

**IMAGE
QUALITY
EXPERIENCE**

IMAGE QUALITY EXPERIENCE

Proefschrift

ter verkrijging van de graad van doctor
aan de Technische Universiteit Delft,
op gezag van de Rector Magnificus
prof. ir. K. C. A. M. Luyben,
voorzitter van het College voor Promoties,
in het openbaar te verdedigen op
dinsdag 30 juni 2015 om 15:00 uur

door

Hani Alers

ingenieur Media en Kennistechnologie
geboren te Farwania, Koeweit.

Dit proefschrift is goedgekeurd door de promotor:

Prof. Ingrid Heynderickx

Copromotor:

Dr. Judith Redi

Samenstelling promotiecommissie:

Prof. Henk Sips,	Technische Universiteit Delft, voorzitter
Prof. Ingrid Heynderickx,	Technische Universiteit Eindhoven, promotor
Dr. Judith Redi,	Technische Universiteit Delft, copromotor
Prof. Huib de Ridder	Technische Universiteit Delft
Prof. Mark Neerinx	Technische Universiteit Delft
Prof. Marcel Reinders	Technische Universiteit Delft
Prof. Wijnand IJsselsteijn	Technische Universiteit Eindhoven
Prof. Andrew Perkis	Norwegian University of Science and Technology

Thesis cover: the cover is an unmanipulated image of an LCD screen showing the title of the book where it is possible to see how RGB sub-pixels form different colors. The cover in a sense simulates the human perception of image quality where some properties of the original image are changed. This is due to the limitations of the capture and reproduction systems (i.e. camera and printer) in an analogy to limitations in the human visual system.

Copyright © 2015 by Hani Alers

ISBN 978-94-6186-499-4

An electronic version of this dissertation is available at <http://repository.tudelft.nl/>.

*To Ingrid Heynderickx,
for her unequivocal support*

Contents

Abstract	5
1. Introduction.....	7
1.1. Image Quality (IQ) Perception.....	8
1.2. IQ Assessment	10
1.3. Understanding Visual Attention (VA)	13
1.4. Visual Attention and Image Quality.....	19
1.5. Research questions and thesis layout.....	21
2. Effects of Task and Image Properties on Visual Attention Deployment in Image Quality Assessment.....	25
2.1. Introduction.....	26
2.2. Analyzing similarities in visual attention deployment: problem setup	29
2.3. Visual attention data collection	33
2.4. Analyzing similarity between saliency maps.....	39
2.5. Impact of experimental conditions on saliency similarity	42
2.6. Saliency changes with scoring task.....	53
2.7. Conclusions.....	55
3. Studying the Effect of Optimizing Image Quality in Salient Regions.....	61
3.1. Introduction.....	62
3.2. Experimental set-up.....	63
3.3. Experimental protocol.....	66
3.4. Results	72
3.5. Discussion.....	77
3.6. Conclusions.....	81
4. Examining the Effect of Task on viewing Behavior in Videos Using Saliency Maps...87	87
4.1. Introduction.....	88
4.2. Methodology	89
4.3. Analyzing the data	92
4.4. Results	93
4.5. Discussion.....	99
4.6. Conclusions	101
5. Quantifying the Importance of Preserving Video Quality in Visually Important Regions	105
5.1. Introduction	106
5.2. Methodology.....	110

5.3. Results.....	117
5.4. Discussion.....	124
5.5. Conclusions.....	131
References.....	132
6 Effect of Image Quality on Disaster Response Applications.....	137
6.1. Introduction.....	138
6.2. Methodology.....	139
6.3. The experimental protocol.....	142
6.4. Results.....	144
6.5. Discussion.....	145
6.6. Conclusions.....	146
7. Thesis discussion.....	149
7.1. Using eye tracking for visual analysis.....	150
7.2. Task effect on viewing behavior.....	153
7.3. Importance of ROI.....	154
7.4. Quality masking by task.....	156
8. Conclusions and Recommendations.....	157
8.1 Thesis conclusions.....	157
8.2 Thesis recommendations.....	157
Acknowledgments.....	161
List of Publications.....	163
Thesis Propositions.....	165
English list of propositions.....	165
Dutch list of propositions (proefschrift stellingen).....	166

Abstract

While the world we live in becomes more saturated with ubiquitous digital displays, and as the threshold for creating digital media continues to drop, image quality is an issue that concerns an increasingly large segment of the population. Higher resolutions, increased dynamic range, and faster frame rates put increasing demands on resources such as disk space and transmission bandwidth. Unfortunately, these resources are also needed for other functionalities of our digital devices and are often in short supply.

To find new ways to optimize the production pipeline of visual media while maintaining a good image quality, more knowledge is required about how we perceive visual content. In this work, we examine how a specific viewing task or content affect the viewing behavior of an observer. We then examine how localized differences in image integrity affect the overall perceived quality. From these results we gain knowledge on how image quality should be optimized for a given viewing behavior. In addition, we show that for specific tasks there is a limit to the required content integrity. We investigate these research questions empirically using eye tracking to scan in real time how the viewing behavior changes under different tasks and for different content, while one of the tasks involved scoring image quality.

Our results show that the viewing task and image content have a significant effect on the viewing behavior. We also find that the region of interest has a 5 times stronger effect on perceived quality in still images than the rest of the image. In videos, this effect is increased to 10 times. This finding can be utilized to optimize digital content once the region of interest is identified. We finally find that certain applications can mask degradations in image quality, making it redundant to allocate extra resources to maintaining content integrity.

1.

Introduction

With the rapidly accelerating advancements in multimedia technologies, we find ourselves increasingly surrounded by numerous images, apps, alerts, adverts, videos, and other stimulating items. All of these items are, in a sense, competing for some of our resources, such as our attention, money, time, or combinations thereof. Deciding which item(s) to attend to and which to ignore is quite a complicated process. Undoubtedly, one such element is the quality of items, since good quality tends to be appreciated by the receiver. When it comes to images and videos, “good quality” entails the visual integrity of the content and how purely it conveys its source in nature and/or the vision intended by the content creator. This is what this book refers to as Image Quality (IQ). Given this definition, we will examine why IQ is such an important notion, and explore how it interferes with visual attention.

1.1. Image Quality (IQ) Perception

As humans, we crave visual stimuli, when looking at a scene in real life, admiring a painting or a printed photo, or looking at an image or a video on an electronic display. When it comes to displays, IQ is not to be ignored. It has been shown before in a number of market studies that IQ (along with cost) is one of the top customer considerations in purchasing a product [1]. Achieving good IQ remains to be a moving target. If we take the displays of mobile phones as an example, it is obvious that a high-end (Nokia) phone with a monochromatic 96x60 pixel display was considered to have good IQ less than a decade ago. Today a typical mobile (smart) phone has a display resolution of 1920x1080 pixels capable of reproducing millions of colors, and still developments on improving IQ by means of, for example, using OLED displays are ongoing. With the advancements in display technologies and related changes in the multimedia supply chains, there are many new variables that affect the eventual images reaching the viewer. Examples include

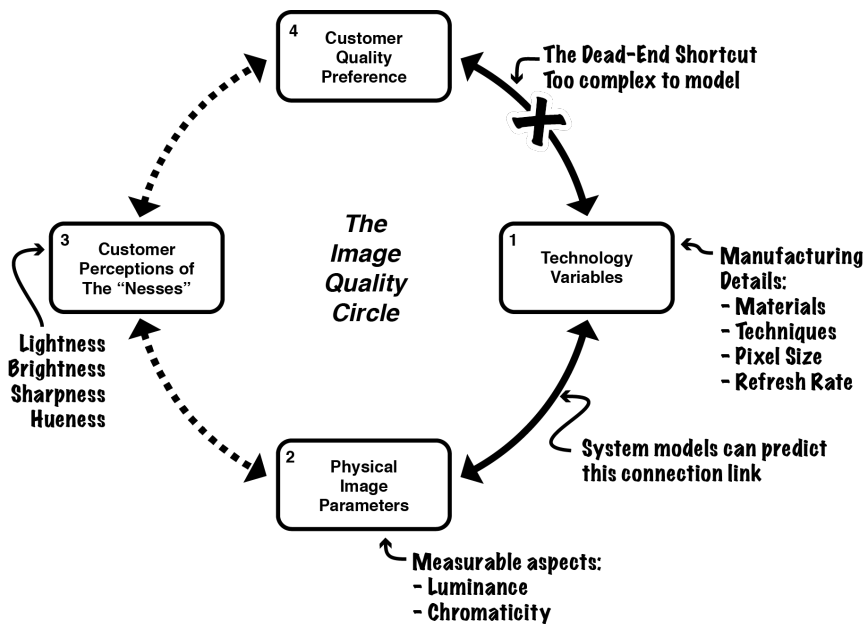


Figure 1.1 The Image Quality Circle breaks down the quality perception process into four different steps represented in the four boxes in the figure. Only the link between Steps 1 and 2 is well understood, while the other two require further research.

variables related to compression algorithms (e.g. JPEG, H.264, HEVC), transmission media (e.g. Internet streaming, Blue-ray disks), and display specifications (e.g. size, resolution, color depth).

It is vital here to highlight the concept of perception in relation to image quality. A lot of work has been put in defining parameters that could objectively describe an image. By calculating the amount of current going through an LED panel, we are able to accurately predict the amount of light it is going to produce, given that we know enough about its manufacturing specifications. We know exactly the resolution and the refresh rate of each display. Moreover, if we are in doubt of our calculations, we can use sophisticated measurement devices (such as colorimeters and microdensitometers) to measure the physical parameters of the images produced by these display devices. However, one should understand that this information only gives us a part of the information when we need to determine what perceived IQ is. In fact, at the point we have measured all the physical properties of the image, perception has not started yet.

Image quality appreciation starts with the image information entering the eye and ends up with the brain forming an opinion regarding the quality of a specific image. Initially, the human visual system (HVS) processes the image information in order to send it to the brain. In this stage, some details of the physical attributes are lost (or simplified) due to limitations of the HVS. For example, the finite resolution of the eye (determined by the number of cone/rod light receptors in the retina) is only capable of capturing details up to a specific limit [1]. Any details beyond that limit are simply filtered out. Similarly, when it comes to subtle differences in image aspects (e.g. brightness or contrast), the HVS can only discern these differences to a specific level of detail known as the Just Noticeable Difference (JND) [2]. Any variations within these JNDs, which may be easily measurable using instruments, are completely imperceptible by humans.

