Task Influence of Scene Content Selected by Active Vision

Marianne A. Lipps

B.S. Imaging Science

Rochester Institute of Technology (2002)

A thesis submitted in partial fulfillment of the requirements for the degree of Master of Science in Imaging Science in the Chester F. Carlson Center for Imaging Science of the College of Science at the Rochester Institute of Technology

August 2004

_____

Signature of the Author

_____

Accepted by Dr. Harvey Rhody

Coordinator, M.S. Degree Program

Thesis Release Permission Form

Chester F. Carlson Center for Imaging Science

College of Science

Rochester Institute of Technology

Rochester, New York

Title of Thesis: Task influence of scene content selected by active vision

I, Marianne A. Lipps, hereby grant permission to the Wallace Memorial Library of the Rochester Institute of Technology to reproduce my thesis in whole or in part. Any reproduction will not be for commercial use or profit.

_____

Signature

_____

Date

Chester F. Carlson Center for Imaging Science
College of Science
Rochester Institute of Technology
Rochester, New York

Certificate of Approval

M.S. DEGREE THESIS

The M.S. Degree Thesis of Marianne A. Lipps has been examined and approved by the thesis committee as satisfactory for the thesis requirement for the Master of Science degree.

_____

Dr. Jeff B. Pelz, Thesis Advisor

_____

Dr. Roxanne Canosa

_____

Dr. Carl Salvaggio

_____

Date

Task Influence of Scene Content Selected by Active Vision

Marianne A. Lipps

Submitted to the Chester F. Carlson Center for Imaging Science of the College of Science in partial fulfillment of the requirements for the Master of Science Degree in Imaging Science at the Rochester Institute of Technology

Abstract

This thesis investigated possible features that are used to guide saccadic eye movements in specific tasks, including a visual search task, answering questions about an image, and freely viewing images. Current eyetracking technology was used to gather eye movement records of subjects as they viewed images.

A classical experiment that shows the influence of task on eye movements conducted by Alfred Yarbus was replicated under natural viewing conditions. In this experiment, 17 viewers were given a set of different instructions before viewing a Russian painting. Eye movement records were compared between tasks, and it was found that the instruction a viewer is given did affect which regions of the image are fixated. Even though the viewing times in the two experiments were drastically different (3 minutes compared to ~20 seconds), the behaviors of the 17 subjects were remarkably similar to the original record published by Yarbus; regions that were 'informative' for the task were fixated. Behavior between the 17 subjects (within one task) was more similar than between the seven tasks (within one subject).

In a second experiment, 23 observers performed a visual search task in images of real-world scenes. Before each trial, the subject was shown a preview image of the target. This image was either pixel-for-pixel exactly as it appeared in the image ('Extracted Object' condition) or was a cartoon icon representation of the target ('Cartoon Icon' condition). On average, the reaction time in finding the target in the Cartoon Icon condition was 3.0 seconds, and less than 2.5 seconds in the Extracted Object condition. This increase in reaction time was caused primarily by the viewer taking longer to initially fixate on the target.

Perceptual saliency and other feature content of the images at fixated and random locations were compared to gain insight into what features the visual system was using to guide, and expedite, visual search in each of the two conditions. Several commonly used metrics were used to measure the performance of each of 18 different topographical feature maps. It was found that feature maps that weight areas according to the color and spatial characteristics of the target perform better than general low-level saliency maps, showing that the visual system can be fine-tuned according to the task. However, a general model of visual attention for search in real-world scenes cannot be created using only low-level stimulus properties.

Acknowledgments

I would first like to thank Dr. Jeff B. Pelz for his guidance as a teacher and advisor, for teaching me about human vision, for encouraging me to give public presentations, and for giving me the opportunity to work in the Visual Perception Lab.

Thank you to Dr. Roxanne Canosa for her previous research on both visual perception in the real world and computational modeling, and to Dr. Carl Salvaggio for advice and guidance on my thesis.

Much appreciation goes to the faculty and staff of the Center for Imaging Science, who taught me all that I know about imaging. Thank you to Catherine Carlson for financial support. I have thoroughly enjoyed the six years I have spent in the Center, and hope to remain part of the community in the future.

A special note of thanks goes to Jason Babcock for teaching me everything about eyetrackers and eyetracking. Thanks also to Constantin Rothkopf for his classification algorithms, and to Mary Ellen Arndt for help with running experiments.

Most importantly, I would like to thank Christopher DeAngelus and my family for their love and support.

Table of Contents

List of Figures

List of Tables

Chapter 1

1 Introduction

1.1 Overview

Art students are often taught that a piece of art with good composition is one that smoothly 'leads the eye' throughout it, toward or away from the focal point of the piece. Color, shape, and line are common elements used to lead a viewer through a painting. "Lines may be used to guide the eye of the spectator or to stop it. Rule: The eye follows the length of a line, and across a gradation from the tone nearest the ground tone to that farthest removed in value [Thompson 1936]." This ability to control the movement of the viewer's eyes is often discussed in critiques:

"[Nine Spaces by Ron Kroutel] (shown in Figure 1) is a thicket of diagonal shapes, which direct the eye past vertical barriers. The viewer's eye enters the picture plane at the bottom left, then zig zags over an impenetrable wall at the back, and soars up to a distant sky. … The forms themselves are almost like arrows, relentlessly pointing the eye upwards. The ellipses in the middle offer a visual pause for the eye as it follows the strong, sharp diagonals to the top of the piece. Next, look at the color; since the warmest colors are reserved for the distant sky, the eye is pulled towards the top of the piece. In the lower left corner, you'll notice a pale blue color that is repeated in the center, then again at the upper right. Kroutel provides the eye a direct route to the top of the painting by inviting the viewer to connect the dots through the use of this light blue paint… [ArtSmart]"

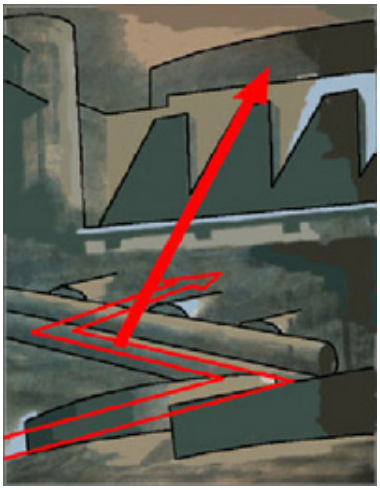


Figure 1: ‘Nine Spaces’ by artist Ron Kroutel with overlay of supposed path of eye movements

However, researchers in human visual perception know that the eye actually behaves much differently than artists claim. Instead of smoothly following lines in a piece of art, the eye actually makes a series of rapid jumps and pauses to various regions in the image. These ‘jumps’ are saccadic eye movements, which reorient the eye to regions where the eye pauses, or fixates, several times per second. When different people look at the same painting, the paths created by the movement of their eyes are never identical. They may fixate on similar regions, but often not in the same temporal order.

A significant amount of research has been done to investigate the behavior of eye movements during various tasks such as looking at pictures, reading, and doing complex, everyday tasks like washing hands or making tea. From these studies, we know that eye movements are neither random nor simply reactive; rather, movement of the eyes is a very active process even though we are typically unaware of it. We also know that fixations land on the most ‘informative’ regions of the image for the task given [Brandt, 1945; Yarbus, 1967]. What features of the region make it most informative? Comparisons between computational models of vision and human fixation patterns show that saccades are not simply directed by low-level (bottom-up) attractors such as contrast, color, edges, high spatial frequencies, as artists may claim, but are influenced by a person’s task, background experience, or interests [Canosa, 2003; Noton and Stark, 1971; Privitera and Stark, 2000].

1.2 Objectives

The objective of this research project is to utilize current eyetracking technology to gain insight into how humans choose areas of interest in a scene (i.e., what to fixate on next) in different tasks. What features of a region make it most "informative" for a specific task? This question will be investigated by analyzing the system as an active imaging system. By examining the image content at locations of fixation, inferences can be made about what types of information this active system is seeking, and how that information is affected by the person's task.

Recent work involving computational models of visual attention involves first building a model, and then seeing how well the model predicts locations of fixations obtained by eyetracking experiments. It is difficult to estimate how well a model performs due to the inconsistency in performance metrics, observer tasks, and images used.

This research project does not attempt to produce a new model to be verified by eyetracking experiments. Rather, it will take a step back and first analyze the information selected by the visual system when viewing images. The results will then provide guidelines for constructing better (or task specific) models of visual attention.

The main objectives of this thesis are:

- Replicate the classical experiment of Alfred Yarbus that shows task influence on viewing behavior
- Conduct an experiment in which observers search for targets in real-world scenes; the target preview is either exactly the same as in the containing image, or a cartoon icon representation of it
- Determine relationship between features of target preview and features of fixated locations in the image
- Investigate factors that influence how humans choose locations of future fixations in order to improve computational models of visual attention

Chapter 2 of this report will give an overview of the human visual system, and present background information about previous research on the task influence of eye movements, scene content selected by active vision, computational models of visual attention, and performance

metrics. Chapter 3 will discuss the eyetracking technology used to perform the experiments. Chapter 4 describes the replication of a classical experiment performed by Alfred Yarbus under more natural conditions. Chapter 5 presents a visual search experiment in which the features of the target preview were varied. Chapter 6 contains general conclusions and a discussion of future work.

Chapter 2

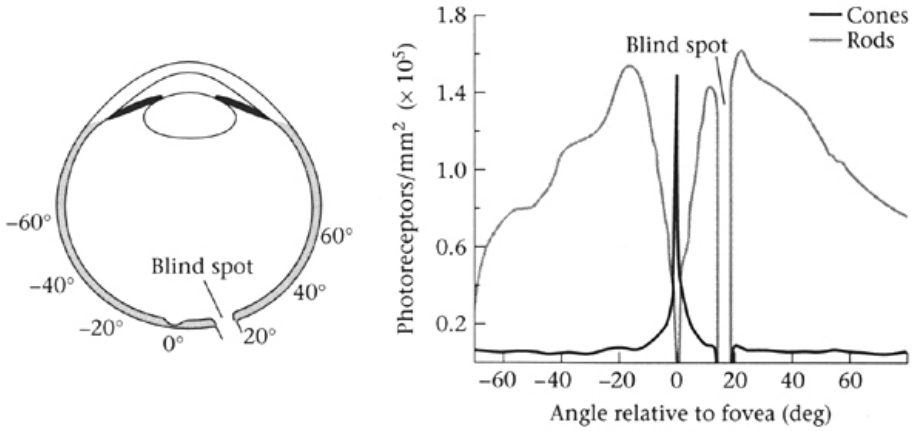
2 Background

2.1 Overview

The following subsections give an overview of the nature of the human visual system and discuss research on saccadic eye movements, fixation patterns, and computational models of visual attention.

2.2 The Foveal Compromise

The human visual system is highly sophisticated, and allows for the subjective perception of a full field of high-resolution vision. This perception is created by sampling the environment through a complex process that occurs below consciousness [Pelz and Canosa, 2001]. In everyday vision, the surrounding environment is sampled spatially and temporally by a pool of sensors located on the retina. There are two types of photosensitive receptors used; the first is the *cone*, which is used for color vision during normal levels of illumination. The *rods*, on the other hand, are highly sensitive and are useful in low levels of illumination. These sensors are not evenly distributed throughout the retina, as in a CCD camera; the cones are clustered on the back of the retina near the optical axis and comprise the *fovea*, while a greater number of rods compose the periphery, as illustrated in Figure 2.



a) b)

Figure 2: (a) Degrees of visual angle relative to the position of the fovea in the left eye. (b) Distribution of rods and cones on the retina of the human eye [Wandell, 1995].

Although we perceive high resolution vision everywhere, it is only in the fovea that we have high-resolution capabilities. The central 1-2 degrees of the visual field is represented with a resolution that is almost 100 times greater than the periphery [Rao, et al., 1996]. In fact, a larger amount of computational resources in the cortex is responsible for processing the central areas in comparison to peripheral areas [Palmer, 1999, page 155].

In order to create the effect of high resolution everywhere with our resolution-limited system, it is necessary to move the detector across the visual field rapidly. The oculomotor system allows humans to move their eyes at speeds up to 700 degrees per second [Rao, et al., 1997]. These eye movements are used to stabilize an image on the retina, follow an object that is moving, or to reorient the eye to gather new information about a scene. On average, a person makes over 100,000 eye movements every day. *Saccades* are rapid, ballistic eye movements that reorient the fovea to new targets that require high acuity or that are of interest for a given task. Saccades take only 150-200 milliseconds to plan and execute, and the actual movement of the eye is completed in only 20 milliseconds + 2 milliseconds per degree of visual angle. During the execution of a saccadic eye movement, perception is suppressed so that blurring of the retinal image is not perceived. *Fixations* occur when the eye pauses at a particular spatial location and typically last about 250 milliseconds in visually engaging tasks such as reading. It is during these brief pauses that high resolution information about the visual environment is collected.

Eye movements are external manifestations of selective visual attention. By studying these eye movements, it is possible to understand how visual attention is deployed in the environment in various conditions and tasks [Pelz and Canosa, 2001].

2.3 Monitoring Eye Movements

The mechanics of the oculomotor system have been studied in depth through experiments in controlled laboratory settings. Typically, eye movements are tracked as observers are asked to perform simple tasks while the head is held stationary. These tasks involve looking at small lights or searching for a specific shape in a field of similar shapes. While these research studies have learned much about the visual system, the findings cannot be applied to visual perception during complex, natural tasks. Several studies have shown significant differences between eye movements when the head is fixed and eye movements when the head is allowed to move freely. For example, it was found that retinal image stabilization decreased when the subject's head was not supported [Skavenski, et al., 1979]. Other research showed that saccades are faster and more accurate when the head is free to move [Collewijn, et al., 1992]. Also under natural conditions, it has been seen that vergence eye movements (counter-rotation of the eyes) are carried out at a higher velocity than previously thought [Steinman, et al., 1990].

2.4 Eye Movements and Picture Viewing

2.4.1 General behavior during picture viewing

The first thorough objective investigation into how people look at pictures was published in 1935 by Guy T. Buswell [Buswell, 1935]. Prior to his research, most information about eye movements was based on subjective and introspective analysis. In his experiments, Buswell recorded eye movements of over 200 participants as they viewed 55 photographs of various types of fine art. He compared eye movements of trained and untrained artists, but found no significant differences. However, he concluded that although no two subjects exhibited the exact same viewing behavior, two general classes of viewing behavior could be formed. The first is represented by a global survey of the image, where subjects made brief fixations, averaging 210 milliseconds, over the main features of the image. The second behavior is characterized by long fixations, averaging 350 milliseconds, over smaller sections of the image. In general, the global fixations were made early, followed by longer fixations as viewing time increased.

When fixation patterns were plotted collectively over a specific image, areas of high fixation density often corresponded to "information-rich" regions in the image. This suggests that observers fixated on the same spatial locations in the image, but not in the same order over time. Generally, people did not randomly explore the images. Instead, they focused on foreground elements including faces and people, and rarely focused on background elements.

2.4.2 Task dependencies of eye movements

In 1967, Alfred Yarbus reported that as a subject viewed I.E. Repin's painting entitled *They Did Not Expect Him*, eye movement patterns changed when different instructions were given, as shown in Figure 3. For example, when the observer was asked to remember the clothes the people in the painting are wearing, or to estimate the age of the people, the most informative regions (as defined by the task) received the most fixations. Eye movements were recorded for three minutes for each instruction, and while this is a very unnatural viewing condition, the results suggest a high degree of task influence on visual behavior.

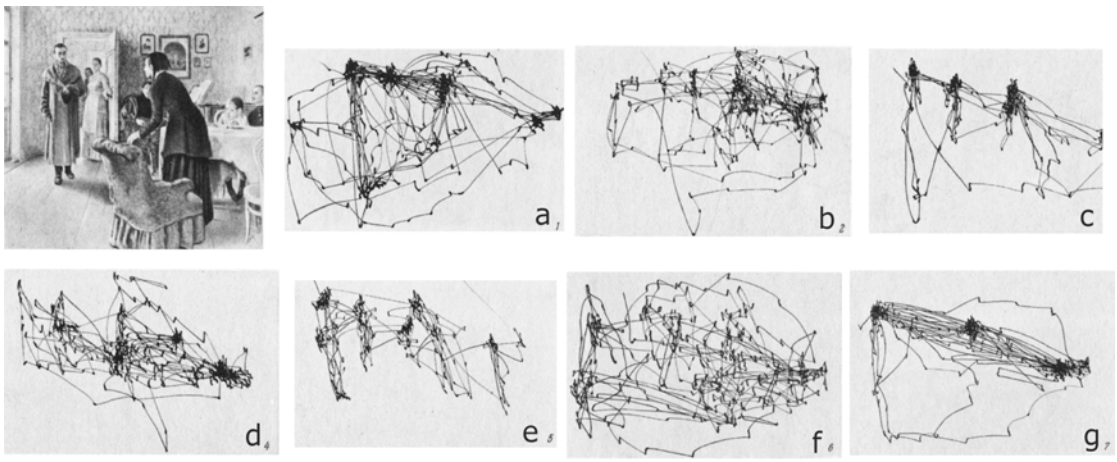


Figure 3: Seven records of eye movements by the same subject while viewing a painting (top left). Each record lasted three minutes. (a) Free examination of the picture. Subsequent records were made after the subject was asked to: (b) estimate the material circumstances of the family in the picture, (c) give the ages of the people, (d) surmise what the family had been doing before the arrival of the "unexpected" visitor," (e) remember the clothes worn by the people, (f) remember the position of the people and objects in the room, (g) estimate how long the "unexpected visitor" had been away from the family [Yarbus, 1967, adapted from Figure 109].

Research presented by Molnar in 1981 also showed that eye movement patterns change depending upon the task given to observers. A group of fine-art students viewed eight classical paintings as their eye movements were recorded. Half of the group was told that they would later be questioned about what they saw. The other half was told that they would be asked about aesthetic qualities of the painting. Molnar found that fixations were much longer for the group making aesthetic judgments. In other research done by Nodine, Locher, and Krupinski in 1991, it was found that composition of images affected eye movement patterns of trained artists. Observers made long fixations and tended to focus on spatial relationships between foreground

objects and background. Untrained viewers made shorter fixations, and focused on semantically important regions of the image.

In a study [Babcock, Lipps and Pelz, 2002] in which subjects were eyetracked as they took digital photographs and later cropped them, it was shown that eye movement behavior differed between the two tasks. During image capture, each subject was asked to take a photograph of a person, sculpture, and interior environment. The amount of time spent looking at the primary object, surrounding environment, and camera differed between subject matter. However, as the subjects cropped the same photographs that they took, the differences found between scenes in the image capture task were not found.

Research by Canosa [Canosa, 2003] also explicitly showed differences in visual routines for subjects performing different tasks. First considering low-level eye movement metrics, it was found that mean fixation duration and saccade amplitude vary between tasks. For example, visually engaging tasks such as reading and sorting blocks elicited short fixation durations of 200-350 milliseconds, as well as small saccade amplitudes of 4-6 degrees of visual angle. Other tasks, such as walking down a hallway or having a face-to-face conversation elicited longer fixation durations as well as larger saccade amplitudes. Differences between fixation locations were also shown. Subjects performed different tasks in the same environment, and it was found that an average of 65% of fixations were on task-relevant objects, or objects that may be potentially useful. Figure 4 shows the amount of time spent looking at various objects in a washroom, across three different tasks: washing hands, filling a cup with water, and combing hair.

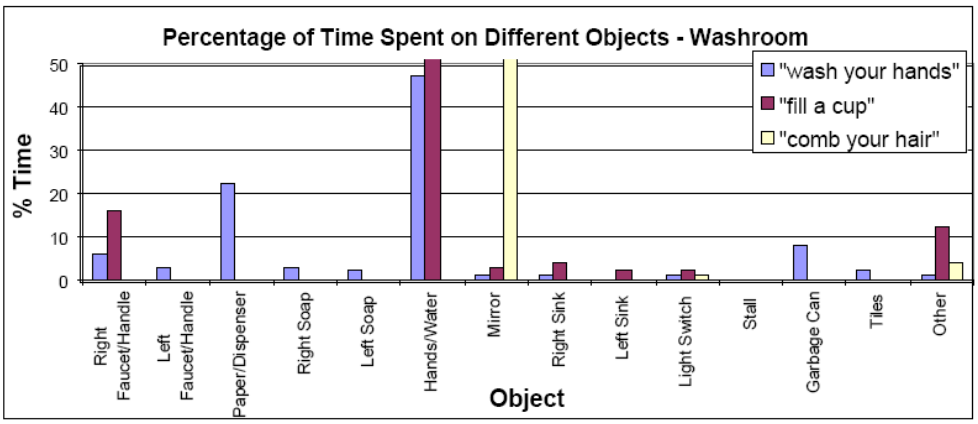


Figure 4: Relative amount of time spent on objects in the washroom environment for all subjects. Note that fixations on hands/water area for “fill a cup” is 52%, and fixations on mirror for “comb your hair” is 90%. [Canosa, 2003, page 151].

These studies suggest that patterns or locations of fixations cannot be predicted by the structure of the stimulus alone. Rather, eye movements are guided by the interaction between the contents of the stimulus with the perceptual goals of the observer.

2.4.3 Idiosyncrasies of eye movements

Noton and Stark found that observers execute very characteristic scan paths, or temporal sequences of fixations, when repeatedly viewing a particular image. However, there is much variation between these scan paths across different observers viewing the same image; this result is in agreement with previous findings that subjects tend to fixate on similar regions but in different temporal sequences [Buswell, 1935]. The authors suggest that perception proceeds in a serial fashion, in which a growing collection of fixated features facilitates recognition.

Another study by Andrews and Coppola in 1999 investigated whether temporal and spatial characteristics of eye movements are idiosyncratic. Subjects were eyetracked under five different viewing conditions for comparison: in a dark room, while looking at simple textured patterns, while looking at a complex natural scene, while performing visual search, and while reading. In general, the visual environment had significant influence on the mean saccade size and fixation duration. It was also found that the mean fixation duration and saccade size when viewing a complex natural scene covaried significantly with those parameters in the absence of visual stimuli (dark room). Similarly, the same spatio-temporal parameters covaried between reading and visual search, but did not correlate with patterns of eye movements in the other visual environments, such as viewing a complex scene. From this analysis, the authors claim that an individual's eye movements in a dark room predict the pattern observed when viewing a complex natural scene, i.e., there is significant endogenous influence on oculomotor control.

2.5 Scene content selected by foveal vision

2.5.1 Statistical analysis of fixated regions of images

A study by Krieger, Rentschler, et al. [2000], investigated statistical properties of fixated regions of images. The 49 images included photographs of people, natural environments, man-made structures, cartoons and abstract artwork, and were presented in grayscale at a resolution of 512 x 512 pixels. The stimuli subtended 18 x 18 degrees of visual angle, and were present for 5 seconds. Eleven subjects' eye movements were recorded using a dual Purkinje-image eyetracker, which required that the subjects head be fixed during the experiment. Subjects were instructed to view each image as carefully as possible, to be able to perform subsequent tests with the images.

For each image, small regions (author does not give size of regions) around fixation points were extracted for statistical analysis. Initial analysis showed that in comparison to randomly selected regions, fixated regions contained higher spatial variance, suggesting that saccadic eye movements avoid regions with little structural content. However, a closer look at the power spectra of the fixated and random regions show no significant difference in structure. Further analysis into the image regions' bispectra yielded some differences. The bispectrum is the Fourier transform of an image which has been multiplied by two shifted copies of itself. The bispectrum of fixated regions was more circular in shape in comparison to the bispectrum of random regions, suggesting that the eye selects regions of an image that contain strong statistical dependencies between frequency components of different orientation. These regions include curves, occlusions, and corners, rather than straight lines and edges.

A similar study by Reinagel and Zador [1999] showed agreement regarding the higher spatial variance (local contrast) of fixated regions, but found conflicting results regarding second-order statistics. Subjects were eyetracked with an IScan RK-416 infrared pupil tracking system as they viewed grayscale images subtending 23 degrees of visual angle. Square image patches (1 x 1 degree) were extracted at gaze positions every 20 milliseconds, a small percentage of which are samples during saccades. When analyzing the spatial correlation between central pixels at the points of fixation and neighboring pixels, they found significantly lower correlation in the fixated regions compared to randomly chosen locations.

Parkhurst and Niebur [2003] repeated the analyses conducted by Krieger, et al., and Reinagel, et al., to determine why conflicting results were found. Four large databases of images (300 images total) were used, containing images of: fractal patterns, natural landscapes, buildings

and city scenes, and home interiors. Four participants were eyetracked (IScan RK-416) while they viewed images subtending 30 degrees of visual angle horizontally and 22.4 degrees vertically. Images were displayed at a resolution of 640 x 480 pixels in 16-bit color. Each image was displayed for 5 seconds, and images were blocked according to image type.

Image patches at points of fixation were extracted and used to create an 'image ensemble' for further analysis. Image ensembles were also created from randomly generated locations for each image. To account for any misleading results caused by a central bias of fixations, image ensembles created by participants' fixations were applied to random images (image-shuffled ensembles). Ensembles were also created using various sizes of round image patches. Average contrast of the image patches were found for each image database, for all image patch sizes, and it was found that the amount of local contrast varied with image type as well as image patch size, as shown in Figure 5. The local contrast of fixated regions was found to be significantly higher than in both the image-shuffled and uniformly random ensembles. Interestingly, the magnitude of this difference also varied with image type. For images that contained a higher amount of local contrast, larger differences were seen between the local contrast fixated regions compared to the image-shuffled. Also, this difference was found to be largest for image patches with a radius of about 1 degree.

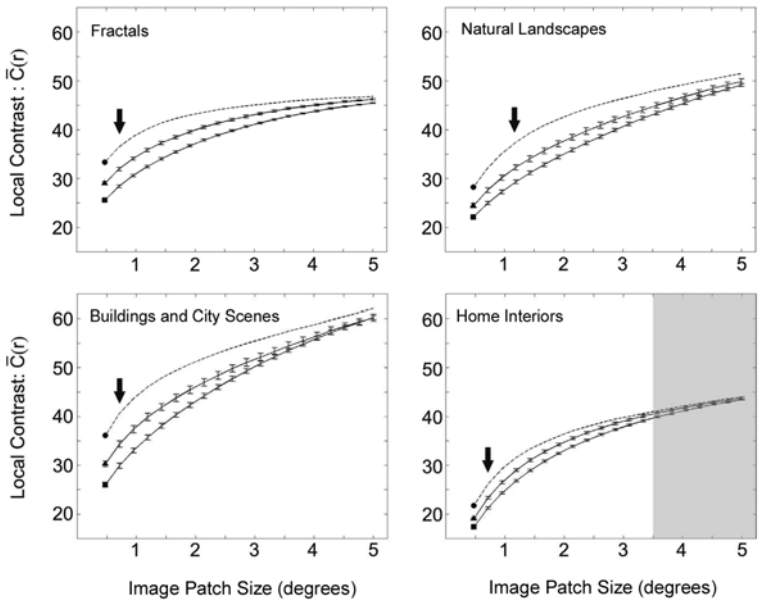


Figure 5: Average contrast in each image database as a function of image patch size in the participant selected image ensemble (dashed line; circle), the uniformly selected image ensemble (solid line; square) and the image-shuffled ensemble (solid line; triangle). Error bars represent 1 standard error of the mean contrast. Arrows indicate maximal difference between participant-selected and the image shuffled ensembles. Regions with non-significant differences between the participant-selected and the image-shuffled ensembles are lightly shaded [Parkhurst and Niebur, 2003].

The two-point correlation between pixels at the center of the image patch and neighboring pixels was calculated. This analysis showed a significantly lower correlation value for fixated regions for all image types, which is consistent with the findings of Reinegal, et al. Again, the largest amount of decorrelation was found between pixels at the center of gaze and pixels at a radius slightly larger than 1 degree. The magnitude of difference in correlation differed from the results of Reinagel, et al.; the authors suggest that this is a result of differences in analysis techniques. Reinagel used points of gaze, including samples during saccades when perception is suppressed, rather than points of fixation. Also, Reinagel averaged over horizontal and vertical orientations only, rather than all orientations.

To compare this to the results of Krieger, et al., spatial autocorrelation was performed with the image patches. This method is the spatial-domain equivalent to the Fourier analysis performed by Krieger. It was found that the correlations resulting from the autocorrelation technique were higher than the two-point correlations. Also, the differences between correlations of the fixated and random image patches were smaller. The authors claim that the difference in analysis techniques and small image sample size are the primary causes of the discrepancy in results. The autocorrelation analysis measures the average correlation between any two points in the image patch separated by a certain distance, whereas the two-point spatial correlation measures only the correlation between the central pixels and others displaced by a certain distance. This also means that the autocorrelation technique is inherently influenced by image patch size. However, the comparison of central pixels at the point of fixation with the surrounding pixels is not appropriate given the accuracy of the eyetracker, which is on the order of 0.5 to 1 degree of visual angle. The authors show that for various image patch sizes, the difference in correlation between the fixated regions and random regions decreases with increasing image patch size.

2.5.2 Discrimination (Classification) Images

Rajashekar, et al. [2002] conducted a visual-search experiment in which subjects searched for a simple shape (shown in Figure 6) that was embedded in random noise. The noise used had a frequency spectrum amplitude that was inversely proportional to the frequency, also called "1/*f* noise." Three subjects were eyetracked using an SRI Generation V Dual Purkinje eyetracker, which has an accuracy of < 10' of arc. Horizontal and vertical eye positions were sampled at 200 Hz.

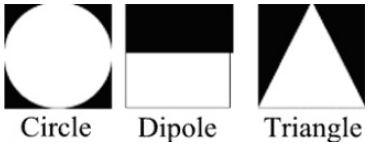


Figure 6: Targets used in visual search experiment. From [Rajashekar, et al., 2002]

At each fixation location, the surrounding patch of 128 x 128 pixels was extracted (4 x 4 degrees of visual angle). The patches of the noise image were averaged to produce a "discrimination image" (also called a "classification image" in the field of psychophysics). This image will reveal any image features that the human visual system uses as a "filter" or "template" to search for the target. The left side of Figure 7 shows the resulting discrimination image when a subject searched for a dipole. Pixel values that were not significantly different from the mean were set to an average gray value in order to more easily visualize pixels that are significantly different from the mean. The right image in Figure 7 shows the discrimination image created from an equal number of random image locations. Interestingly, a structure emerges in the discrimination image that contains features present in the preview target (dark upper region, bright lower region). This effect extends to other targets as well, as shown in Figure 8.

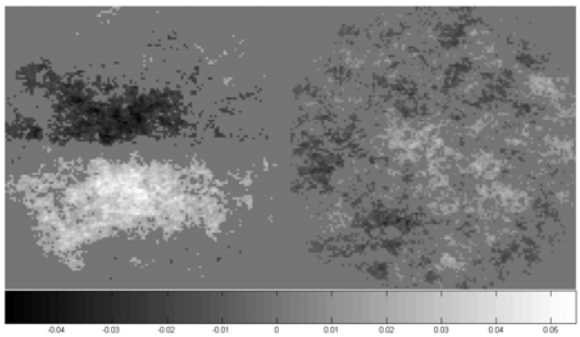


Figure 7: Discrimination image from fixated locations during dipole search (left), discrimination image from an equal number of random locations (right). Pixel values that were not significantly different from the mean have been set to gray. From [Rajashekar, 2002].

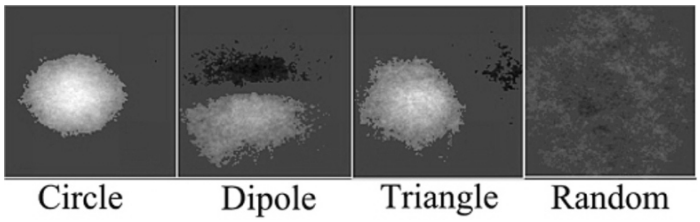


Figure 8: Discrimination images created from fixations (from one observer) made during search for each of the three different search targets, as well as an image produced from random locations. From [Rajashekar, 2002].

The evidence presented in this paper suggests that the visual system may use a linear filtering technique to guide saccadic eye movements to regions of the image that may contain a subset of features in the target, thus suggesting that the visual system is primarily driven by the bottom-up features. In a more recent paper [Rajashekar, et al., 2004], the discrimination images are shown for the other two observers; the images do not contain the exact same spatial structure as those presented above, showing idiosyncrasies between observers' strategies. The authors also expand on the previous analysis by filtering the noise image with each observer's discrimination image. The result is a correlation map, where peaks indicate where the image closely matched the kernel. Clusters of fixations from all three observers were overlaid on the correlation map to qualitatively visualize how well the map predicts locations of fixations. In the examples given, the map predicts locations of fixation very well. The Kullback-Leibler distance was used as a metric to quantify the similarity between the map and a fixation density map. The resulting Kullback-Leibler distances were found to be smaller that those measured from maps created by using the 'random' discrimination image as a kernel. Additionally, the distance was about the same value as measured using a map generated by filtering the original target with the noise image. These results show that in this task, the visual system is not random, but uses a very efficient filter that resembles (in spatial structure) the target. Also, observer's discrimination images are good fixation prediction kernels. However, the design of the experiment did not allow for any interference from top-down or cognitive features. It is difficult to say whether this linear filtering strategy may be used by the visual system when searching complex, real-world scenes.

2.5.3 PCA of natural images

A study by Hancock, Baddeley, et al. [1992], aimed to extract the principal spatial components found in images of natural scenes. In these types of images, nearby pixels will often be part of the same object and are therefore statistically related. Principal components analysis is a common method for analyzing the inter-relations between variables. A set of 15 images of natural scenes was used; these included people, animals, plants and terrain, but did not include any obvious horizons or man-made structures. The images were grayscale, and 64x64 pixel pieces of the original 256x256 pixel images were chosen at random for analysis. The image samples were masked with a Gaussian window in order to remove any effects created by the edges of the image windows. To find the principal components (PCs), the eigenvectors of the correlation matrix of the 4096 variables (pixels) were found using a neural network technique. The resulting PCs are shown in Figure 9. The procedure was also performed on copies of the input images that were rotated by 45 degrees, and resulted in the same PCs also rotated by 45

degrees, suggesting that the horizontal and vertical orientations of the first few components were artifacts only of the image content itself. The PCs were also found to be independent of scale of the input images. The results were verified using a larger set of 40 images, containing some man-made structures as well as more natural images, and produced PCs of similar shape and output variance.

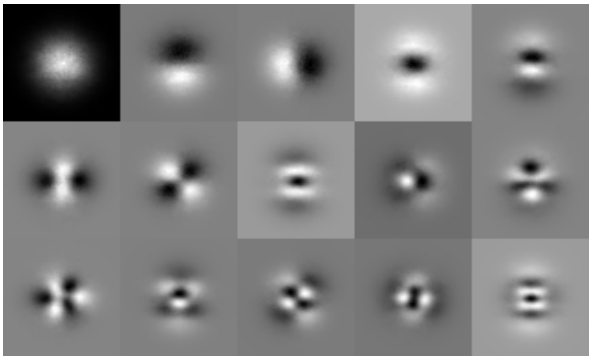


Figure 9: First 15 principal components of natural images, numbered from left to right, top to bottom [Hancock, et al., 1992].

The procedure was also performed on images of text documents of different scales. This resulted in PCs resembling Gaussian-modulated sinusoidal patterns oriented either horizontally or vertically. These PCs also changed with scale of the input image, unlike those of natural images.

The first few PCs of natural images have the general form of receptive fields that are thought to be used in the early stage of visual processing in humans as well as some animals [Palmer, 1999, Chapter 4]. A natural question is whether the early visual cortex is performing such PCA. PCA is an efficient method of detecting the most important information due to the fact that all PCs are mutually orthogonal, or independent. However, the authors claim that it seems more likely that various constraints operating in the primary visual cortex lead it to do something other than PCA.

2.5.4 Weighting of Features

In the experiment by Rajashekar, et al. [2002], presented above, a similar PCA was performed on the image patches at fixation locations. When comparing the first few principal components between the circle search and the dipole search, the spatial features were similar. However, in the dipole search, a horizontal pattern produces a larger PCA score (weight) than the

other features. For the circle search, the first horizontally oriented PC and vertically oriented PC were nearly equally weighted. This suggests that during a simple visual search task in random noise, the visual system may selectively weight 'basis' features or channels that match the target.

In an experiment conducted by Navalpakkam, et al. [2004], observers were eyetracked as they searched for a target amongst distracters. The target consisted of one horizontal line, and one vertical line conjoined in an "L" shape, shown in Figure 10. The distracters were designed to either contain the same amount of features as the target, more of those features, or new features. Relative numbers of fixations falling on each of the three types (Same, More, and New) were compared. It was found that subjects fixated more on the Same distracters than on the More and New type. This indicates that the visual system may actually suppress certain target features when performing a simple visual search task.

Figure 10: Target and distracters used in search experiment. From [Navalpakkam, 2004].

2.6 Models of Eye Movements and Visual Attention

In recent years, a large amount of work has been done in the areas of active vision modeling for use in artificial vision systems. The goal of these models is to locate areas of interest within a scene, and then shift attention (or computational resources) to those areas sequentially over time. Each model begins with low-level information of the visual scene, typically pixels of an image of a natural scene. By processing these images in a way that mimics early visual processes, 'saliency' can be determined at each spatial location of the image. In this sense, these models are purely "bottom-up" in that they do not take into account any high-level or cognitive information about the scene, the viewer's goal or previous experiences. As discussed earlier in this chapter, these factors have a significant influence on a viewer's behavior. A few researchers are beginning to investigate how cognitive or "top-down" information about the scene or viewer's task can be integrated into these models.

2.6.1 Feature integration and retinotopic maps

There exists physiological evidence of the presence of retinotopic feature maps which contain information about the presence and location of certain features. These retinotopic feature maps code different visual properties including redness, greenness, orientation of edges, etc., while retaining spatial location [Treisman, 1980].

2.6.1.1 Saliency Maps

A computational model first suggested by Koch and Ullman [1985] and implemented by Parkhurst, Law & Niebur [2002] as well as Itti & Koch [1998, 2000], utilizes color, intensity, and orientation feature maps in order to create an overall saliency map of a visual scene. Given an input digital image in RGB color space, feature maps for each of the three categories are generated at various resolutions, and then combined to make three feature maps. The three maps are then combined to create a master topographical saliency map. A 'winner-take-all' mechanism locates the area of highest saliency, and directs attention to that location. The saliency of that area is then depressed, or inhibited, and the next area of high saliency is then chosen, and so on. Attention is directed to a new location every 30-70 milliseconds, and attended locations are inhibited for 500-900 milliseconds [Itti, et al., 1998].

This model is inspired by the behavior and neural architectures of the early primate visual system, but all computation is performed in RGB space. Although it is biologically plausible, it does not correspond well with human saccadic eye movements in natural tasks, even when a person is freely viewing a scene [Canosa, 2003]. However, the model is widely used as a basis for many active computer vision systems because it is dependent only on low-level visual information.

