How people look at pictures before, during, and after scene capture: Buswell revisited

Jason S. Babcock*[*], Marianne Lipps, and Jeff B. Pelz

Visual Perception Laboratory, Chester F. Carlson Center for Imaging Science
Rochester Institute of Technology, Rochester, NY, USA 14623

ABSTRACT

A wearable eye tracker was used to record photographers' eye movements while they took digital photographs of *person, sculpture,* and *interior* scenes. Eye movement sequences were also recorded as the participants selected and cropped their images on a computer. Preliminary analysis revealed that during image capture people spend approximately the same amount of time looking at the camera regardless of the scene being photographed. The time spent looking at either the *primary object* or the *surround* differed significantly across the three scenes. Results from the editing phase support previous reports that observers fixate on semantic-rich regions in the image, which, in this task, were important in the final cropping decision. However, the spread of fixations, edit time, and number of crop windows did not differ significantly across the three image classes. This suggests that, unlike image capture, the cropping task was highly regular and less influenced by image content.

Keywords: Eye tracking, visual perception, looking at pictures, image editing, fixation maps

1. INTRODUCTION

The fact that some people take better photographs than others do is reason enough to study how people perform this task. To gain further insight into the picture-taking process it is necessary to see what the photographer sees and to study what he/she pays attention to. An important question to ask is what exactly does the photographer look at while composing the image through the camera's viewfinder? Further, what does the photographer look at in the *real* scene just before image capture, and how might these regions of interest compare to those made by the photographer when editing these images on a computer screen?

In 1935, Guy T. Buswell[1], a professor of educational psychology at the University of Chicago, published a book entitled, *How People Look at Pictures: A Study of The Psychology of Perception in Art.* The work in this book is important to the history of eye tracking because it is the first thorough investigation to record and analyze the eye movements of subjects while they looked at complex scenes. In this investigation Buswell eye tracked over 200 subjects while they viewed 55 photographs of objects ranging from paintings and statuary pieces, to tapestries, patterns, architecture, and interior design. Unlike the eye tracking technology used in early reading studies, the apparatus used in Buswell's experiment was built specifically for the purpose of recording both the horizontal and vertical eye position during the course of image viewing. Buswell recognized that he was applying state of the art technology in order to gain new insight into what people did when they looked at pictures. In his introduction he states, "As is generally the case when a technique is first applied in a new field, this study possesses many of the characteristics of a survey experiment rather than one which tests

[*] Correspondence: jsb1623@cis.rit.edu

carefully formulated hypotheses. The writer is in a much better position to set up such hypotheses now than at the beginning of the study (p 17)[1]."

In many ways the work published here parallels the work of Buswell in 1935. Specifically, a portable eye tracking system has been developed by the Visual Perception Laboratory at Rochester Institute of Technology. This system enables us to record the eye movements of subjects while they perform realistic tasks outside of the laboratory. Unlike the limitations imposed by the eye tracking device used by Buswell, it is now possible to extend his question of how people look at pictures to how people take pictures. Further, with digital photography and image editing software available, it is now possible to study subjects' eye movements while they edit their pictures. With this in mind, the purpose of this paper is to present exploratory results from an experiment that recorded photographers' eye movements while they captured and edited a series of digital images. In order to best understand the motivation for this research, we begin with a discussion on visual acuity and eye movements.

2. BACKGROUND

2.1 Eye Movements and Visual Perception

Evolution has equipped humans with a clever imaging system. Unlike a uniform CCD sensor in a digital camera, the eye's retina is composed of two types of sensors called rods and cones. An important distinction between these sensors is their visual sensitivity. At intermediate levels of illumination, both rod and cone vision is active. As luminance levels decrease, rod sensitivity increases. Conversely, high luminance levels effectively saturate the rods so that only the cone photoreceptors are functioning. In the periphery of the retina, the rods greatly outnumber the cone photoreceptors. The large rod distribution is useful for seeing in low illumination conditions. However, visual acuity in the periphery is quite poor. At the center of the eye the cone photoreceptors are distributed in the region of the retina referred to as the fovea. Here, high-resolution cone photoreceptors are packed tightly together near the optical axis. From the center outward, the distribution of cones substantially decreases past one degree of visual angle. Unlike the rods, each cone photoreceptor in the fovea reports information in a nearly direct path to the visual cortex. In this region of the brain, the fovea occupies approximately five times more neural tissue than the rods. For the most part, the spatial mapping of the retina (cone-center, rod-periphery) is analogous to the spatial mapping of the visual cortex. Given these characteristics, the fovea is responsible for high-resolution vision. Since the oculomotor system allows us to orient our eyes to areas of interest very rapidly, and with little effort, most of us are completely unaware that spatial acuity is not uniform across the visual field.

The temporal nature of eye movements (at a macro-level) can be described as a combination of fixations and saccades. Fixations occur when the eye has paused on a particular spatial location in the scene. To re-orient the fovea to other locations, the eyes make rapid angular rotations called saccades. On average, a person will execute over 150,000 eye movements each day[2]. This active combination of head and eye positioning (referred to as gaze changes) provides us with a satisfactory illusion of high resolution vision, continuous in time and space. When performing everyday tasks, the point of gaze is often shifted toward task-relevant targets even when high spatial resolution from the fovea is not required. Since these 'attentional' eye movements are made without conscious intervention, monitoring them often provides us with a window into cognition[3,4]. While eye movements do not expose the full cognitive processes underlying perception, most of the time they do provide an indication of where attention is deployed.

2.2 Eye Movements and Picture Viewing

While Buswell was unable to quantify differences in eye movements between trained and untrained artists, he did conclude that observers exhibited two forms of eye movement behavior. In some cases, viewing sequences were characterized by a general survey of the image, where a succession of brief pauses was distributed over the main features of the photograph. In other cases, observers made long fixations over smaller sub-regions of the image. In general, no two observers exhibited exactly the same viewing behavior. However, people were inclined to make quick, global fixations early, transitioning to longer fixations (and smaller saccades) as viewing time increased.