From the above discussion, it is clear that IQ is more than a collection of objectively measured technical specifications. Hence, a more human-centered approach is needed to improve imaging products in terms of IQ. As recently as the year 2000, Peter Engeldrum, in his book

“Psychometric Scaling”, pointed out the problems of a disorganized approach to researching IQ where there was even disagreement on whether perceived quality could be measured or only approximated. At the time, IQ research was largely driven by industry in a fragmented manner that “led to confusion and chaos” [1]. Engeldrum introduced the Image Quality Circle (shown in Figure 1.1) as a model to serve as a common starting point that researchers can refer back to when discussing their work. The circle represents the path between the technical variables of the image reproducing system (block 1 in the figure) and the eventual quality opinion of the observer (denominated “customer” in block 4). Engeldrum illustrates here that finding the relation between these two variables is too complex, and proposes another path with two intermediate steps (blocks 2 and 3). We already have a path to establish the physical image characteristics, though understanding how combinations of these characteristics are perceived in terms of attributes and eventually give a quality preference are still illusive affairs. One of the most important contributions that the Image Quality Circle model brought to the field was that it clearly put a distinction between the physical and technological aspects of the process and the human perception aspects. Looking at Figure 1.1, one can clearly see that the last two missing links in the circle are centered around the human element of the puzzle.

1.2. IQ Assessment

The interest in objective image quality assessment (IQA) has been growing at an accelerated pace over the past decade. Objective IQA measures aim to predict IQ perceived by human subjects, who are the ultimate receivers in most image processing applications. To evaluate the accuracy of such measures, large independent databases of images have been created where the quality of the images was scored by a number of observers [3,4]. By averaging the scores from all observers, each image received a, so called, Mean Opinion Score (MOS) which represents the ground truth for the quality of these images. The aim of all objective IQA metrics is to be able to predict the MOS of images as closely as possible.

Depending on the availability of a pristine reference image that is presumed to have perfect quality, IQA measures may be classified into

full-reference (FR), reduced-reference (RR), and no-reference (NR) methods [5-8]. FR measures require full access to the reference image, while NR methods assume completely no access to the reference. RR methods provide a compromise in-between, where only partial information in the form of RR features extracted from the reference image are available in assessing the quality of the distorted image. IQA measures may also be categorized into application-specific or general-purpose methods. The former only apply to some specific applications where the types of distortions are often known and fixed (e.g. JPEG compression). The latter are employed in general applications, where one may encounter diverse types and levels of image distortions.

A considerable number of IQA measures have been proposed in the literature, exhibiting substantial diversity in the methodologies used. Still, they also share some common characteristics. In particular, all of them are rooted from certain knowledge in one or more of the following three categories:

1. knowledge about the image source, which can be either deterministic (when the reference image is fully available) or statistical (when certain statistical image models are employed)
2. knowledge about the distortion channel, which is often associated with some known facts about the specific distortion process that the images underwent, for example, blocking and blurring artifacts in JPEG compression, and blurring and ringing effects in wavelet-based image compression
3. knowledge about the HVS, where computational models are developed based on visual, physiological, and psychological studies.

In general, the available objective IQA approaches utilize either signal fidelity measures (i.e., examining only numerical differences from the original content), or perceptual quality metrics (i.e., also taking into account aspects of the HVS). The signal fidelity measures include the traditional MSE (mean square error), PSNR (peak signal to noise ratio), or similar approaches [9]. These approaches are quite popular and widely used since they are simple, well defined, and have a clear numerical meaning. Some of these metrics have been used to evaluate the quality of picture transmission channels, such as throughput, jitter, noise, and

packet loss rates. However, the same transmission parameters may result in different degradation of visual content, and therefore different perceived IQ. As predictors to perceived IQ, these signal fidelity measures can perform quite poorly since they do not take into account any aspect of the HVS [10, 11]. Since perceived quality is determined by the viewer's perception, it is much more complex than the statistics that a typical network management system can provide. It has been well acknowledged that a signal fidelity measure does not align well with human visual perception of images and videos [9,12-15].

To get a better objective prediction of subjective visual quality scores, a new generation of perceptual quality metrics is being developed. Subjective IQ is a function of visual content where the change of predefined test signals through a system is not necessarily a reliable source of visual quality measurement. In spite of the recent progress in related fields, objective evaluation of IQ in line with human perception still has a great room for improvement [16-20] due to the complex, multidisciplinary nature of the problem. It combines challenges from the fields of physiology, psychology, computer science, and (most importantly) human vision. A better understanding of the HVS mechanisms, and the diversified scope of its applications and requirements, are key elements in improving perceptual quality metrics. Still, there has been some interesting advances in IQA methods lately. For example, a handful of objective IQA measures have been shown to significantly and consistently outperform MSE and PSNR in terms of correlations with subjective quality evaluations [15]. Until now, the area that has achieved the greatest success is FR IQA of gray-scale still images. Several newer algorithms [5-8,21,22], significantly outperformed MSE and PSNR in a series of tests based on several MOS rated image databases.

When it comes to NR metrics, we need a more clever approach than comparing the image to a reference. Most extensively developed in this area are algorithms to measure blockiness in compressed images or videos. One approach is to use a Fourier transform along the rows and columns to estimate the strength of the block edges of the image [23]. An alternative approach proposed a nonlinear-model for NR quality assessment of JPEG images, where the parameters of the model were

determined with subjective test data [19]. Vlachos used cross-correlation of subsampled images to compute a blockiness metric [24]. Some proposed NR metrics are based on computing gradients along block boundaries, where the block edge strength for each frame was computed [25]. The general idea behind such metrics is to evaluate the visibility of each block (artifact) edge. These approaches utilize the fact that the visibility of a block edge may be masked by more spatially active areas around it, or in regions of extremities in illumination (very dark or bright regions) [10, 26].

Just like with the FR metrics mentioned above, some NR metrics also attempt to improve their accuracy with a weighting function simulating visual attention based on proper ties of the HVS [25]. Therefore, in order to understand the value of these new approaches, it is useful to learn more about visual attention itself.

1.3. Understanding Visual Attention (VA)

In our everyday perception of our environment, we pay attention to some things and ignore others. We decide that it would be interesting or necessary to look here but not there. And as we shift our gaze from one place to another, we are doing more than just "looking": we are directing our attention to specific features of the scene in a way that causes these features to become more deeply processed than those features that are not receiving our attention. Most of the time we exhibit divided attention because we need to focus on a number of things at once. For example, as you drive a car, you need to simultaneously attend to the other cars around you, traffic signals, road signs and pedestrians, while occasionally glancing at the navigation system and checking your rear view mirror. Since no mortal entity is limitless, there is naturally a limited amount of attention a person can possess. Therefore, in order for a human to function properly, it follows that one has to be able to prioritize his attention on some stimuli while ignoring others.

One mechanism of selective attention is eye movements. By scanning a scene, the fovea is aimed at places we want to process more deeply. The human eye is constantly moving to take in information from different segments of a scene. A question that one can ask here is whether eye-

movements can be directly linked to attention. The answer to this question is, not always. For example, if you are trying to read a book you are not interested in (just to pass an exam), at some point, you become aware that although you are moving your eyes across the page and "reading" the words, you have no idea what you just read. So, even though your eyes were looking at the words, your attention was dedicated to something else. Still, despite possible misinterpretation, eye movements are the best (currently available) way for measuring where attention is allocated. It is therefore beneficial to examine it further and establish methods to measure and represent it.

One can wonder how human eyes actually scan the scene when looking at images. Tracking the eye movements can help us understand this process. To record eye movements, early researchers resorted to using devices such as small mirrors and lenses that were attached to the eyes [27]. However, modern researchers use camera-based eye trackers that track the position of the eye without attaching anything to the eye, for example by using regular cameras or by using light in the infrared



Figure 1.2. A representation of actual eye scan path (from eye tracking data) from an observer asked to look at the image casually. The arrows represent saccades where the eye is moving from one location to another. The circles represent fixations where the eye is focused on a specific part of the scene.

spectrum (invisible to the human eye) reflected at the retina. Such devices track the eye movements and remotely determine the gaze target, making the process far less intrusive.

Using an eye-tracker results in information as presented in Figure 1.2, overlaid with the original image the viewer was looking at. The eye movements shown by the arrows in Figure 1.2 are called saccades. The saccades are punctuated by pauses, indicated by the circles, where the eye stops momentarily to take in information about a specific part of the scene. These pauses, called fixations, indicate where the person is attending. What determines where we fixate in a scene? The answer to this question is complicated because our looking behavior depends on a number of factors, including characteristics of the scene and the knowledge and task of the observer.

Looking at the scene in Figure 1.3, certain areas stand out because they have high contrast, contain easily recognizable features (e.g. a face or a car), or entice the viewers curiosity like the clock showing the time the picture was taken. These areas have high saliency, as they attract attention based on their stimulus properties. This type of saliency usually is referred to as natural saliency, i.e., saliency measured when the viewer is looking to an image casually, without having a certain task or to respond to a certain question.



Figure 1.3. On the left is an image used for an eye tracking experiment. By averaging the saliency data from several observers and superimposing it as a heat map over the original image (right), it is possible to see where the salient regions of the image are located.

By using eye tracking equipment, it is possible to measure natural saliency. Figure 1.3 on the right shows the scene overlaid by a saliency map that shows which areas of the image are more salient than others. The figure visualizes the saliency map as a heatmap, though in pure terms, a normalized saliency map is a matrix of values between 0-1, that are associated to the corresponding pixel in the image and characterize its probability to be attended (i.e. to attract attention) by an average observer.

