The Influence of Image Content Levels and Looking Type on Eye Movements

Helsinki University of Technology, Department of Automation and Systems Technology,
Media Technology Laboratory
P.O.Box 5500, FIN-02015 HUT, Finland, http://www.media.tkk.fi
E-mail: vsaarin2@cc.hut.fi, mari.laine-hernandez@tkk.fi, Hannu.Saarelma@tkk.fi
* Master's thesis student, ** Instructor, *** Supervisor

ABSTRACT

The purpose of this study was to find out how observing the visual content levels (syntactic, generic semantic, interpretive) and types of looking (task-relevant vs. spontaneous) influence eye movements when viewing images on a computer screen. Subjects' eye movements were measured while they viewed stimulus images on a computer screen. The subjects either looked at the images freely or answered to questions about image content. Eye movement measures were compared between different types of looking.

There were differences in the eye movement measures between looking at different content levels, but most of the differences did not appear with all of the images. The results differed the most during the first second of looking at an image and the difference were usually smaller between different types of looking after that first second. Based on the results, we assume that shorter fixation durations may be an indication of more complex processes in still image viewing. The opposite is true for the number of fixations, which may increase with increasing required understanding.

INTRODUCTION

Background

Visual perception is based on the collaboration between the eyes and the brain. Eyes collect visual stimuli from a scene and the brain forms visual percepts and interprets eye movements. Eye movements are often voluntary, controlled by attention and compose perceptions from the visual scenes /4,11/. Several variables have an influence on eye movements; they differ between individuals (depending on characteristics such as background, knowledge, age and gender), viewing situations and stimuli /9,11/. Today's eye trackers are capable of tracking human's gaze direction and measure eye movements accurately in real time. Eye tracking can be used to collect data of gaze direction and processes of visual attention. Eye tracking can also reveal a person's unconscious processes. Knowledge about eye movements during the observation of images could make it possible to predict the way a person is looking at the images.

Visual content consists of syntactic and semantic layers. This study aims to reveal people's eye movement behaviour when observing different visual content levels. If the viewing behaviour for each content layer is known, it may be possible to predict the viewing events.

The Master's thesis /15/ titled "The influence of image content levels and looking type on eye movements" was conducted as part of a media experience research laboratory installation project.
in the Media Technology Laboratory at Helsinki University of Technology. The laboratory is used in the study of media experience with various media such as digital television content, still images and print media by means of both subjective and objective methods (including eye tracking and psychophysiological measurements). The aims of the study are to install an eye tracking research laboratory, to define the influence of image content levels on eye movements and to find out how eye movements differ between different types of looking.

**Eye Movements and Attention**

Humans do not observe visual scenes with smooth round movements. Instead, the gaze shifts from place to place with rapid voluntary eye movements called saccades and vision collects information with static fixations, which occur between saccades. Saccades direct the fovea to points of interest /17/. On average saccade durations are between 30 and 120 ms /5,6/. No visual information is collected during a saccade, because the eyes do not focus anywhere during the movement /11,14/. Eyes fixate on certain objects between consecutive saccades. Typical fixation duration is 130–600 ms /6,10,11/.

Eye movements are not always conscious, nor selection processes or indicators of shifts of attention, but outcomes of preattentive selection. Observer's attention shifts to a new location before a saccade is initiated. There is normally a process delay between perceiving an object and the first saccade. /10/

Humans search useful areas from a visual scene with active rapid eye movements. The observation happens in cycles: the most interesting areas of the scene are observed first and then the gaze goes through the same areas again. This means that the least interesting areas are not observed at all and the most interesting areas get the attention over and over again /18/.

Eye movements vary depending on the interests of individuals. For example it has been shown that fixations and saccades differ between an active visual search task and passive scene observation /1,9/. Kahneman classified eye movements into three categories according to viewing situation: spontaneous looking, task-relevant looking and orientation of thought looking. Spontaneous looking is free viewing of a visual scene and task-relevant looking is viewing with a certain task in mind. Orientation of thought looking occurs when a person is concentrated on something else than the visual scene /9/.

Eye movement data can be analyzed in various ways. In this study we concentrate on the analysis of fixation duration, number of fixations and fixation duration on ROI (Region of Interest). Average fixation duration is usually thought to be an indicator of problems in search efficiency. Yet, according to Andreassi /1/, longer fixation duration is an indication of more efficient learning when reading. Number of fixations is a measure typically used in visual search tasks; an increase in the number of fixations is thought to indicate a decrease in the efficiency in task performance /6,12/. The total duration of fixations on a region of interest is an indicator of the area's significance. The regions of interest that are used in the analysis can be defined based on image content or eye movement data /13/.