When observers' fixation patterns were plotted collectively for the same image, areas with a higher density of fixations often corresponded to information-rich regions in the image. This result indicated that observers often fixated on the same spatial locations in an image, but not necessarily in the same temporal order. Further, these consistencies revealed that people did not randomly explore pictures. Instead, the eye tended to focus on foreground elements like faces and people rather than background elements like clouds or foliage. While the above results were reported for free viewing situations, Buswell also concluded that the "mental set" obtained from experimental instructions (or reading a paragraph of text about the picture beforehand) significantly influenced how people looked at pictures (p 136)[1].

A decade later, Brandt[5] published a general analysis of eye movement patterns of people looking at advertisements. His study also investigated the role of eye movements in learning strategies, as well as in the perception of art and aesthetics. Like Buswell, Brandt concluded that there were individual differences in eye movements, but in general, these behaviors were similar enough that certain "psychological laws" could be formulated (p 205).

Yarbus[6] later reported that eye movements were not merely visual reflexes tied to the physical features of an image. Instead he believed that the eyes were directed to areas in the image that were "useful or essential" to perception (p 175). In his well-known example, Yarbus recorded the eye movements of subjects while they examined I.E. Repin's, *An Unexpected Visito*r. During free-viewing, eye movement patterns across seven subjects revealed similar areas of interest. However, different instructions, such as estimating the material circumstances of the family, or remembering the clothes worn by the people, substantially changed the eye movement patterns for the person viewing the painting. In general, the most informative regions were likely to receive more fixations.

Since observers generally directed their attention to the same regions in an image, several authors[7,8,9,10] set out to explore how the semantic information in a scene influenced eye movements behavior. Noton and Stark[11,12] analyzed the sequential nature of fixations in an attempt to identify recurring sequences of saccades they termed *scan paths*. In most of these experiments participants viewed black-and-white line-drawings or monochrome-shaded drawings of realistic scenes (in Antes[8], subjects viewed two color photographs- a mask and coastline). The general conclusion was that eye movements were not random, and that many of the fixations across observers did land on informative regions in the picture. Further, while there was variability across subjects, individuals often made the same scan paths to specific regions in the image. It was less clear, however, exactly how the semantic information affected fixation position. Henderson and Hollingworth[13] point out that experimental parameters such as image size, viewing time, and image content make it difficult to compare eye movement results across the above mentioned studies.

In studying the effect of aesthetic judgments in picture viewing, Molnar[14] had nine fine-art students view eight classical pictures ranging from Rembrandt to Chirico. Half of the fine art students were instructed to view the pictures carefully, as they would later be questioned about what they saw. These individuals were designated as the semantic group. He told the other half of the fine art students that they would be asked about the aesthetic qualities of the pictures (labeling them as the aesthetic group). Measures of fixation duration indicated that the aesthetic group made longer fixations than the semantic group. However, there was little difference in the magnitude of saccades between the two groups. The longer fixations for the aesthetic group suggest that more time was needed to make aesthetic judgments about the pictures, but that aesthetic judgments did not influence the angular distance between fixations. Similarly, Nodine, Locher, and Krupinski[15] later found that the composition of the image did influence how trained artists looked at paintings. In this experiment artists' fixation durations were longer, and their eye movement patterns had a tendency to focus on structural relationships between objects and backgrounds. For untrained viewers, fixation durations were shorter, and eye movement patterns focused mainly on pictorial elements that best conveyed objective reality.

The series of experiments started by Buswell in 1935 have focused on the perceptual and cognitive significance of eye movements relating to photographs, line drawings, and artwork already *captured* by others. While these experiments have demonstrated that observers tend to deploy their attention to similar regions in an image, they have not been able to study the kinds of eye movements that occur *before* and *during* image capture. The focus of this research is to connect what we know about still picture viewing (after image capture) to the oculomotor behaviors that occur before and during image capture.

2. METHODS

The research described here relies on the ability to monitor the eye movements of photographers during several phases of the photographic process, from scanning the scene before the photograph is captured to editing the digital image on a computer. Special equipment is necessary to complete each component of the task. The first phase of the process is for the photographer to view an object or scene and decide what to photograph – the *exploratory* phase. The second phase is when the photographer frames the photograph with the camera – the *framing phase*. The final phase requires the photographer to crop the digital image on a computer screen – the *editing phase*.

To ensure that the experimental results are not unduly affected by the instrumentation necessary to track the photographer's eye movements, special care must be taken to minimize constraints on the photographer during each phase of the photographic process so that the photographer can behave as naturally as possible. Most traditional eyetracking instrumentation requires the observer's head to be held stable by some form of head restraint. Even systems that are designed to allow free head movements require a tether between the observer and the eyetracker control instrumentation. In order to monitor the visual behavior of the photographer during the exploratory phase, we made use of a custom-made eyetracker developed at Rochester Institute of Technology. To allow natural behavior during the editing phase (while collecting data about the photographer's gaze position on the image during image editing) it was necessary to track the photographer's head movements in addition to their eye movements.

2.1 Eye Tracking Instrumentation

2.1.1 Wearable Eye Tracking System

A wearable eye tracking system was used to record subject's eye movements while they walked through the building taking photographs. The primary component of the tracker is a pair of modified racquetball goggles as shown in Figure 1. The far left side of the goggles supports an optics module housing an infrared illuminator, miniature CMOS video camera (sensitive only to IR), and a beam splitter (used to align the camera so that it is coaxial with the illumination beam). An external first-surface mirror folds the optical path toward an infrared reflective mirror shown next to the nose bridge of the goggles. This mirror simultaneously directs IR illumination toward the pupil and reflects an image of the eye back to the video camera. When aligned properly, the illumination beam enters the pupil, retro-reflects off the retina, and back-illuminates the eye. The same phenomenon causes 'red eye' in photographs when a camera's flash is aimed at the subject's line of sight.

