value linked to that object. To overcome this limitation, one may divide the object into sub-objects, where each of these are assigned a different ID. With this division, a fine granularity is used, which allows to obtain a more precise value for the saliency of the object. For the sake of simplicity, the aforementioned assumption was adopted throughout the project.

As for the object identification map, as stated previously, after obtaining the saliency map, a mapping between the saliency value and the corresponding object is created. In other words, for each pixel of the saliency map, a query is performed so as to obtain the object located at that specific pixel. In order to reduce the overhead of performing a raycast operation for each pixel, a shader is used so as to render the scene using false colours, resulting in a new image. Each colour will then correspond to a given object. This strategy is very efficient, however some problems related with the rendering process may arise. More specifically, the researcher suspects that some objects may not be present in the object ID map due to resolution concerns. This can be observed in Figure 3.4b, where the bottom traffic light for the cars is not present. The size of the lights is not significant enough to be presented in the final object ID map. Of course, this phenomenon leads to errors when obtaining the final saliency value for the object. The use of other techniques or approaches may help mitigate the problem.

Additionally, the limited range of values each colour channel supports hinders its application for a considerable large number of objects to be identified in the scene. Considering a practical example, for a RGB colour scheme with 8-bit channel depth, at most $(2^8)^3 = 16777216$ unique IDs can be considered. However, in order to compensate for possible errors (floating point precision, or sampling issues such as the one presented above), not all values are used. The proposed implementation uses a base 6 numeral system, in which a colour component is capable of encoding 6 distinct IDs, allowing an approximate 50 pixel value distance between consecutive IDs. $6^3 = 216$ unique IDs are available, which represents a considerable decrease when comparing to the first approach. A compromise between accuracy and available IDs is thus at stake. 216 available IDs is sufficient for the purposes of the dissertation. If one requires a broader range of available IDs, different approaches such as a greater base value can be taken into consideration.

The Itti's model output is a saliency map which dictates the areas of the visual stimulus that are more prone to be visited by the attention mechanisms. In order to obtain such map, the input image is subject to multiple operations such as convolutions and decimations, whose temporal complexity is dependent on the pixel size of that input image. More specifically, the greater the size of the image is, the longer the model takes to fully process the image. The execution times of the model were taken while the simulation was running, and an average of 6 seconds was obtained. Since the objective of the model is to be run in real time so as to keep up with the constant changes of the visual stimulus, these execution times are not acceptable. A size reduction of the input image can be done, however some problems arise when trying to identify the objects. Note that the output map's dimensions is 16 times[5] smaller when compared to the input image. A reduction in size of the input image will lead to even smaller output image, which in turn makes the aforementioned problem in the previous paragraphs more prominent. A GPU implementation of the model can mitigate this problem.

Moreover, according to the Feature Integration Theory, other properties of the visual stimulus

[5] As stated in the description of the model, the dimensions of the output image are the same as those located at scale 4 on the image pyramids. Since each level is 2 times smaller when compared to the previous level, an image at scale 4 has its dimensions reduced by a factor of $2^4 = 16$.

such as the presence of motion and flicker are important to the human visual system when searching for salient regions. The first iteration of the Itti's model [IKN98] only took into consideration the colour, intensity and orientation features. A latter iteration (for example, [IDP04])[6] includes motion and flicker features. For video and real-time processing, this addition leads to better results. However, these features take into account the differences between frames of the video (or the video feed, in case of real time applications). Since an average of 6 seconds is required for the model to return in the virtual environment, the scene would be analysed at a rate of 6 seconds, meaning that any change in terms of motion or flicker occurring between that time interval would not be detected. In fact, the obtained results could be worsen by the inclusion of these features, since a false difference was being considered in the final output of the model. As such, the implementation this dissertation is based upon does not consider motion or flicker. In order to include these features into the model, a reduction in terms of computing time must be firstly achieved.

It is important to notice that the work being presented in this dissertation represents a proof-of-concept. The aforementioned items can therefore be seen as future work to be done in future iterations.

3.2 Top-down Visual Attention

The previous section focused on bottom-up attention, which deals with the properties of the visual stimulus in order to detect salient regions. Nevertheless, volitional processes also take part in the attention process, and such processes are framed in the top-down component, which is represented by the function td in Equation 3.1. Therefore, the purpose of this section is to give to the reader a notion regarding the implementation of the top-down component.

The top-down component of attention is greatly influenced by cognitive processes, which, in turn, is responsible for modulating (or biasing) the output response of the bottom-up component. And, as stated in Chapter 2, top-down mechanisms are based on higher level cognitive processes responsible for encoding information such as the task the person is executing, and its prior knowledge and experience. And such information changes the attention patterns when the person is perceiving its surrounding environment. It is then necessary to analyse how users perceive the scene. The use of HMD to immerse users into the virtual environment as well as the use of eye-trackers represents a key aspect for this analysis, since it allows to examine what objects caught the attention of the user. More specifically, such infrastructure will be used while users are performing a given task (in this case, crossing streets). The methodological approach that was used to obtain data from participants is described in greater detail in Chapter 4. Moreover, an analysis of the collected data is conducted in the same chapter. The current chapter will focus on the procedures that were used in order to process the collected data and model the top-down component of attention.

[6] The implementation located at `http://ilab.usc.edu` already considers these features

3.2.1 Data Modelling

For the definition of the top-down component of the attention model, data collected from real subjects while immersed in the virtual environment is used. Since the virtual environment is used to collect data which incorporates information regarding the behaviour that users have while they are perceiving an environment, a high level of immersion is needed. The model must reflect reality as close as possible, and the virtual environment must be used just as a means to obtain data with minimal obtrusion. From that data, several approaches were considered throughout the development of the dissertation in order to represent the top-down attention component. A brief description of the considered approaches is presented in this section.

A first approach focused on the definition of a grid of equal size cells, in which each cell represents a portion of the virtual environment and is responsible for analysing the objects of the virtual environment that are observed while the subject is located at the cell. The total observation time, and the number of observations for each object are the metrics considered by each cell. The sequence of observed objects is also analysed. With the total observation time and the number of observations, an average of observation time can be deduced. The hypothesis considered by the researcher is that longer observation times in the cell would translate into a greater importance in terms of attention. This takes into consideration that relevant object to the task at hands are observed more frequently when compared to less relevant objects [TWK+10]. In the final saliency map, the object saliency is increased according to that degree of importance. The definition of the size each cell must have and the rigid definition of the frontiers of the cells represent some disadvantages of this modelling approach. In a real scenario, objects' relevance is not firmly defined by "cells", but instead it is a "fluid" process. When focusing on the frontier problem, an attenuation coefficient can be proposed when the entity goes from one cell to another.

Instead of defining a grid of cells of equal sizes and with iso-oriented edges, the definition of "zones" is an alternative. Each "zone" can be of different sizes and can cover different regions of the scenario, and each "zone" can be defined according to the task to be performed at each location. For instance, a simple crosswalk and a crosswalk with an island in the middle can represent two distinct zones, and each of these can have different sizes and have different orientations. Nonetheless, the rigid frontiers of the "zones" can still represent a problem. Even if the entity is located outside the "zone" for a small distance, the model considers the objects have no relevance to the task at hands, which does not reflects the reality. Moreover, the "zones" must be manually defined, which implies that the location where the tasks should be performed must be defined *beforehand* by a third-party. The perception of the surrounding environment may already be biased by top-down processes way before entering the designated "zone".

Instead of focusing on the observation times and the number of observations that are considered for each object, an analysis can be performed by taking the position of the subject when the object was observed. In other words, when the object is perceived, the location of the subject where such observation is carried out is recorded. And such recording is made at a constant rate while the subject is immersed in the virtual environment, meaning that longer observation times

Table 3.1: A summary of the approaches considered for modelling the top-down component of attention

Approach	**Definition of region**	**Definition of object importance**
Grid	The scenario is divided into a iso-oriented grid of cells	Greater average observation times translate into greater importance
Manual region definition	The scenario is divided into regions manually defined	Greater average observation times translate into greater importance
Mixture Model	Each region of the scenario is given by each component of the mixture model	Greater probabilities of a given point belonging to a component related to the object translate into greater importance

translate into a greater amount of records. This is important so as to reflect the fact that objects given a greater importance are observer for longer periods of time [TWK+10]. Having the observation points from multiple subjects, a clustering algorithm is applied to the data. The objective is to detect regions in the scenario where certain objects are given more attention than others. Several clustering algorithms can be applied to the collected data. For instance, the k-means algorithm is a renowned algorithm, in which k clusters are defined. However, the previously stated frontier problem still stands. The ideal approach is to obtain a model in which a fuzzy membership is adopted by the model. A mixture model can be an option, as the membership criterion is defined by a *probability* that an instance is given by a density probability distribution. A Gaussian Mixture Model is an example of a mixture model, and it was the adopted approach for this project. A Gaussian Mixture Model is used to define such regions (each component of the mixture model represents a region of the scenario), and the Expectation-Maximisation algorithm is applied so as to obtain the parameters of the model that best describe the data. Since the parameters of the model are automatically found, the definition of the regions is automatically performed. Moreover, when the degree of membership of an instance is queried, a probability is given. Therefore, the strict frontiers that were given by the previous approach are replaced by fuzzy ones. The probability value is used as a measure of relevance. The mixture model of the object is used to measure how important the object is when the entity is located at a specific location. Greater probabilities lead to greater degrees of importance. Considering the benefits of this approach, it was the adopted one in this dissertation.

A description of the considered approaches for modelling the data collected from the experiments was presented. A definition of equally sized cells was the first proposed approach, but, due to its strict nature in terms of location definition, a manual region description represented the next step. Nevertheless, both previous approaches suffer from the strict border definition. A mixture model definition was the adopted solution for the top-down component of attention. A summary of the presented description is made available in Table 3.1.

3.2.2 Integration between Bottom-up and Top-down Attention

Having both components of the attention model defined, an integration of those components is required. The bottom-up component is based on the model proposed by Itti *et al.*, which takes an image as input and it outputs the saliency values for the observed objects. The top-down component, based on the position and orientation of the user and the data collected from real subjects, modulates the response outputted by the bottom-up component, giving priority for some objects that are observed at that location. Therefore, the next logical step is to formulate a procedure responsible for integrating the information from both components. The current section will focus on such integration.

Equation 3.24 represents the adopted approach to the integration of the components of attention. Let $sal_{bu} \in [0,255]$ represent the saliency value given by the bottom-up component of the attention model, and $sal_{td} \in [0,1]$ the value provided by the top-down component, representing the importance of the object at the current location of the entity. Note that sal is parameterised by the user's position p and orientation o in the world; obj represents the object one wishes to assess the corresponding saliency value. A final normalisation to the value sal is performed so as to frame it in the interval $[0,255]$. For the purposes of ordering the objects of the scene according to their saliency value, such normalisation is not required, as it does not change the ranking of the objects. However, for demonstration purposes, such normalisation in the range $[0,255]$ is conducted in order to create a bitmap image of the resulting saliency map.

$$sal(p,o,obj) \triangleq sal_{bu}(p,o,obj) \times [1 + sal_{td}(p,o,obj)] \tag{3.24}$$

This approach boosts the saliency value according to the importance given by the top-down component, while, at the same time, taking into account how much the object "pops out" from the scene. As suggested by the literature (see Section 2.1.2), a multiplicative gain modulation is applied to the bottom-up response. The multiplication constant is based on the probability returned by the mixture model pertaining the object under analysis, and the gain is not based on some magic constant, but instead, it takes into account how much the object is already salient with no top-down factors.

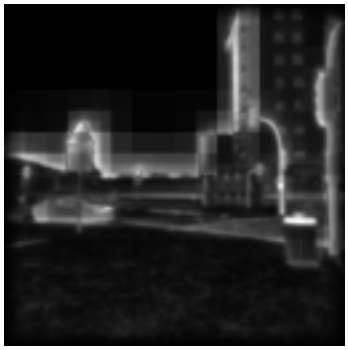
Figure 3.6 represents an example before and after the bottom-up component is modulated by top-down factors. A highlight is given to the pedestrian lights and the car on the left. Notice how the the car become more salient when compared to the bottom-up response alone.

Note that the above equation represents only a proposal for merging the contributions from both components. Nevertheless, a more complex approach may be more adequate to conduct the modulation procedure of the bottom-up response. As future work, the definition of other combination strategies based on different rationales can be conducted.

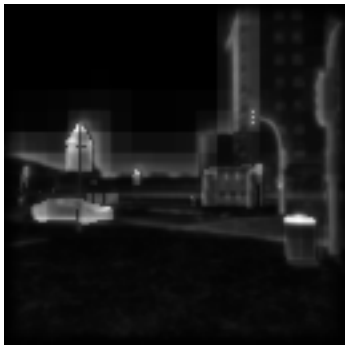
3.2.3 Limitations

As of now, a description of the considered approaches to model the top-down component of attention was presented. Additionally, a proposal for integrating the contributions from both (bottom-up

(a) The original image.



(b) The bottom-up saliency map.



(c) The final saliency map.

Figure 3.6: An example of application of the proposed model of attention.

and top-down) components was reported. Nevertheless, the proposed implementation is a first attempt at modelling attention, and, as such, some limitations can be presented. The current section is focused on outlining some aspects of the implementation concerning the top-down component that, according to the researcher, can be seen as limitations.

As stated before, the equation that is used to combine the contributions from both attention components is an hypothesis proposed by the researcher. In the literature, the researcher could not find a specific expression capable of expressing the modulation process of the bottom-up response from top-down biases. The definition of the equation is based on the idea that some findings seem to support a multiplicative gain modulation procedure (see Section 2.1.2); however, no *specific* procedure appears to be available yet, as findings seem to support no consensus. An in-depth literature in research fields such as neuropsychology and psychology may be needed so as to allow the researcher to have a better understanding regarding the biological processes that occur when the user is perceiving the environment. According to the researcher, a simplistic approach to the integration procedure may represent a considerable limitation of the model herein presented. Other coefficients may be missing in the above equation, and they could be responsible for a more accurate measure of the saliency values. In addition, the literature suggests additive baseline shift as an alternative approach to modulate the response, which can lead to better results in the context of this work. These alternatives can represent possible future work to be conducted in upcoming iterations of the project.

The need to collect data from subjects is another of the limitations of the proposed model. In order to build the top-down component, data from real subjects must be processed. Without such data, in the context of the proposed model, only the bottom-up component is operational. Summoning subjects to participate in the experiment can be a daunting task. Several hours had to be spent by the researcher in order to collect data, and subjects had to be willing to participate in the experiment. Moreover, a minimum number of subjects is required so as to extract meaningful conclusions and apply algorithms such as clustering. The idea of spending approximately 30 minutes and answering to a lengthy questionnaire may prevent more people from participating. Outlining quicker experiments by shortening the questionnaire and the course to follow may represent a possible alternative.

On the other hand, the collected data may not be enough to express the attention behaviours people tend to demonstrate while they wander in an environment. In particular, information such as the user's position and orientation towards the environment may not suffice to model user's attention. Other dimensions may be needed to be collected, and other instruments may be required to be used so as to capture information capable of expressing the attention behaviours in a meaningful manner. The use of questionnaires may represent one of such instruments; however, the development of questions capable of capturing such behaviours in a useful way is one of the downfalls of these tools. Knowledge from statistical fields is paramount for a proper development of these questionnaires.

As for the modelling approach, three different techniques were outlined, including the respective disadvantages each one would carry. As aforementioned, EM-clustering based on Gaussian

Mixture Models was the adopted approach, in which a mixture model is the resulting output for the top-down component of attention and each cluster is represented by a component of the model. This clustering approach assumes however that the collected data can be described by multivariate normal distributions. This assumption may be too strong for the the reality the researcher is trying to model. In addition, the resulting clusters are elliptic-shaped, which represents a restriction imposed by the model. Complex data patterns may not be captured by models based on convex-shaped clusters, which, in turn, may imply loss of information. Different clustering techniques may be used as future work; on top of that, approaches that fall outside the clustering problem may be considered as well. The use of supervised learning techniques may represent a course of action to be taken in subsequent iterations.

A final note pertaining the limitations of the proposed modelling approach concerns the equation used to modulate the output response of the bottom-up component of attention. The attentive reader might have already observed that $sal \geq sal_{bu}$; in other words, no inhibition is performed in the bottom-up response by the top-down component. However, the literature suggests that top-down processes may conduct inhibition mechanisms into the obtained response from bottom-up processes. The integration of inhibition mechanisms may represent a interesting course of research.

3.3 Summary

The description of the implementation of the attention model conducted by the researcher was presented in this chapter. Following the segmentation suggested by the literature, the attention module is divided into two components (bottom-up and top-down), and each of these components were described in terms of implementation in their proper sections.

For the bottom-up component, an implementation of the saliency model proposed by Itti *et al.* was performed. A description of the steps required to transform the input image into a saliency map was conducted. In addition, the resulting image pyramids and respective reduction operations were presented. The relevant implementation details were highlighted; the development of the object identification map, the application of separate kernels, and the use of parallel programming paradigms such as task and data parallelism were the highlighted topics of implementation. Finally, limitations concerning the implementation of the bottom-up component were introduced. The assumption of the same saliency value throughout the object's surface, the limited resolution of the object ID map, the execution time of the component, and the addition of new features such as flicker and motion into the bottom-up process were the main limitations pointed out by the researcher.

As for the top-down component, the considered modelling approaches were presented; such approaches are based on the definition of regions in the environment in which an importance factor is assigned to the elements of the environment. Grid, manual area definition, and mixture models were the considered approaches by the researcher, being the mixture models the adopted approach for this dissertation. For each approach, a description regarding the definition of the regions was

provided, and the definition of the importance factor was presented. A final section regarding the limitations of the modelling approach for the top-down component was provided. The modulation approach, the need to obtain data and conduct experiments, and the use of information other than user's position and orientation were some of the limitations outlined. As stated before, the outlined limitations can be used as starting point for further developments of the proposed work.

Chapter 4

Experiment, Results and Discussion

As stated in Chapter 1, the dissertation proposes a framework capable of defining a function s describing how much objects "pop out" from the scene so that a ranking of the elements of the virtual environment in terms of attention can be obtained. A description of the framework was given in greater details in Chapter 3. Following the architecture suggested in the literature (a review of the literature is provided in Chapter 2), the model is divided into bottom-up and top-down components. The bottom-up response is based on an input image that represents what the user is currently observing from the environment, while the top-down component tries to express the long term cognitive processes that take place while the user is perceiving the environment by modulating the response from the bottom-up counterpart.

Data collected from real subjects is used to support the top-down component, as referred in Section 3.2.1. The behaviour that users show while immersed in the virtual environment is analysed and processed so as to be put to use by the top-down component. Since data collection procedures are required, several elements must be defined to back up the execution of the experiments, namely:

1. the definition of a **test bed**, responsible for providing an infrastructure in which an experimental setup can be outlined with several adjustable parameters. Such setup is then used to immerse users into a virtual environment, allowing the researcher to collect data pertaining the user's behaviour while perceiving the environment.

2. the definition of an **experimental protocol**, so as to allow a systematic approach of conducting the experiments, thus minimising discrepancies among experimental results.

3. the definition of **performance metrics**, which are responsible for assessing the performance of the attention models provided by the framework.

The focus of the current chapter is to provide a deeper thought concerning the aforementioned topics. The test bed and the experimental protocol are presented in greater detail in Sections 4.1

and 4.1.3, respectively. Section 4.2 is responsible for introducing the performance metrics and conducting an analysis regarding the developed attention model. A discussion of the obtained results is also provided in this section, followed by a description of the encountered limitations regarding the experiment and the methodology used for assessing the model.

4.1 Test Bed

In order to model the top-down component of the attention model, data collected from immersed participants is used. As stated in Section 3.2.1, Gaussian mixture models are used to cluster the data; such clustering takes into account the user's location and orientation where the object was observed. In order to retrieve such data, a test bed is defined. Therefore, the current section focus on providing to the reader a thorough description of such infrastructure. A description of the technologies used (eye-trackers, HMD) is given, and an outline of the used virtual scenario is presented as well. Lastly, the experimental protocol is unveiled.

4.1.1 Eye-trackers

While the participant is on the virtual environment, it is crucial to analyse what objects are fixated. In order to achieve this, the position of the pupils are recorded and processed by the eye-tracking cameras attached inside the HMD, near the lens, as shown in Figure 4.1. The eye-tracking cameras are developed by Pupil Labs, and some of the technical specifications are available at their website [1]. The cameras are capable of analysing eye movements as fast as 120 times per second [2], with an accuracy of $\sim 1^{\circ}$ and a precision of $\sim 0.08^{\circ}$. A summary of the technical specifications is made available in Table 4.1. This information was taken into consideration, when developing the code necessary to detect the observed object.

As for the integration with game engines, Pupil Labs provides a Unity plugin to support the integration with this engine. The implementation is made available in a GitHub repository [3]. Alongside with the implementation, some scenes are made available for demonstration purposes, including a calibration scene. This template is used every time an experiment is conducted, just before the participant is immersed in the virtual environment.

