Eye Tracking Observers During Rank Order, Paired Comparison, and Graphical Rating Tasks

Jason S. Babcock[1,2], Jeff B. Pelz[1], and Mark D. Fairchild[2]
[1]Visual Perception Laboratory
[2]Munsell Color Science Laboratory
Rochester Institute of Technology
Rochester, New York

Abstract

In studying image quality and image preference it is necessary to collect psychophysical data. A variety of methods are used to arrive at interval scale values which indicate the relative quality of images within a sample set. The choice of psychophysical technique is based on a number of criteria, including the confusion of the sample set, the number of samples to be used, and attempts to minimize observer effort. There is an implicit assumption that the final result does not depend on the particular method selected. However, it may be the case that viewers adopt different strategies based on experimental methods. Task-dependent eye movements may be a source of variability when comparing results across different psychometric scaling tasks. This research focuses on learning where people center their attention during color preference judgments and examines the differences between paired comparison, rank order, and graphical rating tasks.

Introduction

Image Quality Judgments and Scaling

Experiments focusing on color tolerance for image reproductions and the effect of image content on color difference perceptibility have not investigated how observers look at images during standard psychophysical tasks.[1,2,3] Certain assumptions are made regarding the applicability of data collected in laboratory experiments. Specifically, perceptions resulting from psychophysical experiments are assumed to correspond to visual perceptions in real imaging devices and displays. Selecting the best psychophysical technique is often based on the confusion of the sample set, the number of samples used, and observer effort. Practical situations further dictate which method is most appropriate. For example, softcopy displays make best use of the paired comparison paradigm over rank order due to the impracticality of displaying many images on the screen while maintaining high-resolution. Assuming all other factors are equal, how well does a scale obtained from one technique compare to that

of another? Further, how do we know whether different experimental techniques themselves have any influence on the strategies adopted by observers? This paper examines whether viewing strategies are different across paired comparison, rank order, and graphical rating tasks. Eye movement data has been collected to show which regions (in the five images viewed) received the most "foveal" attention, and whether peak areas of attention are the same across the three tasks.

Eye Movements and Visual Perception

The central region of the eye occupies the area of highest visual resolution with peripheral vision having much lower acuity. This might be surprising, but physiological inspection reveals that the eye's retina is composed of two types of receptors called rods and cones. About 120 million rod photoreceptors occupy the retina (mostly in the periphery). Approximately 5-6 million cones occupy the central portion of the retina called the fovea. Despite the high sampling density of rods in the periphery, visual acuity from rods alone is quite poor due to signal pooling. Cones in the fovea are packed tightly together near the optical axis. The peak distribution of these receptors decreases substantially past one degree of visual angle. Unlike rods, cone photoreceptors in the fovea are not pooled, so the high sampling density is represented in the visual cortex. In this part of the brain the fovea occupies a much greater proportion of neural tissue than the rods. As a means of compensating for the bandwidth limitations in visual processing, humans rapidly shift their high-resolution fovea to areas of interest. In fact, it is estimated that humans shift their eyes over 150,000 times each day.

In the context of viewing static images, eye movements can be described as a sequence of fixations and saccades. A *fixation* indicates that the eye has paused on a particular spatial location in the image. A *saccade* indicates the period when the eyes move to the next fixation point. *Gaze* is the combination of head and eye movements to position the fovea.

The active combination of eye and head movements creates an adequate impression of high-resolution over the

entire visual field, so little attention is paid to the fact that vision is sharpest only in the fovea. Since we use vision as a tool to get information from the scene, point of gaze is closely linked to the stream of attention and perception. Recording eye movements is important to image quality studies because it can show where people direct their attention when viewing images and making image quality judgments.