A second miniature CMOS camera is mounted just above the subject's right eye to record the scene from the subject's perspective. This provides a frame of reference to superimpose a pair of crosshairs that corresponds to the subject's point of gaze. Just above the scene camera is a small semiconductor laser and a two-dimensional diffraction grating. The laser/diffraction grating system projects a grid of points that are used to calibrate the subject's eye movements relative to the video image of the scene. Since the laser is attached to the goggles, the calibration plane is fixed with respect to the head. The laser system greatly increases the accuracy of calibration and minimizes subject calibration time.

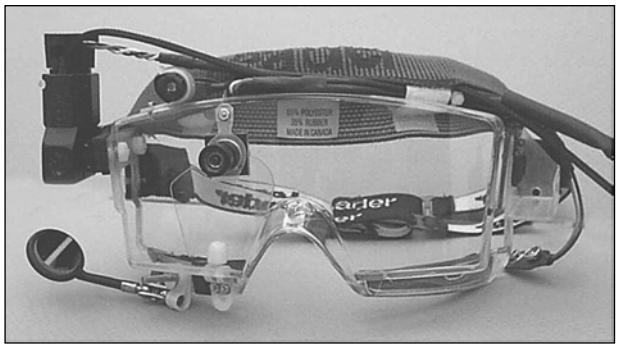


Figure 1 Custom-built headgear for the wearable eye tracker.

The backpack (Figure 2) carries a customized Applied Science Laboratory (ASL) 501 control unit. Eye and scene video-out from the ASL control unit is piped through a picture-in-picture box so that the eye image is superimposed onto the scene image. The combined video image is then recorded onto a digital video camcorder as shown in Figure 3

Figure 2 Wearable eye tracking backpack containing a digital video camcorder, picture-in-picture, ASL control unit, and batteries. The LCD is visible through the backpack to facilitate setup.

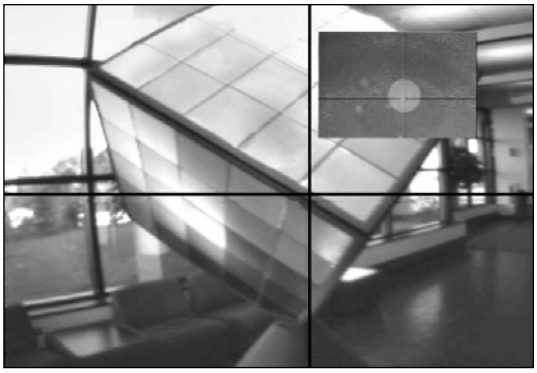


Figure 3 Video image of the scene with the eye image superimposed in the upper right corner. The crosshairs indicate point of gaze.

2.1.2 Integrated Eye and Head Tracking System

For the editing phase, horizontal and vertical eye position coordinates with respect to an LCD monitor were recoded using an ASL Model 501 eye tracker and Polhemus 3-Space magnetic transmitting system. Figure 4 shows a subject wearing the headgear illustrated in Figure 5. Because the system is based on NTSC video signals, gaze position on the LCD monitor is calculated at 60 Hz. However, gaze position values can be averaged over a variable number of video fields to reduce signal noise. Eight video fields were averaged for this task, yielding an effective temporal resolution of 133 msec.

Gaze position (integrated eye-in-head and head-position & orientation) was calculated by the ASL using the eye-in-head signal described above and a head position/orientation signal from a Polhemus Fastrak magnetic field head-tracking system. This system uses a fixed transmitter (mounted above and behind the subject) and a receiver attached to the eyetracker headband (Figures 4 and 5). The transmitter contains three orthogonal coils that are energized in turn. The receiver unit contains three orthogonal Hall-effect sensors that detect signals from the transmitter. Position and orientation of the receiver are determined from the absolute and relative strengths of the transmitter/receiver pair. The position of the sensor is reported as the (x, y, z) position with respect to the transmitter, and orientation as azimuth, elevation, and roll angles.

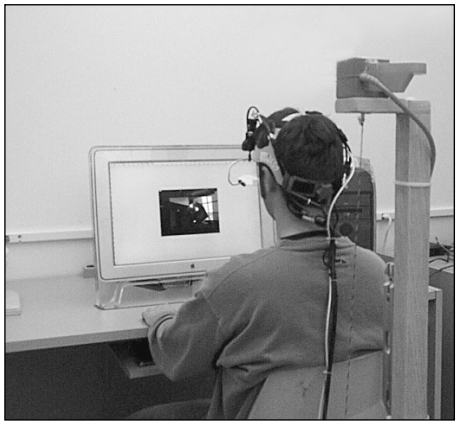


Figure 4 Magnetic transmitting and eye tracking setup.

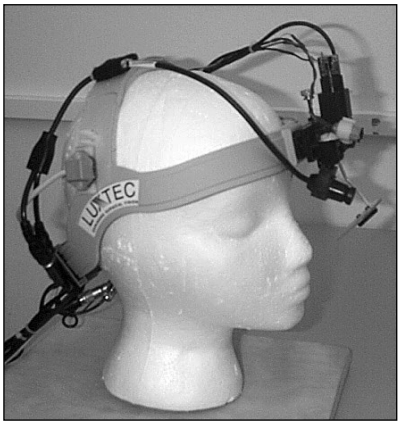


Figure 5 Close-up of headgear with head-position sensor.

The ASL control unit reports gaze position as the X-Y intersection of the line-of-sight with the LCD display. In order to calculate the gaze intersection point on the viewing plane, the position and orientation of the LCD display had to be measured with respect to a fixed point. The position and orientation of the display were defined by entering the three-dimensional coordinates of three points on the plane into the ASL control unit. The Fastrak transmitter is defined as the origin, and the distance to each of the three points is measured and entered manually. The gaze intersection point on the LCD display was collected by a laboratory computer for off-line analysis.