Apart from the software responsible for integrating with Unity, standalone applications are also provided. Among such software, *Pupil Capture* is responsible for communicating with the eye-tracking cameras located within the HMD. Such communication is performed by using a message-based API. Three views are made available by the *Pupil Capture* software, in which each view focuses on a particular processing step of the pupil detection algorithm. These views are:

[1] https://pupil-labs.com/vr-ar/, as of 20th June, 2018

[2] the effective framerate will depend on several factors, such as the resolution of the obtained images and the absolute exposure time

[3] https://github.com/pupil-labs/hmd-eyes, as of 6th March, 2018

Table 4.1: Technical specifications of the eye-trackers

Tracking frequency	120 Hz
Gaze Accuracy	$\sim 1^{\circ}$
Gaze Precision	$\sim 0.08^{\circ}$
Connectivity	Through network message-based A.P.I.
Integration with Unity 3D	Yes

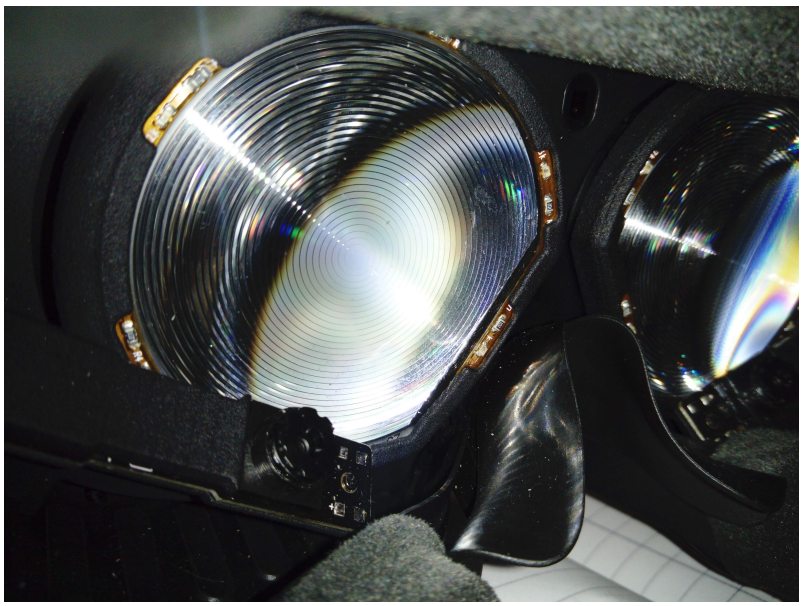


Figure 4.1: Inner side of the *HTC Vive* HMD. The eye-tracker camera is located at the bottom-left size of the image. Also, the Fresnel lens, used to decrease the weight of the optical system, are also visible.

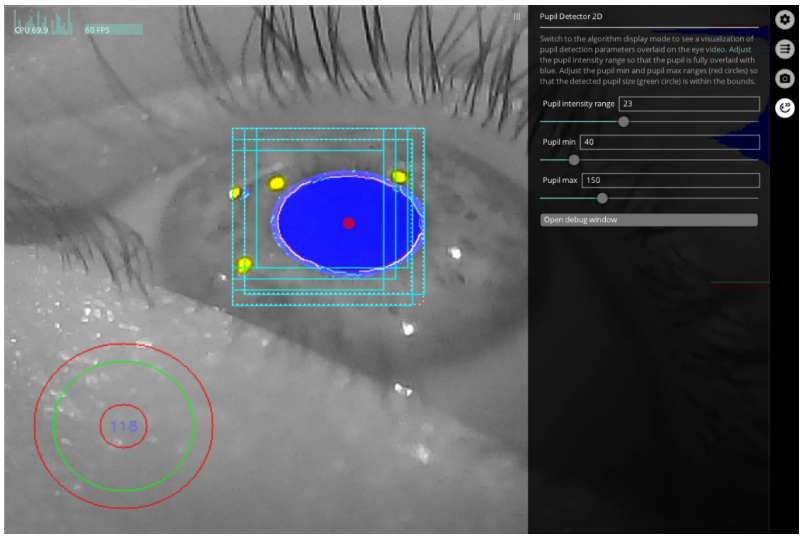


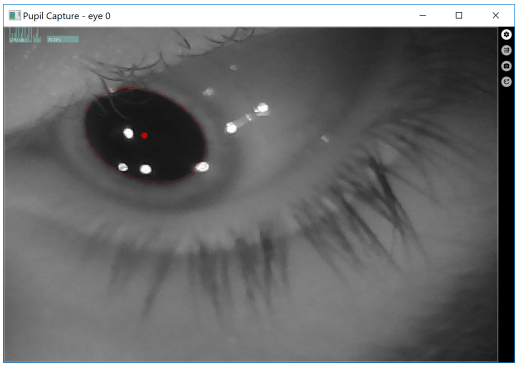
Figure 4.2: Interface of the *Pupil Capture* software. On the left size, a preview of the feed from the cameras is displayed. On the right size, the configuration parameters are available. Note that the parameters are organised into categories (icons on the right side of the image).

- **camera image view** (Figure 4.3a) - shows real-time view of the eye-tracking cameras. If a pupil is being detected, a red dot and an ellipsis are shown to depict the centre of the pupil and the frontier between the pupil and the iris, respectively.

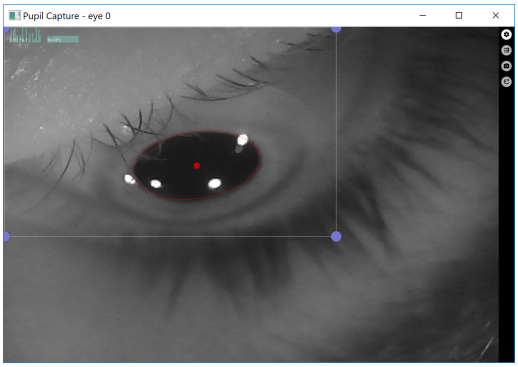
- **R.O.I. image view** (Figure 4.3b) - allows the user to choose the image's region of interest to be analysed by the pupil detection algorithm. A rectangle with four draggable corners appears, and the size and position of the rectangle can be adjusted by dragging such corners.

- **algorithm image view** (Figure 4.3c) - allows the user to perceive the intermediate results output by the detection algorithm such as the areas of the image labelled as "pupil", the image's histogram, and the pupil's diameter.

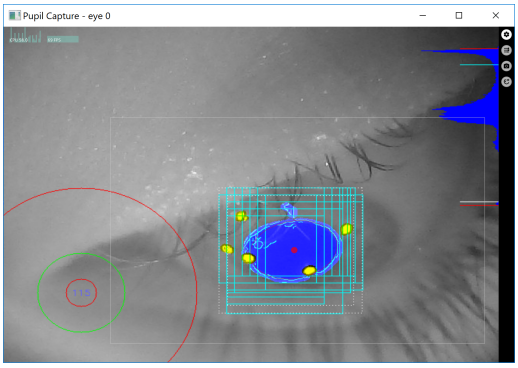
So as to obtain the gaze coordinates, the eye's pupils must be detected, and, therefore, computer vision algorithms are used. Kassner *et al.* give an overall explanation of the used algorithm in their article [KPB14]. The objective of the algorithm is to detect the dark pupil in the infra-red illuminated eye camera images. The received images are converted into a grayscale version, and an initial pupil region is estimated by performing a Haar-like centre surround feature convolution, as suggested by Świrski *et al.* [ŚBD12]. Canny edge detection algorithm is used and thresholds are established on the image histogram. Edges caused by spectral reflections (yellow blobs in Figure 4.2) are removed, and contours are obtained by using connected components. Contours are divided into subcontours and an ellipse fitting is performed so as to obtain candidate pupil ellipses. The results evaluation is based on the ellipse fit of the supporting edges and the ratio of supporting edge length and ellipse circumference. Kassner *et al.* call this ratio "confidence". If the "confidence" is above a defined threshold, the ellipsis is considered to represent the pupil.



(a) Camera image view. If detected, the centre of the pupil and its border are shown as a red dot and a red ellipsis, respectively.

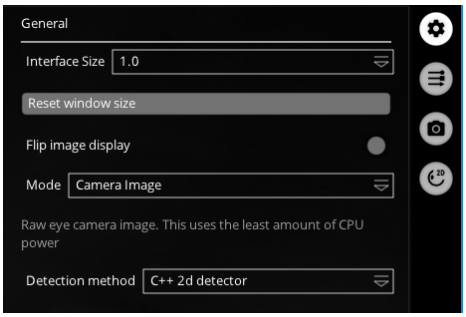


(b) Region of Interest view. A portion of the image can be selected to be analysed by the detection algorithm by dragging each one of the corners of the rectangle. The portion of the image that falls outside the rectangle is not processed by the algorithm.

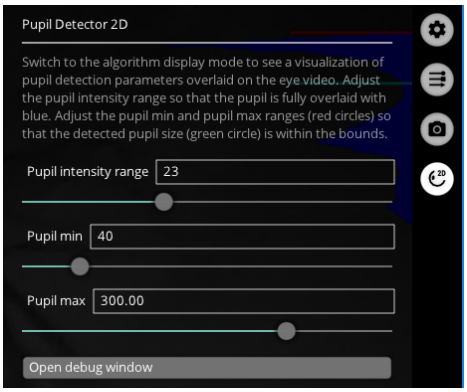


(c) Algorithm view. Intermediary results outputted by the detection algorithm are shown. In the left corner, the the pupil diameter range (red circumferences) and the current pupil diameter (green circumference) are depicted. At the right, the image histogram is rendered. The blue region represents the pupil.

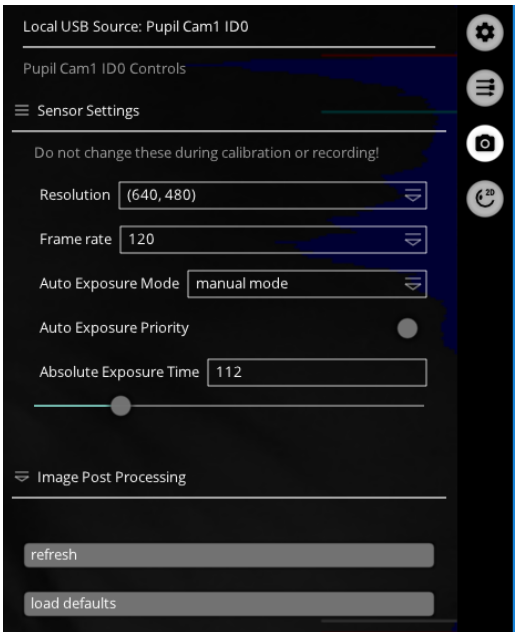
Figure 4.3: The available modes of visualising the eye-tracking image feed.



(a) General settings of *Pupil Capture*. The display mode (Camera Image, ROI, Algorithm) can be chosen.

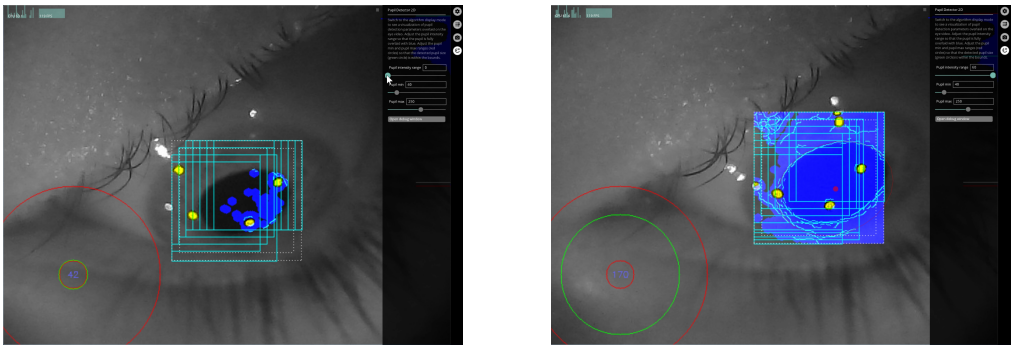


(b) Settings of the pupil detection algorithm. The minimum and maximum pupil diameter can be adjusted. Additionally, the pupil intensity range is also adjustable.



(c) Settings of the input feed. The image resolution, as well as the frame rate and exposure, are some of the adjustable parameters.

Figure 4.4: Parameters available in *Pupil Capture* software.



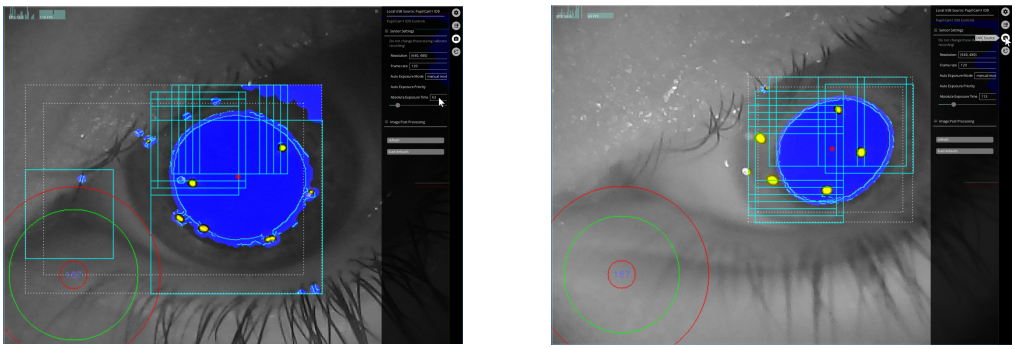
(a) Pupil detection with minimum pupil intensity range (0). Some areas of the pupil are not correctly labelled by the detector.

(b) Pupil detection with maximum pupil intensity range (60). Some portions of the iris are also considered as "pupil".

Figure 4.5: Comparison between different values of pupil intensity range.

Various configuration parameters pertaining the pupil detection algorithm are made available inside the application, as shown in Figure 4.4. For instance, the minimum and maximum size of the pupil are some of those parameters (Figure 4.4b). The definition of a range for the pupil diameter is required since the size of the pupil can vary according to the brightness of the visual stimulus. The proper adjustment of this range may filter out the detection of some false positives. As for the parameter "pupil intensity range", it is responsible for adjusting the degree of tolerance that is used by the pupil detection algorithm when segmenting the image in "pupil" regions. Figure 4.5 demonstrates the effects obtained by varying the values of this parameter. In addition, so as to increase the contrast between the pupil and the iris, the absolute exposure time may be adjusted. The adjustment of this parameter will dictate the amount of time the camera will take to capture a frame. Greater values will dictate greater contrasts, at the cost of the number of frames per seconds, which can decrease. Thus, a balance between good contrast (and thereby, a greater probability of correct pupil detection) and responsiveness is at stake. For this project, a lower bound of 60 FPS was established as the number of frames per second in the virtual environment are located around 70 FPS. In the opinion of the researcher, 60 FPS represent a good compromise. Figure 4.6 gives a comparison between different values of absolute exposure times, in which Figure 4.6a depicts some regions being wrongly classified as belonging to the pupil. A greater exposure time allows a better detection of the pupil region since a greater contrast between the pupil and the iris can be obtained.

Having the parameters of the pupil detection algorithm properly calibrated, the next step is focused on establishing a transformation function between the position of the user's pupils and gaze coordinates in the virtual environment. As such, a calibration procedure must be conducted. After the HMD is put on the participant's head, and after the *Pupil Capture* program is executing and receiving the feed from the cameras, a set of dots is displayed, one at a time, for a predefined period of time, around a circumference, as shown in Figure 4.7b. The subject is asked to look at the centre of such dots, allowing the software to adjust a transform function capable of mapping the eyes' position to virtual world coordinates. After all dots are displayed and presented to the



(a) Captured image with absolute exposure time of 63.

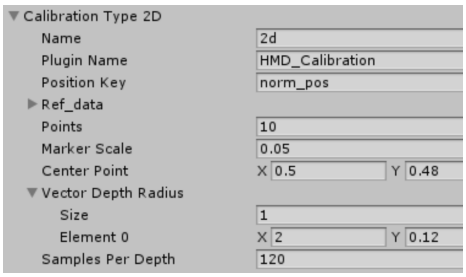
(b) Captured image with absolute exposure time of 113.

Figure 4.6: Comparison between different absolute exposure times.

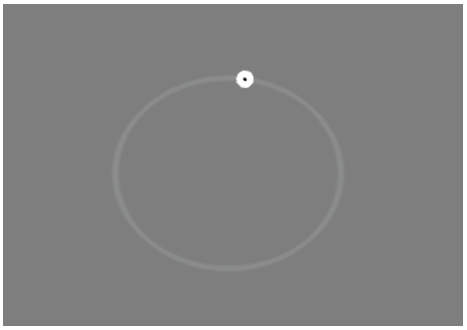
user, the aforementioned function is obtained, and the virtual world is loaded. While the scene is being executed, the gaze coordinates can be queried at any time. Note that the calibration procedure is customisable, with several adjustable parameters located in `PupilSettings` file (see Figure 4.7a) inside Unity. The number of points to be displayed, the displacement between them, and the amount of samples to be ignored while collecting eye position data for each dot are some of those parameters. Note that the calibration procedure must be conducted every time the simulation is executed as the transformation function is not persisted. Additionally, for different people, distinct transformation functions can be obtained.

A debug mode was developed so as to check whether a correct mapping of the gaze coordinates has been obtained. If activated, a square is shown to the used, and it follows the gaze of the user. Moreover, when the user is observing a given object, a box surrounding said object is rendered, thus signalling the fact that the object is being observed. If the gaze goes outside the boundaries of the observed object, the box disappears. Figure 4.8 depicts the execution of the simulation with the aforementioned debug mode activated.

Throughout the tests performed during the implementation phase, the presence of jitter was observed when collecting data pertaining the position of the pupils. The presence of noise, as well as the image resolution of the feed from the cameras represent some of the causes of such jitter. In order to reduce it, *Kalman filters* [Kal60] were implemented and used. These filters can be seen as an estimator that takes a series of (noisy) measurements taken over time and updates its internal model state so as to produce an estimate of the true value of the variables of interest. A joint probability distribution over the considered variables is used and updated as more data is collected. It follows an iterative approach, meaning that the filter can be updated as new measurements arrive. It assumes that the observed and latent variables follow a Gaussian distribution. Kalman filters are updated in two steps: prediction and measurement. The **prediction** stage is responsible for predicting the next state of the system based on current knowledge. Since this is a prediction, some errors may be present. The **measurement** step is responsible for processing the received measurement, which can incorporate some noise, and update the prediction accordingly. To the interested reader, please refer to [Kal60] and [Far12] for more information regarding the equations

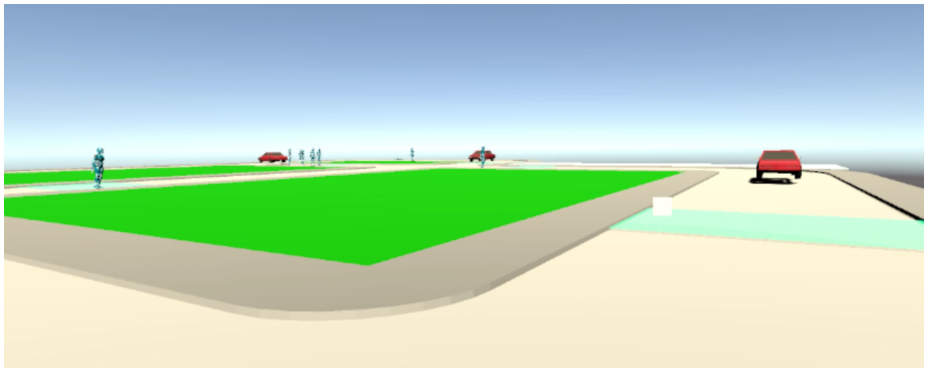


(a) A subset of the available calibration parameters for the calibration scene.

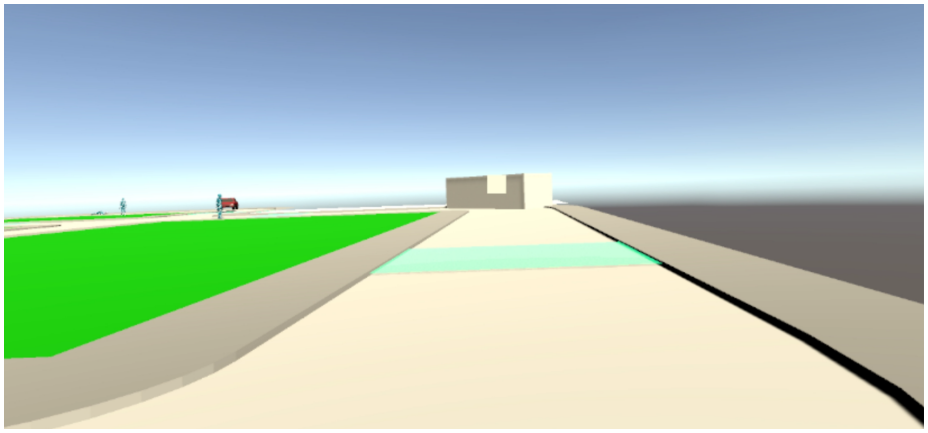


(b) The calibration procedure. The user is asked to look at the centre of the marker. While the user is looking at it, the software collects multiple samples of the eyes' positions in order to adjust a transformation function.

Figure 4.7: The calibration procedure of the eye-trackers, with the respective calibration parameters.



(a) An example in which no object is being observed by the user.



