Saarland University

Faculty of Natural Sciences and Technology I

Department of Computer Science

Master thesis

A study of visual perception in virtual reality interfaces: Analyzing importance of object attributes

submitted by

Sruti Subramanian

submitted

27.09.2016

Reviewers:

Prof. Dr. Jörg Siekmann

Prof. Dr. Babette Park

Eidesstattliche Erklärung

Ich erkläre hiermit an Eides Statt, dass ich die vorliegende Arbeit selbständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel verwendet habe.

Statement in Lieu of an Oath

I hereby confirm that I have written this thesis on my own and that I have not used any other media or materials than the ones referred to in this thesis

Einverständniserklärung

Ich bin damit einverstanden, dass meine (bestandene) Arbeit in beiden Versionen in die Bibliothek der Informatik aufgenommen und damit veröffentlicht wird.

Declaration of Consent

I agree to make both versions of my thesis (with a passing grade) accessible to the public by having them added to the library of the Computer Science Department.

Saarbrücken,_____ _____
(Datum/Date) (Unterschrift/Signature)

Acknowledgement

I would like to express my sincere gratitude to my advisor, Dr. Sergey Sosnovsky, for having provided me with the opportunity and guidance to pursue this thesis. I would like to thank him for all his advice and constant supervision from the very beginning. I am grateful to Prof. Dr. Jörg Siekmann, for being my supervisor and for his support. I am thankful to Yecheng Gu for giving me a great introduction to SafeChild and for all his technical advice and timely help.

I would also like to thank Prof. Dr. Ronald Brünken for his guidance and Prof. Dr.Babette Park for being my supervisor, and for her constant support in conducting the Experiment. I also need to thank Andreas Korbach for his timely guidance.

Finally, I am highly thankful to my family and friends for all their love and support.

Abstract

Virtual Reality is an emerging and an ever evolving area of work. The power of modern devices provide a wide range of input/output interfaces that have been created over the last decade, such as the Microsoft Kinect and the Occulus Rift which are being used for a wide range of applications ranging from education to entertainment. New challenges emerge with the increase in applications. This thesis aims on handling one such challenge.

One main difficulty was the necessity to keep track of what the user has or has not seen within the Virtual Reality environment. Without this fundamental information, it would not be possible to claim whether a particular situation had occurred with the acknowledgement of the user upon seeing a particular object, or merely that by accident. For this exact purpose of tracking a user's attention, eye trackers are widely used. However eye trackers are not commercial products, and hence are not a part of the everyday household gadgets, this thesis is an initial attempt in identifying an alternative solution.

The thesis analyzes several environmental/object attributes of a particular object within an environment, and examines how the modification of those parameters can influence a user's attention towards that specific object. This should help predict a user's visual attention within the environment solely based on the modification of these parameters, and determine the influence of those parameters in driving a user's attention towards the target object.

Table of Contents

1 INTRODUCTION

Virtual reality (VR) has drawn much attention over the years. The fundamental origins of VR can be traced back to "The Ultimate Display" (Sutherland 1996), which was a seminal paper that initially introduced key concepts of immersion in a simulated world, and also that of a complete sensory input and output: "The screen is a window through which one sees a virtual world. The challenge is to make that world look real, act real, sound real, feel real". It can be summarized as offering simulation to users as an interface metaphor to a synthesized world. It has become the innovation agenda for a growing community of researchers and industries. The motivation for such a remarkable research direction is twofold. From an evolutionary perspective it has been obvious that VR technology has always been seen as a way to overcome the limitations of standard human-computer interfaces; and from a revolutionary perspective, virtual reality technology has been the means to open the door to new types of applications that exploit the possibilities offered by presence simulation (Gobbetti and Scateni 1998).

1.1 Motivation

There are different kinds of VR: some restrict interaction of a user with its objects, others are free-world dynamic VR environment that let the user freely move and interact with the VR objects. In many contexts, it is highly essential to provide intelligent support to VR users as a response to their interaction or their performance within the virtual environment. That, however, is a highly challenging task as it is difficult to trace user's attention and track which of the VR objects have been actually noticed and interpreted by the user. Particularly in the case of intelligent and adaptive applications it is highly essential to analyze the user's performance based upon whether certain important objects within the environment have been perceived by the user or not. And in such circumstances eye trackers have proved to be highly helpful. Eye trackers are used in VR for interactive needs as well as for diagnostic purposes. The user's gaze direction, as well as head position and orientation are tracked to allow recording of the user's fixation within the environment. Methods have already been deployed for (1) integration of the eye tracker into a VR framework (Duchowski et al. 2001) (2) Stereo calculation of the user's 3D gaze vector (Matsumoto and Zelinsky 2000). (3) 3D calibration developed to estimate the user's inter-pupillary distance *post-fact* (Duchowski et al. 2002). And (4) eye movement analysis in 3D-space (Bahill, Clark, and Stark 1975). The results obtained indicate that the recorded eye movements provide valuable human factor process

that measures complementing performance statistics used to gauge training effectiveness. However, eye trackers have their own drawbacks.

Almost everyone today has a mobile phone or a computer which they use on a daily basis. Many of the devices which previously did not have computational facilities now have those features, such as television, mobile phone, etc. There is currently an explosion of interfaces for VR, and an emergence of real commercial products exploiting them, such as Microsoft Kinect (Zeng 2012), Microsoft Hololense, and Oculus Rift (Desai et al. 2014). With all these new devices and the fact that we have interfaces that run these devices, the spread of applications has and will grow faster. It could be seen that new use cases and contexts of VR will definitely emerge. VR has become more complex and open ended and some of the existing technologies aim at applying intelligent and adaptive approaches in VR settings, such as intelligent support of user activity. For such tasks, the detection of a user's visual perception is very important. The next section elaborates the underlying problem.

1.2 Problem description

An important example of a VR application that would require capturing a user's visual attention would be traffic simulation. In such a scenario, it would be highly essential to analyze whether the user did not exhibit certain skills while driving or walking because he was unaware of a particular traffic rule, or maybe because he did not see an important traffic element within the environment.

This problem is essential for the VR based adaptive training system SafeChild (Gu, Sosnovsky, and Ullrich 2015). This intelligent tutoring system aims at helping children train pedestrian safety skills by practicing crossing virtual streets in various traffic situations. SafeChild provides different scenarios in which the users are expected to apply specific safety skills. In the context of SafeChild, it is highly significant to analyze why a learner did not exhibit a required skill. It could happen because he was unaware of a specific rule or maybe because he did not see an important traffic element in the environment. Hence, it becomes highly essential to track a user's visual attention in this regard, as a rich stream of visual data enters our eyes every second (Koch et al. 2006).