Eye Movements and Picture Viewing

Buswell provided the first thorough investigation of eye movements during picture viewing.[4] He showed that observers exhibited two distinct viewing behaviors. In some cases viewing sequences were characterized by a general survey of the image. In other cases, observers made long fixations over smaller regions in 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. A number of experiments since Buswell have focused on understanding and modeling the role of eye movements in image perception.[5] In general, these experiments have demonstrated that most observers deploy their attention to the same general regions in an image, but not necessarily in the same temporal order. They have shown that where people look is not random and that eye movements are not simply bottom-up responses to visual information. Further, these experiments indicate that the level of training, the type of instruction, and observer's background all have some influence on the observer's viewing strategies.

Methods

Eye Tracking Instrumentation

An Applied Science Laboratory Model 501 eye tracking system was used in conjunction with a Polhemus 3-Space Fastrak magnetic head tracker (MHT) for all experiments. The headgear houses an infrared LED illuminator, a miniature CMOS video camera (sensitive to IR) to image the eye, and a beam splitter used to align the camera so that it is coaxial with the illumination beam. A second miniature CMOS camera is used to record the scene from the subject's perspective. This provides a frame of reference to superimpose crosshairs that indicate the subject's point of gaze. In addition to a video record, horizontal and vertical eye position was recorded with respect to the monitor plane. The integrated eye-in-head coordinates were calculated at 60 Hz by the ASL control unit using the bright pupil eye image and a head position/orientation signal from the MHT.

Displays

Display size is important in eye movement studies because the accuracy of the track is defined as a function of visual angle. At a constant distance larger displays result in a smaller fraction of fixation uncertainty within an image. For these reasons a 50'' Pioneer Plasma Display (PPD) and a 22'' Apple Cinema Display (ACD) were used to present stimuli. Each display was characterized using one-dimensional lookup tables followed by a 3x3 matrix. Optimal flare terms were estimated using the techniques outlined by Berns, Fernandez and Taplin (in press) and a regression-based channel interdependence matrix was included to further improve the accuracy of the forward models.[6]

In estimating the accuracy of the track across subjects, the average angular distance from the known calibration points and the fixation records was calculated for both 9 and 17-point targets. On-average, the accuracy of the eye tracker was 1°. However eye movements toward extreme edges of the screen did produce deviations as large as 5°.

Images and Stimulus Presentation

Observers performed rank order, paired comparison, and graphical scaling tasks on both displays evaluating the five images in Figure 1. For the plasma display, images were 421 x 321 pixels, subtending 13 x 9° at a viewing distance of 46 inches. For the LCD, images were 450 x 338 pixels with a visual angle of 9.5 x 7° at a distance of 30 inches.

Figure 1. Images used for the psychophysical scaling tasks.

For each image shown in Figure 1, five additional images were created by manipulating lightness, chroma, or hue. The intention was to simulate variability from a set of digital cameras or scanners. Adobe Photoshop was used to perform hue rotations for the *kids* and *firefighters* images and chroma manipulations for the *bug* image. The *wakeboarder* and *vegetables* images were manipulated by linearly increasing/decreasing the slope of CIE L^*_{ab} in the original image.

Nineteen paid subjects, (5 females, 14 males,) ranging from 19-51 years of age participated in the experiment. Eye tracking records from six of the subjects were discarded due to poor calibration, excessive number of track losses, and problems related to equipment failure. Psychophysical data was collected for all 19 observers. Stimulus presentation for the rank order, paired comparison and graphical rating tasks was implemented as a graphical user interface in Matlab. The rank order interface displayed all six manipulations and observers ranked them from 1 to 6, where 1 was most preferred and 6 was least preferred. In

the paired comparison task subjects used the mouse to select the most preferred of the two images displayed on screen. All pairs of the 6 manipulations were displayed in random order. In the graphical rating task subjects used a slider bar to rate the image quality between two imaginary extremes (no anchor pairs were given).

Results

Fixation Duration

Due to limited space, only results for the Pioneer Plasma Display are presented here. The full results and details of the experiment can be found in Babcock 2002.[7]

For 13 subjects and two displays, measures of fixation duration showed that viewers spent about 4 seconds per image in the rank order task, 1.8 seconds per image in the paired comparison task, and 3.5 seconds per image in the graphical rating task.

