12-31-2017

Gaze-Contingent Displays for Interactive Text Enhancement

Divya Bajaj
University of Massachusetts Boston

GAZE-CONTINGENT DISPLAYS FOR INTERACTIVE TEXT ENHANCEMENT

A Thesis Presented by

DIVYA BAJAJ

Submitted to the Office of Graduate Studies,
University of Massachusetts Boston,
in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

December 2017

Computer Science Program

GAZE-CONTINGENT DISPLAYS FOR INTERACTIVE TEXT ENHANCEMENT

A Thesis Presented

by

DIVYA BAJAJ

Approved as to style and content by:

Marc Pomplun, Professor
Chairperson of Committee

Dan A. Simovici, Professor
Member

Craig Yu, Assistant Professor
Member

Dan A, Simovici, Program Director
Computer Science Program

Peter Fejer, Chairperson
Computer Science Department

ABSTRACT

GAZE-CONTINGENT DISPLAYS FOR INTERACTIVE TEXT ENHANCEMENT

May 2017

Divya Bajaj, B.Tech., Punjab Technical University
M.S., University of Massachusetts, Boston

Directed by Professor Marc Pomplun

Eye trackers are used for measuring a person's eye movements, for example, while reading text on a screen. This information can be used by scientists to study the human visual system during object recognition or text comprehension. Moreover, engineers can build gaze-controlled interfaces that trigger a variety of actions when the user looks at pre-specified icons on a screen. In my research, I studied how gaze-contingent displays can be used to enhance the information that is provided by texts. First, I implemented a simple scripting language that allows even users without programming experience to set up gaze-contingent text displays. The language allows to display any text on the screen and define keywords that trigger actions when the user looks at them. These actions can either be a specific sound file being played or a specific bitmap image being displayed at a given position on the screen. Second, using this scripting language, I conducted an experiment on 15 users. They saw displays of 20 written words, which they had to memorize. In a control condition, the display would be static. In another condition, whenever they looked at a word, that word would be spoken, in a third condition, an image associated with the word would be shown, and in a fourth condition, both effects would occur at the same time. Surprisingly, memory performance was reduced by all gaze-contingent effects, whereas subjects believed that especially the image condition was helpful for memorization. The results suggest that gaze-contingent text enhancement is appreciated by its users but, instead of presenting identical information in different forms, should provide additional information related to the attended words.

ACKNOWLEDGEMENTS

I would like to thank all those who have helped me throughout this thesis to complete it and achieve the desired results. Firstly, I would like to thank Professor Marc Pomplun, my advisor, for giving me this great opportunity to do a thesis on the selected topic. I am thankful to him for mentoring me and making it possible for me to complete the thesis, learn and explore many new things in this field during my Master's degree studies. I am really thankful to Professor Dan Simovici for believing in me and permitting me to do the thesis. While thanking professors I could not forget Professor Craig Yu for serving on my thesis committee and suggesting some ideas and real-life problems to which my research work could be applied.

At last I would also like to thank my friends who have always supported me and given me ideas throughout the thesis work. I would also like to thank my parents and family members for supporting me in all my decisions and helping and encouraging me to pursue my Master's degree.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

"It has been said that 80% of what people learn is visual" - Allen Klein.

The eyes, together with about half of the neurons in the human brain, are the crucial organ which provides vision. Vision, in turn, is crucial for identifying objects in this world or recognizing patterns such as words in a text when reading. Humans move their eyes towards a point of interest in order to see that point in the fovea. The fovea is the central part of the retina which contains the highest concentration of our high-acuity, color-sensitive photoreceptor cells, which are called cones. Consequently, in this region, objects can be seen with the highest resolution.

1.1 EYE MOVEMENTS

Humans usually move their eyes while inspecting their surroundings or while reading. Because only the center of the retina, called fovea, provides high resolution of the visual input, the eye needs to be moved quickly and often to identify different objects in the visual field. This motion is termed eye movement. Through these eye movements, the eyes act as a window between the visual scene and the brain. The eyes receive light patterns from

objects in the scene and their retinal neurons create electrical signals that are then transferred through optic nerves to the visual system in the brain for further processing.

1.1.1 TYPES OF EYE MOVEMENTS

Eye movements are classified into five main types, namely saccades, fixations, vergence movements, smooth pursuit movements and vestibulo-ocular movements:

1. Saccade - Fast motion of both the eyes from one position to another in the same direction is known as a saccade. It is used to quickly scan pieces of information in the visual field by sequentially moving their images into the fovea.
2. Fixation – During a fixation, the eyes are gazing at a particular position. This mainly occurs between saccadic movements. Information from the visual field is mainly obtained and processed during fixations, when the eyes are stable.
3. Vergence - This is the movement between the visual axes of the two eyes. It is commonly used to change the gaze target from a far to a near object or vice versa.
4. Smooth Pursuit - Smooth pursuit movement is the motion in which the eyes follow a particular moving object. This motion is smooth as the eyes try to catch a stable view of the object. Human cannot make this type of eye movement without a moving object being present.
5. Vestibulo-Ocular - In this movement, the absolute position of the eyes in space remains stable despite concurrent head movements. During the horizontal, vertical or torsional motion of head, the eyes rotate around the same axis but in opposite direction to that of the head in order to compensate for the head movement and provide stable vision.

Without eye movements, our eyes would be much less useful, or we would have to use head movements instead, which are much slower and require substantially more effort than eye movements. Therefore, loss or abnormalities of eye movements are serious problems.

1.1.2 EYE MOVEMENT DISORDERS

Some disorders related to eye movements are as follows:

1. Nystagmus - A person is said to have nystagmus when he/she has difficulty maintaining stable fixations. Their gaze often keeps drifting away from the inspected object and has to be frequently realigned by saccades. It can lead to poor vision, disturbing one's normal life.
2. Strabismus - This disorder affects approximately 4% of the population in the U.S [1]. In this disorder, the direction of one or both the visual axes differs from the intended gaze position.

Most abnormalities are recognized while going through clinical examination but some of them may require recording of the eye movements for their diagnosis. This tracking is performed by using an eye tracker.

1.2 EYE TRACKERS

An eye tracker is a piece of equipment which is used to determine a person's gaze position in real time. The eye tracker tracks the position of the pupil within the eye, and this information is used to determine the gaze direction and also the sequence in which the eye moves [2].

With the great advancements in the field of technology, eye trackers are now available as mobile devices. They can, therefore, assist the researchers to collect gaze data in the field which was not possible earlier, when they could only record data in research laboratories. Mobile eye tracking also helps, for example, in investigating the behavior of consumers in the supermarket. Eye tracking gives the opportunity to look through the consumer's eyes. So, instead of only listening to the opinion of the customer, it is now possible to perform an analysis with tools such as the heatmap and gaze plot. It helps supermarkets to analyze how people respond to different products, which products are most attractive and why. Using such analyses, they can work on improving and growing their sales.

Eye-tracking techniques have also been used for collecting data from the standpoint of conducting empirical studies of human perception, cognition, and behavior. They are commonly used in fields like psychology, VR research, psycholinguistics, etc.

1.2.1 HISTORICAL BACKGROUND

In the 17th century, eye movements were studied by direct observations. People were interested in knowing and understanding the motion of eyes while reading, watching, talking, etc. In the 19th century, a French ophthalmologist named Louis Emile Javal described that humans never moved their eyes continuously or regularly while reading some text. Instead, they perform little movements very frequently together with minimal stops [3]. These brief stops are known as fixations, and the movements between two points are known as saccades. As noted above, this saccade-fixation behavior is typical for

inspecting static scenes or texts. Only if the eyes are trying to follow a moving target, then they move in smooth pursuit.