1.3 A brief description of the approach

The thesis proposes an approach to analyze several attributes of VR objects with respect to their influence on user's attention towards the objects. For instance attributes such as the

object's size, contrast, distance from the center of the screen, speed, distance to the object, and time of how long the particular object has been observed. The influence of these attributes has been analyzed by conducting an experiment that involved users in recognizing a target object in numerous scenes (images). For the experiment, three main factors were taken into account, and for each scene these factors were slightly modified. Essentially, it was analyzed how the user's performance was influenced by the modification of these factors in each scene, and hence deduced how these factors influenced user's visual attention within the VR environment.

1.4 Thesis Structure

This thesis is structured as follows. Chapter 2 describes the background of this work by introducing SafeChild project and overviewing related research on capturing viewer's visual attention. Chapter 3 presents the details of the initial implementation conducted within the SafeChild environment on extracting visual attributes of virtual objects within the SafeChild environment. The user study design is detailed in Chapter 4. Chapter 5 describes the results obtained from the experiment. Chapter 6 discusses potential direction for future work and concludes the thesis.

2 BACKGROUND AND RELATED WORK

This chapter provides an overview of SafeChild and the related work on visual attention modeling in VR environments. Several other models for predicting and tracking user attention and their limitations have also been described.

2.1 The context of this thesis: SafeChild project

SafeChild is a VR intelligent tutoring system that aims at helping children train pedestrian safety skills (Gu, Sosnovsky, and Ullrich 2015). It comprises of a VR city environment that includes realistic urban architecture and traffic simulation, developed with the Unity3D game engine. The user can engage in typical pedestrian activities such as crossing roads under different traffic conditions. Several parameters within the environment can be adjusted such as: car speed, traffic density and walking speed of the user. An example of a typical SafeChild environment is shown in Figure 1.

Figure 1: SafeChild Environment (picture taken from safechild.celtech.de).

2.1.1 Architecture

The overall architecture of SafeChild consists of four typical Intelligent Tutoring System components (Corbett, Koedinger, and Anderson 1997): the Domain Model describing the knowledge to be taught, the Student Model representing current state of learning of each individual child, the Pedagogical Model that defines how to support training, and the Interface Model that serves road-crossing exercises, as well as controls and monitors user's interaction with the rest of the system.

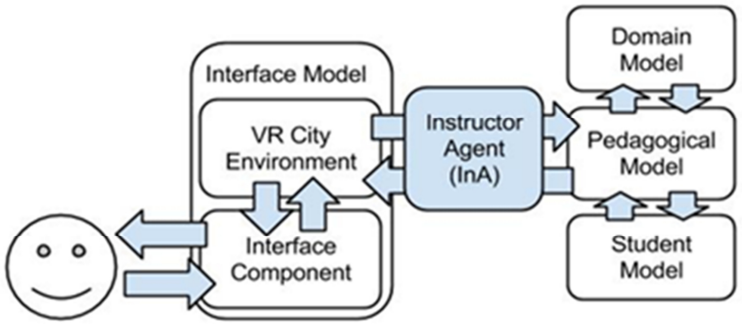

Figure 2: SafeChild Architecture ((Gu et al, 2015)).

2.1.2 Child pedestrian safety skills

SafeChild focuses on several safety skills that are categorized into two groups namely: the "basic skills", which are less demanding cognitively and should be easier for children to apply and master, and the "advanced skills" that involve meta-cognitive processes, more complex planning, decision making procedures and maintaining the awareness of others. For SafeChild, the hierarchical organization of skills becomes an additional source of information to elaborate student modeling (by propagation) and adaptation strategies (e.g., by task sequencing). The complete hierarchy of skills is shown in Figure 3.

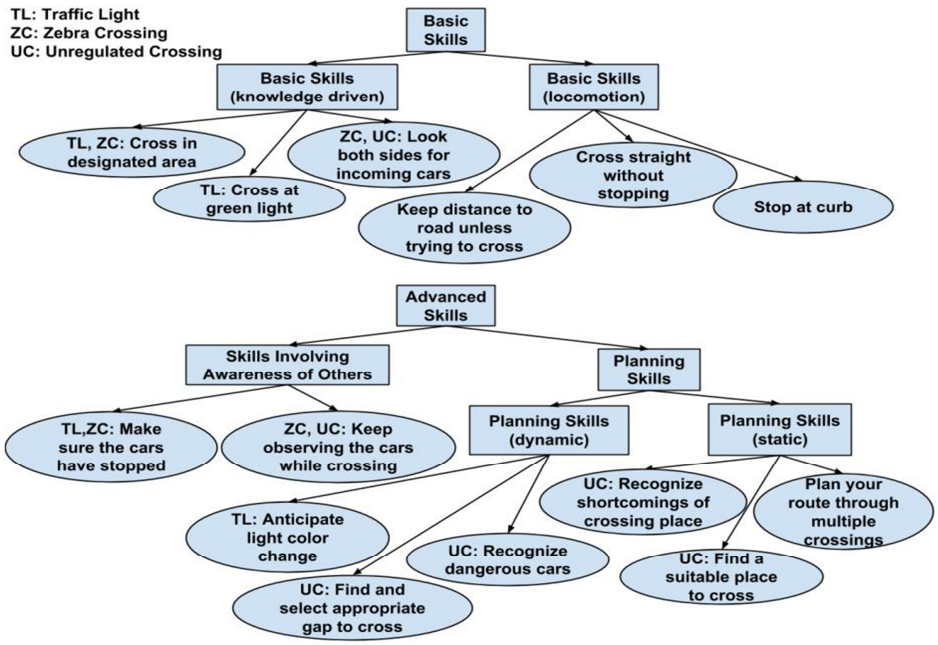

Figure 3: Skills in SafeChild ((Gu et al, 2015)).

2.1.3 Virtual traffic exercise

For the domain of child pedestrian safety, there are several behavior rules that children need to know and corresponding skills that they need to master in order to become safe pedestrians. There are three high-level skills that are mandatory: (I) making judgments about safety of a

crossing place (II) identifying traffic that could pose a threat and (III) integrating information from different aspects of the traffic situation. These high-level skills require a number of underlying abilities that are especially challenging for young children since their cognitive apparatus is still developing.

The challenge in such circumstances was to keep track of what the user has or has not seen, because this information was essential to decide whether a particular incident took place with the user's acknowledgement or not. For instance, in the case of a user crossing the road it would be necessary to know whether he saw the traffic light, the fast moving vehicles, and also if he was aware of the basic pedestrian rules. For the above reasons it was highly essential to track a user's visual perception in this regard.

2.2 Relevant visual attention models

There are several visual attention models used for the purpose of detecting a user's visual attention. High-level cognitive and complex processes such as object recognition or scene interpretation rely on data that has been transformed in such a way as to be tractable. A few of the following models were used to compute saliency maps for any image or input video (Rothenstein and Tsotsos 2008).