In addition to the data stream containing the gaze intersection on the LCD, the ASL creates a video record with a cursor overlay indicating gaze with respect to a scene camera attached to the eyetracker headband. Figure 6 shows a single frame from the video record.

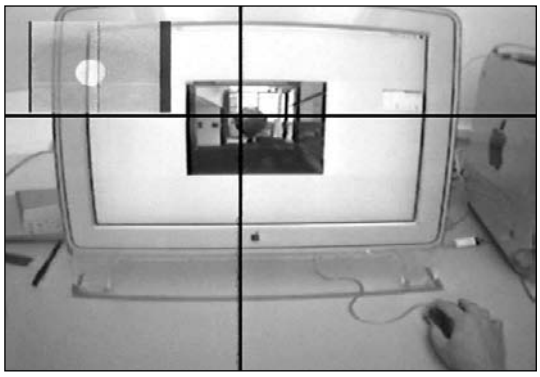


Figure 6 Video frame from the editing task. Cross hairs indicate gaze on screen.

Because the scene camera is not coaxial with the line of sight, calibration of the video signal is strictly correct for only a single distance. All gaze points are in the plane of the LCD display, and subjects typically do not change their distance from the display substantially, so the parallax error is not significant in this task, though it can be significant in tasks not constrained to a near-vertical plane. Note that the gaze intersection point calculated by the ASL (by integrating the eye-in-head and head position/ orientation signals from the transmitter and receiver) is not affected by parallax. The scene camera is used only during calibration when the distance to the scene is fixed. After initial calibration, the gaze intersection is calculated by projecting the eye-in-head position onto the display, whose position and orientation are known.

The ASL was calibrated for each subject before each trial session. Calibrating the ASL required three steps, 1) entering the position of the three reference points on the calibration plane as described above, 2) locating nine calibration points on the ASL controller, and 3) recording the subject's pupil and first Purkinje centroids as each point in the calibration target was fixated.

2.2 The Tasks

Subjects performed two primary tasks. The first task was to capture images with a digital camera. The second task was to edit (specifically to crop) the captured images on a computer. In addition, observers were asked to rate the cropped images and then fill out a brief questionnaire at the end of the experiment.

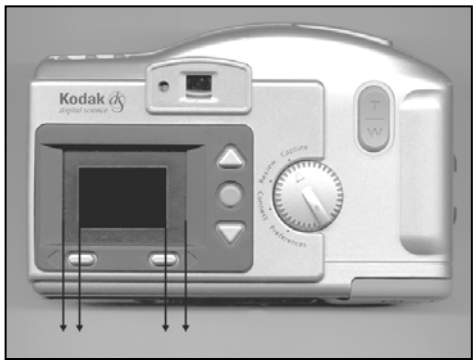


Figure 7 The camera's LCD preview panel was masked to limit the field-of-view during image capture.

2.2.1 Task 1: Image Capture

A Kodak DC-210 digital camera was used for image capture. The original size of the LCD panel was 1.63 x 1.20 inches, but for this experiment the LCD was masked off by a border of approximately 0.25 inches (see Figure 7). A mask closely resembling the appearance of the LCD was chosen so that subjects would be unaware of the alteration. The masking was done so that the photographer would capture more of the scene than was visible on the LCD (see Figure 8 for examples). The extra image area would later give subjects a second chance to manipulate their composition on the computer display. Subjects were not notified about the image editing task until after they finished taking the nine photographs.

After calibrating the wearable eye tracker, subjects were handed a 5.5 x 8.5 inch, double-sided, mock brochure describing the various types of research taking place in the Center for Imaging Science building. The inner spread of the brochure contained three empty rectangles meant to display, 1) a photograph of any person in the Imaging Science building, 2) a photograph of the large color sculpture on the first floor, and 3) a photograph of the stairway in the building atrium. The *primary objects* in these scenes define their class, and are referred to in the following text as *"person, "sculpture,"* and *"interior."* Figure 8 shows sample images of the three classes. The dark, transparent border indicates the image area not seen by the photographer through the LCD view finder in the digital camera.

The nine observers participating in this study were students and faculty from the Center for Imaging Science, so they were familiar with the people, offices and floors in the building. Participants were instructed on how to use the digital camera and told that they would be using it to take pictures for a brochure about the Imaging Science building. Their instructions were to take three horizontal pictures of each of the primary objects described above (totaling 9 images). All subjects were instructed to use the LCD panel to compose their images because it was not possible to track the eye movements directly through the camera's optical viewfinder.

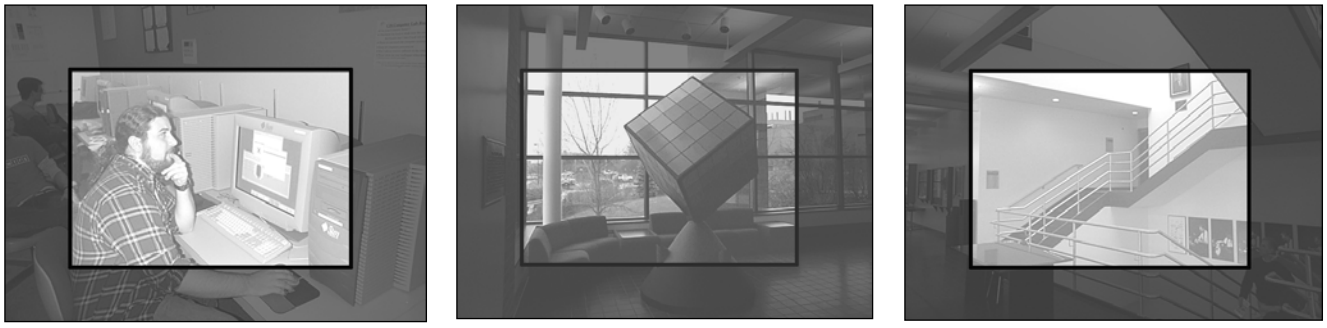


Figure 8 Sample pictures of the three image classes: *"person," "sculpture,"* and *"interior"*. The dark, transparent border indicates the portion of the image masked off in the LCD view finder in the digital camera as shown in Figure 7.