In the late 19th century and until the mid-20th century, researchers focused on eye tracking systems employing the "old" tracking method. This type of eye tracker was first built by Edmund Huey [4]. He used a special contact lens which had a hole for the pupil. Movement of eyes were recorded by an aluminum pointer which was attached to the lens. Huey showed that few saccades are classified as regressions, i.e., are returning to previously scanned words, and not all the words are fixated in a sentence.

The general, fundamental understanding regarding eye movements was retrieved within this period. With the advancements in technology in the late 20th century, experts used contact lenses and video for tracking and studying the motion of the eyeballs. Three significant improvements achieved in the mid of the 20th century were:

1. More precise and convenient tracking equipment for the eye movements,
2. Enhancement of eye tracker performance & monitoring power using computer-based techniques, and
3. Processing and keeping the record of a considerable amount of information generated by eye movement recording.

1.2.2 TYPES OF EYE TRACKER

Eye trackers are classified into three main categories:

1. Eye-attached tracking – In this type, tracking of movement of eye positions is done by using magnetized contact lenses which are tracked by magnetic sensors.
2. Optical tracking – The projector in the (remote) eye tracker projects near-infrared light onto the eye, which is then reflected back onto the sensors of the eye tracker. These sensors are responsible for taking high-resolution pictures of the eye. An image processing algorithm is then applied to detect the pupil in the camera image. Finally, from the pupil position the gaze position on a screen is computed.
3. Electric potential measurement – This type of tracking is achieved by positioning electrodes on the skin around the eyes. These electrodes measure the electric potentials of the eye muscles that generate eye movements. This information is then used to compute the eye movement parameters.

1.2.3 TECHNOLOGY AND TECHNIQUES

Video-based trackers are the most widely used eye trackers; here the movements of one or both eyes are recorded for one or more stimuli presented on a computer screen. These trackers often use non-collimated infrared or near infrared to create corneal reflections, i.e., specular reflections from the front of the eyeball. Using the vector between pupil center and corneal reflection center for eye tracking is more robust than using pupil position alone, because it is nearly invariant to slight movements of the subject's head. Proper functionality of the system is achieved by calibration, in which the subject typically has to visually track a marker that moves to different positions on the screen. For each position,

the pupil position (or pupil-corneal reflection vector) is measured as "ground truth." Afterwards, interpolation techniques are used to determine the gaze position for any newly measured pupil position or difference vector.

Video-based eye trackers typically consist of a camera, light emitters and a system for image processing. The monitoring of eye movements is achieved by a computer vision system. This system identifies the pupil and possibly the corneal reflection in the camera image and achieves robust tracking performance by comparing this information with previously captured frames.

There are mainly two types of technique used for eye tracking: Active vs. passive eye tracking. In active eye tracking, typically infrared (IR) light is projected onto the eye so that a high-contrast image can be obtained. Humans are not able to see IR which makes it less distracting and a good choice to use in experiments. Furthermore, for indoor experiments, it eliminates most of the ambient light in the room. This helps in maintaining consistent environmental conditions which minimize the number of variables that potentially interfere with accurate eye tracking. On the other hand, it makes it difficult to use outdoors because of interference by sunlight, which includes IR light in a wide range of wavelengths. IR is also a dangerous choice. It can cause serious damage, if eyes are exposed to high-intensity IR light for a long duration. However, in industrial eye trackers, the IR light intensity is restricted to be harmless even for long experiment durations. Passive eye tracking, on the other hand, simply uses natural sunlight for illuminating the

eye. This method does not require an illuminator but leads to greater variability in lighting conditions that often reduces the reliability of gaze measurements.

Further, active light techniques are classified into bright and dark pupil techniques. The eye performs as a retroreflector if the illumination and optical path are coaxial. Therefore, a bright pupil is created when the light reflects off the retina. On the other hand, when they are not coaxially aligned, we see a dark pupil.

Bright pupil tracking is often preferred over dark pupil tracking as it creates a better contrast between the iris and the pupil and also decreases other forms of interference. But there are limitations as well. It cannot be used for outdoor activities as it gets affected by external IR to create artefacts or misleading results.

1.2.4 APPLICATIONS OF EYE TRACKING

Eye tracking is used in many different areas such as psychology, neuroscience, industrial engineering, computer science, educational research, etc. Researchers use the technique of eye-tracking to study eye movements of a person while they are engaged in various actions.

The interest in applying eye tracking (ET) methods increased with technological progress and rising ET performance and accessibility. Eye trackers have existed for several decades, but their use did not exceed laboratory experiments. Now, the devices are becoming reliable and affordable enough to re-consider their application in real-world Human

Computer Interaction (HCI). Many recent studies have focused on appropriate interaction techniques that accommodate gaze movements into HCI conveniently and rationally [5].

Eye Tracking and Assistive Technology

Many people all around the world have some form of disability. To work and live normally, they require some specialized equipment or assistance. This type of equipment is collectively termed Assistive Technology. One of the typical examples is a software which converts text to speech. This software is commonly used by people with vision and speech disabilities. The technique of eye tracking is also embedded in many assistive devices to enhance their users' work. They are widely used by neuro-disabled patients to communicate by moving their eyes. For instance, they can look at a virtual keyboard and can trigger different keys by looking at them for a minimum duration. Eye tracking can also be used along with a computer to select an expression from a menu. For example, while examining a patient with a neurological disorder, it could be used for a face-to-face conversation. The patients use this device by creating a remote message via a communication network. Through that message, the examiner can diagnose the stage of illness in the patient.

New interaction techniques can be implemented by combining eye moments with brain-computer interfaces (BCIs). BCIs are devices that are used to send commands from the neurons in the brain to a computer or vice-versa. BCIs primarily focus on augmenting, assisting, and repairing human cognitive/sensory-motor functions. Serious research on BCI began in the 1970s at the University of California Los Angeles (UCLA) and was focused

initially on applications of neuroprosthetics that concentrate on re-instating damaged sight, hearing and movement [6].

Use of Eye Tracking in Neuroscience and Psychology

The mobility of the eye is a delicate function that is related to the central nervous system. Hence, the disorders and illnesses that affect the brainstem, the cerebral cortex or the cerebellum often have a substantial effect on eye movements. Consequently, the analysis of dysfunction of eye movements can provide information about which part of the brain is damaged. It is also a reliable marker for dementia and a considerable number of other diseases related to the brain. Eye movements in healthy subjects can inform researchers how they shift their visual attention when performing a given task. This helps to better understand the functioning of the human visual system, and it can also inform the development of new computer vision systems, which are still inferior to human vision in many aspects.

E-learning and Eye Tracking

In the last several years, different technologies, for instance, collaborative software, screen casting, cloud computing, e-portfolios, virtual classrooms and various devices like mobile devices, webcams, audio/video systems and smart boards have been used to facilitate e-learning development. They are also utilized to increase the effectiveness and accessibility of e-learning platforms. Some previous studies revealed that eye tracking methods could improve the effectiveness and usability of e-learning systems [7].

Eyes Tracking and Object Selection

Object selection by eye movements is challenging. The problem in gaze-controlled HCI is that it easily tells the computer what icon, letter, or button the user is looking at. However, it is difficult for an interface to decide whether the user actually wants to activate that function or is just looking at it to inspect the options that are available, which can lead to inadvertent selections and triggering of interface functions, known as the "Midas touch problem." It could be resolved by having the user blink to trigger the action associated with the item they are looking at, or by setting a minimum dwell time needed to trigger the event. However, such blinking is unnatural and often found to be tedious by users.