2.2.1 Model of user interest

This particular model was implemented to capture user's visual attention using an "interest module" that would find the objects currently being looked at and record the time. Three slightly different approaches were implemented for determining the apparent level of interest of the user in a given object (Fuchs, Kedem, and Naylor 1980). In the first model, when the screen coordinate of the gaze point corresponds to an object or objects, the tally for that object is incremented by one. The interest level equals the tally. In the second model, the elapsed time since the given object was seen is multiplied by a constant K2 and subtracted from a constant K1, times the tally of glances for the object. In the third model whenever there is a fresh look at an object the old value is decayed by the proper amount and then incremented by a constant (Fresh look constant).

2.2.2 Saliency-based model

Most attention models are inspired by cognitive concepts. These models use three feature channels namely color, intensity and orientation. It has shown to correlate with human eye movements. A given input image is subsampled into a Gaussian pyramid and each pyramid

level σ is decomposed into channels for Red(R), Green(G), Blue(B), Yellow(Y), Intensity(I), and local orientation(O_θ). In each channel, maps are summed across scale and normalized again. These maps are linearly summed and normalized once more to yield the"conspicuity maps". Finally the conspicuity maps are linearly combined once more to generate the saliency map (Itti, Koch, and Niebur 1998).

2.2.3 Bayesian Models

This model combines the sensory evidence along with the prior constraints involved. Prior knowledge and sensory information are combined probabilistically according to Baye's rule to find the object of interest. It also entails the assumption that the distribution of a feature at a point on the target does not change with location (Borji, Sihite, and Itti 2012).

2.2.4 Decision theoretic model

This model evolves to produce decisions about the states of the surrounding environment that are optimal in a decision theoretic sense. The overarching point is that visual attention should be driven by optimality with respect to the end task. The Bayesian computation is a special case of the Decision theoretic model. Saliency computation in the entire decision theoretic approach boils down to calculating the target posterior probability. Decision theoretic models have been very successful in computer vision applications such as classification while achieving high accuracy in fixation prediction (Horvitz and Lengyel 1997).

3 THE APPROACH

This chapter explains the details of the proposed approach; it outlines the main parameters of graphical objects influencing their prominence that are considered for capturing and analyzing user's visual perception. It also includes a brief description of the initial implementation.

3.1 Brief description of the approach

The overall idea behind this work is supported by the assumption that there exists a range of parameters within a VR environment that contribute towards capturing user's visual attention. The approach is carried out by means of sequentially modifying a few environmental parameters and analyzing how these modifications have influenced the user's performance in identifying the particular target object within the environment. Several such parameters have been described below.

3.2 Bottom-up parameters of an object

Bottom-up cues are mainly based on characteristics of a visual scene. Regions of interest that attract our attention in a bottom-up manner must be sufficiently distinctive with respect to the surrounding environment (Desimone and Duncan 1995). "Bottom up attention is fast, involuntary and most likely feed forward". A prototypical example of the bottom-up attention is looking at a scene with one horizontal bar among several vertical bars where attention is immediately drawn towards the horizontal bar (Treisman and Gelade 1980). While many models fall in this category, they can only explain a small fraction of eye movement since a majority of fixations are driven by the task (Henderson and Hollingworth 1999). Several object parameters falling in this category that could be highly influential have been described below.

3.2.1 Distance / Size

The distance to a particular object from the location of the user could essentially be one of the most important parameters involved in such an analysis. The larger the distance, the smaller the object appears. Whereas on the other hand, the closer an object is to the player, the larger the object appears, and hence it is more likely that the player has observed the object.

3.2.2 Speed

Speed is another important factor that can potentially influence visual prominence of an object. However, there are several aspects of speed that should be taken into account. The

absolute speed of an object within the environment is less important than its angular velocity within the user's visual cone, as that also depends on the distance to the object and the speed of the user himself. Direction of the object moving can be influential as well. While higher speed essentially means shorter time to see the object that can be translated into lower prominence, the moving object on a static scene is more prominent than the objects standing still. Hence, when considering such attribute of a target object as speed, one should take into account speed of its neighboring objects as well.

3.2.3 Contrast

The color of a particular object, in conjunction with that of its background may also influence whether it catches the viewer's attention. For instance, a red object on a grey background is probably more likely to be observed than a white object on a grey background. The prominent colors of objects ensure that they stand out from the rest of the scene, and catch the user's attention.

3.2.4 Time Observed

The amount of time an object is present within the viewer's visual cone, or within his view may also be another important parameter. If the viewer's gaze is fixed towards a particular object for a longer period of time, the probability of the viewer having seen that object could be comparatively higher.

3.2.5 Distance from the Center of the screen

The position of an object with respect to the center of the screen could be another important parameter. An object that is located in the center of the screen is more likely to capture viewer's attention in comparison to an object that is situated towards the corner of the screen.

3.3 Top-down parameters of an Object

Top-down attention is slow, task driven, voluntary and closed-loop. One of the most famous examples of top down attention guidance (Hayhoe and Ballard 2005) showed that eye movements depend on the current task with the following experiment: subjects were asked to watch the same scene under different conditions (Itti et al. 2001). Eye movements differed rapidly for each case. The factor of deciding where to look relies on the target object that is to be found.

3.3.1 Scene context

It was observed that targets that appeared in repeated configurations relative to some background (distractor) objects were more quickly detected (Joubert et al. 2008). Semantic associations among objects in a scene (eg. A computer is often placed on top of a desk) or contextual cues have also been shown to play a significant role in the guidance of eye movement (Hwang, Wang, and Pomplun 2011). Gist representations have become increasingly popular in computer vision since they provide rich global yet discriminative information useful for many applications such as search in large-scale scene datasets available today.

3.3.2 Task

Task has a strong influence on deployment of attention. It has been claimed that visual scenes are interpreted in a need based manner to serve task demands (Triesch et al. 2003). It has been showed that there is a strong relationship between visual cognition and eye movements when dealing with complex tasks. Subjects performing a visually guided task were found to direct a majority of fixations towards task relevant locations (Hayhoe and Ballard 2005). It is often possible to infer the algorithm a subject has in mind from the pattern of her eye movements.

The prevailing view is that the bottom-up and top-down attention is combined to direct the attention behavior. An integration method can explain when and how to attend to a top down visual item or to skip it for the sake of a bottom up cue.

3.4 Initial implementation

The entire implementation within this thesis was carried out using the Unity3D game engine. Unity3D is a powerful cross-platform 3D game engine and a user friendly development environment. It has been developed by Unity Technologies and is used to create video games for PC, consoles, mobile devices and websites.

The initial implementation for this thesis was carried out within the framework of the SafeChild platform. The target objects which were considered were important traffic elements such as traffic lights, traffic signs, zebra crossings, and cars. As the user was interacting within the environment and crossing virtual roads and coming in contact with a target object, the following parameters with respect to that particular object were extracted from the environment: distance, time observed, size, and speed.

Figure 4 depicts a display box showing to a user, values of these parameters for a nearby traffic light. These parameters were simultaneously logged to an external log file along with the time stamp. This log file could have been used for further analysis. This was the initial implementation intended to demonstrate the possibility to obtain the mentioned parameters of virtual objects from within the environment.