(b) An example in which a car is being observed by the user. Note the box surrounding the object, indicating it is being observed.

Figure 4.8: A snapshot of the simulation with debug mode activated. The square represents the position of the user's gaze.

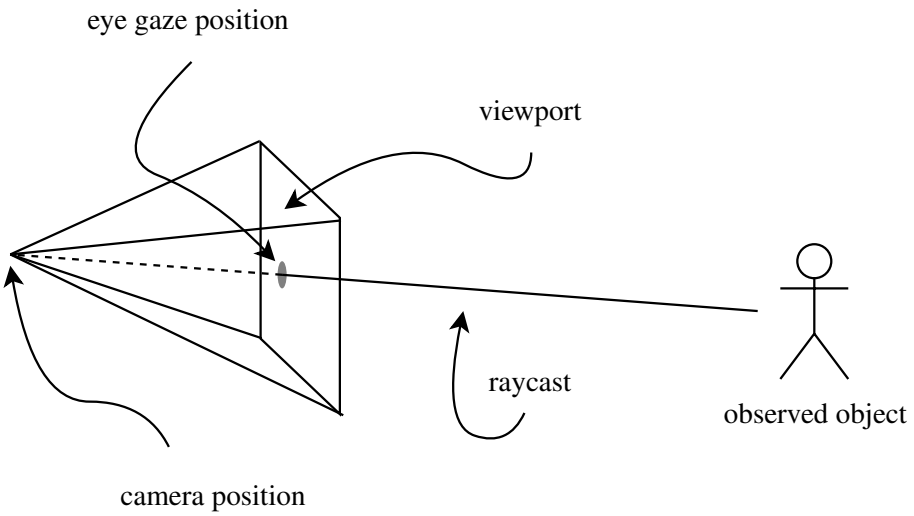


Figure 4.9: A graphical explanation of the raycast mechanism. This procedure is used to detect the observed object by the user.

responsible for updating the system state. Since the implementation of Kalman filters requires the manipulation of matrices, such manipulation was provided by an external library[4] compatible with the framework `.NET 3.5`.

One question that may arise is related with how the objects are detected, given the coordinates of the user's gaze. The attentive reader may realise that multiple steps are required in order to retrieve the object the user is looking at. First, a transform function from the eyes' positions to viewport coordinates is needed. The calibration step is responsible for constructing such map as accurately as possible. Using this map, it is possible to know the 2-D coordinates in the viewport the user is currently looking at. The next step is thus projecting those coordinates into the 3-D environment. This can be achieved by applying a technique called raycast, which allows one to determine the intersection point of a line with a surface. With the position of the camera and the 2-D coordinates of the user's gaze, a ray is cast from the position of the camera and passing through the gaze position. If the user is looking at an object, the ray will collide with it and the observed object is known. Unity already implements this technique, giving the option of filtering the objects that are not relevant for the analysis (for instance, the ground). Figure 4.9 gives a graphical overview of the above explanation.

So as to obtain the observed object, collision detection mechanisms are used. Such collision can be detected by colliders, which can be seen as volumes surrounding the object and are used by collision detection algorithms so as to detect and, eventually, respond to collisions between elements. Most of the times, objects present a very complex geometry, which makes the process of detecting collisions quite difficult and thus expensive. Since most of the times a precise collision detection is not needed, a simpler approximation of the object's geometry is used for collision purposes. Shapes with a simpler geometry such as parallelepipeds, spheres, and cylinders are

[4] https://numerics.mathdotnet.com/, as of 20th June, 2018

used as a surrogate of the object's geometry, working as colliders. The object is then surrounded by a collider, and the latter is used when one wishes to detect whether a collision with other objects took place. In Unity, colliders are represented as components that can be attached to game objects. Particular attention was given to colliders due to the fact that objects located at great distances from the subject are more difficult to be detected since they appear smaller to the user. Smaller objects are more prone to incorrect detection, since this requires a higher precision measurements from the eye-trackers. In order to minimise the detection error, the size of the colliders is dynamically adapted according to the distance between the corresponding object and the player. Greater distances lead to greater collider sizes. Objects placed near the user have colliders with almost the same size as the original size of the collider. The piecewise-defined Function 4.1 was used to obtain a scale factor for the collider size, which establishes a relationship between the distance between the object and the user, and the size of the collider. The slope was defined by empirical observation. Smaller objects have a greater slope value (the traffic signs have $coeff = 2$), whereas bigger ones have a smaller value (for instance, the cars have $coeff = 1$). To prevent a situation where the size increases indefinitely, the maximum distance of 60 meters is imposed, meaning that the collider stops growing after the distance threshold is attained. Such maximum distance is also empirically established. One may wonder why such empirical exercise was not based on the error parameters reported by Pupil Labs. Indeed, they cause an impact in the obtained results pertaining the position of the pupil; however, other factors such as errors in the calibration, and relative displacements between the user's head and HMD can bring unquantifiable errors in the final measurements, hence the adopted approach.

$$f(x) = \begin{cases} \frac{coeff}{100} * x + 1 & 0 \leq x \leq 60 \\ \\ \frac{coeff}{100} * 60 + 1 & x \geq 60 \end{cases} \tag{4.1}$$

In conclusion, a focus on the use of the eye-trackers was given in this chapter. A brief description of the technical characteristics was provided and the integration with game engines such as Unity was highlighted. The software bundled with the eye-trackers was reported, including the *Pupil Capture*. Within this software, a pupil detection algorithm is applied to the received image feed of the cameras, and as such, a description of the configuration parameters was presented. With a proper adjustment of the parameters, the calibration procedure was explained. Due to noisy measurements, the use of estimators such as Kalman filters was proposed in the implementation. Finally, so as to detect the observed object, the raycast algorithm was used together with adaptive colliders, in which the size of these elements is adjusted according to the distance between the observer and the object.

4.1.2 Scenario

So as to conduct the experiments and validate the developed framework, a virtual environment was developed. Requirements such as a high degree of realism and responsiveness were considered

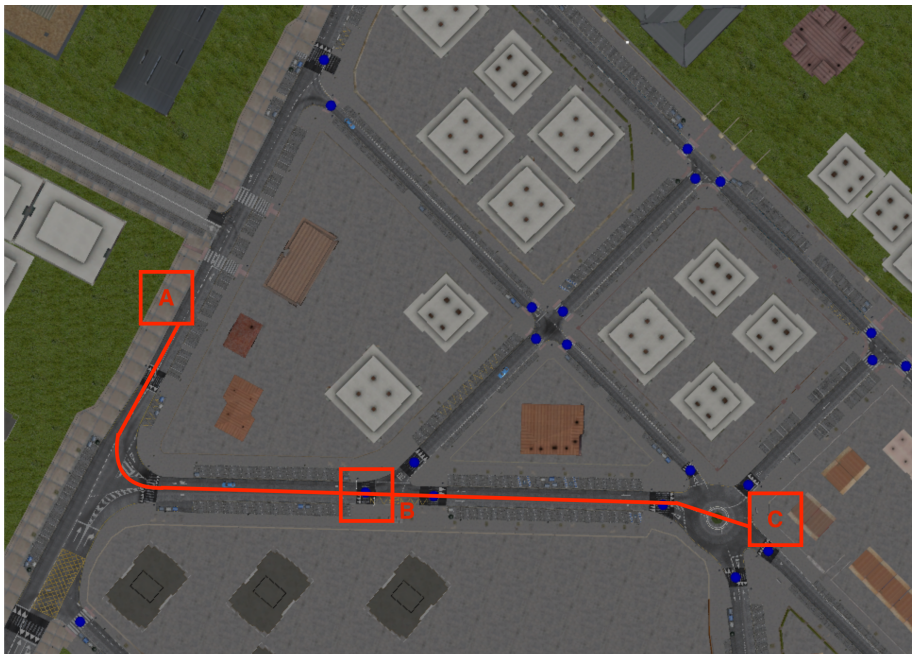


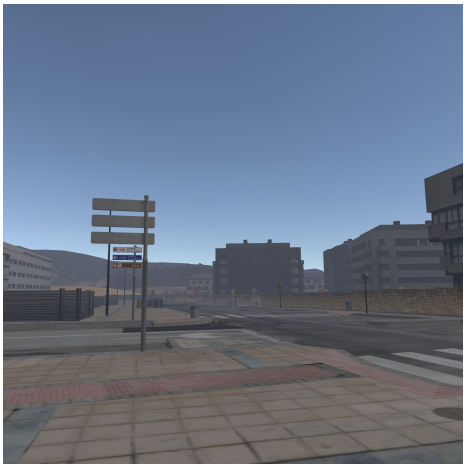
Figure 4.10: Aerial view of the scenario. A path is defined which goes through the outlined regions. Regions A and C are the extremities of the adopted course, and region B is located at the crosswalk with the traffic lights.

key features, as limitations of the environment should cause the least amount of impact into the obtained data. As the outcome of the dissertation is framed in the project SIMUSAFE, a Unity scenario of a Spanish city provided by ITCL (Technological Institute of Castilla y León) was used. Urban furniture such as trash cans, traffic signs, and buildings are also present in the scenario. Dynamic elements (in particular, cars and pedestrians) were made with the purpose of populating the virtual environment. Dynamic traffic lights were added in specific roads of the environment, in which cars and pedestrians take into account its current state. Figure 4.12 shows some of the elements present in the virtual environment. Since the scenario is quite extensive, a portion of the scenario is used. So as to further narrow the scope of the project, a focus on crossing streets is given. Therefore, three regions are outlined based on three types of crossings, representing common crossing situations in urban environment:

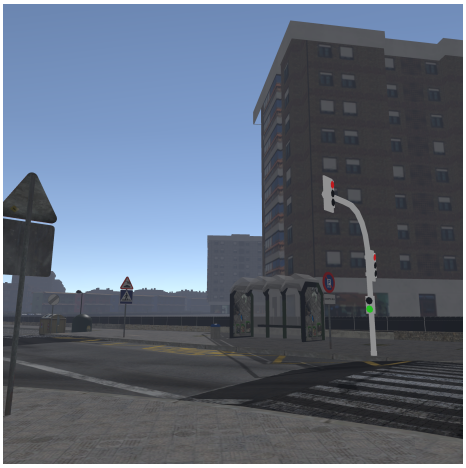
- crosswalk with no traffic light (region **A**)
- crosswalk with traffic light (region **B**)
- crosswalk at roundabouts (region **C**)

Additionally, a path visiting the aforementioned regions is outlined as well. An aerial view of the scenario is depicted in Figure 4.10. In this map, the outlined regions are presented, including the selected path. An in-scene view of the aforementioned regions is available in Figure 4.11.

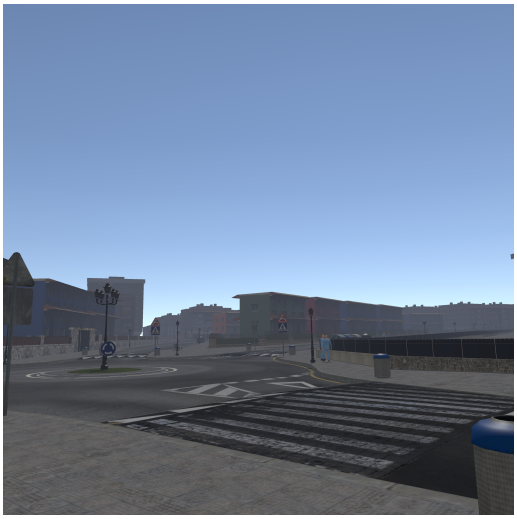
As for the responsiveness component of the virtual environment, an avatar is created, which can be controlled by the user in order to navigate freely throughout the scene. This avatar can



(a) Crosswalk *without* traffic light (region **A**).



(b) Crosswalk *with* traffic light (region **B**).



(c) Roundabout (region **C**).

Figure 4.11: Snapshots of the regions of the virtual environment.

(a) Car.

(b) Pedestrian.

(c) Traffic light and traffic sign.

Figure 4.12: Some examples of the elements available in the scene.

Figure 4.13: Controller with touchpad being pressed. In order to move the avatar, the touchpads on both controllers must be pressed while, at the same time, moving them in a swing-like movement.

be seen as the projection of the user's presence in the virtual environment. Some mechanics were implemented to allow users to control the avatar. Using the `VRTK` [5] framework, HTC Vive controllers can be used to move and rotate the avatar's body. The avatar can be moved by pressing and holding the touchpad (as shown in Figure 4.13) on both controllers and swing them back and forth, so as to imitate the arms' natural movement when walking. The speed of the avatar in the environment is dictated by how much the controllers are swung. In order to rotate the avatar, the user must rotate his body as well while holding the controllers. It is worth noting that the head can be rotated independently from the body. Therefore, the user can walk straightforward while looking sideways, allowing a greater degree of freedom of movement.

At any moment, the position of the avatar is known. In order to obtain some information regarding the interactions conducted by the user, a geometric approach is used to test whether the user is currently located at one of the aforementioned regions. In fact, two approaches were used to defined the regions:

- **axis-aligned rectangular shape** - to define a rectangle, the two opposite vertices of the rectangle $\{(x_{min}, y_{min}), (x_{max}, y_{max})\}$ must be supplied. This forces the edges to be axis-aligned with the assumed coordinate system. If $(x_{min} < x < x_{max}) \wedge (y_{min} < y < y_{max})$ holds, then the point (x, y) is within the rectangle with vertices $\{(x_{min}, y_{min}), (x_{max}, y_{min}), (x_{min}, y_{max}), (x_{max}, y_{max})\}$.

- **non-axis-aligned rectangular shape** - this approach does *not* force the edges of the rectangle to be aligned with the adopted coordinate system axes. As such, 3 vertices (A, B, and D)

[5] https://vrtoolkit.readme.io/, as of 16th April, 2018

must be defined [6]. If $(0 < \overrightarrow{AM} \cdot \overrightarrow{AB} < \overrightarrow{AB} \cdot \overrightarrow{AB}) \wedge (0 < \overrightarrow{AM} \cdot \overrightarrow{AD} < \overrightarrow{AD} \cdot \overrightarrow{AD})$ holds, then point M is located within the rectangular shape defined by the vertices $ABCD$.

In conclusion, a description of the used scenario was presented in this section. A listing of the elements available within the environment was given. Taking into account that a focus on street crossing was given during the project, the definition of three regions (crosswalk with *no* traffic light, crosswalk *with* traffic light, roundabout) was provided. In addition, the definition of a path visiting the aforementioned regions was given. So as to give a contextual framing to the reader, an aerial view of the scenario presented. A virtual projection of the participant was implemented into the environment, which can be controlled by interacting with the HTC Vive controllers. Finally, a geometric approach to the definition of regions was outlined.

4.1.3 Experimental Protocol

In order to support the attention model, data is collected from participants that are immersed in a controlled virtual environment. While immersed, they are guided to follow the path defined in Section 4.1.2. At the same time, several metrics pertaining the user's behaviour in the VE are gathered. In this subsection, the adopted experimental protocol is outlined, and a definition of the collected metrics, as well as the questionnaire used at the end of each experiment, is presented.

An experimental protocol was defined to ensure a systematic approach during experiments, which consisted of the following points:

1. **Explain the experiment** - the topic of research and the objectives of the dissertation are mentioned and the objective of the experiment is described, so that the participant can have an understanding regarding his contribution to the dissertation.

2. **Show an aerial view of the virtual scene** - the regions the participants will have to visit are outlined in this map. The objective of this step is to allow a first contact from the participant with the scene.

3. **State the conditions of the experiment** - all the equipment that is going to be used is shown to the participant. At this phase, the feed obtained from the eye-trackers is also shown, allowing the participants to know the type of data that is going to be analysed.

4. **Ask the participant to sign the consent form** - if the participant agrees to take part of the experiment, the researcher shall ask the former to sign the consent form before proceeding to the next phase. The used consent form is available in Appendix A.1.

5. **Execute the scene** - the HMD (Figure 3.1b) is put on the participant's head, and all straps are properly adjusted without causing any injuries to the participant. Also, the controllers are given (Figure 3.1a) to the participant as well. Throughout the experiment, the researcher will check if the the participant is feeling any discomfort, any symptom of cybersickness.

[6] The fourth vertex C is constrained by the remaining vertices of the rectangle, and thus it is not required to be defined.

(a) **Execute a dry-run** - the first iteration of the execution is focused on showing the mechanics to the participant. This is especially important for those who are not experienced with VR equipment. In addition, the interpupillary distance of the HMD is adjusted, and the parameters of the pupil detection algorithm are also adjusted accordingly.

(b) **Execute the calibration procedure of the eye-tracker** - the calibration scene is executed. The participant is asked to not move the HMD, in order to ensure a proper calibration. Moreover, the participant is also asked to look at each of the dots that appear throughout the calibration procedure.

(c) **Check if the calibration procedure was successfully performed** - after the execution of the calibration procedure, when the virtual scene is loaded, a small square representing the position of the user's gaze appears in front of the subject. The user is asked to look at specific locations (for instance, a car), and the researcher checks if the square falls within the vicinity of the chosen location. Additionally, if the calibration was correctly performed, a box involving the observed object appears, thus indicating the object was successfully detected.

i. *If the calibration failed* - if the square is misaligned, the subject is asked to repeat the calibration step. Relative displacements between the user's head and HMD caused by small movements of the head are the most probable cause of misalignment. As such, the researcher checks if the HMD is properly fitted on the subject's face, and the calibration procedure is executed again.

(d) **Follow the assigned path** - the subject is assigned a path to follow. As the subject is following the path, metrics such as his position in the virtual world and the observed object are collected.

6. **Fill the questionnaire** - when the path is completed, and after removing the equipment from the subject, the questionnaire is presented to the subject, and he is asked to answer it.

7. **Thank for the participation** - the researcher thanks the participant for the cooperation.

At any moment, during the experience, the participant may ask any questions. Also, if the participant wishes to pause or leave the experiment, he may do so without any harm to him. Some people might feel some discomfort (such as headache, nausea, among other symptoms) when they are placed in a VR session, a phenomenon called cybersickness [Kol95, Bar04]. Throughout the experiment, the researcher must ask frequently if the participant is comfortable, and, if not, the experiment must be suspended. The experiment may be resumed, if the participant agrees; otherwise, the experiment is terminated.

While the participant is immersed, he is asked to follow a given path. The researcher gives orientation to the participant as to the route he must follow. The objective is to arrive at a predefined destination and return at the starting point without being run over. Several crosswalks are made available, which can be used by subjects to cross streets. However, no enforcement is imposed as

to follow the traffic rules. Note that, for the sake of simplicity, only two-way streets are considered alongside the path.

If all the participants were to follow the path from the same starting point, some problems may arise with the collected data. For instance, the participants will most likely get tired when arriving at the end of the path, possibly compromising the collected data and biasing the obtained conclusions. In order to minimise such biases, the starting point of the path is randomised from participant to participant. Regions **A** and **C** are the selected starting points. If the subject starts at region **A**, he is asked to go to region **C** and return, following the path depicted in Figure 4.10. Otherwise, if region **C** is chosen as the starting point, the user is asked to go to region **A** and return, following the same path as before. Note that the subject is asked to cross the crosswalk at region **B** at least once per session, independently of the starting point. As the experiments are made, the next available permutation is chosen. `Random.org` [7] was used so as to obtain the permutations, which are presented in Figure 4.15.

$$A \Rightarrow B \Rightarrow C \Rightarrow B \Rightarrow A \tag{4.2}$$

$$C \Rightarrow B \Rightarrow A \Rightarrow B \Rightarrow C \tag{4.3}$$

Figure 4.14: The available directions of the chosen path.

$$4.3 \Rightarrow 4.3 \Rightarrow 4.2 \Rightarrow 4.2 \Rightarrow 4.2 \Rightarrow 4.2 \Rightarrow 4.3 \Rightarrow 4.3 \Rightarrow 4.3 \Rightarrow 4.2 \hookleftarrow$$

Figure 4.15: The obtained permutations. At the end, the sequence of permutations is repeated.

Before the end of the experiment, a questionnaire (see Appendix A.2 to observe the used questionnaire) is presented to the participant, being divided by the following sections:

- **sociodemographic questions**, such as age, sex, the possession of a driver's license, the most used means of transportation, and experience with VR equipment.

- **pedestrian behaviour questions**, focusing on the behaviour of pedestrians at crosswalks. How often the participant crosses the street when a red light is in place, or how often the street is crossed without paying attention to both sides of the street are examples of such questions. A validated questionnaire for the U.S. population proposed by Deb *et al.* (see [DSD+17]) was used as basis for this part of the questionnaire.

[7] `https://www.random.org/sequences/`, as of 20th June, 2018

- **questions pertaining the regions of the virtual environment**, in which the degree of importance given to the cars, other pedestrians, or even to signage is inquired, and the perceived objects' observation time is queried.

- **questions pertaining the immersiveness of the virtual environment**, based on the questionnaire proposed by Tcha-Tokey *et al.* (see [TTCLER16]).