2.2.2 Task 2: Image Editing

After completing Task 1 (averaging, about 20 minutes) subjects reported back to the laboratory. The wearable eye tracker was taken off, and the person was informed that they would participate in a second task giving them the option to edit the images they had just photographed. Subjects were allowed to take a five minute break while the digital images from the camera were transferred to the computer and digital thumbnails were created.

After calibrating the ASL, participants went through a training session on how to use the crop tool in Adobe Photoshop. For this experiment, the cropping window was constrained so that the aspect ratio always matched the aspect ratio of the masked LCD viewfinder in the camera. While practicing on the test image, subjects were told that they could resize and move the crop window anywhere in the image as long as the window did not fall outside the image area. They were also told that if the default crop window framed the image exactly the way they wanted, they could simply double-click the mouse and everything outside of the default window would be cropped. Note that subjects were not told that the image area inside the default crop window was the image area they saw through the LCD view finder during image capture.

To start the actual task, subjects viewed the digital thumbnails of the nine images they had just photographed. Beginning with the *sculpture* class, subjects were asked to select the best of the three sculpture images for final editing. Subjects were then instructed to close their eyes while the experimenter loaded the full resolution image and positioned the crop window to its default location. This sequence was repeated for the *interior* and *person* image classes.

3. RESULTS

The results will be considered first for the capture phase, followed by the image editing phase, and finally the relationship between the two phases.

3.1 Task 1: Digital Image Capture

One measure of behavior before and during image capture is to compare the total time that the photographer spends looking at the scene before taking a photograph to the total time spent composing and taking the photograph. Figure 9 shows these values averaged across five photographers. The values for "scene" represent the sum of all gaze fixations on the *primary object* and *surround*[†] before lifting the camera. The bar representing "camera" is the total time spent looking at the back of the camera (mostly at the LCD) even when the photographer framed the object but did not take a photograph. Note that subjects were instructed to take three photographs of each primary object, so the times shown in Figure 9 indicate the time spent on all three photographs of each scene. It is evident from the figure that subjects spent approximately the same amount of time looking at the scene as they did framing and taking the photographs; about 42 seconds for all three pictures, or 14 seconds per photograph.

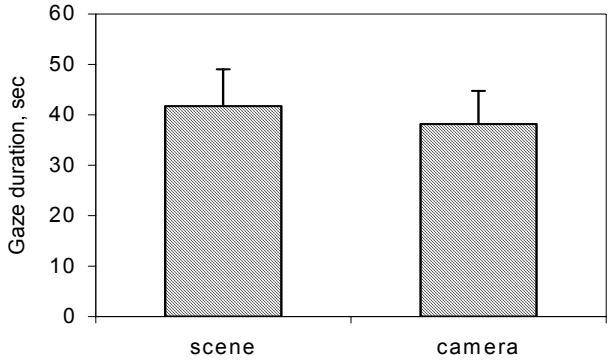


Figure 9 Gaze duration before and during image capture for all photographs. Average gaze duration (in seconds) on the *scene* before lifting the camera, and on the *camera* as the photograph was framed on the camera's LCD panel. Error bars represent one standard error of the mean for five subjects.

The analysis in Figure 10 averages gaze duration across the three classes: *person, sculpture*, and *interior*. This figure addresses the question of how the primary object being photographed influences photographers' behavior. As is evident in Figure 10, different classes significantly influence the time spent looking at the scene, but have less effect on the time spent looking behind the camera (i.e. framing and taking the photograph).

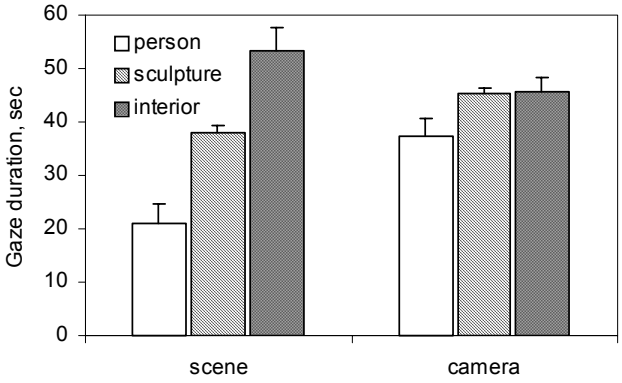


Figure 10 Gaze duration before and during image capture for the person, sculpture, and interior. Average gaze duration (in seconds) on the *scene* before lifting the camera, and on the *camera* as they framed the photograph on the camera's LCD panel. Error bars represent one standard error of the mean for five subjects.

While it is difficult to draw inferences on image class based on the small sample, we note that the trend in the total gaze duration on the scene shows a marked increase with the extent of the thing being photographed. The content in the *person* class is smallest, and takes up a relatively small portion of the frame. The *sculpture* takes up a larger portion of the frame, and the *interior* still more.

A finer-grained analysis of photographers' behavior was made by breaking down the total time spent looking at the scene into the time spent looking 1) at the primary object and 2) at the surround. This determination was straightforward in the *person* and *sculpture* classes, but more subjective in the *interior* class. The distinction is only clear after the photograph is composed and taken, as time spent on the primary object and surround regions reflects, to some degree, the decision-making process of what to include in the final photograph. In the interior case, the composing situation might be considered more "abstract". Thus, when extracting data from the video tapes, the criterion used in distinguishing the primary object from the surround was as follows:

[†]As mentioned earlier, the *primary object* is defined specifically as the "person", "sculpture", or "interior", and the *surround* is defined as any region that is not a *primary object* or the digital camera.