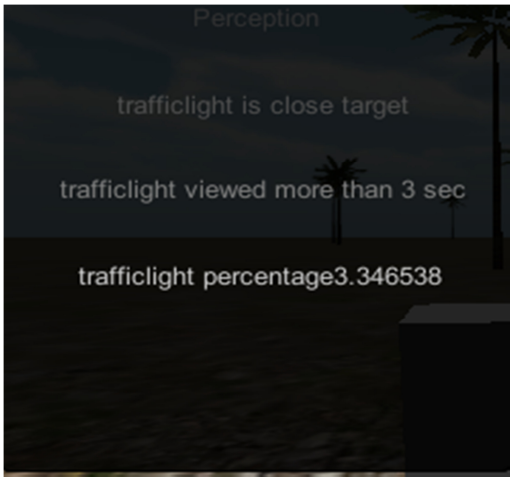

Figure 4: Display box with respect to a traffic light target object.

Following this, to analyze, which of those parameters were influential in drawing user's attention, a user study was further conducted.

3.5 Conclusion

This chapter described several parameters of virtual objects that can potentially influence user's attention within a VR environment. The chapter also describes the initial implementation of the approach. The implementation demonstrated that it is possible to extract several such parameters from a VR environment with respect to specific target objects.

4 STUDY DESIGN

The previous chapter explained parameters that could be potentially taken into consideration to implement this approach. This chapter discusses the detailed structure of the user study that has been conducted to put the proposed approach to use with the selected set of parameters.

4.1 The task

The study has been designed around an object recognition task where a cohort of human subjects had to identify a particular target object from a sequence of images (static scenes). These are referred to as static scenes because unlike other interactive VR environments which also comprise of dynamic objects in the scenes, here everything is stationary and non-interactive. To make this study relevant to SafeChild all the static scenes were designed in a typical traffic based context. The task was to identify a common target. For the purpose of this experiment a unique yellow and black colored road sign was designed, this is shown in Figure 5.

Figure 5: Yellow and black colored target road sign.

Users with driving experience might be better equipped in terms of recognizing real road signs due to observing them more often, remembering how they look and where they ought to be located within a typical traffic environment. Therefore, we attempted to make the task impartial to all the users by considering the uniquely designed road sign as a target instead of common road signs potentially more familiar to more experienced traffic participants.

4.2 The Idea

The basic motivation behind this study was to observe which object attributes were significant in terms of driving user's attention towards a target object within the environment. For the purpose of this study three main factors were considered: size of the object, the object's contrast with respect to its background, and the position of the object with respect to the center of the screen. Further, for each static scene these three factors were sequentially varied over a range of values (high, medium, low). With such a sequential variation, a total of twenty-seven scenes were designed. The design mechanism that was adopted to differentiate the three variations is described in Table 1. The scenes were further classified in terms of either a busy environment or a less busy environment, which differed in the total number of traffic elements that were present within the scenario. The details involved in the classification of these two environments are shown in Table 2. Hence, ultimately, twenty-seven such scenes were designed individually for the busy environment and the less busy environment, resulting in a total of fifty-four static scenes.

	Size	**Contrast**	**Position**
High	1.5-3 cm	White/Sunny	Middle(18-20 coms from edge)
Medium	0.7-1 cm	Orange/dark/red/default sky	7-11 cms
Low	0.5 cm	Yellow/greenish/yellow/dark and shadowy	Edge (0-2 cms)

Table 1: sequential variation of the three factors.

	Busy environment	**Less busy environment**
Vehicles	3-4	1-2
Road signs	4-6	2-3
Total traffic elements	7-10	3-4

Table 2: classification of busy and less busy environment.

4.3 Sequential variation

The above considered parameters were sequentially varied for each scene, over a range of high, medium or low. The details involved for each variation with respect to the corresponding object attribute is described below.

4.3.1 High

When the parameters were tuned to high, the size of the target road sign was considerably large in size, measuring around 1.5cms to 3cms on the screen. Similarly, the factor of contrast was sufficiently high. Since the yellow and black colored road sign was considered as the target, a background color of white showed high contrast. Likewise, the overall contrast of the scene was very bright and sunny. When positioning the third parameter (distance to the center of the screen), the target object was positioned precisely in the center of the screen, that is around 18 to 20cms from the edge of the screen.

Figure 6: Scene with high target size, high contrast, and high positioning to the center.

Figure 6 shows a "busy" scenario in which all the three parameters are set to high. Hence target sign is present in a large size with good contrast, and positioned towards the center of the screen.

4.3.2 Medium

When the parameters were tuned to medium, the size of the target road sign was of a considerable size, which was neither too large nor too small. In such a scenario the target was partially hidden, and the overall amount of the target that was visible measures to a size of about 0.7cms to 1cm on screen. Similarly, the parameter of contrast would be highly moderate. The background contrast was either an orange or a dark red color. Likewise the overall contrast of the scene was a moderate default blue sky. Similarly when positioning the third parameter (distance to the center of the screen), the target object was positioned

somewhere between the center and the edge of the screen, with a measurement of nearly 7 to 11cms from the edge of the screen.

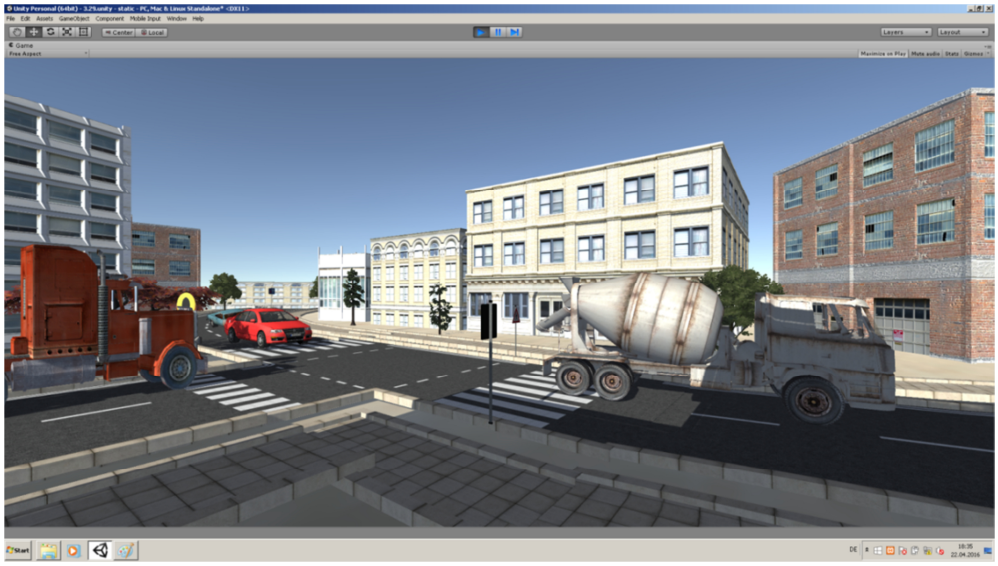

Figure 7: Scene with medium target size, medium contrast, and positioning towards the center.