Although the amount of time subjects spent looking at images for each of the three tasks was different, peak areas of attention, as indicated by fixation density maps, show a high degree of similarity. Foreground objects in the *kids* and *firefighters* images clearly received a higher density of fixations than other objects (e.g. the right firefighter, and the kids' faces). Fixation peaks were also similar across the three tasks for the *wakeboarder* and *bug* images (not shown here).

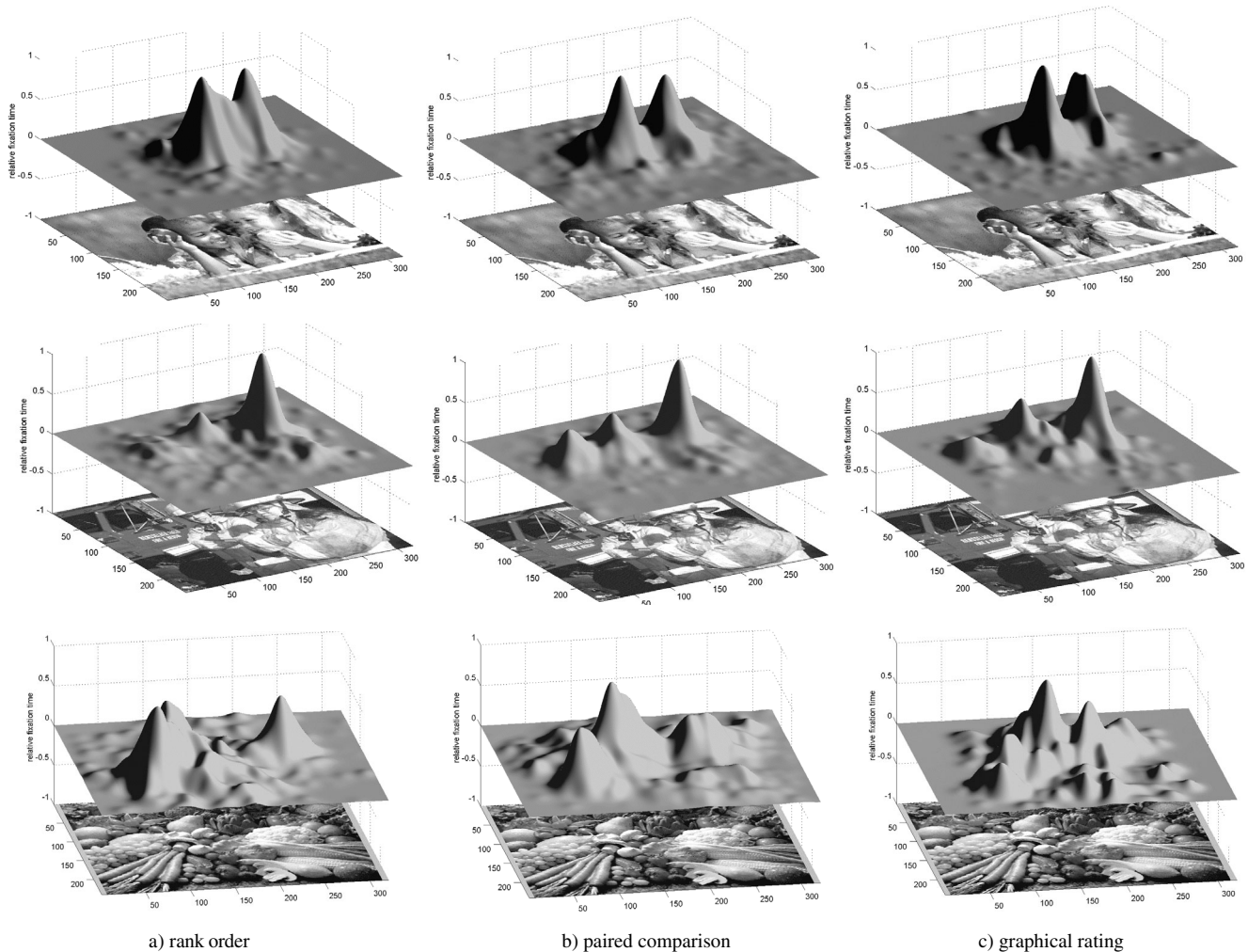a) rank order   b) paired comparison   c) graphical rating

Figure 2. Graphs show normalized fixation density across 13 subjects for the rank order, paired comparison, and graphical rating tasks for the kids, firefighters, and vegetables images (viewed on a 50" plasma display).