Recent work has expanded on the model to incorporate more top-down influence to mask or boost saliency of various regions. Walther [2002] uses a modulated saliency map to determine the location and size of regions that are likely to be objects in a scene. Instead of using the final topographical saliency map, Walther uses the most influential feature map as a mask to modulate saliency. The results show promise in quickly identifying regions of probable objects in which computational resources can then be allocated for further object recognition.

2.6.1.2 Conspicuity Map

A model developed by Canosa [2003] built upon the salience model and incorporates high-level information and task-specific constraints. The first piece added to the salience model was a preprocessing step that converts the input RGB image into a representation in terms of the early physiological responses of the human visual system. The RGB image was first converted to XYZ tristimulus values, which take into account the spectral properties of the display device as well as the color-matching functions of the CIE Standard Colorimetric Observer. The XYZ values were then converted into long (L), medium (M) and short (S) wavelength cone responses, as well as rod responses. The L, M, and S images were used to create two color-opponent channels and one achromatic channel. The achromatic channel was also weighted with the rod response, and used for subsequent spatial processing. The two color channels were combined to create one color feature map.

In parallel, the oriented edge and "proto-object" maps were computed from the achromatic channel. The orientation map was created through convolution with four oriented Gabor filters of different spatial frequencies (via Gaussian Pyramid). At this point, the responses of different spatial frequencies were weighted according to the contrast sensitivity function of the human visual system. The proto-object map was another addition to the salience model, and found potential objects in the scene by detecting texture from edge densities. First, a local estimate of the background was subtracted (figure/ground segmentation), the image was then thresholded and a Canny edge operator was applied. Morphological operations were performed to fill holes and smooth boundaries. The result was a binary mask that was used to enhance regions of the saliency map that are potential objects, and inhibit uniform regions.

In Canosa's model, the color (C), intensity (I), and edge (E) maps were combined with equal weighting to create the basic saliency map, referred to as the CIE map. The CIEP map included the proto-object map (P) as an additional channel as well as a binary mask. Finally, the C_Map was the CIEP map with unique weightings for the different channels, given by Equation (1) below. The weights were found via a genetic algorithm trained on data collected from eyetracking experiments; they are image specific and are used only to show that the performance of the model can be improved when the channels are weighted. Figure 11 shows an example input image of a room containing a copy machine, computer, and office supplies. The CIE map found the ceiling light, wall, floor, and some objects on the desk to be most salient, whereas the

C_Map allotted high conspicuity values to the computer monitor and other machines more likely to be fixated on.

$$C\_Map = \left(w_1 \cdot C + w_2 \cdot I + w_3 \cdot E + w_4 \cdot P\right) \cdot w_5 \cdot P \tag{1}$$

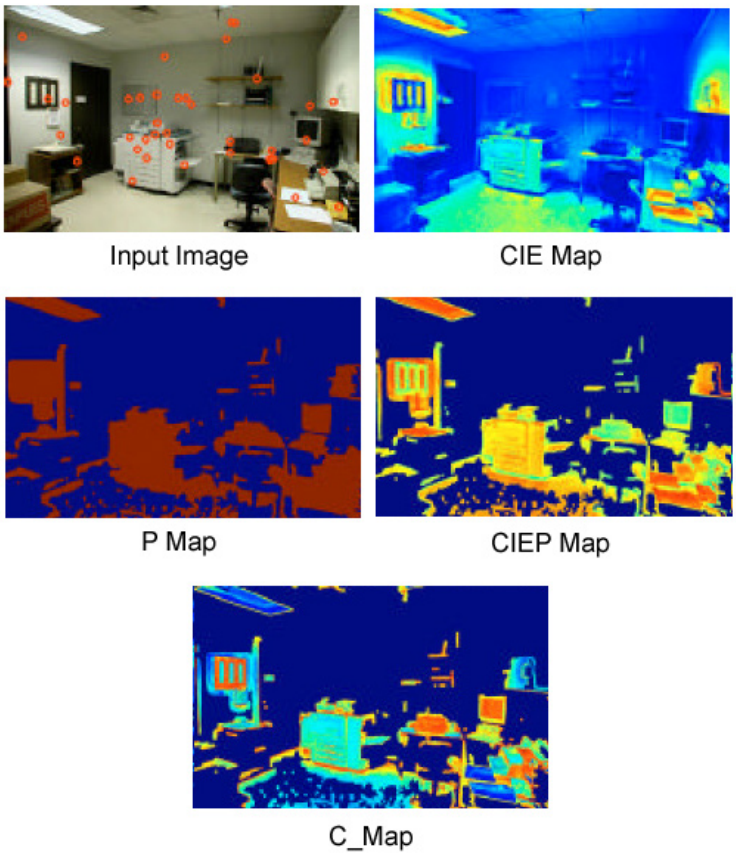


Figure 11: Example of input image and corresponding maps. Fixation locations for one subject during freeview are overlaid on the input image. Adapted from [Canosa, 2003], pg. 208.

The four maps were compared to fixation data from 11 subjects who each freeviewed 152 images. The conspicuity values of the fixation locations (1x1 degree window) were averaged, and then divided by the mean value of the map for each image to create the F/M (fixation to map) ratio. An F/M ratio near 1 means that the areas of high values in the map did not correlate with fixation locations, since a set of random locations would perform just as well. As shown in Figure 12, the low-level saliency map (CIE) produced an F/M ratio near 1. Both the P map and CIEP map produced F/M ratios which were considerably higher. The weighted C_Map produced the highest mean F/M ratio, which was greater than 2. Because the C_Map enhances the

computed conspicuity of potential objects in the image, it naturally preserved any location bias revealed in the fixation data, unlike other models that impose an artificial central location bias.

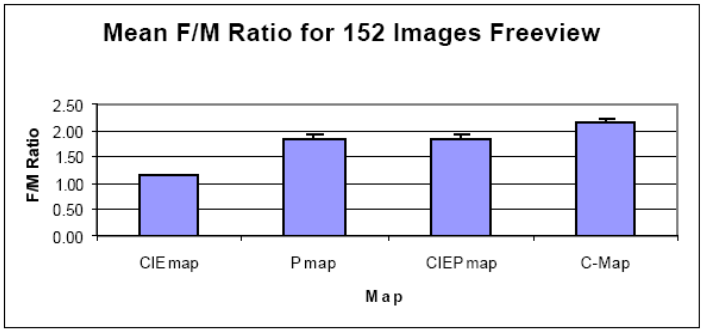


Figure 12: Mean F/M ratios for four different maps, averaged across 152 images. Taken from [Canosa, 2003], pg. 206.

Figure 13 shows F/M ratios for the C_Map of different locations of the image, compared across tasks. Three different instructions were given as the subjects viewed the image of a hallway scene: "Put something in the garbage can," "Find a bathroom," and "The fire alarm just went off." The F/M ratios were computed using only one of nine sections of the C_Map. In this example, there is a strong central bias, demonstrated by the high F/M ratio for section 5 for all tasks. However, the section of the image containing the exit sign produced a high F/M ratio for the "Fire alarm" task. Also, sections 4 and 6, where bathroom doors are likely to be found, produced high F/M ratios for the "Find a bathroom" task. It is suggested that having a large database of empirically determined expected locations for various tasks would provide a way of introducing top-down information into computational models of attention. Expected locations of novel scenes and tasks could be derived from existing information about similar scenes and tasks.

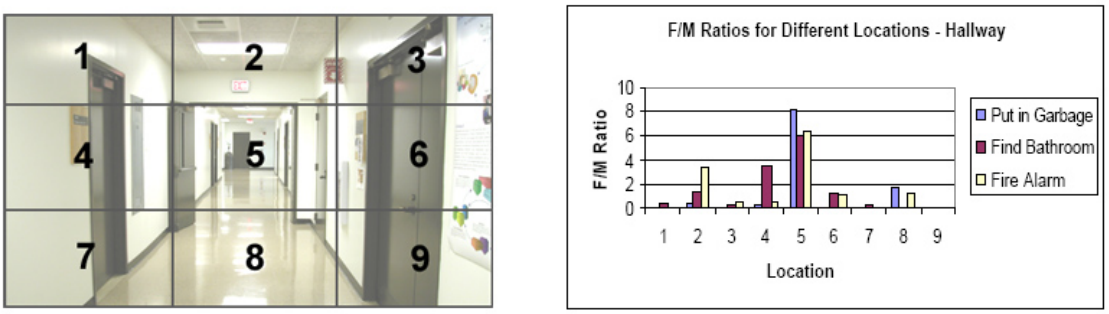


Figure 13: Task influences on fixation locations for the hallway scene. The numbered sections of the image are shown on the left, and the F/M ratios of those sections are shown in the graph on the right. Instructions heard before a subject viewed the image included: "Put something in the garbage can," "Find a bathroom," and "The fire alarm just went off." Adapted from [Canosa, 2003], pg. 223.

2.6.1.3 Region-based importance maps

Wilfried Osberger [1998] has developed a model of assigning perceptual importance to regions of an image. The model provides a robust prediction of a viewer's locus of attention across a wide range of image content. Evidence has shown that visual attention is directed toward objects rather than location, so the model begins with a rough segmentation of a scene into homogenous regions that are likely to be objects or parts of objects. Computations are performed on each region in the image, and an importance factor for various features is assigned. These features include the region's size, shape, intensity contrast, color contrast, foreground/background content, location, amount of skin tone, etc. For each feature, an importance map is produced, which are combined via a weighted average to create a master importance map. To define the weights of each feature in the final map, an eyetracking experiment was conducted to find the regions that are most perceptually important to humans. The features receiving the highest weights included location, foreground/background, and amount of skin tone. The algorithm was trained on one set of images, and tested on a different set. When the testing set was compared to human fixation locations, Osberger reports that 75% of viewers' fixations occurred in the 30% area estimated as most perceptually important by the algorithm.

The importance map algorithm is very modular and can easily incorporate additional features such as motion, which expands the algorithm's application into motion video sequences. Also, weights can be easily adjusted so that the algorithm can be applied to various tasks, such as a visual search task in which the features of the target are known. It also gives insight into high-level perceptual features that guide saccadic eye movements, including the location of the region, whether it is part of a foreground object, and whether it contains skin tones. These features proved to be more important in guiding eye movements than a region's intensity contrast with neighboring regions, and other features used by the saliency model.

2.6.1.4 Oriented spatial filter decomposition

Rao, Zelinsky, et al., [1996] have constructed a model of saccadic targeting in visual search in order to gain insight into possible computational mechanisms that underlie the guiding of saccadic eye movements. An eyetracking experiment was conducted in which subjects searched for an object in a natural scene (items on a table). The subject was shown a preview image of the target for 1 second, and then instructed to determine whether the target object was present in the following scene. Eye movement records of four subjects showed that several

saccades were typical, each moving closer to the target object. Rather than using only bottom up saliency models to generate scan paths, this model uses bottom-up scene representations in conjunction with previously memorized top-down object representations. This iconic object representation is a feature vector containing the responses of an image patch to a set of oriented spatial filters (derivatives of Gaussians) at several different spatial scales. Saliency images are made that compare the similarity between regions of the image with the known target feature vector. The algorithm begins by analyzing the saliency image of the largest scale, or coarsest spatial resolution. A weighted population averaging scheme chooses a location that is the center of gravity between possible target locations. The process is iterated with smaller scales (increasing spatial resolution), until the target is fixated.

The model assumes a coarse-to-fine matching mechanism, inspired by the variation in resolution of the retina. Comparison between the model's saccades and subjects' eye movements show "remarkably good correspondence," producing very similar histograms of placement of the first three fixations.

2.6.1.5 Color histogram backprojection

A study by Swain, Kahn and Ballard [1992] explores low resolution cues that are suitable for guiding saccades, and have shown that color cues are an efficient method of locating known objects by computational algorithms. An experiment was conducted in which a computer vision system performed visual search using only low-resolution color histograms as a guide in finding the target. The technique is called histogram backprojection and effectively re-indexes the colors of an image with its ratio histogram value. The ratio histogram is defined as the color histogram of the target divided by the histogram of the image. The backprojection process creates a saliency map of sorts, in which colors that are not present in the target are deemphasized. The map is smoothed and dilated via convolution with a mask which is a circle of the same area as the object. A spatially variant sensor first locates the peak value of the map and performs further processing of high resolution color information to verify whether the object matches the target. If not, the sensor chooses the next highest peak to evaluate. The algorithm proceeds in a series of fixations and saccades, and its performance is measured by the number of saccades needed to find the target object. The image used consisted of 32 un-occluded target objects consisting of cereal boxes, items of clothing, and other products. Out of the 32 items, 29 were correctly identified by the algorithm. Most were detected in 1-3 saccades, and the maximum number of saccades needed was 6.

2.6.1.6 Other algorithmically defined regions-of-interest (ROIs)

Privitera and Stark [2000] conducted a survey of a wide variety of image processing algorithms that can be used to define regions-of-interest in images. These algorithms include simple convolutions with various masks of different shapes, including an x-like shape, a center-surround receptive field, Laplacian of the Gaussian, and Gabor masks of different orientation. Also used were measures of spatial frequency, energy, entropy, contrast, as well as transforms including discrete wavelet and discrete cosine transforms. The goal of the study was to propose an engineering (rather than biologically plausible) approach to modeling vision and eye movements.

The algorithms were each applied to a set of test images for which human fixation data was collected. Comparisons between the first seven fixations and the seven most important regions (defined by each algorithm) were compared both spatially and temporally using a defined similarity metric. Different human subjects looking at the same picture yielded a spatial similarity value of 54%, which is used as a guideline when comparing the performance of algorithms to humans. Temporally, the similarity value fixated regions between subjects for one image was only 28%, which agrees with previous findings [Buswell, 1935; Yarbus, 1967] showing that humans tend to look in similar places but not in the same temporal order.

After comparing the performance of the algorithms, it was found that the wavelet transform was most efficient in matching human ROIs (producing a high spatial similarity value) for several classes of images, most likely due to its implicit multiresolution analysis. The symmetry metric performed well on general images such as paintings, and the contrast metric correlated well for images of terrain. The discrete cosine transform performed poorly when compared to human fixation data.

2.6.2 Performance Metrics

2.6.2.1 Chance-Adjusted Cumulative Probability (CPa)

The chance-adjusted cumulative probability (CPa) metric was developed by Parkhurst, et al. [2000] in an effort to make a distribution- and scale-independent metric for assessing the performance of a saliency model. The absolute magnitude of the stimulus dependence of visual selection is difficult to estimate given that the range of salience values are scaled or normalized on a per-image basis. In the database of images used by Parkhurst, et al. (which contained natural landscapes, buildings and city scenes, home interiors and fractal images), the mean salience for

an image was usually between 30 and 35 units on a scale of 0 to 100. The histograms of saliency values per image were in general positively skewed, resembling a Raleigh or log-normal distribution.

For each fixation on an image, the cumulative probability of observing a salience value less than the value at the fixation location is calculated. For each image, the cumulative probability of observing a salience value below the mean value of the map was also calculated, giving the cumulative probability expected by chance. This value was usually above 0.5 because the distributions were positively skewed. The chance-adjusted cumulative probability (CPa) is given by Equation (2). This metric was then used by Parkhurst, et al., to estimate the magnitude of the influence of bottom-up stimulus properties on visual selection. A value of 0 indicates that fixations were essentially random, and a value of 1.0 indicates that fixations fell only on regions of very high salience as defined by the model.

$$CPa = \frac{(\textit{observed cumulative probability} - \textit{chance cumulative probability})}{(1.0 - \textit{chance cumulative probability})} \qquad (2)$$

2.6.2.2 F/M ratio

The metric used by Canosa [2003] is also designed to be distribution- and scale-independent. For each fixation location, the value of the salience map is extracted. That value, F, is then divided by the mean salience value of the entire map, M. This value is referred to as the F/M ratio. If the ratio is near 1, then the model did not match locations of observed fixations more than expected by chance. This metric is somewhat sensitive to the mean of the salience map, but not to the range that the map is scaled to. For example, a map for a specific image has a mean value of 20. An observed (fixated) location has a value of 60, producing an F/M ratio of 3. Now imagine a second map, with the same shape of distribution as the first image, but shifted toward a higher mean of 40. An observed value of 80 would produce an F/M ratio of 2. In both cases, both the difference in salience units and the chance-adjusted cumulative probability remain the same.

To compare the behavior of the CPa and F/M Ratio metrics, six salience histograms were simulated. These included normal, log normal, Rayleigh, hyperbolic tangent, and uniform distributions, as shown in Figure 14. Figure 15 shows a comparison of the CPa and F/M Ratio metrics for each distribution. For each possible saliency value between 0 and 100, the metrics were computed and plotted to make one line per distribution. Values for the CPa converge to 1 as the saliency value increases, whereas the maximum possible F/M ratio is completely dependent

on the mean saliency. The rate at which the CPa increases is dependent on the shape of the distribution; this slope changes if the distribution is not uniform, whereas the rate of increase of the F/M Ratio is constant within each distribution.

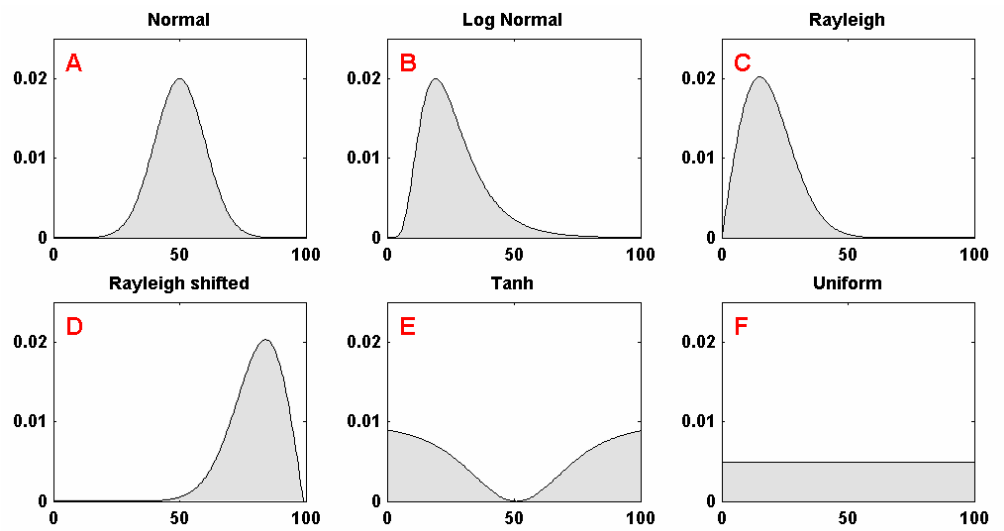


Figure 14: Simulated histograms of salience values

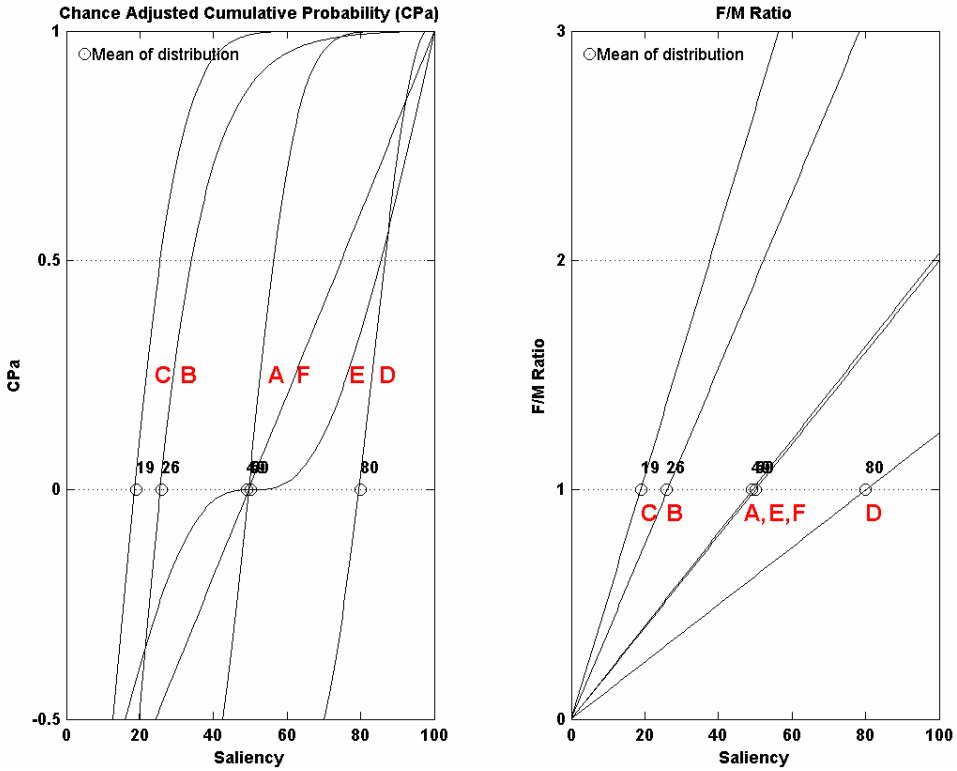


Figure 15: Comparison of the behavior of CPa and F/M Ratio metrics for the histograms shown in Figure 14. The mean salience of each distribution is marked with a circle.

2.6.2.3 ROC Curve Area

A common analysis tool for signal detection experiments in the field of psychophysics is the Receiver Operating Characteristic Curve. This curve shows the number of true positives (or "hits") compared to the number of false positives ("false alarms") for a particular binary threshold that separates two distributions, as shown in Figure 16. These distributions are typically thought of as a noise distribution, and a signal-plus-noise distribution. Example ROC curves are shown in Figure 17. For threshold A the number of hits is large while the number of false alarms is small, indicating that the two distributions can be discriminated with few errors. For threshold B, the number of hits is only slightly higher than the number of false alarms, meaning that the signal and noise distributions can not be discriminated accurately. As distributions are more easily distinguishable, the area under the ROC curve increases.

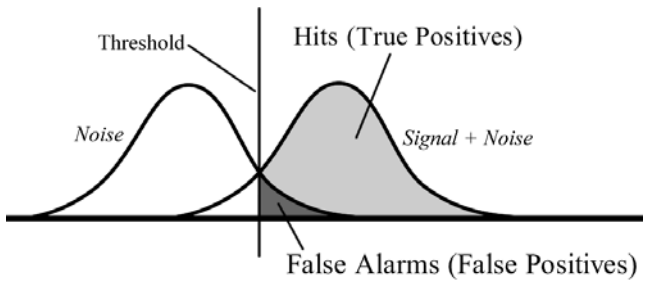


Figure 16: Example probability distributions separated by a threshold

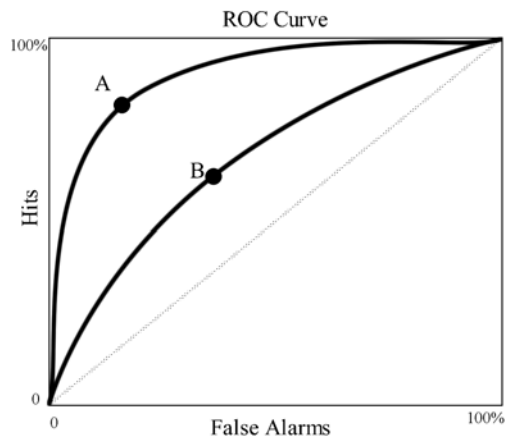


Figure 17: Example ROC Curves

For use as a performance metric of predictive models, [Tatler, et al., 2004] the distribution of saliency values that are not fixated can be thought of as the noise distribution. The distribution of saliency values at fixated locations can be considered the signal plus noise

distribution. Given a set of fixation locations, and an equal number of randomly chosen locations ("non-fixated"), a hit is defined as when the saliency value at a fixated location is above some threshold. A false alarm is when the saliency value at a non-fixated location is also above the threshold. Assuming the goal of the model is to assign high saliency values to regions where fixations are likely to occur, while assigning low values to regions where fixations are not likely to occur, the area of the ROC curve will be high if the model accurately predicts locations of fixation.

2.6.2.4 Chance-Adjusted Saliency "Accumometric" (CASA)

The Chance-Adjusted Saliency Accumometric (CASA), used by Carmi and Itti [2004], is defined as the weighted sum of the binwise differences between the frequency of salience values of human fixations and random fixations. In other words, the histogram of salience values of random locations in an image (or video frame) is subtracted from the histogram of salience values of actually fixated locations. This difference histogram is then weighted by the value of the salience at that bin. For this metric, salience maps are normalized to range from 0 to 1. The weighting step serves to attenuate any differences in low salience values. A high CASA value indicates that a.) human fixations differ from random, and that b.) human fixations fall on areas of high salience.

Figure 18 shows three examples; in each example the histogram of random fixations is the same. The first column is an example where the human fixation locations corresponded with high salience values, producing a high CASA value of 0.29. In the second column, the histogram of human fixations is uniform, producing a CASA value of 0.15. The last column shows an example where human fixations fall mostly on medium to low salience values. From this histogram, it seems that the example map does not do a very good job of assigning high values to locations where people are likely to look; however, because of the artificial weighting of each bin, the CASA value is very close to that of the second example. The weighting step creates ambiguity, and may therefore not be an appropriate metric of the performance of a prediction model. Additionally, the number of bins used will change how certain values are weighted, and will thereby affect the CASA value, as illustrated in Figure 19.

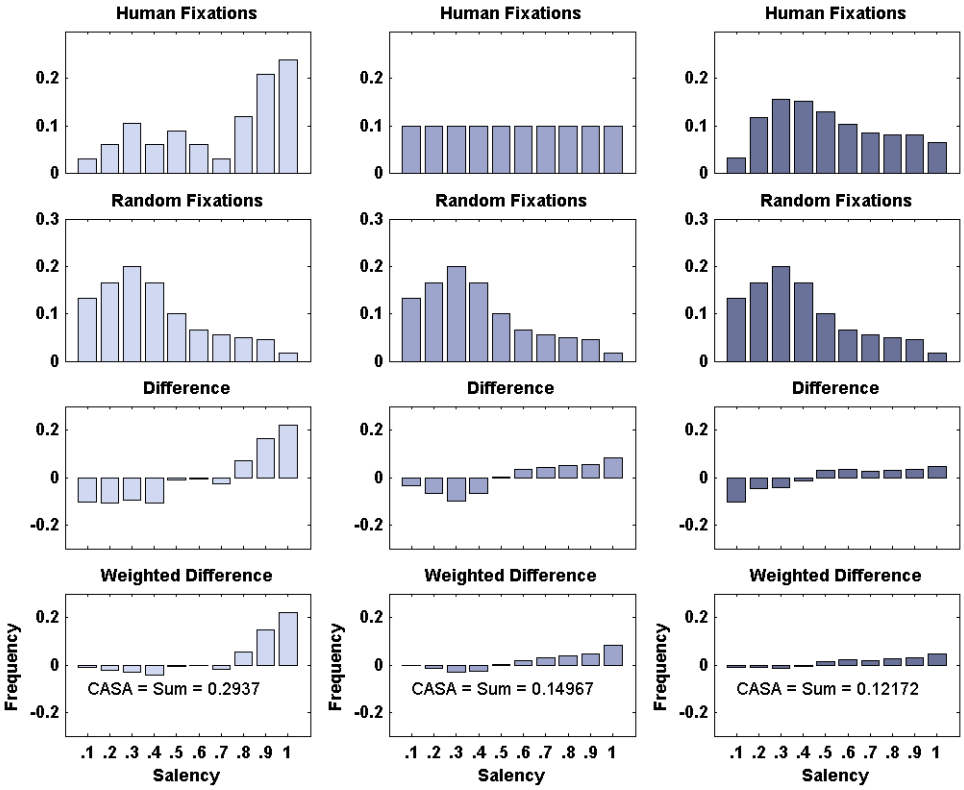


Figure 18: Examples of computation of CASA values. In each column, the first row represents the histogram of saliency values at locations of human fixations. The second row is the histogram using uniformly random locations. The third row is the difference between the Human and Random histograms. The last row is the difference histogram weighted by the saliency value of that histogram bin.

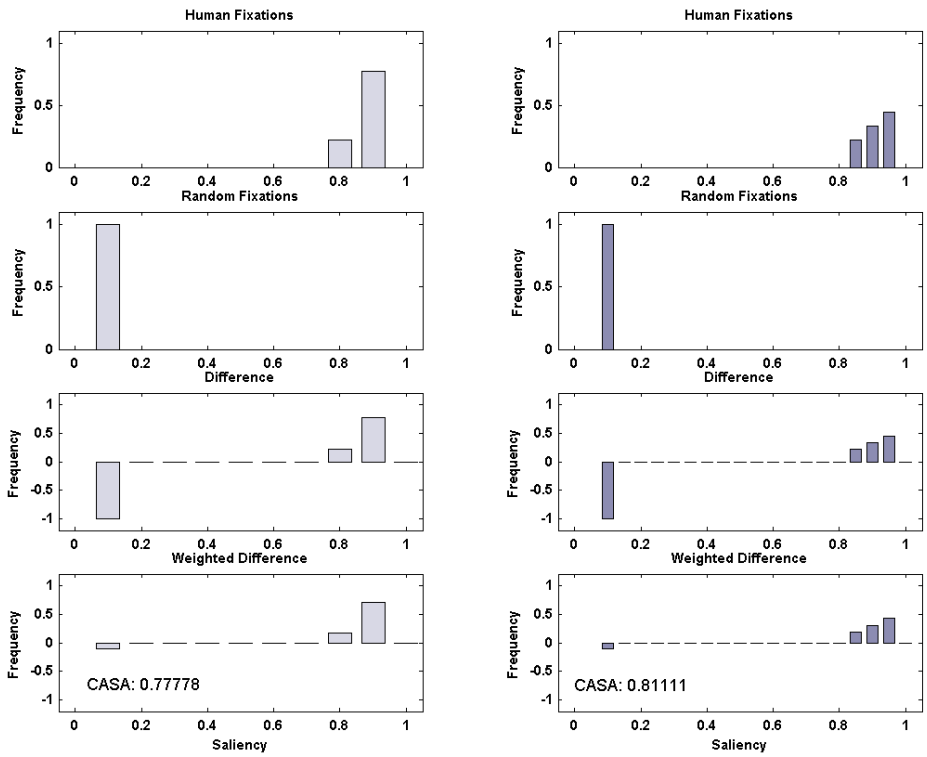


Figure 19: Example in which the number of bins used to create the histograms influences the resulting CASA value

Chapter 3

3 Approach

3.1 Overview

Research summarized in Chapter 2 shows that eye movements are highly tailored to the observer's perceptual goal, environment, motivation, experience, and expectations. Therefore, creating one general model of visual attention is not feasible at this point in time. Models that mimic human visual routines can be made more flexible and adaptive by incorporating prior knowledge of the task and environment, and the corresponding characteristic eye movement behavior.

In this research study, properties of regions of an image that are chosen by saccadic eye movements will be studied and compared in specific tasks. In [Canosa, 2000], it was shown that different tasks could be grouped according to the temporal characteristics of eye movements. Similarly, relationships between features of fixated regions and specific tasks will be investigated.

This chapter will give an overview of the eyetracking methods used in the experiments discussed later. Methods of analyzing image content at locations of fixation will also be described.

3.2 Bright Pupil Configuration

The eye movement records of the experiments presented in this thesis were obtained using a bright pupil video-based eyetracking system. This type of system makes use of the fact that the back of the retina is highly reflective in the near-infrared. By illuminating the eye with

an infrared (IR) LED that is coaxial with an IR-sensitive video detector, an image of the eye can be obtained in which the pupil appears as a bright circle, illustrated in Figure 20. The illumination is also brightly reflected off of the cornea (first Purkinje reflection). The difference in position of the center of the pupil and the corneal reflection varies with rotation of the eye; this vector difference is used to calculate the line of gaze. Tracking the position of the pupil alone is not sufficient for an accurate track; if the camera moves with respect to the head, the system will mistake that shift for a shift of gaze. The position of the pupil with respect to the corneal reflection does not change significantly if the eye camera is moved.

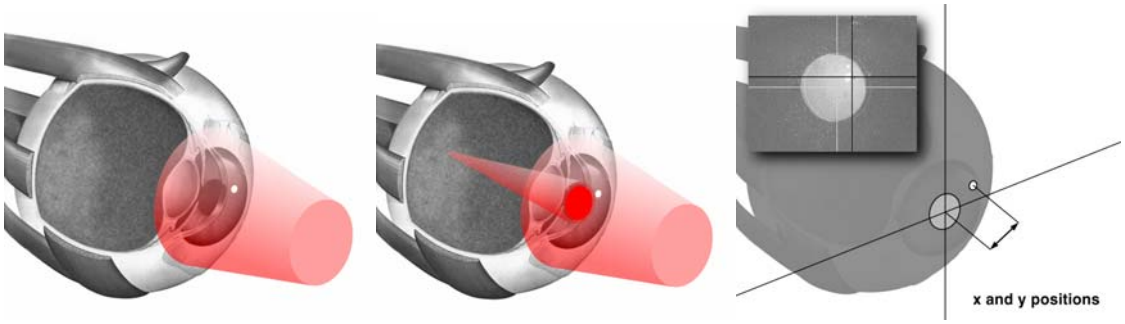


Figure 20: a) An infrared source illuminates the eye. b) When aligned properly, the illumination beam enters the eye, retro-reflects off of the retina and back-illuminates the pupil. c) The center of the pupil and corneal reflection are detected and the vector difference computed by the ASL control box. (Illustrations from [Babcock 2002], with permission.)

3.3 Video-based Eyetracking

The Applied Science Laboratories (ASL) Model 501 video-based eyetracking system, shown in Figure 21, was used for the experiments presented in this thesis. The system contains a head-mounted optics module, which is comprised of a miniature IR-sensitive CMOS video camera, an infrared LED illuminator, and a beam splitter used to align the illumination beam and camera so they are coaxial. The optical path of the IR illumination beam is folded via a mirror that reflects IR and passes visible light. This mirror directs the illumination toward the pupil and simultaneously reflects an image of the eye to the CMOS camera.

A second miniature CMOS camera records video of the scene from the subject's perspective. A small laser and two-dimensional diffraction grating was added to the headgear and is used for calibration purposes (explained further below).

The video signals from both the eye and scene camera are sent to the ASL control unit, which outputs a digital data stream of horizontal and vertical positions of the eye at a rate of 60 Hz. A video record is also recorded in which black crosshairs are superimposed over the scene

video indicating the subject's point of gaze. A small version of the eye camera video is superimposed using a picture-in-picture video mixer. This reference provides information about blinks, track losses, and extreme eye movements. The output video record is recorded to MiniDV tapes using a JVC video deck.

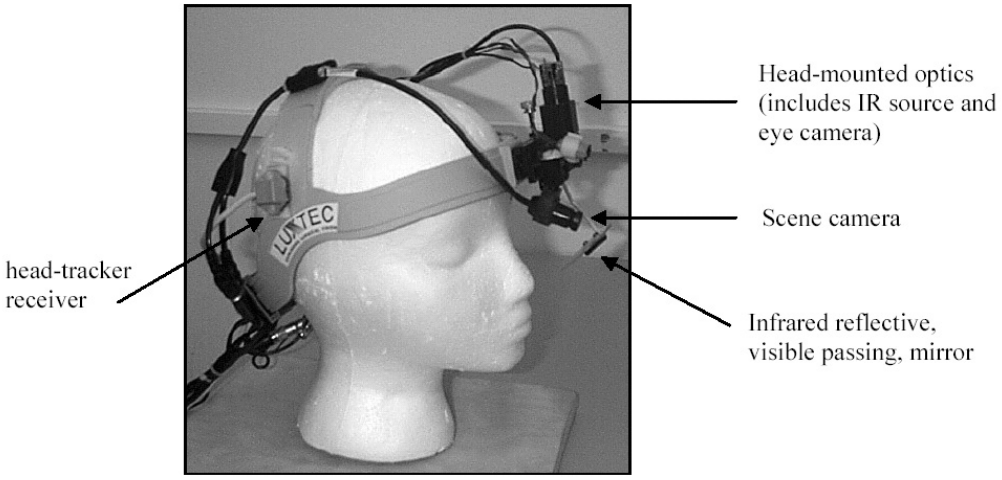


Figure 21: Applied Science Laboratories Model 501 eyetracker.

3.4 Integrated Eye and Head Tracking

The Applied Science Laboratories eyetracking system also provides support for the use of a head-tracking system. By integrating both the position of the eye in the head, and the head in space, the intersection of gaze on a predefined plane can be calculated. This allows for the tracking of eye movements without constraining the subject's head.

A Polhemus 3-Space Fastrak magnetic head tracker (MHT) was used in conjunction with the video-based eyetracking system. The MHT uses a fixed transmitter, mounted behind the subject, and a small receiver that is attached to the ASL headgear. Position (x, y, and z) and orientation (azimuth, elevation, and roll) of the receiver are reported with respect to the position of the transmitter.

When using the MHT, the output data stream from the ASL control box reports the integrated eye-head data, the intersection of gaze on a 2-D plane, at a rate of 60 Hz. This plane corresponds to the image display used in the experiments.

3.5 Eye and Head Calibration

Before each experiment, the eye and head trackers were calibrated. The calibration process can be described as 4 steps:

1. Optical alignment of the cameras and illumination source with the subject's eye and scene. Also, one laser point (produced by the diffraction grating) is aligned with a calibration point on the screen. This provides feedback to the subject when asked to hold his or her head still during calibration.
2. Definition of the nine calibration points (shown in Figure 22) with respect to the scene video image
3. Recording the vector difference between the center of the pupil and corneal reflection as the subject fixates on each of the nine calibration points
4. Recording the position and orientation of the subject's head as he/she fixates on the central fixation point

After these steps are performed, the calibration accuracy is assessed by examining the video record and the coordinates of the point of gaze on the image plane. It is possible that the video record (eye gaze only) may be correct while the integrated point-of-gaze coordinates are incorrect. If the headgear shifts on the subjects head, the integrated data will be offset. The eye-in-head position will still be correct since the relative positions of the pupil and corneal reflection do not change if the camera is moved with respect to the head (as described earlier).

Once the calibration is finished and verified, the laser is turned off and the subject is free to move his/her head naturally. Subjects were re-calibrated if needed during the experiment.