First, all but one photographer approached the atrium from the second floor of the building. As soon as the stairwell was in sight subjects began making fixations not typical of the normal stairwell usage. This distinction is based on previous examinations of video tape where subjects were performing a completely different task, in which case they used the stairs only to get from one floor to the next. Subjects were coded as looking at the primary object when they looked around at the ceiling, railings, and walls in a manner that did not indicate that they were using these fixations to help them go up or down the stairs. In some cases, subjects made eye movements to other people passing by on the stairs. These fixations were coded as gaze durations in the surround. Further, it was characteristic of subjects to look at individual steps as they walked up or down the stairs. Such eye movements appear to be used for navigation and were also recorded as surround gaze durations. In cases where the photographer stepped back far enough (typically after they had walked down to the first floor) to include most of the atrium, all fixations within the entire visible area (relative to the camera monitoring the scene) were recorded as fixations on the primary object.

Figure 11 shows the amount of time spent looking at the primary object, surround, and camera across the three image classes (*person, sculpture,* and *interior*). The underlying cause of the increasing total time spent looking at the "scene" in Figure 10 is evident in Figure 11. Consider first the difference between the *person* and the *sculpture* classes. There is no significant difference between the times spent looking at the primary objects in these two cases. For the sculpture class, the increase in gaze duration in the "scene" (shown in Figure 10) is due completely to an increase in the time spent looking at the surround, not the primary object. In the person class, photographers spent approximately the same amount of time looking at the person as the surrounding regions (an average of ~10 sec for the three photographs of each class), while the gaze duration in the surround was nearly three times longer in the sculpture class (~30 sec). Unlike photographing the person, where gaze duration through the view finder evidently draws more attention, the photographer is more concerned with the 'negative space' when photographing the sculpture. In examining the video records, it was quite common for photographers to walk around the sculpture to try different angles of composition. This behavior is not true when photographing the person, probably because the face, not the back or side of the head, is of greatest importance in the portrait.

Next we consider the difference between the gaze duration on the scene in the sculpture and interior classes. The scene gaze duration is ~40% longer in the interior class. It is clear from Figure 11 that the entire increase is due to longer gaze duration on the primary object in the *interior* class. Recognizing the more subjective classification between primary object and surround gaze location in the *interior* class, it is important to consider whether the difference could be due to misclassifying object or surround fixations. This is not likely, as the average time the photographers spent viewing the scene regions (the sum of the gaze duration on the primary object and surround) increased 40% (~15 sec) between sculpture and interior classes (shown in figure 10).

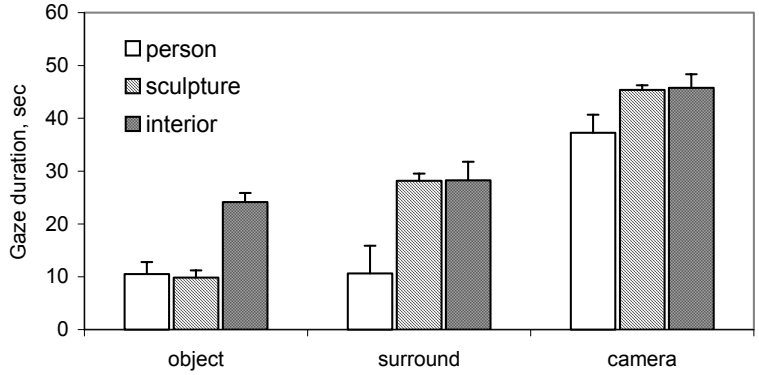


Figure 11 Gaze duration before and during image capture for the three scenes. Average gaze duration (in seconds) on primary object, surround, and the camera LCD panel. Error bars represent one standard error of the mean for five subjects.

3.2 Task 2: Image Editing

3.2.1 Duration and Number of Crop Windows

The videotaped records of each editing task were analyzed to determine the time spent editing each image, and the number of intermediate crop windows during the editing session[‡]. Unlike the dramatic differences between image classes observed during image capture, there was no significant difference in edit time or the number of crop windows. The trend toward longer edit times and more crop windows seen in Figure 12 (left) is not significant ($P > 0.5$).

While the average time spent per crop window was nearly constant across subject classes (7.2, 7.1, and 8.0 seconds per window for *person, sculpture*, and *interior*, respectively), the variables were only weakly correlated ($R^2 = 0.22$), and differed between subject classes. The far left graph in Figure 13 shows the relationship for all three image classes. The slope of a linear fit pooled across classes was 3.3 seconds per window. The three graphs on the right in Figure 13 illustrate how that value differed by class. The slope of the linear fit was 5.3 seconds per window for the *person* class, 4.4 seconds per window for the *sculpture* class, and only 1.0 second per window for the *interior* class.

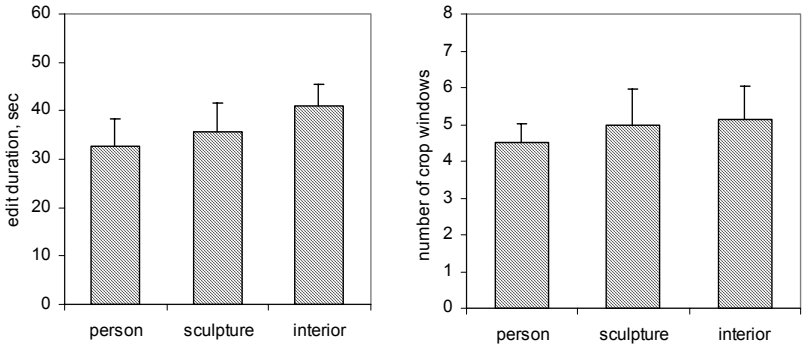


Figure 12 Average editing times (left) and the number of crop windows (right) used by subjects during image editing. Error bars represent one standard error of the mean for eight subjects.