**Image Content Levels**

Images can represent many things and be analyzed on various syntactic and semantic levels (Figure 1). Jaimes and Chang /7/ developed an overview structure of these content levels for the indexing of images. Jaimes and Chang's Index Pyramid consists of four syntactic and six semantic levels, where the width of each level gives an indication of the amount of knowledge
required to describe image content on that level. Even though the lower levels consist of upper levels, single levels can be seen as individual parts.

![Image of the Index Pyramid by Jaimes and Chang](image)

**Figure 1** The Index Pyramid by Jaimes and Chang. /7/.  

Syntactic levels define the visual elements such as colours and lines and are in close relation to visual perception. Semantic levels are concentrated on visual concepts and they define the meanings of the visual elements and of their arrangements. As can be seen from the Pyramid, semantics can be observed on general, specific or abstract level. For example an image of an apple can be seen as an instance of apple (general), as a specific apple such as the Macintosh apple (specific), or it may generate a mental image of forbidden fruit (abstract). /7/

Jörgensen /8/ for one, divides image attributes into three different groups; perceptual, reactive and interpretive attributes. Perceptual attributes are closely related to visual stimuli, e.g. the colour "red". Reactive attributes are peoples' personal reactions to images, e.g. uncertainty, confusion and liking the image. Interpretive attributes include features like sentiment, abstract concepts and content elements like action and function. Because interpretive features include abstract concepts, we merge Jörgensen's attributes and Jaimes and Chang's syntactic, general, specific and abstract layers and use interpretive attributes instead of the abstract layer. In this study we examine the influence of these content layers on eye movements. /8/

Because the syntactic levels have a fairly simple structure, syntactic level processes are typically fast. Factors that influence observation at the syntactic level include contrast, size, shape, colour and movement. Things that draw the observer's attention include lightness, orientation and line ends /5,12,13/. Interesting semantic level elements include different human and animal shapes, as well as figures that do not fit in the picture. In general things that evoke powerful feelings, like anger or threat, are looked at the most /14/.

**METHODS**

**Eye Tracking Equipment**

The SMI eye tracking equipment used in the experiment is composed of a helmet, the SMI-computer and attachment cables. Two cameras are attached to the helmet: a seeing camera and an eye camera (Figure 2). The seeing camera records the scene and the eye camera records eye
positions at a 50 Hz sampling rate. The eye camera is an infrared camera, so the recorded video is greyscale. Infrared is used in the eye camera to reduce the red-eye effect, a very common effect in photography, usually appearing when visible light reflects back to the camera from the retina /3/.

Figure 2  The SMI iViewX-HED head mounted eye tracker.  
Figure 3  Test environment. The stimulus monitor is on the left and the operation monitor on the right. The eye tracking helmet is attached to a headrest.

The SMI iViewX-HED head mounted eye tracker measures eye movements in relation to the subject's head. It is therefore difficult to define the gaze path on a specific surface if the head is moving in the environment. To enable eye tracking on a certain plain, the subject's head has to be either tracked with a magnetic head tracker /3/ or held totally still. In this study the eye tracking system was used with a headrest (Figure 3), which kept the head still in relation to the stimulus monitor.

The SMI eye tracker uses the corneal and pupil reflection relationship technique. Infrared light is reflected into the user's eye. Several reflections occur from the boundaries of the lens and cornea. They are called Purkinje images. This technique uses the first Purkinje image and the dark pupil /4,11/. The changes in gaze direction are calculated based on the measured changes in the relation of the first Purkinje image and the pupil /4/. The subjects viewed test images from the stimulus monitor at a 500 mm distance. The stimulus monitor was a 22" ViewSonic widescreen LCD display. The resolution of the images was 720 × 576 pixels and their physical size was 186 mm × 149 mm, so the viewing angle was 21,1° × 16,9°. The eye tracking resolution with the SMI eye tracker was 0,1° and accuracy 0,5–1,0°.

The test images were presented with Presentation 9.81 software. The eye tracker was operated with the iViewX software on the operator monitor, which was placed next to the stimulus monitor (see Figure 3). The program detects pupil and corneal reflections from the eye camera video and measures gaze direction based on these two points. The eye tracking program records eye coordinate data, but it also creates a video with an eye cursor, which defines the gaze focus. /3/

In order to calibrate and track the gaze direction accurately, the eye tracking helmet should be attached to the subject's head tightly. The eye camera should have a clear view of the eye through the eye mirror. If the angle between the eye and the eye camera is too steep, eyelashes or eyelids can get in the way of the camera, preventing it from getting a clear view. Although in
previous eye tracking studies eye glasses were reported as the biggest individual problem, eye tracking worked well with the SMI equipment with eye glasses and contact lenses /16/.