Previous work has already shown that the observer's task affects the visual attention deployment in a specific scene. The earliest example is the work of Yarbus [27] performed in 1967. It involved a series of recordings of observers viewing a painting called *The Unexpected Visitor* (Figure 1.4). Yarbus asked the same individual to view the painting seven times, each time with a different instruction before starting to view the image. These instructions asked the viewer to make a series of judgments about the scene depicted, to remember aspects of the scene, or simply to look at it freely. The data illustrated compellingly that simply altering the instructions given to the observer, and thus their task while

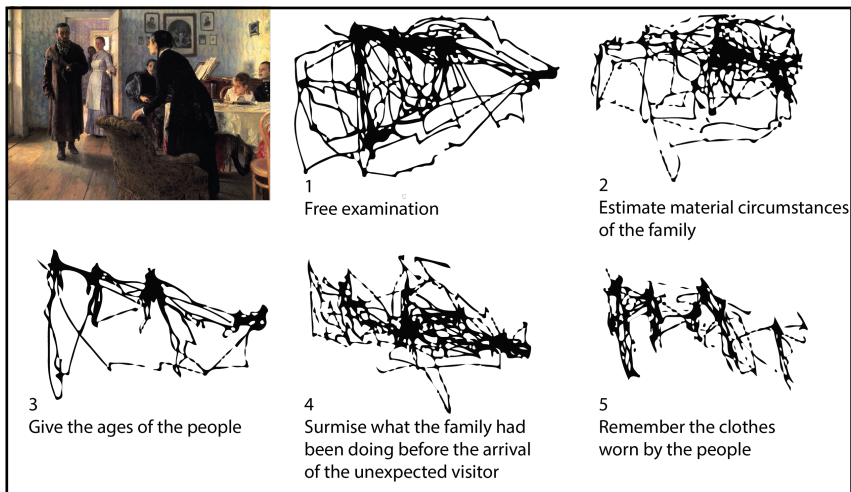


Figure 1.4. Yarbus [28] asked the same observer to look at the “unexpected visitor” painting several times, each for a duration of 3 minutes. The observer was given a different task before each viewing as described above. An eye tracking device allows us to see where the viewer was looking with each of the assigned tasks.

viewing, had a profound effect on the viewing behavior of the observer, as shown with the different eye movement paths in Figure 1.4. As Yarbus observed: “Depending on the task in which a person is engaged (i.e., depending on the character of the information which he must obtain) the distribution of the points of fixation on an object will vary correspondingly, because different items of information are usually localized in different parts of an object” [27]. This example shows that the demands of the task override the scene’s natural saliency.

Our ability to quickly comprehend a scene even when it is presented briefly or off to the side, is an important skill, but there is a great deal of evidence indicating that when it comes to determining specific details, focused attention is necessary. This has been illustrated in a number of ways.

There are even studies showing that task can completely block the perception of the observers. Arien Mack and Irvin Rock [28] demonstrated this effect using the procedure shown in Figure 1.5. The observer’s task is to indicate which arm of the cross is longer, the horizontal or the vertical. Then, after a few iterations of the trial, a small test object, which is within the observer’s field of clear vision, is added to the display. When observers are then given a recognition test in which they are asked to pick the object that was presented, they are unable to do so. This shows that concentrating their attention to the vertical and horizontal arms apparently made observers blind to the unattended test object.

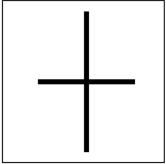
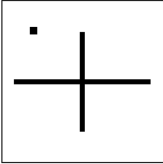
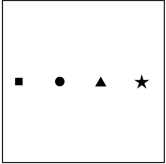
What is shown on screen		6 more trials ...		
Given task	Indicate longer arm: horizontal or vertical?		Indicate longer arm: horizontal or vertical?	Which object did you see?

Figure 1.5. Inattentional blindness experiment [29]: participants were asked to judge whether the vertical or horizontal arms are longer in each trial. After a few trials, a geometrical object appeared on the screen together with the arms. The participants showed difficulty in recalling the geometrical shape of the object.

In another example, Daniel Simons and Christopher Chabris [29] created a situation in which one part of a scene is attended and the other is not. They made a 75-second film that showed two teams of three players each. One team was passing a basketball around, and the other that was guarding that team by following them around as in a basketball game. Observers were told to count the number of passes, a task that focused their attention on one of the teams. After about 45 seconds, one of two events occurred. Either a woman carrying an umbrella or a person in a gorilla suit walked through the "game," an event that took 5 seconds.

After seeing the video, observers were asked whether they saw anything unusual happening or whether they saw anything other than the six players. Nearly half (46%) of the observers failed to report that they saw the woman or the gorilla. In another experiment, when the gorilla stopped in the middle of the action, turned to face the camera, and thumped its chest, half of the observers still failed to notice the gorilla (Figure 1.6). These experiments demonstrate that when observers are given a task that consumes their entire attention, they can fail to notice another event, even when it is right in front of them. This opens the door for many potential questions regarding attention, viewing task, and IQ perception.



Figure 1.6. By asking viewers to focus on the basketball being passed around in a video, some completely missed a man in a gorilla suit walking across the scene and pausing in the middle to beat his chest [30].

1.4. Visual Attention and Image Quality

As explained in Section 1.3, there is no doubt that VA plays a central role in the HVS. However, there are different (and sometimes contradicting) views regarding the role of VA in IQA. Some research has shown improvements in IQ prediction by incorporating VA information in their algorithms [30-32]. On the other hand, further research argues that applying VA data in perceptual quality metrics is not a trivial affair and requires some better understanding of how the HVS works [33]. So far, the mechanism for incorporating VA data in these IQA metrics has been to simply use the saliency map of the image as a weighting map for the IQA metric values. This meant that the quality of the areas of the image with a higher saliency value would have a higher contribution to the overall MOS value of the image.

Let us take the image in Figure 1.7. as an example. This particular image was created by applying a strong (lossy) JPEG compression to the original image resulting in a low bitrate file. Note how the sky in the background suffers from clear color banding artifacts that betrays the files low bitrate. However, the statue in the center of the image contains more details that mask the artifacts in the image and make them less noticeable. Assuming that the statue is the most salient region of an image, an IQA metric that uses simple weighting of VA data may give this image a high quality score despite its low bitrate. The artifacts in the sky will be weighed down by the low saliency they have. At the same time, assuming that the metric accounts for artifact masking, the statue area will have a higher quality score which will be weighed up by its high saliency. This results in a high overall quality score.

Here one starts to wonder, is it correct to give an image like the one in Figure 1.7. a high quality score? It is true that the statue is the most salient region in the image, while the artifacts in the sky are quite visible, and so also may attract attention. Will the viewers notice these artifacts even though they mostly give attention to the statue? And if they do, how much will that affect their judgment of the quality of the image? After all, the part of the image that they are most interested in seems to be in good quality. So will they give the image a high or low MOS? We simply do not know the answers to these questions. Additionally, this makes it clear that

a simple weighting of VA data in IQA metrics is quite a naive approach, which does not represent the complex process of subjective IQ evaluation.

Another important aspect that should be examined more carefully is how VA data pertain to the task of the viewers. As we saw in Section 1.3, the task of the observer can completely change his VA deployment on the same scene. So what type of VA data should be used for IQA? Some IQA metrics that reported improved results using recorded VA data (via an eye tracker), observed that greater improvement was found with VA recorded in task-free viewing than in the cases of subjects being asked to assess the picture quality [34]. On the other hand, all MOS scores collected by researchers have been collected while giving the viewers a task to score the quality of the images. Does that mean that we have changed the viewing behavior of the observers? And to what extent has their viewing behavior been changed? It is vital to examine these issues closely since they have a direct effect on the reliability of the MOS score databases which in turn form the basis for modeling and evaluating IQA

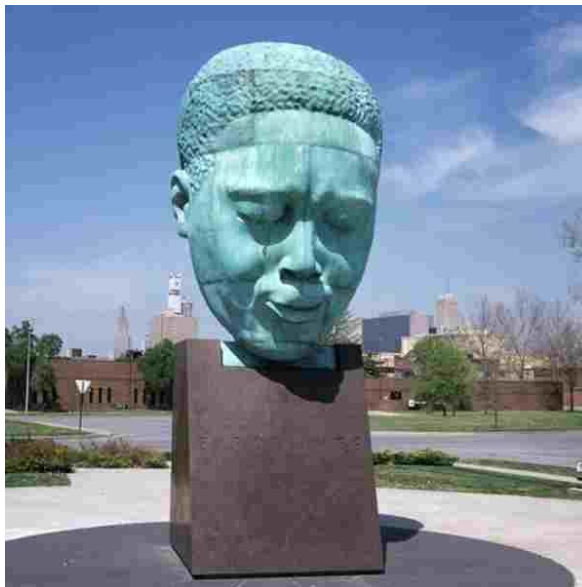


Figure 1.7. An image suffering from high JPEG compression resulting in a relatively low bitrate file. Note how the artifacts are more visible on the sky in the background, while the details on the statue make the artifacts less visible.

metrics. The answers to these questions will also help to guide the efforts in VA modeling. Most existing computational VA models are bottom-up (i.e. based upon contrast evaluation of various low-level features in images) in order to determine which locations stand out from their surroundings [35]. As to the top-down (or task-oriented) attention, there is still a call for more focused research, although some initial work has been done [36,37].

Besides the task of the observer, one should also consider how artifacts in the image can affect the VA (as we discussed with Figure 1.7). Some research has argued that distortions in image compression (e.g. with JPEG artifacts) and transmission (e.g. from packet loss) change the subject's eye fixation and the associated duration [38], while other work has indicated that there is no obvious difference in the saliency maps obtained for a distorted video sequence and its corresponding pristine version [39]. This shows that the influence of the stimulus IQ on VA is still an open issue for research as well.

1.5. Research questions and thesis layout

So far we have seen how important IQ is, and learned a bit about IQA metrics. We also found out that VA can help improve the performance of these metrics and had a look about its inner workings. Still, we observed that the relation between VA and IQ is not completely understood. Due to the vast complexity of the HVS and human perception, more work is needed to examine this relation. In the following we will formulate a few research questions that embody the direction this work is going to head towards. The research questions of this thesis are:

1. How does the task given to the observer and quality level of the stimulus affect their viewing behavior? And how is that different between images and videos?
2. How does the observer evaluate the overall quality of a stimulus if different parts of the scene convey a different level of quality? And how does that differ between images and videos?
3. Can the task given to the observer mask the perception of artifacts in the scene?

This thesis contains a collection of chapters (Chapter 2 - Chapter 6) that take us on a journey to examine these questions from different angles. We start with Chapter 2, which looks at how the task and quality level can affect VA in images. In Chapter 3 we examine how the global image quality of still images is determined when salient parts are shown at a different quality level than the background regions. Chapters 4 and 5 explore the same questions as Chapters 2 and 3 respectively, but in this case addressing video content. Great care and effort has been taken to keep the methodology and test equipment as similar as possible in order to be able to compare the results between still images and videos. Studying the relation between task load and artifact perception is handled in Chapter 6. Subsequently, reflecting back on all the previous chapters, Chapter 7 discusses the main findings and how they relate to each other. Finally Chapter 8 shortly presents the main conclusions of this thesis.