The objective of this questionnaire is to supplement the collected data. With the data obtained by the questionnaire, a description of the sample can be obtained. In addition, some aspects pertaining the behaviour of users at crosswalks are also queried, as it is important to have an idea of the user's perception of his behaviour. The purpose of the questions concerning the regions of the virtual environment visited by the subjects is, once again, to get a glimpse of the user's perception concerning the virtual environment at those locations. The questions regarding the immersiveness of the virtual environment are important so as to assess the extent the virtual environment can be considered as a proxy of the real environment. The possibility of using the VE to collect data expressing some aspect of the user's behaviour in the real environment is crucial for the developed project.

Some remarks pertaining the questionnaire should be highlighted. In particular, some adaptations of the questionnaire proposed by Tcha-Tokey *et al.* were performed. The complete questionnaire is composed by 87 questions, which, for the purpose of the project, was considered to be quite extensive, taking into account the substantial questionnaire made by the researcher. As such, a shorter version made available in their paper ([TTCLER16]) was used. However, some adaptations of their questionnaire were made, as some questions were not adequate for the purposes of the project. For instance, the question "*I felt confident selecting objects in the virtual environment*" does not make sense in the developed virtual environment, as no emphasis was given to selecting objects in the virtual environment. In fact, no object can be selected by the user while immersed, therefore, the question was not considered into the questionnaire. Moreover, the question "*Personally, I would say the virtual environment is impractical/practical*", also present in the shorter version of their questionnaire, was adapted into "*The virtual environment was realistic*". The researcher was interested in the subject's perception concerning the realism of the scenario. As for the remaining questions, they were kept intact and considered into the questionnaire.

Some remarks are also in order for the second and third sections of the questionnaire. The questions pertaining the pedestrian behaviour at crosswalks are based on the questionnaire proposed by Deb *et al.* in their article (see [DSD+17]). Two versions are made available: a longer version composed by 50 items, and a shorted version, which is a subset of 20 questions based on the longer version. The shorter version was used in order to minimise the length of the researcher's questionnaire. Although this questionnaire was based on the US population, no issues were immediately detected when adapting it to the context of the Portuguese population. As for the questions regarding the user's perception of the elements of the VE, they were based on information needs of the researcher.

In conclusion, the experimental protocol used to conduct the experiments was presented in this section. An explanation of the steps that compose the protocol was provided. Moreover, the approach used to orient the participant in he virtual environment was given. A path was outlined, which goes through three predefined regions. The starting point of the path is randomised between experiments, so as to compensate eventual fatigue at the end of the path. Finally, a description of the questionnaire and the sections which composes it was supplied. Four sections were outlined: sociodemographic questions, questions pertaining pedestrian behaviour, questions regarding the user's perception at the defined regions of the virtual environment, and questions pertaining the immersiveness of the virtual environment. Some observations concerning the questionnaire were exposed as well. The questionnaire used in the conducted experiments is available in Appendix A.2.

4.2 Results and Discussion

With the definition of the test bed, and the outlining of the experimental protocol, the next step is focused on processing the collected data and infer some conclusions from it. Therefore, the current section will focus on such data analysis. A focus will be given at two components of the project:

- **participants' experience** - a descriptive analysis of the obtained sample is conducted. Additionally, the immersiveness of the virtual environment is discussed.
- **attention model** - the performance of the attention model is analysed. A performance metric is defined and used to assess the potential of the attention model, followed by a discussion of the obtained results.

A final section will focus on limitations pertaining the aforementioned components.

4.2.1 Participants' Experience

Throughout the development of the experimental setup, four experimental series were considered. Each series is characterised by a set of procedures that must be followed while conducting the experiments. Any changes to be performed in the questionnaire must result in a creation of a new series, as those changes may result in different interpretations from participants. As such, the obtained results from the new version of the questionnaire cannot be used together with the results from previous versions, otherwise, erroneous results may be obtained. In addition, each time a modification is performed in the experimental setup, a new series is created as well. Throughout the development of the project, no changes were needed to be implemented into the experimental protocol.

In fact, changes in the questionnaire were the main reason that new series were created. In the first series, three sections of the questionnaire were dedicated to each region of the environment, in which each section asked the participant some questions pertaining his perception of

the environment. However, the researcher got a negative feedback from the participants; they complained about the length of the questionnaire, saying that it had a considerable size; indeed, participants were spending 15 to 20 minutes solely on answering the questionnaire. In response to this feedback, the questionnaire was shortened, by dedicating only one section to all regions of the environment. The remaining changes that culminated into the remaining series were due to interpretation issues of the questions. For the first three series, 5 subjects participated in the experiment. For the remainder of the discussion in this section, the data collected from the first three series is not considered. It is worth noting that not all sections of the questionnaire were used for the final result analysis. In particular, the sections regarding the users' pedestrian behaviour at crosswalks and the users' perception of the environment at the designated regions of the VE were ignored in the analysis presented in this section, as such questions seem to be not relevant when taking into account the objectives the researcher proposed for the dissertation. In fact, the attention model does not focus solely on the users' behaviour when crossing streets; as for the users' perception at the designated regions of the VE, throughout the experiments, the researcher observed that different interpretations were given to the questions, which can lead to inconsistencies into the obtained results. Ultimately, a compromise between the length of the questionnaire (and, thus, the amount of time participants must spend answering it) and the amount of information that can be collected from it is at stake. The researcher preferred an approach in which more questions than the minimum required are exposed to the participant. The excess of information can be trimmed, but the lack of it cannot be recovered. If a question is missing, less information can be retrieved after all participants have answered the questionnaire.

In total, each session of the experiment lasted approximately 30 minutes. 27 subjects participated in the experiment with ages between 17 and 35, with an average of 23.04 years and a standard deviation of 3.7 years. $2/3$ of the participants were male, and the remaining portion were female. 20 out of 27 said to be students, while the remaining claimed to be researchers, professors or PhD students. With 37%, most of the participants had the 12^{th} grade as their completed educational level, followed by bachelor degree (25.9%), master's degree (25.9%), PhD (7.4%) and finally 9^{th} grade (3.7%). Almost everyone had a driving license (26 out of 27 participants). As for the most used means of transportation, car is the preferred one, as shown in Figure 4.17. Note that participants were allowed to choose more than one option. Participants were also asked about any vision problems they have. 16 of them alleged to have some sort of vision problems, with myopia, astigmatism and hypermetropia being the selected problems. The majority of the participants (23) was right handed, while the remaining alleged to be left handed. No participants said to have any locomotion problems. Participants were also asked about their experience with VR equipment, and the distribution of their responses concerning this question is made available in Figure 4.16.

A particular focus will be given to the responses from the participants about the sense of immersiveness they had in the virtual environment. This is important so as to ensure that the obtained results are close to those one could have gotten if the experiment had been performed in a real scenario. The objective of section 9 of the questionnaire is then to assess the extent of the impact that the use of HMD has in the execution of the experiment. A summary of the responses

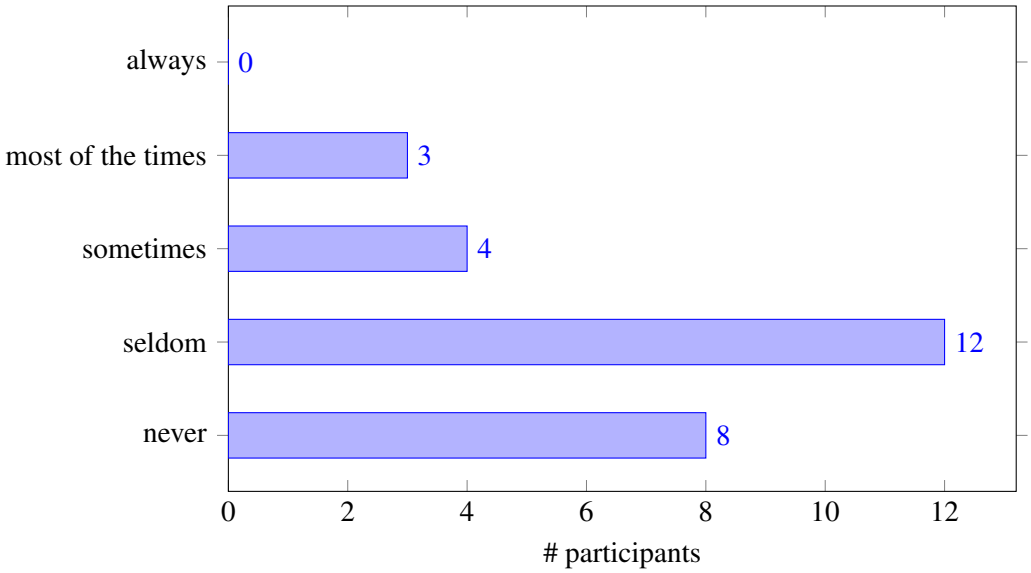


Figure 4.16: Participants' experience with VR equipment.

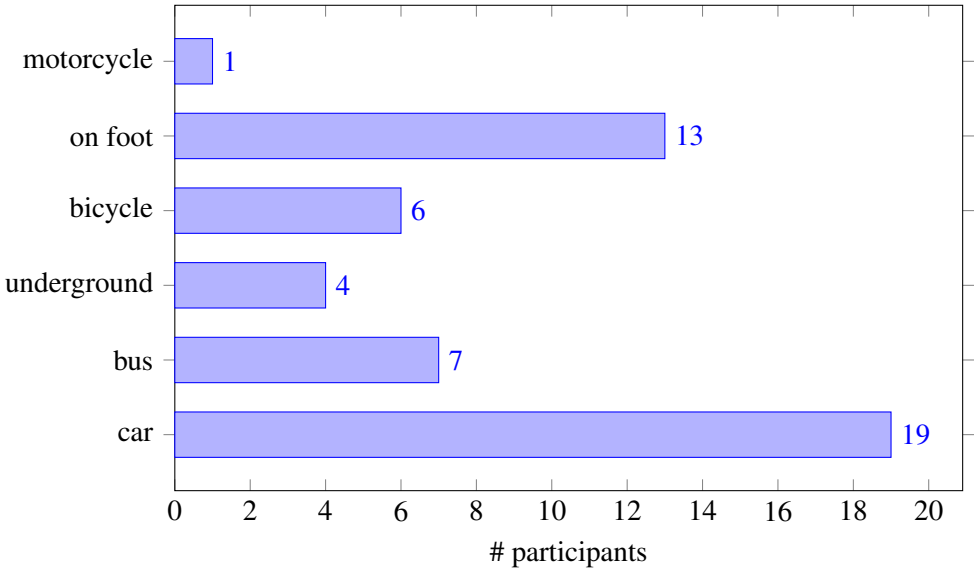


Figure 4.17: Distribution of answers for the question related to the means of transportation used by the participants.

Table 4.2: Questions from section 9 of the questionnaire.

ID	Component	Question
9.1	Presence	The virtual environment was responsive to actions that I initiated.
9.2	Engagement	The sense of moving around inside the virtual environment was compelling
9.3	Immersion	I felt stimulated by the virtual environment
9.4	Flow	I felt I could perfectly control my actions
9.5	Usability	I thought the interaction devices (headset, gamepad and/or keyboard) were easy to use
9.6	Emotion	I enjoyed being in this virtual environment
9.7	Judgement	The virtual environment was realistic
9.8	Experience consequence	I suffered from fatigue during my interaction with the virtual environment
9.9	Technology adoption	If I use again the same virtual environment, my interaction with the environment would be clear and understandable for me

Table 4.3: Statistics of the responses from participants to the questions pertaining the user experience in the virtual environment.

Question	Median	Mode	IQR
9.1	8	8	1
9.2	8	8	2
9.3	9	9	2
9.4	8	9	**2,5**
9.5	**10**	**10**	1,5
9.6	**10**	**10**	1
9.7	8	9	**2,5**
9.8	2	1	2
9.9	9	**10**	2

of the participants is made available in Table 4.3.

In general, the participants claimed the virtual environment was responsive to their actions. They also tended to agree to the fact that the sense of moving in the virtual environment was compelling. Most subjects felt stimulated by the virtual environment as well. As for the responsiveness of the actions, people claimed to have total control over the actions they could perform. In addition, when asked if the participants agreed the VE was realistic, they would corroborate the statement in general. However, a consensus was difficult to achieve in these question, as show by the IQR values for questions `9.4` and `9.7`. This can reveal a weak aspect of the virtual environment. The peripherals were considered to be easy to use and almost every participant enjoyed to be in the VE, as shown by the median value for the questions `9.5` and `9.6`. Users were also asked whether they felt some discomfort during the experiment, and most of them gave a negative response. The median value for question `9.8` tends to support such observation. Nevertheless, some exceptions occurred, with cybersickness being the main factor. Some endeavours may be put forth in order to minimise these occurrences in future iterations of the project. Lastly, users were asked if future interactions with the VE would be simple and easy to be done, which they are likely to agree.

In the final section of the questionnaire, participants were allowed to give their opinion regarding the conducted experiment. The positive aspects of the experiment, as well the negative ones, were inquired. Participants were also asked what improvements could be done to the experiment.

The following points represent a summary of the positive aspects of the experiment considered by the participants:

- First contact with Virtual Reality
- Scene fidelity
- Controls' responsiveness
- Avatar responsiveness
- Clear objective of the experiment without stating what is being monitored

Throughout the experiments, the researcher highlights the fact that the realism of the scene and the system responsiveness were the most praised features of the experiment. Moreover, for those who never experienced VR, they classified the experiment as a gratifying moment for themselves.

On the other hand, as for the negative points highlighted by the participants, an outline of such points is given:

- Cybersickness
- Sometimes unrealistic avatar movement and control
- Limitations of HMD, such as blurred peripheral area
- Burning sensation in the eyes

- The behaviour the participant had in the virtual environment may not match the one he has in reality
- Slope between the street and the walkway in the virtual environment sometimes hinders the movement of the avatar. Also, the avatar can get stuck between objects

Indeed, a few number of participants (3 out of 27) felt some considerable discomfort during the experiment. One of them was unable to complete the experiment, with only the initial test completed. As for the majority of the participants, they felt no to mild discomfort during the experiments. In order to minimise such occurrences, a small break was done between the the first and second executions of the experiment. A burning sensation in the eyes was a common complaint as well. This was largely due to the heat emanated from the eye-tracking equipment. Such equipment was turned off during the break, so as to cool it down. As for the VE, the avatar occasionally could get stuck between objects and in the slope between the street and the walkway. Note that this problem would not arise near the crosswalks, because the slope does not exist. Eventually, if the participant insisted on moving forward, the avatar would get loose and continue the movement.

Finally, some suggestions were provided by the participants to enhance the experiments. A summary of those responses is provided by the following list:

- Cars and pedestrians with random "skins" (different colours, diversified shapes)
- Pedestrians with different facial expressions
- Add other stimuli, such as audio
- Different mechanics (instead of balancing the arms to move the avatar, use a different source of input such as legs)
- Cars and pedestrians with greater "intelligence", so as to bring closer the gap between VE and reality
- Better collision detection between the elements of the VE
- Greater number of cars
- Correct unrealistic situations such as pedestrian getting run over by cars, and pedestrian gatherings
- The fact that the use of eye-tracking technology is stated before the experiment is executed can influence the participant's behaviour and, inherently, the data collected

The addition of additional stimuli such as audio was one of the most requested items by participants. According to them, sound is a very important aspect of attention when they are crossing a street and walking on the street. Although visual stimulus was quite realistic in the virtual environment, the participants felt the need of listening to cars. One of the participants specifically stated that, when walking on the street, it pays more attention to sound stimuli rather than visual ones.

Participants also suggested a greater number and variety of pedestrians and cars, as similar looking pedestrians and cars were seen as quite unrealistic. Moreover, the participants stated that they could easily cross the road without getting run over, thus requiring little to no precaution. Finally, some participants brought up an aspect pertaining the experimental protocol that, according to the researcher's perspective, requires a deeper reflection. Before any of the experiments are started, it is stated that eye-trackers are going to be used to monitor the participant's gaze throughout the experiment. Nevertheless, some of them said that participants should not know that their eyes are being tracked. Knowing that they are being observed can lead to abnormal behaviour, which can lead to erroneous results. Throughout the experiments, the researcher could not have referred the eye-trackers, however, one may not agree to the fact that their eyes are being monitored. This is specially true if one is using retinal scanning for security purposes, which may not give consent to monitor the eyes. In conclusion, consensual concerns were the main reason for the researcher to state the fact that the participants' eyes were being monitored throughout the experiments.

In conclusion, this section started by giving an overview regarding the various series that were developed. In total, 4 series were generated, mainly due to changes in the questionnaire. The result discussion focused on the 4^{th} series (the last one), which is comprised by 27 participants. A descriptive analysis of the sample was conducted, and a focus was given to the immersion ability of the virtual environment. Collected responses seem to support the idea that the developed virtual environment is immersive, with responsiveness and realism being the main praised aspects. Therefore, the obtained results seem to support the claim that the VE can be seen as a proxy of the real environment. Nevertheless, more robust testing procedures are required so as to ascertain such claim. In addition, participants were asked their opinion regarding the experiment. Positive aspects such as scene fidelity and controls' responsiveness were some of the metioned aspects. As for negative aspects, cybersickness was one of the points put forth by the participants. Suggestions such as the addition of other stimuli such as audio and a greater diversity of elements were provided so as to improve the VE. Note that the negative points highlighted by the participants are of utmost importance, as they can serve as a starting point for future work. Additionally, extensions of the project can be performed based on the provided suggestions.

4.2.2 Attention Model

In the previous section, a critical analysis was conducted on the virtual environment. The objective of such analysis was to assess the extent to which the VE could be considered as a proxy of the real environment. The objective of the current section is to perform an analysis concerning the developed attention model. A focus on the performance of the model will be given in this section. Additionally, a description of the evaluation procedure will be stated, and the obtained results will be presented. Then, a discussion of the results will be provided. The section ends with a description of the possible limitations concerning the adopted evaluation procedure.

As previously, the results taken from the subjects that participated in the 4^{th} series were considered for the analysis of the model. Therefore, 27 subjects comprise the sample used to support the top-down component of the attention model, and to evaluate the performance of the model.

Table 4.4: Information recorded while the participant is immersed in the VE.

Field	Data Type	Description
timestamp	d	the time from the start of the simulation
body position	t(d, d, d)	the position of the avatar in the virtual world
body orientation	t(d, d, d)	the orientation/rotation of the avatar in the virtual world
body linear velocity	t(d, d, d)	the "instantaneous" linear velocity of the body
body angular velocity	t(d, d, d)	the "instantaneous" angular velocity of the body
head position	t(d, d, d)	the head position of the avatar in the virtual world
head orientation	t(d, d, d)	the head orientation of the avatar in the virtual world
head linear velocity	t(d, d, d)	the "instantaneous" linear velocity of the head
head angular velocity	t(d, d, d)	the "instantaneous" angular velocity of the head
gaze coordinates	t(d, d)	the viewport coordinates of the gaze

Data type description:
d - double
t - tuple

For each participant, no less than two sessions were conducted. The first ones were used so as to allow the participant to get used to the mechanics of the VE, and check whether the eye-trackers were working properly; the data collected from the last session is the one that is considered for validation purposes and to support the top-down component of attention. For each session, a log of the activities performed by the participant is recorded, and each entry of the log is composed by the pieces of information presented in Table 4.4. Such information can be broadly divided into 3 categories:

- **metadata** - the *timestamp* is the only field among those present in the table that fits into this category. Its value represents the amount of elapsed seconds from the beginning of the simulation.
- **body information** - information such as the position in the VE and its orientation is recorded.
- **head information** - since the avatar's head can move and rotate independently from its body, a new set of information pertaining the head's state is also recorded into the log.
- **gaze information** - the information collected pertains the coordinates of the user gaze in viewport coordinates. This is the information returned by the eye-tracker pupil detection software.

As for the granularity at which data is collected, a new entry is added to the log every time a frame is rendered; in other words, the number of frames per second dictate the frequency the log is updated.

If an object of the scene is detected to be observed by the participant, the information presented in Table 4.5 is added to the entry. The object ID is a unique number assigned by Unity, and it can

Table 4.5: Additional information recorded when an object is observed by the participant.

Field	Data Type	Description
object ID	i	the object identification number
object name	s	the object name
semantic information	s $\cup \varnothing$	additional information pertaining its current state
object position	t(d, d, d)	the object position in the virtual world
object rotation	t(d, d, d)	the object orientation/rotation in the virtual world
object linear velocity	t(d, d, d)	the "instantaneous" linear velocity of the object
object angular velocity	t(d, d, d)	the "instantaneous" angular velocity of the object
euclidean distance	d	the distance between the object and the user

Data type description:
d - double
i - integer
s - string
t - tuple

be used to distinguish different objects of the same type (for instance, two different trash cans of the same type). Moreover, any information regarding the current state of the object is also included into the logs. For the purposes of the dissertation, such information is classified as "semantic information". In particular, when considering the adopted virtual scenario, the traffic lights include such semantic information. If the user observes the traffic lights, their current state (in this case, what light was light at the time of observation) is added to the entry.

Some notes regarding the collected data should be highlighted. The position of the elements (body, head, observed object) take into account the global coordinate system located at the centre of the virtual world with coordinates $(0, 0, 0)$. Therefore, all coordinates have the same coordinate system. As for the obtained "instantaneous" velocity, it is based on the evolution of values between consecutive log entries. Finally, not all data is used in the context of the dissertation; indeed, the head position and orientation, and the object name and semantic information represent the used information. The processing of the remaining information can be a topic of research for future work.