Figure 7 shows a "busy" scenario in which all the three parameters are set to medium. Hence the yellow and black target sign is moderately blocked, revealing a major portion of the sign with a moderate contrast, and positioned between the center and the edge of the screen.

4.3.3 Low

When the parameters were tuned to low, the size of the target road sign is of a smaller size. Similarly, the factor of contrast was very low. The background contrast for the case of low was either a yellow or greenish yellow color that camouflaged well with the yellow and black colored road sign. Also in case of such low contrast the sign was placed in dark and shadowy areas of the screen. When positioning the third parameter, the distance to the center of the screen to low, the target object was positioned somewhere towards the corner of the screen, measuring about 0 to 2cms away from the edge of the screen.

Figure 8: Scene with low target size, low contrast and positioning towards the center.

Figure 8 shows a "busy" scenario in which all the three parameters are set to low. Hence, the target sign is blocked to a greater extent, revealing only a small portion of the sign, with a very low contrast, and positioned towards the edge of the screen.

4.4 Eye-tracking

On completing the task, the eye tracker recording was analyzed to see the user's performance in each scene. By means of this analysis, it was possible to monitor how the variations in the three factors (size, contrast, position) influenced the user's performance within each scene. From the eye tracker it was possible to obtain three important parameters denoting the user's performance within each scene: time to the first fixation on the target, time to the first mouse click on the target, and number of previous fixations made within the scene. And these three parameters in conjunction with the three factors (size, contrast, position) that were varied, helped in the computation of the final results. The three significant parameters that were obtained from the eye tracker are described below. Figure 9 shows the tobii eye x eye tracker mounted at the bottom of the monitor.

Figure 9: Tobii eye x eye tracker.

4.4.1 Time to the first fixation

This denotes the amount of time in seconds that the user took to fixate on the target object within the particular scene.

4.4.2 Time to the first mouse click

This parameter denotes the amount of time that the user spent before clicking on the target object within the static scene.

4.4.3 Number of previous fixations

This parameter denotes the number of previous fixations that the user has made within the environment before fixating on the target.

4.5 Questionnaire

All users filled out a questionnaire prior to the experiment in which they had to answer the following questions.

- What is your gender?
- Do you have a driver's license?
- Do you play video games?
- Do you have vision problems?
- Do you wear spectacles?
- Do you wear contact lenses?

The questionnaire helped to further analyze whether the above attributes of a user influenced their performance in the experiment.

4.6 d2-R test

In addition to the questionnaire the users had to take a d2 test to measure their concentration or focused attention. The test consists of the letters d and p, which are arranged in 14 rows of 57 marked above and / or below with 1 to 4 lines[1].

The role of the user was to strike off in 20 seconds as many letters d which had more or less than two strokes in a row. The examiner gives the start signal and calls after every 20 seconds to move to the next character row. The test takes including instructions about 8 minutes.

The following characteristic values are calculated at the d2-R, the most important is the concentration power.

- KL - concentration power: Correct operations minus error (BZO - AF - VF).
- BZO - Number of processed targets: the last painted target (d with two strokes) per line (summed over all rows).
- F% - Percentage error: relative frequency of errors in machining (100*(AF + VF)/BZO).
- AF - error of omission: overlooked or missed targets (up to the last painted target object; false negative).
- VF - Likelihood of error: mistakenly painted targets (false positive).

These values could have been further used to analyze the influence of a user's concentration level in their performance in the experiment. Figure 10 shows a small sample of a d2-R test.

[1] https://www.testzentrale.de/shop/test-d2-revision.html

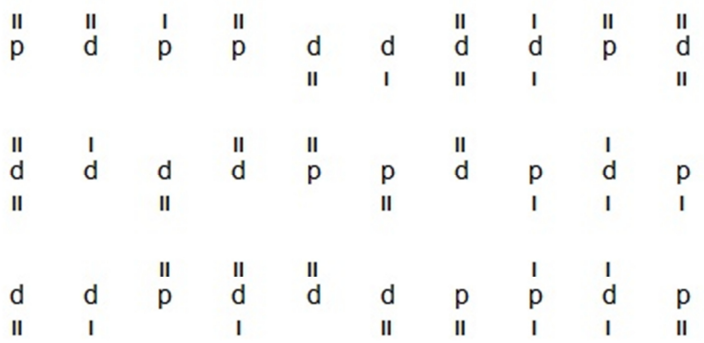

Figure 10: d2-R test

5 EVALUATION AND RESULTS

This chapter presents the results of the conducted experiment.

A total of 23 users participated in the user study. Their diversity ranged from their nationality to their field of education. They were students from the following disciplines: Education Technology, Computer Science, Computer and communication technology and Bioinformatics. The results of the questionnaire are shown in Table 3.

Gender	driver's license	Play video games	vision problems
Male	Yes	No	No
Male	Yes	Yes	No
Male	Yes	Yes	No
Male	Yes	Yes	No
Female	Yes	No	No
Male	Yes	Yes	Yes
Female	No	Yes	No
Male	Yes	No	No
Female	Yes	No	No
Female	Yes	No	Yes
Female	Yes	Yes	No
Female	Yes	Yes	No
Female	No	Yes	Yes
Female	Yes	No	No
Female	Yes	Yes	Yes
Female	Yes	No	No
Female	Yes	No	Yes
Male	No	Yes	No
Male	No	Yes	No
Female	No	No	No
Female	Yes	No	Yes
Female	Yes	No	Yes
Male	No	No	No

Table 3: Results of the questionnaire.

The final analysis revealed that there was no significant difference in results obtained when comparing the different categories of users. Table 4 depicts the descriptive d2-R test results that were obtained.

F%	BZO	AF	VF	KL
55.2	201	107	4	90
45	220	88	11	121
65.1	189	119	4	66
28.3	240	68	0	169
64.4	188	120	1	67
93.2	161	147	3	11
23.6	258	50	11	197
9.5	294	14	14	266
63.4	191	117	4	70
66.3	187	121	3	65
30	240	68	4	168
65.2	187	121	1	65
73.6	178	130	1	45
145.2	126	182	1	-57
121.4	140	168	2	-30
82.6	172	136	6	30
62.8	188	118	0	70
131.1	132	173	0	-41
33	209	63	6	130
235.2	105	201	46	-142
113.2	144	161	2	-19
71.9	153	110	0	96
36.2	221	80	0	141

Table 4: d2-R Test Results

Table 5 represents the descriptive results obtained based on the input data from the Tobii eye x recording. It can be seen that for all three factors describing visual characteristics of the target object, the attention decreased as the values of the parameters decreased. In other words, as the size of the object, its centrality and the contrast lowered, the time to fixate on the object and click on it, as well as the number of previous fixations went up. The following

subsection presents these results in more details and accompanies them with the outcomes of statistical tests verifying corresponding hypotheses.

