HOW THE TASK OF EVALUATING IMAGE QUALITY INFLUENCES VIEWING BEHAVIOR

Hani Alers[1], Lennart Bos[1], Ingrid Heynderickx[1,2]

[1]Delft University of Technology, Delft, The Netherlands;
[2]Philips Research Laboratories, Eindhoven, The Netherlands

ABSTRACT

Image quality scores collected in subjective experiments are widely used in image quality research, particularly in the design of objective quality assessment algorithms. It is therefore of vital importance to make sure that the collected scores reflect viewers' opinions in real-life situations. However, just by giving the viewers the task of assessing quality, there is a risk that their behavior has changed from what it would be in a natural viewing condition. We here investigate this difference in behavior by examining the gaze response in both conditions, i.e. free looking and scoring quality, with the help of eye-tracking equipment. Even though the observed behavior shows similarities between the two conditions, there are also significant differences which should be taken in consideration in future image quality research.

Index Terms— Image Quality, Region of Interest, Perception, JPEG, Compression, Task Effect on Saliency.

1. INTRODUCTION

Research in image quality assessment and image quality optimization algorithms has always considered subjectively collected image quality scores to be the ground truth [1,2]. The performance of an objective quality-assessment algorithm is measured by how closely it judges the images compared to the Mean Opinion Score (MOS), which basically reflects the average quality score that viewers give an image. However, it has also been known for a long time that the task given to a viewer can have a great impact on how they look at an image [3,4]. The question that one should ask is whether the task of image quality assessment (IQA) changes the viewing behavior of the observers, and if so then how.

To investigate this question, a database of images with a clear region of interest (ROI) is created and then degraded to different levels of quality. We first focus on images with a clear ROI in order to have a distinct and clear gaze response from the viewers. These images are then shown to two groups of viewers, one where they score the quality of the images, and the other where they simply look freely at the images. By using an eye tracker, saliency maps are created for each task. Similar studies conducted earlier compared saliency maps of general image content between the two viewing conditions [5,6]. However, by creating a new database emphasizing images with a clear ROI, we are capable of specifically analyzing changes in the unique ROI, which helps us to explain the differences found in a clear and explicit manner. Our hypothesis states that viewers give more attention to the ROI during free looking than during scoring, since during scoring their attention may be focused on compression artifacts or other quality degradation aspects in parts of the image outside the ROI. They would hence pay less attention to the ROI compared to people who are looking freely.

This paper starts by explaining the experimental setup in Section 2. This is followed by a detailed description of the experimental protocol for both conditions in Section 3. Section 4 explains how the eye-tracker data are processed to create saliency maps and identify the ROI of the images used. The resulting data are then used in Section 5 where we compare the differences in the gaze response between the two conditions. Section 6 attempts to explain the results and use them to shed light on what viewers do in each viewing condition. The paper ends in Section 7 with some conclusions and recommendations for future work.

2. METHODOLOGY

2.1 Stimuli

The stimuli used in the experiment were created from 40 original images. Each image was further processed to produce 4 different versions, which resulted in a total of 160 stimuli used in the experiment. Considering the goal of the experiment, we only chose images that contained a clear ROI in the form of a face, an animal, or an object that clearly stood out from the rest of the image. Images were cropped to 600 by 600 pixels in order to have a standard size for all the images.

Each original image was degraded towards four different versions with the JPEG compression implementation of the `imwrite` function, defined in MATLAB. The four different levels of compression used to process the images ranged between 10 (low quality) and 100 (high quality).

978-1-4577-1334-7/11/$26.00 ©2011 IEEE

2.2 The eye tracker

An eye-tracking system, as shown in Figure 1, was used to determine the gaze location of the users while viewing the images. The system used in the experiment was the iView X system developed by SMI. In order to track the eye, this system identifies the location of the pupil from the reflections of infrared light on the retina. To this end, it uses an infrared light source mounted above the lens of an infrared camera. Since infrared falls outside the human visual sensitivity spectrum, the viewer is not distracted by the light emitted from the system. The REDIII camera used by the eye-tracker has a sampling rate of 50 Hz and a tracking resolution of ± 0.1 deg. The gaze position tracking accuracy is ± 1 deg. Viewers were asked to place their head on a head rest as recommended by the eye tracking system's manual in order to avoid head movements and get the highest accuracy. The head rest kept the viewer at a distance of 60cm from the screen, which represented a typical viewing distance and fell in the system's recommended operating distance of 40-60 cm. The height of the head stand was adjusted to suit the viewer and insured a comfortable and non-confining seating position while performing the experiment. The eye tracker was always calibrated using a 13-point grid, where the points were equally distributed over the full display screen used in the experiment.