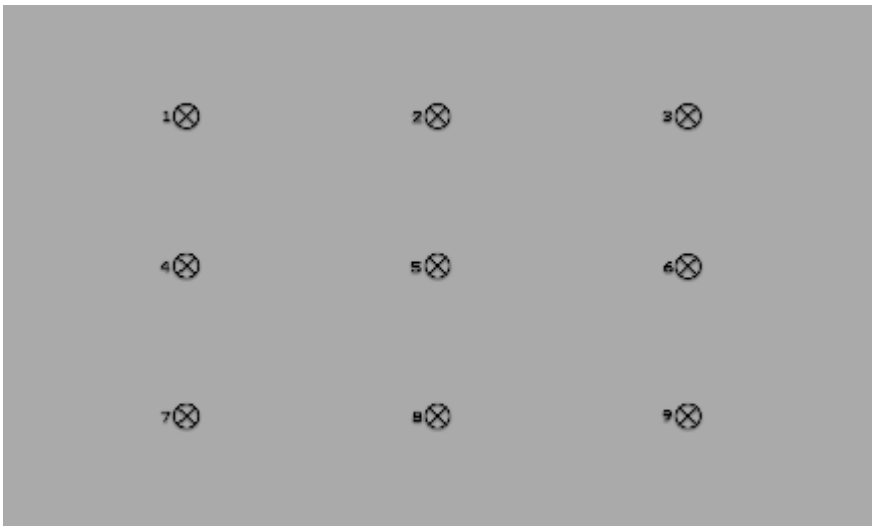


Figure 22: Set of nine points shown on the image plane during the calibration process

3.6 Fixation Location Accuracy

The accuracy of the output fixation locations is dependent on several factors. First, because the video system tracks the retro-reflection through the pupil and the corneal reflection, differences between each subject's retina and shape of cornea may affect the tracking algorithm's performance. Secondly, using the MHT in conjunction with the eye tracker introduces noise if the subject moves his or her head very quickly or very frequently during the experiment, although the classification algorithm discussed below was designed to be less sensitive to these sources of noise. In a previous study in which this system was used [Babcock, 2002], angular deviation between fixation points and target calibration points (during calibration checks) was found to range between 0.4 to 1.1 degrees across 26 subjects, with an average deviation of 0.7 degrees of visual angle. Additional uncertainty of fixation location may be introduced by the algorithm used to classify fixations and saccades, discussed below. An eye movement record was discarded if the final angular deviation was greater than 1 degree, or if more than 10% of the samples were classified as blinks or track losses.

3.7 Fixation, Saccade, and Blink Classification

As mentioned earlier, the horizontal and vertical positions of the point-of-gaze on the image plane is recorded at 60 Hz. Fixations, saccades, blinks, and any track losses were extracted from this data stream using an adaptive-velocity threshold method developed by Constantin Rothkopf. The velocity threshold chosen to differentiate between fixations and saccades changes dynamically with the amount of noise in the signal, which varies between subjects. Figure 23 shows an example of raw data samples from the eyetracking system, along with locations of classified fixations. The location of each fixation is the average location of the samples classified as being part of that fixation. This classification process outputs the times that each fixation began and ended as well as location on the image plane.

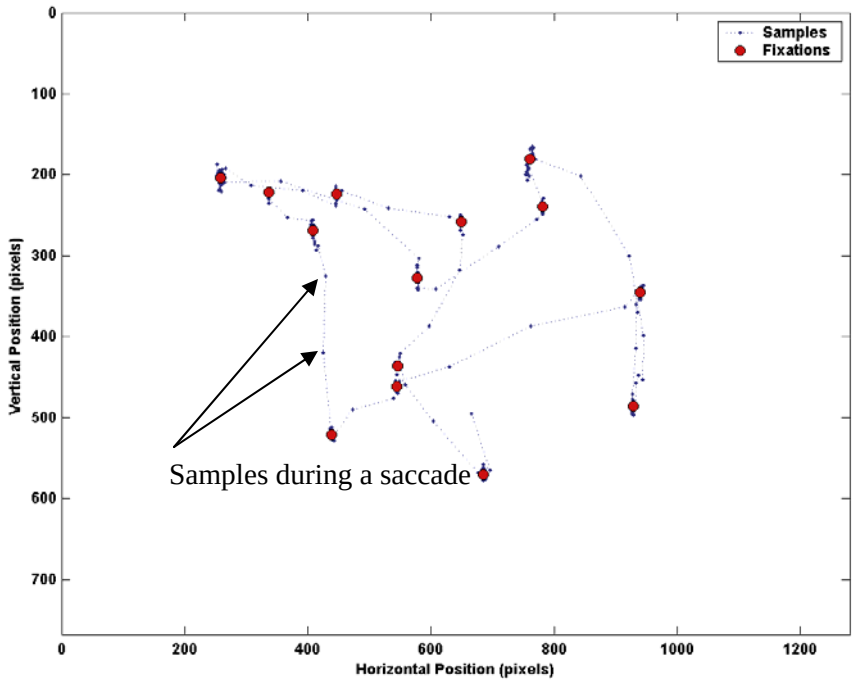


Figure 23: Example of raw data samples (small blue dots) and fixation locations as determined by the adaptive-velocity threshold method (large red dots).

3.8 Offset Correction

In the event that the eyetracking headgear shifted on the subject's head, the output point-of-gaze coordinates will be offset in the direction opposite to the movement of the headgear. This offset can be seen when the fixation locations are plotted over the image. Fixation location data from calibration checks, plotted over the nine calibration points, provide information about any global offset caused by a shift in the headgear. Also, for each subject and each experiment, the output fixation locations were superimposed over the image using Matlab. These graphs were then compared to the video record of the experiment to determine any offsets.

Offsets were manually corrected by adding the appropriate value to the horizontal and vertical fixation positions, as illustrated in Figure 24.

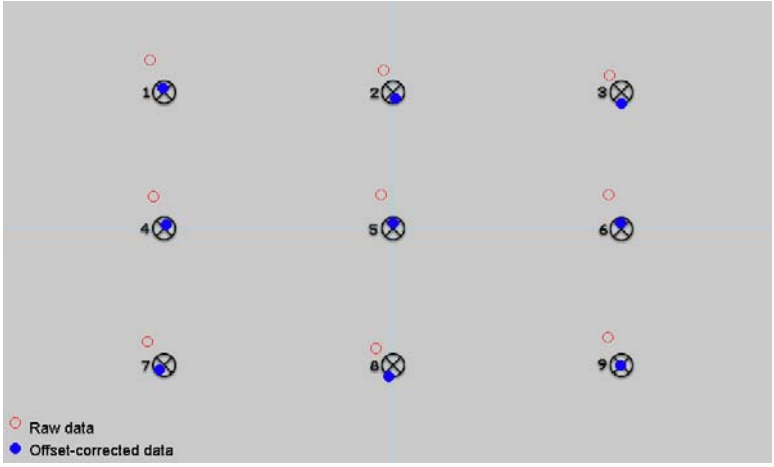


Figure 24: Example data from a calibration check during an experiment. Open, red circles represent the raw eye-head integrated data that was offset up and to the left due to movement of the headgear. Blue, closed circles represent the offset-corrected coordinates.

3.9 Stimulus Display

For the experiments presented in Chapters 4 and 5, images were displayed on a 50" Pioneer Plasma display. Subjects were seated approximately 40" from the display, which subtended 50 degrees of visual angle horizontally, and 35 degrees vertically. Images were presented in 24-bit color at a resolution of 1280 by 768 pixels. One degree of visual angle corresponds to 25 x 21 pixels for an average viewer distance.

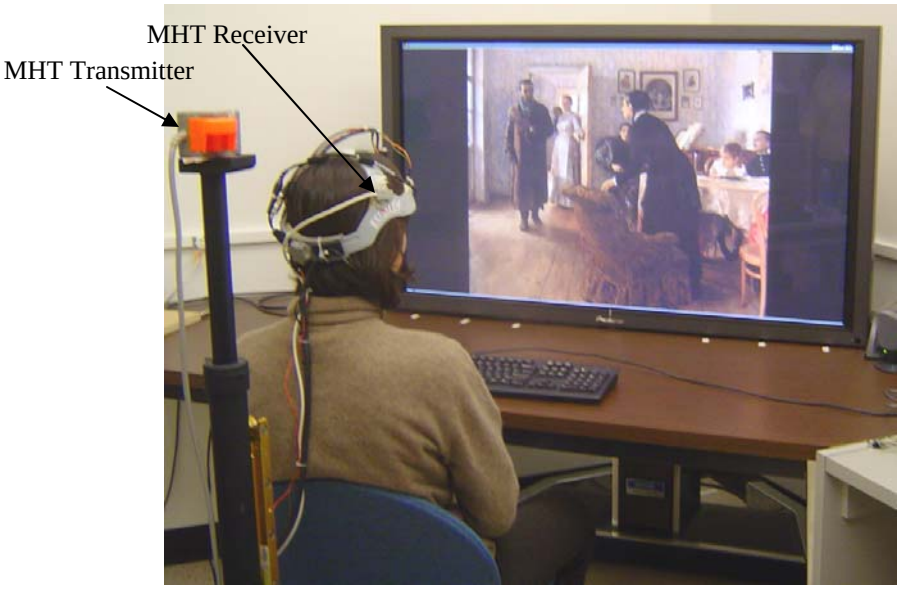


Figure 25: Experimental setup showing the Pioneer 50" Plasma display, the eyetracking headgear, magnetic head tracker transmitter and receiver.

Chapter 4

4 Yarbus Revisited

4.1 Overview

This chapter discusses the work of Alfred Yarbus, the importance of his work, and the replication of his classical experiment that shows the influence of task on an observer's viewing behavior. Spatial and temporal characteristics of fixations are analyzed to find whether the results of the replicated experiment support the claims of Yarbus.

4.2 Eye Movements and Vision

The doctoral work of Alfred Yarbus was published in 1965 in Moscow. In 1967, it was translated to English by Basil Haigh and Lorrin Riggs (editor) and published in the book, "Eye Movements and Vision" [Yarbus, 1967]. This collection of work spans a great number of topics. Yarbus researched and characterized the mechanics of the oculomotor system, including the velocities and durations of different types of eye movements. He also showed photographic records of corrective saccades, curved saccades, and eye movements of patients with disorders such as nystagmus and glaucoma. Most impressively, Yarbus constructed a series of miniature optical devices, or suction "caps," that were attached to the eye in order to project images on the retina and record eye movements.

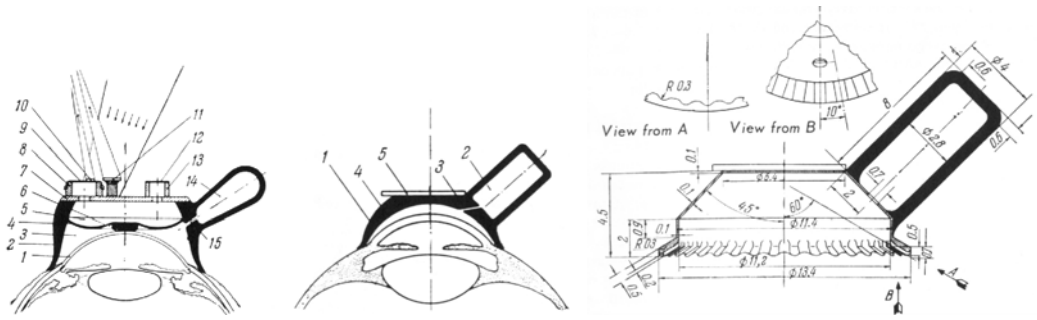


Figure 26: Diagrams of suction “caps” used. From [Yarbus, 1967].

An entire chapter is devoted the perceptual effects of stimuli presented stationary to the retina. These stimuli included blank fields, fields of high luminance and color contrast, and also flickering objects.

Eye movement patterns were recorded for observers viewing both stationary and moving objects, including simple stimuli, text, and optical illusions. The last chapter of Eye Movements and Vision shows records of observers viewing complex objects such as paintings and photographs. While he was not the first researcher to conduct this sort of analysis (see Buswell, 1935), he was one of the first to investigate the relationship between a person’s eye movement pattern and his or her attention or thoughts.

In Figures 107-112 of his book, Yarbus shows photographic records of people viewing I.E. Repin’s painting, “They Did Not Expect Him” (1884). This is a very politically significant painting portraying a Russian revolutionary returning from exile. All of Yarbus’ observers were highly educated and very familiar with the painting. In Figure 109, shown below, Yarbus shows a set of seven eye movement patterns of one subject as he viewed the painting; before each three-minute viewing the subject was given a different instruction. In the first viewing, the subject was not given a specific instruction, but only asked to look at the painting. Before each of the subsequent six viewings, the instructions were: “Estimate the material circumstances of the family in the picture,” “Give the ages of the people,” “Surmise what the family had been doing before the arrival of the ‘unexpected visitor,’” “Remember the clothes worn by the people,” “Remember the position of people and objects in the room,” and “Estimate how long the ‘unexpected visitor’ had been away from the family.”

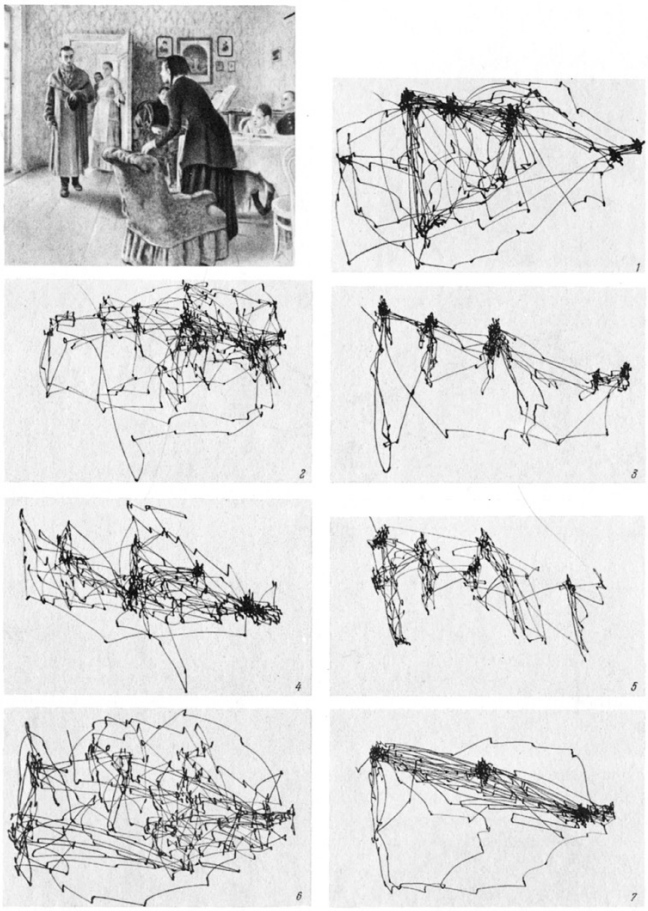


Fig. 109. Seven records of eye movements by the same subject. Each record lasted 3 minutes. The subject examined the reproduction with both eyes. 1) Free examination of the picture. Before the subsequent recording sessions, the subject was asked to: 2) estimate the material circumstances of the family in the picture; 3) give the ages of the people; 4) surmise what the family had been doing before the arrival of the "unexpected visitor"; 5) remember the clothes worn by the people; 6) remember the position of the people and objects in the room; 7) estimate how long the "unexpected visitor" had been away from the family.

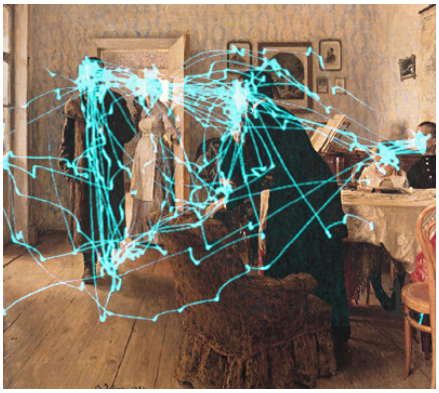
Figure 27: Figure 109 from [Yarbus, 1967].

There is a striking difference in the eye movement patterns between the different conditions. These differences are more pronounced than between records of seven different subjects freely viewing the painting without instruction. From this experiment, Yarbus concluded that the eyes "fixate on those elements of an object which carry or may carry essential or useful information [Yarbus, 1967, page 211]." The eyes are not reactively drawn to salient, low-level properties of the image such as bright regions or edges. Instead, the elements fixated are those that provide most information for the task at hand. As the task changes, so does the 'informativeness' of certain regions, thereby changing the observers viewing behavior. Furthermore, the patterns and locations of eye movements give insight into what the observer was thinking; Yarbus was one of the first researchers to recognize eye movements as an external manifestation of cognitive processes.

Other than being qualitatively different, Yarbus also noted a “cyclic” pattern of eye movements. Acknowledging that the three-minute view time was more than long enough to fixate on the important regions of the picture, he noted the fact that once these regions were fixated, the subject did not move on to examine the secondary elements and details in the picture. Instead, the observer looked at these primary regions again and again. Yarbus concluded from other freeview experiments that this “cycle” can last from a few seconds to many tens of seconds.

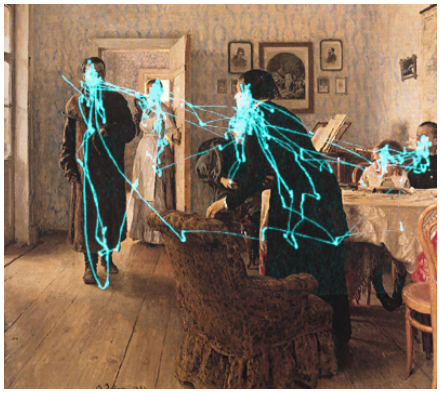
Below are visualizations made by overlaying the original photographic records on an image of Repin’s painting.

Original Image

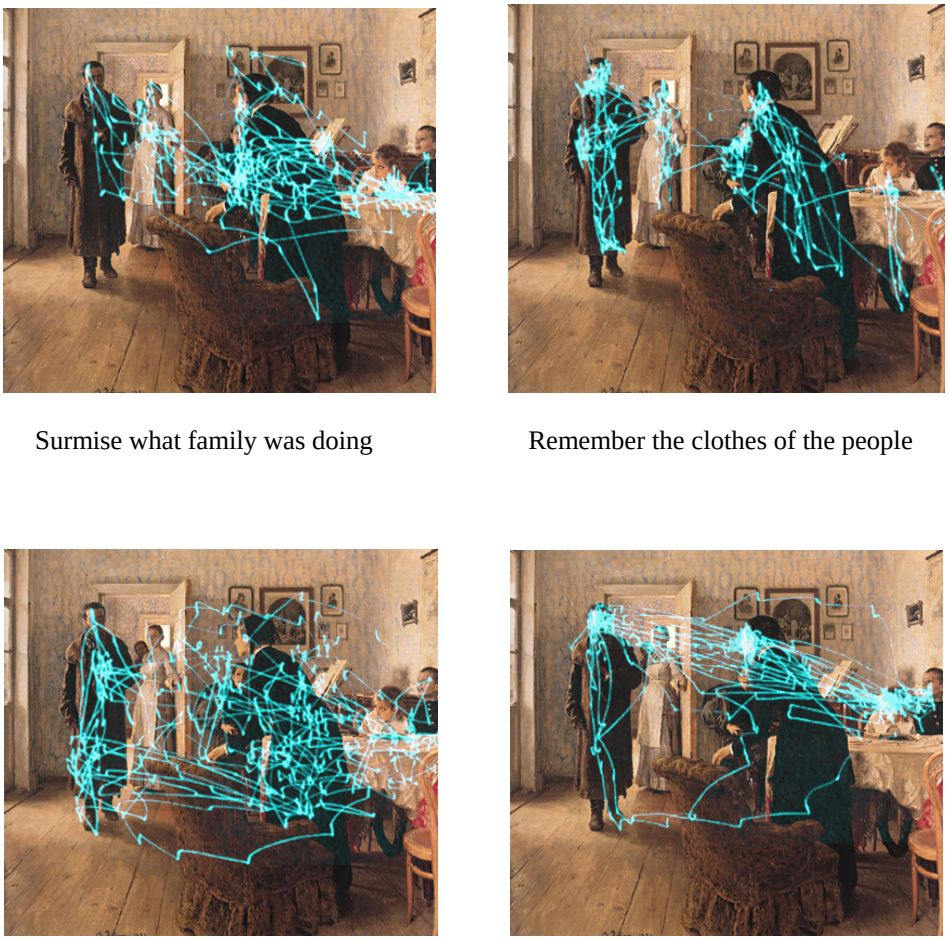


Freeview

Estimate material circumstances



Give the ages of the people



Surmise what family was doing

Remember the clothes of the people

Remember the position of people

Estimate how long away

Figure 28: Visualizations of original photographic records overlaid on top of the stimulus

4.3 Implications

The results from this experiment are very significant in that they demonstrate a "top-down" component of visual selection, showing that the human visual system is an active system. An observer's cognitive goal and past experiences interact with the visual stimulus in order to execute an appropriate behavior. The system is not passive; it does not randomly or uniformly sample the visual environment.

4.4 Yarbus' Methods

Yarbus' Figure 109 is very well known and often cited in literature concerning eye movements and behavior. The fact that is often not recognized or overlooked is that the records published are for only one observer. The text "Eye Movements and Vision" does not make any

reference to other observer's performing these tasks or producing similar results. The original Russian journal article [Yarbus, 1961] contains the same set of records as shown in Figure 109 of the translated text.

Also, each of the photographic records was made as the subject viewed the picture for three minutes at a time. During the recording, the subject's eyes were anaesthetized and his eyelids were taped open with heated strips of adhesive plaster. A small suction device ("cap") holding a small mirror was then firmly attached to the sclera. Light was projected onto the mirror, and reflected onto a piece of photographic film. The subject's head was constrained using chin and forehead rests.

Fig. 23. Position of lids held by strips of adhesive plaster in work with the P_1 cap.

Figure 29: Figure of configuration of eyelids for recording eye movements. From [Yarbus, 1967, page 44].

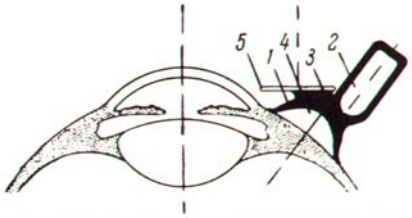


Fig. 13. The P_1 suction device or "cap."

Figure 30: Suction device containing small mirror used to record eye movements. From [Yarbus, 1967, page 30].

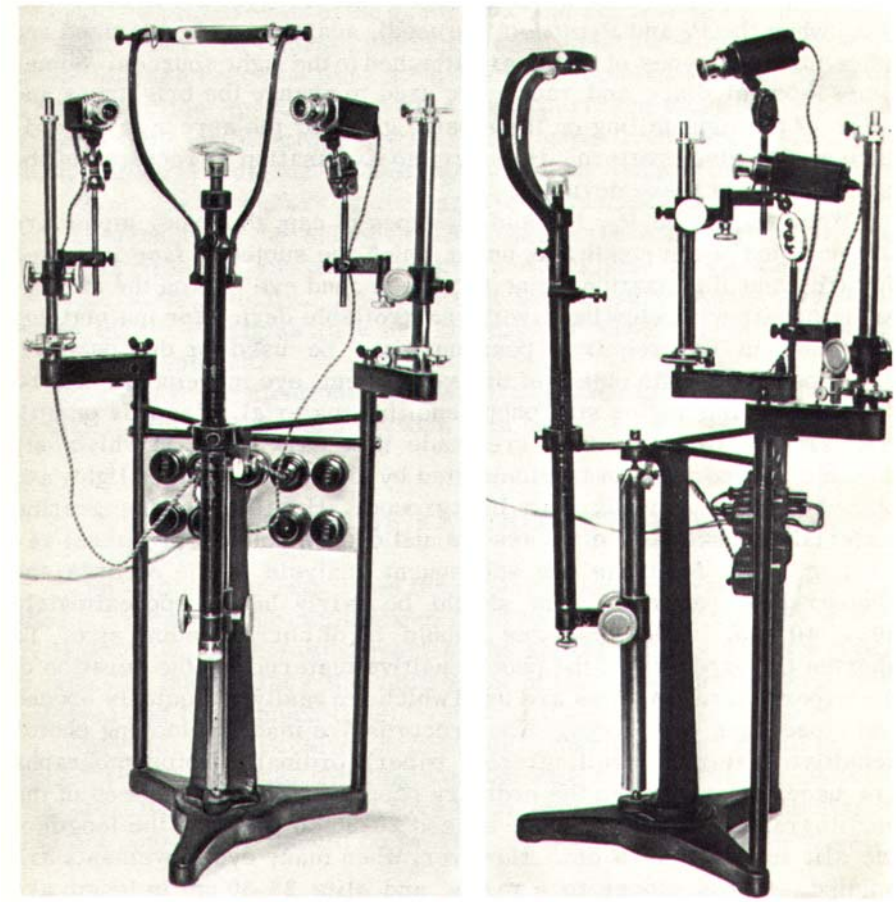


Fig. 21. The apparatus used in recording eye movements.

Figure 31: Apparatus used in recording eye movements. The setup contains chin and forehead rests, light sources, and a control panel. From [Yarbus, 1967, page 41].

4.5 Replication of Experiment

Given the significance of Figure 109, and the fact that the records were made for only one subject using a constrained, and likely painful, experimental setup, it is important to know if many subjects show the same task-influenced eye movement patterns when the experiment is replicated under more natural conditions.

For this experiment, each subject freely viewed a set of 57 digital images of paintings, photographs, and drawings while being eyetracked. The subject's head was unconstrained, and images were presented on a 50" Plasma display (see Chapter 3 for more information on the experimental setup). On this display, the image of Repin's painting subtended approximately 35 by 35 degrees of visual angle, when seated 40 inches from the display, as illustrated in Figure 32.

The viewing was self-paced; the subject pressed the spacebar to move on to the next image. Between every 10 images, the subject was asked to look at a sequence of 9 points to check the accuracy of the track. Subjects were recalibrated if needed (see Chapter 3). Following the calibration check, a screen with written instructions was presented. These instructions are the same as those reported in Yarbus' book, and were also presented in the same order. (For the no-instruction task, Repin's painting was shown randomly within the first 10 images without any instruction; it was simply another painting within the set of random images.) The only exception is that "Estimate the material circumstances…" was reworded as "Estimate the financial circumstances…" because during the pilot experiments some subjects did not readily understand the meaning of the original wording. When the subject had read the instruction, he or she pressed the spacebar to view Repin's "They did not expect him." The subject then performed the task, which sometimes involved answering questions out loud. When the subject completed the task, the spacebar was pressed to freely view the next random image. Including the calibration checks, instruction screens, and repetitions of Repin's painting, there were 78 images displayed during the experiment. On average, the experiment lasted 15 minutes. Twenty-five subjects performed the experiment; seventeen were successfully eyetracked.

Figure 32: Experimental setup for replication of Yarbus' experiment.

A subset of subjects (12 of the 17) also freely viewed Repin's painting for a forced three minutes following the first experiment. At this point, the subject had seen the painting 8 times.

One subject's eye movements were recorded as he viewed Repin's painting in three-minute intervals for each of the seven tasks. For this experiment, no other images were shown between the tasks. This experiment lasted 21 minutes, plus the time the subject spent reading the instruction screens.

4.6 Results

Given that the results published by Yarbus were very qualitative, and lacked any information about the temporal sequence of viewing, it is difficult to make a direct, quantitative comparison between the results of the present experiment and Yarbus'. In the following sections, several qualitative and quantitative methods are used to compare the eye movement patterns of subjects across the different tasks and viewing conditions.

4.6.1 Self-terminated

Figure 33 shows eye movement records of two subjects as they performed each of the seven tasks during this experiment. Subject A, shown on the left, represents a typical subject. For this subject, the view times ranged from 6 to 92 seconds. Although the view times are significantly less than 3 minutes, the eye movement patterns are remarkably similar to those published by Yarbus. Fixations in the "Freeview," "Ages," and "How long away" tasks fell primarily on faces and figures. The "Financial" and "Position" tasks elicited more spatially distributed patterns of fixations. Subject B represents an atypical subject, whose eye movement patterns are not drastically different between tasks. Also, the view times were short, ranging from 5 to 19 seconds. It is important to note that the subject did complete all tasks and answered the questions out loud; it is not the case that these records are a result of subjects not following directions.

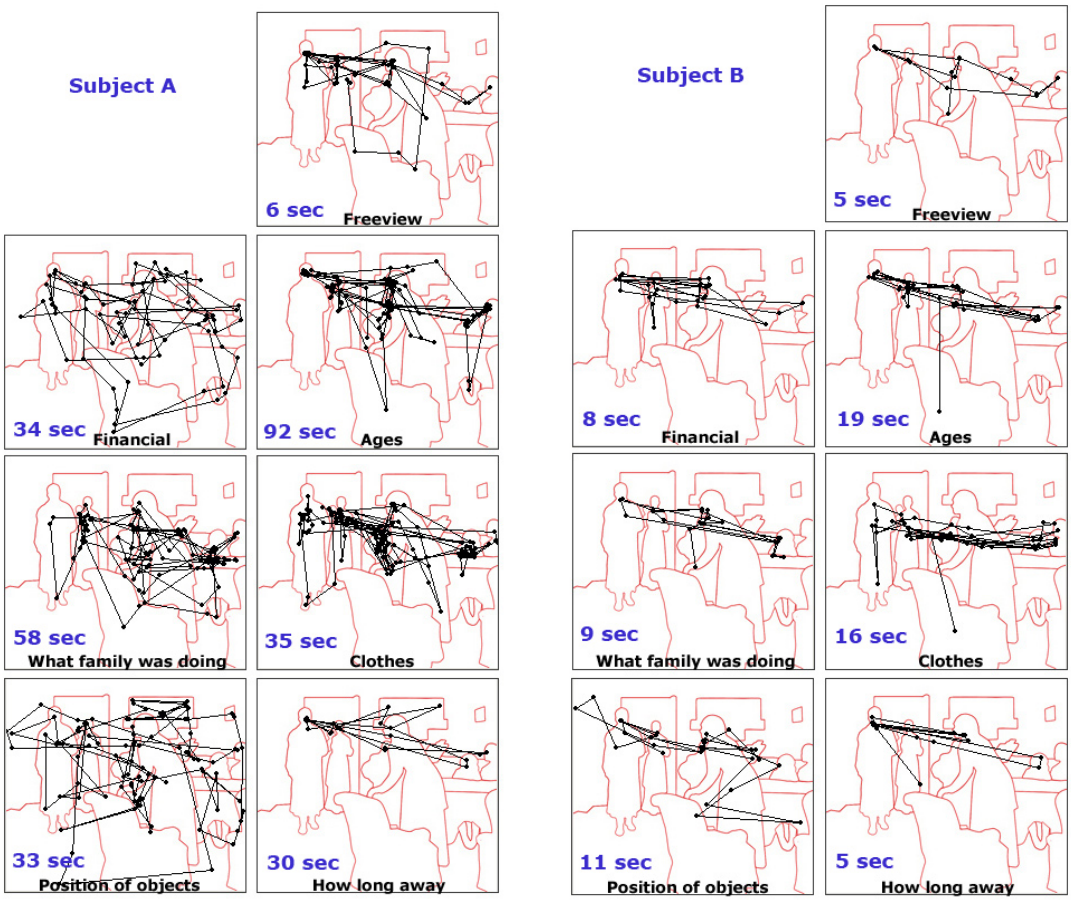


Figure 33: Eye movement records from two subjects as they performed each of the seven tasks. Subject A represents a typical subject whose view patterns resemble those published by Yarbus. Subject B represents an atypical subject, whose view patterns do not differ significantly between task.

Figure 34 shows the amount of time each subject viewed Repin's painting for each task. For the "Freeview" task, subjects' viewed the painting for, on average, only 9 seconds. Nineteen seconds was the average time subjects spent answering the question, "Estimate the financial circumstances." When asked to give the ages of the people, the painting was viewed for an average of 50 seconds, which is significantly longer than any other task. For the tasks "Surmise what the family had been doing", "Remember the clothes," and "Remember the position of people and objects," the image was viewed for an average of 25, 24, and 29 seconds, respectively. The last task, "Estimate how long the visitor has been away," was completed in an average of 15 seconds.

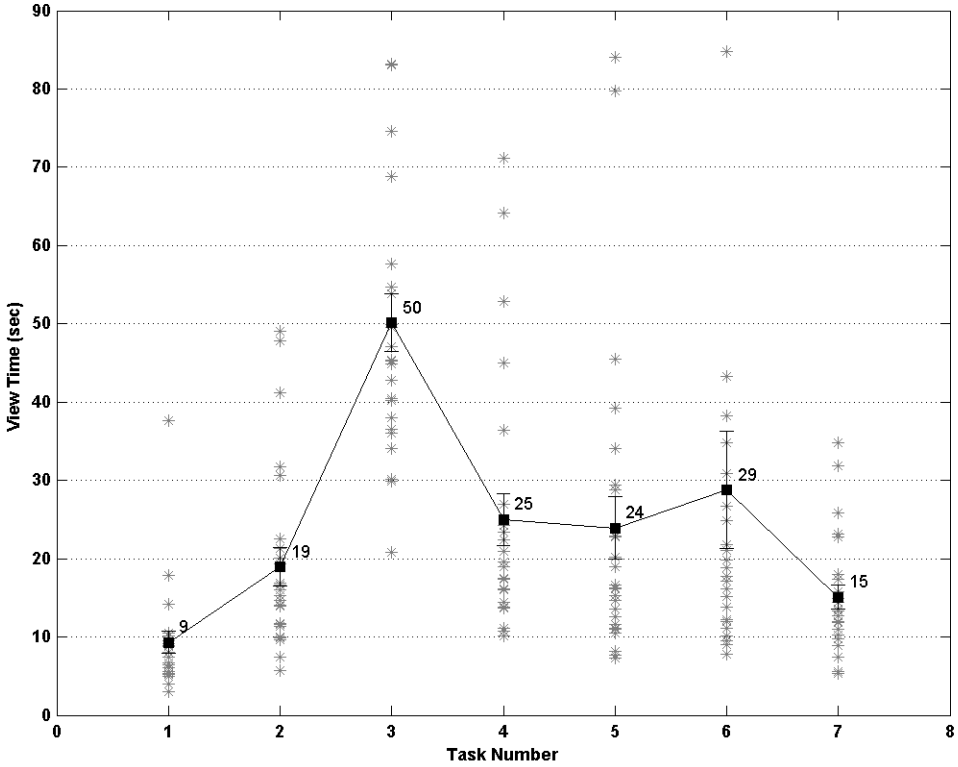


Figure 34: View times for each of the seven tasks in the order in which they were performed. Black rectangles mark the average across 25 subjects. Error bars represent one standard error of the mean. Tasks: 1: Freeview; 2: Financial circumstances; 3: Give the ages; 4: Surmise what family was doing; 5: Remember the clothes; 6: Remember the position of people and objects; 7: Estimate how long the visitor was away

Figure 35 shows histograms of fixation durations for 17 subjects for each task. Listed on each graph is the distribution's mean, median, standard deviation, and number of fixations. Each distribution resembled a Raleigh distribution in that they are all positively skewed. The task "Give the ages of the people" elicited the longest fixations, which were 376 milliseconds on average. About 14% (133) fixations were longer than 1 second in duration. This distribution also had the largest standard error of 495 milliseconds. When judging the "Financial Circumstances" of the family, fixations were typically 250 milliseconds, which was the shortest average of all the tasks.

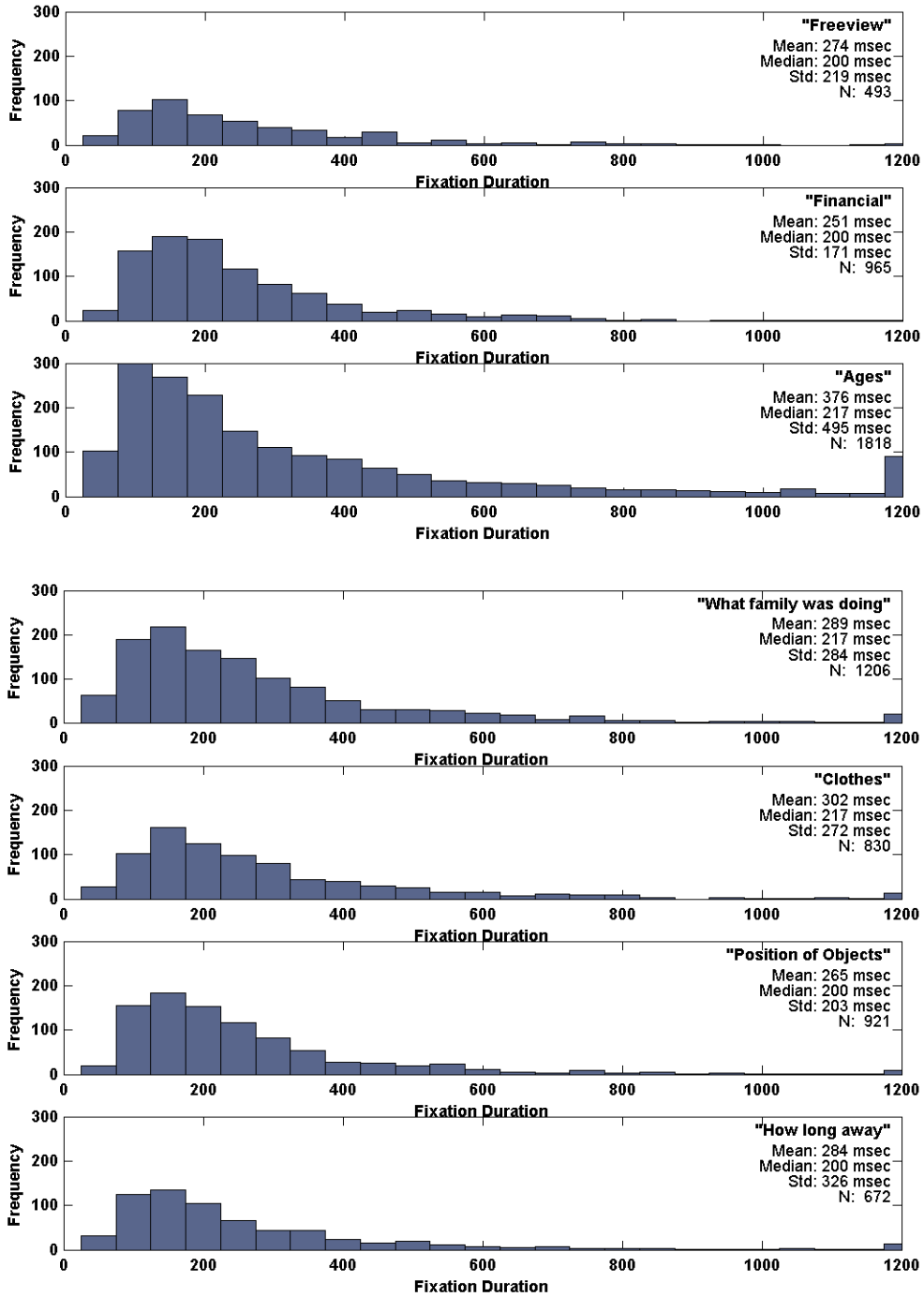


Figure 35: Histograms of fixation durations for each task across 17 subjects.

The image of Repin's painting was segmented into 22 different regions (e.g., the man's face, the man's figure, chairs, floor, etc.). The regions and associated labels are shown in Figure 36. The total gaze duration in each of these regions was found for every subject. These durations were then normalized by the viewing time for that particular subject and task to produce the percentage of time spent viewing each region. The results for each task are shown in Figure 37, and are averaged across all subjects.