In one study [8], object selection was studied in two different ways – using a button and using a dwell time threshold. In the first approach, the user gazes at the object that he or she wants to select and the presses a button to conform the selection. In the second approach, to select an object using dwell time, the user is required to gaze an object for a sufficiently long duration that can be adjusted for individual users. The study finds the dwell time approach to be much more accessible and preferable provided that the dwell time could be made brief for experienced users.

1.3 VISUAL WORKING MEMEORY

The term Visual Working Memor*y* is used to explain the storage and handling of visual information in human memory. It is an essential function to perceive the identity of objects and observe where these objects are situated in space at any given point of time. The capacity for holding items in visual working memory is roughly 4 visual items for adults,

whereas 3-year olds have the ability of 1.3 visual items and 4 year olds have the capacity of 1.8 visual items [9]. Visual working memory is directly proportional to age during early development.

1.3.1 WORKING MEMORY COMPONENTS

A popular model of Working Memory, the multicomponent model, consists of three elements which use the information of what the eyes see and ears hear to perform actions [10].

1. Phonological loop: This is how a person keeps sounds active in working memory by "replaying" them in their mind.
2. Visuospatial sketchpad: This is a crucial part of working memory which is accountable for a person's short-term storage of visual and spatial information. It is also responsible to store the location or speed of visual elements.
3. Central Executive Function: It controls the information flow from and to phonological loop and visuospatial sketchpad. This helps people to self-monitor their initial reactions to the environment.

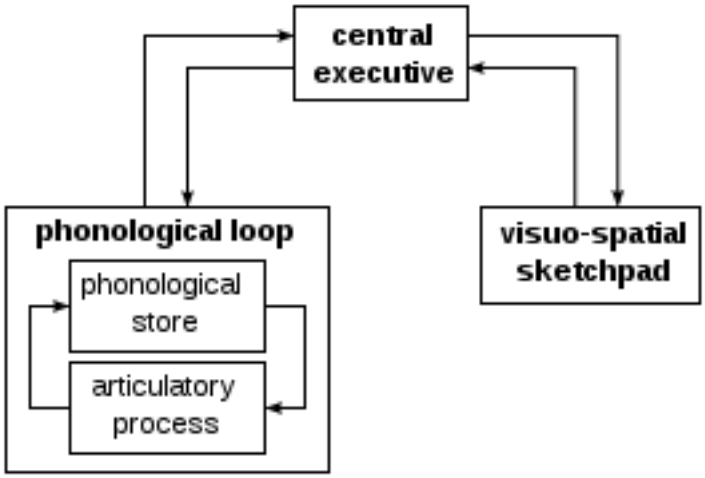

Figure 1 - The working memory model [11]

1.3.2 WHY IS VISUAL WORKING MEMORY IMPORTANT?

A person's visual working memory underlies their capability to recall the specifications of any given object, symbol or form etc. When it gets difficult for someone to remember the information and store it in their working memory, it also gets complicated for them to process that information and transition it into their long-term memory. While reading, it is very difficult for a person to remember previously read words and understand complex sentences if they are struggling with their visual memory skills. For such tasks, Visual Sequential Memory is particularly important, which is a person's ability to recall symbols, forms or objects in a certain order. In general, a person who has poor visual working memory might also have difficulties with remembering phone numbers or keep track of the locations of objects and steps taken in a complex task.

1.3.3 THE ROLE OF VISUAL WORKING MEMORY IN VISION

Visual working memory tracks the locations and identities of moving objects, which allows people to maintain temporal continuity in a time changing environment. During short periods of fixation, visual information is acquired. These periods are separated by saccadic eye movements which suppress visual processing while shifting the retinal image in order to enable a stable perception of the visual scene. For this purpose, some kind of memory is needed to match the information perceived before a saccade with that perceived afterwards. Recent studies show that visual working memory automatically stores the target of an upcoming eye movement and afterwards compares this information with the freshly fixated object to allow a seamless integration of the new visual information into the perception of the visual environment.

Furthermore, eye movements tend to be aimed toward the objects which match the current content of visual working memory. If the saccade target matches the current content of visual working memory, then the saccade is executed even faster. Visual working memory might not be really a memory system but it can act as a visual representation system which can store information over short delays.

1.4 GAZE-CONTINGENT DISPLAYS

Gaze-contingency describes a technique in which a screen changes in response to where a subject is looking. Gaze-contingent displays try to balance out the capacity of information being displayed against the visual information processing on the observer's side through real time eye-movement sensing.

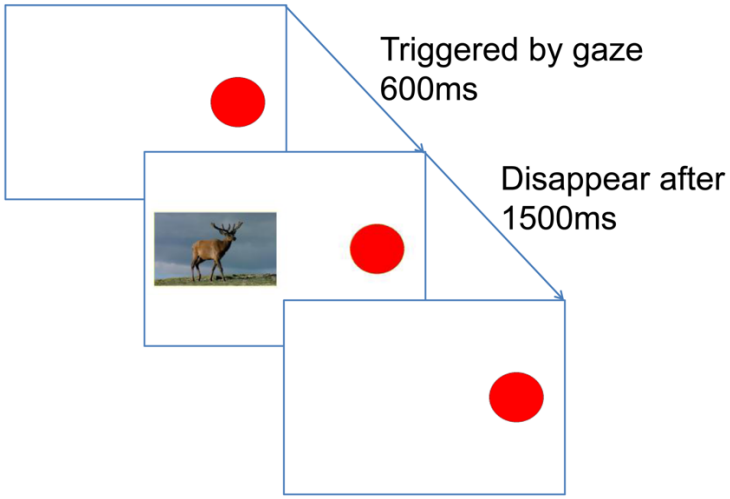

Figure 2 – Gaze contingent display that shows the picture of an animal for 1500 ms when the user looks at the red dot. Here, a dwell time threshold of 600 ms is chosen to prevent inadvertent triggering of this action [12].

Figure 2 shows an example of a gaze-contingent display. When the red dot is fixated for 600 ms, the image of an animal pops up and disappears after another 1500 ms. A gaze-contingent display's content can be modified by different display processing techniques based on the information of the location where the observer is focusing. These displays are used, for instance, to minimize the bandwidth in video telephonic applications or in graphical applications where data sets are way too complex to be displayed all at once. There are two types of gaze-contingent applications: one is screen-based and the other is model-based. The screen-based gaze-contingent applications deal with pixel handling whereas model-based applications deal with the manipulation of graphical objects or models.

1.4.1 SCREEN-BASED DISPLAYS

It is essential to differentiate between the causalities that are affecting perception and those which are affecting performance while doing the evaluation of gaze-contingent displays. To enhance perception it can be possible to degrade the display to the point where its effect becomes unnoticeable and which also does not degrade the performance. Then the smallest window that still allows task performance with normal efficiency indicates the size of the visual span, i.e., the area around fixation from which information is obtained. Most of the gaze-contingent displays lag behind with respect to user's gaze by a small, constant time. This lag depends on the time it takes for a gaze position to be measured, which is 16.7 ms for a 60 Hz video-based tracker, plus the subsequent time to refresh the gaze-contingent display, which can be up to 8.3 ms for a 120 Hz LCD display. Therefore, both eye trackers and monitors with high fame rates are preferable for gaze-contingent applications.

1.4.2 MODEL-BASED DISPLAYS

Model-based displays typically reduce the resolution of computer-rendered displays in those areas that are not currently attended by the viewer, as determined by an eye tracker. These changes can be made because they are not perceived by the user. They are achieved by making changes to the geometry model before rendering, often using Level of Detail (LOD) rendering wherein an object's screen area coverage is used as a metric to select one of several pre-computed reduced resolution meshes for display [13]. Applications like animation rendering and virtual reality use this technique. It is also used in object simplification by reducing resolution meshes.