In order to test the performance of the model, the participants are split into two groups: training and test. 70% of the participants go to the first group, whereas the remaining part (30%) goes to the second one. Having 27 participants, 8 (approximation to the nearest integer) of them are used as test data. A random selection of those 8 participants is performed. The adopted proportion of data splitting (70/30) is common in the machine learning domain when subsampling the data.

The training data is used to fit the mixture model, which, in turn, is queried by the top-down component of attention. To build the model, only entries of the logs with observed objects were considered. For each observed object in all training set, the position of the observer is taken into account. Additionally, the orientation of the participant's head is also taken into account. However, since almost all elements of the virtual environment are at the level of the subject, only the yaw component of the head orientation was considered, so as to allow a simpler approach

to the data modelling procedure. For the model, a maximum number of 40 mixture components was established, and the effective number of components was established using variational inference methods [BKM17] (in fact, the number of components is the same, however the weights of some components is converged to zero). Such number was based on the heuristic proposed by Formann [Dol02, For84], which suggests 5×2^n to be the minimum number of instances, being n the dimensionality of the instances. Since, for each combination (object, semantic information), 3 variables are being considered (head position - x, head position - z, head orientation - y), $5 \times 2^3 = 40$ is the minimum sample size suggested by the adopted heuristic. Moreover, the researcher observed that no object seem to require more than 40 components to properly cluster the data, as most of the components had a component weight close to zero. An implementation of these methods is made available for Python by Scikit-Learn [8] [PVG+11]. A maximum of 200 iterations for the Expectation-Maximisation algorithm was imposed so as to allow convergence. As output, for each observed object, a mixture model is obtained and fed into the attention model. Note that some of the objects might not be considered by the model due to the fact that the minimum number of instances was not met. The adopted clustering method requires a minimum number of instances equal to the number of components. This means that, for each combination (object, semantic information), if the number of instances is lower than 40, such combination is considered to be not relevant for the top-down component of attention. If one considers the conducted experiments, some of the traffic signs directed at cars are the objects that were not considered by the top-down component. This might be explained by the fact that such objects may be not relevant to subjects while they take the role of pedestrians or by the fact that such observations are outliers. Moreover, it was interesting to notice that some states of the traffic light of the cars were not considered, namely the yellow state of the top traffic light. This may highlight the fact that, when the object's state is considered, a greater number of participants may be required so as to make sure that such state is perceived by a sufficient amount of subjects.

The test data from the selected 8 participants is aggregated. Such data is then filtered out in order to contain only entries with observed objects. An average of 2500 entries is obtained for each participant after the filtering procedure. From the test data, $2500 \times 0.3 = 750$ entries are then selected. Note that all participants made the same path, and, based on this observation, the researcher assumed that all data could be treated as if it belongs to a single person that performed the course 8 times, hence the rationale behind the 750 value from above. One could use $20000 \times 0.3 = 6000$ (that is, 30% over the total amount of data), however Unity could not handle such vast amount of information. Some optimisations to the developed work may be needed to apply a greater amount of information.

The same virtual environment is used again but in non interactive mode. The head position and orientation when each entry was obtained are used to obtain a saliency map. After obtaining the map, the objects located in the map are sorted in decreasing order of saliency. At this point, one can establish a mapping between the saliency map and the observed object from the correspondent entry of the log. If the object was detected while the saliency map was obtained, the saliency order

[8] http://scikit-learn.org/stable/modules/mixture.html, as of 20th June, 2018

Table 4.6: Number of observations pertaining the validation phase of the model.

	Number of Observations
Observations with detected objects	417
Observations with observed objects	748
Total number of observations	750
Average number of observed objects	9

for that object is obtained. A **performance metric** for the model can be outlined based on the saliency ranking for the observed object. Let *r* represent such ranking; if the observed object has a ranking of *r*, this means that the saliency of the object and the most salienct observed object are *r* positions apart. This performance metric assumes that the user is always looking towards the most salient region at any moment. This same analysis was conducted for the bottom-up component alone, and for the integration with the top-down modulation. This approach allows to extract some preliminary conclusions regarding the increase of performance when top-down component is incorporated.

As stated in Table 4.6, 750 analysis were conducted, and, out of those, 748 contained observable objects. In other words, 2 scene observations contained no objects to analyse. As the scene is non-deterministic, the following may have happen: while the experiment was conducted with the subjects, objects such as pedestrians might have been detected and these were the only observable objects at that moment; however, when the tests were conducted, for the same position and head orientation, no pedestrians were present. Therefore, this explains the difference between the total number of observations and the number of observations with observed objects.

As for the difference between the number of observations with observed and detected objects, since non-static objects (cars, pedestrians) could be present at the time of the experiment, some observations during the test phase might have not included the observed object. For example, the participant might have observed a car, but, during the testing phase, no cars were observed for the same position and orientation of the participant. Out of 748 observations, 417 contained, in fact, the object observed by the participant at that location. That is, 55.7% of the conducted observations were thus used to assess the performance of the attention model. It is worth noting that an average of 9 object instances were obtained per observation, truncated to the unity.

As aforementioned, due to the non-deterministic nature of the scenario, some objects that were observed during the experiment might not be observed during the validation phase. An analysis pertaining the objects that were observed during the experiment but not during the validation was conducted. Pedestrians and cars were the objects that accounted the most for the number of missed observations; indeed, these are non-static objects, which, as stated above, can lead to the situation where the objects were observed during the experiments but they were not present at the same location during the validation phase. In addition, traffic signs (such as stop sign, and turn right ahead) were also responsible for the number of missed observation, although in a smaller number. Errors from the eye-tracker readings and the resolution of the object identification mapping can be

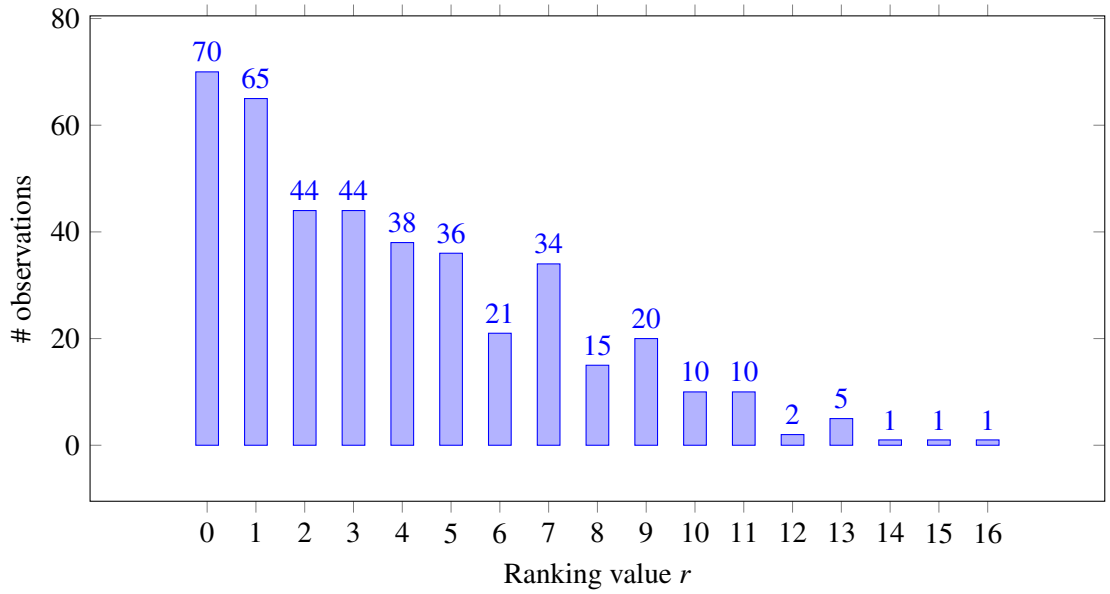


Figure 4.18: The performance values for the attention model.

responsible for such misses. The first problem can result in outliers in the obtained results, which can lead to erroneous results. The problem related with the object mapping is more relevant when the object is of smaller dimension or is located at a greater distance from the observation point; when projected into the viewport, these objects fill a small region of the observable image, which can result into an erroneous colour value in the object identification map due to the small occupied area, thus leading to an incorrect identification of the object.

Considering the aforementioned performance metric, Figure 4.18 illustrates the obtained results. $70/417 \approx 17\%$ of the observations had a correct match between the observed object and the most salient object in the scene. If one analyses the incorrect classifications, when the objects were ranked by decreasing order of saliency, more than 50% of the conducted observations put the saliency ranking distance from the most salient object as far as 3 places from the expected ranking ($(70+65+44+44)/417 \approx 0.53$). It is interesting to observe the number of observations for $r = 0$ and $r = 1$ are somewhat close. This may indicate that a substantial amount of near-misses may have occurred. Moreover, knowing that, in average, 9 objects are observed when perceiving the scene, at least $2/9 \approx 22\%$ [9] of the objects of the scene are misclassified in terms of saliency ranking for $(417-70)/417 \approx 83\%$ of the analysed observations. The observed object was ranked in terms of saliency as far as 16 places from the most salient. These results seem to suggest that the performance of the model is low, with only 17% of the analysed observations being correctly examined, if one considers the adopted performance metric. When focusing on the incorrectly examined observations, the median is located at 3, meaning that half of the conducted observations put the ranking displacement as far as 3 positions. At least 22% of the objects of the scene are misclassified in 83% of the conducted observations, in average.

[9] Rationale behind the numerator value: at least, two objects are misranked - the object that should be the most salient, and the object that is taking the position of the former; the remaining objects may be correctly ranked.

Another comparison was made regarding the difference of performance when the top-down component is removed from the attention model. The same performance metric r was considered, and both alternatives – attention model with bottom-up component, and attention model with bottom-up and top-down components – were analysed for the same scenes. Since the same scenes are used, a comparison of performance between them can be done. An interesting finding was that the model had exactly the same performance whether the top-down component was included into the model or not. In fact, the same results expressed in Figure 4.18 can be used for both situations, for when the bottom-up component of attention is considered alone, and when the top-down is incorporated. This may suggest the following:

- the adopted modulation procedure is not adequate. That is, Equation 3.24 may be lacking information that is required to obtain a better saliency value for the object. The inclusion of other features such as motion and flicker may increase the performance of the model.

- the adopted performance metric is not sufficient to evaluate the impact that the proposed top-down component have in the final result. The current metric considers the ranking distance between the observed object and the most salient object. Other approaches may take into account the amount of time the object was observed by the user, for instance.

- the top-down factor may be acting as a redundant component, adding no significant disturbance into the bottom-up output. Nevertheless, when analysing the ranking of the objects' saliency individually according to both alternatives, a difference in rankings is indeed obtained, but such difference did not impact the ranking of the most salient object. As before, other performance metrics may be more suited to give further conclusions.

Therefore, some conclusions may be drawn. The model seems to present a poor performance, with 17% of the conducted observations with a correct labelling as the most salient object in the scene. Note, however, that the adopted performance metric may not be capable of assessing the performance of the model in a relevant way. In fact, assuming that the user is always looking to the most salient region at any moment may represent a strong assumption, as the user may conduct exploratory activities. As future work, the proposal of new performance metrics can reveal to be an interesting effort to pursuit. In addition, using the same performance metric, no performance differences were obtained when integrating a top-down component based on mixture models. This may indicate that modelling the top-down bias via mixture models is not adequate, or it may also indicate that dimensions considered by the model (head position and orientation) are not fit to model the top-down bias. In particular, the results may admit the fact that the equation responsible for integrating the attention biases from the bottom-up and top-down is incomplete; a linear combination may not encapsulate the attention mechanisms that take place in the human brain. On the other hand, if one assumes a correct approach towards the top-down component and also assumes that the performance metric is adequate, the obtained redundancy of results may suggest that the proposed mixture model may act as a proxy for the attention as a whole. Indeed, if no difference in performance is obtained when the mixture model is added, one may suggest

that the component alone may act as a substitute for the attention model as a whole. This can be useful, as the performance impact of the top-down component is far less when compared to that of the bottom-up component alone. Note, however, that the researcher stresses the fact that further analysis is required so as to make such extrapolation more concrete.

4.3 Limitations

Some limitations identified by the researcher pertaining the adopted experimental methodology and the developed attention model were already presented in the previous section. Nevertheless, the current section provides a more extensive listing of the identified limitations. Such listing is focused on aspects of the conducted experiments, and on the methodologies applied to the evaluation of the attention model.

4.3.1 Experiment

A focus on the limitations pertaining the conducted experiments is done in the current subsection. Motion sickness, concerns regarding the eye-trackers, the presence of outliers and systematic errors, the lack of experimental data, the adopted approach to attention modelling, and concerns regarding the questionnaire are the limitations highlighted by the researcher.

Cybersickness

The obtained feedback from the participants can be useful to identify possible limitations regarding the conducted experiments. The cybersickness was a concern raised by some participants. As aforementioned, some of them felt mild to strong discomfort during the experiment. Obviously, no participant should feel any discomfort; the objective of the experiment is not to induce soreness in the participants. Moreover, if the participants are feeling discomfort, the sense of immersiveness is severely affected, which can also compromise the results to some extent. Some studies regarding the causes of cybersickness have been conducted [Bar04, Kol95, MS92, LaV00], which can be used to support further developments of the work presented in the dissertation in order to minimise motion sickness.

Eye-tracker Transparency

An interesting aspect that was raised by one of the participants is related to the fact that, before participants were asked to sign the consent the form, they were alerted that they would be monitored; in particular, it was stated that the eye-trackers would monitor eye movement throughout the experiment. Knowing that their activity is being monitored can lead to misleading results, since participants may adjust their behaviour into a more expected one. Nevertheless, nothing was said about the dimensions that were being monitored; the participants were only given directions so as to follow the defined course, but nothing was said regarding what objects were being monitored.

This can minimise such impact. Ultimately, a compromise between transparency and result accuracy represents what is at stake. Since users may not agree to their eyes being recorded (due to security reasons, for instance), the researcher favoured transparency.

Outliers, Systematic Errors

The existence of outliers in the collected data is another of the possible limitations. Data collected from the initial and final stages of the experimental session may reflect abnormal behaviour, as subjects are starting to move around in the environment or arriving at the final destination, respectively. Ignoring data collected from these experiment phases is a procedure that can help to mitigate the problem. Systematic and random errors from measurement instruments (eye-tracker, VR tracking equipment) can also impact the analysis of the results. In fact, the presence of jitter in readings taken from the eye-trackers was quite visible during testing. The use of Kalman filters in the project allowed the reduction of such jitter; nevertheless, the total removal was impossible. Moreover, minor shifts of the HMD caused by head movement and facial expressions can cause misalignments in the readings from the eye-trackers. These errors were minimised by limiting the slack of the HMD so as to reduce the movement.

Insufficient Data, Different Variables, and Different Scenarios

Another limitation may concern the amount of collected data, as well as the considered variables to be studies. Although an average of 2500 observation points were collected per person, this number may not be sufficient to capture attention models. Adding new participants may be an interesting approach; in particular, the inclusion of participants with different background and different sociodemographic contexts may reveal interesting findings pertaining attention behaviour. Also, the inclusion of different scenarios may reveal interesting knowledge pertaining attention patterns. For this work, three major areas (crosswalk, crosswalk with traffic lights, roundabouts) were considered. The inclusion of a greater variety of scenarios may also be an interesting approach.

Object as Unit of Attention

The object was regarded as the unit of attention. That is, objects were individually identified in order to trigger some response when the user is looking at a specific object. This was done so as to make the model and its analysis computationally tractable. However, some problems may arise. For instance, there might be objects the users were looking at, but they were made as non detectable. Indeed, objects such as buildings and light poles were not made to be detectable, as, according to the researcher, such objects were seen as background elements or they were barely detectable; in addition, they were considered to be not relevant for pedestrian activities such as following a given path. In order to assess whether the ignored objects were perceived by participants, a question asking for objects they could remember was put in the questionnaire. Most of the listed objects were indeed made as detectable in the VE; however, some participants stated

that they were able to remember to have observed buildings, light poles, and even benches, which were some objects that were not implemented as detectable. One could have made these objects as detectable; however, increasing the number of detectable objects can lead to an increase in noise in the collected data. Nevertheless, the inclusion of such objects may be done in future iterations.

Questionnaire and Respective Analysis

Finally, a critical analysis of the questionnaire must be conducted. This tool may be of utmost importance when capturing information that may be difficult to measure while in the virtual environment, such as the participant's perception of the environment. Initial approaches of the experiment segmented the environment into pre-defined zones (for example, a region encompassing a crosswalk with traffic lights), and the participant's attention would be carefully analysed while inside the zone. At the end of the experiment, questions focusing on the participant's perception in such regions would be presented. The amount of seconds and the amount of times the subject thought to have seen the object was an example of such questions. However, the researcher struggled to develop questions capable of encapsulating the attention behaviour in a meaningful manner. Moreover, the inclusion of the aforementioned questions produced a lengthy questionnaire, which caused some complaints to be pointed out by some participants; as such, those questions had instead their focus on the entire session, instead of focusing on each region. Also, the questions that were used to assess the user experience with the VE were based on a small version of the validated questionnaire. The use of the extended version in future iterations of the work can be very interesting so as to have a solid perception regarding the immersiveness of the VE. With the addition of new questions capable of representing the perception of the user in the environment, new approaches of discussing the results can be used, such as ANOVA analysis. Furthermore, the use of novel techniques of information processing can lead to new findings (for instance, the extraction of behaviour patterns), which can be pursued in future work.

4.3.2 Evaluation of the Model

A focus on the limitations pertaining the evaluation of the attention model is done in the current subsection. The lack of hyperparameter optimisation techniques, the non-deterministic nature of the validation procedure, the adopted data subsampling strategy, and the robustness of the adopted performance metric are the limitations highlighted by the researcher.

Hyperparameter Optimisation

As for the validation of the attention model, some drawbacks can be outlined regarding the adopted methodology. Firstly, no systematic approach was used to determine the number of components for the mixture model used to represent the top-down component of the attention model. Variational inference methods were used so as to estimate such number; however, these methods require an upper bound to be provided. When establishing such value, the minimum number of points to be clustered is also established. With this idea in mind, the minimum number of points a cluster

must be established. Formann [Dol02, For84] suggested that number to be, at least, 5×2^n instances, where n represents the number of variables that characterises the instances. Nevertheless, this is just a "rule-of-thumb"; in fact, no proven methodology was found pertaining the minimum number of instances to perform clustering. Moreover, the adjustment of the hyperparameters of the clustering algorithm is another limitation of the proposed solution, as no hyperparameter optimisation techniques were used. Nevertheless, according to the researcher, the values used for the hyperparameters seem to suffice for an adequate solution. The maximum number of iterations for the Expectation-Maximisation algorithm allowed a solution to be converged, and the adopted covariance matrix type [10] enabled components to be more independent from one another. The addition of optimisation techniques can be of utmost interest as future work.

Non-deterministic Nature of the Validation, and Data Subsampling

Another limitation pertaining the evaluation of the model is related to its non-deterministic nature. Such nature is caused by two key aspects: getting a random division of the data into training and test sets, and the non-deterministic nature of the scene. The first problem may be responsible for the introduction of biases into the obtained results, depending on how such division is made. As stated above, data was partitioned using a 70/30 proportion schema with a hold-out approach. The adopted data splitting procedure can cause an impact into the final results of the model and its evaluation. k-fold cross-validation techniques may be used in future iteration of the work as to minimise bias introduced by a simple random division of data.

On the other hand, the non-deterministic nature of the scene can be responsible for the fact that a large portion of non-static objects (cars and pedestrians) was not detected during the validation phase. Also, the fact that dynamic objects evolve in different ways across participants, some subjects may perceive the objects, while others may not. A solution to this problem would be to store the state of every dynamic object each time a new entry was added to the log, which translates into considerable amounts of information to be saved. Also, different participants interact in different ways with the environment, which leads again to a non deterministic observation of the environment (some participants can quickly arrive at the end of the path, while others can take a substantial amount of time; which means that the environment is not deterministic even if the path of dynamic objects is fixed and predefined). The definition of contained scenarios can represent a reasonable approach to the problem. The addition of new validation methods capable of coping with dynamic scenes can represent an interesting path for future research.

Non Robust Performance Metric

As aforementioned, the adopted performance metric may lack robustness. The performance metric takes only into account the difference between the saliency of the observed object by the participant and the most salient object at the time the observation is performed by the validation process; then, a list of visible objects is obtained and ordered according to their saliency and the the ranking

[10] "full" type, in which each component has its own general covariance matrix

position of the object observed by the participant is obtained. This performance metric assumes that the observations collected throughout the experiments as ground truth; also, it assumes that the user is constantly looking at the most salient object while immersed in the VE. This may represent a strong assumption, as the user may conduct exploratory activities while he is in the VE; in addition, the data collected from the eye-trackers may contain errors, which ultimately can lead to erroneous performance values. This can be a simple metric, and more robust validation techniques may reveal to be more fruitful, as more strong conclusions may be inferred; nevertheless, the adopted performance metric allow the reader to infer some preliminary conclusions regarding the performance of the model.