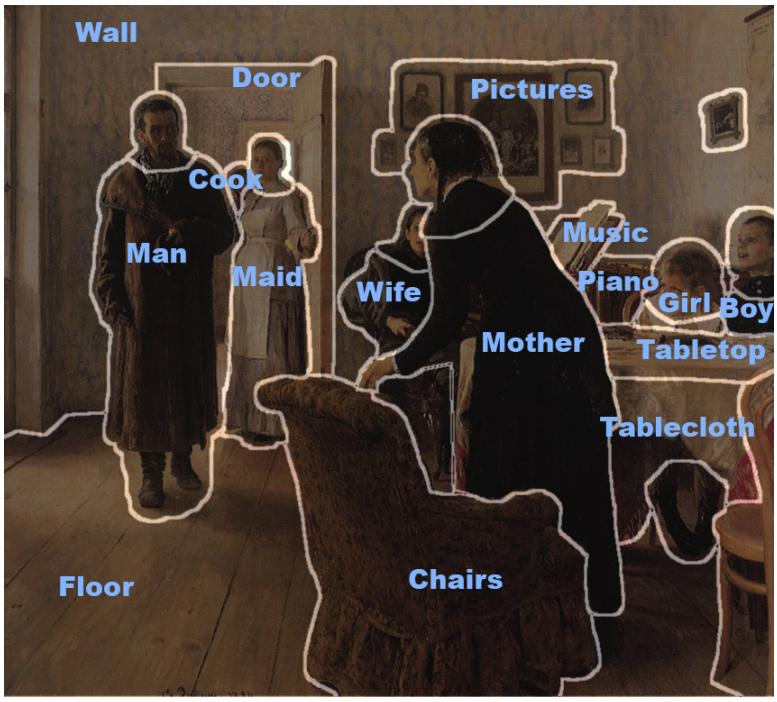


Figure 36: Image segmented into 22 regions with associated labels. The faces and figures of each person are two separate regions, although not labeled in this illustration.

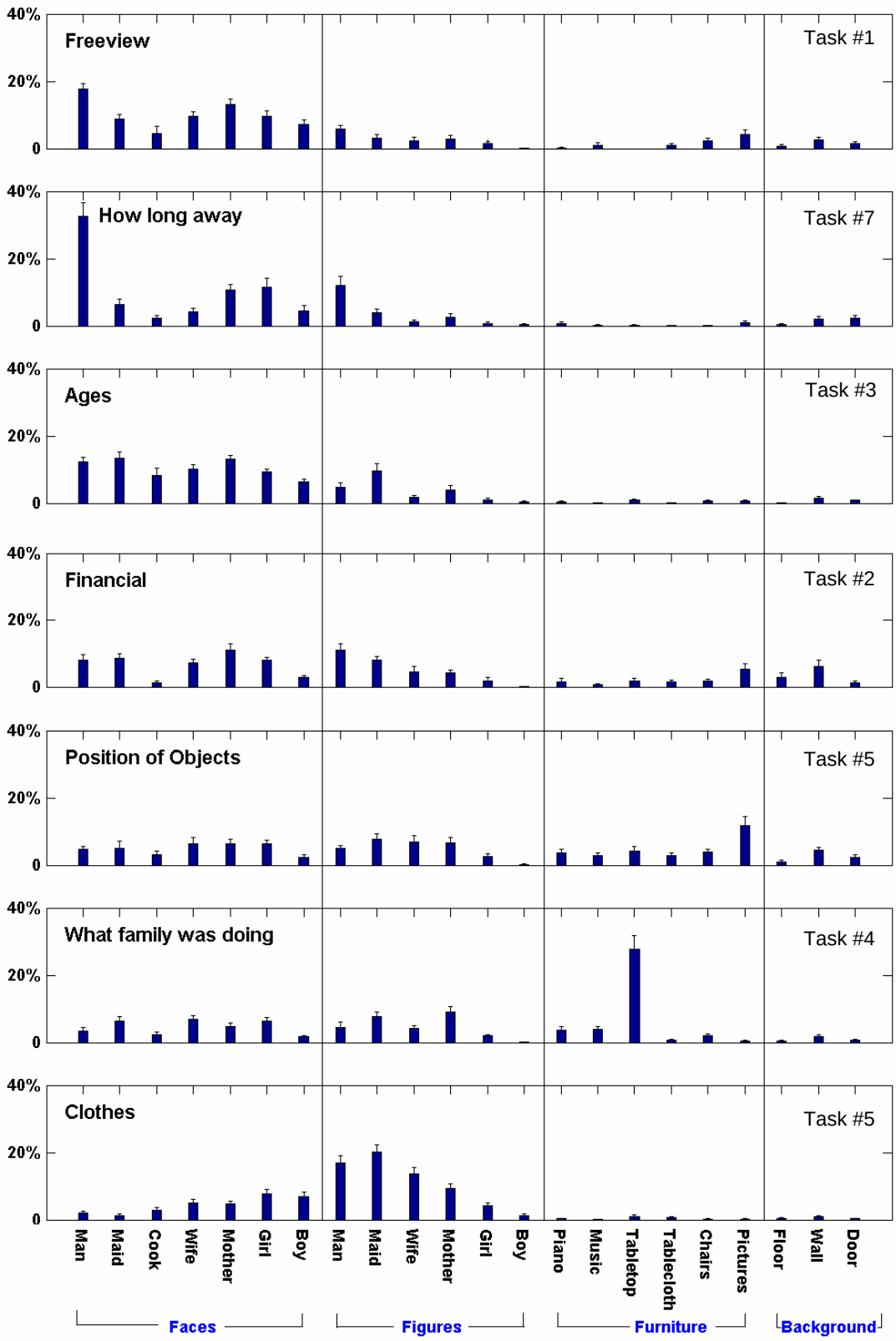


Figure 37: Percentage of time spent viewing each region. Error bars represent standard error of the mean across 17 subjects.

During the freeview task, 20% of the time was spent fixating on the man's face, followed by the faces of the mother, wife, maid, and children. This result agrees with Yarbus' observer's behavior: the most fixations fell on the faces of the people followed by the figures.

The distribution of gaze durations for the task "Estimate how long the 'unexpected visitor' was away" is very similar to that of the freeview task. Again the faces received the most fixations. In this task, more of the viewers' time, around 35%, was spent looking at the man's face and figure. For this task, Yarbus noted particularly intensive movements between the faces of the children and man, in hopes to gain information from the expressions of the children.

The task "Give the ages of the people" resulted in a more uniform distribution across the faces. Yarbus reported that for this task, all of the observer's attention was concentrated on the faces, with few saccades between faces.

For the task, "Estimate the financial circumstances of the family," the faces were again the most attended-to regions. However, the clothing and furniture received more fixations than in previous tasks. Yarbus' observer paid particular attention to the women's clothing, armchair, and tabletop.

For the task, "Remember the position of people and objects in the room," every region received fixations in most cases, producing a more uniform distribution of gaze durations throughout the image. The pictures on the wall were examined for a larger fraction of time (12%) than in other tasks. Yarbus reported that for this task, his observer examined the whole room and all of the objects.

When asked to "surmise what the family had been doing before the arrival of the unexpected visitor,'" the tabletop proved to be the most informative, in that viewer's typically spent 30% of their time looking at it. The piano and sheet music also received a larger percentage of viewing time than in other tasks. These are also the regions that Yarbus' subject attended to the most: "the observer directed his attention particularly to the objects arranged on the table, the girl's and the woman's hands, and to the music (Yarbus, 1967, page 192)."

Lastly, when asked to "remember the clothes worn by the people," viewer's typically spent the majority of time examining the figures in the image, as did Yarbus' subject. It is worth noting that in every task, the faces of the people in the image were fixated.

The fraction of time spent in each of the regions in the image can be thought of as a 22-element feature vector. As a measure of between-subject variability within each task, the Euclidean distance between each possible pair of vectors was found. The average distances for

each task are show in Figure 38. Also shown is the within-subject/between-task distances, averaged across all 17 subjects; this value was 0.39. For the first six tasks, the between-subject distance is lower than the between-task distance, indicating that the eye movement patterns across subjects for that task were more similar than the patterns of one person performing different tasks. "Give the ages" task (#3) was the most similar between subjects, with an average distance of 0.26. The average distance for the "How long away" task (#7) was not significantly different than the within-subject distance, suggesting that observers used different viewing strategies to perform that task.

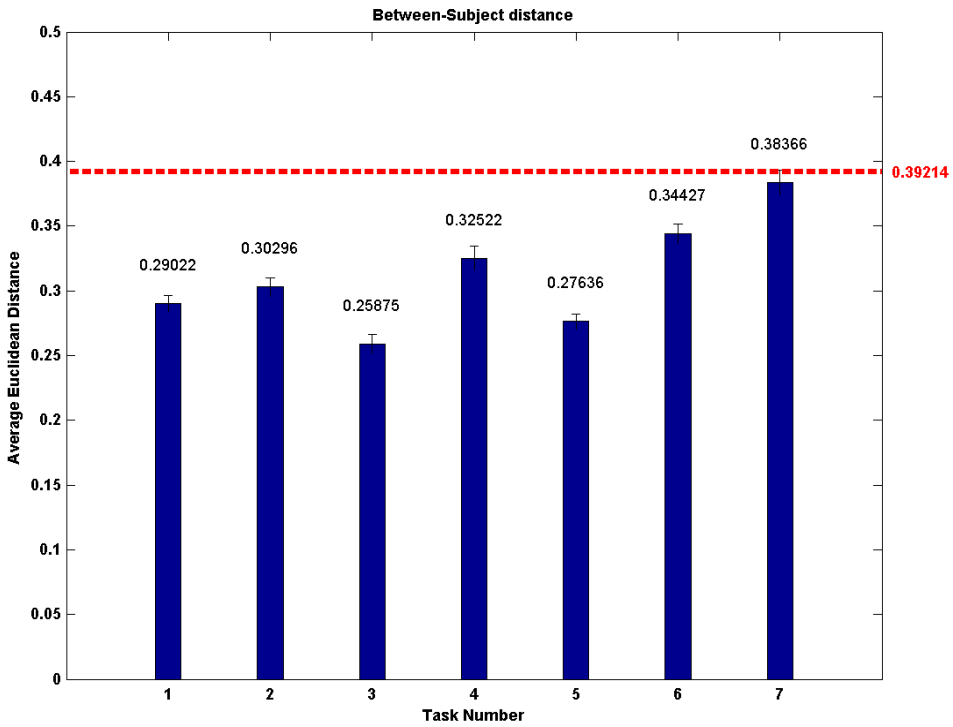


Figure 38: Between-subject distance for each task defined by the average distance between region histogram vectors. Error bars represent one standard error of the mean. The thick dotted red line represents the within-subject (between-task) distance, averaged across 17 subjects. The lines above and below the red line represent one standard error of the mean. Tasks: 1: Freeview; 2: Financial circumstances; 3: Give the ages; 4: Surmise what family was doing; 5: Remember the clothes; 6: Remember the position of people and objects; 7: Estimate how long the visitor was away

The correlation coefficient between each pair of vectors was also computed. The average correlation coefficient for each task is shown in Figure 39. Also shown is the within-subject/between-task correlation, averaged across all 17 subjects (shown as the red dotted line). The high correlation coefficients for the "Give the Ages" task (#3) and "Remember the Clothing" (#5) indicate that subjects' behavior within these tasks were most similar. The "Remember the Position of People and Objects" task (#6) resulted in an average correlation coefficient of 0.16, which was below the average between-subject correlation coefficient of 0.23.

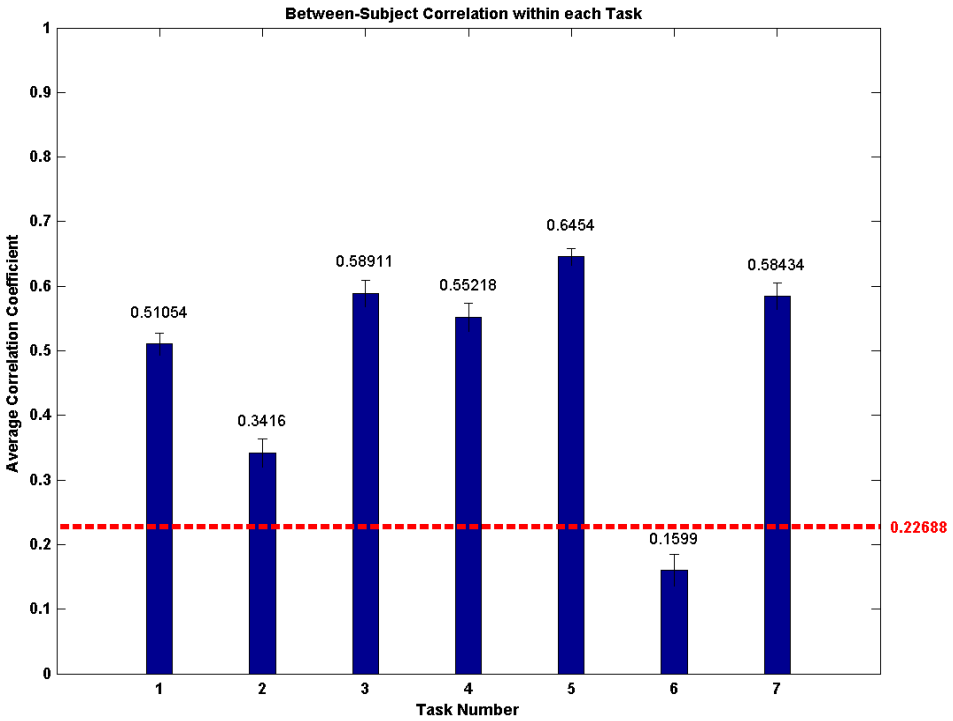


Figure 39: Between-subject correlation for each task defined by the average correlation coefficient between region histogram vectors. Error bars represent one standard error of the mean. The thick dotted red line represents the within-subject (between-task) correlation coefficient, averaged across 17 subjects. The lines above and below the red line represent one standard error of the mean. Tasks: 1: Freeview; 2: Financial circumstances; 3: Give the ages; 4: Surmise what family was doing; 5: Remember the clothes; 6: Remember the position of people and objects; 7: Estimate how long the visitor was away

For each of the 17 subjects, there are seven feature vectors corresponding to each of the seven tasks, yielding 119 vectors. A 22x119 matrix was constructed, and Principal Components Analysis (PCA) was performed on the data. PCA can be used to display the data in the most informative way. The first principal component (PC) represents the dimension in which most of the variability of the data lies. The second is orthogonal to the first principal component, and represents the dimension in which the next largest amount of variability is contained, and so on. For this data, 28% of the variability is explained by the first PC; an additional 16% is explained by the second, and 11% by the third. Figure 40 shows each of the 119 vectors projected onto the first and second PCs, and Figure 41 shows the projections on the first and third PCs. Each point represents one subject viewing the image for one task.

The first component seems to separate the data according to the fraction of time spent looking at faces. In the tasks "How long away" and "Freeview," subjects spent almost all their time examining faces, whereas for the tasks "What the family was doing," and "Remember the Clothes," subjects' attention shifted to other objects in the scene. The second component may separate by which objects were attended to the most. "What the family was doing" is a fairly distinct cluster, possibly caused by the amount of time spent on one single object, the tabletop. The third component differentiates between the tasks "Ages" and "How long away." The averages of these vectors, shown above in Figure 37, differ primarily in the amount of time spent on the man's face. Similarly, "What the family was doing" and "Clothes" differ from the other tasks in that one or two regions receive the majority of fixations.

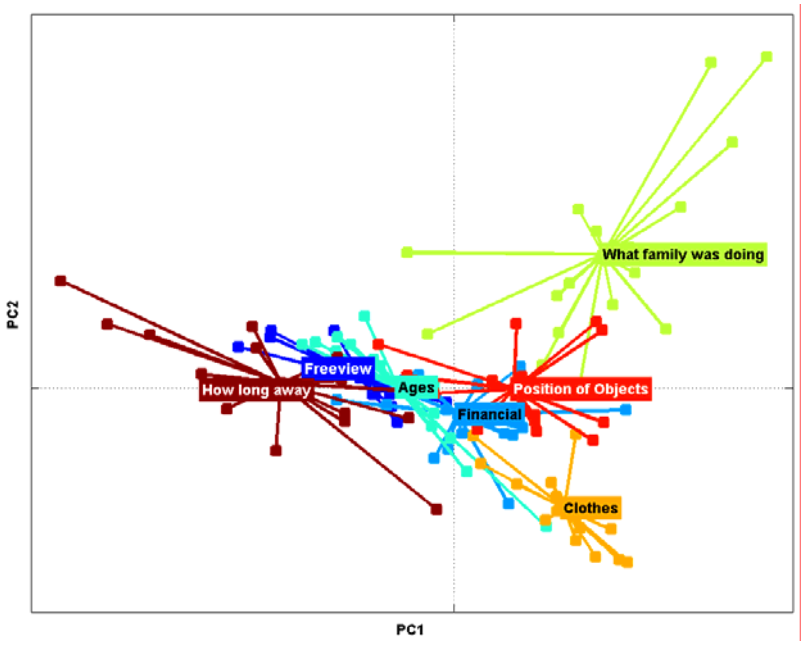


Figure 40: Data projected onto first and second principal components. Each point represents one subject performing one task. Twenty-eight percent of the variability is explained by the first PC; and additional 16% is explained by the second.

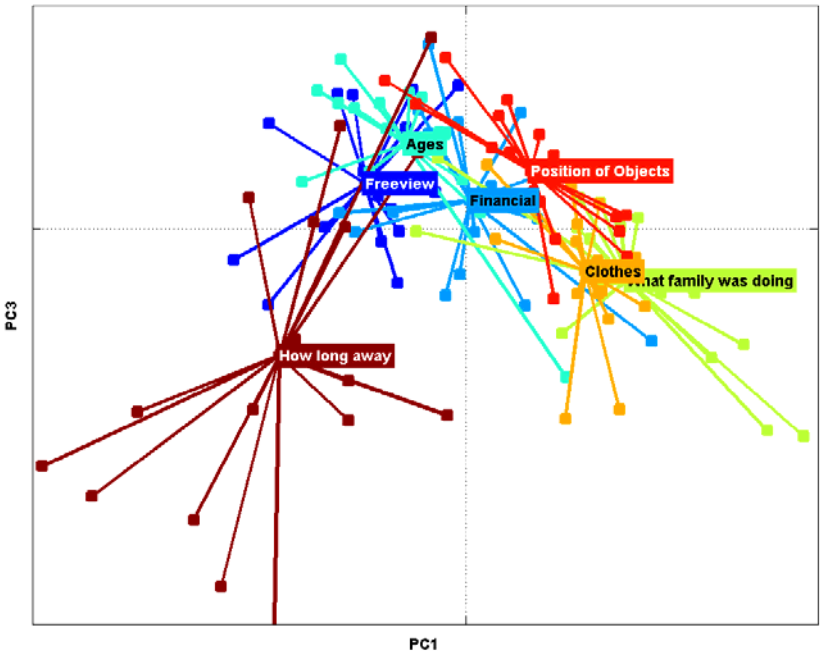


Figure 41: Data projected onto first and third principal components. Each point represents one subject performing one task. An additional 11% of the variability is explained by the third component.

4.6.2 Enforced three-minute view, all tasks

Figure 42 shows the eye movement records for the subject who viewed the painting for 3 minutes for each task. Figure 43 shows the fraction of time the subject spent viewing each of the 22 regions. When compared to the average behavior during the self-terminated condition, shown in Figure 37, there are many similarities, which may support Yarbus' idea of eye movement patterns being composed of cycles. For example, for the "How long away" task, the subject spent most of his time looking at the man's face, and the expressions on the faces of the mother, girl, and boy, which were also the most fixated regions in Figure 37. The subject reported that he was trying to guess whether both of the children recognized the man. For the "Financial" and "Position of Object" tasks, the histograms are more uniform, as they were for the self-terminated condition. For the "Ages" task, this subject's behavior was also similar in that almost all fixations fell on the faces, although he did spend some time examining the pictures on the wall. The histogram for "Clothes" is also similar, but the subject did spend almost an equal amount of time looking at the people's faces as he did at their clothing.

When guessing "What the family was doing," the subject did examine the table top, but unlike in the self-terminated task where subjects spent almost 30% of their time, he did not spend more than 10% of the time viewing it. Instead, more time was spent examining the faces of the people in the room; the subject reported that by the end of the task, he was developing a story about the characters in the scene, their personalities, and what they were doing.

For the "Freeview" task, although the subject did spend time looking at the faces in the scene, many of the other objects, or "secondary details" were attended to. The subject was particularly interested in the furniture and pictures on the wall. It is possible that during the first portion of viewing, the faces were fixated, and by the end of the 3 minutes, the subject had moved on to the secondary elements; this behavior would not support Yarbus' assertion that a viewer would revisit the same regions over and over.

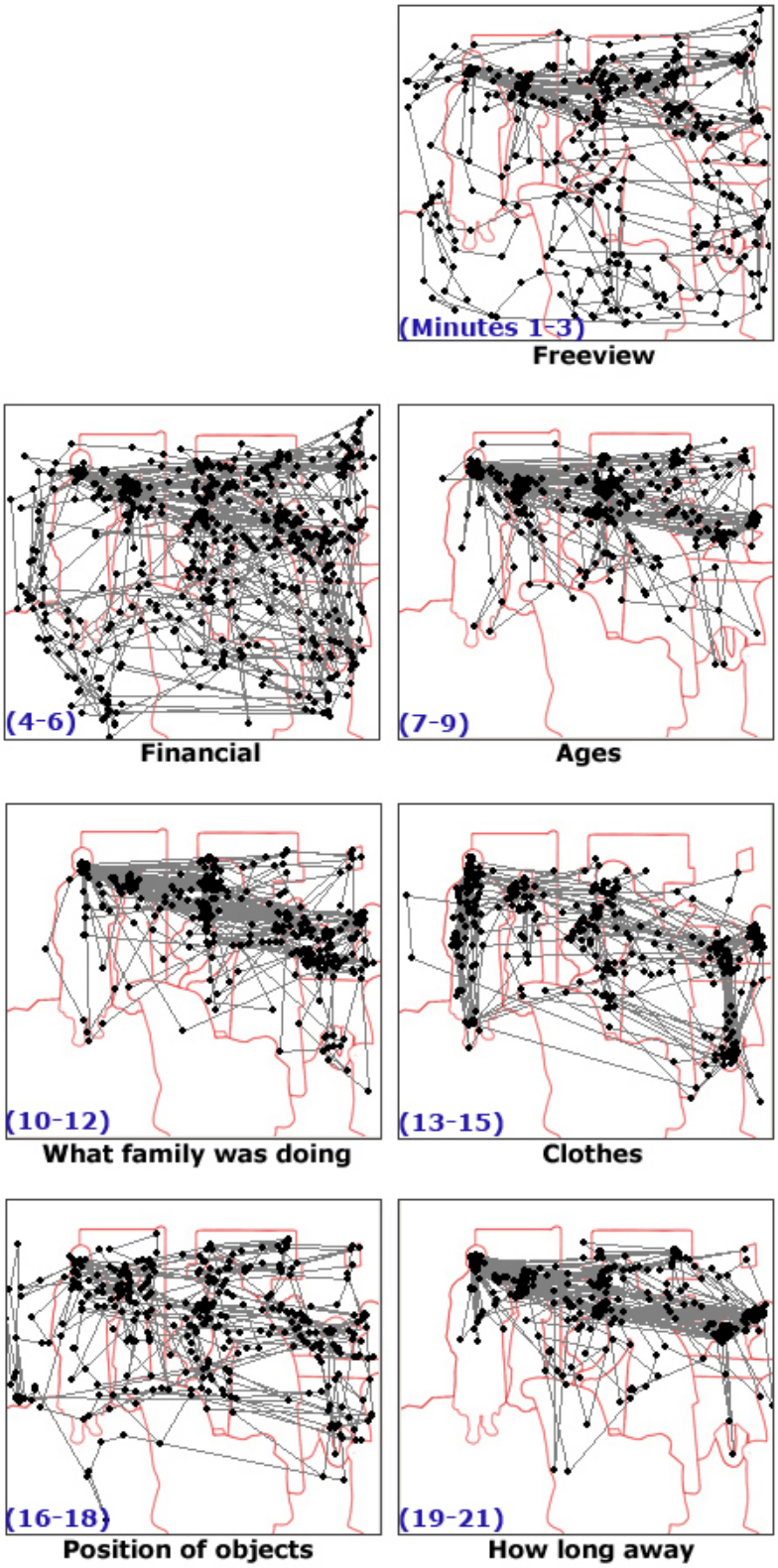


Figure 42: Eye movement records of one subject. For each task, the subject viewed the painting for 3 minutes. The minutes in which each task was performed is labeled in the corner of each record.

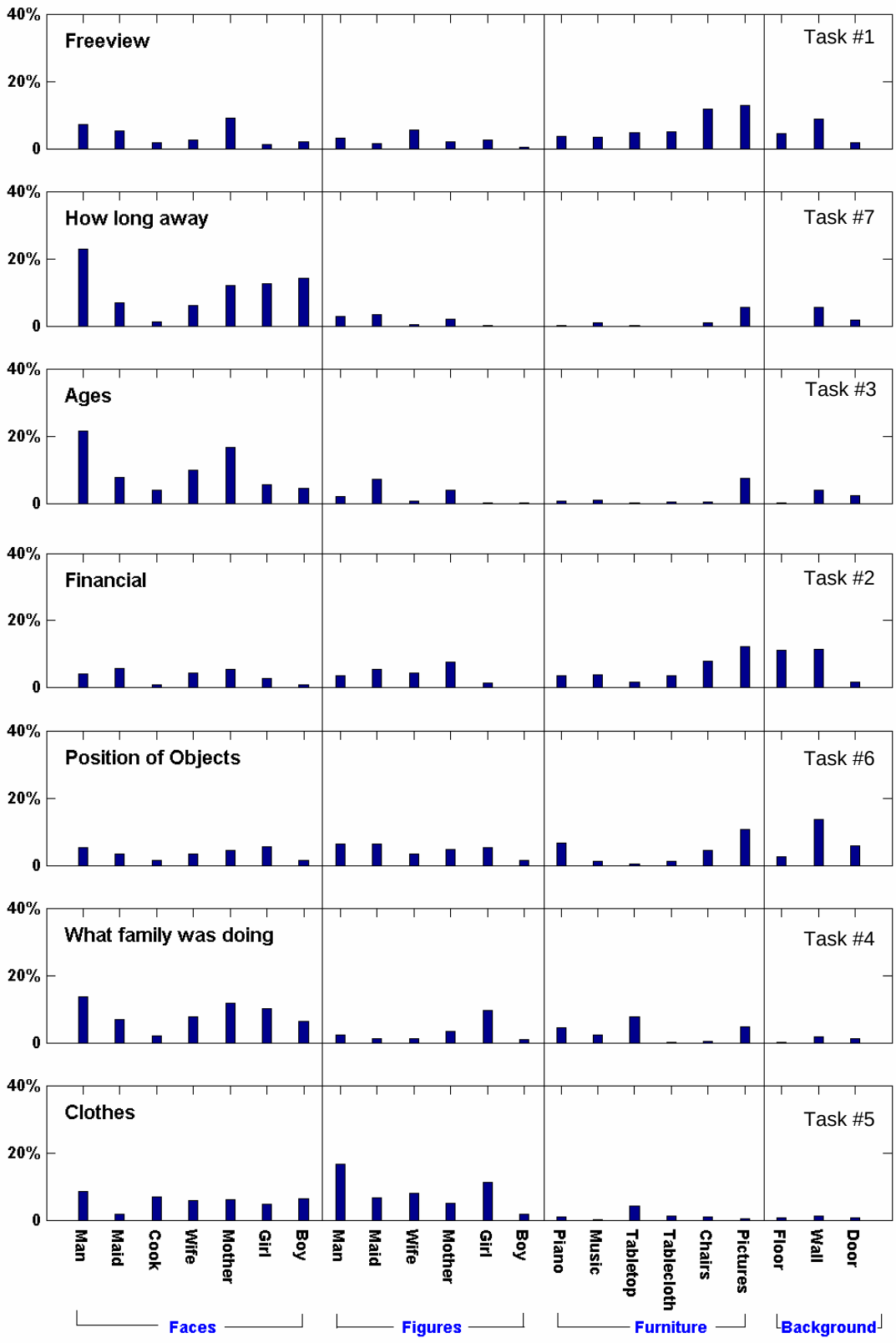


Figure 43: Percentage of time one subject spent viewing each region as he viewed the image for 3 minutes per task.

To determine if the subject's viewing behavior was indeed cyclic, or whether it changed over time, the temporal order of regions fixated (i.e., shifts of gaze) for the "Freeview" task is shown in Figure 44. These regions are grouped into three areas of interest: Faces, Figures, and Background/Furniture. The reason for this grouping is that these are the elements Yarbus referred to as primary (faces and figures) and secondary (other foreground items and background). During the first 20-30 seconds of viewing, the subject made many short fixations on each of the faces, most of the figures, and almost all the background elements. For the next 30 seconds, the subject seems to change behavior. Gaze durations become longer, and are spent on the background elements for 10 seconds, then to the man's figure, and then to the man's face. During the next 30 seconds, the background elements are again examined. After that, there are again short fixations distributed among the faces, figures, and background, which is similar to the behavior at the beginning of the task. For the rest of the viewing time, the background elements and figures are examined with long gaze durations.

The six figures following Figure 44 show this visualization for the rest of the tasks. For the "Financial" task, the behavior seems to be consistent: short fixations distributed between the areas over time. For the "Ages" task, the faces are fixated during the first 40 seconds. During the rest of the trial, the figures and background elements are examined. During the first half of the "What the family was doing" task, many short gaze durations were distributed throughout all the regions, followed by specific examination of the tabletop and girl. The subject spent almost all the remaining time shifting his gaze between all of the faces in the scene. In the "Clothes" task, again we see a change in viewing behavior halfway through. During the first half, the faces and clothes are each examined in turn. During the second half, the background elements are examined, and gaze shifts frequently between different regions. The pattern of gaze shifts in the "Position of Objects" task resembles that of "Freeview" in that the first 30-40 seconds were spent fixating on all of the objects in the scene. Following that, the pictures on the wall were examined for almost 10 seconds, then a figure, and then a face. The second half was again spent shifting gaze between all types of regions except for one long gaze on the wall/floorboards. In the "How long away" task, there were many transitions between each of the faces in the scene throughout the viewing, specifically between the boy's and the girl's faces. Toward the end, the subject looked between the man's, maid's, and mother's faces several times.

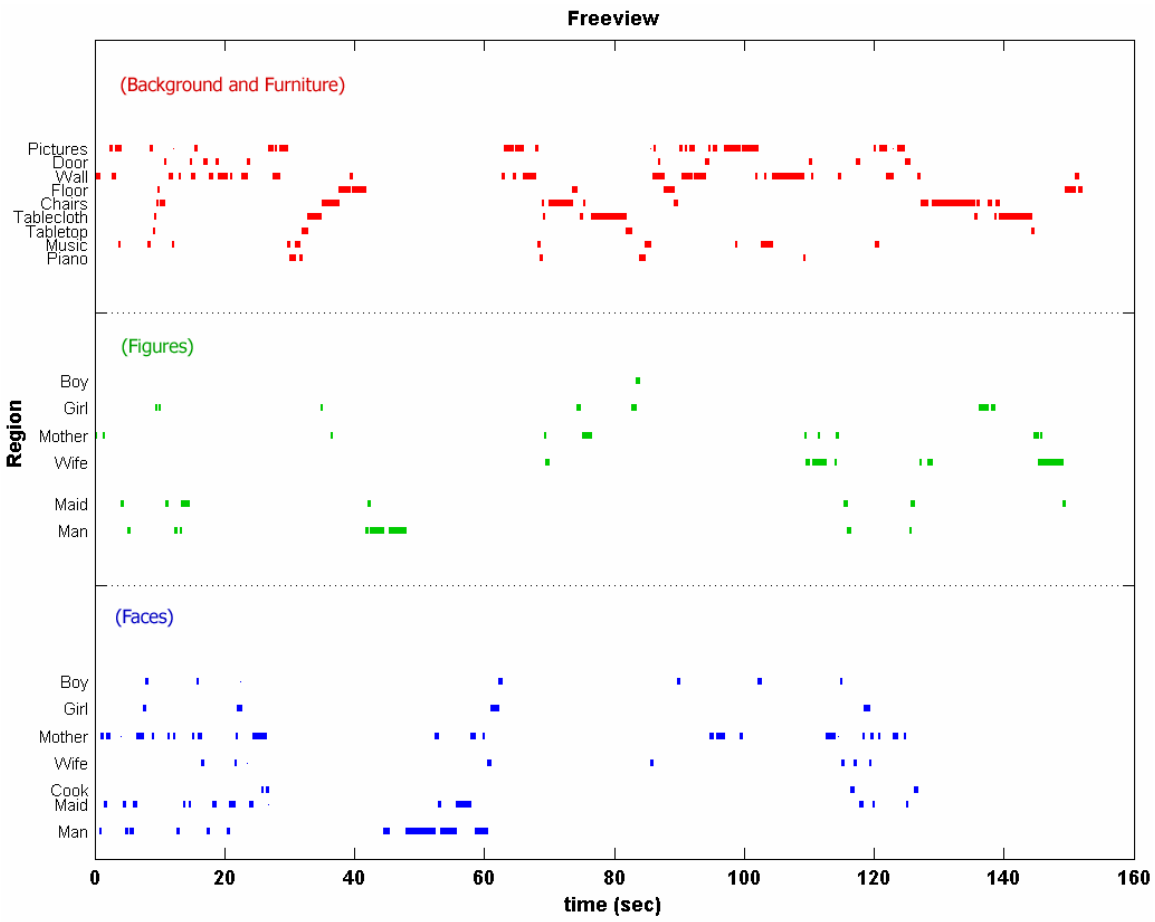


Figure 44: Temporal order for one subject during the "Freeview" task. The length of each line represents fixation duration. Each vertical level represents a different region. The bottom section, shown in blue, shows all of the Face regions. The regions in the middle section, shown in green, are the Figures. The top section, shown in red, shows all other regions. (The time does not extend to the full 3 minutes, or 180 seconds, because the time during blinks and saccades has been removed.)