1.4.3 EYE-MOVEMENT ANALYSIS FOR ENHANCING MEMORY

Eye movements are of great interest while viewing advertisements including on TV and websites as they provide very useful information regarding the patterns of visual attention:

1. Television Advertisements**:** Advertisements getting displayed regarding various drinks during sporting events are not greatly effective to catch the viewer's attention or influencing subsequent memory. Advertisement skipping occurs when a traditional advertisement is being displayed in between parts of the primary event shown on TV. Advertisements which include subtitles tend to improve memory for particular pieces of information provided in the advertisement. Eye-movement analysis can reveal which advertisement forms and placements are most effective at catching the viewer's attention.

2. Internet Advertisements**:** Advertisements that appear in the form of banners on websites are generally avoided by the viewers. The significance of internet advertisements or their relationship with the content appearing in the advertisement does not affect the viewers who are viewing a particular item or piece of information on the website as they seem to be irrelevant to them. Therefore, the memory for the advertisements getting displayed on the Internet are poor. Eye tracking can provide insight into the effectiveness of banners at getting the viewer's attention.

There are various factors which control the eye movements with regard to advertisements. One such factor is the visual characteristics of the advertisement (e.g., size or color) and what the viewer is looking for in the advertisement. Facial images in online banner advertisements or TV have a very powerful impact as they maximize the viewer's attention to the information provided in the advertisement. This further increases the viewer's ability to remember the contents being displayed in the advertisement like the name of the brand and the message which the advertisement is trying to convey. One famous example – although no advertisement but rather the opposite - are the written warning labels on the tobacco or alcohol advertisements, which generally used to be black and white text. However, once these warnings are replaced by graphical images, the attention of the viewer and the memory regarding such warnings are enhanced.

1.5 MOTIVATION OF THE PRESENT STUDY

Given that it is possible to change the content of a display in real-time based on the user's current gaze position, it should also be possible to build interactive, gaze-controlled human-computer interfaces that enable better memorization of information than static displays do. One possible way to achieve this would be to enhance a text display by showing the same information given in the text also in visual (images) or auditory (spoken words) form, or both. Here, the eye tracker can indicate which word the user is attending to so that the additional resources for that word are shown at that time. This way the user is free to explore the display in any order and at any pace he or she desires, while always receiving multimodal information for only the currently relevant word.

In the following Chapter 2, I will review relevant literature indicating how such an interface could plausibly enhance memory performance. Subsequently, I will describe how I prepared the study and generated tools for other researchers to build similar interfaces and conduct related research. Finally, I will report how the experiment was conducted.

CHAPTER 2

BACKGROUND AND METHOD

2.1 LITERATURE REVIEW

The chapter discusses the existing scholarship in the field of multimodal interactive systems and how they link to memorization of information. Memory is responsible for storing our day-to-day instances of events and previous ones encountered in our life. Whatever we see or hear may stay in our mind for a long time or vanish away immediately. Depending on its type and importance, information is stored in one of the following memory systems (see also Figure 3):

1. Immediate memory: The information is stored only for a few milliseconds in this memory. For example, whenever we look around and do not pay attention, then the visual points that we saw almost immediately vanish from the memory.
2. Working memory: It is used to store information for a relatively short duration, usually only until the time we reuse this information, for example, until a task is completed.
3. Long-Term memory: It contains different facts, events or stories which are important to us beyond the current task or situation and are closely attached to our life.

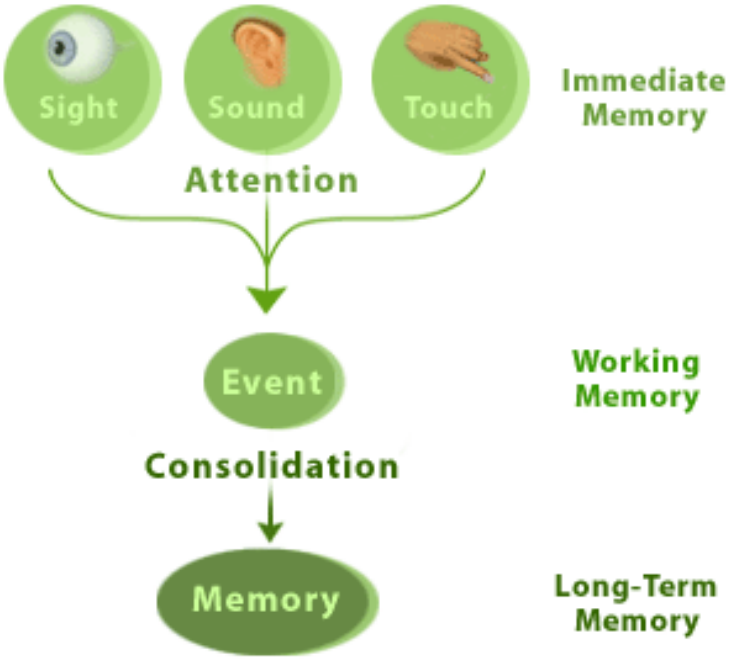

Figure 3 - Types of Memory [14]

Information is not something that we can touch or detect with a machine. It is all around us. Humans have created languages, pictures, art or mathematics to help convey information among each other. For instance, if there are three t-shirts of different colors and you are asked to identify my favorite t-shirt, then it would be a difficult task for you. But, if you know my favorite color, you could probably give it a try and answer correctly. Here, knowing my favorite color acts as a piece of information. This is only a simple example that represents the significance of information in our life.

Information is widely represented as text, image, and audio. A single picture might have a lot of information to depict. It is widely accepted that an object studied through an image

is better remembered than an object considered through a verbal expression. There is a well-known expression regarding this, "An image is worth a thousand expressions" [15]. Kids understand and enjoy stories from books which contain more images than text. The universal nature of some images makes them the open source of knowledge. For instance, even though a traveler cannot understand the local language, he or she can nonetheless find the ladies' or men's room practically in any country by looking at images/markers indicating this information. Several other reasons are illustrated in Figure 4.

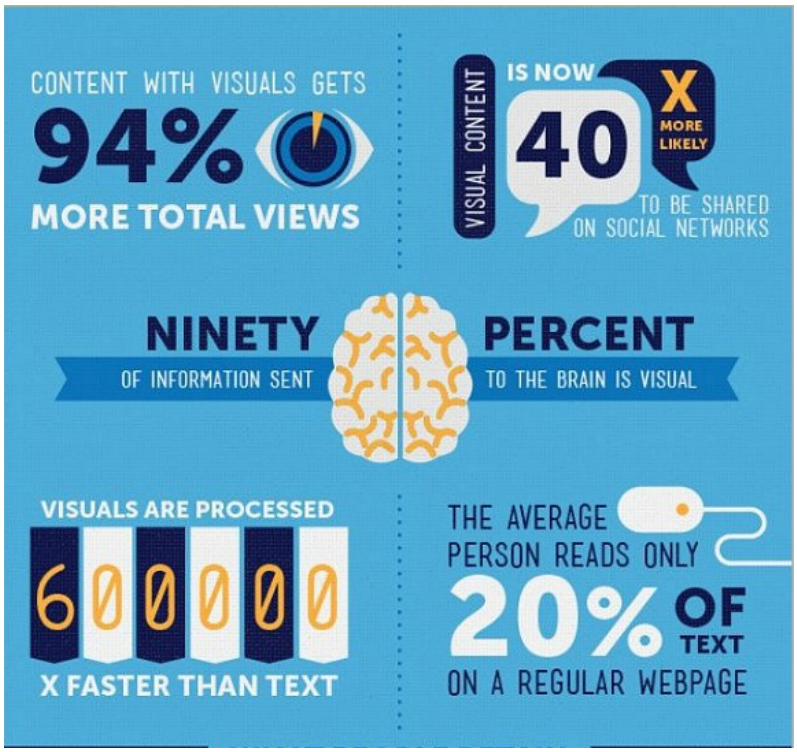

Figure 4 - Why visual content is better than text [16]

MIT neuroscientists discovered that the human brain can process images that the eyes see for as little as 13 milliseconds [17]. This is much faster than a person's blink of an eye

which is 300-400 milliseconds long. Visual images typically stay in our memory for a longer time than textual expressions because our brains have evolved to process visual images as the primary source of information. In fact, about half of the neurons in the human brain are dedicated to visual processing.