4.4 Summary

In this chapter, a description of the test bed used to collect data to support the attention model and assess its performance was provided. Additionally, the experimental protocol adopted in the conducted experiments was detailed. As for the evaluation of the attention model, a performance metric was proposed and the obtained results were discussed. Finally, a listing of the limitations regarding the experiments and the evaluation process was presented.

The description of the test bed started with a technical overview of the used eye-trackers. The integration of the eye-trackers with game engines such as Unity was detailed; a plugin is made available by Pupil Labs containing some demo scenes and the implementation required to communicate with the equipment. Additionally, the software bundled with the equipment was inspected; the pupil detection algorithm was examined and the respective calibration parameters were explored. Kalman filters were implemented and used so as to obtain a better estimation of the real values for the gaze coordinates, as the collected data contained jitter. The calibration procedure responsible for fitting a transform function capable of mapping the eyes' position into virtual world coordinates was explained. Finally, the raycast mechanism was described so to illustrate its pertinence to detect the observed object. As for the virtual scenario, an overview was provided; the outline of regions and the path to be followed by participants was given, and the mechanics to be used by participants to interact with the avatar were introduced. Three regions were outlined and a path traversing such regions was proposed; the *HTC Vive* controllers are used to interact with the avatar located in the VE. Regarding the experimental protocol, the steps that were followed throughout the experiments were introduced; furthermore, the description of the questionnaire presented to participants at the end of the experiments was presented.

The discussion of the obtained results was divided into two components. The first one focused on the experiment and the obtained sample. A description of the sample was provided, with 27 subjects with ages between 17 and 35 having participated in the experiments. The responses to the questionnaire were also inspected. Results seem to suggest that the developed VE presents a high degree of realism and responsiveness. Participants were also aspect for positive and negative aspects regarding the experiment. The scene fidelity and cybersickness were some of the highlighted elements, respectively. Also, suggestions for improvement were given by participants as well, in

which the inclusion of cars and pedestrians with different apparel was one of those suggestions. Nevertheless, the researcher stresses out the fact that further research is required to extract more robust results pertaining the user experience of the VE. The other component of the discussion approached the performance of the attention model. A description of the variables provided by the data collected was provided, and the process used to evaluate the model was outlined. A performance metric based on the ranking of the object observed by the participant was introduced. Results seem to suggest a poor performance of the model, with approximately 17% of the observations with detected objects being correctly classified in terms of being the most salient object. As for the top-down modulation, the proposed component and respective integration seem to have no impact into the final result, thus being a redundant element of the model; such redundancy may hint the fact that the modelling approach or the integration approach were not adequate, or the attention model may be replaced by a mixture model. However, a different testing approach may be useful as to dissipate eventual doubts. A listing of the encountered limitations concerning the conducted experiment and the used methodology for evaluating the attention model was presented.

Chapter 5

Conclusions

The current chapter provides the reader with an overview of the developed work and draws conclusions highlighting the main contributions achieved. It also proposes future work directions opening up a number of possible ways to further develop the method herein described. The work concludes with some final remarks.

5.1 Main Contributions

An outline of the contributions from the work developed by the researcher are presented, which can be framed in four main points.

Computational attention model

The proposal of a computational model of attention represents the main contribution of the dissertation. Following up the suggestion proposed in the literature, the attention model is divided into two components: *bottom-up*, responsible for modelling attention based on the properties of the stimulus, and *top-down*, in which cognitive processes are emulated so as to modulate the response of the bottom-up component. The bottom-up component is based on the model proposed by Itti *et al.*, which established the theoretical foundations for saliency-based computational models of attention. As for the top-down factor, a Gaussian mixture model is used so as to provide spatial contextual information into the final saliency result. The proposal of an integration approach of both components of attention represents another major contribution of the developed work. A parallel implementation in C# of the bottom-up attention model is also a contributions of the work, which can be used in game engines that support the aforementioned programming language (for instance, Unity3D). As for the top-down component, a program developed in Python, which uses the `Scikit-learn` library, performs clustering on data collected from experiments so as to output the mixture models.

Test bed with an immersive VE

A test bed was implemented in order to conduct the experiments required to populate the attention model and to validate the performance of such model. A virtual environment was defined in Unity3D, in which a 3D model based on a Spanish city was used. The objective of the environment was to create an immersive VE, in which users were "placed" inside the scenario. Therefore, VR solutions such as *HTC Vive* system were used. The development of an avatar was imperative, in order to implement the projection of the user in such a virtual environment.

Moreover, the integration of the eye-tracker technology within the developed framework was crucial, so as to obtain information regarding the user's gaze while immersed in the virtual environment. The ability to analyse to which object of the environment the user was paying more attention was essential for the project.

Furthermore, future research projects may use the aforementioned facilities, since the test bed can also be used for purposes other than attention modelling. In fact, since the test bed is based on an urban environment, experiments regarding user's behaviour in an urban settings can be conducted.

Experimental protocol

In order to conduct the experiments in a systematic fashion, an experimental protocol was outlined. The protocol stated a set of steps to be followed throughout the experiment. Before the execution of the experiment, an explanation of the developed work and respective objectives are provided to the participant. If he agrees to participate, he is asked to sign the consent form. At least, the scenario is executed twice, so as to give the participant an adaptation period to the VE; the data collected from the last execution is used for further processing. Moreover, a questionnaire was also defined. Such questionnaire was answered by the participants just after the session in the VE. The purpose was to retrieve sociodemographic information pertaining the selected sample of participants. Also, information pertaining their behaviour as pedestrians in the real environment was queried as well. Finally, questions regarding how subjects perceived the environment and their user experience in the VE were analysed.

Validation Process

In order to assess the user experience of the VE and the performance of the attention model, the obtained results and respective discussion were introduced. The responses given to the questionnaire were used to assess the user experience of the VE. A high degree of fidelity and responsiveness may suggest a notable level of immersion, thus allowing the virtual environment to act as a proxy of the real environment. Nevertheless, motion sickness was one of the negative aspects highlighted by the participants. As for the attention model, a performance metric based on the saliency ranking of the observed object was proposed and used. The results seem to suggest a poor performance of the model. Moreover, the integration of a mixture model-based top-down component into the

attention model seems not to increase the performance of the model, as the same results are obtained. Such a trend may hint the fact that the adopted strategy for modulating the bottom-up response is either not appropriate, or the use of mixture models is not adequate to model the top-down component, or the adopted validation strategy is not as robust. On the other hand, such a redundancy may indicate the fact that a mixture model can act as an approximation of the attention model as a whole. Therefore, further research needs to be conducted so as solid conclusions can be further realised.

5.2 Future Work

Some of the limitations pertaining the model and the experimental setup were identified throughout the project and discussed in their proper sections in this report. As such, those limitations can be considered as a starting point for further developments aiming at improved results and enhanced performances. Some challenging problems are also suggested as future work. Note that the proposed listing is only a recommendation and therefore is not limited to the topics herein proposed. Other lines for future research can also be outlined. The developed work can be considered as the grounds upon which future contributions can be further devised, implemented, tested and validated.

Short-term Future Work

- **GPU Implementation** – the current implementation is CPU-based, making use of abstractions such as data and task parallelism so as to increase the performance of the model. Nevertheless, the implementation takes, approximately, 6 seconds to compute the final saliency map; this makes the application in real-time systems impractical. Due to the nature of the operations performed by the attention model (array multiplications and additions) and due to the high parallel nature of the model (in particular, the bottom-up component), GPU-accelerated computing can provide a boost in terms of performance.

- **Model Refinement and Component Integration** – from the obtained results, the proposed modelling approach seem to be inadequate to represent the top-down factors. Additionally, the adopted procedure responsible for modulating the bottom-up response may not suffice to express the attention response. The inclusion of new features such as motion, depth, and flicker into the attention model may help to achieve better results. The proposal of new integration approaches may yield superior outcomes.

- **VE Refinement** – the responses provided by the participants seem to suggest that the used VE possesses a high degree of realism and responsiveness. Nonetheless, some observations were made by the participants regarding their experience with the environment, which can be used as starting point for further refinements. Cybersickness represents a problem that needs to be tackled in future iterations of the VE. Also, adding diversity into the elements of the environment (cars with different shapes and colours, and a greater variety of pedestrians)

can increase the realism component. In order to devise a more immersive environment, other stimuli such as sound are planned to be added. The reader is referred to Section 4.2.1 for a more complete listing of the comments highlighted by the participants.

- **Validation Methodologies** – in the dissertation, a validation methodology based on the ranking of the observed object was proposed. However, as the attention is a complex mechanism to be assessed, such evaluation may not be adequate enough to evaluate the model. The design of new validation processes may represent an interesting course of investigation.

- **Mixture Model Refinements** – as stated previously, no performance improvement was obtained when the proposed top-down component was included in the model. In fact, the same results were obtained, which may suggest the fact that this component is acting as a redundant element for the final result and, therefore, the attention model as a whole may be approximated by a mixture model. Further validation regarding this observation may reveal a pertinent research path, as such approximation may be relevant to speed up operations.

Long-term Future Work

- **Integration with Cognitive Models** – agent architectures such as BDI (Beliefs, Desires, Intentions) try to mimic the human reasoning process in a computationally tractable way. However, such architectures do not focus on how the input perceptions are processed by the human brain. The application of attention models to the received input before it is processed by the architecture would bring closer the gap between the architecture and reality. More, the ability of the cognitive architecture to provide feedback into the attention model, thus potentially changing the the response to input of the latter, would make the architecture even more realistic. Therefore, the integration of attention models into these architectures would represent an interesting research agenda in the field of multi-agent systems.

- **Behaviour Analysis** – the information collected from the conducted experiments may be used in applications that go beyond attention modelling. Indeed, the extraction of behaviour patterns may be performed in the information collected from the questionnaires, for instance. Furthermore, the data collected from the interactions of the participants with the VE may be of utmost interest to analyse user behaviour in a controlled and secure way. This is especially the case for critical application domains in which certain activities are hard to be tested and evaluated in real-life settings, such as road safety and transportation systems.

- **Real-time computing** – in robotics, real-time computing is paramount. However, systems currently developed have usually limited computing power. Therefore, the ability to filter out the input these systems receive from the surrounding environment represents a promising research area in Robotics. The development of a robotic system in which attention models are used may provide a substantial contribution to this research field.

5.3 Final Remarks

The work herein detailed represents an important contribution to the SIMUSAFE project, an European-funded endeavour whose purpose is to improve safety on roads based on concepts such as simulation games, virtual environments, behavioural modelling, and multi-agent systems. The understanding of the human attention mechanisms is paramount when one wants to assess risk perception and decision-making of road users. Additionally, it is the intention of the researcher to submit a paper to a journal reporting on the main contributions of the work developed throughout the dissertation. As for the pursuit of new findings related to computational attention models, the enrolment of the researcher in a doctoral program is a possibility.

References

[ABKS99] Mihael Ankerst, Markus M. Breunig, Hans-Peter Kriegel, and Jörg Sander. OPTICS. In *Proceedings of the 1999 ACM SIGMOD international conference on Management of data - SIGMOD '99*, pages 49–60, New York, New York, USA, 1999. ACM Press.

[AHB+17] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering. *CoRR*, 2017.

[AL10] Tamar Avraham and Michael Lindenbaum. Esaliency (Extended Saliency): Meaningful Attention Using Stochastic Image Modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(4):693–708, apr 2010.

[AV07] David Arthur and Sergei Vassilvitskii. k-means++: The Advantages of Careful Seeding. In *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1027–1035, Philadelphia, PA, USA, 2007. Society for Industrial and Applied Mathematics.

[AVLH98] Lourdes Anllo-Vento, Steven J. Luck, and Steven A. Hillyard. Spatio-temporal dynamics of attention to color: Evidence from human electrophysiology. *Human Brain Mapping*, 6(4):216–238, 1998.

[BAA11] Ali Borji, Majid N. Ahmadabadi, and Babak N. Araabi. Cost-sensitive learning of top-down modulation for attentional control. *Machine Vision and Applications*, 22(1):61–76, jan 2011.

[Bar04] Judy Barrett. Side Effects of Virtual Environments: A Review of the Literature. Technical report, DSTO Information Sciences Laboratory, Australia, 2004.

[BCC99] N. I. Badler, D. M. Chi, and S. Chopra. Virtual human animation based on movement observation and cognitive behavior models. In *Proceedings Computer Animation 1999*, pages 128–137. IEEE Comput. Soc, 1999.

[Ber06] Pavel Berkhin. A Survey of Clustering Data Mining Techniques. In Jacob Kogan, Charles Nicholas, and Marc Teboulle, editors, *Grouping Multidimensional Data: Recent Advances in Clustering*, pages 25–71. Springer-Verlag, Berlin/Heidelberg, 2006.

[BI13] Ali Borji and Laurent Itti. State-of-the-art in visual attention modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(1):185–207, 2013.

[BKM17] David M. Blei, Alp Kucukelbir, and Jon D. McAuliffe. Variational Inference: A Review for Statisticians. *Journal of the American Statistical Association*, 112(518):859–877, apr 2017.

[BT09] N. D. B. Bruce and J. K. Tsotsos. Saliency, attention, and visual search: An information theoretic approach. *Journal of Vision*, 9(3):1–24, mar 2009.

[CB01] Sonu Chopra Khullar and Norman I. Badler. Where to Look? Automating Attending Behaviors of Virtual Human Characters. *Autonomous Agents and Multi-Agent Systems*, 4(1):9–23, 2001.

[CJ98] Marvin M. Chun and Yuhong Jiang. Contextual Cueing: Implicit Learning and Memory of Visual Context Guides Spatial Attention. *Cognitive Psychology*, 36(1):28–71, jun 1998.

[Com92] Pierre Comon. Independent component analysis. *Higher Order Statistics*, pages 29–38, 1992.

[CPGV97] Charles E. Connor, Dean C. Preddie, Jack L. Gallant, and David C. Van Essen. Spatial Attention Effects in Macaque Area V4. *The Journal of Neuroscience*, 17(9):3201–3214, may 1997.

[DD95] Robert Desimone and John Duncan. Neural Mechanisms of Selective Visual Attention. *Annual Review of Neuroscience*, 18(1):193–222, mar 1995.

[Dol02] Sara Dolnicar. A Review of Unquestioned Standards in Using Cluster Analysis for Data-Driven Market Segmentation. In *CD Conference Proceedings of the Australian and New Zealand Marketing Academy Conference 2002*, pages 2–4, Melbourne, 2002. Deakin University.

[DOS17] Ana De Abreu, Cagri Ozcinar, and Aljosa Smolic. Look around you: Saliency maps for omnidirectional images in VR applications. In *2017 Ninth International Conference on Quality of Multimedia Experience (QoMEX)*, pages 1–6. IEEE, may 2017.

[DSD+17] Shuchisnigdha Deb, Lesley Strawderman, Janice DuBien, Brian Smith, Daniel W. Carruth, and Teena M. Garrison. Evaluating pedestrian behavior at crosswalks: Validation of a pedestrian behavior questionnaire for the U.S. population. *Accident Analysis & Prevention*, 106:191–201, sep 2017.

[EE74] B. A. Eriksen and C. W. Eriksen. Spatial attention effects in macaque area v4. *Perception and Psychophysics*, 16(1):143–149, 1974.

[EKSX96] Martin Ester, Hans-Peter Kriegel, Jorg Sander, and Xiaowei Xu. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In *KDD-96 Proceedings*, 1996.

[EZW97] Stephen Engel, Xuemei Zhang, and Brian Wandell. Colour tuning in human visual cortex measured with functional magnetic resonance imaging. *Nature*, 388(6637):68–71, jul 1997.

[Far12] Ramsey Faragher. Understanding the Basis of the Kalman Filter Via a Simple and Intuitive Derivation [Lecture Notes]. *IEEE Signal Processing Magazine*, 29(5):128–132, sep 2012.

REFERENCES

[FD07] Brendan J. Frey and Delbert Dueck. Clustering by Passing Messages Between Data Points. *Science*, 315(5814):972–976, feb 2007.

[For84] Anton K. Formann. *Die Latent-Class-Analyse : Einfuhrung in Theorie und Anwendung*. Beltz, Weinheim, Germany, 1984.

[GA16] Brian F. Goldiez and Anastasia Angelopoulou. Serious Games - Creating an Ecosystem for Success. In *8th International Conference on Games and Virtual Worlds for Serious Applications (VS-GAMES)*, pages 1–7. IEEE, sep 2016.

[GHV09] Dashan Gao, Sunhyoung Han, and Nuno Vasconcelos. Discriminant saliency, the detection of suspicious coincidences, and applications to visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(6):989–1005, 2009.

[Gol09] E. Bruce Goldstein. *Sensation and Perception*. Cengage Learning, 8 edition, 2009.

[GRP+14] João S. V. Gonçalves, Rosaldo J. F. Rossetti, João Tiago Pinheiro Neto Jacob, Joel Gonçalves, C. Olaverri-Monreal, António Coelho, and Rui Rodrigues. Testing Advanced Driver Assistance Systems with a serious-game-based human factors analysis suite. In *IEEE Intelligent Vehicles Symposium Proceedings*, number 4, pages 13–18. IEEE, jun 2014.

[GV09] Dashan Gao and Nuno Vasconcelos. Decision-Theoretic Saliency: Computational Principles, Biological Plausibility, and Implications for Neurophysiology and Psychophysics. *Neural Computation*, 21(1):239–271, jan 2009.

[Har75] John A. Hartigan. *Clustering Algorithms*. John Wiley & Sons, New York, New York, USA, 1975.

[HB05] Mary Hayhoe and Dana Ballard. Eye movements in natural behavior, 2005.

[HSD+13] Aleshia T. Hayes, Carrie L. Straub, Lisa A. Dieker, Charlie E. Hughes, and Michael C. Hynes. Ludic Learning: Exploration of TLE TeachLivE(TM) and Effective Teacher Training. *International Journal of Gaming and Computer-Mediated Simulations*, 5(2):20–33, apr 2013.

[Hun06] Hans-Werner Hunziker. *Im Auge des Lesers: foveale und periphere Wahrnehmung - vom Buchstabieren zur Lesefreude*. Transmedia Stäubli Verlag, Zürich, 2006.

[HW79] J. A. Hartigan and M. A. Wong. Algorithm AS 136: A K-Means Clustering Algorithm. *Journal of the Royal Statistical Society*, 28(1):100–108, 1979.

[HZ07] Xiaodi Hou and Liqing Zhang. Saliency Detection: A Spectral Residual Approach. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, jun 2007.

[IB15a] Laurent Itti and Ali Borji. Computational models: Bottom-up and top-down aspects. *CoRR*, oct 2015.

[IB15b] Laurent Itti and Ali Borji. Computational models of attention. *CoRR*, oct 2015.

[IDP04] Laurent Itti, Nitin Dhavale, and Frederic Pighin. Realistic avatar eye and head animation using a neurobiological model of visual attention. In Bruno Bosacchi, David B. Fogel, and James C. Bezdek, editors, *Proceedings of SPIE 48th Annual International Symposium on Optical Science and Technology*, page 64, jan 2004.

[IK00] Laurent Itti and Christof Koch. A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research*, 40(10-12):1489–1506, jun 2000.

[IK01a] Laurent Itti and Christof Koch. Computational modelling of visual attention. *Nature Reviews Neuroscience*, 2(3):194–203, mar 2001.

[IK01b] Laurent Itti and Christof Koch. Feature combination strategies for saliency-based visual attention systems. *Journal of Electronic Imaging*, 10(1):161, jan 2001.

[IKN98] Laurent Itti, Christof Koch, and Ernst Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(11):1254–1259, 1998.

[Itt00] Laurent Itti. *Models of Bottom-Up and Top-Down Visual Attention*. Phd thesis, California Institute of Technology, 2000.

[Itt04] Laurent Itti. Automatic Foveation for Video Compression Using a Neurobiological Model of Visual Attention. *IEEE Transactions on Image Processing*, 13(10):1304–1318, oct 2004.

[Itt07] Laurent Itti. Visual salience. *Scholarpedia*, 2(9):3327, 2007.

[Jam90] William James. *The principles of psychology*. Holt, New York, New York, USA, 1890.

[Joh14] Saint John Walker. Big Data: A Revolution That Will Transform How We Live, Work, and Think. *International Journal of Advertising*, 33(1):181–183, jan 2014.

[Kal60] R. E. Kalman. A New Approach to Linear Filtering and Prediction Problems. *Journal of Basic Engineering*, 82(1):35–45, 1960.

[KC14] Fumi Katsuki and Christos Constantinidis. Bottom-up and top-down attention: Different processes and overlapping neural systems. *The Neuroscientist*, 20(5):509–521, oct 2014.

[Kol95] Eugenia M. Kolasinski. Simulator Sickness in Virtual Environments. Technical report, U.S. Army Research Institute, 1995.

[KOS11] Elena Kokkinara, Oyewole Oyekoya, and Anthony Steed. Modelling selective visual attention for autonomous virtual characters. *Computer Animation and Virtual Worlds*, 22(4):361–369, jul 2011.

[KPB14] Moritz Kassner, William Patera, and Andreas Bulling. Pupil: An Open Source Platform for Pervasive Eye Tracking and Mobile Gaze-based Interaction. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing Adjunct Publication - UbiComp '14 Adjunct*, pages 1151–1160, New York, New York, USA, apr 2014. ACM Press.