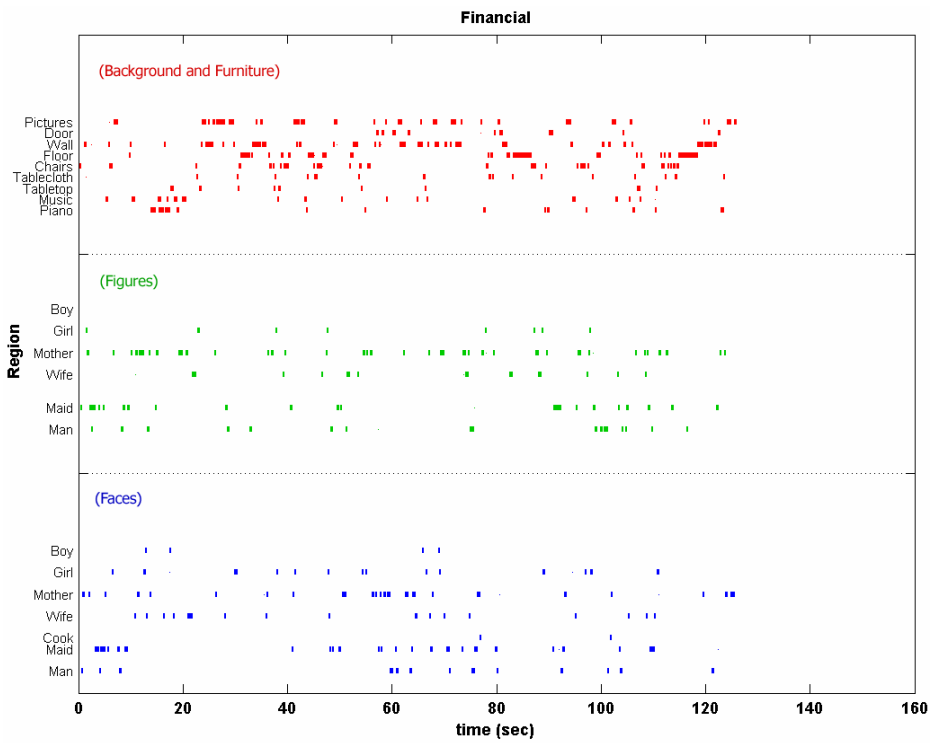


Figure 45: Temporal order of fixations during “Financial” task

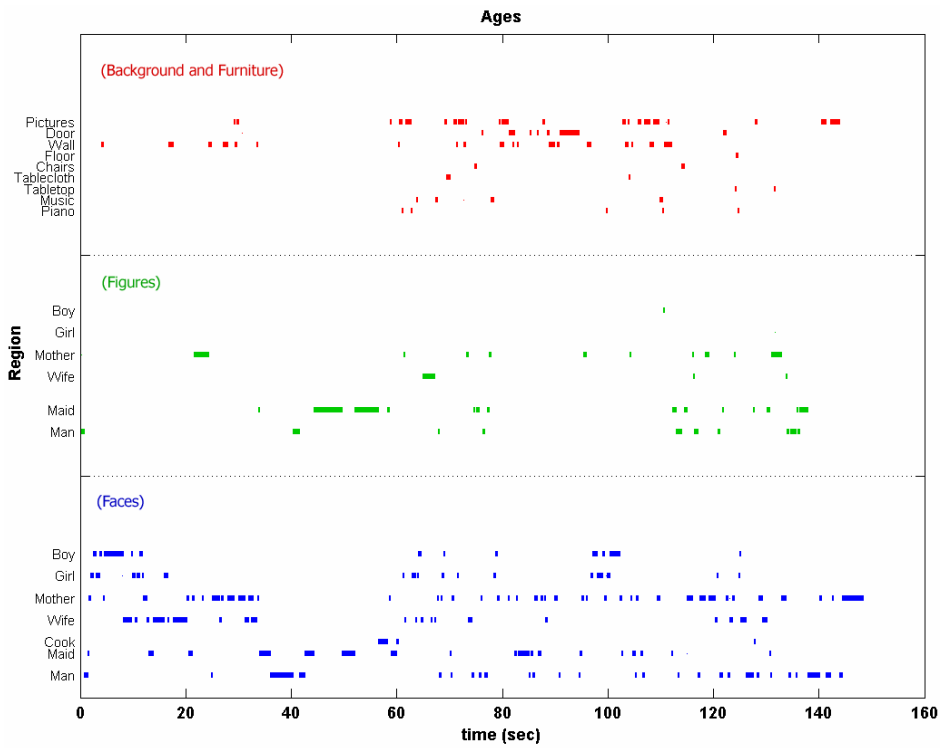


Figure 46: Temporal order of fixations during “Give the Ages” task

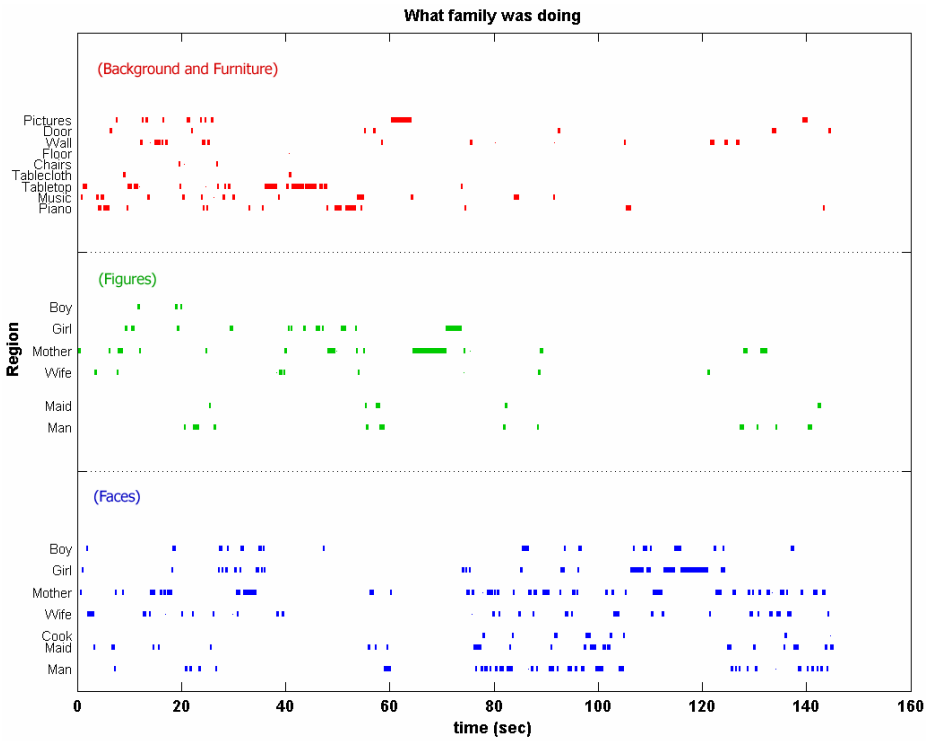


Figure 47: Temporal order of fixations during “What the family was doing” task

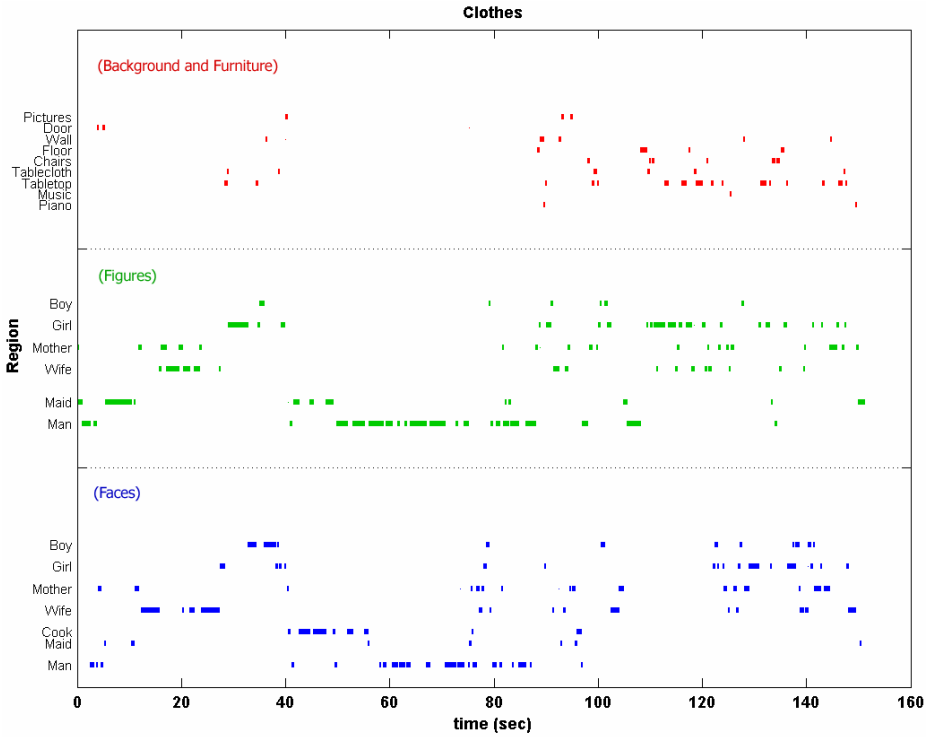


Figure 48: Temporal order of fixations during “Remember the clothes” task

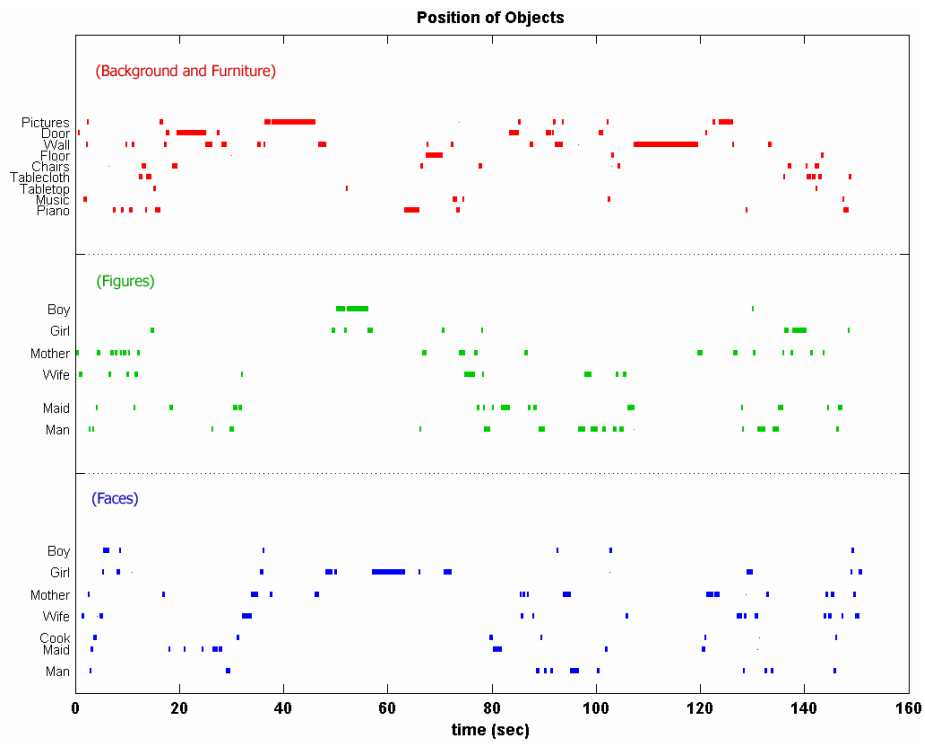


Figure 49: Temporal order of fixations during “Remember the position of people and objects” task

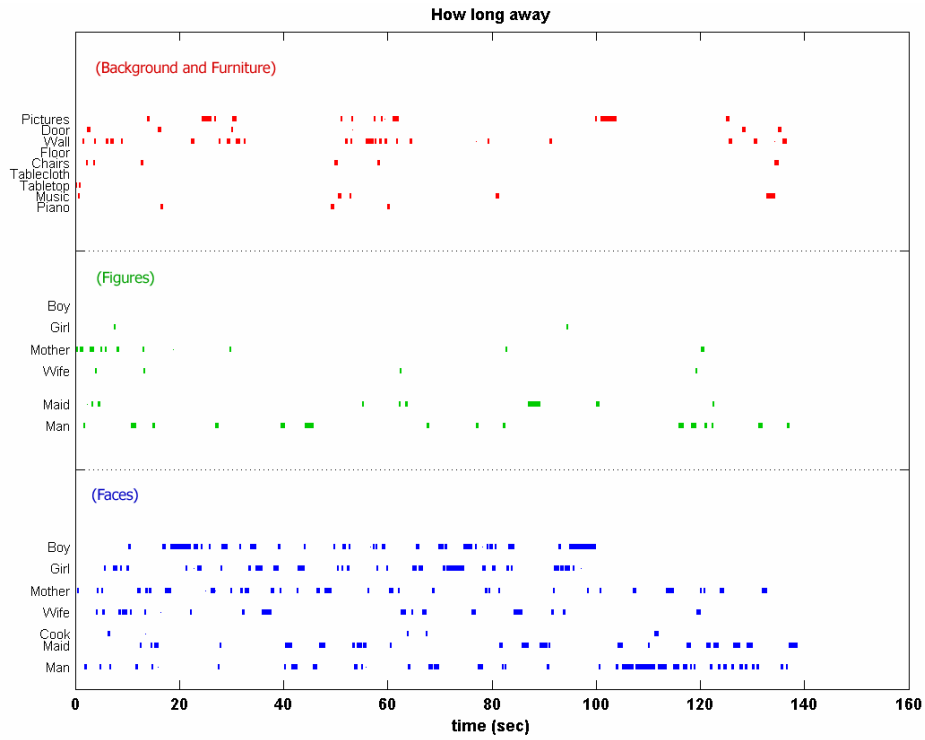


Figure 50: Temporal order of fixations during “How long away” task

4.7 Conclusions

This chapter presents the work of Alfred Yarbus and his classical experiment that showed the influence of task on viewing behavior. In his experiment, Yarbus had one subject view a painting 7 times for 3 minutes each time; before each viewing the observer was given a different instruction. From records of the spatial pattern of eye movements, Yarbus' concluded that the observer's fixations fell on regions that were most important or "informative" for the particular task. This experiment was replicated using current eyetracking technology that is not as restrictive or uncomfortable as the methods used by Yarbus.

Seventeen subjects were eyetracked as they performed each of seven tasks that Yarbus used in his experiment while viewing I. E. Repin's painting, "They did not expect him." These tasks were: 1) Freeview (no instruction); 2) "Estimate the financial circumstances of the family in the picture;" 3) "Give the ages of the people;" 4) "Surmise what the family had been doing before the arrival of the 'unexpected visitor;'" 5) "Remember the clothes worn by the people;" 6) "Remember the position of people and objects in the room;" and 7) "Estimate how long the 'unexpected visitor' had been away from the family."

Viewing time was self-terminated. The average view time for each task was 9, 19, 50, 25, 24, 29, and 15 seconds, respectively; these times are significantly less than the three-minute view time of Yarbus' subject. The task, "Give the ages of the people," elicited the longest fixations, with an average across all subjects of 376 milliseconds; fourteen percent of fixations during this task were longer than 1 second. The shortest average fixation duration was 250 milliseconds in the task "Estimate the financial circumstances of the family."

Figure 37 shows the amount of time spent viewing each of 22 regions per task, averaged across 17 observers. These results are in agreement with Yarbus' findings. Although the viewing times were much different, observers attended to specific regions of the image – the same regions that Yarbus' observer fixated on. In each of the tasks, the faces of the observers were fixated. Yarbus' single observer showed behavior typical of most subjects performing these tasks. However, some subjects' view patterns were not dramatically different between tasks, as shown in Figure 33.

The Euclidean distance between these 22-element feature vectors was used as a measure of both within-subject and between-subject variability. The average within-subject (between task)

distance was found to be 0.39. All tasks except for “How long away” resulted in between-subject average distances that were less than the within-subject distance, shown in Figure 38. The average distance for the “Give the ages” task was 0.26, which was the smallest of all tasks. This indicates that the eye movement patterns were most similar between the subjects, compared to the other tasks. The “How long away” task showed the most variability between subjects, suggesting that observers used different viewing strategies to complete this task.

Yarbus’ observation of a “cyclic” behavior of eye movements was also investigated. One subject performed each of the 7 tasks with an enforced three-minute view time. The temporal sequences of fixations are shown in Figure 44 through Figure 50. The viewing behavior for the “Financial” task was consistent across the viewing in that the subject frequently shifted his gaze between faces, figures, and background objects. Other tasks, including “Remember the Clothes,” show a distinct change in behavior. For this task, the subject began the task by examining the clothes and faces of the people in the scene, with some fairly long gaze durations. Halfway through the viewing time, the subject began to examine furniture and background elements. Yarbus’ subject did not show this behavior of moving attention to secondary elements, but instead spent the entire three minutes re-fixating on the ‘informative’ regions. This cyclic behavior may in part be a result of the uncomfortable setup of his experiment, which may have made his subject conscious of where he was looking. In doing so, the observer may have changed his viewing behavior to adhere strictly to the given instructions.

Chapter 5

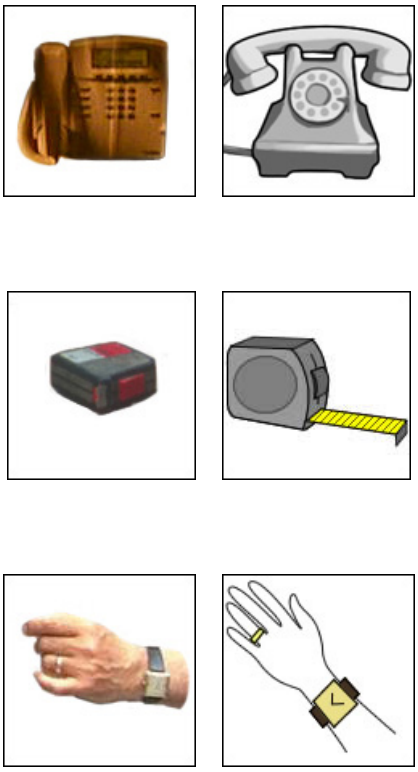
5 Visual Search Experiment

5.1 Overview

A visual search experiment was designed to investigate how the characteristics of a target preview affect the allocation of attention by the visual system. Several types of saliency maps are presented and used to examine differences in scene content at locations of fixation. Features that are considered include variants of color content, spatial structure, intensity, and edge content.

5.2 Methods

A visual search experiment was designed in which subjects searched for known objects in images of real-world scenes (shown in the Appendix). The experiment consisted of 60 trials, and was separated into two blocks of 30 trials. For one block, the target preview was pixel-for-pixel exactly the same as it was in the image to be searched. This condition will be referred to as the "Extracted Object" condition. For the other block of trials, the target preview was a cartoon representation of the object, which was varied in orientation, color, size, or number of details. This condition will be referred to as the "Cartoon Icon" condition. Figure 51 shows examples of Extracted Object and corresponding Cartoon Icon targets (see appendix for all sets of targets).



(a) (b)

Figure 51: Examples of targets used in the search experiments. Some subjects were presented with the Extracted Object version of the target (a), and others were presented with a Cartoon Icon representation (b).

Before each block, the subject read a short set of instructions and saw one example of a target and image pair. The subject then pressed the spacebar to view the first target for as long as needed. When the subject was ready, he or she pressed the spacebar to display the image to be searched. As soon as the target was found, the spacebar was pressed again. At that time the next target preview was shown, and so on.

The images were displayed at a resolution of 1280 x 768 pixels in 24-bit color. They were presented on a large-field plasma display which subtended approximately 50 x 35 degrees of visual angle (see Chapter 3 for more information). The target preview in each condition was shown on a white square, which subtended 5 x 5 degrees of visual angle. Within each block, images were randomized, and block order was balanced across subjects, as shown in Table I.

Table I: Experimental Conditions for different sets of subjects, balanced for order of presentation and image set

Set A1

	Target Condition	Images
Part 1:	Extracted Object	Images #1-30
Part 2:	Cartoon Icon	Images #31-60

Set A2

	Target Condition	Images
Part 1:	Cartoon Icon	Images #1-30
Part 2:	Extracted Object	Images #31-60

Set B1

	Target Condition	Images
Part 1:	Extracted Object	Images #31-60
Part 2:	Cartoon Icon	Images #1-30

Set B2

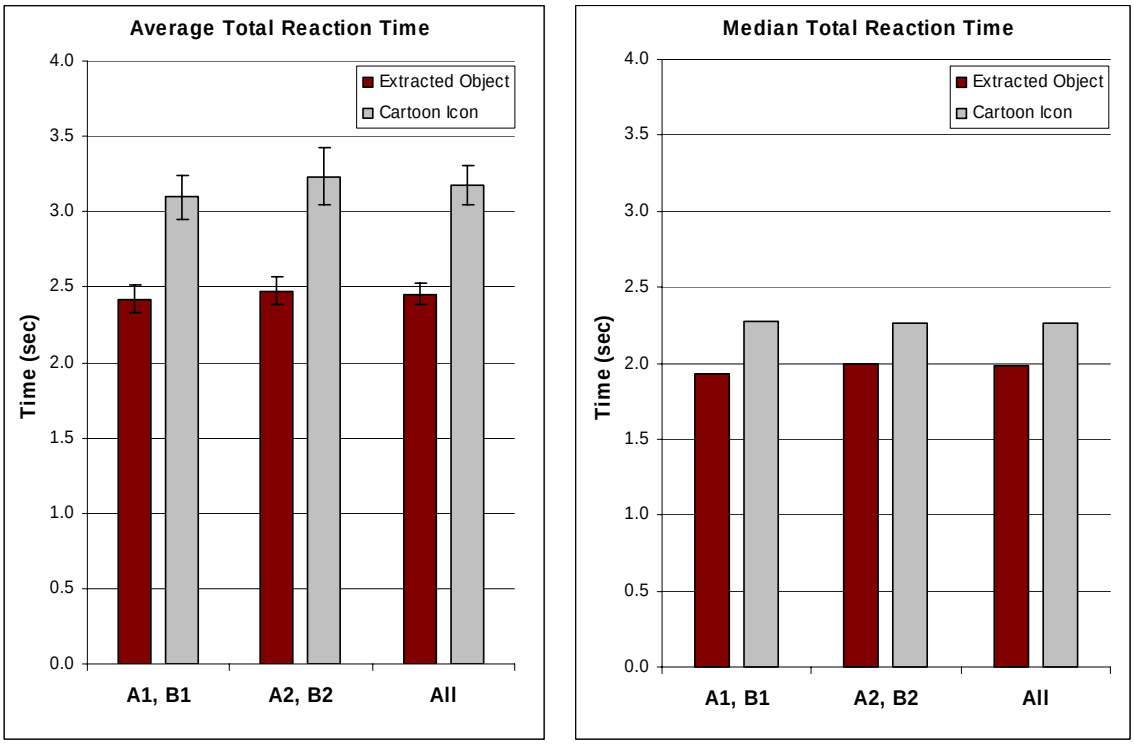
	Target Condition	Images
Part 1:	Cartoon Icon	Images #31-60
Part 2:	Extracted Object	Images #1-30

Subjects were eyetracked using an Applied Science Laboratories Model 501 video-based eyetracker in conjunction with a Polhemus Magnetic Head Tracker. By using both, the intersection of a person's line of gaze on the display plane can be computed at 60 Hz without constraining the person's head (see Chapter 3 for more information). Twenty-three subjects were successfully eyetracked.

5.3 Reaction Time Results

The following results from the experiment contain reaction times from 53 of the 60 images. Images in which some subjects could not find the target in less than 20 seconds were discarded. Additionally, each subject's first trial was discarded. Eye movement data was used to verify that the subjects fixated on the true target. Figure 52 shows the average and median of all reaction times in each target condition pooled across all subjects and all images. The total reaction time is calculated as the time between the image appearing on the screen and the subject pressing the spacebar. On average, subjects found the target in 2.5 seconds when presented with the Extracted Object, and found the target in 3.1 seconds in the Cartoon Icon condition. This difference in time is seen regardless of which block condition was presented first. The difference in median times was smaller, but still consistent. The median reaction times were 2 and 2.3 for Extracted Object and Cartoon Icon conditions, respectively. Figure 12 shows the median reaction

times in each condition for each subject. Twenty of the 23 subjects had a longer median reaction time in the Cartoon Icon condition. Figure 13 shows a histogram of reaction times in each target condition. A Wilcoxon Rank-Sum test at alpha level of 0.05 gives a p-value of 5.4×10^{-8}, meaning the means of the two distributions are significantly different.



(a) (b)

Figure 52: Average (a) and median (b) reaction times for each target condition across all subjects and images. Error bars represent one standard error of the mean. Subjects in sets A1 and B1 were presented with the Extracted Object block first, then Cartoon Icon. Sets A2 and B2 saw the Cartoon Icon block first, then Extracted Object.

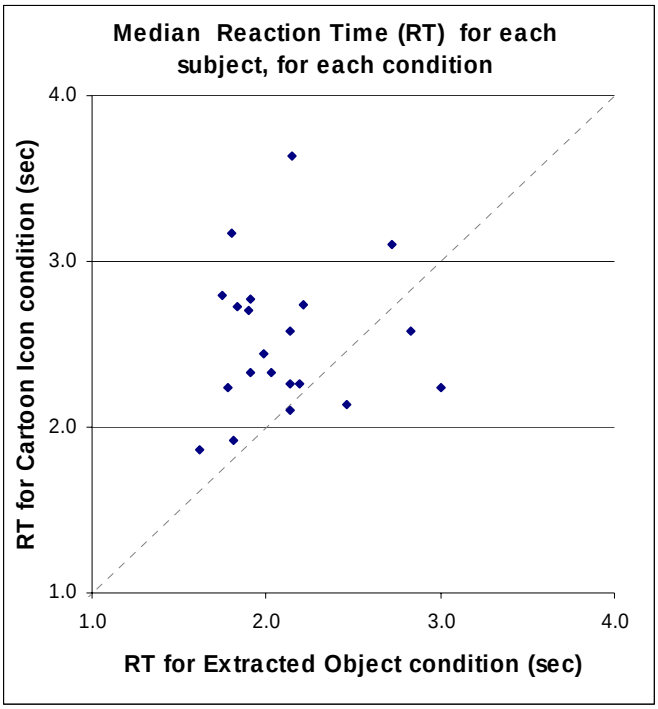


Figure 53: Median reaction times for the two conditions, per subject. Twenty of twenty-three subjects had a longer median reaction time for the Cartoon Icon condition.

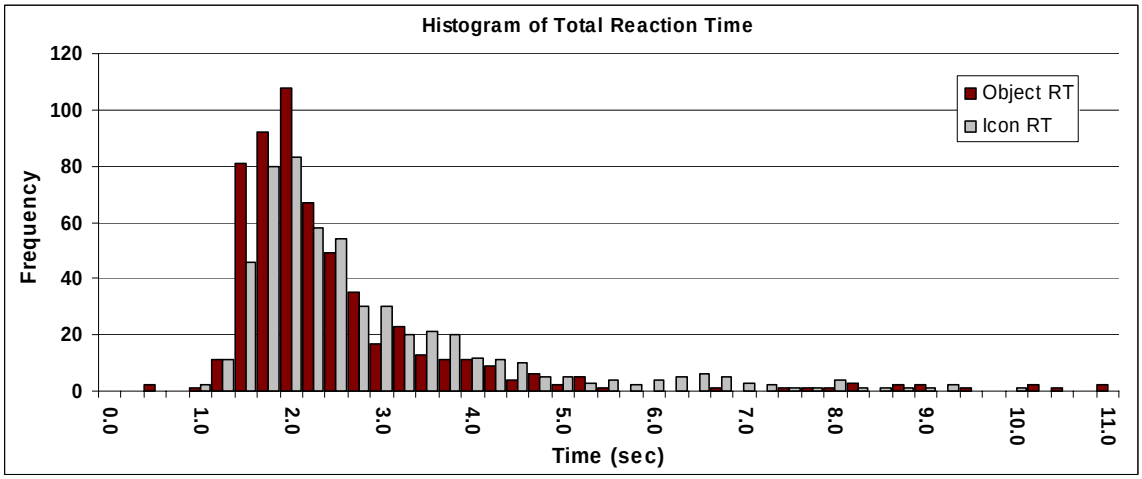


Figure 54: Histogram of reaction times in each target condition

Figure 55 shows the average reaction times in each condition for 53 images. In some images, reaction times were nearly equal (points falling close to line with a slope of 1). However, many images have much longer Cartoon Icon reaction times. In some cases, it proved easier to find the target object in the Cartoon Icon condition than in the Extracted Object condition.

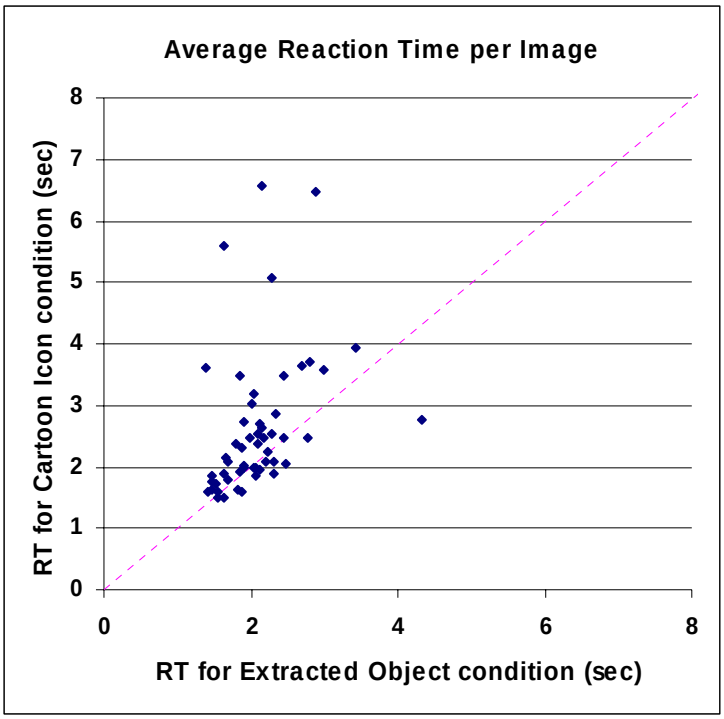


Figure 55: Average reaction times for the two conditions for 53 images.

One hypothesis was that the increased amount of time needed to find the target in the Cartoon Icon condition was a result of subjects being uncertain that the item in the image was in fact the target. In this case, a person would take the same amount of time to initially fixate on the target, but spend more time 'deciding' or 'reacting' in the Cartoon Icon condition. However, further analysis of the eye movement data showed otherwise. On average, it took subjects longer to initially fixate on the target in the image in the Cartoon Icon condition, as shown in the first pair of columns in Figure 56. The time between first target fixation and the pressing of the spacebar (including subsequent target fixations and fixations elsewhere) did not show as large of a difference between conditions. However, both differences were significant at a level of $\alpha = 0.05$ ($p = 0.015$ in both cases). This suggests that knowing the target's exact features helps expedite visual search.

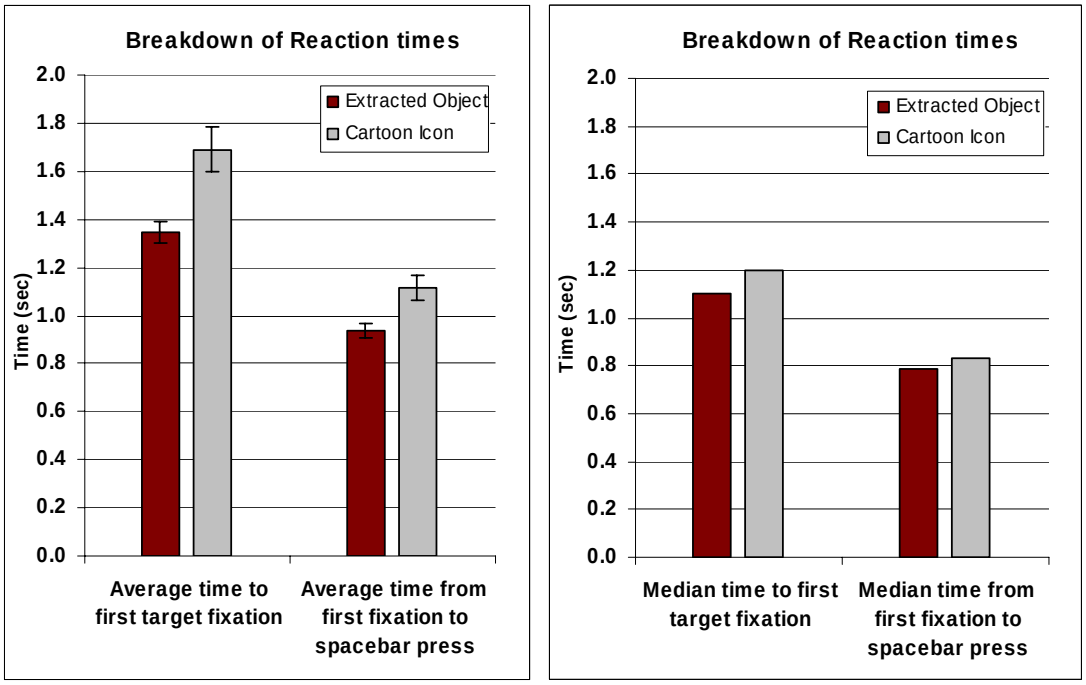


Figure 56: Breakdown of reaction times during visual search. (a) shows the average time it took before a subject first fixated on the target object in the image, and the average time between the first target fixation and when the spacebar was pressed. Error bars represent standard error. (b) shows the medians of the same data.

A breakdown of reaction times for Image 51 is shown in Figure 57. In this image, the target was a purple sunburst shape on a woman's shirt. The Cartoon Icon representation was of a similar shape, but in grayscale, shown in Figure 58. Figure 57 shows that knowing the color features of the target improved the reaction time. In this case, the average time until the first target fixation was 1.2 seconds in the Extracted Object condition, and 2.4 seconds in the Cartoon Icon condition. Again, the time between first target fixation and spacebar press were not significantly different. Figure 59 and Figure 60 show eye movement records of two subjects performing the visual search task for this image under each condition. When shown the Extracted Object target preview, the subject found the target immediately, in one saccade. The subject who saw the Cartoon Icon preview took longer to find the target.

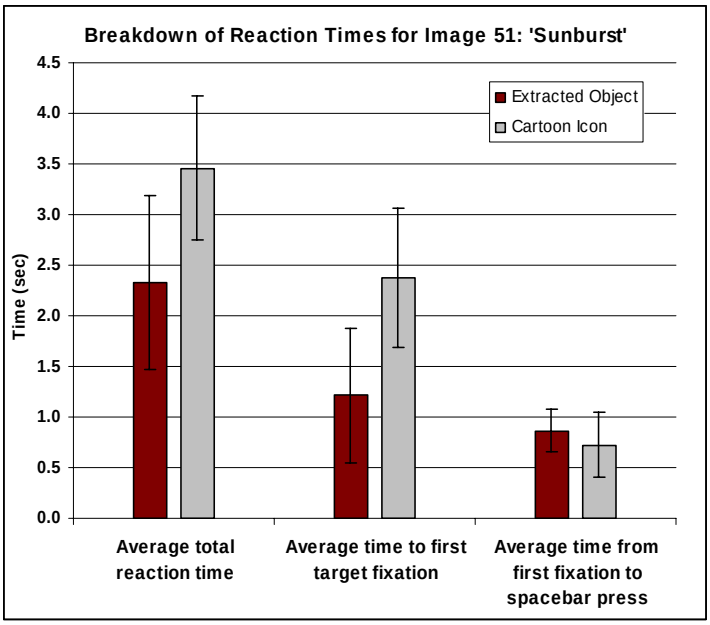
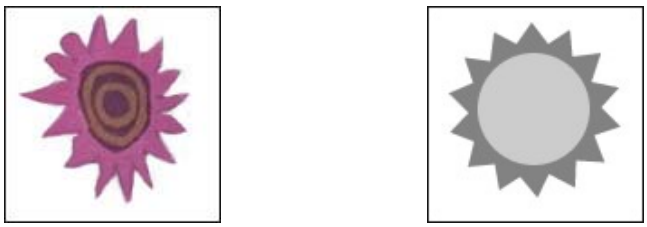


Figure 57: Breakdown of Reaction Times for Image 51: Sunburst. Error bars represent standard error for 7 subjects in the Extracted Object condition, and 8 subjects in the Cartoon Icon condition.



(a) (b)

Figure 58: Target previews for Image 51. (a) Extracted Object. (b) Cartoon Icon.

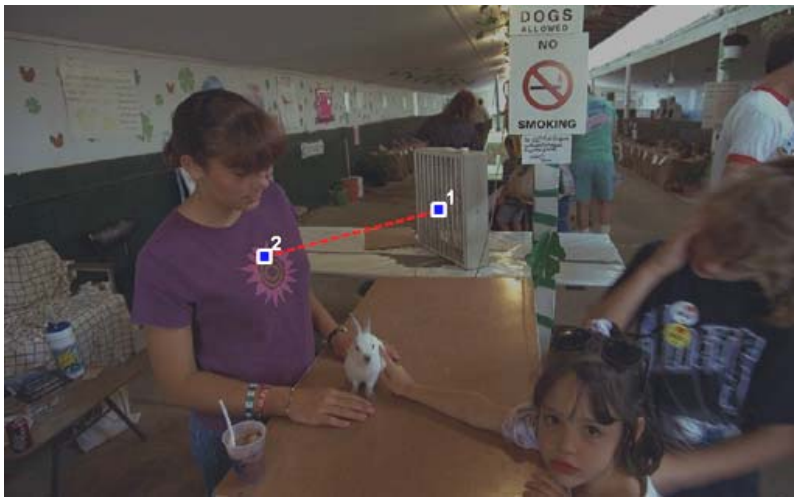


Figure 59: Eye movement record of one subject during the visual search task. The subject was shown the Extracted Object preview image of the target (Figure 58 a), and found the target in one saccade.

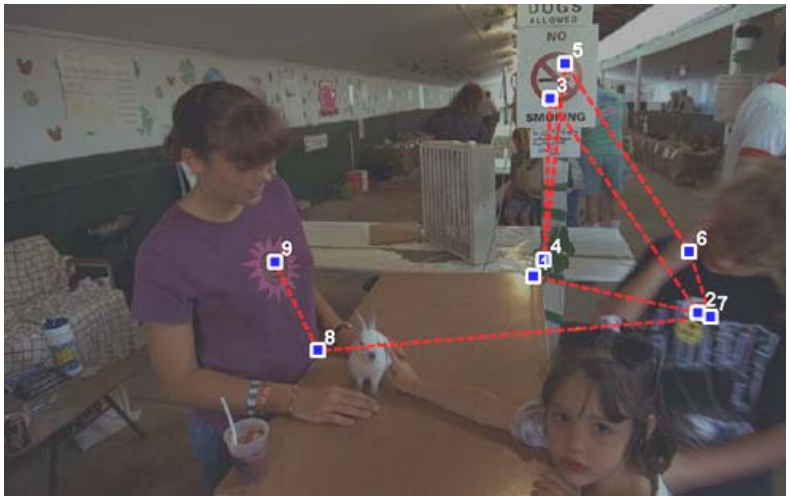


Figure 60: Eye movement record of one subject during the visual search task. The subject was shown the Cartoon Icon preview image of the target (Figure 58 b), and did not find the target immediately.

5.4 Topographical Feature Maps

To investigate what features are used by the visual system to perform the visual search task in the different conditions, feature content at fixated locations was extracted. A set of topographical feature maps was created and used to compare different types and combinations of features. The following section describes the creation of each map.

5.4.1 List of Maps

[1] Original Image in ***RGB*** color space values ('RGB')

Figure 61: Example image

[2] Image in 1976 ***CIELab*** color space values ('Lab'). The RGB image was converted via a Matlab function, "srgb2lab."

---

[3] ***Intensity***, ('I_RGB'), where

$$I = \frac{(R + G + B)}{3}$$

[4] ***Intensity***, ('I_Lab'), where

$$I = L^*$$

[5] ***Intensity***, ('I_Cone'), where I is defined by a model of rod and cone responses to the stimulus. From [Canosa, 2003].

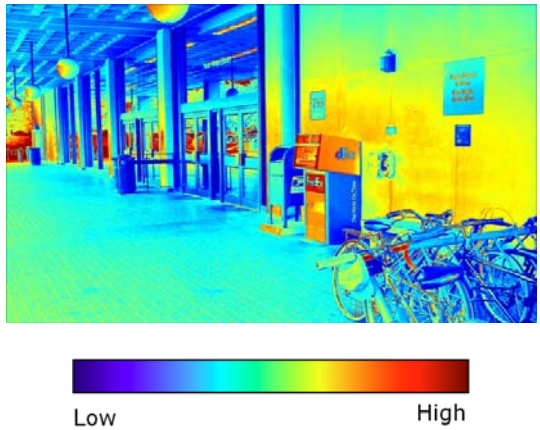


Figure 62: Example ‘I_rgb’ map

---

[6] ***Colorfulness***, (‘C_RGB’) where

$$C = |R - G| + |B - (R + G)|$$

[7] ***Colorfulness***, (‘C_Lab’), where

$$C = |a^*| + |b^*|$$

[8] ***Colorfulness***, (‘C_Cone’), where C is defined by a model of rod and cone responses to the stimulus. From [Canosa, 2003].

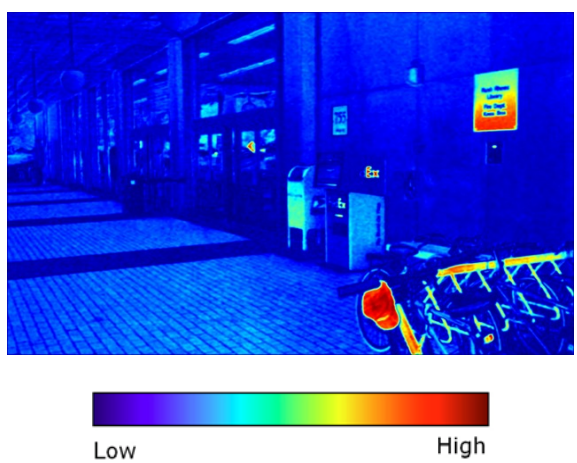


Figure 63: Example ‘C_rgb’ map

---

[9] ***Oriented edge*** ('Edge') content, defined by [Canosa, 2003]

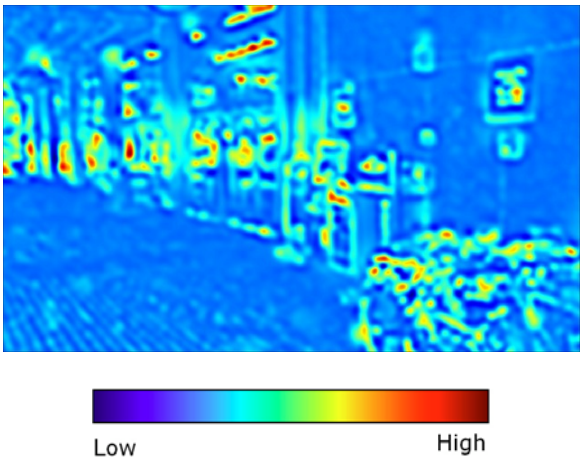


Figure 64: Example 'Edge' map

[10] ***Proto-object*** map ('P_object'). This is a binary map that defines regions that are likely to be objects using figure/ground segmentation techniques. See Chapter 2 for a more detailed description. From [Canosa, 2003].