Fixation density maps for four of the five images were similar across the rank order, paired comparison, and graphical rating tasks. This similarity was quantified by calculating the 2-D correlation between tasks (i.e. rank order vs. paired comparison, rank order vs. graphical rating, etc.) using Equation 1. The 2-D correlation metric is sensitive to position and rotational shifts and provides a first-order measure of similarity between two grayscale images. Table 1 presents the correlations calculated between fixation maps for all pairs of the three scaling tasks for the Plasma display.

$$r = \frac{\sum_m \sum_n (A_{mn} - \overline{A})(B_{mn} - \overline{B})}{\sqrt{\left(\sum_m \sum_n (A_{mn} - \overline{A})^2\right)\left(\sum_m \sum_n (B_{mn} - \overline{B})^2\right)}} \quad (1)$$

Table 1. Correlation (r) between rank order, paired comparison, and graphical rating fixation maps for the Pioneer Plasma Display.

Pioneer Plasma	wakeboarde r	vegetable s	firefighter s	kids	bug
rank order : paired comp	0.90	**0.74**	0.90	0.9 4	0.9 3
rank order : graphical rating	0.86	**0.62**	0.92	0.9 3	0.9 2
paired comp : graphical rating	0.93	**0.82**	0.87	0.9 2	0.9 0

Table 1 shows that the *vegetables* image produced the lowest overall correlation between the three tasks, and that rank order fixation maps compared to graphical rating fixation maps were most different. This result is likely due to the spatial complexity of the image and the variety of objects with distinct memory colors. Highlight regions on the mushrooms and cauliflower objects were clipped for boosts in lightness. These objects seemed to attract a high degree of attention, but not with the same weight. Because the *vegetables* scene had over 20 distinct objects, it is also likely that observers moved their eyes toward different regions out of curiosity, causing unique fixation maps across tasks.

The *bug* and *kids* images resulted in the highest overall correlations across tasks. This result is likely related to the fact that semantic features were located mostly in the center of the image and that surrounding regions were uniform with low spatial frequency and moderate color changes. Since flesh tones are important to image quality peak fixation density was expected for faces in the *wakeboarder*, *firefighters*, and *kids* images.

Circling Regions Used to Make Preference Decisions

It is natural to ask whether regions of interest could be identified by simply asking viewers to physically mark or circle important regions in the image rather than track their eye movements. One question is whether regions with a higher number of fixations correspond to regions identified by introspection. To make this comparison, subjects were given a print-out (at the end of the experiment) showing

the five images in Figure 1. Directions on the sheet instructed observers to: "Please circle the regions in the image you used to make your preference decisions." Each participant's response was reconstructed as a grayscale image in Adobe Photoshop. Circled regions were assigned a value of 1 digital count and non-circled areas were assigned a value of 0 digital counts. Figure 3a shows an example across 13 subjects for the *kids*, *firefighters*, and *vegetables* images.