[KR87] Leonard Kaufman and Peter J. Rousseeuw. Clustering by Means of Medoids. In *Statistical Data Analysis Based on the L1-Norm and Related Methods. First International Conference*, pages 405–416, Amsterdam, Netherlands, 1987. North Holland / Elsevier.

[KTM+16] Sumie Kurita, Yuichi Takei, Yohko Maki, Suguru Hattori, Toru Uehara, Masato Fukuda, and Masahiko Mikuni. Magnetoencephalography study of the effect of attention modulation on somatosensory processing in patients with major depressive disorder. *Psychiatry and Clinical Neurosciences*, 70(2):116–125, feb 2016.

[KU87] Christof Koch and Shimon Ullman. Shifts in Selective Visual Attention: Towards the Underlying Neural Circuitry. In *Matters of Intelligence*, pages 115–141. Springer Netherlands, Dordrecht, 1987.

[LAG+89] G W Leibniz, Roger Ariew, Daniel Garber, Richard Arthur, David Blumenfeld, Stuart Brown, Daniel Cook, Alan Gabbey, Nicholas Jolley, Harlan Miller, and M A Stewart. *Philosophical Essays*. Hackett Publishing Company, Indianapolis, Indiana, 1989.

[LaV00] Joseph J. LaViola. A discussion of cybersickness in virtual environments. *ACM SIGCHI Bulletin*, 32(1):47–56, jan 2000.

[LCC13] Ning-Han Liu, Cheng-Yu Chiang, and Hsuan-Chin Chu. Recognizing the Degree of Human Attention Using EEG Signals from Mobile Sensors. *Sensors*, 13(8):10273–10286, aug 2013.

[LG14] Jia Li and Wen Gao. *Visual Saliency Computation*. Lecture Notes in Computer Science. Springer International Publishing, Cham, 2014.

[LJ13] Chang-Wook Lim and Hyung-Won Jung. A study on the military Serious Game. In *Advanced Science and Technology Letters*, pages 73–77. Science & Engineering Research Support soCiety, dec 2013.

[Lur73] Aleksandr R. Luria. *The Working Brain: An Introduction To Neuropsychology*. Basic Books, New York, New York, USA, 1973.

[MAPK14] Konstantinos Mykoniatis, Anastasia Angelopoulou, Michael D. Proctor, and Waldemar Karwowski. Virtual Humans for Interpersonal and Communication Skills' Training in Crime Investigations. In R. Shumaker and S. Lackey, editors, *Virtual, Augmented and Mixed Reality. Designing and Developing Virtual and Augmented Environments. 6th International Conference, VAMR 2014, Held as Part of HCI International 2014. Proceedings: LNCS 8525*, pages 282–292, Cham, Switzerland, 2014. Springer International Publishing.

[Mat] George Mather. The Visual Cortex.

[MB12] Andrew McAfee and Erik Brynjolfsson. Big data: the management revolution. *Harvard business review*, 90(10):60–6, 68, 128, oct 2012.

[MC06] David Michael and Sande Chen. *Serious Games: Games That Educate, Train, and Inform*. Thomson Course Technology, Boston, MA, 2006.

[MR05] Oded Mainon and Lior Rokach. *The Data Mining and Knowledge Discovery Handbook*. Springer-Verlag, Berlin, Heidelberg, 2005.

[MS92] Michael E. McCauley and Thomas J. Sharkey. Cybersickness: Perception of Self-Motion in Virtual Environments. *Presence: Teleoperators and Virtual Environments*, 1(3):311–318, jan 1992.

[ODK99] K. M. O'Craven, P. E. Downing, and N. Kanwisher. fMRI evidence for objects as the units of attentional selection. *Nature*, 401(6753):584–7, oct 1999.

[Oli05] Aude Oliva. Gist of the scene. *Neurobiology of Attention*, pages 251–256, 2005.

[ORK+97] K M O'Craven, B R Rosen, K K Kwong, A. Treisman, and R L Savoy. Voluntary attention modulates fMRI activity in human MT-MST. *Neuron*, 18(4):591–8, apr 1997.

[PI07] Robert J. Peters and Laurent Itti. Beyond bottom-up: Incorporating task-dependent influences into a computational model of spatial attention. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, jun 2007.

[PM00] Dan Pelleg and Andrew W. Moore. X-means: Extending K-means with Efficient Estimation of the Number of Clusters. In *Proceedings of the Seventeenth International Conference on Machine Learning*, pages 727–734, San Francisco, CA, USA, 2000. Morgan Kaufmann Publishers Inc.

[PMKA07] Yannis Paloyelis, Mitul A Mehta, Jonna Kuntsi, and Philip Asherson. Functional MRI in ADHD: a systematic literature review. *Expert Review of Neurotherapeutics*, 7(10):1337–1356, oct 2007.

[PO01] Christopher Peters and Carol O'Sullivan. A Memory Model for Autonomous Virtual Humans, 2001.

[PVG+11] Fabian Pedregosa, Gael Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[Ren00] R. a. Rensink. Scene Perception. *Encyclopedia of Psychology*, 7:151–155, 2000.

[RSFL15] Antonio J. Rodríguez-Sánchez, Mazyar Fallah, and Aleš Leonardis. Editorial: Hierarchical Object Representations in the Visual Cortex and Computer Vision. *Frontiers in Computational Neuroscience*, 9, nov 2015.

[ŚBD12] Lech Świrski, Andreas Bulling, and Neil Dodgson. Robust real-time pupil tracking in highly off-axis images. In *Proceedings of the Symposium on Eye Tracking Research and Applications - ETRA '12*, pages 173–176, New York, New York, USA, 2012. ACM Press.

[SEKX98] Jorg Sander, Martin Ester, Hans Peter Kriegel, and Xiaowei Xu. Density-based clustering in spatial databases: The algorithm GDBSCAN and its applications. *Data Mining and Knowledge Discovery*, pages 169–194, 1998.

[Sha48] C E Shannon. A Mathematical Theory of Communication. *Bell System Technical Journal*, 27(3):379–423, jul 1948.

[Sin14] Sam Sinai. *A Study in Human Attention to Guide Computational Action Recognition*. Master's dissertation, Massachusetts Institute of Technology, 2014.

[SKS09] Sungkil Lee, G.J. Kim, and Seungmoon Choi. Real-Time Tracking of Visually Attended Objects in Virtual Environments and Its Application to LOD. *IEEE Transactions on Visualization and Computer Graphics*, 15(1):6–19, jan 2009.

[SM89] McKay Moore Sohlberg and Catherine A. Mateer. *Introduction to cognitive rehabilitation: Theory and practice*. Guilford Press, New York, New York, USA, 1989.

[Sol88] Robert L. Solso. *Cognitive psychology*. Allyn and Bacon, Boston, 2nd edition, 1988.

[SORJM15] Jonathan Stevens, Eric Ortiz, Lauren Reinerman-Jones, and Douglas Maxwell. Approach to Examine Efficacy of Game-Based and Virtual Simulation Training. In *Proceedings of the Conference on Summer Computer Simulation*, pages 1–7, San Diego, CA, USA, 2015. Society for Computer Simulation International.

[SSM08] R.J. Sternberg, K. Sternberg, and J. Mio. *Cognitive Psychology*. Cengage Learning, Belmont, CA, 2008.

[Sze10] Richard Szeliski. *Computer Vision: Algorithms and Applications*. Springer-Verlag, Heidelberg, Berlin, 2010.

[TCK+95] John K. Tsotsos, Scan M. Culhane, Winky Yan Kei Wai, Yuzhong Lai, Neal Davis, and Fernando Nuflo. Modeling visual attention via selective tuning. *Artificial Intelligence*, 78(1-2):507–545, oct 1995.

[TG80] Anne M. Treisman and Garry Gelade. A feature-integration theory of attention. *Cognitive Psychology*, 12(1):97–136, jan 1980.

[TIR05] John K. Tsotsos, Laurent Itti, and Geraint Rees. A Brief and Selective History of Attention. In *Neurobiology of Attention*, pages xxiii–xxxii. Elsevier, 1 edition, 2005.

[Tre03] Stefan Treue. Visual attention: the where, what, how and why of saliency. *Current Opinion in Neurobiology*, 13(4):428–432, aug 2003.

[TTCLER16] Katy Tcha-Tokey, Olivier Christmann, Emilie Loup-Escande, and Simon Richir. Proposition and Validation of a Questionnaire to Measure the User Experience in Immersive Virtual Environments. *The International Journal of Virtual Reality*, 16(1):33–48, 2016.

[TWK+10] Benjamin W. Tatler, Nicholas J. Wade, Hoi Kwan, John M. Findlay, and Boris M. Velichkovsky. Yarbus, Eye Movements, and Vision. *i-Perception*, 1(1):7–27, apr 2010.

[Wan95] Brian A. Wandell. The Photoreceptor Mosaic. In *Foundations of Vision*, chapter 3. Sinauer Associates, Sunderland, MA, US, 1995.

[WB04] Jennifer L. Wilkinson-Berka. Diabetes and retinal vascular disorders: role of the renin–angiotensin system. *Expert Reviews in Molecular Medicine*, 6(15), jul 2004.

[WCJ11] Jinjun Wang, Jian Cheng, and Shuqiang Jiang. *Computer Vision for Multimedia Applications*. IGI Global, 2011.

[WDGK16] Ryan Wang, Samuel DeMaria, Andrew Goldberg, and Daniel Katz. A Systematic Review of Serious Games in Training Health Care Professionals. *Simulation in Healthcare: The Journal of the Society for Simulation in Healthcare*, 11(1):41–51, feb 2016.

[Wol00] Jemery M. Wolfe. Visual Attention. In K.K De Valois, editor, *Encyclopedia of Neuroscience*, pages 335–386. Academic Press, San Diego, CA, 2nd edition, 2000.

[Won15] Ka-Chun Wong. A Short Survey on Data Clustering Algorithms. In *2015 Second International Conference on Soft Computing and Machine Intelligence (ISCMI)*, pages 64–68, Los Alamitos, CA, USA, nov 2015. IEEE Computer Society.

[WP06] Michel Wedel and Rik Pieters. Eye Tracking for Visual Marketing. *Foundations and Trends® in Marketing*, 1(4):231–320, 2006.

[WP08] Michel Wedel and Rik Pieters. *Visual Marketing: From Attention to Action*. Lawrence Erlbaum Associates, New York, NY, USA, 2008.

[Wun93] Wilhelm Max Wundt. *Grundzuge de physiologischen Psychologie*. W. Engelman, Leipzig, 4th edition, 1893.

[XBK+15] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard Zemel, and Yoshua Bengio. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. *CoRR*, feb 2015.

[XW05] Rui Xu and D. WunschII. Survey of Clustering Algorithms. *IEEE Transactions on Neural Networks*, 16(3):645–678, may 2005.

[Yar67] Alfred L. Yarbus. Eye Movements During Perception of Complex Objects. In *Eye Movements and Vision*, pages 171–211. Springer US, Boston, MA, 1967.

[YJW+16] Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. Image Captioning with Semantic Attention. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4651–4659. IEEE, jun 2016.

[ZH07] Zhiwen Yu and Hau-San Wong. A Rule Based Technique for Extraction of Visual Attention Regions Based on Real-Time Clustering. *IEEE Transactions on Multimedia*, 9(4):766–784, jun 2007.

Appendix A

Experiment Appendices

A.1 Consent Declaration

The following document represents the form all participants signed should they agreed to participate in the experiment under the stated conditions. It requires a signature of the participant, and the date of participation is also recorded. Note that it is based on the Declaration of Helsinki, which defines the set of ethical principles to be taken into consideration during experiments involving human beings.

Declaração de Consentimento

(Baseada na declaração de Helsínquia)

No âmbito da realização da tese de Mestrado, enquadrada no curso de Mestrado Integrado em Engenharia Informática e Computação da Faculdade de Engenharia da Universidade do Porto, intitulada **An Approach to Attention Modelling in Virtual Environments**, realizada pelo estudante **António David Casimiro**, orientada pelo Prof. Rosaldo Rossetti e sob a co-orientação do Prof. João Jacob, eu, abaixo assinado,

_____

(nome do participante)

, declaro que compreendi a explicação que me foi fornecida acerca do estudo em que irei participar, nomeadamente o carácter voluntário dessa participação, tendo-me sido dada a oportunidade de fazer as perguntas que julguei necessárias

Tomei conhecimento de que a informação ou explicação que me foi prestada versou os objectivos, os métodos, o eventual desconforto e a ausência de riscos para a minha saúde, e que será assegurada a máxima confidencialidade dos dados.

Explicaram-me, ainda, que poderei abandonar o estudo em qualquer momento, sem que daí advenham quaisquer desvantagens.

Por isso, consinto participar no estudo e na recolha de dados (nomeadamente de imagens) necessários, respondendo a todas as questões propostas.

Porto, ____ de _____ de 2018

_____

(Assinatura do participante ou seu representante)

A.2 Questionnaire

The following document represents the questions that were asked before the end of the experiment. The questionnaire was done using the Google Forms platform, and the online version of the questionnaire was used, so as to ease data collection. Considering the target population in this project, the questionnaire is presented in Portuguese to favour an easier understanding by the subjects. As stated in section 4.1.3, this questionnaire was inspired by the work of Deb *et al.* ([DSD+17]) and Tcha-Tokey *et al.* ([TTCLER16]).

Questionário de sessão experimental

Meta-dados a serem preenchidos pelo orientador da experiência (v. 4)

*Obrigatório

1. **Series ID** *

2. **Subject ID** *

Pretende-se com este questionário analisar padrões de comportamento e de atenção enquanto o utilizador atravessa passadeiras. Para tal, será utilizado um cenário virtual como meio de obtenção das respostas.

Note que todos os dados que providenciar serão tratados com total confidencialidade.

0. Dados sociodemográficos

3. **0.1. Idade (anos)** *

4. **0.2. Sexo** *
   Marcar apenas uma oval.
   - ◯ Masculino
   - ◯ Feminino
   - ◯ Outra: _____

5. **0.3. Profissão** *

6. **0.4. Nível de escolaridade** *
   Marcar apenas uma oval.
   - ◯ Iliterário
   - ◯ 1º ciclo (4º ano)
   - ◯ 2º ciclo (6º ano)
   - ◯ 3º ciclo (9º ano)
   - ◯ Ensino Secundário (12º ano)
   - ◯ Licenciatura
   - ◯ Mestrado
   - ◯ Doutoramento
   - ◯ Outra: _____

7. **0.5. Possui carta de condução** *
Marcar apenas uma oval.

- ◯ Sim
- ◯ Não

8. **0.6. Se possui carta de condução, há quanto tempo a possui (anos)?**

_____

9. **0.7. Com que frequência conduz por semana?** *
Marcar apenas uma oval.

	0	1	2	3	4	5	6	7	
Nunca	◯	◯	◯	◯	◯	◯	◯	◯	Todos os dias

10. **0.8. Qual(Quais) o(s) meio(s) de transporte que mais utiliza quando se desloca?** *
Marcar tudo o que for aplicável.

- ☐ Carro
- ☐ Autocarro
- ☐ Metro
- ☐ Bicicleta
- ☐ A pé
- ☐ Outra: _____

11. **0.9. Possui algum dos seguintes problemas de visão?**
Marcar tudo o que for aplicável.

- ☐ miopia
- ☐ hiperopia
- ☐ astigmatismo
- ☐ presbiopia
- ☐ estrabismo
- ☐ daltonismo
- ☐ Outra: _____

12. **0.10. Qual a mão dominante?** *
Marcar apenas uma oval.

- ◯ Destro
- ◯ Esquerdino (canhoto)
- ◯ Ambidestro

13. **0.11. Apresenta limitações motoras?** *
Marcar apenas uma oval.

- ◯ Não
- ◯ Necessito de bengala
- ◯ Necessito de cadeira de rodas
- ◯ Tenho dificuldades em caminhar, mas não preciso de dispositivos de suporte
- ◯ Outra: _____

14. **0.12. Quantas passadeiras atravessa, em média, por dia?** *

_____

15. **0.13. Com que frequência usa equipamentos de Realidade Virtual?** *
Exemplos de equipamento: Oculus Rift, HTC Vive, Playstation VR
Marcar apenas uma oval.

- ◯ Nunca
- ◯ Raramente
- ◯ Às vezes
- ◯ Muitas vezes
- ◯ Sempre

1. Análise de comportamento

Com que frequência...

16. **1.1. Atravesso a estrada mesmo quando o semáforo para os pedestres está vermelho** *
Marcar apenas uma oval.

	1	2	3	4	5	6	
Nunca	◯	◯	◯	◯	◯	◯	Sempre

17. **1.2. Atravesso a estrada na diagonal** *
Marcar apenas uma oval.

	1	2	3	4	5	6	
Nunca	◯	◯	◯	◯	◯	◯	Sempre

18. **1.3. Atravesso a estrada fora da passadeira, mesmo se existir uma a menos de 50 metros** *
Marcar apenas uma oval.

	1	2	3	4	5	6	
Nunca	◯	◯	◯	◯	◯	◯	Sempre

19. **1.4. Uso caminhos não destinados a pedestres para poupar tempo** *
Marcar apenas uma oval.

	1	2	3	4	5	6	
Nunca	◯	◯	◯	◯	◯	◯	Sempre

2. Análise de comportamento

Com que frequência...

20. **2.1. Quando o tráfego está congestionado, aproveito para atravessar entre os veículos parados** *

Marcar apenas uma oval.

	1	2	3	4	5	6	
Nunca	◯	◯	◯	◯	◯	◯	Sempre

21. **2.2. Atravesso mesmo quando vejo os veículos a aproximar porque penso que eles vão parar** *

Marcar apenas uma oval.

	1	2	3	4	5	6	
Nunca	◯	◯	◯	◯	◯	◯	Sempre

22. **2.3. Ando na via dos ciclistas, mesmo quando existe um passadiço** *

Marcar apenas uma oval.

	1	2	3	4	5	6	
Nunca	◯	◯	◯	◯	◯	◯	Sempre

23. **2.4. Atravesso sem olhar porque estou com pressa** *

Marcar apenas uma oval.

	1	2	3	4	5	6	
Nunca	◯	◯	◯	◯	◯	◯	Sempre

3. Análise de comportamento

Com que frequência...

24. **3.1. Tomo consciência de que atravessei várias estradas e cruzamentos sem ter prestado atenção ao tráfego** *

Marcar apenas uma oval.

	1	2	3	4	5	6	
Nunca	◯	◯	◯	◯	◯	◯	Sempre

25. **3.2. Esqueço-me de olhar antes de atravessar porque estava distraído** *

Marcar apenas uma oval.

	1	2	3	4	5	6	
Nunca	◯	◯	◯	◯	◯	◯	Sempre

26. **3.3. Atravesso sem olhar porque estou a falar com alguém** *

Marcar apenas uma oval.

	1	2	3	4	5	6	
Nunca	◯	◯	◯	◯	◯	◯	Sempre

27. **3.4. Esqueço-me de olhar antes de atravessar porque eu quero ter com alguém situado no outro lado da estrada** *
Marcar apenas uma oval.

	1	2	3	4	5	6	
Nunca	◯	◯	◯	◯	◯	◯	Sempre

4. Análise de comportamento
Com que frequência...

28. **4.1. Zango-me com pedestres, condutores, ciclistas, ... e grito-lhes** *
Marcar apenas uma oval.

	1	2	3	4	5	6	
Nunca	◯	◯	◯	◯	◯	◯	Sempre

29. **4.2. Atravesso devagar a estrada para irritar o condutor** *
Marcar apenas uma oval.

	1	2	3	4	5	6	
Nunca	◯	◯	◯	◯	◯	◯	Sempre

30. **4.3. Zango-me com pedestres, condutores, ciclistas, ... e faço-lhes um gesto** *
Marcar apenas uma oval.

	1	2	3	4	5	6	
Nunca	◯	◯	◯	◯	◯	◯	Sempre

31. **4.4. Zango-me com o condutor e bato no veículo dele** *
Marcar apenas uma oval.

	1	2	3	4	5	6	
Nunca	◯	◯	◯	◯	◯	◯	Sempre

5. Análise de comportamento
Com que frequência...

32. **5.1. Agradeço a um condutor que pára e me deixa atravessar** *
Marcar apenas uma oval.

	1	2	3	4	5	6	
Nunca	◯	◯	◯	◯	◯	◯	Sempre

33. **5.2. Quando ando acompanhado com outros pedestres em vias estreitas, eu ando em fila única de forma a não incomodar os restantes pedestres** *
Marcar apenas uma oval.

	1	2	3	4	5	6	
Nunca	◯	◯	◯	◯	◯	◯	Sempre

34. **5.3. Circulo no lado direito do passeio para não incomodar os restantes pedestres** *
Marcar apenas uma oval.

	1	2	3	4	5	6	
Nunca	◯	◯	◯	◯	◯	◯	Sempre

35. **5.4. Deixo um carro avançar, mesmo se eu tiver prioridade, caso não haja outro veículo atrás deste** *
Marcar apenas uma oval.