References

1. Engeldrum, P. G. (2000). *Psychometric scaling: a toolkit for imaging systems development*. Imcotek Press.
2. Qin, S., Ge, S., Yin, H., Xia, J., Liu, L., Teunissen, C., & Heynderickx, I. (2007). JND of Image Attributes for Natural Images. In *Proceedings SID*, 38(1), (pp. 326-329).
3. Delft Image Quality Lab "<http://ii.tudelft.nl/iqlab/>"
4. Laboratory for Video and Image Engineering "<http://live.ece.utexas.edu/research/quality/subjective.htm>"
5. Wang, Z., Bovik, A. C., Sheikh, H. R., & Simoncelli, E. P. (2004). Image quality assessment: from error visibility to structural similarity. *Image Processing, IEEE Transactions on*, 13(4), 600-612.
6. Lin, W., & Jay Kuo, C. C. (2011). Perceptual visual quality metrics: A survey. *Journal of Visual Communication and Image Representation*, 22(4), 297-312.
7. Hemami, S. S., & Reibman, A. R. (2010). No-reference image and video quality estimation: Applications and human-motivated design. *Signal processing: Image communication*, 25(7), 469-481.
8. Moorthy, A. K., & Bovik, A. C. (2011). Visual quality assessment algorithms: what does the future hold?. *Multimedia Tools and Applications*, 51(2), 675-696.
9. Eskicioglu, A. M., & Fisher, P. S. (1995). Image quality measures and their performance. *Communications, IEEE Transactions on*, 43(12), 2959-2965.
10. Karunasekera, S. A., & Kingsbury, N. G. (1995). A distortion measure for blocking artifacts in images based on human visual sensitivity. *Image Processing, IEEE Transactions on*, 4(6), 713-724.

11. Limb, J. O. (1979). Distortion criteria of the human viewer. *Systems, Man and Cybernetics, IEEE Transactions on*, 9(12), 778-793.
12. Girod, B. (1993, October). What's wrong with mean-squared error?. In *Digital images and human vision* (pp. 207-220). MIT press.
13. Mannos, J., & Sakrison, D. J. (1974). The effects of a visual fidelity criterion of the encoding of images. *Information Theory, IEEE Transactions on*, 20(4), 525-536.
14. Tian, D., & AlRegib, G. (2004, October). FQM: a fast quality measure for efficient transmission of textured 3D models. In *Proceedings of the 12th annual ACM international conference on Multimedia* (pp. 684-691). ACM.
15. Wang, Z., & Bovik, A. C. (2009). Mean squared error: love it or leave it? A new look at signal fidelity measures. *Signal Processing Magazine, IEEE*, 26(1), 98-117.
16. Eckert, M. P., & Bradley, A. P. (1998). Perceptual quality metrics applied to still image compression. *Signal processing*, 70(3), 177-200.
17. Pappas, T. N., Safranek, R. J., & Chen, J. (2000). Perceptual criteria for image quality evaluation. *Handbook of image and video processing*, 669-684..
18. Video Quality Experts Group. (2000). Final report from the video quality experts group on the validation of objective models of video quality assessment. *VQEG*, Mar..
19. Wang, Z., Bovik, A. C., & Lu, L. (2002, May). Why is image quality assessment so difficult?. In *Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on* (Vol. 4, pp. IV-3313). IEEE.
20. Winkler, S., & Mohandas, P. (2008). The evolution of video quality measurement: from PSNR to hybrid metrics. *Broadcasting, IEEE Transactions on*, 54(3), 660-668.
21. Chandler, D. M. (2013). Seven challenges in image quality assessment: past, present, and future research. *ISRN Signal Processing*, 2013.
22. Sheikh, H. R., & Bovik, A. C. (2006). Image information and visual quality. *Image Processing, IEEE Transactions on*, 15(2), 430-444.
23. Wang, Z., Bovik, A. C., & Evan, B. L. (2000). Blind measurement of blocking artifacts in images. In *Image Processing, 2000. Proceedings. 2000 International Conference on* (Vol. 3, pp. 981-984). IEEE.
24. Vlachos, T. (2000). Detection of blocking artifacts in compressed video. *Electronics Letters*, 36(13), 1106-1108.
25. Wu, H. R., & Yuen, M. (1997). A generalized block-edge impairment metric for video coding. *Signal Processing Letters, IEEE*, 4(11), 317-320.
26. Yuen, M., & Wu, H. R. (1998). A survey of hybrid MC/DPCM/DCT video coding distortions. *Signal processing*, 70(3), 247-278.
27. Yarbus, A. L. (1967). *Eye movements and vision* (Vol. 2, No. 5.10). L.

- A. Riggs (Ed.). New York: Plenum press.
28. Mack, A., & Rock, I. (1998). *Inattentional blindness*. The MIT Press.
 29. Simons, D. J., & Chabris, C. F. (1999). Gorillas in our midst: Sustained inattentional blindness for dynamic events. *Perception-London*, 28(9), 1059-1074.
 30. Liu, H., & Heynderickx, I. (2009, November). Studying the added value of visual attention in objective image quality metrics based on eye movement data. In *Image Processing (ICIP), 2009 16th IEEE International Conference on* (pp. 3097-3100). IEEE.
 31. Lu, Z., Lin, W., Yang, X., Ong, E., & Yao, S. (2005). Modeling visual attention's modulatory aftereffects on visual sensitivity and quality evaluation. *Image Processing, IEEE Transactions on*, 14(11), 1928-1942.
 32. Moorthy, A. K., & Bovik, A. C. (2009). Visual importance pooling for image quality assessment. *Selected Topics in Signal Processing, IEEE Journal of*, 3(2), 193-201.
 33. Ninassi, A., Le Meur, O., Le Callet, P., & Barba, D. (2007, September). Does where you gaze on an image affect your perception of quality? Applying visual attention to image quality metric. In *Image Processing, 2007. ICIP 2007. IEEE International Conference on* (Vol. 2, pp. II-169). IEEE.
 34. Larson, E. C., Vu, C., & Chandler, D. M. (2008, October). Can visual fixation patterns improve image fidelity assessment?. In *Image Processing, 2008. ICIP 2008. 15th IEEE International Conference on* (pp. 2572-2575). IEEE.
 35. Buschman, T. J., & Miller, E. K. (2007). Top-down versus bottom-up control of attention in the prefrontal and posterior parietal cortices. *science*, 315(5820), 1860-1862.
 36. Hopfinger, J. B., Buonocore, M. H., & Mangun, G. R. (2000). The neural mechanisms of top-down attentional control. *Nature neuroscience*, 3(3), 284-291.
 37. Navalpakkam, V., & Itti, L. (2006). Top-down attention selection is fine grained. *Journal of Vision*, 6(11), 4.
 38. Vu, E. C. L., & Chandler, D. M. (2008, March). Visual fixation patterns when judging image quality: Effects of distortion type, amount, and subject experience. In *Image Analysis and Interpretation, 2008. SSIAI 2008. IEEE Southwest Symposium on* (pp. 73-76). IEEE.
 39. Le Meur, O., Ninassi, A., Le Callet, P., & Barba, D. (2010). Do video coding impairments disturb the visual attention deployment?. *Signal Processing: Image Communication*, 25(8), 597-609.

2.

Effects of Task and Image Properties on Visual Attention Deployment in Image Quality Assessment

Abstract

It is important to understand how humans view images, and how their behavior is affected by changes in properties of the viewed images and the task they are given, particularly the task of scoring the image quality. This is a complex behavior that holds great importance for the field of image quality research. This work builds upon 4 years of research work spanning three databases studying image viewing behavior. Using eye tracking equipment, it was possible to collect information of human viewing behavior of different kinds of stimuli and under different experimental settings. This work performs a cross-analysis on the results from all these databases using state of the art similarity measures.

The results strongly show that asking the viewers to score the image quality significantly changes their viewing behavior. Also muting the color saturation seems to affect the saliency of the images. However, change in image quality was not consistently found to modify visual attention deployment, neither under free looking, nor during scoring. These results are helpful in gaining a better understanding of image viewing behavior under different conditions. They also have important implications on work that collects subjective image quality scores from human observers.

2.1. Introduction

Mean opinion scores (MOS) obtained in subjective image and video quality experiments are to date the only widely accepted measure of perceived visual quality in terms of reliability [1]. However, to make quality assessment practically implementable in real-life applications, e.g., in post-processing chains of television sets, objective quality models are needed [2]. These models usually start from image or video signal features and are then trained to predict the MOS, a process that may be improved by including human vision characteristics, such as masking or visual attention (VA). Focusing on the latter, many researchers have successfully shown an interaction between visual quality preferences and VA deployment [3, 4]. As a result, many attempts have been presented in literature to integrate VA information into objective quality metrics [5-15], yet with mixed results in terms of accuracy improvement.

In part, the lack of a clear consensus on the extent to which visual attention is beneficial to visual quality assessment may be related to the fact that different types of VA information have been used in the different studies. VA information obtained by tracking the eye-movements of people while either freely observing images (e.g., [9]) or scoring their quality [5] was interchangeably used in literature. Furthermore, visual attention data were either recorded or extrapolated through models (e.g. [16-18]) from both unimpaired and quality impaired images. Because of its intrinsic nature, both viewing task and quality level of the image may significantly alter the deployment of visual attention. Consequently, the type of information to be integrated in the quality metrics may be very diverse. It is therefore interesting to verify and quantify to what extent visual attention deployment changes depending on the experimental conditions under which it is captured. This work aims at doing so by analyzing visual attention data obtained through eye-tracking of image observation under a number of different tasks and visual quality conditions.

Visual attention is a prominent characteristic of the human visual system (HVS) and as such it has been investigated for a long time. When observing a scene, the human brain exploits visual attention mechanisms to reduce the complexity of the visual information to be processed by the

visual system [19]. Scene awareness is built by shifting the eye gaze from one part of the scene to the next, gradually learning more about it one piece at a time. Since the processing resources of the brain are limited, the visual stimuli are constantly competing for these resources and the most relevant stimuli in a given context are favored over the less relevant ones.