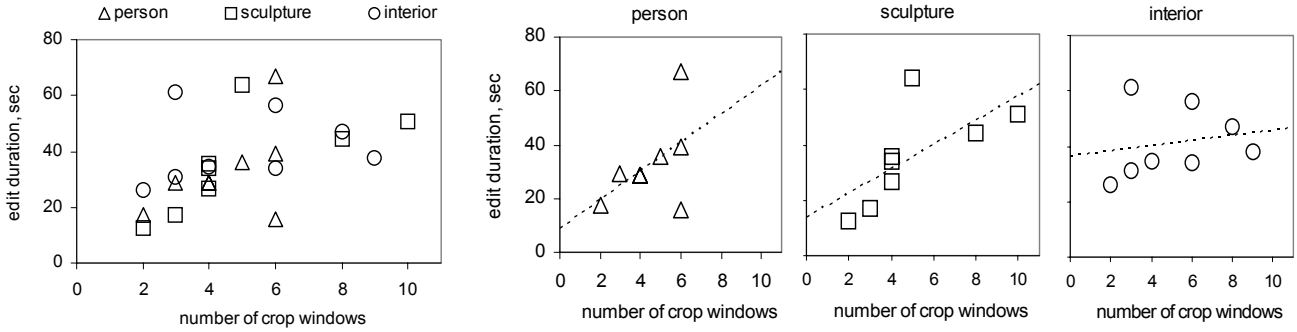
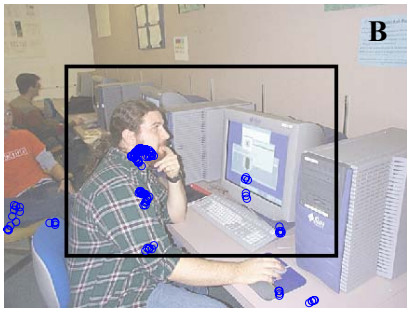


Figure 13 Total duration of edit session *vs.* the number of crop windows used during the editing task, pooled across *person, sculpture*, and *interior* classes (far left). Total duration of edit session *vs.* the number of crop windows used during the editing task for each *person, sculpture*, and *interior* image (right three).

3.2.2 Fixation Density Analysis

The previous section focused on the time photographers spent editing their digital images, and on the number of edit windows before the final crop was selected. In this section we consider the spatial distribution of gaze fixations during the editing process. Figure 14 A shows the fixation position for subject CS during the total editing time for the *person* photograph. The smaller figures show subsections of those fixations for the four intermediate crop windows (Figures 14 B, C, D, and F). The image without the black crop window (Figure 14 E) shows fixations while the subject was moving and manipulating the position of the final window.

[‡] Intermediate crop windows are defined as the portion of editing time where the crop window was stationary (i.e. the subject was not manipulating the size of the window or moving it).



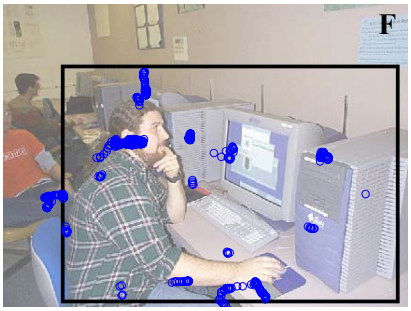


Figure 14 A Shows fixation densities on a *person* image across the entire editing session for subject CS. Figures 14 B, C, D, and F show fixations while the crop window is stationary. Figure 14 E shows fixations while the subject was dragging the window down and to the right.

Note that Figure 14 B is the default crop window, which is the same portion of the scene that the photographer saw through the LCD viewfinder of the digital camera just before taking the photograph. In editing this image, subject CS decided to expand the crop window from its default position. The fixation records suggest that the person's arm and hand were important parts to include in the final crop. Figure 14 D shows that the window was moved downward to include these features. The following fixations in Figures 14 D, E and F seem to depict the subject's decision of whether or not to exclude the background people in the final image. In comparison to the default crop window, it appears that CS decided to include more of the computer and the person's arm. In general, the eye movement patterns for other subjects revealed that where people looked during the cropping task often correlated with their decision to include or exclude those regions in the final image.

Results from the image capture task showed significant differences in looking at the primary object versus the surround. Because by definition the entire *interior* scene after capture can be considered the primary object, the same object/surround analysis could only be performed between the *person* and *sculpture* scenes. Unlike the differences shown in the image capture task, where subjects spent more than twice as long looking at the surround in the *sculpture* scene, there was no significant difference between the time spent fixating on the primary object and surround between the *person* and *sculpture* classes.

In order to compare the variability of eye movements across the different image classes for editing, we computed the mean horizontal and vertical eye position for all fixations occurring on each image for each *person, sculpture,* and *interior* class across eight subjects. Next, the radial distance from the mean fixation position to each individual fixation was calculated for all images. This is illustrated in Figure 15.

To estimate the fixation spread relative to the mean, the standard deviation of the radial distances across all fixations was then computed. Finally the mean standard deviation across the eight subjects for each of the three classes was compared. The graph in Figure 16 shows that there was no significant difference between the spread of eye movements (relative to the mean fixation position) for the *person, sculpture,* and *interior* classes for the cropping task.

Figure 15 Example of the spread from the mean fixation position across all fixations on the *person* image for subject CS. Such computations were performed for each image class, across eight subjects.

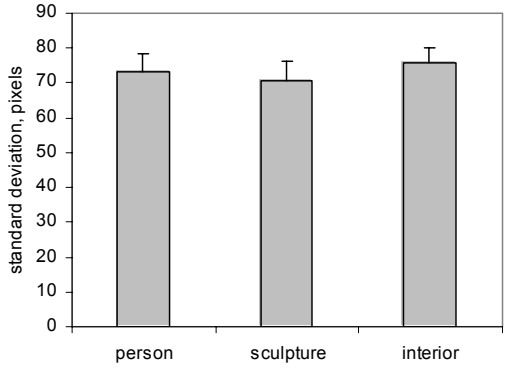


Figure 16 Standard deviation of the spread from the mean fixation position in pixels for each of the image classes. Error bars represent one standard error of the mean for eight subjects.