	Time to first fixation			Time to first mouse click			Number of previous fixations		
	High	Medium	Low	High	Medium	Low	High	Medium	Low
Size	M=1.31 SD=.14	M=1.38 SD=.18	M=1.89 SD=.22	M=2.13 SD=.24	M =2.23 *SD =.21*	M=3.04 SD=.24	M=4.29 SD=.54	M =4.45 SD = .72	M=5.68 SD=1.79
Contrast	M=1.35 SD =.17	M = 1.44 SD = .22	M=1.77 SD=.21	M=2.30 SD=.21	M=2.46 SD=.25	M=2.71 SD=.19	M=4.67 SD=.70	M=4.32 SD=.44	M=4.99 SD=1.05
Position	M=1.09 SD=.19	M=1.54 SD=.17	M=1.92 SD=.19	M=2.14 SD=.21	M=2.56 SD=.25	M=2.76 SD=.24	M=3.29 SD=.42	M=4.90 SD=.66	M=5.99 SD=.81

Table 5: Descriptive statistics.

5.1 Size of the target

It was observed that there was a statistically significant difference with respect to the factor "size". Subjects took more time to find the target in scenarios where the size of the target was set to low, compared to high or medium. This was the case with respect to all the three parameters that were taken into account. It was noted that users also had more number of previous fixations in the scenarios where the size was varied to low.

5.1.1 ANOVA results for parameter: Time to first fixation

A repeated measures ANOVA determined that mean "parameter" differed statistically significantly between the levels of factor *(F(2, 44) = 103, P <.05)*. Subjects have spent more time to fixate on targets of smaller size (*T(small): M = 1.89(SD = .22); T(medium): M = 1.38 (SD =.18); T(large): M = 1.31 (SD = .14)).*

The amount of time taken to fixate on the target increased sequentially as the size of the target was varied from high to medium to low. The total number of seconds taken to fixate on the target was high when the size was set to low compared to that of high or medium. This is depicted in Figure 11.

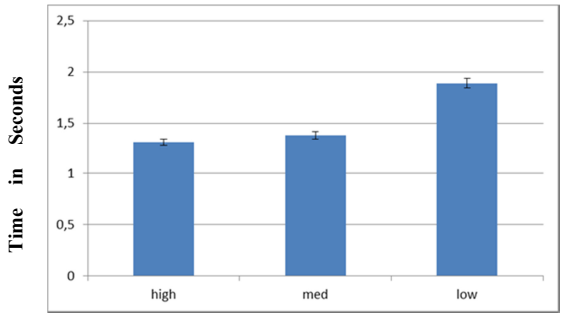

Figure 11: Size - Time to first fixation.

5.1.2 ANOVA Results for Parameter: Time to first mouse click

A repeated measures ANOVA determined that mean "parameter" differed statistically significantly between the levels of factor *(F(2, 44) = 323, P < .05).* Subjects have spent more time to mouse click on targets of smaller size (*T(low): M = 3.04 (SD =.24); T(medium): M =2.23 (SD =.21); T(large): M =2.13 (SD =.24)).*

Considering this second parameter, the total number of seconds taken to click on the target was higher when the size was set to low compared to that of high or medium.

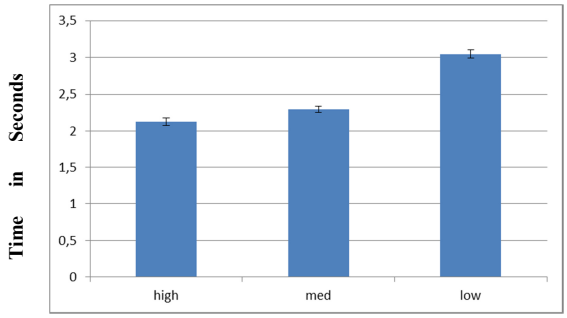

Figure 12: Size –time to first mouse click.

5.1.3 ANOVA results for Parameter: Number of previous fixations

A repeated measures ANOVA determined that mean "parameter" differed statistically significantly between the levels of factor *(F(2, 44) = 15.0, P < .05).* Subjects have spent more previous fixations in scenarios where the size was *low(T(low): M = 5.68 (SD = 1.79); T(medium): M =4.45 (SD = .72); T(large): M = 4.29 (SD = .52)).*

The total number of previous fixations made before fixating on the target was also higher when the size was set to low compared to that of high or medium.

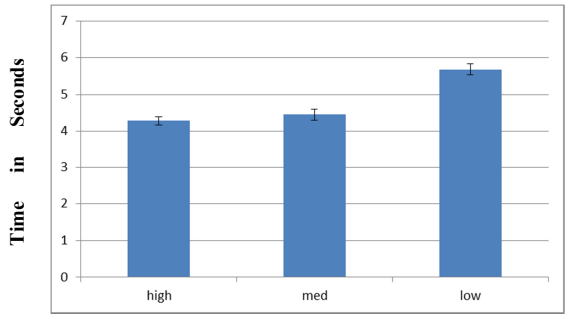

Figure 13: Size- Number of previous fixations.

5.2 Contrast

It was observed that there was a statistically significant difference with respect to the factor contrast; the users took more time to find the target in cases where the contrast of the target was set to low, compared to that of high or medium. This was the case with respect to all the three parameters that were taken into account. It was noted that users also had more number of previous fixations in the scenarios where the contrast was varied to low.

5.2.1 ANOVA results for Parameter: Time to first fixation

A repeated measures ANOVA determined that mean "parameter" differed statistically significantly between the levels of factor *(F(2, 44) = 31.6, P < .05).* Subjects have spent more time to fixate on target in scenarios of low contrast*(T(low): M = 1.77 (SD = .21); T(medium): M = 1,44 (SD = .22); T(large): M = 1.35 (SD = .17)).*

The total number of seconds taken to fixate on the target was high when the contrast was set to low compared to that of the contrast being high or medium. The amount of time taken to fixate on the target increased sequentially as the factor contrast was varied from high, to medium to low.

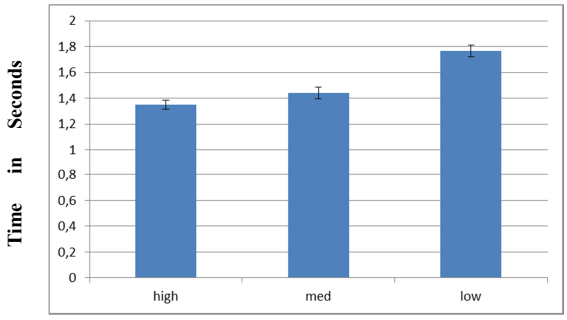

Figure 14: Contrast- Time to first fixation.