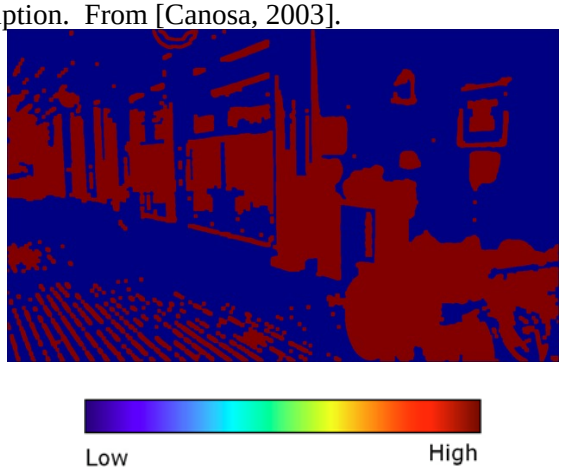


Figure 65: Example 'P_object' map

[11] ***Saliency*** (‘CIE_RGB’), where

$$CIE\_RGB = C\_RGB + I\_RGB + Edge$$

[12] ***Saliency*** (‘CIE_Lab’), where

$$CIE\_Lab = C\_Lab + I\_Lab + Edge$$

[13] ***Saliency*** (‘CIE_Cone’), where

$$CIE\_Cone = C\_Cone + I\_Cone + Edge$$

[14] ***Conspicuity*** (‘CIEP_Cone’), from [Canosa, 2003] where

$$CIEP\_Cone = (C\_Cone + I\_Cone + Edge) \cdot P\_object$$

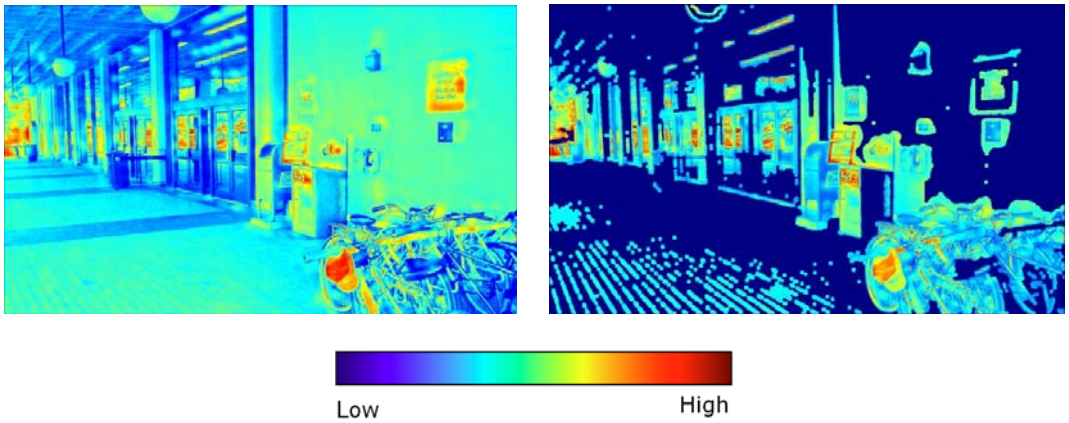


Figure 66: Example ‘CIE_rgb’ (left) and ‘CIEP_cone’ (right) maps

[15] ***Histogram backprojection*** – Extracted Object

Beginning with the image in L*a*b* color space, the image is quantized and indexed into 1000 colors as follows. Each axis in L*a*b* space was divided into ten sections, as shown in Figure 67; each voxel was given an index number. For each pixel in the image, the voxel in which the pixel's L*a*b* value resides was found. An indexed image was created by inserting the voxel index number into the pixel location.

The same process was performed for the Extracted Object target. Using this indexed target, a one-dimensional histogram was created, which then was used as a look-up table. Each pixel in the indexed image was replaced by the histogram value of the corresponding color (index value).

Figure 68 shows an example image and a target. The target, a tree, contains mostly green and a lesser amount of brown. In the output backprojected image, green regions will be given a high value, and brown regions a smaller value. All other regions will be set to zero, as shown in Figure 69.

[16] ***Histogram backprojection*** – Cartoon Icon

This map is made in the same way as the previous one, except that the color histogram of the Cartoon Icon target is used. For some Cartoon Icons, a color that appears in the target does not appear in the image.

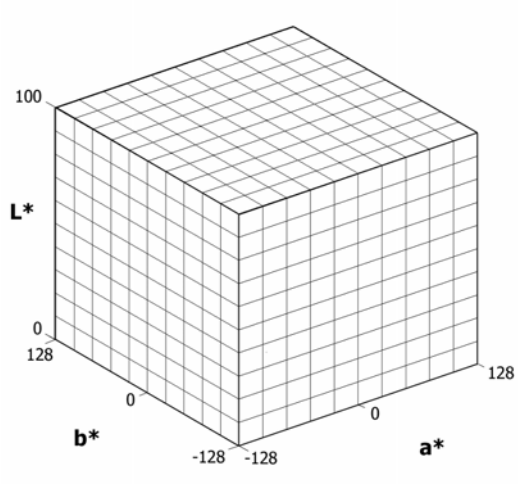


Figure 67: Color cube used to create indexed images in L*a*b* space

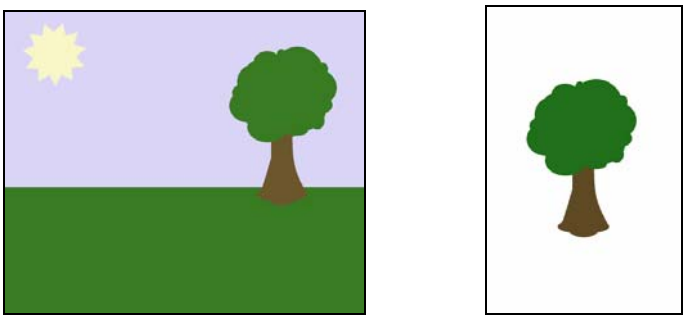


Figure 68: Example image and target.

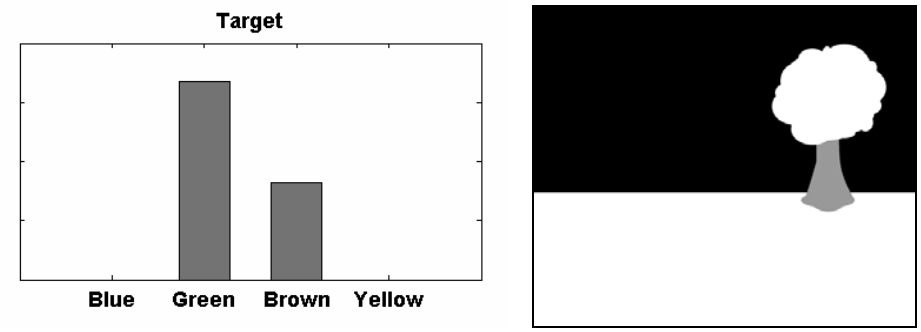


Figure 69: Target histogram and resulting backprojected image. Green regions in the image were assigned a high value (white), brown regions were assigned a lower value (gray), and regions of other colors were set to zero (black).

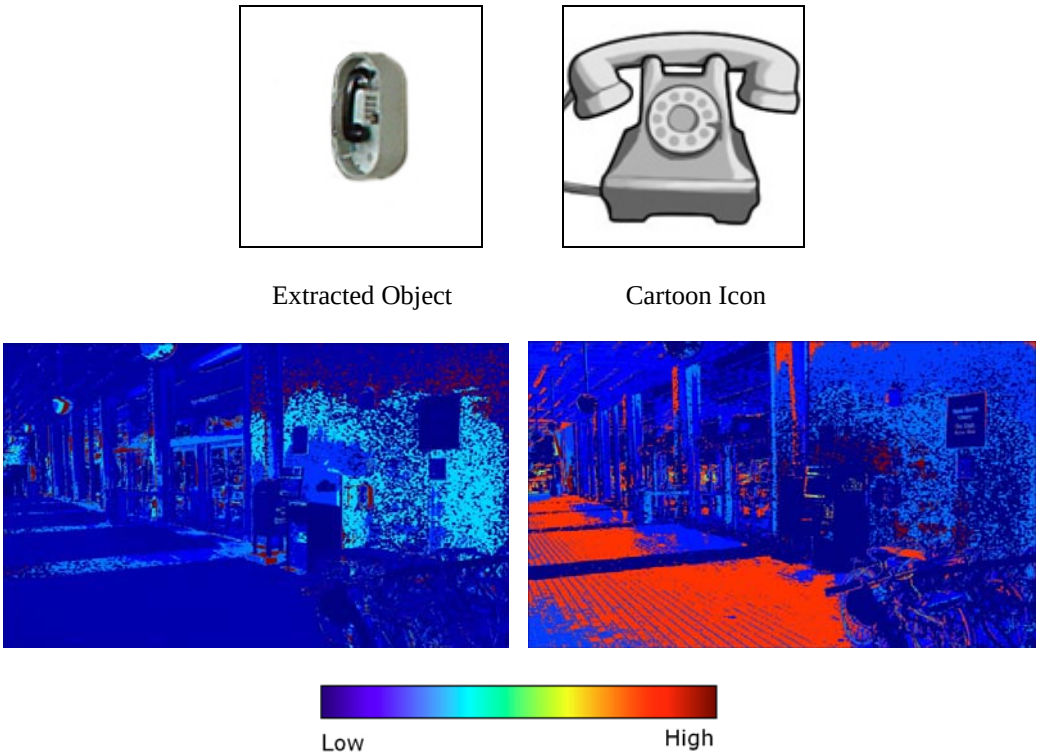


Figure 70: Example targets, ‘Hist_object’ (left), and ‘Hist_icon’ (right) maps.

---

[17] ***Ratio histogram backprojection*** – Extracted Object

This algorithm is similar to the backprojection process described above. Instead of assigning the target histogram values to corresponding pixels, values of a ratio histogram are used. The ratio histogram is simply the target histogram divided by the image histogram. An example for the tree image is shown in Figure 71. By using the ratio histogram values as a look-up table, the colors that are heavily present in the image are suppressed, and the colors that are less common are weighted more. The resulting image is shown in Figure 72.

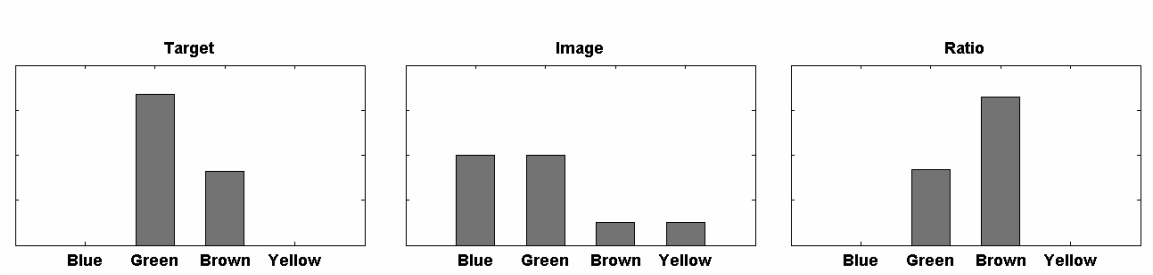


Figure 71: Target, image, and ratio histograms. The ratio histogram is made by dividing the target histogram by the image histogram.

Figure 72: Result of ratio histogram backprojection.

[18] ***Ratio histogram backprojection*** – Cartoon Icon

This map is made in the same way as the previous one, except that the color histogram of the Cartoon Icon target is used. It is possible, although unlikely, that a color that appears in the target does not appear in the image.

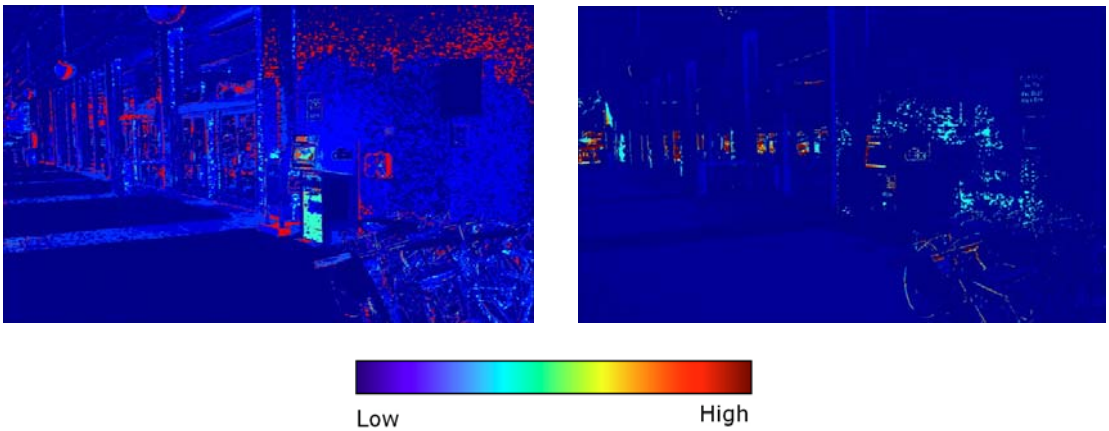


Figure 73: Example 'Ratio_object' (left) and 'Ratio_icon' (right) maps. Targets are shown in Figure 70.

[19] ***Spatial Cross-correlation*** – Extracted Object

The L* channel of the image is convolved with the L* channel of the Extracted Object target. This process produces large values where the image closely matches the spatial structure of the target. The absolute value of the image is used as the final map so that large, negative correlations are also considered.

[20] ***Spatial Cross-correlation*** – Cartoon Icon

This process is the same as above, except that the Cartoon Icon target is convolved with the image.

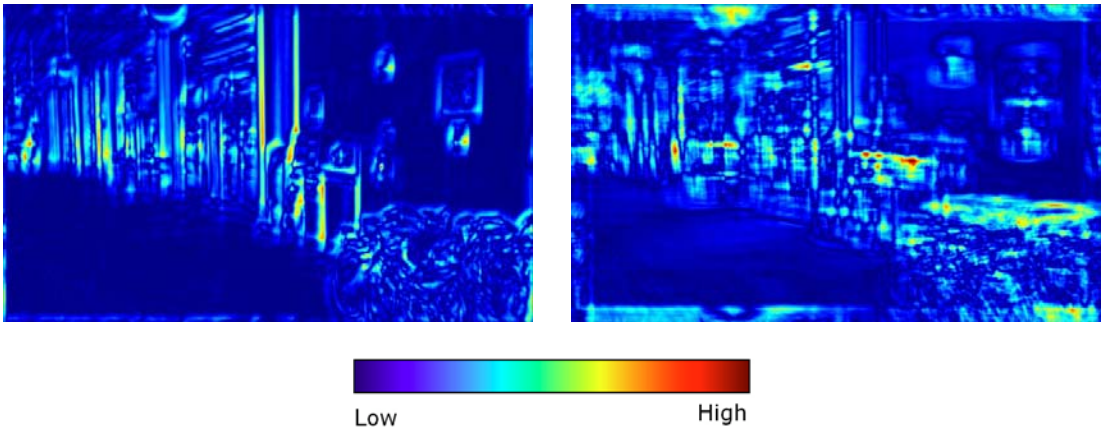


Figure 74: Example 'Spatial_object' (left) and 'Spatial_icon' (right) maps. Targets are shown in Figure 70.

5.5 Extraction of Features at Fixated Locations

For each of the maps listed above, content at fixated locations in each of the two visual search conditions was extracted. Fourteen subjects' eye movement records were used for this analysis; there were 7 subjects for each target preview condition. Circular image patches of a radius of 25 pixels, which corresponds to a radius of approximately 1 degree (depending upon the distance between the subject and the display), were extracted. In current literature, the maximum saliency value of each patch is used. Effects of using the maximum value, instead of using the average value, are explored. Four performance metrics, as described in Chapter 2 (page 24), are used to compare results. These include: Chance-adjusted Cumulative Probability (CPa) [Parkhurst, 2000]; F/M Ratio [Canosa, 2003]; ROC Curve Area [Tatler, et al, 2004]; and Chance-Adjusted Saliency 'Accumometric' [Carmi and Itti, 2004]. Additionally, Chance-adjusted Cumulative Probability and F/M Ratios were also computed using fixation durations as weights. For the metrics requiring a set of non-fixated points, locations were chosen at random from a pool of locations that were fixated across all subjects and all conditions (~8000 fixations). The distribution of all fixations is shown in Figure 75. Uniform random locations were not used due to the central biasing of eye movements and salience in natural images [Parkhurst, 2000; Canosa, 2003].

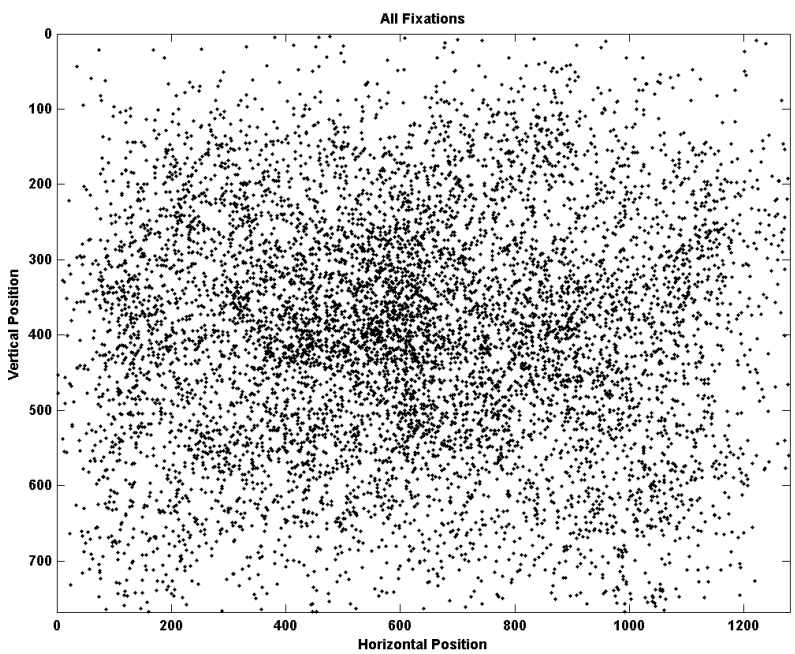


Figure 75: All fixated locations throughout the entire visual search experiment

5.5.1 Effects of using the maximum versus average value at fixation locations

Figure 76 shows the histogram of all map values for the CIE_rgb map of Image 24 (shown as the shaded gray curve). Also shown on the graph is the histogram of maximum values of the patches extracted at fixation locations (solid red line). The solid black line represents the maximum values of patches at non-fixated locations. These locations were selected at random from the pool of fixated locations shown in Figure 75. Note that even though these locations are random, the histogram does not approximate that of the entire map due to the act of selecting only the maximum value around each location.

The red dotted line is the distribution of values at fixated locations when the average around each location is used. Similarly, the black dotted line is the distribution of values at random, non-fixated locations when the average around each location is used. Here the distribution of random locations more closely approximates the distribution of the entire image. However, the separation between the distributions at fixated and random locations is reduced by the act of selecting the average value around each fixation location.

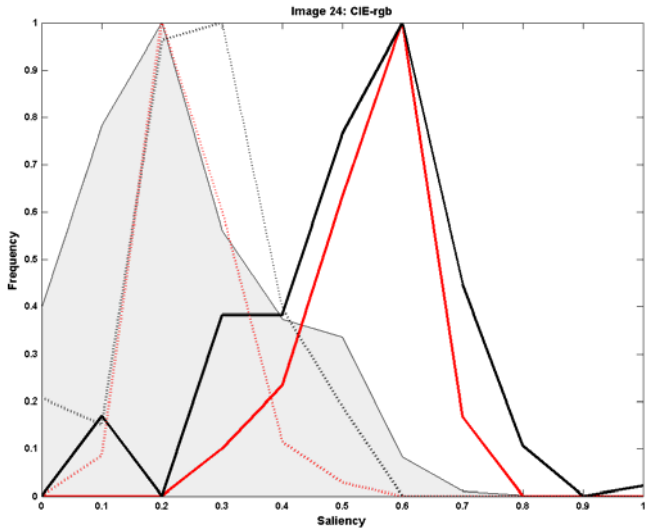


Figure 76: Histogram of feature map values for the CIE_rgb map for Image 24 (shown in gray); using maximum map values at fixated locations (solid red line); using maximum map values at random non-fixed locations (solid black line); using average map values at fixated locations (dotted red line); using average map values at random non-fixated locations (dotted black line)

Table II: Performance values for Image: 24, Map: CIE_rgb

	Using Average	Using Max	Difference	% Difference
CPa	-0.06	0.88	0.94	106
F/M Ratio	1.03	1.59	0.56	35
ROC Curve Area	0.45	0.61	0.16	27
CASA	-0.01	0.05	0.06	129

Table II shows the value of each performance metric for the example image and map shown in Figure 76. The percent difference between the conditions was calculated as the difference divided by the value when using the maximum value of the patch. For example, (0.88 - -0.06) / 0.88 = 1.06 = 106% difference. When using the average of each patch, the resulting CPa value is near zero, indicating that the probability of selecting those feature map values is no greater than chance. However, when using the maximum of each patch, the CPa is 0.88. This value is very high, indicating the selection of those feature map values is much greater than chance. However, after selecting the maximum of each patch, comparing to the mean of the entire map is not truly comparing to chance. Instead, comparing to the maximum values at random locations may be more appropriate.

The same trend is seen in each of the other three metrics. The ROC Curve Area metric shows the smallest increase between these two conditions. When using the average patch value, the ROC Curve Area for this image is 0.45, which means that the distributions of values at fixated and non-fixated are nearly indistinguishable. When using the maximum value, the ROC Curve Area increases to 0.61.

Figure 77 shows the same visualization for Image 57. In this example, the distributions of map values at fixated and non-fixated locations are more distinct. In comparison to Image 24, it is expected that the performance metrics would indicate that this map does a good job assigning high saliency values to locations where people look. Each of the performance metrics, shown in Table III, is indeed higher than the one shown in Table II. Comparing the values when using the average versus maximum of each patch, the increase in CPa and F/M Ratio is not as drastic as in the previous example, primarily because this map has a higher mean value. The ROC Curve Area again is the least affected by using the average versus maximum of each patch because it is not affected by where the distributions lie on the saliency axis. The ROC Curve is reduced somewhat when using the average value because that act of averaging will reduce the separation between

the two distributions. The CASA metric is also, ideally, unaffected by a shift in the distributions along the saliency axis; however in this case it is more sensitive than the ROC Curve Area metric to whether the maximum or average of the patch is used.

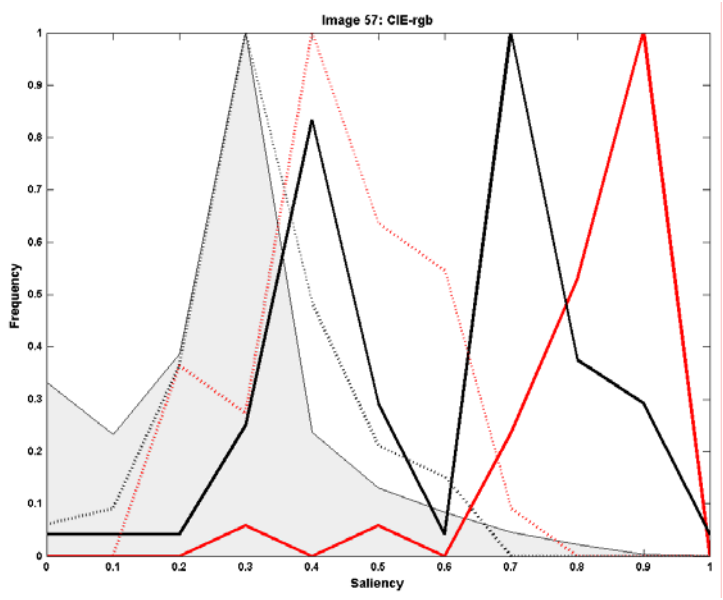


Figure 77: Histogram of feature map values for the CIE_rgb map for Image 57 (shown in gray); using maximum map values at fixated locations (solid red line); using maximum map values at random non-fixed locations (solid black line); using average map values at fixated locations (dotted red line); using average map values at random non-fixated locations (dotted black line)

Table III: Performance values for Image: 57, Map: CIE_rgb

	Using Average	Using Max	Difference	% Difference
CPa	0.61	0.97	0.37	38
F/M Ratio	1.50	2.09	0.59	28
ROC Curve Area	0.74	0.81	0.08	10
CASA	0.10	0.18	0.08	45

5.5.2 Effects of using Different Color Spaces

Figure 78 shows a comparison of the performance of the three types of Intensity maps. The performance metrics were calculated when using the average of each patch surrounding a fixation location (first column of Figure 78) as well as when using the maximum value (second column). As described above, each of the metrics showed an increase when the maximum value of the patch was used. For this condition, the F/M Ratio, shows a large increase in performance for the I_cone map. The CASA metric also shows a slight increase. These differences are most likely due to the fact that the mean of the I_cone map is usually much lower than either of the other two maps, as shown in Figure 79. For maps scaled from 0 to 1, the mean value of the I_cone map across 60 images was 0.25; the means for the I_rgb and I_lab maps were 0.40 and 0.43, respectively.

Table IV show the results of paired t-tests between each map for the condition in which the average value of the patch was used. When comparing the I_rgb, I_lab, and I_cone maps, the CPa metric shows better performance by the I_lab map, whereas the F/M Ratio and ROC Curve Area metrics show better performance for the I_cone map. The CASA metric shows no significant difference in performance across the three maps.

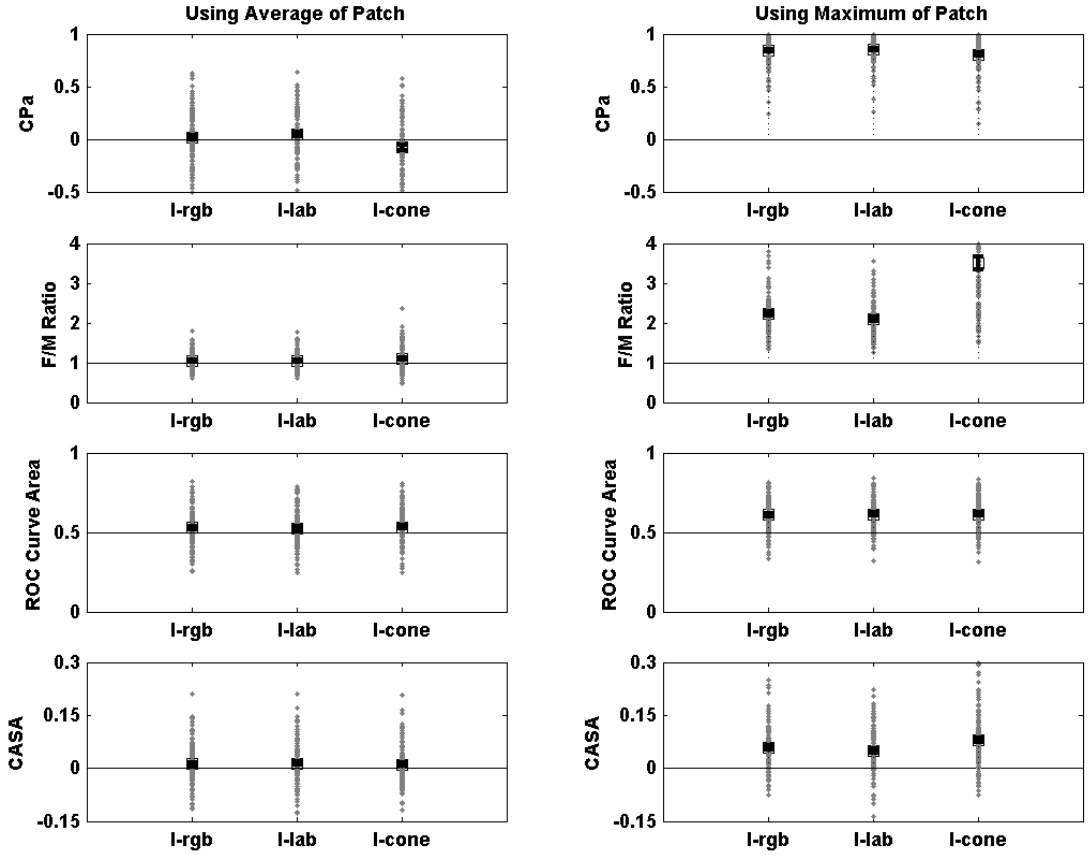


Figure 78: Comparison between the three types of Intensity maps using four performance metrics. The first column shows results when the average of the patch surrounding the fixation location is used; the second column shows the results when the maximum value is used. Each gray point represents the value for one image and one search target condition; the square represents the mean, and error bars represent one standard error of the mean.

Table IV: Results of paired t-test between the performances of each map (when the average of each patch was used) A value of 1 indicates a significant difference between the means.

CPa

	I_lab	I_cone
I_rgb	1	1
I_lab		1

ROC Curve Area

	I_lab	I_cone
I_rgb	0	0
I_lab		1

F/M Ratio

	I_lab	I_cone
I_rgb	0	1
I_lab		1

CASA

	I_lab	I_cone
I_rgb	0	0
I_lab		0

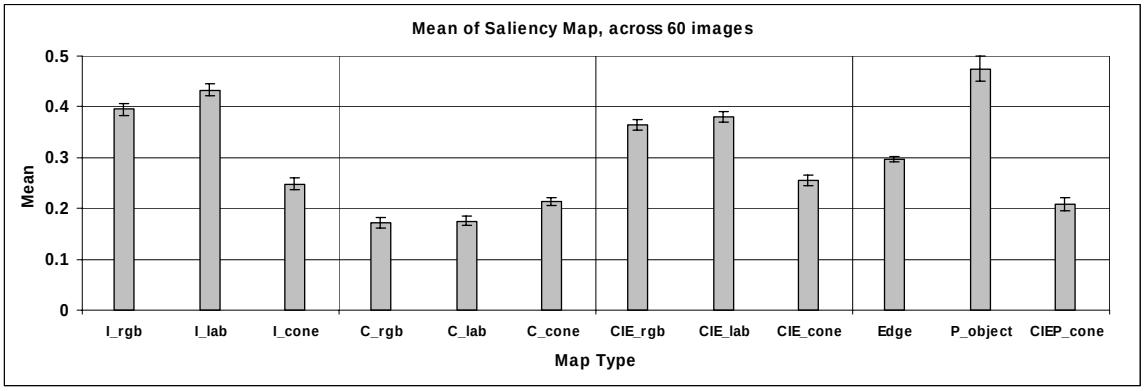


Figure 79: Means of saliency maps after scaling from 0 to 1, averaged across 60 images. Error bars represent standard error of the mean.

The same analysis was performed on the three types of Colorfulness maps, as well as the three types of CIE maps. The results are shown in Figure 80 and Figure 81. Results from paired t-tests are shown in Table V and Table VI. Three of the four metrics show little difference in performance between the C_rgb and C_lab maps; the F/M Ratio metric shows that the C_rgb map performed better than the C_lab. The CPa metric shows performance is below chance, on average, for the C_cone map when using the average value of each patch; this difference in performance between the maps is not shown in the CPa values when the maximum of the patch is used. The F/M Ratio metric also shows a lower value for the C_cone map compared to the others; this is a result of the C_cone map having a higher mean compared to the other two, shown in Figure 79. Each of the metrics shows that the C_cone map's performance value was significantly less than the others'.

When comparing the three types of CIE saliency maps created by adding the colorfulness, intensity, and oriented edges maps, the ROC Curve metric shows no significant difference between the maps. The F/M Ratio metric shows an increase in performance for the CIE_cone map when using the maximum value of each patch, possibly due to the fact that this map has a smaller mean value, on average, than the others (shown in Figure 79). Interestingly, when using the average value of each patch, the CPa metric shows a decrease in performance for the C_cone map. This may be caused by differences in the shape of the distributions. If the distribution of the C_cone has a small mean and a long tail, as in Figure 14 B, the rate at which the CPa value increases is lower than it would be for a distribution such as in Figure 14 A or D.

The results of the paired t-tests show that when using the CPa metric, the CIE_lab map performed better than the CIE_rgb map, which performed better than the CIE_cone map.

However, the F/M Ratio metric showed that the CIE_rgb map performed better than the CIE_lab map. Each of the metrics shows that the CIE_cone map's performance value was significantly less than the others'.

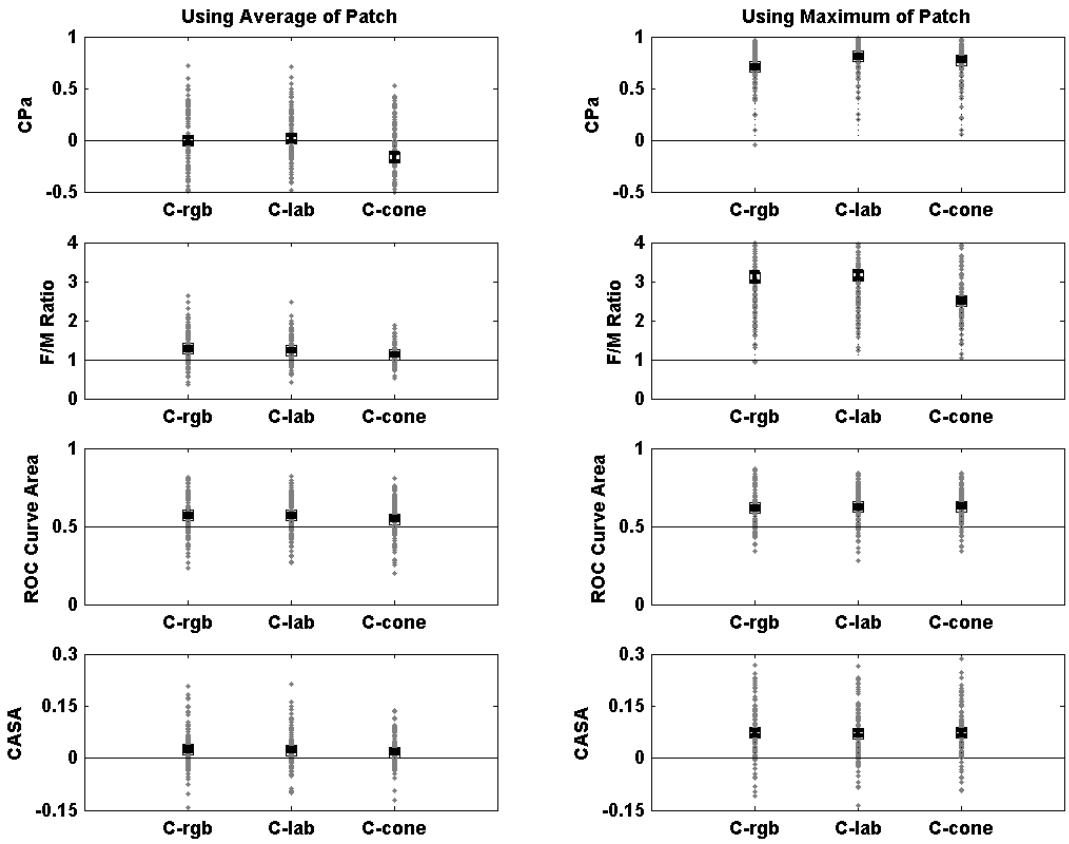


Figure 80: Comparison between the three types of Colorfulness maps using four performance metrics. The first column shows results when the average of the patch surrounding the fixation location is used; the second column shows the results when the maximum value is used. Each gray point represents the value for one image and one search target condition; the square represents the mean, and error bars represent one standard error of the mean.

Table V: Results of paired t-test between the performances of each map (when the average of each patch was used) A value of 1 indicates a significant difference between the means.

CPa

	C_lab	C_cone
C_rgb	0	1
C_lab		1

ROC Curve Area

	C_lab	C_cone
C_rgb	0	1
C_lab		1

F/M Ratio

	C_lab	C_cone
C_rgb	1	1
C_lab		1

CASA

	C_lab	C_cone
C_rgb	0	1
C_lab		1

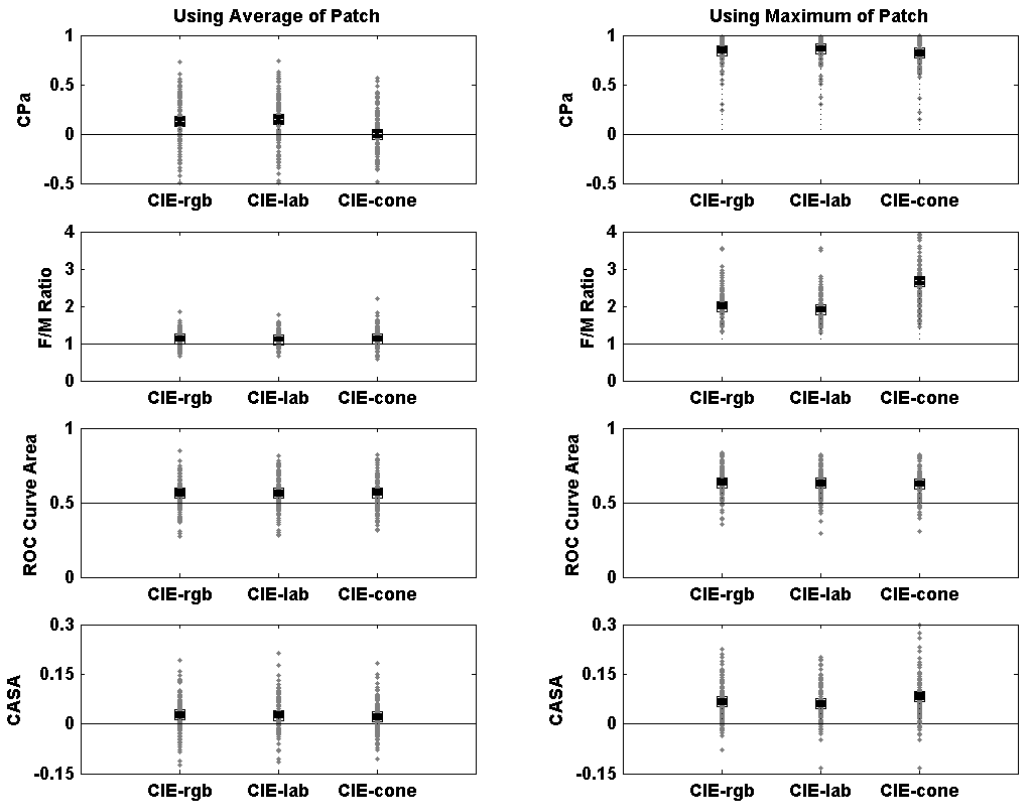


Figure 81: Comparison between the three types of Saliency (CIE) maps using four performance metrics. The first column shows results when the average of the patch surrounding the fixation location is used; the second column shows the results when the maximum value is used. Each gray point represents the value for one image and one search target condition; the square represents the mean, and error bars represent one standard error of the mean.

Table VI: Results of paired t-test between the performances of each map (when the average of each patch was used). A value of 1 indicates a significant difference between the means.

CPa

	CIE_lab	CIE_cone
CIE_rgb	1	1
CIE_lab		1

ROC Curve Area

	CIE_lab	CIE_cone
CIE_rgb	0	0
CIE_lab		0

F/M Ratio

	CIE_lab	CIE_cone
CIE_rgb	1	0
CIE_lab		0

CASA

	CIE_lab	CIE_cone
CIE_rgb	0	1
CIE_lab		0

5.5.3 Effect of weighting by fixation duration

For each image and each target condition, the CPa value was calculated by averaging the CPa values for each fixation. A second calculation was also performed, in which the CPa value for each fixation was weighted by its duration. That way, if a subject spent a long time looking at locations of high salience values, and fixated for shorter amounts on areas of lower values, then weighting by the duration will show an increase in the performance metric for that image. This weighting was also done for the F/M Ratio metric. Figure 82 shows histograms of the differences between the unweighted and weighted CPa and F/M Ratio metrics across all conditions (2), all maps (18), across all images (60), for a total of 2x18x60 = 2160 trials. This calculation was performed for the condition when the average value surrounding each fixation location was used (shown in the left column), as well as when the maximum was taken (right column). The histograms show that the mean difference in each case is very close to zero, but positive. A t-test was performed on each distribution at a significance level of 0.01 to test whether the mean is zero. For both CPa distributions, the hypothesis that the mean was zero was rejected ($p = 4.8\text{x}10^{-10}$ and $p = 0.0$). This was also the case for the F/M Ratio when using the average of each patch ($p = 2.9\text{x}10^{-13}$). For the F/M Ratio metric when using the maximum of each patch, the hypothesis that the mean is zero cannot be rejected due to the large standard deviation. The p-value for this test was 0.12. When using the average of each patch, there were some instances where weighting by the fixation duration increased (or decreased) the CPa value by more than 0.5 units, which can be very significant given that this difference is larger than the average CPa value for all images for

the CIE saliency maps (shown in Figure 81). When the maximum value of the patch is used, the differences in CPa values are less dramatic, as shown by the smaller standard deviation (0.1 compared to 0.18). For the F/M Ratio, increases larger than 1.5 units were found.