An effective tool for studying visual attention is eye tracking. Eye-trackers record eye movements of observers attending at scenes or images, delivering then a profile of the viewing behavior in the form of a collection of fixations and saccades. Fixations represent the viewing location at a moment when the pupil has seized to move, while a saccade is an abrupt, rapid, small movement of the pupil while the eye shifts the viewing location from one fixation to the next. The analysis of the duration of both fixation and saccades is already useful in the analysis of viewing behavior (e.g., [26, 31]). However, since visual perception is active only during fixations and is largely suppressed during saccades [20], often fixation data are further analyzed to better understand visual attention. Fixation paths [21] can reveal important insights in the spatial nature of visual attention deployment. The further transformation of fixation data into saliency maps [22, 23], representing the probability that a certain location in the image content gets attention, can also bring detailed information on the spatial deployment of visual attention. In particular, discrepancies in the saliency distribution between images as obtained under different viewing conditions (e.g., while scoring image quality or freely observing the image) can indicate dissimilarities in VA deployment due to the change in viewing condition. The analysis of eye-tracking data is therefore a rich source of information for our purpose to detect to what extent VA data collected under different tasks and visual quality conditions are (in)consistent.

Reasons for possible inconsistency can be found by looking at the basic functioning of VA. Two processes contribute to the deployment of visual attention: bottom-up attention and top-down attention. In general, bottom-up is rapid, saliency-driven, and task independent, while top-down is slower and task dependent [22]. Such dependency has been studied extensively in the past. Already in 1935 G. Buswell [24] proved, by means of eye tracking, that the task had a substantial effect on how viewers

looked at the image. Buswell even referred to comments by other researchers pondering over this issue as early as 1907 [25]. In 1967, a famous experiment by Yarbus involved asking an observer to look at the painting “The Unexpected Visitor” by I.E. Repin [21]. The observer was given 7 different tasks while looking at the painting and the eye movement patterns were recorded. The results showed a clear difference in viewing behavior. For example, when given the task of judging the ages of the people in the photo, the observer concentrated on their faces, while when asked about what they were doing the observer shifted the focus to what they were holding in their hands. These conclusions are not surprising since the given tasks basically convert the viewing process from bottom-up to top-down. In a similar way, a visual quality scoring task might alter the natural deployment of visual attention when observing an image. Some evidence in this sense has been shown already [26]. However, no consistent meta-analysis has been carried out so far across data collected from different experiments that shows the effect of a visual quality scoring task on (top-down) visual attention.

Visual attention mechanisms might be altered by the presence of impairments in the image as well. Bottom-up attention is deployed in the very first stage of the observation of a scene, and drives the selection of eye gaze locations (fixations) based on the visual (physical) characteristics of the scene. Color, texture and motion contrast strongly influence this selection, in a way that is largely independent of the semantic value of the elements placed at that location [22]. Visual impairments due to signal distortions (e.g., blockiness due to compression, noise) introduce singularities in the image; as a consequence, it is possible that their visible presence alters the natural VA deployment, and the resulting saliency distribution. A few studies have reported preliminary information in this sense, yet without a clear consensus. In the work by Vuori and others [27] the quality of the judged image was shown to have an impact on the saccades’ duration. Researchers in [28] and [29] looked at this aspect from the computational saliency point of view. Some researchers [3] showed that saliency maps of unimpaired images obtained from free-looking eye-tracking data were poorly correlated to the maps derived from the image quality scoring of slightly impaired versions of the same images. This effect was shown, though, to decrease with the increase of the amount of impairment visible in the images. In videos, quality was shown to have an impact on the

dispersion of fixations (increasing with the decrease of video quality) and to be positively correlated with the duration of the fixations [30]. Furthermore, the correlation among saliency maps corresponding to eye-movements recorded while scoring videos affected by packet-loss artifacts was found to increase along with the quality [31], whereas this was not the case for videos affected by blocky artifacts only [32]. In general, no clear conclusion can be drawn from the above studies.

A factor contributing to the lack of agreement in the results presented above could be the fact that different indicators were used to detect an effect of visual quality and viewing task on visual attention (correlation of saliency maps, duration of fixations, dispersion of fixations, etc.). This research builds further on earlier work [3, 9, 23, 33] and aims at analyzing the impact of task and quality on visual attention, by (1) using a collection of databases of VA data collected at Delft University of Technology over four years of research and (2) investigating trends and attention deployment shifts through a large and consistent set of saliency similarity measures [23]. By using four different similarity measures and a collection of diverse datasets, we aim as well at comparing the soundness of the different similarity indicators and at giving recommendations on which to use to more precisely unveil trends in visual attention data.

The remainder of this chapter is organized as follows. Section 2.2 describes the problem and data analysis setup. Details on experimental visual attention data collection are provided in Section 2.3. Section 2.4 gives more details on how we implemented the similarity measures to compare the saliency maps. Section 2.5 starts with an overview of the results using different similarity measures. These results are then used to closely examine how scoring task and quality losses affect visual attention deployment. Section 2.6 looks again at the data using different analysis techniques to discern how task changes viewing behavior. Finally, the conclusions of this research are summarized in Section 2.7.

2.2. Analyzing similarities in visual attention deployment: problem setup

To analyse the effect of factors such as task and visual quality on visual attention, we first define the concepts of *reference* (control) and *test*

viewing situation. In a typical experiment, eye movements of a number of observers are first recorded for different images in a *reference* setting, e.g., during task-free image observation. Then, one or more factors are introduced to modify the reference condition (e.g., the viewing task, an impairment of the visual quality of the images used in the reference condition, or a combination of the two), and the eye movements are recorded again with the new setup. We will refer to this experimental condition as the *test* condition.

To observe the effect of a (set of) factor(s) on visual attention, we process two collections of eye-movement data recorded via an eye-tracker. Given a set of images, in the most general setting we have, for every image I in the dataset, a collection of eye movement data $\mu_R^{(I)}$, recorder under the reference condition, and a collection of eye movement data $\mu_T^{(I)}$, obtained under the test condition, i.e. under the effect of the factor(s) of interest. We then study similarities between the two collections μ_R and μ_T . This can be approached in multiple ways, from the analysis of frequency of fixations and saccades [26, 28, 34] to a more complex analysis of the spatial deployment of fixations. In this study, we privilege the latter, for two main reasons: (1) a spatial analysis can reveal shifts in the locations attended, perhaps due to the presence of quality impairments, and (2) VA information is often integrated in objective quality metrics as a local weighting factor (pixel-by-pixel or region-by-region) for metric values [5-15]; as a consequence, its spatial distribution is of major interest for visual quality research.

We study the spatial deployment of visual attention by means of saliency maps. These maps [22] are a visual representation of the probability that a location of the scene is attended by the average observer. Although originally intended to represent spatial deployment of bottom-up visual attention, in this study we are going to use the term “saliency map” to indicate the distribution of gaze probability resulting from bottom-up and top-down attention jointly. To create saliency maps from the raw eye-tracking data $\mu^{(I,k)}$, $k = 1, \dots, K$, each corresponding to a different image I and observer k , the following procedure can be applied:

1. Extract the set of fixation locations on the image $\mathbf{F}^{(l,k)} = [(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)]$, where n is the number of fixations included in $\mu^{(l,k)}$

2. Create the fixation map for observer k

$$FM^{(l,k)}(x, y) = \begin{cases} 1 & \text{if } (x, y) = (x_i, y_i) \in \mathbf{F}^{(l,k)} \\ 0 & \text{otherwise} \end{cases}$$

3. Create a global fixation map over all observers:

$$FM^{(l)}(x, y) = \frac{1}{K} \sum_{k=1}^K FM^{(l,k)}(x, y)$$

4. Apply a Gaussian patch having a standard deviation σ of the amplitude of the fovea (about 2° of visual angle) to each fixation point in $FM^{(l)}(x, y)$ to obtain the saliency map, $SM^{(l)}(k, l)$:

$$SM^{(l)}(k, l) = \sum_{j=1}^T \exp\left[-\frac{(x_j - k)^2 + (y_j - l)^2}{\sigma^2}\right], \quad T = \{(x, y) | FM^{(l)} \neq 0\}$$

where T is the total number of fixations over all observers. Note that in this formulation, no temporal information is considered (e.g. the duration of the fixations or their order).

As a result, each element of $SM^{(l)}$ expresses the probability that the average observer attends location (k, l) in image l over the observation period. Thus, given two saliency maps $SM_R^{(l)}(x, y)$ and $SM_T^{(l)}(x, y)$, the impact of a specific (combination of) factor(s) on visual attention can be assessed by evaluating (dis)similarities among the two distributions $S(SM_R^{(l)}(x, y), SM_T^{(l)}(x, y))$

By now, many ways to quantify similarity among saliency maps have been reported in the literature [23, 34, 35, 36]. Some widely used examples of those so-called similarity measures are: linear correlation coefficient (LCC), Kullback-Leibler divergence (KLD), normalized scanpath saliency (NSS) [34], and structural similarity index (SSIM) [37]. The LCC is traditionally the most commonly used measure. A value of $LCC = 1$ indicates identical maps, while $LCC = 0$ indicates uncorrelated maps. The same holds for the SSIM measure. This measure was originally introduced as a full-reference objective quality metric, but has recently shown its merits in comparing saliency maps as well [23]. The rationale behind its usage lays in the ability of SSIM to capture structural

Table 2.1. An overview of all the data used in this research with details about the number of participants and the experimental setup. In terms of databases: LC refers to the “complete LIVE dataset”, LS to the “LIVE subset dataset” and RS to the “ROI image set”. In terms of viewing tasks: FL refers to “free looking” and SC to “scoring”. In terms of stimuli quality and attributes: ORIG refers to the “original stimuli”, BW to “grayscale stimuli”, DIS to the “distorted stimuli”, LQ to “low quality”, MQ to “medium quality” and HQ to “high quality”. Note that the data set RS FL DIS LQ is used once as test data (i.e., Test-Data-7) and once as reference data (i.e., Reference-Data-4)