2.3 The experimental setup

The images were displayed on a 19-inch CRT monitor with a resolution of 1024x768 pixels and an active screen area of 365x275mm. The experiment was controlled from a remote computer with its monitor positioned so that it was not viewable by the participant (see Figure 1). In order to avoid outside elements interfering with the results, the experiment was carried out in the User-Experience Lab located in the Electrical Engineering, Mathematics and Computer Science (EEMCS) faculty building at the Delft University of Technology. Only the experimenter and the viewer were present during the experiment. The lab also gave us the ability to control the light level independently of the outside lighting conditions. The illumination was kept at 70 lux measured vertically at the position of the screen.

2.4 The participants

The experiment had a total of 60 participants. They were collected from the faculty of Computer Science at the Delft University of Technology, and were either students or staff members. It is therefore estimated that all participants possessed some experience with the type of degradation and artifacts caused by JPEG compression. When asked whether they suffered from any vision problems, they all expressed having sound (corrected) vision. This was considered sufficient to ensure that they were able to observe the differences in image quality. All participants were naive to the purpose of the experiment.

3. THE EXPERIMENTAL PROTOCOL

As mentioned before, the experiment included 2 separate phases. Phase1 required people to examine images and give them a score based on their quality, while participants in phase2 were only asked to look at the images without a predefined task. The participants were divided over the two phases with 20 participants in phase1, and 40 in phase2. A different number of participants was used since each set of data was analyzed differently. All participants in phase1 judged all 160 stimuli (in four separate sessions), and so we obtained 20 different quality scores and saliency maps per stimulus. The participants in phase2 saw each image content only once, albeit at a different quality level. As such, they only saw 40 stimuli, and the combination of all 160 stimuli was seen by a group of 4 participants. As a result, we obtained 10 saliency maps per stimulus in phase2. This is further detailed below in the separate description of each of the phases of the experiment.

The participants were informed that they would carry-out an experiment on image quality research. They were told that the position of their gaze would be recorded using an eye-tracking device. This was followed by a quick test to check whether the eye tracker locked on the participant's pupil. Those who passed this check were asked to start the experiment. In order to ensure consistency, the instructions for the experiment were given to the participants through the computer screen, together with examples of how to perform each step. This was followed by a short training which involved showing two images to the participants in the free-looking condition. The training session for the scoring condition showed the participants five images spanning the quality range of the images used in the experiment, and gave them a chance to practice on using the scoring screen. When the training was completed, the subjects were allowed to ask questions about any unclear points. Once they were ready to start, the experimenter started the eye-tracker calibration process, and then started showing them the stimuli.

3.1 Phase1: Scoring Image quality

As mentioned above, each participant in phase1 was shown all 160 compressed stimuli. The experiment was split in 4

Fig. 1. The experimental setup showing the participant using the chin-rest while looking at the images on the screen with the eye tracker recording the gaze data.

sessions requiring the participants to evaluate 40 images in each session. Every session contained one compressed version of each original image content. The system chose the image at random ensuring that at the end of the session, the participant saw one version of each of the 40 original image contents in the database. In the subsequent sessions, the participant was shown one of the remaining versions of each image, such that at the end of the fourth session all versions were seen once by each participant. The order in which the stimuli were shown in each session was also chosen randomly by the system. Between the sessions, the participants were given a short break where they could take their head off the chin-rest and have something to drink. This was done to avoid strain developing in the neck and back muscles, and in order not to exhaust the eyes of the participants.

The experiment followed the single-stimulus protocol set by the ITU [7]. The participant was shown a gray screen (R,G, and B values set to 127) with a white dot in the center. He or she was asked to focus their gaze on that dot while it remained on the screen for 3 seconds. The eye-tracking data collected during these three seconds were later used to correct for shifts from the eye-tracker calibration position. Subsequently, a randomly selected image was displayed at the screen centered on the gray background. Participants were allowed to examine the image until they decided on the quality score. They could then use the left mouse button to go to the scoring screen, where they saw a horizontal slider bar separated into 10 equal segments with the words "Poor" on the left and "Excellent" on the right. The slider could be controlled by moving the mouse towards the intended score. A click on the left mouse button saved the score and took the participant again to the gray screen with the white dot in the center. The system then chose another image from the database randomly, but with the constraint that it was not created from the same original content as any of the previously scored images in the session. These steps were repeated until the participant scored 40 different images. After a short break, the participant started the following session by first completing the 13-point calibration step described earlier, followed by another 40 stimuli to be scored. This process was repeated in 2 more sessions taking each participant through the entire database of 160 stimuli. The amount of time needed by participants to complete all these steps varied since the viewing time duration was not fixed, but on average, it took them about 45 minutes.