Auditory memory processing plays an important role when text is read out loudly. While reading, memory links are made using visual pathways. We remember an object because we have already seen it. People with photographic memory are particularly good at making such connections. For other people, however, completely relying on the visual memory leads to suboptimal memorization because it leaves the memory for other modalities unused. Therefore, finding alternative ways of remembering things is important. While reading out loud, memory pathways form auditory links along with the visual links enhancing memory capacity. People who prefer memorization this way are generally called auditory learners.

Moreover, in terms of auditory information, audio is better for memorization than written text. For instance, in some websites, when we try to log in using our credentials, the website asks for answering a captcha. This captcha is represented either in a textual or auditory form. Using audio is more comfortable and easier to understand for the user than text, which requires more effort and can sometimes be confusing. Also, while text chatting, writing a long text takes enormous efforts whereas conveying that message through verbal notes can make the task easier and more efficient.

It is important to notice that the brain processes various information modalities in different ways. It stores the audio information in a more temporary way whereas it processes the visual information in a different way that is more suitable for long-term memorization. The first study showing that memory works better with images than expressions was reported in the late 19th century by Kirkpatrick [18]. Later, Standing showed that people could remember thousands of distinctive images with great accuracy [19]. Perceptually, expressions are weaker than images. With this visual difference, images have an advantage in terms of memory storage.

However, there are some situations where expressions can be memorized more easily than images. The verbalization technique serves as a more effective, easier and faster way in these situations. These techniques tend to work faster because there is no need for encoding the information. To make the information more meaningful, techniques like acrostics, acronyms or rhymes are used. For example, most of us can imagine how a rainbow looks in our head, but we might not be able to recall the sequence of colors in it. However, remembering the name Vib-g-yor is a verbal technique. Using this technique, we can easily remember all the colors in order: Violet, Indigo, Blue, Green, Yellow, Orange, and Red.

2.2 TECHNICAL PREPARATION

A scripting language was created for gaze-contingent displays. Programs in this language are ASCII files with the extension ".GCL", which stands for Gaze Control Language. The most attractive thing about this language is that it is very convenient to use. Even non-computer scientists can utilize this language for presenting sounds and images for specific

words when they are encountered by the user's gaze. The scripting language has all the instructions for generating the interactive display. For example:

- `showtext "words.txt"`
- `if "elephant" then show "elephant.jpg" at (100, 200)`
- `if "apple" then play "apple_pronunciation.wav"`

I implemented an interpreter for this language in MATLAB in a function named "GCL_Execute.m". It utilizes the PsychToolBox framework, which is used by many vision researchers because it allows precise timing of stimulus presentation and contains many other useful functions for psychophysical experimentation. When GCL_Execute.m is started with a GCL filename as its parameter, it reads the corresponding scripting file word by word. The Function strfind is used to find keywords such as "showtext", "show" and "play". Whenever it sees the word "showtext" it will read the following filename (here: "words.txt"), then open the corresponding text file, and display the text contained in it on the screen.

Whenever the program reads an "if", then it reads on to find out whether it is a "show" or "play" command, and then reads its parameters. While doing this, it creates a list of relevant words and the actions to take whenever the word is being looked at. And once it reaches the end of the file, the program closes the file and starts the display; first the calibration, and then the actual eye-tracking display. The following are the main functions used in the GCL interpreter:

EyeLinkDoTrackerSetup(el); el: EyeLink default values

- This is used to calibrate the eye tracker

EyelinkDoDriftCorrection(el); el: EyeLink default values

- Performs a final check of calibration using drift correction.

DrawFormattedText(w, [line1 line2 line3 line4 line5 line6 line7 line8 line9]);

- Draws a string of text from line 1 to line 9 into PsychToolBox window 'w'

imwrite(img,'img.jpg');

- Writes image img to the file img.jpg.

imread('img.jpg');

- Reads the image from the file imp.jpg.

EyeLink();

- EyeLink uses different functions to perform various functions for an eye tracker.

audioread(filename);

- This command reads the audio file specified in its parameter.

To display interactive objects at the user's gaze position in real time, we need some computation with words in the file. For this study, I decided to find out the center pixel of each word and compare that with the current gaze position. Finding the center pixel was tricky with the words in textual format. Therefore, I decided to convert all the words into images. I made another folder which contains image templates of all the letters in the English alphabet and digits from 0 to 9. These templates were concatenated to produce any English word. Now it was a lot easier to work with the pixel coordinates of each word. The entire word was treated as a rectangle. Once the fixation landed within the rectangle, the corresponding action was triggered. In this way, these words were correlated with the eye positions measured by the eye tracker to create an active gaze-contingent display.

2.3 SUBJECTS

At the University of Massachusetts, there were a total of 15 students recruited for the experiment. Ages of all the students ranged between 20 and 36, and all of them had normal or corrected-to-normal vision.

2.4 APPARATUS

Eye movements were recorded on a different computer using an SR Research EyeLink 1000 desktop system (see Figure 5). The sampling frequency of this system is 1000 Hz. The average calibration error measure after all the calibrations was 0.5^{o} of visual angle. A ViewSonic LCD monitor with the screen resolution of 1,024 × 768 pixels and the refresh rate of 75 Hz was used to present the stimuli. The participants were seated in front of the monitor using a chinrest to eliminate head moments in a dimly lit room. The distance of all the participants from the screen was approximately 70cm.

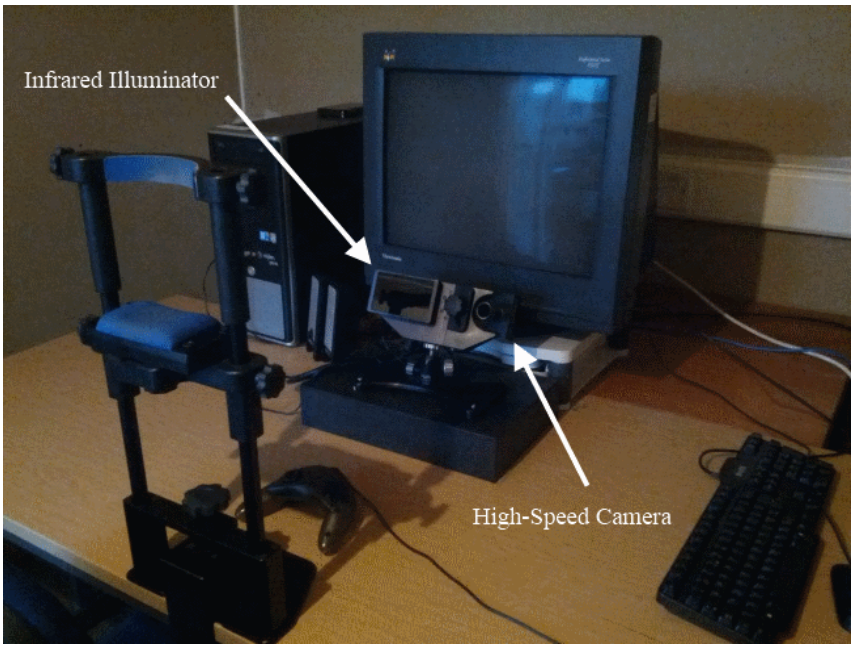

Figure 5 – The EyeLink 1000 eye tracker [20]

2.5 PROCEDURE

As a first step, calibration and validation were performed using the 9-point grid technique in which the subject has to look at each point of a 3×3 grid. Participants were then shown a series of four screens. Each screen started with written instructions, followed by a set of 20 words shown in one of the four following experimental conditions:

1) Words appeared as plain text without any image and audio being invoked at the time of gaze (see Figure 6).