4. DISCUSSION AND CONCLUSIONS

It is interesting to note the observed asymmetry between image capture and image editing. For this set of conditions it appears that oculomotor behavior during image capture is dependent on the type of image being photographed. However, no specific differences in the cropping task were observed across the *person, sculpture,* and *interior* classes.

4.1 Task 1: Image Capture Summary

Data pooled across five subjects showed that the amount of time spent composing the image through the LCD view-finder (for the photographs) was nearly the same regardless if it was a *person, sculpture,* or *interior*. However, the time spent looking at the surround versus the primary object was different for these classes. Subjects photographing the *person* and *sculpture* classes spent about the same amount of time looking at the primary object, but they spent more than twice as long looking at the surround in the *sculpture* class. In comparing the *interior* and *person* classes, the time spent looking at the surround and the primary object was equally distributed, but the gaze duration was longer overall for the *interior* class. The differences in behavior, as revealed through eye movements, are evidently task-related. Further investigation may reveal that people look at specific features of a scene (in a characteristic manner) depending on the type of scene being photographed. However, we must also consider that subjects may have spent more time than they would under other circumstances because of the expectation that the final image would be used in a brochure. For example, taking candid snapshots might significantly change the photographer's behavior. Because there was only one image in each of the three classes, the descriptors (*person, sculpture,* and *interior*) are, to some degree, arbitrary. Further study is required to identify the relevant characteristics in a scene that make-up archetypical photographs.

4.2 Task 2: Image Editing Summary

Data pooled across eight subjects showed no significant differences in the amount of editing time for the three image classes. Further, the average number of intermediate crop windows used in the editing task was not significantly different for the *person, sculpture,* or *interior* images. The number of crop windows was weakly correlated with the total edit time. Averaged across scene classes, the total edit time increased 3.3 seconds per crop window. This slope was

shown to change depending on image class, but the correlation was weak for all classes. There was no significant difference in the average time spent looking at the surround and primary object for the *person* and *sculpture* classes, and there was much variability between subjects. To determine if the spread of eye movements was significantly different between the *person, sculpture* and *interior* classes, we compared the radial distances of all fixations to the mean fixation position for all images. The average standard deviation from the mean of all fixations was 73 pixels. This deviation did not differ significantly between image classes. For the *person* and *sculpture* images, mean fixation position was often spatially aligned with semantic regions in the image such as the face and center portion of the sculpture. The results from the editing phase agree with other reports that eye movements are image-dependant. However, the spread of fixations, edit time, and number of crop windows did not differ significantly across the three image classes. This suggests that the image-editing task is highly regular and independent of image class.

Like Buswell, we have used new research tools to study what people look at before, during and after image capture. The preliminary results reported here allow us to formulate better hypotheses now than at the beginning of the study. The next step is to expand this analysis across more observers using different scenes of the same class. It is not yet possible to compare scan paths made through the viewfinder to the scan paths made during image editing. The next step is to develop new instrumentation that allows us to track the eye movements of subjects looking through an optical viewfinder. This tool will enable us to directly compare the decisions photographers make when composing their images to the decisions they make while cropping these same images.

4. ACKNOWLEDGMENTS

This work was supported in part by a grant from the New York State Science, Technology, and Academic Research program, and by Eastman Kodak.

5. REFERENCES

1. G. T. Buswell, *How People Look at Pictures: A Study of The Psychology of Perception in Art*, The University of Chicago Press, Chicago, 1935.

2. R. A. Abrams, "Planning and Producing Saccadic Eye Movements," *Eye Movements and Visual Cognition: Scene Perception and Reading*, K. Rayner, p.66, Springer-Verlag New York Inc., New York, 1992.

3. S. P. Liversedge and J. M. Findlay, "Saccadic eye movements and cognition," *Trends in Cognitive Sciences*, **4**, pp. 6-14, 2000.

4. J. B. Pelz, R. L. Canosa, D. Kucharczyk, J. Babcock, A. Silver, and D. Konno, "Portable Eyetracking: A Study of Natural Eye Movements," *Human Vision and Electronic Imaging V*, B.E.Rogowitz and T. N. Pappas, SPIE Proc. **3659**, pp. 2000.

5. H. F. Brandt, *The Psychology of Seeing*, Philosophical Library, New York, 1945.

6. A. L.Yarbus, *Eye Movements and Vision* (B. Haigh , Trans.) Plenum Press, New York, 1967.

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

8. J. R. Antes, "The time course of picture viewing," *Journal of Experimental Psychology*, **103**, pp. 62-70, 1974.

9. G. R. Loftus and N. H. Mackworth, "Cognitive determinants of fixation location during picture viewing," *Journal of Experimental Psychology: Human Perception and Performance*, **4**, pp. 565-572, 1978.

10. J. M. Henderson, P.A. Weeks, and A. Hollingworth, “The effects of semantic consistency on eye movements during complex scene viewing,” *Journal of Experimental Psychology: Human Perception and Performance*, **25**, pp 210-228, 1999.

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

12. D. Noton, and L. Stark, ”Eye movements and visual perception,” *Scientific American*, **224**, pp. 34-43, 1971.

13. J. M. Henderson and A. Hollingworth, “Eye Movements During Scene Viewing: An Overview,” *Eye Guidance in Reading and Scene Perception*, G. Underwood, pp. 269-293, Elsevier, New York, 1998.

14. F. Molnar, “About the role of visual exploration in aesthetics,” *Advances in Intrinsic Motivation and Aesthetics,* H.I. Day, pp. 385-413, New York, Plenum Press, 1981.

15. C. F. Nodine, P. J. Locher, and E. A. Krupinski, “The role of formal art training on perception and aesthetic judgment of art composition,” *Leonardo*, **26**, pp. 219-227, 1991.