Figure 3b shows fixation density and the regions of importance circled for individual subjects. In some cases, participants circled areas that received very few fixations.
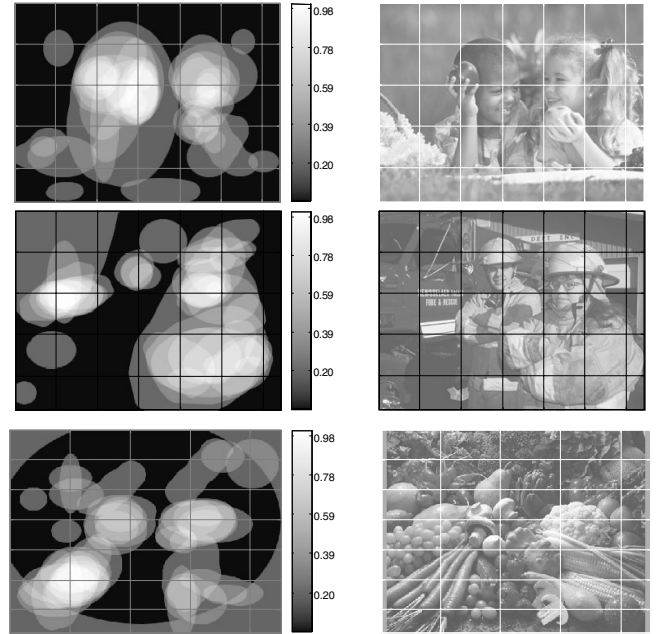

Figure 3a. Circles regions (across 13 subject) used to make preference decisions.
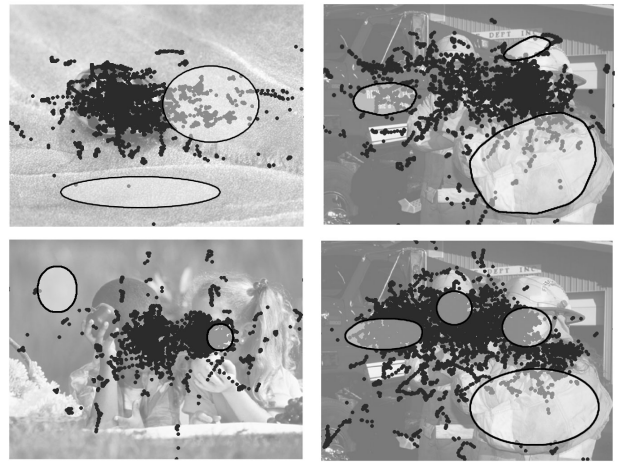

Figure 3b. Black markers indicate fixations compiled across the six manipulations for both displays from a single individual. Circles indicate regions in the image that were important to the observer's preference decision.

The upper-left image in Figure 3b shows a subject's eye movement record collapsed across all observations of the *bug* image for both displays. The circled areas are superimposed over the image indicating that the bottom portion of the leaf was important in the preference decision. However, very few fixations occurred in those regions. Inconsistencies between self-report and eye movement records were also evident in three other subjects looking at the *firefighters* and *kids* images. It is evident that subjects' peak areas of attention do not necessarily agree with introspective report.

Scale Values

Figure 4 graphs the interval scale values as a function of image manipulation (difference from the original image, i.e. ΔL*, Δh°, etc.) for the *kids*, *firefighters*, and *vegetables* images. Graphical rating data across all subjects was put on a common scale by subtracting the mean value from each observer's rating and dividing that result by the observer's rating scale standard deviation.[8] Rank order and paired comparison data were converted to frequency matrices and then to proportion matrices. Because there was unanimous agreement for some pairs, zero-one proportion matrices resulted. All values that were not zero-one were converted to standard normal deviates and the scale values were solved using Morrisey's incomplete matrix solution. As suggested by Braun and Fairchild, confidence intervals were computed using 1.39/sqrt(N), where N is the number of observers.