3.2 Phase2: free looking

Here the viewers were not given any task and were only asked to view the images in a casual manner. The data collected from this phase of the experiment was later used to subjectively identify the natural ROI of the images. To avoid any deviation in the measured ROI due to a learning effect resulting from viewing the same image content multiple times, participants only viewed one version of each image.

Phase2 was performed concurrently with phase1, taking place at the same lab and using the same equipment and setup. Participants were told to simply look at the images as if they were viewing a photo album. Every image shown in the experiment was preceded by a gray screen with the white dot in the center of the screen similar to that used in phase1. Participants were instructed to focus on the white dot while it appeared on the screen, which again gave us a uniform starting gaze position for all images and provided us with data which could be used to correct for shifts from the tracker's calibration position.

After completing the training, the participants went through the 13-point calibration step as before and then started viewing the images. Each image was displayed on the screen for 8 seconds followed by the gray screen. Basically, each participant saw a selection of stimuli as if he or she has completed one session of phase1 (requiring approximately 10 minutes). As a result, every 4 participants saw the entire set of 160 images presented at a random order, while each of them only saw one compressed version of each original content. By the end of phase2 we gathered the free looking gaze data from 10 participants for each version of the compressed images.

4. PROCESSING EYE TRACKER DATA

4.1 Creating the saliency map

To visualize the eye tracker data, a height map was used where the height at a given coordinate indicates the total duration of the fixations of all test subjects to that coordinate. To construct this map, each fixation location gives rise to a gray-scale patch whose activity is distributed using a Gaussian function. The width (σ) of the Gaussian patch is approximates the size of the fovea (about 2° visual angle). A mean saliency map that takes into account all fixations of all subjects is calculated as follows:

A B

Fig. 2. The image on the left shows a visualization of the eye tracking data on the displayed image. Once the ROI threshold is chosen, we get the mask shown in the image on the right.

$$S_i(k,l) = \sum_{j=1}^{T} \exp[-\frac{(x_j - k)^2 + (y_j - l)^2}{\sigma^2}]$$

where $S_i(k, l)$ indicates the saliency map for stimulus I_i of size MxN pixels (i.e. $k\epsilon[1, M]$ and $l\epsilon[1, N]$), (x_j, y_j) indicates the spatial coordinates of the *jth* fixation ($j=1…T$), T is the total number of all fixations over all subjects, and σ indicates the standard deviation of the Gaussian. The intensity of the resulting saliency map is linearly normalized to the range $[0, 1]$. Figure 2.A shows an example of a height map or saliency map with the warmer colors representing high saliency. With this visualization it was possible to manually observe that the saliency maps clearly showed the highest concentration of fixations around the expected ROI of the images: people, animals, faces, and other objects in the center or foreground.

4.2 Identifying the Region Of Interest (ROI)

The saliency maps described were used to determine the ROI of an image. This was accomplished by considering all values in the saliency map above the saliency value of 0.2 to belong to the ROI. Figure 2.B shows how applying the threshold on the saliency map of Figure 2.A creates an ROI mask which nicely encloses the most salient points in the image. Similar ROI masks were created for all 40 image contents used in the experiment.

4.3 Measuring the fixation-ratio

In this experiment we aimed to analyze how the task of image quality assessment changed the viewing behavior of the observer. The eye-tracking data under the free-viewing condition clearly showed that observers spent the majority of the time looking at the ROI. We then defined for each condition a "fixation-ratio" being the sum of the duration of the fixations inside the ROI divided by the total duration of all fixations. For example, if the fixation-ratio of a given participant was 0.7 for a particular stimulus, this would mean that this participant fixated for 70% of the time inside the ROI for that particular stimulus and for 30% on other areas in the image outside the ROI.

4.4 Comparing the scoring and free viewing data

As explained above, the viewing time per stimulus was fixed at 8 seconds in Phase2. However, the sessions in Phase1 had variable viewing times, because the participants could stop viewing the image and continue to the scoring screen at any time. To make the two sessions comparable, the average viewing time participants used in Phase1 was used as the limit to the viewing time in Phase2. That is to say, since the average viewing time in Phase1 was approximately 5 seconds, the data analysis of Phase2 included only data recorded during the first 5 seconds. The data gathered in the remaining 3 seconds was simply discarded.