5.2.2 ANOVA results for Parameter: Time to first mouse click

A repeated measures ANOVA determined that mean "parameter" differed statistically significantly between the levels of factor *(F(2, 44) = 66.1, P <.05).* Subjects have spent more time to mouse click on the target where the contrast was low *(T(low): M = 2.71 (SD = .19); T(medium): M = 2.46 (SD = .25); T(large): M = 2.30 (SD = .21)).*

The total number of seconds taken to click on the target was higher when the contrast was set to low compared to that of high or medium.

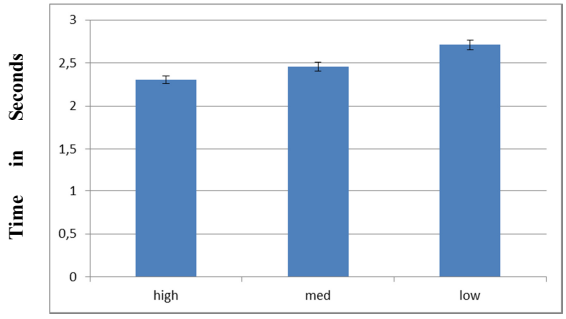

Figure 15: Contrast- Time to first mouse click.

5.2.3 ANOVA results for Parameter: Number of previous fixations

A repeated measures ANOVA determined that mean "parameter" differed statistically significantly between the levels of factor *(F(2, 44) = 4.89, P <.05).* Subjects have spent more previous fixations in cases where the contrast was set to low *(T(low): M = 4.99 (SD = 1.05); T(medium): M = 4.32 (SD = .44); T(large): M = 4.67 (SD = .70)).*

The total number of previous fixations made before fixating on the target was also higher when the contrast was set to low compared to that of high or medium.

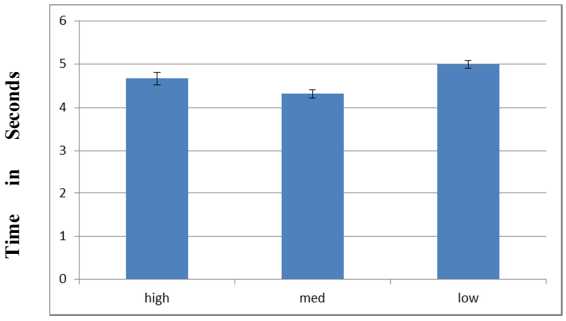

Figure 16: Contrast Number of previous fixations.

5.3 Position

It was observed that there was a statistically significant difference with respect to the factor position, the users took more time to find the target in cases where the position of the target

was set to low compared to that of high or medium. This was the case with respect to all the three parameters that were taken into account. Similarly, the number of previous fixation were also high where the position was set to low.

5.3.1 ANOVA results for Parameter: Time to first fixation

A repeated measures ANOVA determined that mean "parameter" differed statistically significantly between the levels of factor *(F(2, 44) = 184, P <.05).* Subjects have spent more time to fixate on the target in cases where the position was set to low *(T(low): M = 1.92 (SD =.19); T(medium): M = 1.54 (SD = .17); T(large): M = 1.09 (SD = .19)).*

The total number of seconds taken to fixate on the target was high when the position was set to low compared to that of high or medium.

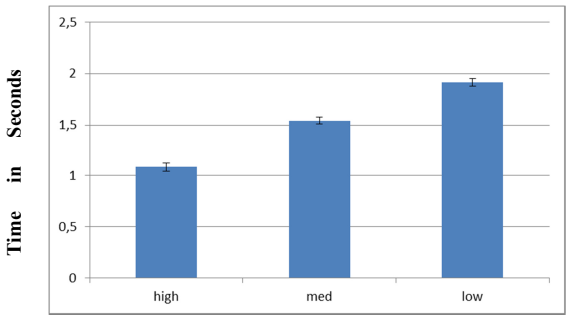

Figure 17: Position- Time to first fixation.

5.3.2 ANOVA results for Parameter: Time to first mouse click

A repeated measures ANOVA determined that mean "parameter" differed statistically significantly between the levels of factor *(F(2, 44) = 90, P <.05).* Subjects have spent more time to click on the target when the position was set to low *(T(low): M = 2.76 (SD =.24); T(medium): M = 2.56 (SD =.25); T(large): M = 2.14 (SD = .21)).*

The total number of seconds taken to click on the target was higher when the position was set to low compared to that of high or medium.

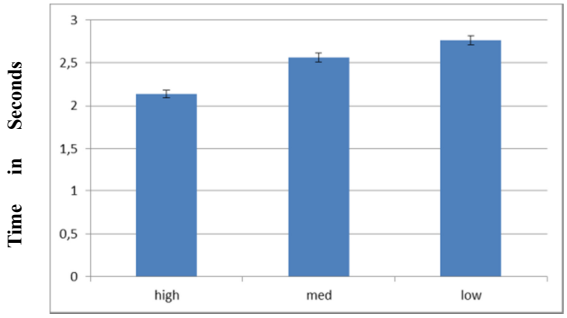

Figure 18: Position- Time to first mouse click.

5.3.3 Parameter: Number of previous fixations

A repeated measures ANOVA determined that mean "parameter" differed statistically significantly between the levels of factor *(F(2, 44) = 158, P <.05).* Subjects have had more previous fixations when the position was set to low *(T(low): M = 5.99 (SD = .81); T(medium): M = 4.90 (SD =.66); T(large):M = 3.29 (SD =.42))*.

The total number of previous fixations made before fixating on the target was also higher when the position was set to low compared to that of high or medium.

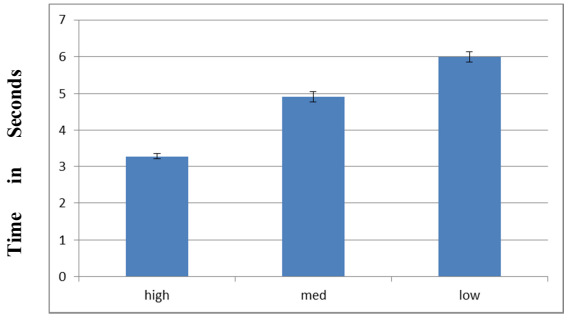

Figure 19: Position-Number of previous fixations.

5.4 Results

In conclusion it is observed that, for all the factors with respect to each of the considered parameters, the variation to low resulted in users taking a considerably longer amount of time to fixate on the target and to click on the target. Similarly, the number of previous fixations were also high in scenarios where the factors (size, contrast, position) were varied to low.

6 CONCLUSION AND FUTURE WORK

6.1 Conclusion

As the main result of the experiment and this thesis, it is concluded that in all the scenarios where the factors: size, contrast and position of the target object were set to low, it was harder for subjects to find the target. This was observed with respect to the three parameters: Time to the first fixation on the target, Time to the first mouse click on the target, Number of previous fixations before fixating on the target. There was also an increased number of previous fixations for these scenarios as well.