Surrounding lights may hinder the detection of pupil and corneal reflections. It is therefore desirable to reduce the illumination level of the environment by turning off any lights hitting the eye or the eye mirror directly. Calibration has to be carried out accurately with every test subject, because all inaccuracy occurring during calibration causes error to the final measurement results.

**Test Subjects**

A total of 30 (27 men, 3 women) subjects participated in the study. All the participants were students and staff members of Helsinki University of Technology. Their ages ranged between 21–47 years, averaging 24 years. All the subjects had normal or corrected to normal vision (5 of them wore glasses, 2 contact lenses). The subjects were divided in random into two groups, group A and B, which both included 15 subjects.

**Test Material**

The test images ([Figure 4](#)) were chosen from a stock photography service to represent some of the most common types of magazine photograph categories. These categories were action (test images 1, 2, 11 and 12), close-up of a person (test images 3, 4, 13 and 14), landscape (test images 5 and 15), urban scene (test images 6 and 16), building (test images 7 and 17), group of people (test images 8 and 18), interior (test images 9 and 19) and food (test images 10 and 20).
Figure 4  Test images 1–20.

Test Procedure

The purpose of the experiment was to compare the influence of image content levels and types of looking on the eye movements. This was achieved by asking the subjects questions in order to draw their attention to certain levels of visual content (i.e. syntactic, generic and interpretive (abstract)) and simulate different types of looking (i.e. spontaneous and task-relevant looking).

Each subject saw all twenty test images twice. On the first round their task was to look at ten of the test images freely (spontaneous looking) and the other ten in order to answer questions regarding their generic or interpretive semantic content (task-relevant looking). On the second round the subjects saw all twenty test images and answered a question regarding their syntax (task-relevant looking). Each question was presented to the subject before seeing an image, and the subject answered the question after seeing the image. The test procedure is summarized in Table 1.

In the spontaneous looking condition the images were presented for five seconds each, with a one-second break between the images. In the task-relevant looking condition the images were
also presented for five seconds each, followed by an unlimited time for answering the question. The duration of the test was approximately 20 minutes per subject.

Table 1  Test procedure.

<table>
<thead>
<tr>
<th>Test group</th>
<th>Test images</th>
<th>Question/Task</th>
<th>Visual content level / Type of looking</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>1–10</td>
<td>What is there in the image?</td>
<td>General / Task-relevant</td>
</tr>
<tr>
<td>A</td>
<td>11–20</td>
<td>Look at the images freely</td>
<td>– / Spontaneous</td>
</tr>
<tr>
<td>A</td>
<td>1–20</td>
<td>Name three main colours in the image</td>
<td>Syntax / Task-relevant</td>
</tr>
<tr>
<td>B</td>
<td>1–10</td>
<td>Look at the images freely</td>
<td>– / Spontaneous</td>
</tr>
<tr>
<td>B</td>
<td>11–12</td>
<td>What is happening in the image?</td>
<td>Interpretive / Task-relevant</td>
</tr>
<tr>
<td>B</td>
<td>13–14</td>
<td>What is the person’s mood?</td>
<td>Interpretive / Task-relevant</td>
</tr>
<tr>
<td>B</td>
<td>15–19</td>
<td>How is the mood in the image?</td>
<td>Interpretive / Task-relevant</td>
</tr>
<tr>
<td>B</td>
<td>20</td>
<td>How does the food look like?</td>
<td>Interpretive / Task-relevant</td>
</tr>
<tr>
<td>B</td>
<td>1–20</td>
<td>Name three main colours in the image</td>
<td>Syntax / Task-relevant</td>
</tr>
</tbody>
</table>

Data Analysis

ROI (Region of Interest) analysis was carried out in order to find out how the subjects observed certain areas in the images. We defined certain areas from the test images that we considered significant for different types of looking (Figure 5). These areas were mainly people, human faces and large uniform colour regions such as blue sky.

![Figure 5](image)

Regions of interest in test images 4 and 14.

Fixation analysis included the calculation of numbers of fixations and fixation durations. We analyzed the fixations for two different time windows: 0–1 seconds and 0–5 seconds. A t-test was carried out to analyze the differences between different types of looking. The difference was significant if the p-value was less than 0.05.
RESULTS AND DISCUSSION

ROI Analysis

The main objects of the images were looked at the most during interpretive looking and the least during syntax looking, while the opposite was true for background areas and other large uniform areas. Fixation values such as the number of fixations and fixation duration of general looking lied in between the values for syntax and interpretive looking.