5. RESULTS

5.1 Differences in spatial behavioral

Our hypothesis stated that the fixation-ratio would be higher for the free–looking condition than for the scoring task, since we expected that the viewers who were scoring would try to seek compression artifacts or other quality

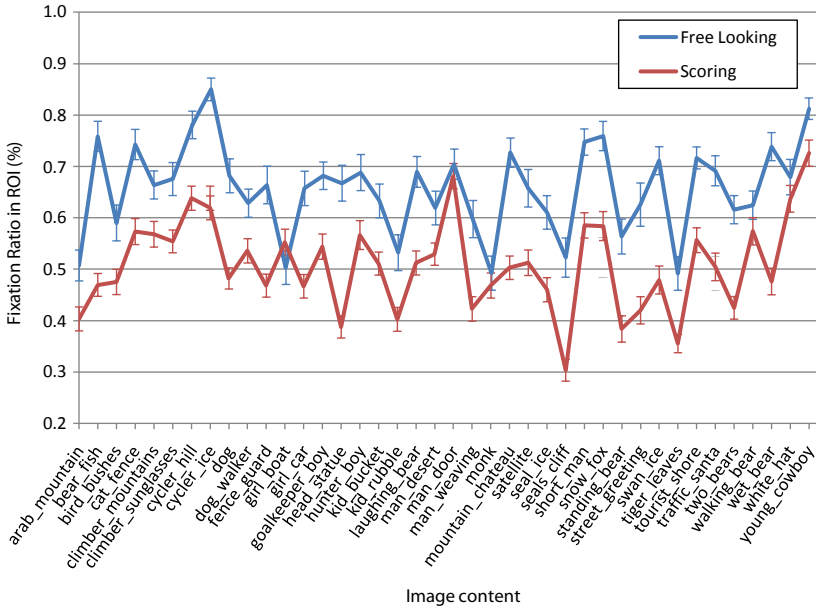

Fig. 3. The fixation-ratios for the used stimuli grouped by image content for both scoring and free viewing tasks

Fig. 4. Although the girl_boat image is highly degraded, the dark background areas significantly mask the compression artifacts.

degradations over the whole image, and not only in the ROI. They would hence pay less attention to the ROI compared to people who were looking freely. An analysis of variance (ANOVA) was performed on the calculated fixation-ratios to statistically determine the effect of the task. In this case a between-subject test was performed, since each test subject participated in only one of the two tasks. The results of the ANOVA analysis confirm our hypothesis: the between-subject test shows a significant difference in the task ($F=31.6$, $df=1$, $p=0.001$), which means that people who were asked to score the image quality indeed looked differently to the ROI than people who looked freely. Figure 3 plots the values of fixation-ratio per image for both tasks separately. It shows that there is a lot of variation amongst the 42 different images, but for almost all of them the fixation-ratio is higher for the free-looking task than for the scoring task. This means that people who are given the task to score the quality of an image tend to look significantly less to the ROI and more to the surrounding than people without a task. The only exception is the image girl_boat. Figure 4 shows that particular image in its lowest quality version (created using the `imwrite` MATLAB function with the compression value of 13). As one can see in the figure, the image contains an exceptionally dark background which masks the compression artifacts. This may have compelled the observers to pay more attention to the region of interest than to the (hidden) artifacts in the background.

Task	Duration inside ROI	Duration outside ROI
Free looking	503.47	322.88
Scoring	402.38	316.20

Table 1 Mean duration (in ms) of the fixations inside and outside the region of interest per task.

5.2 Differences temporal behavioral

The collected data also show an interesting temporal aspect of the task on the viewing behavior. Table 1 shows the mean fixation duration in milliseconds for both the fixations inside and outside the ROI for both tasks separately. As can be seen from the table, the mean durations of the fixations outside the ROI are almost similar ($F=0.53$, $df=1$, $p=0.47$) for both tasks, while they appear to be significantly different ($F=80.92$, $df= 1$, $p<0.001$) for the fixations inside the ROI. The fixations inside the ROI are much longer for the free-looking task than for the scoring task, while the fixations outside the ROI last approximately just as long for both tasks. This suggests that not only the fixation location, but also the fixation duration is significantly affected by the task given to the observer.

Analyzing the viewing behavior over time reveals another interesting trend. Figure 5 shows the fixation-ratio as a function of viewing time. It is noteworthy that for both tasks the fixation-ratio peaks after about half a second, after which it decreases and stabilizes. Yet, the fixation-ratio remains higher for the free looking task than for the scoring task.