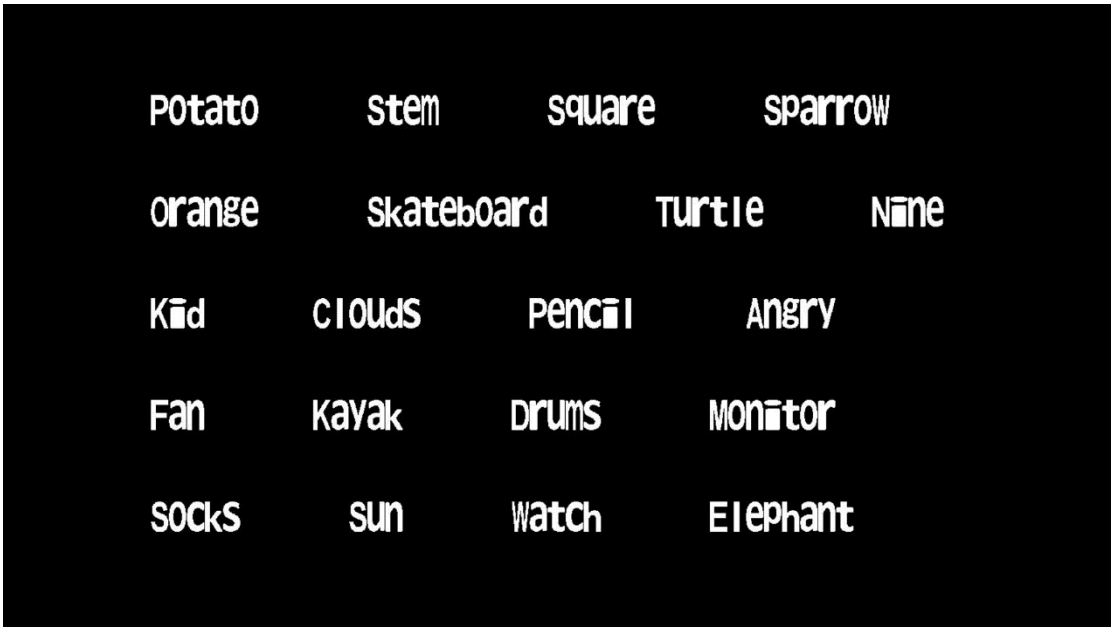

Figure 6 - Condition with no effect

2) An image corresponding to each word was displayed on its left side for a maximum of ten seconds while the subject was gazing at the word (Figure 7).

Figure 7 - Condition with image effect

3) Audio pronunciation of each words was played while gazing the word (Figure 8).

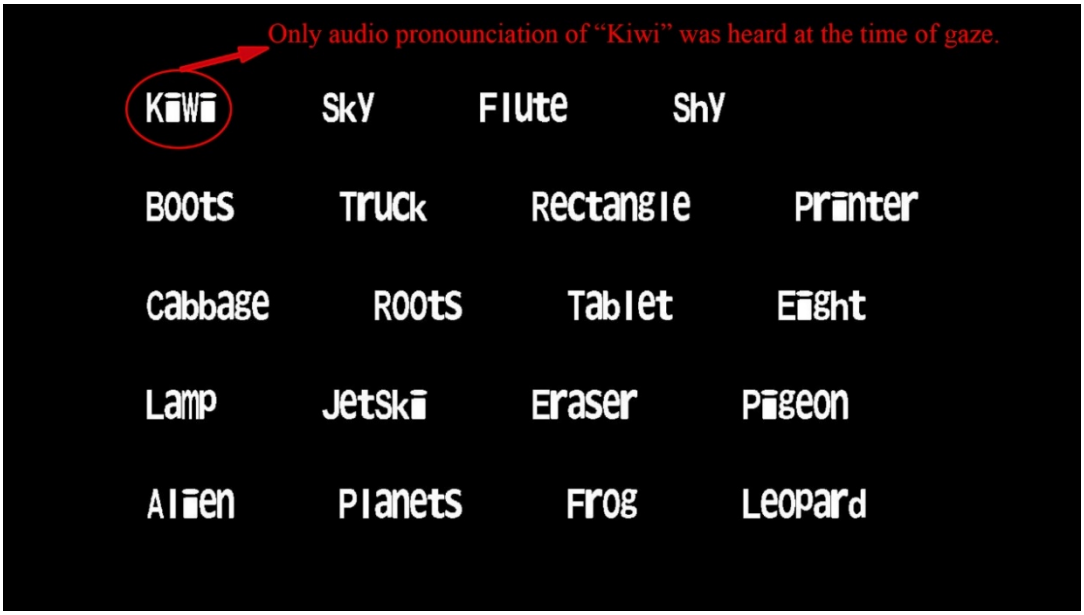

Figure 8 - Condition with audio effect

4) Both image and audio effects were invoked at the time of the gaze (Figure 9).

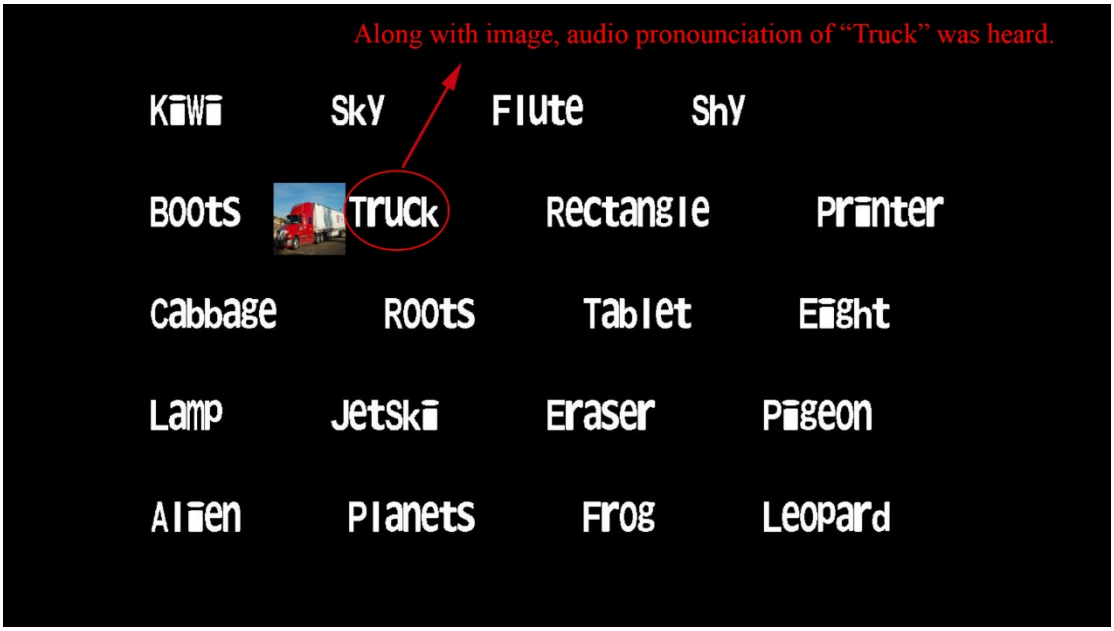

Figure 9 - Condition with image and audio effect

	Set A	**Set B**	**Set C**	**Set D**
1	Girl	Boy	Kid	Alien
2	Apple	Banana	Orange	Kiwi
3	Chair	Table	Fan	Lamp
4	Shoe	Slipper	Socks	Boots
5	Carrot	Tomato	Potato	Cabbage
6	Car	Bicycle	Skateboard	Truck
7	Flower	Leaf	Stem	Roots
8	Water	Rain	Clouds	Sky
9	Boat	Ship	Kayak	Jet Ski
10	Star	Moon	Sun	Planets