Eye Tracking Data	Participants	Stimuli	Stimuli Quality or Attributes	Task
Reference-Data-1: LC FL ORIG	20	29 images from LIVE Database	Full quality	Free Looking
Test-Data-1: LC FL BW	20	29 images from LIVE Database	Grayscale converted	Free Looking
Test-Data-2: LC SC DIS	20	29 images from LIVE Database	JPEG distorted	Scoring Quality
Test-Data-3: LC FL DIS	10	29 images from LIVE Database	JPEG distorted	Free Looking
Reference-Data-2: LS FL ORIG	20	6 images from LIVE Database	Full quality	Free Looking
Test-Data-4: LS SC DIS LQ	14	6 images from LIVE Database	Heavy JPEG, Blur, or noise distortions	Scoring Quality
Test-Data-5: LS SC DIS MQ	14	6 images from LIVE Database	Medium JPEG, Blur, or noise distortions	Scoring Quality
Test-Data-6: LS SC DIS HQ	14	6 images from LIVE Database	Slight JPEG, Blur, or noise distortions	Scoring Quality
Reference-Data-3: RS FL DIS HQ	10	40 Images with clear ROI	JPEG distorted	Free Looking
Test-Data-9: RS SC DIS HQ	20	40 Images with clear ROI	JPEG distorted	Scoring Quality
Test-Data-7: RS FL DIS LQ	10	40 Images with clear ROI	JPEG distorted	Free Looking
Reference-Data-4 RS FL DIS LQ	10	40 Images with clear ROI	JPEG distorted	Free Looking
Test-Data-8: RS SC DIS LQ	20	40 Images with clear ROI	JPEG distorted	Scoring Quality

similarity among two images. In comparing saliency maps, we are interested in checking whether the structure of the saliency distribution has changed: in this sense, SSIM offers a valuable tool to quantify the extent to which the structure of the saliency in $SM_R^{(l)}(x, y)$ is preserved in $SM_T^{(l)}(x, y)$. The NSS returns a value greater than zero if the correspondence between two saliency maps is greater than what can be

expected by chance. A NSS value of zero means that there is no such correspondence, and a value of less than zero means that there is anti-correspondence between the saliency maps. Finally, the KLD is a measure of divergence of two distributions. The further away from zero the value is, the more dissimilar are two maps, in this particular case the maps being two-dimensional distributions of saliency. More details on the definition of these measures and how to calculate them can be found in the literature [23]. Each of these similarity measures has its advocates, but so far evidence in literature is too limited to clearly favor one similarity measure over the others. They capture different properties while being coherent in predicting the similarity between saliency maps [38]. For this reason, all measures are deployed in our analysis to give a multifaceted yet consistent analysis of effects of task and visual attention deployment throughout different eye-tracking databases. For the further investigation with a large-scale analysis, as a convention in the literature [3], [38], SSIM and NSS are employed.

2.3. Visual attention data collection

Our analysis is deployed on an ensemble of three databases that spans a wide range of stimuli and test conditions. All data used in this study can be retrieved from the Delft Image Quality Lab repository [39]. Although the environmental conditions differed slightly from one experiment to the next (small variations in lighting condition and viewing distance, exact form of the scoring scale), all experiments were conducted in the same lab, using the same equipment. In this section, we describe the equipment used to collect the data as well as the experimental methodology used to collect each of the three databases. An overview of all data used and the details about the experimental setup are given in Table 2.1; samples of the image content are shown in Figure. 2.1.

2.3.1 The eye tracker and related equipment

All experiments were carried out with a SMI REDIII camera at sampling rate of 50 Hz and a tracking resolution of ± 0.1 deg. The iView X system developed by SMI provided the framework for the data recording, and the stimuli were shown via the Presentation Software from NeuroBehavioral Systems. During the experiment, viewers were asked to place their head

on a head rest in order to avoid head movements and get the highest accuracy. The height of the head rest was adjusted to suit the viewer and ensured a comfortable and non-confining seating position while performing the experiment. The stimuli were displayed on a CRT monitor with a resolution of 1024x768 pixels and an active screen area of 365x275mm. In order to avoid outside elements interfering with the results, the experiment was carried out in the User-Experience Lab of Delft University of Technology, which provides an experimental environment compliant to ITU BT.500 recommendations [40].

2.3.2 General experimental protocol

The protocol essentially consisted of a short introduction, after which the eye tracking system was calibrated by means of a 13-point calibration grid. For experiments involving multiple viewing sessions, calibration was repeated at the beginning of each new session. In all experiments, participants were briefed about the intent of the experiment and then went through a short training session, showing the participants a few example pictures and asking them to score a few stimuli when



Figure 2.1. Subsets of the images involved in the eye-tracking experiments. LC (also named as complete LIVE dataset) is based on the entire 29 original images of the LIVE database [41]; LS (also named as LIVE subset dataset) is based on 6 original images chosen from the LIVE database; RS (also named as ROI image set) is based on 40 original images with clear region of interest [30].

appropriate. In the experiments collecting free looking VA data, participants were asked to look at the images as they would do when exploring new image content. The images were simply displayed one after the other for a time of 8 to 10 seconds (as will be detailed later in this chapter). All scoring sessions followed a no reference, single stimulus protocol [40]. In all experimental sessions, the images were shown to each participant in a different random order. Table 2.1 illustrates an overview of all the data used in this research, including details about the number of participants, the stimuli and the viewing tasks.

2.3.3 Complete LIVE dataset (LC)

The first dataset we examine is the LC data, also known as TUD LIVE eye-tracking database [9]. The dataset includes eye-tracking data in the form of saliency maps corresponding to the observation of the 29 original images provided by the LIVE database [41] under different experimental conditions, namely:

1. Reference-Data-1: LC FL ORIG free looking of the unimpaired images
2. Test-Data-1: LC FL BW free looking of grayscale versions of the unimpaired images
3. Test-Data-2: LC FL DIS free looking of impaired versions of the unimpaired images
4. Test-Data-3: LC SC DIS image quality scoring of impaired versions of the unimpaired images

2.3.3.1 Data collection for unimpaired images (LC FL ORIG and LC FL BW)

Forty students, being twenty-four males and sixteen females, inexperienced with eye-tracking recordings, were recruited as participants. After passing the Ishihara Test for Color Blindness, they were assigned to two groups of equal size, each with twelve males and eight females. The unimpaired versions of the LIVE images were evaluated along with grayscale versions of the same images; the latter were obtained after gray-scale conversion, using only the luminance-component of the original content. The test stimuli were divided over two

groups of mixed original and grayscale images; thereby each group saw half of the images in a random order. Each stimulus was shown for 10s followed by a mid-gray screen during 3s. The participants were requested to look at the images in a natural way (“view it as you normally would”).

2.3.3.2 Data collection for impaired images (LC FL DIS and LC SC DIS)

To create the images employed in conditions (3) and (4), the unimpaired images were altered by applying JPEG compression with the quality parameter Q varying between 5 and 40, to cover a broad range of quality. To avoid that the recorded saliency was biased by viewing a scene multiple times, each of the unimpaired images was compressed at only one particular quality level (see [9] for more detail). A group of 20 observers, consisting of 12 males and 8 females, scored the quality of the impaired images on an ACR scale [40]. Since all unimpaired images were only compressed at one particular quality level, they were viewed only once per subject, and for each subject in a different random order. Each image was shown for a fixed time of 10s.

2.3.4 LIVE SUBSET dataset (LS)

The second dataset we consider was designed to evaluate the effect of different types of visual impairments on viewing behavior. The TUD interactions dataset [3], here referred to as LIVE SUBSET or LS, includes various distorted versions of 6 original images chosen from the LIVE Database Release [41] (see Figure 2.1). These images were impaired by applying three types of distortions, namely JPEG compression, White noise and Gaussian Blur; for each original image and distortion type, three different quality levels were selected.

A total of 14 observers assessed all images in the dataset, while scoring their quality using the Single Stimulus method with a continuous numerical scale [40]. The scoring scale ranged from 0 to 10, where “0” represented very low quality and “10” indicated very high quality. The observation time was not constrained. To limit memory effects, all 54 stimuli were divided in 18 groups of 3 stimuli, sharing both content and distortion type, and so differing only in quality level. The experiment was then structured in 3 sessions, where in each session only one image from

each of the 18 groups was selected. The selection was done randomly, but in such a way that each of the 3 images per group was presented in only one of the 3 sessions. In the remainder of this paper, images and corresponding eye-tracking data from the LS dataset are divided into three test conditions according to the three quality levels of the images, i.e.:

1. Test-Data-4: LS SC DIS LQ low quality, for which images present strong visible impairments
2. Test-Data-5: LS SC DIS MQ medium quality, for which images present moderately visible impairments
3. Test-Data-6: LS SC DIS HQ high quality, for which the quality of the images is just minimally compromised

Each of the three subsets includes 18 images. As the experiment itself did not include a free looking session to be used as reference condition, in the following analysis we use a subset of Reference-Data-2: LC FL ORIG, namely that part including eye-tracking data obtained from free looking at the 6 unimpaired contents from which the images included in LS were derived (see Table 2.1).

2.3.5 ROI image set (RS)

The ROI (Region of Interest) dataset (also known as TUD Task Effect dataset [33]) was built to study the joint effects of quality and task on visual attention. The stimuli used in the experiment were created from 40 original images. All images in this database were chosen to contain a clear ROI in the form of a human face, an animal, or an object that clearly stood out from the rest of the image. The size of the images was 600 × 600 pixels (see examples in Figure 2.1). Each image was processed to produce 4 different versions, for a total of 160 stimuli. The four different versions of each image were obtained by compressing it (through JPEG compression) at four different quality levels, with quality factors ranging between Q=10 and Q=100. Two different experiments were carried out based on this image material, the first one involving free looking of the stimuli, and the second one involving a quality scoring task.

2.3.5.1 Free looking data collection (RS FL DIS HQ and LQ)

The free-looking experiment had a total of 40 participants. They 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. Participants were given a fixed observation time of 8 seconds. In the following, the data derived from the free-looking observation of the stimuli is divided into two subsets, according to their compression level:

1. Reference-Data-3: RS FL DIS HQ, including eye tracking data corresponding to the two versions of each image with the two highest quality levels
2. Test-Data-7: RS FL DIS LQ, including eye tracking data corresponding to the two versions of each image with the two lowest quality levels

2.3.5.2 Image quality scoring data collection (RS SC DIS HQ and LQ)

20 participants took part in the experiment, each of which judged all 160 stimuli. The experiment was split in 4 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 (as in the data collection for the LS dataset). The order in which the stimuli were shown in each session was also chosen randomly by the system. Images were rated on a 10-point continuous quality scale (identical to that used for collecting the LS SC DIS sets), and the viewing time was unconstrained. As for the free-looking data, the scoring data for the RS image set is divided into two subsets, according to their compression level:

1. Test-Data-9: RS SC DIS HQ, including eye tracking data corresponding to the two versions of each image with the two highest quality levels

2. Test-Data-8: RS SC DIS LQ, including eye tracking data corresponding to the two versions of each image with the two lowest quality levels, hereafter indicated as low quality images (LQ)

As a result, RS SC DIS HQ and RS FL DIS HQ hold data corresponding to the same images, but in the first case collected during a scoring task, and in the second during free looking of the images. The same holds for RS SC DIS LQ and RS FL DIS LQ.