6. DISCUSSION

The results in Section 5 clearly show a difference in viewing behavior between the tasks of free looking and scoring. Figures 3 and 5 both show that the viewers tend to fixate more on the ROI when they are looking freely at the images. Additionally, Figure 5 also shows that this effect persists at least for the first five seconds of viewing which we examined in our analysis. This suggests that when scoring, the attention of the viewer is less focused on the natural ROI of the image. It has, instead, been adapted to the new task of judging image quality.

Fig. 5. The fixation-ratios plotted against the viewing duration for both scoring and free viewing tasks

As explained in Section 2, the images used in this experiment were specifically chosen to have a clearly defined ROI in the foreground and none in the background. Keeping that in mind while looking at Table 1, one may suggest that the fixation duration outside the ROI (the right column) is the fixation duration when randomly scanning the image without focusing on a specific object. This shows that whether scanning for interesting objects in the image (while free looking) or scanning for clues of image quality (while scoring), these scanning fixations have a similar duration. However, when looking inside the ROI there is a clear and statistically significant difference in the fixation duration. The viewers are more interested in the content of the ROI when free looking, spending a longer time on each fixation to understand and admire the image. While when scoring, they are less interested in the content of the image and are distracted by their search for clues to the quality level of the image.

On the other hand, even when scoring, the fixation duration is still higher inside the ROI. This shows that even though the scoring task does have a significant effect on the viewing behavior, people still spend more time on the interesting parts of the image than on the background region. In a similar way, Figure 5 shows that, regardless of the task, the ROI clearly grasps the people's attention in the first (500 ms) of viewing. The spike in the fixation ratio then gradually drops until it stabilizes a few seconds later. This means that observers still give more attention to the ROI of images even when scoring quality. To explain this, it may be helpful to look at earlier work [8] which claimed that when scoring image quality, the observer gives higher weight to the quality level of the ROI than to the quality level of the background of the image. This explains why the ROI keeps gripping the viewers' attention even while scoring.

7. CONCLUSIONS

In this paper we compared viewing behavior for the free-looking and quality-scoring tasks. By tracking the eye movements of the observer, we found significant differences in location and duration of their gaze. Therefore, although subjective scores are considered to be the ground truth of image quality assessment [1,2], one should keep in mind that the viewer is not looking in a natural way at the images while giving these scores. This should, in turn, put some reservations on how to go on and use these subjective scores in future research. Future work should also focus on finding new ways of measuring subjective quality without influencing the viewing behavior of the observers. For researchers interested in having a closer look at the data gathered in this experiment, a database of all the images used in this experiment together with the created saliency maps and original content can be downloaded from the Delft IQlab website [9].

8. REFERENCES

[1] H.R. Sheikh, M.F. Sabir and A.C. Bovik, "A statistical evaluation of recent full reference image quality assessment algorithms", IEEE Transactions on Image Processing, vol. 15, no. 11, pp. 3440-3451, Nov. 2006.

[2] H. Liu and I. Heynderickx, "Visual Attention in Objective Image Quality Assessment: based on Eye-Tracking Data", IEEE Transactions on Circuits and Systems for Video Technology

[3] YARBUS AL.1967 "Eye movements during perception of complex objects". In L. A. Riggs, Ed., Eye Movements and Vision, Plenum Press, New York, Chapter VII, 171-196.

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

[5] A. Ninassi, O. Le Meur, D. Barba, P. Le Callet and A. Tirel, Task Impact on the Visual Attention in Subjective Image Quality Assessment, Proc. European Signal Processing Conference (EUSIPCO'06), Florence, Italy, 2006.

[6] C. Vu, E.C. Larson, and D.M. Chandler, "Visual fixation patterns when judging image quality: effects of distortion type, amount, and subject experience," IEEE Southwest Symposium on Image Analysis and Interpretation, 2008.

[7] Recommendation BT.500-10, "Methodology for the subjective assessment of the quality of television pictures", ITU-R (2000).

[8] H. Alers, H. Liu, J. Redi and I. Heynderickx, "Studying the risks of optimizing the image quality in saliency regions at the expense of background content ", IS&T/SPIE Electronic Imaging 2010, Image Quality and System Performance VII, Jan 2010.

[9] H. Alers, H. Liu, J. Redi and I. Heynderickx, "TUD Image Quality Database: Eye-Tracking Release 2", http://mmi.tudelft.nl/iqlab/eye_tracking_2.html.