Total fixation durations were measured for each region of interest and type of looking. In the following we go over the ROI analysis more precisely for test images 4 and 14 (Figure 5), the results can be seen in Figure 6. Large uniform colour areas like the girl's purple cap were looked at the most during syntax looking and least during general looking (abstract looking was not measured). On the contrary, the girl's face drew most of the attention for general looking and least attention for syntax looking.

![Figure 6](image-url)

**Figure 6** Fixation durations on ROIs for test image 4.

For test image 14 interpretive looking was observed instead of general looking. The results of the ROI analysis for image 14 can be seen in Figure 7. Interpretive looking focused on the expressive parts of the image, such as the boy's face. The boy's shirt and the rest of the image were looked at less with interpretive looking than with spontaneous or syntax looking. The combined results of images 4 and 14 indicate that the main regions of images (i.e. the objects on the foreground) were looked at the most with interpretive looking, second-most with general and least with syntax looking. Contrary to this, the background regions of images were looked at the most with syntax looking and least with interpretive looking. The most significant differences, however, were found between interpretive and syntax looking (shirt p=0.008, face p=0.000 and the rest of the areas p=0.003).
Figure 7  Fixation durations on ROIs for test image 14.

The results for the other test images were similar to the results obtained for test images 4 and 14. The differences were significant between general and syntax looking and also between spontaneous and syntax looking. The differences between different types of looking were not uniform among all test images. Significant differences appeared in some cases, but not in all images at the same time.

**Fixation Analysis**

Average fixation duration was the shortest for interpretive looking and the longest for syntax looking. The differences between different types of looking were, however, not significant for all test images. We presume that the average fixation duration increases when moving towards the top of the index Pyramid (see Figure 1) and the number of fixations increases when moving towards the bottom of the Pyramid. Average fixation durations were slightly shorter for the entire five seconds of viewing than for the first second (Figure 8).

Figure 8  Average fixation duration with different types of looking for time windows 0–1 seconds and 0–5 seconds.
The average number of fixations for different types of looking can be seen in Figure 9. During the first second of viewing, the number of fixations was the largest with interpretive looking and the smallest with syntax looking. During the entire 5-second viewing time, however, the average number of fixations per second was the smallest with interpretive looking. This observation does not fit in any model.

![Average number of fixations](image)

**Figure 9**  Average number of fixations with different types of looking.

Average fixation duration and number of fixations increased and decreased slightly between different index steps of the Pyramid, even though the differences of neighbouring steps were not always significant. The differences in the number of fixations per second and average fixation duration between different types of looking were larger during the first second of viewing than when measured for the entire five seconds of viewing.

**CONCLUSIONS**

**Eye Tracking System**

We were able to successfully conduct eye tracking and measure gaze paths with the SMI iViewX-HED head mounted eye tracker, but the data analysis was difficult and required a lot of manual work. Better data analysis software would make the analysis easier and also possible for inexperienced users.

SMI device adjustment has to be done very accurately and requires experience. Even though the experiments were conducted in laboratory conditions, the surrounding lights were the biggest problem for the eye tracker. However, once the calibration had been carried out successfully, the eye tracking device was quite stable and only a few problems occurred during the tests.

The test was successfully carried out with 28 out of 30 participants, which means that the failing proportion in this experiment was only 7%. In some other studies the failing proportion was much worse; for example Schnipke and Todd measured a success rate of only 37.5% /16/. Eye tracking records from two of the test images were discarded due to poor pupil measurement.
These images were very dark, which presumably made the observer's pupils dilate and become too large for tracking.

**Experiment**

Differences occurred between different test images and it was not possible to conclude, which image types differed from each other. Yet, it was possible to interpret regions of interest. It can be concluded that informative areas get the most attention in general.

According to Jaimes and Chang [7], "at the lower levels of the Pyramid, more knowledge and information is required to perform indexing". The fixation duration during the first second therefore seems to indicate that fixation duration decreases with more understanding required from the scene. This is true only if the content level indexing process is the same as viewing certain content level. It may be possible that shorter fixation durations are indicators of more complex processes in still image viewing. The differences between 0–5s are not as significant as during the first second of observation. The opposite is true for the number of fixations; number of fixations increase with required understanding increasing.

Fixation duration and number of fixations differed more between different types of looking during the first second of looking than during the entire five-second trial. It is possible that the image content levels influence viewing only in the beginning of a trial. We presume that viewing is most efficient in the beginning as the observer views a totally new scene and tries to get as much information as possible. Later, when the scene is already familiar, there is no reason for active scanning. This may be why the content level effect occurred only during the first second of viewing. This hypothesis requires further research including more accurate measurement of fixation durations and numbers of fixations in relation to time.

Significant differences were found also between spontaneous and task-relevant looking, but the differences were not uniform, because eye movements differed between individual task-relevant looking types.

**REFERENCES**