2.4. Analyzing similarity between saliency maps

2.4.1 Setting a benchmark: the saliency Empirical Similarity-Limit

To measure the effect of changes in visual attention with, for example task, a benchmark is needed. How dissimilar do two saliency maps need to be to prove that the change in testing condition had an effect on the viewing behavior? As a term of comparison, we use in this paper the Empirical Similarity Limit (ESL, [23, 42]). The idea behind the establishment of an ESL is that, even under the same viewing conditions and looking at the same image, there are always differences in how individuals deploy their visual attention. So, if we record the eye-movements of two groups of people under the same experimental conditions, the resulting saliency maps will not be identical, due to inter-observer variability. We can, however, set the similarity of those saliency maps, collected under the same conditions, as an upper limit for saliency map similarity, accounting in this way for inter-observer variability.

To determine the Empirical Similarity Limit, we took the eye-tracking data of all observers for a given (reference) condition, and divided them into two disjoint groups, each containing the data of half of the observers. Then, for each group, we computed the corresponding saliency maps for all images, following the procedure outlined in Section 2.2. This produced two saliency maps per image (and reference condition), each representing the saliency distribution of the image as observed by one of the two subgroups of observers. We then measured the similarity between these two maps with the four similarity measures described in section 2.2: LCC, KLD, SSIM and NSS. In this way, we could measure how dissimilar the viewing behavior was because of inter-observer

variability, and use it as an empirical limit for (dis)similarity between saliency maps recorded under different experimental conditions. This process was repeated 50 times by randomly changing the composition of the subgroups, in order to ensure robustness of the estimate. By calculating the average similarity value over the 50 runs and the confidence interval for each of the images, an Upper Empirical Similarity Limit (UESL) [23] was obtained. This UESL value equals the limit of similarity the saliency data can achieve without a change in viewing conditions. Note that, since the KLD has a reversed scale, the Similarity Limit value in this case is the lowest value that the similarity data is expected to reach. This KLD limit is therefore called the Lower Empirical Divergence Limit (LEDL). In the rest of this paper UESL and LEDL are referred to as similarity limits.

2.4.2 Calculating similarity for test data

In order to measure the difference in viewing behavior between test data (data collected when some experimental factors were changed) and the corresponding reference data, we computed, per image, the similarity between the saliency map obtained under reference and test conditions. When both the reference and the test saliency maps were obtained based on the same number of participants, the similarity values could be calculated directly. However, when one dataset was based on more participants than the other (e.g., LS FL ORIG based on 20 participants and LS SC DIS LQ based on 14 participants), the comparison between saliency maps would be unfair. As a consequence, saliency maps for the dataset with a higher number of participants were created by randomly selecting a subsample of participants, so that their number was equal to that of the participants in the other condition. In the case of the LS datasets, therefore, the saliency maps for the reference conditions (LS FL ORIG) were computed based on a random subset of 14 participants out of the initial 20. This process was then repeated 10 times with 10 random subsets, and an average value for the similarity was used in order to avoid any bias by chance in the data. Table 2.2 lists all the pairs of reference data versus test data analyzed in this study. The table also identifies which factors the comparison aims to examine. It should be noted that the comparison of saliency between free looking and scoring on the reference data is not

obvious from Table 2.2, since the participants were not deliberately requested to rate the quality of the original (full quality) images. This comparison, however, could be reasonably speculated from the comparison between Test-Data-6(LS SC DIS HQ) and Reference-Data-2 (LS FL ORIG), and between Test-Data-9 (RS SC DIS HQ) and Reference-Data-3 (RS FL DIS HQ), where high quality images without visible artifacts were used.

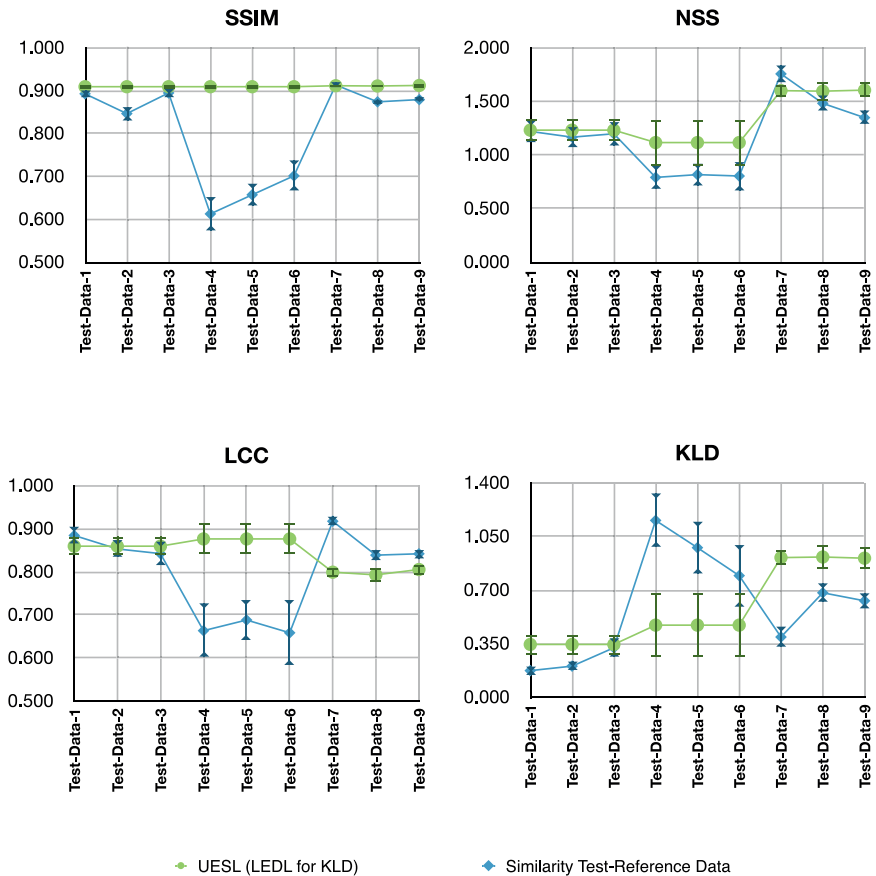


Figure 2.2. An overview of the similarity analysis for all tested subsets using all four similarity measures (i.e., SSIM, NSS, LCC and KLD); the green line (with circles) represents the Similarity Limit based on the reference data, while the blue line (with diamonds) represents the similarity of the test data with the reference data. Higher values represent higher similarity except for the KLD where the reverse is true. The error bars indicate the 95% confidence interval.

2.5. Impact of experimental conditions on saliency similarity

2.5.1 Overview of results and similarity measures

Figure 2.2 gives an overview of the similarity between reference and test data for all combinations mentioned in Table 2.2. Points in the graphs represent the averaged value of the similarity over all images in that specific set, while the error bar represents the 95% confidence interval. The two values given for each point on the horizontal axis represent the ESL based on the reference data (green circles) and the similarity value between the test data and its corresponding reference data (blue diamonds, see Table 2.2). The green circle points represent therefore how similar saliency maps collected under identical conditions are, while the blue diamond points show similarity resulting from comparing two different sets of saliency maps, one of which collected under test conditions.

In principle, it should be expected to have the blue diamond points either overlapping or below the green circle points (except for the KLD since it is inverted), for which we expect the test data to diverge more from the

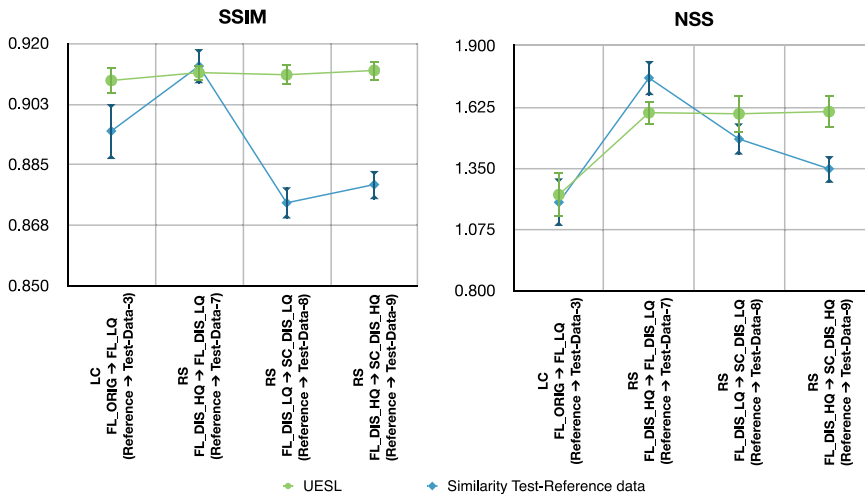


Figure 2.3. Similarity analysis (based on SSIM and NSS) of test conditions for which only quality of the images (Test-Data-3 and Test-Data-7) or viewing task (Test-Data-8 and Test-Data-9) were changed when collecting the VA data. The error bars indicate the 95% confidence interval.

Table 2.2. An overview of the tested effect, including a list of all tested-datasets and their corresponding reference-dataset, and the difference in task and/or distortion between both datasets.