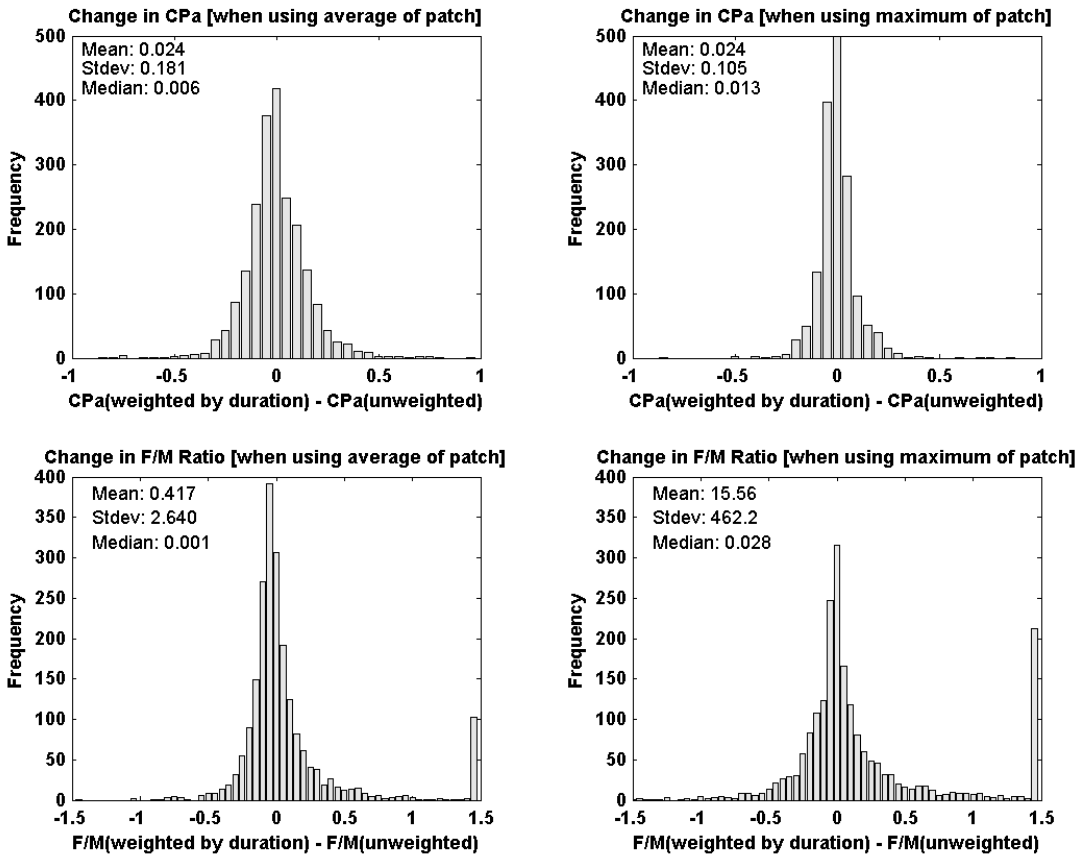


Figure 82: Histograms of the change in each metric (across 2160 trials) when fixation duration is included as a weight during the calculation. The left column shows the condition in which the average feature map value surrounding fixated locations is used, and the right column shows the condition in which the maximum feature map value is used.

5.5.4 Influence of target preview

For the following analyses, the feature content at fixated locations is compared between the Extracted Object and Cartoon Icon conditions. A paired t-test, at a significance level of α = 0.01, was used to determine if there was a significant difference in the amount of the particular feature at fixated locations between the target conditions. The means of the maps used, averaged across 60 images, are shown in Figure 83. The means for the four maps that used the histogram backprojection process are very low; these maps are essentially zero except where there are colors similar to those in the target.

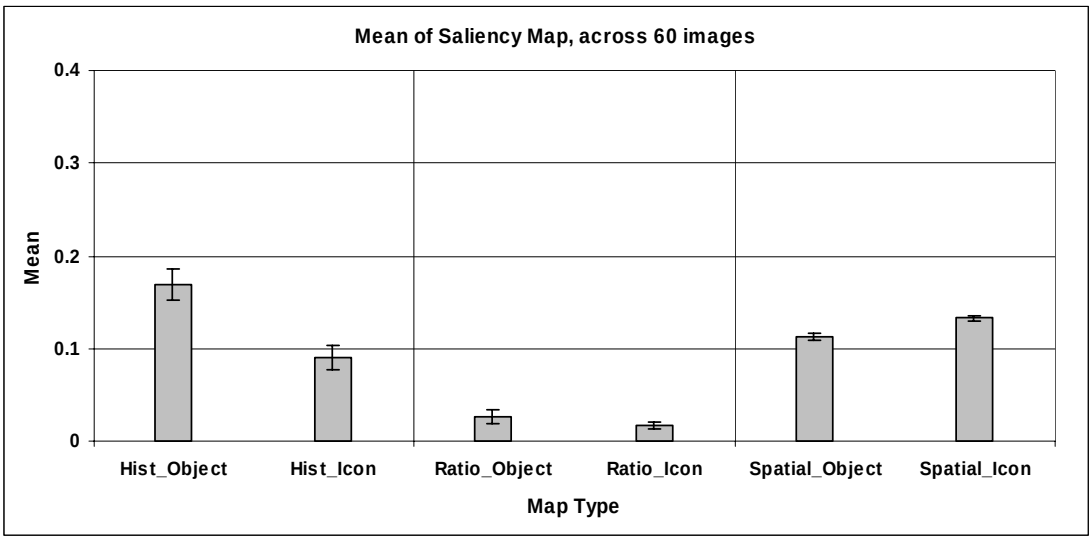


Figure 83: Means of saliency maps after scaling from 0 to 1, averaged across 60 images. Error bars represent standard error of the mean.

5.5.5 Influence of spatial structure of target preview

Table VII lists the average of each performance metric for the Spatial_object feature map. The values are given for both target preview conditions, along with the result of the paired t-test using a significance level α = 0.01. Also, the left table lists the results when the average value of the map surrounding each fixation is used; the right lists the results when the maximum surrounding each fixation is used.

For the Spatial_object map, if the search strategy was to fixate at locations that closely match the target preview spatially, it is expected that the metrics will show a high value in the Extracted Object condition compared to the Cartoon Icon condition. When using the average of

each patch, each metric shows the performance to be slightly above chance. Both the CPa and F/M Ratio metrics show high values in the Extracted Object (when using the maximum of each patch): 0.84 and 4.1, respectively. In addition, weighting by fixation duration produced an increase in the F/M Ratio metric by almost 1 unit. The F/M Ratio metric is the only metric that shows a significant difference between the two search conditions (when using the maximum of each patch), which would indicate that subjects fixated more on locations with a spatial structure that was similar to the true target after they had seen the Extracted Object target preview.

Table VIII shows the average of each of the performance metrics for the Spatial_icon feature map. When using the average of each patch, each of the metrics showed the performance to be slightly above chance. When using the maximum of each patch, all but the ROC Curve Area metric show an increase in performance. Although none of the metrics show a significant difference between the target preview conditions, most show that the average for fixations in the Cartoon Icon condition is slightly higher. This indicates that for some images, the spatial structure of the Cartoon Icon target preview influenced the search strategy.

Table VII: Performance metrics for the Spatial_object map, averaged across 60 images.

	Using Average			**Using Maximum**		
	Object	**Icon**	**Significant difference**	**Object**	**Icon**	**Significant difference**
CPa	0.08	0.08	No	0.84	0.83	No
F/M Ratio	1.25	1.24	No	4.09	3.95	**Yes**
ROC Curve Area	0.60	0.59	No	0.65	0.63	No
CASA	0.02	0.02	No	0.08	0.07	No
CPa weighed	0.14	0.14	No	0.88	0.87	No
F/M weighted	1.32	1.33	No	5.02	4.66	**Yes**

Table VIII: Performance metrics for the Spatial_icon map, averaged across 60 images.

	Using Average			**Using Maximum**		
	Object	**Icon**	**Significant difference**	**Object**	**Icon**	**Significant difference**
CPa	0.01	0.02	No	0.80	0.81	No
F/M Ratio	1.14	1.16	No	3.27	3.30	No
ROC Curve Area	0.56	0.57	No	0.58	0.59	No
CASA	0.01	0.01	No	0.04	0.05	No
CPa weighed	0.02	0.04	No	0.81	0.82	No
F/M weighted	1.13	1.16	No	3.27	3.32	No

5.5.6 Influence of color of target preview

Table IX shows the results for the Hist_object map, and Table XI lists the results for the Hist_icon map. When using the maximum of each patch, the CPa and F/M Ratio metrics show high values; however, when using the average of each patch, the CPa value is below chance. The CPa value is below zero in this case because of the typical shape of the Ratio Histogram maps. Figure 84, Figure 85 and Figure 86 show the behavior of the CPa and F/M Ratio metrics for this distribution. Because the map is mostly zero, many fixations will also have an average patch value near zero or below the low mean value. The CPa value for these fixations will be negative; in this example, a fixation with a salience value of zero corresponds to a CPa value of -1.35. Since there will be few fixations with an average patch value in the higher saliency ranges, the large negative CPa values will outweigh any positive values when the average CPa value is computed. Considering the F/M Ratio metric, a fixation with a salience value of zero corresponds to an F/M Ratio value of zero. Because this metric converges at zero, fixations at areas of very low salience will not outweigh those at higher salience values as much as in the CPa metric case.

The ROC Curve Area and CASA metrics for the Hist_object are each above chance. Again, weighting by the duration of each fixation produces an increase in both the CPa and F/M Ratio metrics. None of the metrics for either map show a significant difference between the two target preview conditions, indicating that the strategy used by the visual system during these search tasks is not significantly influenced, on average, by the color of the target preview. However, there are many sets of targets in which the Cartoon Icon contains colors similar to the Extracted Object. Eight images were selected in which the color of the Cartoon Icon is grayscale, while the Extracted Object is color. The average performance values for the Hist_object map are listed in Table X

For the Hist_icon map, the average F/M Ratio for the Extracted Object condition is 17.2, and 14.2 for the Cartoon Icon condition. These values are very large due to the fact that in some cases, if the Cartoon Icon contains colors that are not found in the image, the Hist_icon map has a mean very close to zero and a distribution similar to the one shown in Figure 84. For the Hist_icon map, all of the metrics show values near or below chance. Again, none of the metrics for either map show a significant difference between the two target preview conditions. When using the average of each patch and not weighting by fixation duration, each metric shows that

there is a significant difference in performance between the two conditions indicating that when shown a colored target preview, the visual system is more likely to fixate on regions of similar colors than in the condition when the target is grayscale.

Table IX: Performance metrics for the Hist_object map, averaged across 60 images.

	Using Average			**Using Maximum**		
	Object	**Icon**	**Significant difference**	**Object**	**Icon**	**Significant difference**
CPa	-0.13	-0.15	No	0.75	0.77	No
F/M Ratio	2.18	1.98	No	9.47	9.12	No
ROC Curve Area	0.62	0.59	No	0.63	0.62	No
CASA	0.06	0.05	No	0.06	0.05	No
CPa weighed	0.11	0.06	No	0.86	0.85	No
F/M weighted	3.33	3.10	No	11.78	11.47	No

Table X: Performance metrics for the Hist_object map, averaged across 8 images.

	Using Average			**Using Maximum**		
	Object	**Icon**	**Significant difference**	**Object**	**Icon**	**Significant difference**
CPa	0.13	0.02	**Yes**	0.89	0.87	**Yes**
F/M Ratio	1.98	1.52	**Yes**	2.11	2.08	No
ROC Curve Area	0.65	0.59	**Yes**	0.64	0.61	No
CASA	0.04	0.03	**Yes**	0.06	0.06	No
CPa weighed	0.30	0.18	No	0.91	0.88	No
F/M weighted	3.01	2.53	No	2.15	2.11	No

Table XI: Performance metrics for the Hist_icon map, averaged across 60 images.

	Using Average			**Using Maximum**		
	Object	**Icon**	**Significant difference**	**Object**	**Icon**	**Significant difference**
CPa	-0.49	-0.54	No	0.63	0.63	No
F/M Ratio	0.93	0.97	No	17.20	14.19	No
ROC Curve Area	0.47	0.46	No	0.56	0.55	No
CASA	-0.01	-0.01	No	-0.01	0.02	No
CPa weighed	-0.47	-0.53	No	0.65	0.66	No
F/M weighted	0.97	1.02	No	20.98	14.57	No

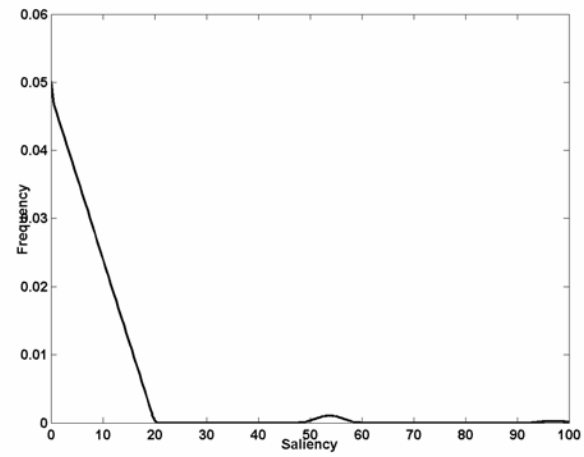


Figure 84: Example histogram of a map created using the histogram backprojection process.

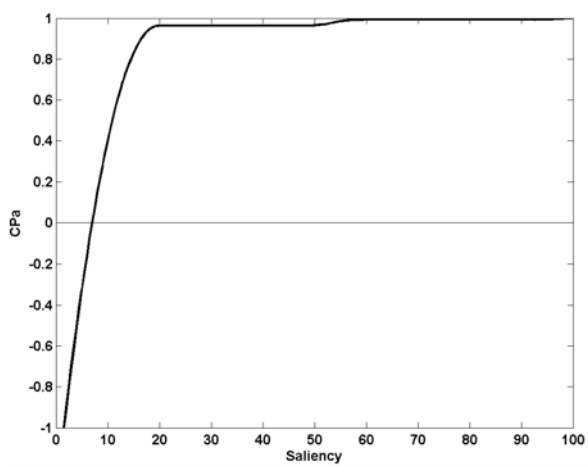


Figure 85: CPa values for any saliency value using the distribution in Figure 84

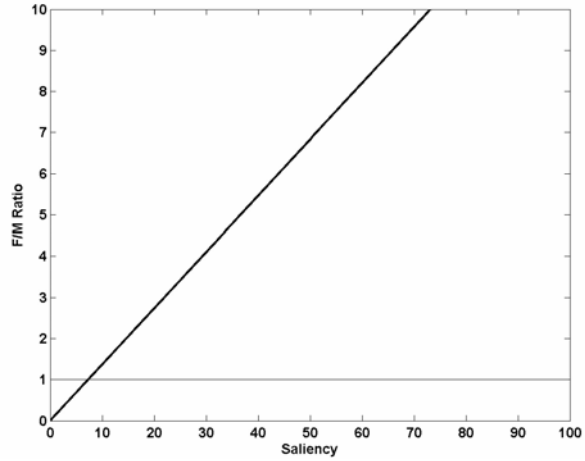


Figure 86: F/M Ratio values for any possible saliency value using the distribution in Figure 84

Table XII and Table XIII show the results for the Ratio_object and Ratio_icon maps, respectively. Again there are very high F/M Ratios resulting from the low mean value of the maps. For the Ratio_icon map, all metrics except the F/M Ratio indicate performance is at or below chance; they also show better performance by the Ratio_object map than by the Ratio_icon map. For each map, there was no significant difference in the performance between search conditions.

Table XII: Performance metrics for the Ratio_object map, averaged across 60 images.

	Using Average			**Using Maximum**		
	Object	**Icon**	**Significant difference**	**Object**	**Icon**	**Significant difference**
CPa	0.06	0.05	No	0.59	0.58	No
F/M Ratio	5.37	4.80	No	104.62	81.93	No
ROC Curve Area	0.61	0.61	No	0.63	0.62	No
CASA	0.01	0.02	No	0.05	0.04	No
CPa weighed	0.20	0.17	No	0.72	0.69	No
F/M weighted	10.59	9.67	No	184.37	132.71	No

Table XIII: Performance metrics for the Ratio_icon map, averaged across 60 images.

	Using Average			**Using Maximum**		
	Object	**Icon**	**Significant difference**	**Object**	**Icon**	**Significant difference**
CPa	-0.20	-0.18	No	0.34	0.34	No
F/M Ratio	2.59	2.50	No	546.00	425.86	No
ROC Curve Area	0.41	0.39	No	0.50	0.52	No
CASA	0.00	0.00	No	0.02	0.01	No
CPa weighed	-0.20	-0.19	No	0.37	0.38	No
F/M weighted	3.32	4.15	No	671.57	719.09	No

Next, the performance of the Ratio Histogram Backprojected maps method was compared to the normal Histogram Backprojected maps. Table XIV lists the differences in average performance of the Ratio_object and Hist_object maps for the Extracted Object condition. A positive value indicates that the performance of the Ratio_object map was greater than that of the Hist_object map. Similarly, Table XV lists the chance in performance between the Ratio_icon and Hist_icon maps. For the maps weighted by the color of the Extracted Object target, the CPa and F/M Ratio metrics show that the Ratio_object maps performed better when using the average value around fixation locations. However, when using the maximum value around each fixation location, the CPa metric shows that the Hist_object map performed better (as shown by the negative values of -0.16 and -0.19). This is also the case when the map is weighted by the color

of the Cartoon Icon; when taking the average value around the fixation location, the CPa metric shows better performance by the Ratio_icon map, but when the maximum value around the fixation location is used, the Hist_icon map shows better performance. The ROC Curve Area showed little difference between the maps, and the CASA values indicate better performance by the Hist_object map.

Table XIV: Change in performance when using the Ratio Histogram Backprojection versus normal Histogram Backprojection using the Extracted Object target.

	Using Average	
	Object	**Significant**
CPa	0.19	No
F/M Ratio	3.20	**Yes**
ROC Curve Area	-0.01	No
CASA	-0.05	**Yes**
CPa weighed	0.09	No
F/M weighted	7.26	**Yes**

Table XV: Change in performance when using the Ratio Histogram Backprojection versus normal Histogram Backprojection using the Cartoon Icon target.

	Using Average	
	Object	**Significant**
CPa	0.36	No
F/M Ratio	1.53	**Yes**
ROC Curve Area	0.01	No
CASA	0.01	No
CPa weighed	0.33	No
F/M weighted	3.13	No

5.5.7 Relationships between feature content and reaction time

In this section, the performance of 8 feature maps will be examined for three images. In the first example, Image 51, the median reaction time in the Cartoon Icon condition was 3.62 seconds, compared to 1.4 seconds in the extracted object condition. For Image 25, the Cartoon Icon target preview was a different color than the true target. For Image 53, nearly half of the subjects could not find the target, regardless of target preview and in some cases subjects took up to one minute to find the target. For each example, a chart shows the performance of each map in both the Extracted Object and Cartoon Icon conditions. The relative values of performance metrics were used to infer possible search strategies used in the different search conditions, which are listed in a table following each image. Also listed is the map that best predicted fixation locations according to each metric.

Figure 87: Image 51

Table XVI: Search strategies and map performance compared across metrics for Image 51

Image 51: Sunburst Object RT: 1.4 s Icon RT: 3.62 s

Metric	Best performance	Search strategy, according to metric
CPa	CIEP_cone	Object: Fixate on colors similar to target color, avoid similar shapes
		Icon: Fixate on salient areas, no influence from target color or spatial similarity
F/M Ratio	Ratio_object	Object: Fixate on colors similar to target color, avoid similar shapes
		Icon: Fixate on colors similar to cartoon icon
ROC Curve Area	CIEP_cone	Object: Fixate on colors similar to target color, avoid similar shapes
		Icon: Fixate on salient areas with similar spatial content
CASA	CIEP_cone	Object: Fixate on salient areas with colors similar to target color, avoid similar shapes
		Icon: Fixate on salient areas with similar spatial content

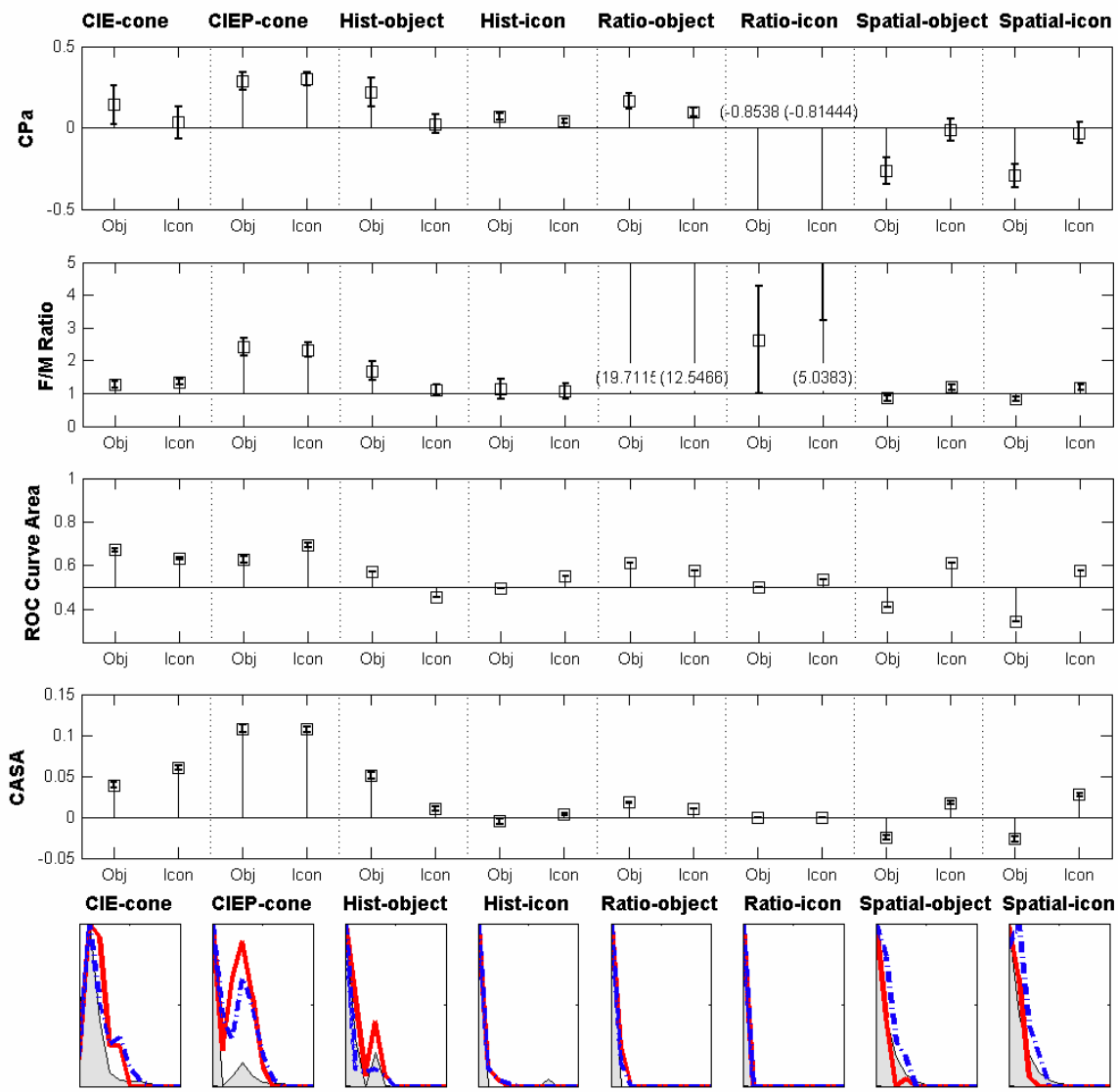


Figure 88: Performance of eight feature maps for Image 51. The last row shows the feature map histogram (shaded gray), the histogram of values at fixated locations during the Extracted Object search condition (red solid line) and Cartoon Icon condition (blue dotted line).

Figure 89: Image 25

Table XVII: Search strategies and map performance compared across metrics for Image 25

Image 25: Blue moon Object RT: 2.17 s Icon RT: 2.48 s

Metric	Best performance	Search strategy, according to metric
CPa	CIEP_cone (conspicuity)	Object: Fixate on salient areas that contain blue and are spatially similar to the target; avoid yellow
		Icon: Fixate on salient areas that are yellow, not blue, that are spatially similar to cartoon icon
F/M Ratio	Ratio_object	Object: Fixate on the colors that are unique to the target; suppress target colors that are heavily present in the rest of the image; avoid yellow
		Icon: Fixate on salient regions that are yellow
ROC Curve Area	Hist_object, Spatial_object	Object: Fixate on blue regions that are spatially similar to the target; avoid yellow
		Icon: Fixate on salient, yellow regions
CASA	CIEP_cone (conspicuity)	Object: Fixate on salient, blue regions that are spatially similar to the target
		Icon: Fixate on salient, yellow regions

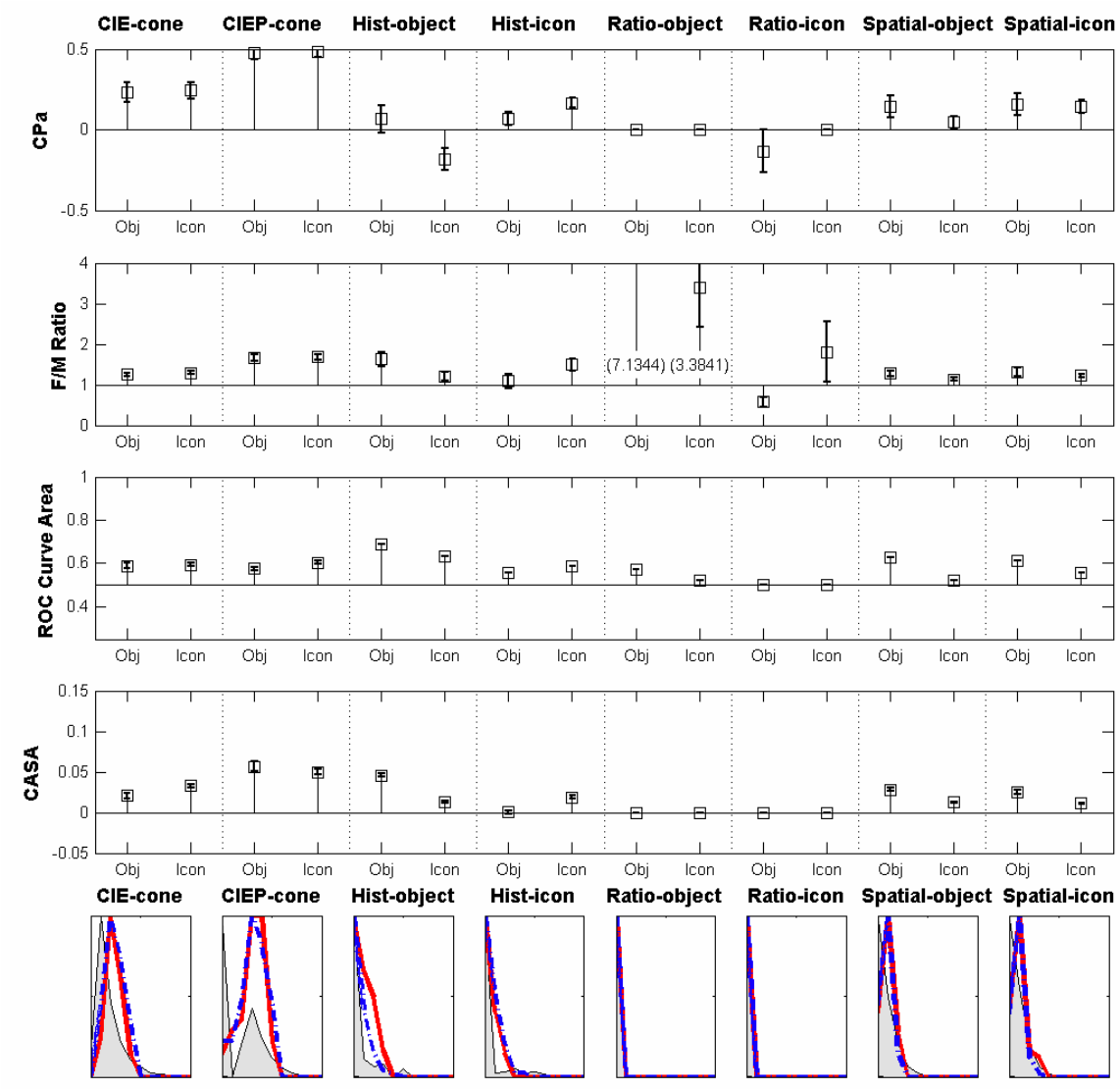


Figure 90: Performance of eight feature maps for Image 25. The last row shows the feature map histogram (shaded gray), the histogram of values at fixated locations during the Extracted Object search condition (red solid line) and Cartoon Icon condition (blue dotted line).

Figure 91: Image 53

Table XVIII: Search strategies and map performance compared across metrics for Image 53

Image 53: Teapot Object RT: 22 s Icon RT: 22 s ; (not found by some subjects)

Metric	Best performance	Search strategy, according to metric
CPa	CIEP_cone, Hist_icon	Object: Fixate on salient regions that are possible objects that are also gray or gold, avoid spatially similar objects
		Icon: Fixate on salient regions that are possibly objects that are gray or gold
F/M Ratio	CIEP_cone	Object: None – search randomly or in a serial pattern throughout the image
		Icon: None – search randomly or in a serial pattern throughout the image
ROC Curve Area	CIEP_cone, Spatial_icon and object	Object: None – search randomly or in a serial pattern throughout the image
		Icon: Fixate on salient regions that are possibly objects that are similar in spatial structure to the icon
CASA	CIEP_cone	Object: None – search randomly or in a serial pattern throughout the image
		Icon: Fixate on salient regions that are possibly objects

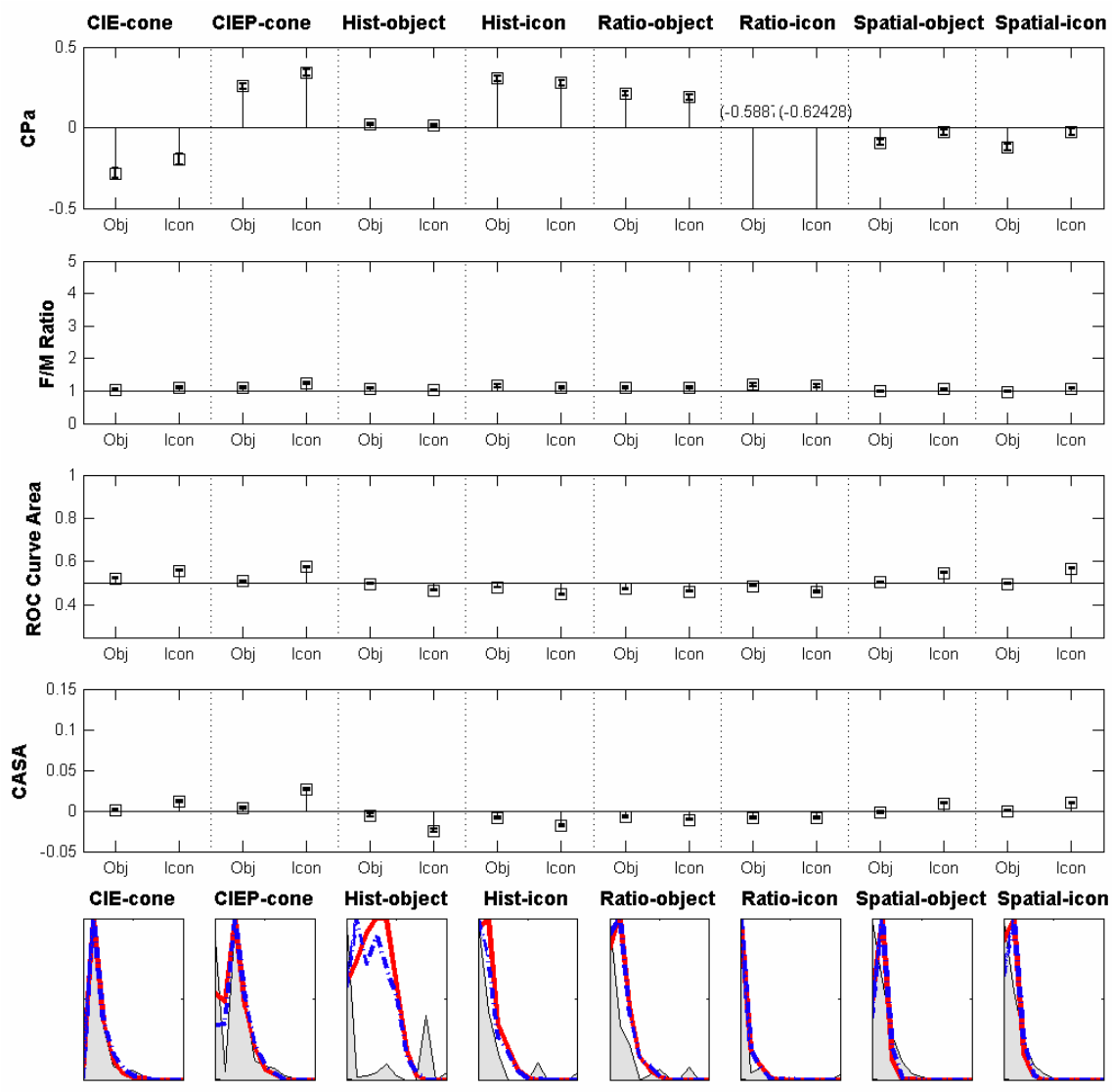


Figure 92: Performance of eight feature maps for Image 53. The last row shows the feature map histogram (shaded gray), the histogram of values at fixated locations during the Extracted Object search condition (red solid line) and Cartoon Icon condition (blue dotted line).

5.5.8 Relative performance of maps in each condition

Table XIX shows the performance of each of the maps side-by-side, for the condition in which the average of each patch was used. The top section lists the results from the Extracted Object target preview condition, and the bottom lists results from the Cartoon Icon condition. Values that are at or below chance level are shaded. The highest value for each metric is shown in bold. For the Extracted Object condition, the Hist_icon and Ratio_icon performed below chance according to almost all metrics. This is expected since subjects in this condition did not see the Cartoon Icon target preview. According to the CPa metric, the Spatial_object map performed best, indicating that the content selected by the visual system during this task was more similar to the target in shape than in color. The F/M Ratio, on the other hand, showed better performance of the maps weighted by the target's color, rather than its spatial structure. The ROC Curve Area and CASA metrics both show that the Hist_object map performed best, although the ROC Curve Area was nearly equal for the Hist_object, Ratio_object, and Spatial_object maps.

For the Cartoon Icon condition, most of the metrics show the performance of the maps weighed by the icon's colors to be below chance, suggesting that subjects did not consistently use the Cartoon Icon's color in a search strategy as was seen in the Extracted Object condition. This strategy would not be efficient given that the subject could not trust that the Cartoon Icon would be the same color as the true target. Again the CPa metric showed best performance by the Spatial_object map, followed by the Ratio_object map. The F/M Ratio metric value was very high for the Ratio_object map, which is also the map that produced the highest ROC Curve Area. Again the ROC Curve Area was nearly equal for the Hist_object, Ratio_object, and Spatial_object maps. Unlike the other metrics, the CASA value was highest for the Hist_object map. These results show that subjects' fixations, on average, correlate to regions with colors that are similar to the true target, even though they had been shown a Cartoon Icon before the trial.

Although Table X, Figure 88 and Figure 90 show that there are differences in map performance between the two search conditions for some images, averaging across all images shows no overall difference.

Table XIX: Relative performance of maps for the two target preview conditions. Shading indicates values at or below chance level.

Extracted Object Condition

	Hist_object	Hist_icon	Ratio_object	Ratio_icon	Spatial_object	Spatial_icon
CPa	-0.13	-0.49	0.06	-0.20	**0.08**	0.01
F/M Ratio	2.18	0.93	**5.37**	2.59	1.25	1.14
ROC Curve Area	**0.62**	0.47	0.61	0.41	0.60	0.56
CASA	**0.06**	-0.01	0.01	0.00	0.02	0.01
CPa weighted	0.11	-0.47	**0.20**	-0.20	0.14	0.02
F/M Ratio weighted	3.33	0.97	**10.59**	3.32	1.32	1.13

Cartoon Icon Condition

	Hist_object	Hist_icon	Ratio_object	Ratio_icon	Spatial_object	Spatial_icon
CPa	-0.15	-0.54	0.05	-0.18	**0.08**	0.02
F/M Ratio	1.98	0.97	**4.80**	2.50	1.24	1.16
ROC Curve Area	0.59	0.46	**0.61**	0.39	0.59	0.56
CASA	**0.05**	-0.01	0.02	0.00	0.02	0.01
CPa weighted	0.06	-0.53	**0.17**	-0.19	0.14	0.04
F/M Ratio weighted	3.10	1.02	**9.67**	4.15	1.33	1.16

5.5.9 Relative performance of all maps

Figure 93 compares the average performance of each map collapsed across the two search conditions. Each of the four metrics is shown for the condition in which the average of each patch was used. The CPa value shows the best performance by the P_object and CIEP_cone maps. This indicates that during the search task, subjects fixated on areas that were perceptually salient and also in regions that are likely to be objects or parts of objects. Performance was above chance for the Ratio_object and Spatial_object maps, suggesting that areas of similar color or spatial structure to the target were fixated.

The F/M Ratio metric shows very high performance for the two Ratio Histogram Backprojected maps. These values may be an overestimation resulting from the very low mean of these types of maps. The P_object, CIEP_cone and Hist_object were also above chance, suggesting that salience, texture, and colors similar to the target guided saccadic eye movements in this search task.

The ROC Curve Area metric shows that the Edge, P_object, Hist_object, Ratio_object, Spatial_object performed nearly equally, and better than the other maps. This suggests that areas of colors and spatial structure that were similar to the target were fixated more than areas of high intensity or bright colors. The maps weighted by the color of the Cartoon Icon showed a performance below chance.