11	Paper	Pen	Pencil	Eraser
12	Cellphone	Laptop	Watch	Tablet
13	Guitar	Piano	Drums	Flute
14	Fish	Crocodile	Turtle	Frog
15	Circle	Triangle	Square	Rectangle
16	Happy	Sad	Angry	Shy
17	Lion	Tiger	Elephant	Leopard
18	Two	Five	Nine	Eight
19	Mouse	Keyboard	Monitor	Printer
20	House Fly	Bat	Sparrow	Pigeon

Table 1: The four sets of words used in the study

For each subject, all four conditions were shown in random order. Moreover, each conditions was paired with a set of words randomly chosen from the sets A, B, C, and D (see Table 1). Randomization of words and display conditions was done to increase the effectiveness of the experiment, whose purpose was to determine the effect of display condition on memory performance. If always the same set of words had been paired with the same condition, then we could not have been sure whether the differences we observed were actually caused by the condition or by the differences between the sets of words.

Participants were asked to look at each screen for two minutes and remember the maximum number of words. After every screen, they were asked to dictate all the words they remembered from the last viewed screen. An experimenter wrote down all the dictated

words to analyze the memory performance of each subject for different viewing conditions. At the end of the experiment, participants were asked which of the four conditions they preferred for learning the words.

CHAPTER - 3

RESULTS

Table 2 shows the main results from the experiment. In the experiment, two variables that I measured were relevant for assessing the gaze-contingent presentation conditions: First, the number of correctly reproduced words as a measure of memory performance, and second, the subjects' report of their preferred conditions for learning the displayed words. In the following sections, these variables will be statistically analyzed and the results discussed.

Subject	**Age**	**Sex**	**Native Speaker?**	**Text Only**	**Image**	**Audio**	**Both**	**Preferred**
1	36	M	No	9/20	7/20	8/20	7/20	-
2	23	F	No	9/20	6/20	6/20	12/20	Image
3	25	F	No	14/20	8/20	9/20	12/20	-
4	32	M	Yes	10/20	12/20	4/20	12/20	Image
5	23	F	No	11/20	13/20	8/20	9/20	Image
6	23	F	No	4/20	7/20	8/20	7/20	Audio
7	34	M	No	16/20	13/20	20/20	8/20	-
8	23	F	Yes	12/20	12/20	11/20	11/20	Image
9	20	F	Yes	14/20	6/20	13/20	13/20	Image
10	26	M	No	16/20	14/20	14/20	10/20	Image
11	31	M	No	13/20	5/20	10/20	10/20	Image
12	23	F	No	15/20	6/20	12/20	20/20	Both
13	23	F	No	17/20	10/20	11/20	11/20	Both
14	26	M	No	11/20	16/20	10/20	10/20	Image
15	23	F	No	10/20	12/20	8/20	12/20	Image

Table 2: Summary of individual subjects' results.

3.1. ANALYSIS OF MEMORY PERFORMANCE

Memory performance in each of the four experimental conditions (text only/images/audio/images and audio) was measured as the number of correctly reproduced words. As there were 20 words to be remembered, the memory scores ranged from 0 (no memorization) to 20 (perfect memorization). Figure 10 shows the results, indicating that – surprisingly – the "text only" condition led to the highest average memory score (12.0), followed by "images and audio" (10.9), "audio" (10.1), and "images" (9.8).

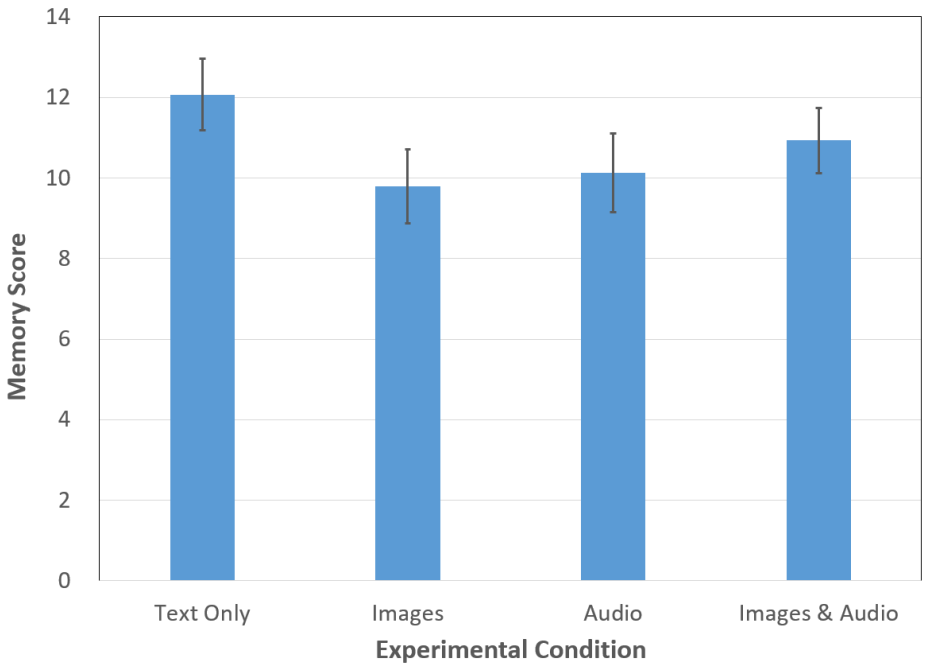

Figure 10 - Mean memory scores measured for each of the four experimental conditions. Error bars indicate standard error of the mean.

To determine whether these numbers reflected systematic differences in memory performance across the experimental conditions, paired t-tests were performed. They showed that subjects performed significantly better in the "text only" condition than in the "audio" condition ($t(14) = 2.55$, $p = 0.023$). Here, p indicates the probability that there is no difference in memory performance between the two conditions and that the deviation in performance scores was only due to random variations in measurement. Typically, in experimentation on human subjects, probabilities below 5% are considered to be meaningful indicators of actual effects. Therefore, in this case, with $p = 2.3\%$, the measured difference is significant.

The difference between the "text only" and "images" conditions was found to be marginally significant ($t(14) = 1.93$, $p = 0.074$), closely missing the significance level of 5%. All other pairwise comparisons between conditions did not reveal and significant or marginally significant effects on memory performance.

3.2 ANALYSIS OF SUBJECT'S REPORTS

After completing the experiment, subjects were asked about which of the experimental conditions they preferred for learning the words. Out of the 15 subjects, 12 provided a preference for one of the conditions. Figure 2 illustrates the results: None of the subjects favored the "text only" condition, nine preferred the "images" condition, one the "audio" condition, and two the "images and audio" condition.