6.2 Future Work

This thesis analyzed the environmental parameters in the VR world and sequentially varied those parameters in order to determine the significance of its influence on visual perception in that regard. However, the work that was performed in this thesis was solely on the basis of static scenes or merely plain images.

Hence, the future work is planned in the following directions:

- A concrete model could be developed which could help to be certain of whether an object within an environment was seen by the user or not.
- Analysis could be performed on other factors such as speed, which can be obtained from a dynamic or a user controllable environment.

Bibliography

Bahill, A.Terry, Michael R. Clark, and Lawrence Stark. 1975. "The Main Sequence, A Tool for Studying Human Eye Movements." *Mathematical Biosciences* 24: 191–204. doi:10.1016/0025-5564(75)90075-9.

Borji, Ali, Dn Sihite, and Laurent Itti. 2012. "An Object-Based Bayesian Framework for Top-Down Visual Attention." In *AAAI Conference on Artificial Intelligence*, 1529–35. doi:10.3389/fpsyg.2012.00151.

Corbett, Albert T., Kenneth R Koedinger, and John R. Anderson. 1997. "Intelligent Tutoring Systems." *Science (New York, N.Y.)* 228 (4698): 456–62. doi:10.1126/science.228.4698.456.

Desai, Parth Rajesh, Pooja Nikhil Desai, Komal Deepak Ajmera, and Khushbu Mehta. 2014. "A Review Paper on Oculus Rift-A Virtual." *International Journal of Engineering Trends and Technology (IJETT)* 13 (4): 175–79. doi:10.14445/22315381/IJETT-V13P237.

Desimone, Robert, and John S Duncan. 1995. "Neural Mechanisms of Selective Visual Attention." *Annual Review of Neuroscience* 18 (1): 193–222. doi:10.1146/annurev.ne.18.030195.001205.

Duchowski, Andrew T., Eric Medlin, Nathan Cournia, Anand Gramopadhye, Brian Melloy, and Santosh Nair. 2002. "3D Eye Movement Analysis for VR Visual Inspection Training." In *Proceedings of the Symposium on Eye Tracking Research & Applications - ETRA '02*, 103. doi:10.1145/507072.507094.

Duchowski, Andrew T., Eric Medlin, Anand Gramopadhye, Brian Melloy, and Santosh Nair. 2001. "Binocular Eye Tracking in VR for Visual Inspection Training." In *Proceedings of the ACM Symposium on Virtual Reality Software and Technology - VRST '01*, 1. doi:10.1145/505009.505010.

Fuchs, Henry, Zvi M Kedem, and Bruce F Naylor. 1980. "On Visible Surface Generation by a Priori Tree Structures." *SIGGRAPH Comput. Graph.* 14 (3): 124–33. doi:10.1145/965105.807481.

Gobbetti, E, and R Scateni. 1998. "Virtual Reality: Past, Present and Future." *Studies in Health Technology and Informatics* 58: 3–20. doi:10.3233/978-1-60750-902-8-3.

Gu, Yecheng, Sergey Sosnovsky, and Carsten Ullrich. 2015. “SafeChild: An Intelligent Virtual Reality Environment for Training Pedestrian Safety Skills.” In *Design for Teaching and Learning in a Networked World: 10th European Conference on Technology Enhanced Learning*, 141--154. Springer International Publishing. doi:10.1007/978-3-319-24258-3_11.

Hayhoe, Mary, and Dana Ballard. 2005. “Eye Movements in Natural Behavior.” *Trends in Cognitive Sciences* 9 (4): 188–94. doi:10.1016/j.tics.2005.02.009.

Henderson, John M, and Andrew Hollingworth. 1999. “High-Level Scene Perception.” *Annual Review of Psychology* 50: 243–71. doi:10.1146/annurev.psych.50.1.243.

Horvitz, Eric, and Jed Lengyel. 1997. “Perception, Attention, and Resources: A Decision-Theoretic Approach to Graphics Rendering.” In *1997, Proceedings of the 13th Conference on Uncertainty in Artificial Intelligence (UAI-97*, 238–49. https://eprints.kfupm.edu.sa/57448/1/57448.pdf.

Hwang, Alex D., Hsueh Cheng Wang, and Marc Pomplun. 2011. “Semantic Guidance of Eye Movements in Real-World Scenes.” *Vision Research* 51 (10). Elsevier Ltd: 1192–1205. doi:10.1016/j.visres.2011.03.010.

Itti, Laurent, Christof Koch, and Ernst Niebur. 1998. “A Model of Saliency-Based Visual Attention for Rapid Scene Analysis.” *IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE* 20 (11): 1575–80. doi:10.1109/TPAMI.2012.125.

Itti, Laurent, Christof Koch, Watt Way, and Los Angeles. 2001. “Computational Modelling of Visual Attention.” *Nature Reviews. Neuroscience* 2 (3): 194–203. doi:10.1038/35058500.

Joubert, Olivier R, Denis Fize, Guillaume a Rousselet, and Michèle Fabre-thorpe. 2008. “Early Interference of Context Congruence on Object Processing in Rapid Visual Categorization of Natural Scenes.” *Journal of Vision* 8 (13): 1–18. doi:10.1167/8.13.11.Introduction.

Koch, Kristin, Judith McLean, Ronen Segev, Michael A. Freed, Michael J. Berry, Vijay Balasubramanian, and Peter Sterling. 2006. “How Much the Eye Tells the Brain.” *Current Biology* 16 (14): 1428–34. doi:10.1016/j.cub.2006.05.056.

Matsumoto, Y., and A. Zelinsky. 2000. “An Algorithm for Real-Time Stereo Vision

Implementation of Head\npose and Gaze Direction Measurement." In *Proceedings Fourth IEEE International Conference on Automatic Face and Gesture Recognition (Cat. No. PR00580)*, 499–504. doi:10.1109/AFGR.2000.840680.

Rothenstein, Albert L., and John K. Tsotsos. 2008. "Attention Links Sensing to Recognition." *Image and Vision Computing* 26 (1): 114–26. doi:10.1016/j.imavis.2005.08.011.

Sutherland, Ivan E. 1996. "The Ultimate Display." In *Proceedings of the Congress of the Internation Federation of Information Processing (IFIP)*, 21:506–8. doi:10.1109/MC.2005.274.

Treisman, Anne M, and Garry Gelade. 1980. "A Feature-Integration Theory of Attention." *Cognitive Psycology* 12 (1): 97–136. doi:10.1016/0010-0285(80)90005-5.

Triesch, J, DH H Ballard, M M Hayhoe, and B T Sullivan. 2003. "What You See Is What You Need." *Journal of Vision* 3 (1): 86–94. doi:10.1167/3.1.9.

Zeng, Wenjun. 2012. "Microsoft {K}inect Sensor and Its Effect." *IEEE Computer Society* 19 (2): 4–10. http://dx.doi.org/10.1109/MMUL.2012.24.

List of Tables

LIST OF FIGURES