The CASA metric shows the best performance by the P_object map. The CIEP_cone and Hist_object maps show better performance than the rest of the maps, suggesting that areas of color similar to the target, or that are likely to be objects, received more fixations than areas of bright colors or many edges.

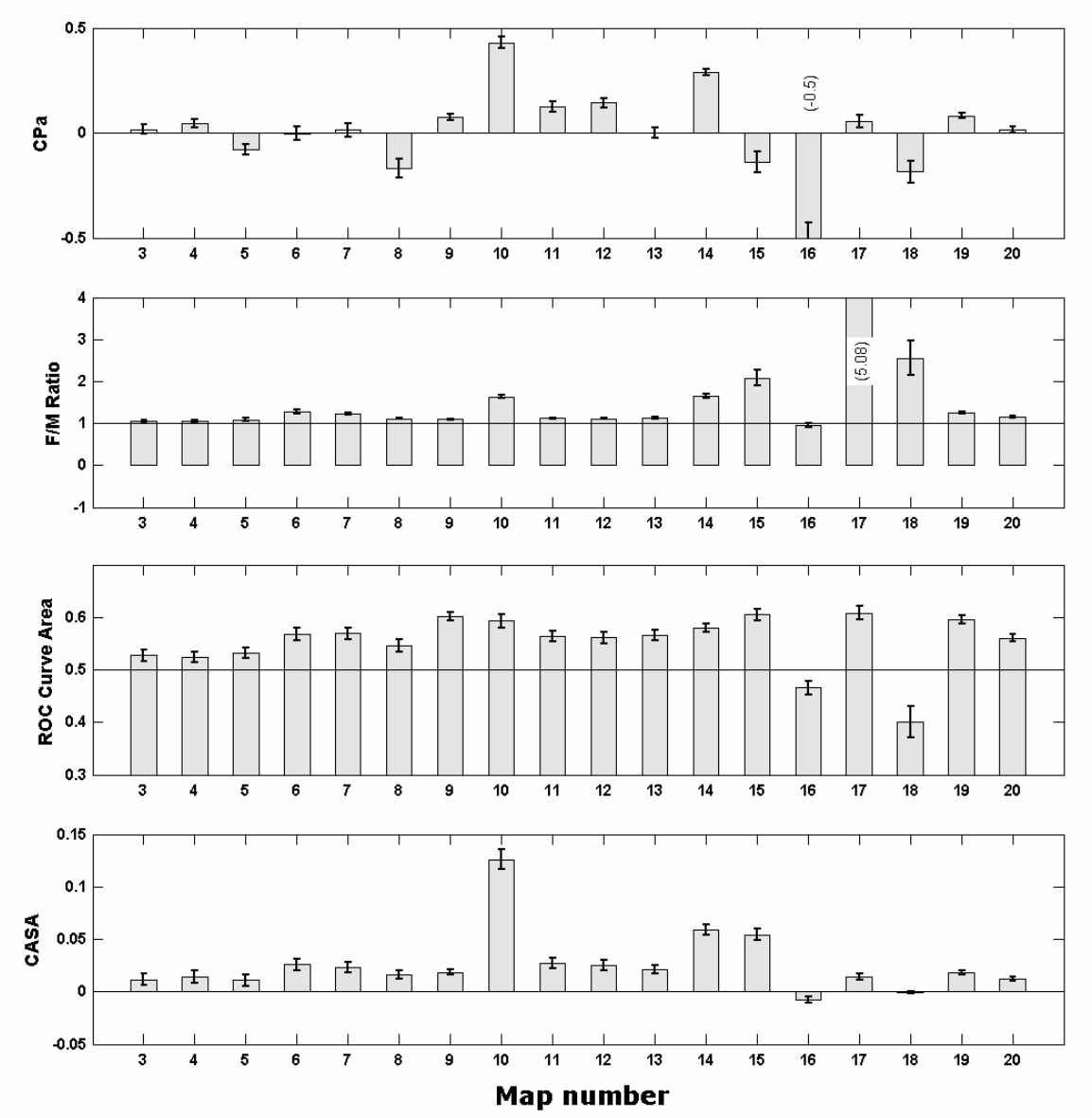


Figure 93: Performance metric value (averaged across all images and both search conditions) for each map: 3.) I_rgb, 4.) I_lab, 5.) I_cone, 6.) C_rgb, 7.) C_lab, 8.) C_cone, 9.) Edges, 10.) P_object, 11.) CIE_rgb, 12.) CIE_lab, 13.) CIE_cone, 14.) CIEP_cone, 15.) Hist_object, 16.) Hist_icon, 17.) Ratio_object, 18.) Ratio_icon, 19.) Spatial_object, 20.) Spatial_icon

5.6 Discrimination images

Research by Rajashekar [2002, 2004] has shown evidence that the visual system uses a form of spatial matched filtering when searching Gaussian noise. Using the eyetracking data from this experiment, patches of an intensity image (defined as the L* channel) at fixated locations during visual search in real-world scenes were extracted to determine if an underlying spatial structure may attract attention. Also, if subjects did use a spatial structure to guide fixations, are there any differences between the two search conditions?

Figure 94: Image 24 and corresponding Extracted Object and Cartoon Icon target previews

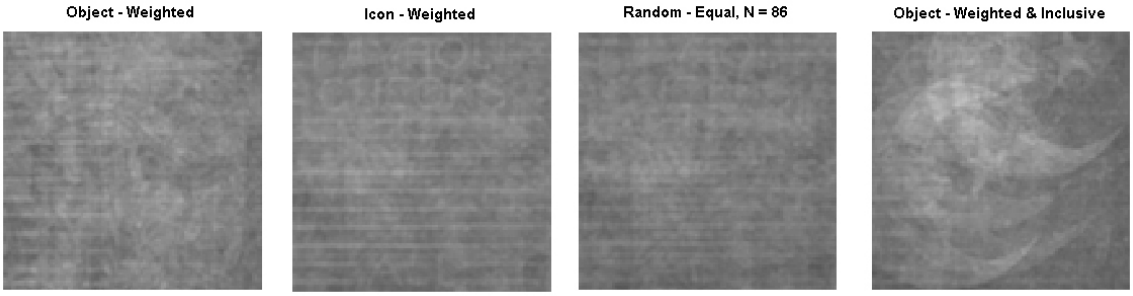


Figure 95: Example discrimination images for Image 24

Figure 94 shows Image 24 used in the experiment, along with its Extracted Object (a blue moon with a face) and Cartoon Icon (yellow crescent shape) target previews. Figure 95 shows the resulting average patches for each condition, and for random locations. The final image in Figure 95 includes patches at fixations on or near the target. The averaging process took into account the duration of the fixation at that location. For both the Object and Icon conditions, a horizontal striped pattern emerges.

Distinct spatial structures that resemble the target, such as those shown by Rajashekar, et al. [2002] were not found in this analysis for several reasons. For each image, there could be as few as 30 fixation locations. This number of samples is not large enough, and some subjects' locations may overlap. In natural images, the 'noise' may not average out as in Rajashekar's experiments. When searching images with large uniform regions, such as a sky, any region with objects may receive more fixations. In that case, the discrimination image may resemble those objects, since there was a high probability that they would be fixated. The horizontal striped patterns in Figure 95 do not indicate that when searching for a moon, subjects' search strategy is to look for horizontal stripes; rather, they are present in the average patches because the image contains many horizontal lines in the neon signs, bricks, and building structure.

Even if more samples were available, it is possible that a structure would not be present in the discrimination image because the visual system is altering its search 'kernel' while acquiring high-level information as time progresses. For example, when searching a kitchen scene for a water faucet, high-level information about the context of the scene will guide eye movements toward likely locations for that object.

Additionally, as mentioned by Rajashekar, et al. [2002], these structures may be "washed out" if the image patches do not overlap perfectly. For example, fixations may occur on different areas of the same target. When patches around those fixations are averaged, the surrounding structure is replicated and may form 'ghost' images, as shown in the last image in Figure 95. Both subject variability and inaccuracy of the eyetracker used will facilitate this problem.

5.7 Conclusions and Discussion

This chapter presented a visual search experiment in which subjects searched for an object in a real-world scene. There were two search conditions: the subject was either shown the exact, pixel-for-pixel target before beginning the trial (Extracted Object condition), or a cartoon icon representation of the target (Cartoon Icon condition). Results showed an increase in reaction time for the Cartoon Icon condition; this was found to be due to both an increase in the amount of time between the presentation of the image and the first fixation on the target and a slight increase in time between fixating on the target and pressing the spacebar to end the trial. This indicates that knowing the exact target features expedites the visual search task.

A series of topographical feature maps were generated in order to investigate what features in the scene were used to guide saccadic eye movements in the visual search task. These maps were also used to compare differences between the content selected by the visual system between the Extracted Object and Cartoon Icon search conditions. Four performance metrics used in current literature were used to measure how well high values of the feature maps correspond to locations of fixations. The performance values of each map can then be used to compare the relative amounts of each feature that was selected by the visual system.

Circular patches of a radius of ~1 degree were extracted from each map at each fixation location. In one condition, the average value of the patch was used to compute performance metrics, in another, the maximum value within each patch was used. Two of the metrics, CPa and F/M Ratio, are very sensitive to which condition was used in the calculation. The ROC Curve was less sensitive since it compares fixated values to values of random locations, and not just to the mean of the entire map.

Intensity, Colorfulness and Saliency (a combination of intensity, colorfulness, and oriented edge content) maps were computed using three different color spaces: RGB, 1976 CIE L*a*b*, and a color space that mimics the rod and cone responses of the visual system. After comparison, it is unclear which color space may be the most effective for building saliency maps since the results from the four performance metrics were not consistent. For the Intensity maps, the CPa metric showed increased performance for the I_lab map, whereas the F/M Ratio and ROC Curve Area metrics showed better performance by the I_cone map. The CASA metric showed no difference between the three. For the Colorfulness maps, three of the four metrics showed no difference between the C_rgb and C_lab maps, but all four showed lower performance

for the C_cone map. For the Saliency maps, the CPa value showed better performance by the CIE_lab map, whereas the F/M Ratio metric showed better performance by the CIE_rgb map. The ROC Curve Area metric showed no difference between the three.

The CPa and F/M Ratio metrics were also computed in a way that takes into account the duration of each fixation. That way, if a subject spends a long time fixating on high feature values, and fixates a shorter time on areas of low values, then the weighting by the duration will show an increase in performance. A paired t-test between 2160 trials showed that there is a positive increase in both the CPa and F/M Ratio metrics (when using the average of each patch) as a result of weighting by duration. This increase was also found when the maximum of each patch was used for the CPa metric. For the F/M Ratio in this condition, there were over 200 trials in which the ratio increased by more than 1.5 units.

Although an increase in map performance may be gained by weighting by fixation duration, it is unclear as to whether this is an appropriate way to gain insight into the extent to which stimulus features guide visual attention. Given that the decision to move the eyes to a particular location is made during previous fixations, it is unclear whether the duration of the fixation is correlated to the influence of those features when they are in the periphery.

The influence of the target preview on the selection of features was investigated using feature maps that were weighed by the target's color and spatial structure. Although the metrics do not agree with regard to the relative performance of the six maps used, most agreed that maps weighed by the color of the Extracted Object target performed best. The CPa metric showed best performance by the map weighted by the spatial structure of the Extracted object for both search conditions. The ROC Curve Area was very similar between the maps weighted by color as well as the one weighted by the structure of the target, indicating these features, on average, had equal influence on search strategies. Some metrics showed better performance by the Ratio_object map than for the Hist_object map, which would indicate that if the image contains a large amount of colors similar to the ones in the target, signals in the visual system from those features would be suppressed, and regions of colors that are unique to the target would be searched for instead. This method of suppression when distracters are present was also shown in a simple visual search experiment by Navalpakkam [2004];

After averaging over all images, there was no significant difference between the performances of each map between the Extracted Object and Cartoon Icon search conditions, suggesting that the characteristics of the target did not, on average, influence the deployment of visual attention during the search task. However, there were target pairs that did not differ much

in color or spatial structure. As shown in Table X, Figure 88 and Figure 90, there exist some cases in which the target preview did influence the search strategy. For example, in some cases subjects looked at regions similar to the target in a particular feature dimension; in others, they may have avoided regions of similar features. When averaging across all images, these differences average to zero. This shows that the behavior of the visual system cannot be generalized by any one model or weighting method; the behavior is highly dependent on the task and scene.

The better performance of maps that are manipulated on a task-by-task basis again demonstrates the ability of the visual system to fine tune its routines actively. As shown in Table X, more fixations fell on regions of colors that were similar to the target when the subjects had viewed the Extracted Object instead of a grayscale Cartoon Icon. This indicates that when given that extra information about the color of the target the visual system used it to optimize its search strategy. This means that prior knowledge about the target also affects the deployment of attention. For the cases where the system performed in a manner opposite to what was predicted (e.g., avoiding features that were similar to the target), this indicates that other, high-level features (such as expected location within the context of the scene) are interacting with the low-level stimulus features to more efficiently guide eye movements. Additionally, the experiment in which 'discrimination images' were created showed that the visual system, during a search task in real-world scenes, does not simply use a spatial matched filter.

When comparing the performance of 18 feature maps, collapsed across the two search conditions, it is difficult to determine the relative amounts of features selected by the visual system because the four metrics do not always agree. The P_object and CIEP_cone maps were shown to perform above chance by each of the metrics. Three of the four metrics show that the Hist_object and Ratio_object maps, which are weighted by the color of the Extracted Object target color, were also well above chance. For the maps weighted by the color of the Cartoon Icon, all but the F/M Ratio are at or below chance. The high value of the F/M Ratio is an overestimation as a result of the very low mean value of the Ratio_icon map, which in some cases is very close to zero if the image does not contain much of the color of the Cartoon Icon.

The maps that are typically used to measure the influence of low-level features on selective attention include those that measure colorfulness, intensity, oriented edges. They are also combined to make a general "CIE" saliency map. Each of these maps, on average, performed very close to chance level. Only the ROC Curve metric showed the color and edge maps to perform as well as some of the weighted maps.

The goal of the visual search experiment and extraction of features described in this chapter was to determine what types of features are used to guide selective attention during a search task. A second goal was to determine how the target preview affected search strategies, as defined by the relative amounts of features at fixated locations. While the selected features used to make topographical maps is not an exhaustive list, it was found that the locations of fixations correlate to areas of the image that are similar in color or spatial structure to the target, rather than areas with bright colors, high intensity, or strong edges. Additionally, fixations correlate to regions that are possible objects or foreground areas, as defined by the P_object map. The relationship between the features used to guide eye movements depend not only on which performance metric is used, but also the feature content of the target and the scene.

The performance metrics used were shown to behave differently with respect to the distribution of the feature map, and the manner in which they are calculated. Using the maximum value of a feature map within a one-degree radius of a fixation location is not appropriate for the CPa and F/M Ratio metrics. Although the CPa and F/M Ratio metrics both in some way compare the map value at fixation to the mean map value, the shape of the distribution may cause the metrics to disagree. In cases in which the map is mostly zero, with a few locations of high values, the CPa value may be very negative (below chance) for map values just below the mean, while the F/M Ratio gives a value only slightly below chance. On the other hand, if the mean is close to zero, a high map value will give a very large F/M Ratio, whereas the CPa metric will converge to 1.

The ROC Curve Area and CASA metrics are less sensitive to whether the maximum or average map value surrounding a fixation location is used, since they are designed to examine the difference between the distributions of map values selected by vision or selected by random. The CASA value also has bounding values, but the number of bins used to create histograms will affect the weighting vector, which will affect the final CASA value. Additionally, two different distributions of fixated map values may result in very similar CASA values. The CPa has an upper limit, whereas the F/M Ratio has a lower limit. The ROC Curve Area is also ideal because it is limited in range; however, the result may be dependent on the number of fixations used since the calculation requires a curve to be fit.

The ideal metric depends upon how saliency maps are designed. If they are designed to consist of mostly low values with a limited selection of high values, then the F/M Ratio and CPa values may not be appropriate. The ROC Curve Area metric is the most conservative among the four metrics used and works well for distributions of different shapes.

Chapter 6

6 Conclusions and Recommendations

6.1 Overview

The goals of this thesis project included utilizing eyetracking technology to gain insight into what scene content guides eye movements in specific tasks. Chapter 2 gave an overview of the human visual system and its behavior as an active, flexible imaging system. Also presented was an overview of current work on computational models that seek to explore the bottom-up influence of visual stimuli on attention, as well as metrics designed to measure that influence. Chapter 3 presented the tools used to perform the experiments performed during this research project.

6.2 Influence of task on eye movements

Chapter 4 introduced the work of Alfred Yarbus and his classical experiment that showed the influence of task on eye movement patterns [Yarbus, 1967]. Yarbus eyetracked only one subject. His experiment was replicated with 17 subjects under more natural conditions in this work. Subjects viewed a copy of I.E. Repin's "They did not expect him" only for as long as needed. Additionally, subjects' heads were not constrained. Unlike Yarbus' experiment, temporal information was also collected.

Results similar to Yarbus' were found; the eye fixated on 'informative' regions for each task. The amount of time spent viewing different regions of the image was compared between the tasks. It was shown that for each of the tasks, the faces were fixated. For some tasks, subjects

spent the majority of the time examining the people in the scene; in others, objects such as furniture or pictures on the wall were examined more.

Subjects showed variability not only in the amount of time they spent performing each task, but also in the percentage of time spent on each region of the scene. However, differences in behavior between tasks were greater than differences between subjects (within each task), indicating that the influence of the task was greater than any individual differences.

Replication of this experiment verified Yarbus' often-cited results showing that eye movement patterns are highly influenced by a person's goal or task, even as the visual stimulus remains constant.

Chapter 5 presented a visual search experiment in which subjects searched for an object in a real-world scene. Two target preview conditions were used: the subject was either shown the exact, pixel-for-pixel target before beginning the trial (Extracted Object condition), or a cartoon icon representation of the target (Cartoon Icon condition). Results showed an increase in reaction time for the Cartoon Icon condition; this was found to be primarily due to an increased amount of time between the presentation of the image and the first fixation on the target, and not an increase in time between fixating on the target and pressing the spacebar to end the trial.

The difference in reaction time between the two conditions showed that the features of the target preview influenced the pattern of eye movements during the search, which resulted in a longer reaction time. This also indicates that knowing the exact features of the target facilitated the performance of the visual system in this task.

6.3 Image features at the point of gaze

A series of topographical feature maps were generated in order to investigate what features in the scene were used to guide saccadic eye movements in the visual search task. These maps were also used to compare differences between the content selected by the visual system between the Extracted Object and Cartoon Icon search conditions.

It was found that areas of the scene that contained low-level image features such as bright colors, high intensity, and high-contrast oriented edges did not correlate well to the locations that were fixated by subjects during the experiment. Rather, regions that are possible objects or foreground elements showed a better correlation. Additionally, regions of the image that were

similar to the search target either in color or spatial content showed a strong correlation to where people looked.

When comparing the features at the point of gaze between the two search conditions, little difference was found when results were averaged across the image set. This indicates that on average, the target preview does not affect the content selected by the visual system (i.e., the search strategy) in the same manner across all trials. However, in some cases it was found that the content at fixated locations closely matched features of the target only in the Extracted Object condition. In other cases, features similar to the target were avoided because the image contained a great deal of those features.

A few conclusions can be drawn from the results of this experiment. First, the behavior of the visual system cannot be explained only by bottom-up responses to low-level information from the environment. Secondly, the regions of the scene that are task-relevant or 'informative' areas are fixated and given more access to computational resources in the brain. These informative regions may contain similar low-level features as those of the target, but they are also constrained by high-level information such as the context of the scene, recognized objects, segmented foreground, or semantics. Third, the instruction, whether verbal or visual, can affect the deployment of attention and therefore patterns of eye movements. However, the effect that a visual target preview has on search strategies is not consistent across targets or real-world scenes. Lastly, the visual system is not simply a mathematical pattern recognition system that uses color and spatial features of the target preview to operate in a visual search task. In real-world scenes, more complex features are extracted and used to guide eye movements, such as expected location within the context of the scene.

6.4 Recommendations and future work

6.4.1 Temporal analysis

A logical progression of this research is temporal analysis of eye movements and the deployment of visual attention. In current literature, there is conflicting evidence regarding the influence of low-level features in guiding eye movements, and how that influence changes over time. It would be interesting to know if in the Cartoon Icon condition, the features of the icon are initially used to target areas to be fixated before being disregarded or suppressed as new strategies evolve as higher-level information about the scene is accumulated.

It is possible that in some cases in which the target is not in an expected location, the system, upon getting the "gist" of the scene within a few hundred milliseconds [Biederman, 1981], first chooses areas where the object is likely to be found. When it is not found, the search strategy may then evolve and become more influenced by low-level features such as target color or spatial structure.

6.4.2 Comparison to Freeview

Eyetracking data should be collected for subjects freely viewing the images used in the visual search experiment. Comparison of the performance of the different types of feature maps may indicate differences in oculomotor behavior between visual search and freeview tasks.

6.4.3 Feature sets and weightings

In order to further explore the role of target color or spatial structure, new Cartoon Icon target previews should be designed. This may include a set that are monochrome versions of the Extracted Object. In the experiment presented above, there were sets of target previews that did not differ much in color or shape. This may reduce the differences seen between conditions. When designing new targets or scenes, it is important to not control every variable to the extent that the search task becomes unlike what people do in the real world. The types of scenes chosen for this experiment mimic what people may do on a daily basis: search for keys on a table, an item in the grocery store, or a face in a crowd.

The features explored in this research project were very limited in scope. Additional sets of features should be explored. For example, the spatial correlation maps were computed at only one spatial resolution. The spatial features of the target may also be influential at other spatial scales. Also, this project did not investigate the relative weighting of combinations of features. For example, when creating the CIE saliency maps, should intensity, color and oriented edge maps be given equal weighting?

The feature maps used in this project were not dynamically modified to take into account any change in feature detection as a result of the region's location in the periphery during the fixation(s) prior. More work could be done to model how these different types of features are represented in the periphery in order to make models of attention more biologically plausible.

6.4.4 Central biasing

A topic that was only briefly mentioned is central biasing. Some models of saliency incorporate an artificial weighting that inhibits regions that are not in the central portion of the visual field (or image). It has been shown that both locations of fixation and locations of high salience gravitate toward the center of the field/image (see Figure 75). What is not clear is whether the locations of fixations are a result of salient regions being in the center of the image or of a central bias of the oculomotor system. An experiment to explore this question would involve one in which subjects viewed images of natural or real-world scenes that do not contain objects or focal points in the center, as they typically are in photographs used in eyetracking experiments.

6.4.5 Performance metrics

The performance metrics used in this analysis each have a unique behavior that depends on the probability distributions of the feature maps. Choosing the most appropriate performance metric begs the question of how predictive models of visual attention should be designed. For example, feature maps could be designed so that they are mostly zero with selected locations of high values. They may also be designed to follow a specific distribution that is tailored to work well with a specific metric. When comparing one feature across conditions or to random, the ordinal value of the feature (such as contrast or the spatial correlation) should be used. When comparing across features, a performance metrics that takes into account the shape of the distribution is more appropriate.

Another consideration is the process of calculating each metric. In this project, circular patches with a radius of approximately 1 degree of visual angle were extracted, and either the average or maximum of each patch was used. The effect of different patch sizes and image resolution on each metric should be explored.

Two metrics in this project took into account the duration of each fixation. A slight increase in performance of feature maps was found, although it is not clear if the fixation duration is correlated with the influence of the feature when it is in the periphery. An analysis which uses the duration of one or more fixations prior could be compared.

6.4.6 Improved models of visual attention

The results of the experiments in this research project can be used as guidelines for the development of new models of visual attention. At this point it is impossible to create a general predictive model which builds upon only bottom-up information while excluding any high-level information or influence from a person's experience or interests. However, research of what scene content is extracted by this active system may shed light onto what types of processes are being used by the brain. In this project, it was verified that adding a mask that inhibits responses from salient regions in the background improves performance [Canosa, 2003]. During the visual search task, color and spatial information were shown to play a role in guiding eye movements for certain scenes. For a search task in which target features are known *a priori*, incorporating both color and spatial information about the target, perhaps via a weighted mask, may improve the performance of the model.

References

Andrews, T.J., and Coppola, D.M. (1999). Idiosyncratic characteristics of saccadic eye movements when viewing different visual environments. *Vision Research*, 39, pp. 2947-2953.

ArtSmart at NextMonet, discussion on Composition. Critique on 'Nine Spaces' by Ron Kroutel. Accessed September 15, 2003 http://www.nextmonet.com/artsmart/CB101.jhtml

Babcock, J., Lipps, M. & Pelz, J.B. (2002) How people look at pictures before, during, and after image capture: Buswell revisited. *Human Vision and Electronic Imaging VII*, Rogowitz & Pappas (Ed), Proceedings of SPIE Vol. 4662, pp. 34-47.

Babcock, J.B. (2002). Eye tracking observers during color image evaluation tasks. Master's Thesis. New York: Rochester Institute of Technology.

Babcock, J.S., Pelz, J.B., & Fairchild, M.D. (2003), Eye tracking observers during color image evaluation tasks, *Human Vision and Electronic Imaging VIII*, Rogowitz & Pappas (Ed), Proceedings of SPIE Vol. 5007

Biederman, I. (1981). On the semantics of a glance at a scene. In M.Kubovy and J.R. Pomerantz (Eds.), *Perceptual Organization* (pp. 213–253). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Brandt, H.F. (1945). *The Psychology of Seeing*. New York: Philosophical Library.

Buswell, G.T. (1935). *How people look at pictures: A study of the psychology of perception in art.* Chicago: Univ. Chicago Press.

Canosa, R.L. (2000). Eye movements and natural tasks in an extended environment, Master's Thesis. New York: Rochester Institute of Technology.

Canosa, R.L. (2003). Seeing, sensing, and selection: Modeling visual perception in complex environments. Doctoral Dissertation. New York: Rochester Institute of Technology.

Collewijn, H., Steinman, R.M., Erkelens, C.J., Pizlo, Z., van der Steen, J. (1992). Effect of freeing the head on eye movement characteristics during three dimensional shifts of gaze and tracking. In Berthoz, A., Graf, W., Vidal., P.P. (Eds.), *The Head-Neck Sensory Motor System* (Chapter 64). Oxford University Press.

Carmi, R., Itti, L. (2004). Disentangling bottom-up from top-down influences on attentional allocation in dynamic scenes. Vision Science Society Annual Meeting (VSS04).

Hancock, P. J. B., Bradley, R. J., and Smith, L. S. (1992).The principal components of natural images. *Network* , 3, 61-70.

Henderson, J.M. & Hollingworth, A. (1998). Eye movements during scene viewing: an overview. In G. Underwood (Ed.), *Eye Guidance in Reading and Scene Perception* (269-293). New York: Elsevier.

Itti, L., Koch, C., and Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20 (11), 1254-1259.

Itti, L., and Koch, C. (2000). A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research*, 40, 1489-1506.

Koch, C., and Ullman, S. (1985). Shifts in selective visual attention: Towards the underlying neural circuitry. *Human Neurobiology*, 4, 219-227.

Krieger, G., Rentschler, I., Hauske, G., Schill, K, and Zetzsche, C. (2000). Object and scene analysis by saccadic eye-movements: An investigation with higher-order statistics. *Spatial Vision*, 13 (2-3), 201-214.

Kowler, E., Pizlo, E., Zhu, G., Erkelens, C.J., Steinmann, R.M., Collewijn, H. (1992).Coordination of head and eye during the performance of natural (and unnatural) visual tasks. In Berthoz, A., Graf, W., Vidal., P.P. (Eds.), *The Head-Neck Sensory Motor System* (Chapter 65). Oxford University Press.

Land, M.F., Mennie, N., Rusted, J. (1999). The roles of vision and eye movements in the control of activities of daily living. *Perception*, 28, 1311-1328.

Mackworth N.H., & Morandi, A.J. (1967). The gaze selects informative details within pictures. *Perception and Psychophysics*, 2, 547-551.

Mannan, S., Ruddock, K., Wooding, D. (1996). The relationship between the location of spatial features and those of fixations made during visual examination of briefly presented images. *Spatial Vision*, 10, No. 3, 165-188.

Navalpakkam, V., Rebesco, J., Itti, L. (2004). Modeling the influence of knowledge of the target and distractors on visual search. Vision Science Society Annual Meeting (VSS04).

Noton, D. and Stark, L. (1971). Scanpaths in saccadic eye movements while viewing and recognizing patterns. *Vision Research*, 11, 929-942.

Osberger, W. & Maeder, A.J. (1998). Automatic Identification of Perceptually Important Regions in an Image. Proc. 14th Int. Conf. on Pattern Recognition, Brisbane, Australia, 701-704.

Palmer, S.E. (1999). *Vision Science: Photons to Phenomenology*. Cambridge, MA: MIT Press.

Parkhurst, D., Law, K., and Niebur, E. (2002). Modeling the role of salience in the overt and covert shifts of visual attention. *Vision Reseach*, 42, 107-123.

Parkhurst, D., and Niebur, E. (2003). Scene content selected by active vision. *Spatial Vision*, 16 (2),125-154.

Pelz, J.B. and Canosa, R., (2001). Oculomotor Behavior and Perceptual Strategies in Complex Tasks, *Vision Research*, 41, 3587-3596.

Privitera C., and Stark L. (2000). Algorithms for Defining Visual Region-of-Interest: Comparison with Eye Fixations. IEEE Transactions on Pattern Analysis and Machine Intelligence. 22 (9), 970-982.

Rajashekar, U., Cormack, L., and Bovik, A. (2002). Visual Search: Structure from Noise. Proceedings of the ACM SIGCHI Eye Tracking Research & Applications Symposium, New Orleans, LA.

Rajashekar, U., Cormack, L., and Bovik, A. (2004). Point-of-gaze analysis reveals visual search strategies Proceedings of SPIE, Human Vision and Electronic Imaging IX. 5292, 296-306.

Rao, R., Zelinsky, G., Hayhoe, M., and Ballard, D. (1996). Modeling saccadic targeting in visual search. *Advances in Neural Information Processing Systems*, Cambridge, MA: MIT Press.

Rao, R., Zelinsky, G., Hayhoe, M., and Ballard, D. (1997). Eye Movements in Visual Cognition: A Computational Study. National Resource Laboratory for the Study of Brain and Behavior. Dept. of Computer Science, University of Rochester, NY.

Reinagel, P. and Zador, A.M. (1999). Natural scene statistics at the center of gaze. *Network: Computation in Neural Systems*, 10, 341-350.

Skavenski, A. et al. (1979). Quality of retinal image stabilization during small natural and artificial body rotations in man. *Vision Research*, 19.

Steinman, R.M., Kowler, E., and Collewijn, H. (1990). New directions for oculomotor research. *Vision Research*, 30 (11), 1845-1864.

Swain, M.J., Kahn, R.E., and Ballard, D.H. (1992). Low resolution cues for guiding saccadic eye movements. Proc., Conference on Computer Vision and Pattern Recognition. Urbana, IL.

Tatler, B., Baddeley, R., and Gilchrist, I. (2004) Visual correlates of fixation selection: Effects of scale and time. *Vision Research* (submitted).

Thompson, Daniel V., Jr. (1936). *The Practice of Tempera Painting*. New Haven: Yale University Press. 132-134.

Treisman, A. and Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology*, 12 (1), pp. 97-136.

Walther, D., L. Itti, M. Riesenhuber, T. Poggio, and C. Koch. (2002). Attentional Selection for Object Recognition – A Gentle Way. In: *Biologically Motivated Computer Vision*, Second International Workshop (BMCV 2002), Tübingen, Germany, 472-479.

Wandell, B.A. (1995). *Foundations of Vision*. Sunderland, MA: Sinauer.

Yarbus, A.L. (1961). Eye movements during examination of complex objects. *Biofizika*, 6.

Yarbus, A.L. (1967). *Eye Movements and Vision* (B. Haigh, Trans.). New York: Plenum Press. (Original work published in 1956).

Appendices

Images


Extracted Object


Cartoon Icon


Extracted Object


Cartoon Icon


Extracted Object


Cartoon Icon


Extracted Object


Cartoon Icon


Extracted Object


Cartoon Icon

Extracted Object

Cartoon Icon

Extracted Object

Cartoon Icon

Extracted Object

Cartoon Icon

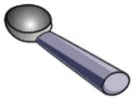
Extracted Object

Cartoon Icon

Extracted Object

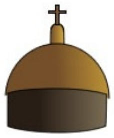
Cartoon Icon

Extracted Object

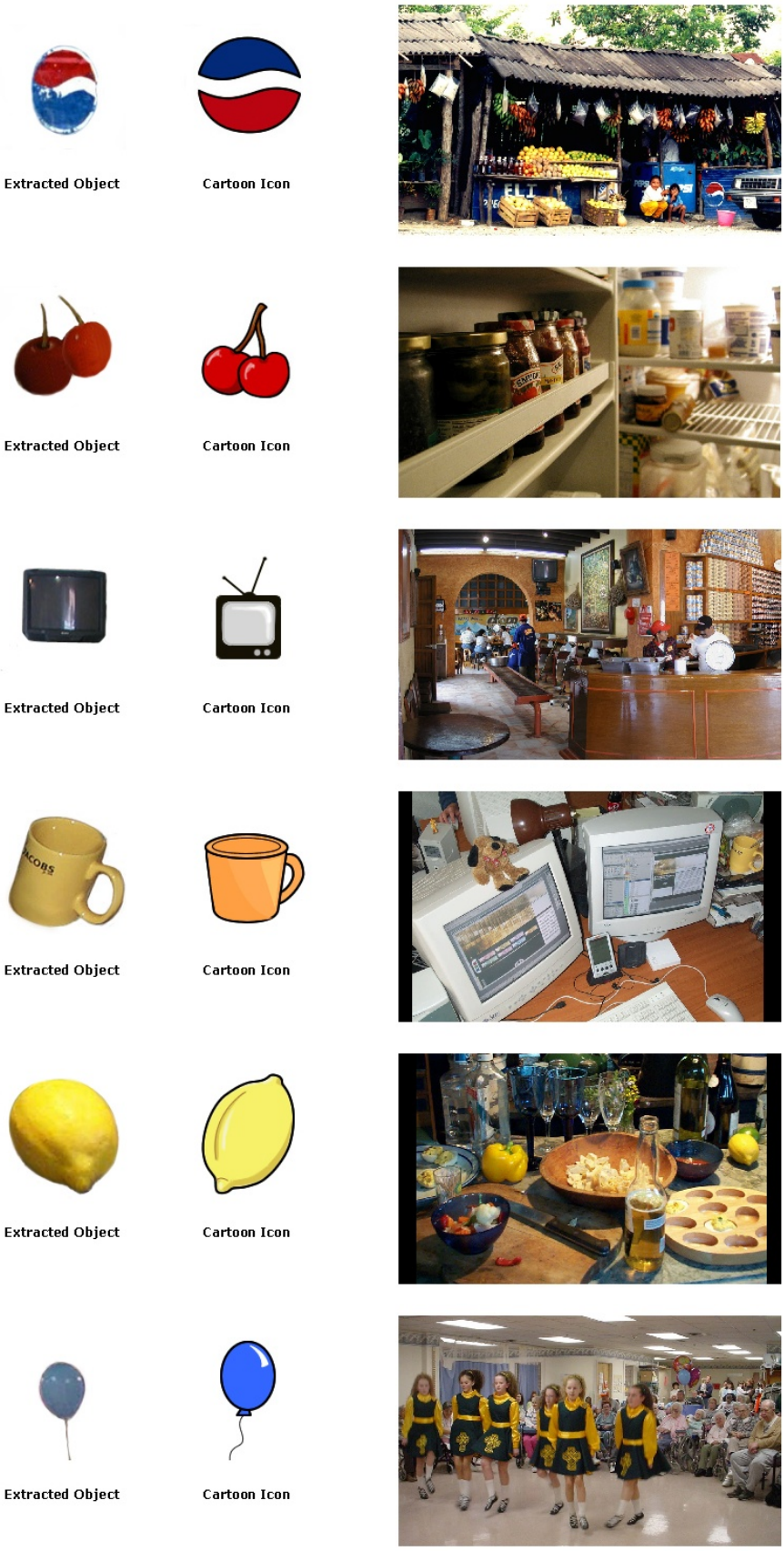


Cartoon Icon

Extracted Object



Cartoon Icon


Extracted Object
Cartoon Icon
Extracted Object
Cartoon Icon
Extracted Object
Cartoon Icon
Extracted Object
Cartoon Icon
Extracted Object
Cartoon Icon
Extracted Object
Cartoon Icon

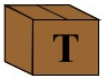
Extracted Object

Cartoon Icon

Extracted Object

Cartoon Icon

Extracted Object



Cartoon Icon

Extracted Object

Cartoon Icon

Extracted Object

Cartoon Icon

Extracted Object

Cartoon Icon

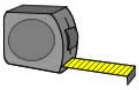
Extracted Object

Cartoon Icon

Extracted Object

Cartoon Icon

Extracted Object



Cartoon Icon

Extracted Object

Cartoon Icon

Extracted Object

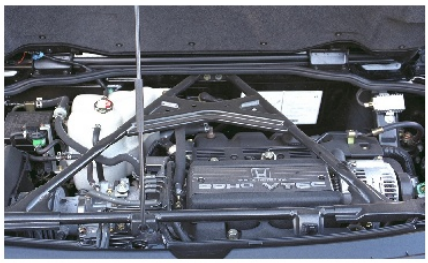
Cartoon Icon

Extracted Object

Cartoon Icon

Extracted Object

Cartoon Icon



Extracted Object

Cartoon Icon

Extracted Object

Cartoon Icon

Extracted Object

Cartoon Icon

Extracted Object

Cartoon Icon

Extracted Object

Cartoon Icon

Extracted Object

Cartoon Icon

Extracted Object

Cartoon Icon

Extracted Object

Lily

Cartoon Icon

Extracted Object

Cartoon Icon

Extracted Object

K

Cartoon Icon

Extracted Object

Cartoon Icon

Extracted Object

Cartoon Icon

Extracted Object

Cartoon Icon

Extracted Object

Cartoon Icon

Extracted Object

ADAM

Cartoon Icon

Extracted Object

Cartoon Icon

Extracted Object

la Ferté

Cartoon Icon

Extracted Object

Cartoon Icon

Extracted Object

Coca

Cartoon Icon

Extracted Object

Cartoon Icon

Extracted Object

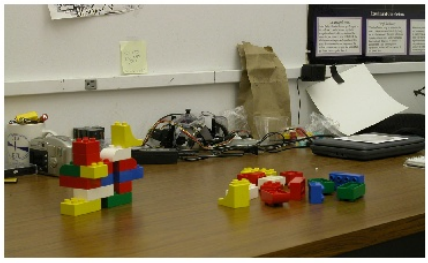
Cartoon Icon

Extracted Object

Cartoon Icon

Extracted Object

Cartoon Icon



Extracted Object

Cartoon Icon

Extracted Object

Cartoon Icon

Extracted Object

Cartoon Icon

Extracted Object

Cartoon Icon

Extracted Object

Cartoon Icon

Extracted Object

Cartoon Icon