Conclusions

This paper examined fixation duration, locus of attention, and interval scale values across rank order, paired comparison, and graphical rating tasks. In judging the most preferred image, measures of fixation duration showed that observers spent about 4 seconds per image in the rank order task, 1.8 seconds per image in the paired comparison task, and 3.5 seconds per image in the graphical rating task. Spatial distributions of fixations across the three tasks were highly correlated in four of the five images. Peak areas of attention gravitated toward faces and semantic features, which supports pervious eye movement studies. Introspective report was not always consistent with where people foveated, implying broader regions of importance than indicated from eye movement plots. Psychophysical results across these tasks generated similar, but not identical, scale values for three of the five images. The differences in scales are likely related to statistical treatment and image confusability, rather than eye movement behavior. There appears to be a relationship between scale variability and eye movement variability, but this result must be further investigated with a larger number of observers.
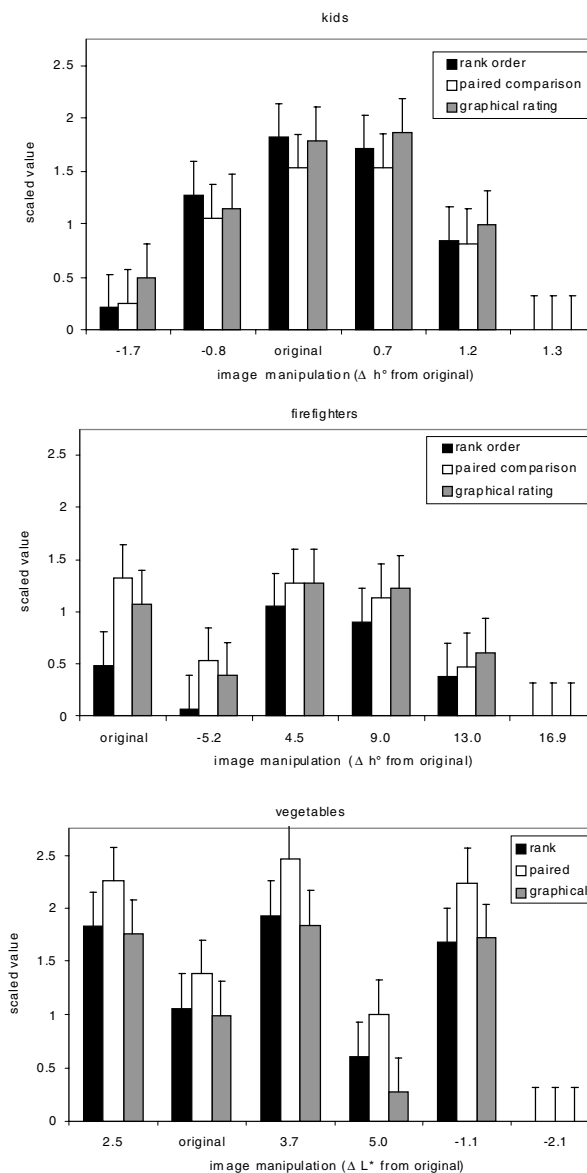

Figure 4. Interval scale values as a function of image manipulation for the Plasma display.

References

1. M. Stokes, Colorimetric Tolerances of Digital Images, Master's Thesis, Rochester Institute of Technology, Rochester, NY (1991).
2. S.R. Fernandez, Preferences and Tolerances in Color Image Reproduction, Master's Thesis, Rochester Institute of Technology, Rochester, NY (2002).
3. S.P. Farnand, The Effect of Image Content on Color Difference Perceptibility, Master's Thesis, Rochester Institute of Technology, Rochester, NY (1995).
4. G.T. Buswell, How People Look at Pictures, Univ. Chicago Press, Chicago (1935).
5. J.M. Henderson, A. Hollingworth, Eye movements during scene viewing: an overview. In G. Underwood (Ed.), Eye

Guidance in Reading and Scene Perception, pp. 269-293, Elsevier, New York (1998).

6. R.S. Berns, S.R. Fernandez, L. Taplin, Estimating black level emissions of computer-controlled displays, Color Res. Appl., (2002, in press).

7. J.S. Babcock, Eye Tracking Observers During Color Image Evaluation Tasks, Master's Thesis, Rochester Institute of Technology, Rochester, NY (2002).

8. P. Engeldrum, Psychometric Scaling: A Toolkit for Imaging Systems Development, Imcotek: Winchester, MA (2000).

Biography

Jason Babcock is a Research Scientist at the Visual Perception Laboratory in the Chester F. Carlson Center for Imaging Science at RIT. He recently completed an M.S. Color Science thesis investigating visual behavior during image-quality evaluation and chromatic adaptation tasks. He received a B.S. in Imaging and Photographic Technology from RIT in 2000. His research interests include the development of portable eye tracking systems, color perception, and the study of image quality and aesthetics.