	1	2	3	4	5	6	
Nunca	◯	◯	◯	◯	◯	◯	Sempre

6. Cenário virtual

Tendo em conta a experiência que teve no ambiente virtual, por favor, responda às próximas questões

36. **6.1. Qual o grau de importância que atribui aos seguintes elementos enquanto atravessava as passadeiras?** *
Marcar apenas uma oval por linha.

	1: Sem importância	2	3	4	5	6: Muito importante
carros	◯	◯	◯	◯	◯	◯
pessoas no lado oposto da estrada	◯	◯	◯	◯	◯	◯
pessoas no mesmo lado da estrada	◯	◯	◯	◯	◯	◯
sinais de trânsito dirigidos aos condutores	◯	◯	◯	◯	◯	◯
semáforo dos pedestres	◯	◯	◯	◯	◯	◯
semáforo dos condutores	◯	◯	◯	◯	◯	◯
largura da via	◯	◯	◯	◯	◯	◯
rotunda	◯	◯	◯	◯	◯	◯

37. **6.2. Com que segurança se sentiu ao atravessar as passadeiras em cada uma das seguintes zonas?** *
Marcar apenas uma oval por linha.

	1: Nada seguro	2	3	4	5	6: Totalmente seguro
Passadeira SEM semáforo	◯	◯	◯	◯	◯	◯
Passadeira COM semáforo	◯	◯	◯	◯	◯	◯
Rotunda	◯	◯	◯	◯	◯	◯

38. **6.3. Em média, durante quanto tempo é que pensa que observou cada carro? (em segundos)** *

_____

39. **6.4. Em média, durante quanto tempo é que pensa que observou cada pessoa? (em segundos)** *

_____

40. **6.5. Em média, durante quanto tempo é que pensa que observou os sinais de trânsito dirigidos aos condutores? (em segundos)** *

41. **6.6. Durante quanto tempo é que pensa que observou o semáforo dos condutores? (em segundos)** *

42. **6.7. Durante quanto tempo é que pensa que observou o semáforo dos pedestres? (em segundos)** *

43. **6.8. Durante quanto tempo é que pensa que esteve neste cenário? (em minutos)** *

44. **6.9. Que objectos se lembra? (por favor, use a vírgula como separador)** *

9. Imersibilidade do ambiente virtual

As próximas questões focam-se na experiência que teve, de uma maneira em geral, com o ambiente virtual

45. **9.1. O ambiente virtual foi responsivo às minhas acções** *

Marcar apenas uma oval.

	1	2	3	4	5	6	7	8	9	10	
Discordo completamente	⬭	⬭	⬭	⬭	⬭	⬭	⬭	⬭	⬭	⬭	Conc comp

46. **9.2. A sensação de movimento providenciada pelo ambiente virtual foi convincente** *

Marcar apenas uma oval.

	1	2	3	4	5	6	7	8	9	10	
Discordo totalmente	⬭	⬭	⬭	⬭	⬭	⬭	⬭	⬭	⬭	⬭	Concordo totalmente

47. **9.3. Senti-me estimulado pelo ambiente virtual** *

Marcar apenas uma oval.

	1	2	3	4	5	6	7	8	9	10	
Discordo totalmente	⬭	⬭	⬭	⬭	⬭	⬭	⬭	⬭	⬭	⬭	Concordo totalmente

48. **9.4. Tive total controlo sobre as minhas acções** *
Marcar apenas uma oval.

	1	2	3	4	5	6	7	8	9	10	
Discordo totalmente	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	Concordo totalmente

49. **9.5. Os periféricos foram fáceis de usar** *
Marcar apenas uma oval.

	1	2	3	4	5	6	7	8	9	10	
Discordo totalmente	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	Concordo totalmente

50. **9.6. Gostei de ter estado no ambiente virtual** *
Marcar apenas uma oval.

	1	2	3	4	5	6	7	8	9	10	
Discordo totalmente	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	Concordo totalmente

51. **9.7. O ambiente virtual estava realista** *
Marcar apenas uma oval.

	1	2	3	4	5	6	7	8	9	10	
Discordo totalmente	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	Concordo totalmente

52. **9.8. Sofri desconforto durante a sessão experimental** *
Marcar apenas uma oval.

	1	2	3	4	5	6	7	8	9	10	
Nenhum desconforto	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	Muito desconfo

53. **9.9. Caso tivesse que utilizar o ambiente virtual de novo, a minha interacção com ele seria fácil e clara para mim** *
Marcar apenas uma oval.

	1	2	3	4	5	6	7	8	9	10	
Discordo totalmente	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	Concordo totalmente

10. Para finalizar...

Questões gerais acerca da experiência realizada

54. **10.1. Na sua opinião, quais foram os aspectos positivos da experiência? (opcional)**

55. **10.2. Na sua opinião, quais foram os aspectos negativos da experiência? (opcional)**

56. **10.3. Sugestões para melhorar a experiência (opcional)**

Com tecnologia
Google Forms

A.3 Responses

The distribution of the responses for the questionnaire's section "Immersion of the virtual environment" is made available in this section. For each question, a bar chart is shown. The purpose is to complement the description made in section 4.2.1.

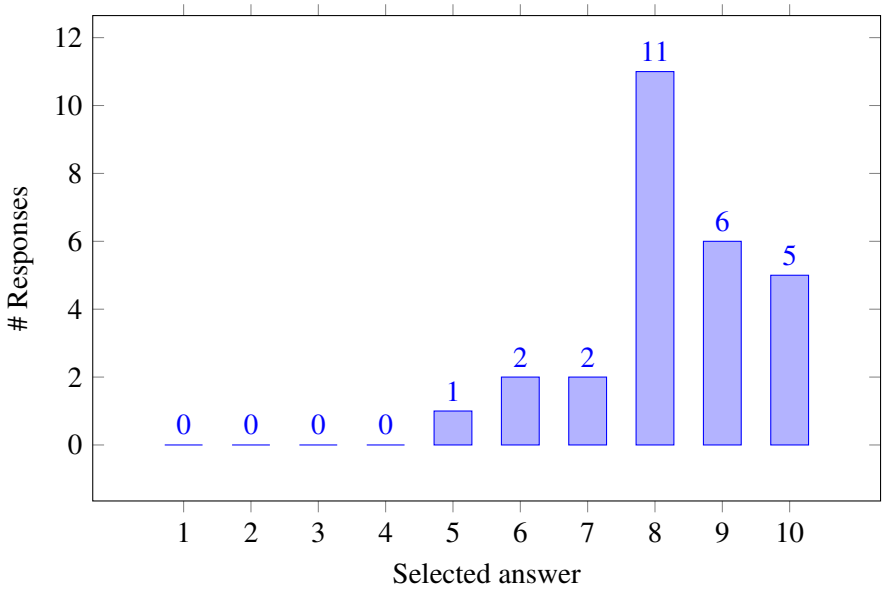


Figure A.1: Responses to question 9.1 - "The virtual environment was responsive to actions that I initiated".

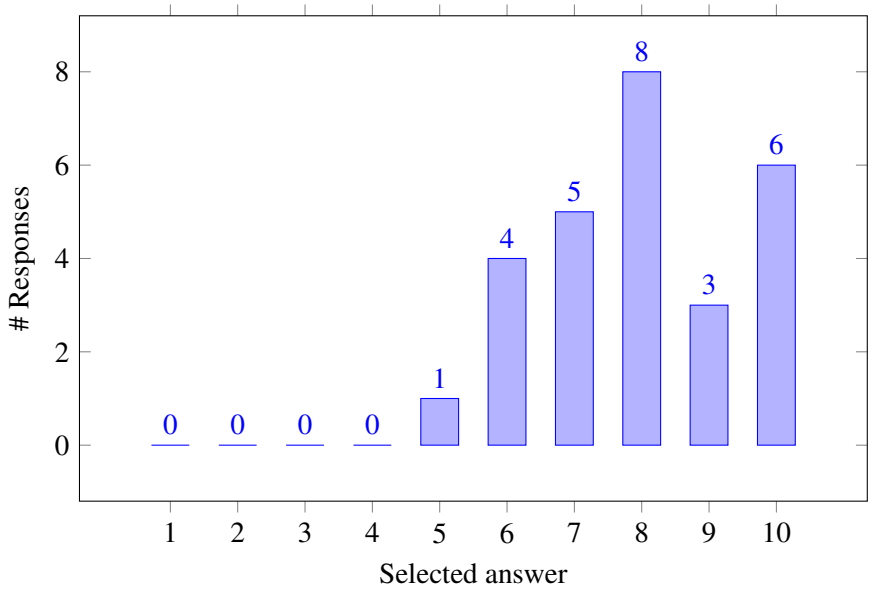


Figure A.2: Responses to question 9.2 - "The sense of moving around inside the virtual environment was compelling".

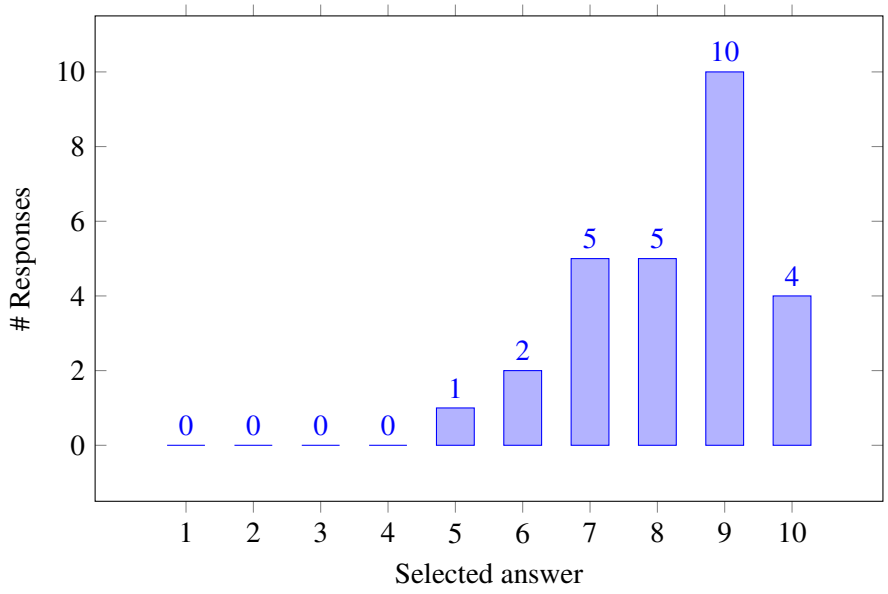


Figure A.3: Responses to question 9.3 - "I felt stimulated by the virtual environment".

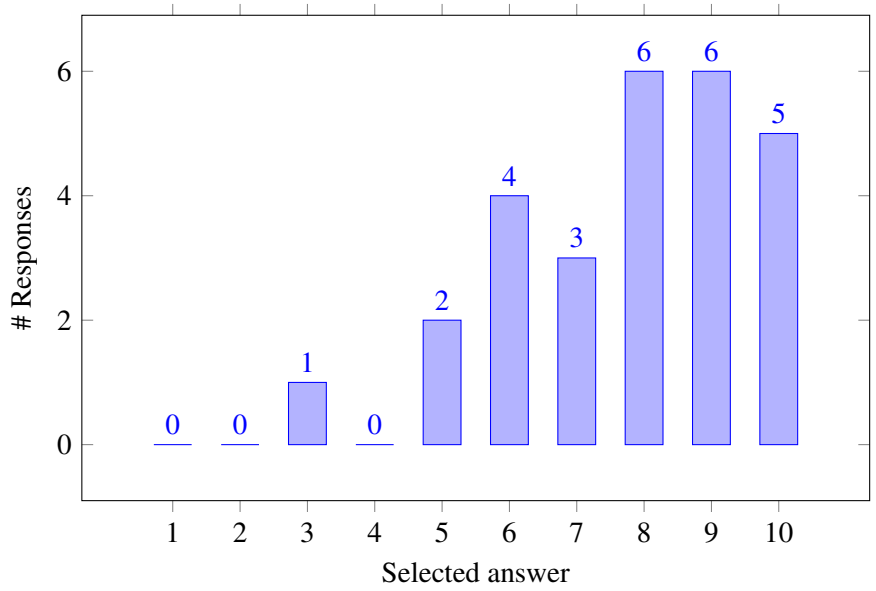


Figure A.4: Responses to question 9.4 - "I felt I could perfectly control my actions".

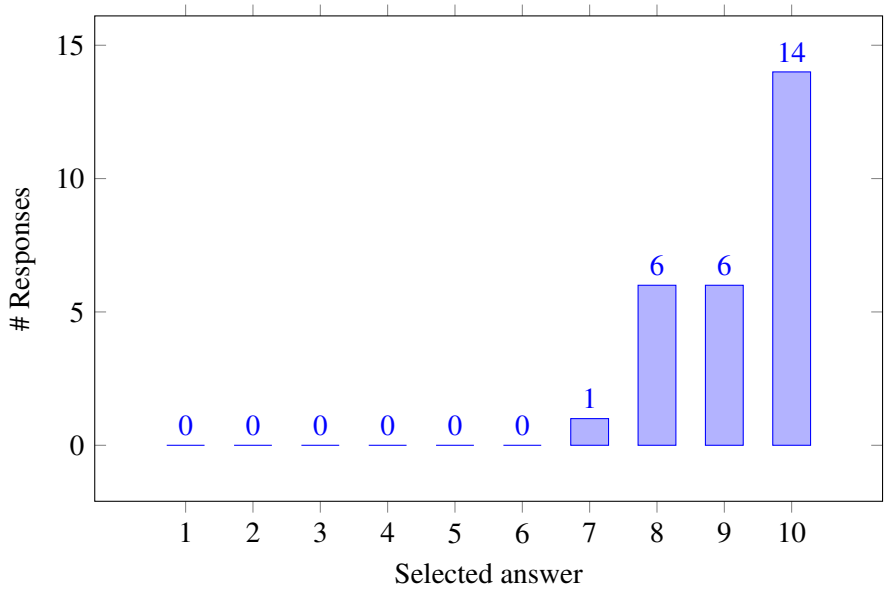


Figure A.5: Responses to question 9.5 - "I thought the interaction devices were easy to use".

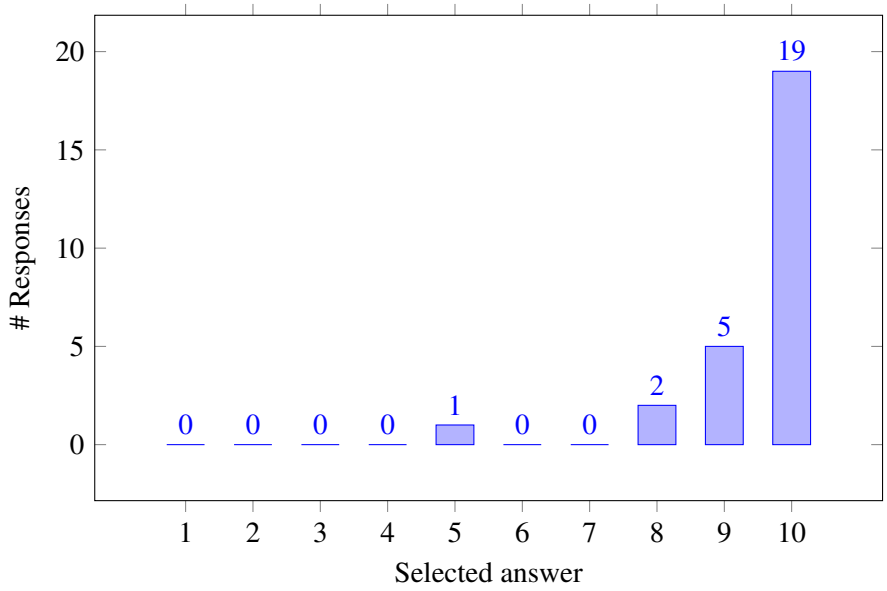


Figure A.6: Responses to question 9.6 - "I enjoyed being in this virtual environment".

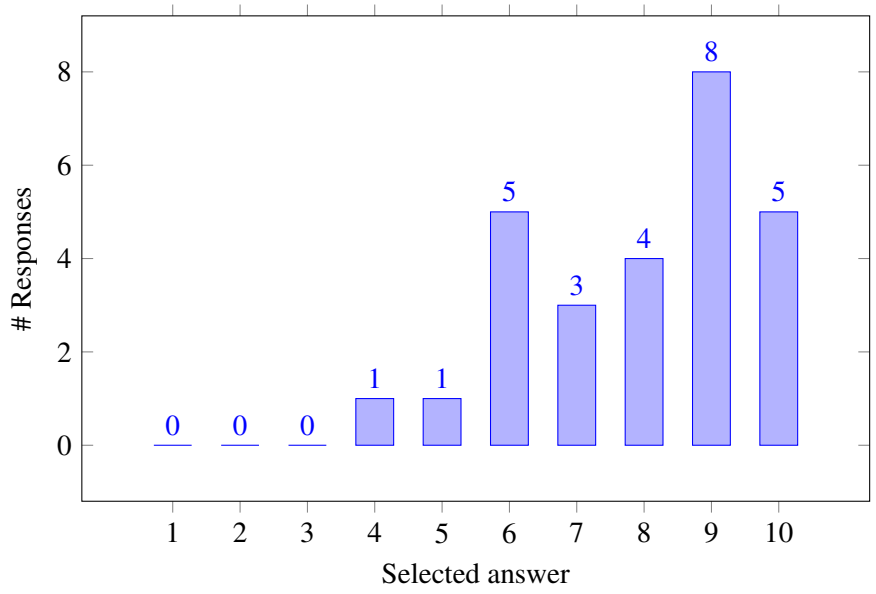


Figure A.7: Responses to question 9.7 - "The virtual environment was realistic".

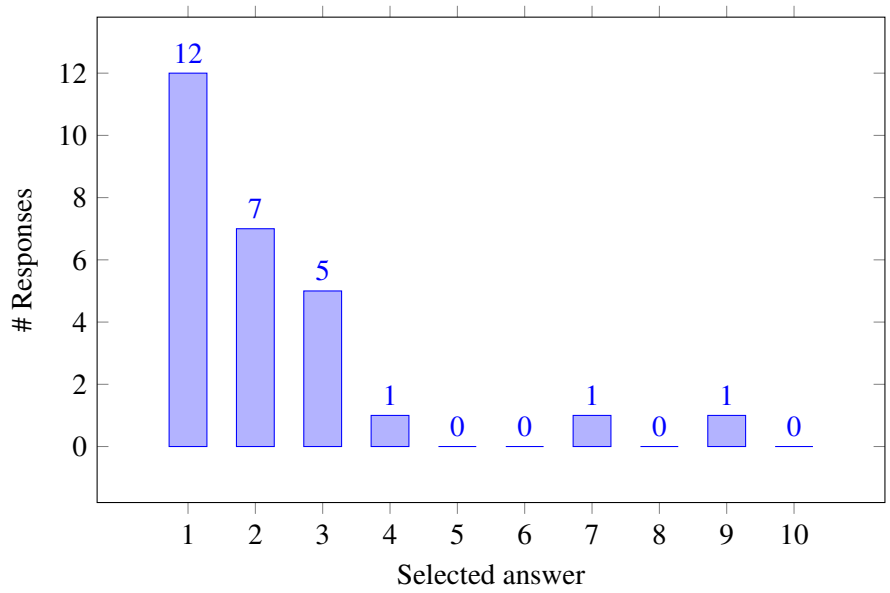


Figure A.8: Responses to question 9.8 - "I suffered from fatigue during my interaction with the virtual environment".

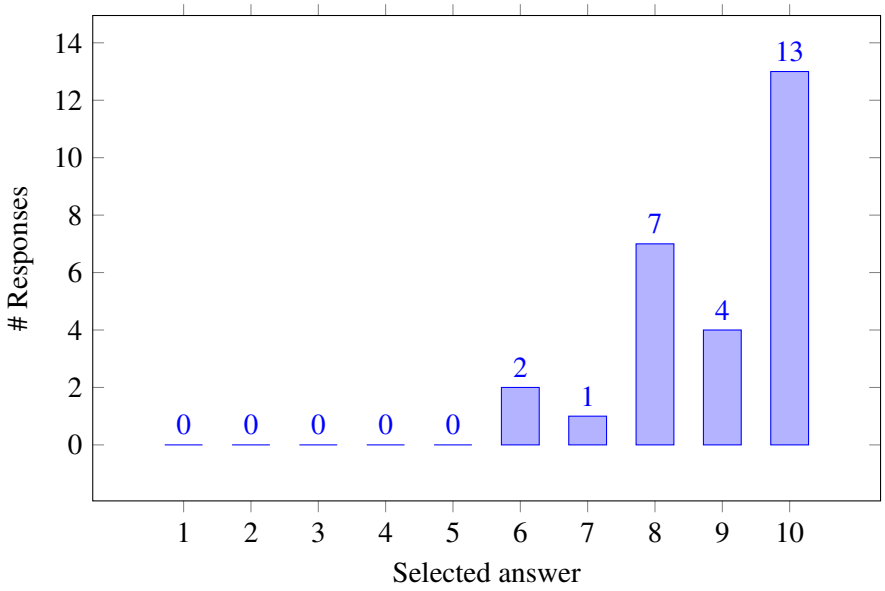


Figure A.9: Responses to question 9.9 - "If I use again the same virtual environment, my interaction with the environment would be clear and understandable for me".