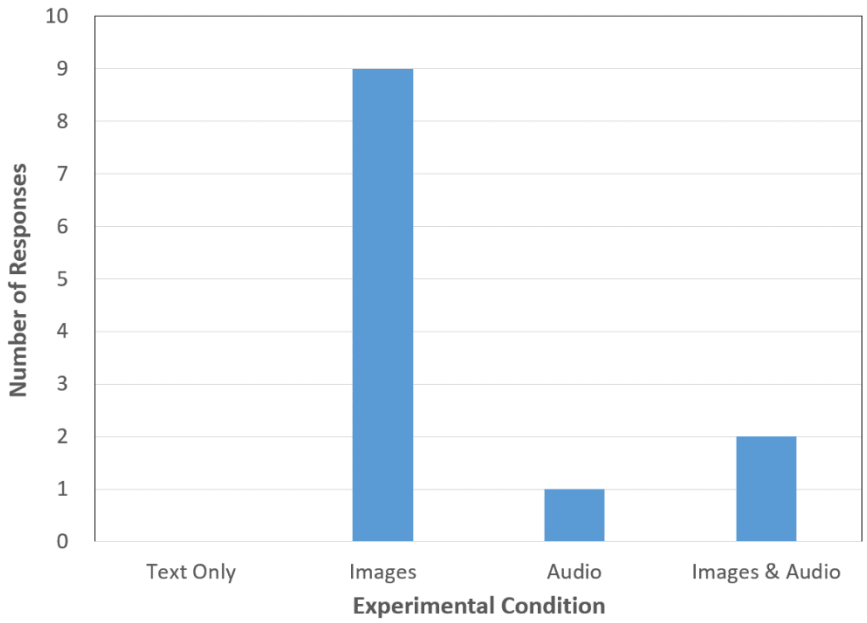

Figure 11 - Number of subjects' responses favoring each of the four experimental conditions for learning the given words.

Similar to the statistical analysis of memory performance, it is also important to test to probability of the subject's preference responses to differ significantly across conditions. For such purposes, a χ^2 ("chi-squared") test was performed. If subjects had had no actual preference for any of the conditions, then on average, each of the four conditions should have received three votes. The χ^2 test showed that the probability p of the actual distribution to occur in that scenario is $p = 0.0008$. Therefore, it is justified to say that there was a significant preference for the "images" condition.

3.3 DISCUSSION

At first sight, the results obtained are counterintuitive. Instead of improving memory performance, the additional, gaze-contingently presented information reduced it. Furthermore, the condition that most subjects preferred for learning – the "images" condition – turned out to yield the lowest memory performance.

A plausible explanation is that the subjects' responses were biased towards those conditions that they found most pleasant to work with. Many subjects reported that they enjoyed the "images" condition the most while they did not like the spoken words in the "audio" and "images and audio" conditions. Some subjects explained that the spoken words interfered with their own "inner voice" reading of them, which was perceived as unpleasant and not helpful for memorization. Because the "images" condition was more pleasant to the subjects, they may have voted for it. However, their enjoyment of the images may actually have distracted them from their task – the memorization of the given words.

Consequently, there are two things that can be learned about gaze-controlled Human-Computer Interaction from this study. First, introspective evaluation of interfaces may be misleading when testing their effectiveness with regard to a specific task. More objective measures need to be developed and analyzed. Second, presenting the same information in multiple modalities can interfere with memory performance. Such information can distract users from their task. Instead, the gaze-contingent technique should be used to present additional information to the user when his or her gaze indicates interest in a given item shown on the screen.

CHAPTER 4

OUTLOOK

In the future, the GCL scripting language that I developed and implemented can be used for further studies, even by non-computer scientists. The purposes of such studies would not have to be purely scientific. For example, the language could be also be used to present poems interactively. It would be a different and possibly effective way to make users understand difficult text in the poem or enhance the experience of the poem by additional visual and auditory effects triggered by the viewing of specific words. In the same way, the language could also be used for children's books to make them more interactive and exciting. Children can probably learn things more quickly this way and also enjoy the experience more.

Such applications can also be used when people want to learn a new language. The GCL scripts would make it possible for the computer to speak the word whose pronunciation the learner may not know. If they look at a word long enough, the word could be spoken by the computer. Or, if they do not know the meaning of the word, a picture could be shown, telling them the meaning of the word without using another language.

Another impressive application of this language could be self-teaching musical instruments. This can be achieved by integrating artificial intelligence along with the GCL

language. For instance, if a person wants to learn how to play the piano, he or she could gaze at a particular keynote on a virtual instrument shown on a screen to initiate two different functionalities at a time: First, he or she could hear the sound of that keynote. Second, key strokes which are used to play that keynote are pressed simultaneously. In this way, users can practice keynotes without external help. The loop of this keynote breaks when user looks back at a particular point on the screen.

In the modern world, teaching can be made more accessible and interesting using this technique. While the learner is reading through a text, related images or audio can make the learning more efficient. For example, in a lecture on parts of a human body using ET technology, students can explore details of a specific body part by simply looking at it.

With the use of ET techniques in e-learning, it is possible to account for learner behavior in real-time. The data collected through ET devices shows the person's interest level and focus of attention, relaxation, problem-solving strategies and emotions. It is clear that using ET in e-learning, the learner pays more attention to the learning system and also has an increased level of motivation.

In the future, research can be done to embed eye-tracking techniques in cell phones which can be used for various applications. For example, with the help of this project, one can use music karaoke applications more interactively. One functionality of such an application could work if music stops automatically when the user’s eye gaze diverges from the song’s current word.

The present work can be considered a first, small step toward the development of these future interfaces. What we can learn from the current results is that the gaze-contingently presented content needs to add information to the screen rather than provide the same information in a different way. Using this basic guideline, my scripting language GCL should be developed further to enable more sophisticated, effective, and enjoyable human-computer interfaces.

REFERENCES

[1] https://www.aapos.org/terms/conditions/100

[2] Robert Gabriel Lupu, Florina Ungurenau, 2013, A Survey of eye tracking methods and applications, http://www12.tuiasi.ro/users/103/071-086_006_Lupu_.pdf

[3] Louis Emile Javal, https://en.wikipedia.org/wiki/Louis_Emile_Javal

[4] Huey, Edmund, The Psychology and Pedagogy of Reading (Reprint), MIT Press 1968 (originally published 1908).

[5] Sarada. T, Monisha. S, Eye Movement-Based Human-Computer Interaction Techniques, 2015

[6] Brain- Computer Interface, https://en.wikipedia.org/wiki/Brain-computer_interface

[7] Victor M. Garcia-Barrios, Christian Gutl, Alexandra M. Preis, Keith Andrews, Maja Pivec, Felix Modritscher, Christian Trummer, AdELE: A Framework for Adaptive E-Learning through Eye Tracking, 2004

[8] Woodrow Barfield, Thomas A. Furness, Virtual Environments and Advanced Interface Design, 1995

[9]http://www.advancedvisiontherapycenter.com/blog/e_961/Signs_of_a_Vision_Problem/2016/11/Visual_Working_Memory.htm

[10] https://en.wikipedia.org/wiki/Memory

[11] http://www.scholarpedia.org/article/Working_memory#The_multicomponent_model_of_working_memory

[12] http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0030884

[13] Hunter Murphy, Andrew T. Duchowski, Hybrid Image-/Model-Based Gaze-Contingent Rendering, 2007, ftp://ftp.tuebingen.mpg.de/pub/kyb/roland/16-murphy-paper.good.pdf

[14] https://brainconnection.brainhq.com/2013/03/12/how-we-remember-and-why-we-forget/

[15] https://en.wikipedia.org/wiki/A_picture_is_worth_a_thousand_words

[16] https://www.prepare1.com/visual-content-amplifies-social-media/

[17] http://news.mit.edu/2014/in-the-blink-of-an-eye-0116

[18] Kirkpatrick, E.A. (1894). An experimental study of memory. Psychological Review, 1, 602-609.

[19] Standing, L. (1973). Learning 10,000 pictures. Quarterly Journal of Experimental Psychology, 25, 207-222

[20] http://wiki.psychwire.co.uk/?page_id=180