Tested-Data	Reference-Data	Tested effect
Test-Data-1: LC FL BW	Reference-Data-1: LC FL ORIG	Colored versus grayscale images
Test-Data-2: LC SC DIS	Reference-Data-1: LC FL ORIG	Task + JPEG compression quality
Test-Data-3: LC FL DIS	Reference-Data-1: LC FL ORIG	JPEG compression quality
Test-Data-4: LS SC DIS LQ	Reference-Data-2: LS FL ORIG	Task + Quality level with miscellaneous distortions
Test-Data-5: LS SC DIS MQ	Reference-Data-2: LS FL ORIG	Task + Quality level with miscellaneous distortions
Test-Data-6: LS SC DIS HQ	Reference-Data-2: LS FL ORIG	Task + Quality level with miscellaneous distortions
Test-Data-7: RS FL DIS LQ	Reference-Data-3: RS FL DIS HQ	JPEG compression quality
Test-Data-8: RS SC DIS LQ	Reference-Data-4: RS FL DIS LQ	Task
Test-Data-9: RS SC DIS HQ	Reference-Data-3: RS FL DIS HQ	Task

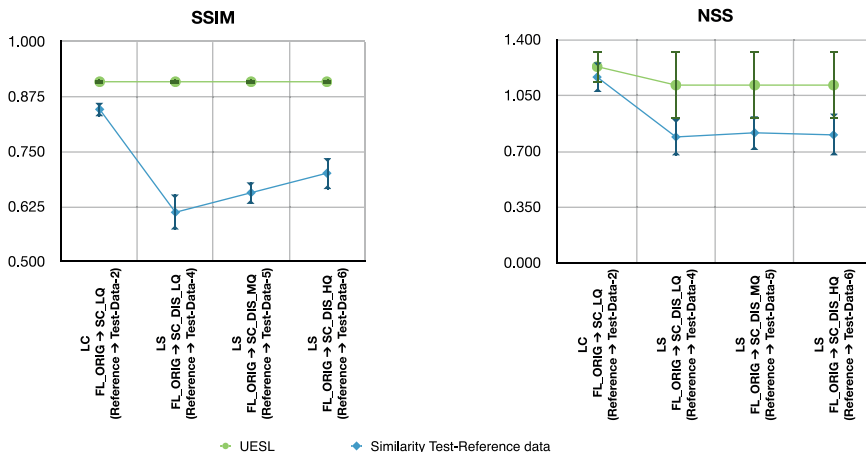


Figure 2.4. Similarity analysis (based on SSIM and NSS) of the Test-Data sets obtained by varying both task and quality level in the experimental conditions. The error bars indicate the 95% confidence interval.

reference data than the reference data diverge from itself. However, the results obtained with the LCC and KLD similarity measures indicate that for some comparisons (e.g. for Test Data 7, 8 and 9) the reference data are less similar to themselves than to the test data. It is obvious that this result is unlikely a true representation of how the viewing behavior is affected by the test condition, since the tested factor is not likely to make the similarity between the test data and the reference data higher than the theoretical upper limit as defined as UESL in [23]. These unreasonable findings in LCC and KLD can be considered as noise in these measures, which might occur since a purely pixel-based metric as LCC does not necessarily properly capture the characteristics of the saliency distributions [38].

To examine the influence of image content on the similarity values, it is useful to look at the confidence interval of the mean similarity values. Wide confidence intervals indicate that similarity between reference and test data considerably varies across image content in the dataset. The SSIM similarity measure seems to have the lowest dependency on image content, thus possibly capturing differences in viewing behavior independent on the specific content of the image. When searching for significant differences between UESL and the similarity between reference and test data, the SSIM and KLD seem to be most sensitive (most differences between green circle points and blue diamond points

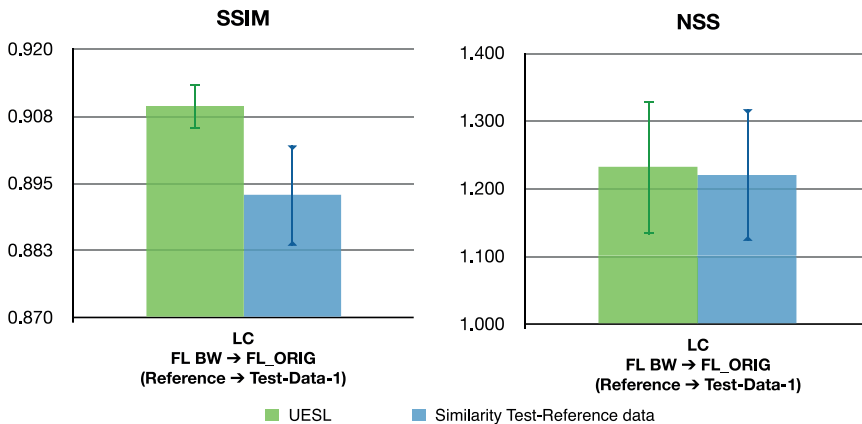


Figure 2.5. Similarity analysis (based on SSIM and NSS) of the LS Test-Data sets using NSS and SSIM. The error bars indicate the 95% confidence interval.

Table 2.3. Results of paired-samples T-tests performed comparing the similarity between test and reference data with their respective UESL. Rows indicating statistical significance are shown in bold font. M indicates the mean difference in the pair, SD the standard deviation of the differences, and N gives the number of pairs in the T-test.

		NSS				
		M	SD	N	t	p
Tested-Data-1: LC FL BW	Reference-Data-1: LC FL ORIG	0.0117	0.0699	29	0.903	0.374
Tested-Data-2: LC SC DIS	Reference-Data-1: LC FL ORIG	0.0656	0.0780	29	4.533	< 0.001
Tested-Data-3: LC FL DIS	Reference-Data-1: LC FL ORIG	0.0339	0.0710	29	2.571	0.016
Tested-Data-4: S SC DIS LQ	Reference-Data-2: LS FL ORIG	0.3090	0.3007	18	4.360	< 0.001
Tested-Data-5: LS SC DIS MQ	Reference-Data-2: LS FL ORIG	0.2749	0.2167	18	5.383	< 0.001
Tested-Data-6: LS SC DIS HQ	Reference-Data-2: LS FL ORIG	0.3526	0.4475	18	3.343	0.004
Test-Data-7: RS FL DIS LQ	Reference-Data-3: RS FL DIS HQ	-0.1514	0.3215	80	-4.315	< 0.001
Test-Data-8: RS SC DIS LQ	Reference-Data-4: RS FL DIS LQ	0.1136	0.3252	80	3.203	0.002
Test-Data-9: RS SC DIS HQ	Reference-Data-3: RS FL DIS HQ	0.2569	0.3443	80	6.839	< 0.001
		SSIM				
		M	SD	N	t	p
Tested-Data-1: LC FL BW	Reference-Data-1: LC FL ORIG	0.0165	0.0268	29	3.330	0.002
Tested-Data-2: LC SC DIS	Reference-Data-1: LC FL ORIG	0.0627	0.0449	29	7.521	< 0.001
Tested-Data-3: LC FL DIS	Reference-Data-1: LC FL ORIG	0.0146	0.0204	29	3.854	0.001
Tested-Data-4: S SC DIS LQ	Reference-Data-2: LS FL ORIG	0.2994	0.0876	18	14.493	< 0.001
Tested-Data-5: LS SC DIS MQ	Reference-Data-2: LS FL ORIG	0.2518	0.0573	18	18.654	< 0.001
Tested-Data-6: LS SC DIS HQ	Reference-Data-2: LS FL ORIG	0.2054	0.0777	18	11.223	< 0.001
Test-Data-7: RS FL DIS LQ	Reference-Data-3: RS FL DIS HQ	-0.0013	0.0261	80	-0.447	0.656
Test-Data-8: RS SC DIS LQ	Reference-Data-4: RS FL DIS LQ	0.0369	0.0212	80	15.967	< 0.001
Test-Data-9: RS SC DIS HQ	Reference-Data-3: RS FL DIS HQ	0.0330	0.0207	80	14.609	< 0.001

Table 2.4. One-way ANOVA statistical analysis testing for significant differences among the various datasets within a given image database; the first three rows refer to LC, the next three rows to LS, and the final row to RS. Rows indicating statistical significance are highlighted bold. The first three columns give the ANOVA statistics, whereas the next four columns represent post-hoc results

	NSS						
	F	df	Sig	M	95%		p
Tested-Data-1 VS Tested-Data-2	0.28	2.84	0.750	0.054	-0.118	0.226	0.735
Tested-Data-1 VS Tested-Data-3				0.022	-0.150	0.194	0.949
Tested-Data-2 VS Tested-Data-3				-0.032	-0.204	0.140	0.898
Tested-Data-4 VS Tested-Data-5	0.05	2.51	0.950	0.261	-0.231	0.179	0.949
Tested-Data-4 VS Tested-Data-6				0.129	-0.218	0.192	0.987
Tested-Data-5 VS Tested-Data-6				0.026	-0.179	0.231	0.987
Tested-Data-7 VS Tested-Data-8	33.80	2.00	< 0.001	0.275	0.155	0.394	< 0.001
Tested-Data-7 VS Tested-Data-9				0.408	0.289	0.528	< 0.001
Tested-Data-8 VS Tested-Data-9				0.134	0.014	0.253	0.024
	SSIM						
	F	df	Sig	M	95%		P
Tested-Data-1 VS Tested-Data-2	23.35	2.84	<0.001	0.046	0.027	0.065	< 0.001
Tested-Data-1 VS Tested-Data-3				-0.002	-0.021	0.017	0.966
Tested-Data-2 VS Tested-Data-3				-0.048	-0.067	-0.029	< 0.001
Tested-Data-4 VS Tested-Data-5	6.67	2.51	0.030	-0.045	-0.103	0.014	0.170
Tested-Data-4 VS Tested-Data-6				-0.889	-0.148	-0.030	0.002
Tested-Data-5 VS Tested-Data-6				-0.044	-0.103	0.014	0.173
Tested-Data-7 VS Tested-Data-8	85.60	2.00	<0.001	0.039	0.032	0.047	< 0.001
Tested-Data-7 VS Tested-Data-9				0.034	0.027	0.042	< 0.001
Tested-Data-8 VS Tested-Data-9				0.005	-0.003	0.013	0.251

are significant, see section 2.5.2). Taking all the above into consideration, SSIM and NSS are considered as the most useful measures to perform a more detailed analysis. The SSIM similarity measure has a very stable performance illustrated by the small confidence intervals and consistent results. It is also capable of detecting the largest number of significant differences among the data sets. NSS also seems to perform well, only with a wider variance in the data representing different image content. Therefore the in-depth analysis presented in section 2.5.2, investigating the effect of task and distortion on VA, only looks at the SSIM and NSS results.

2.5.2 Detailed statistical analysis

Figures 2.3, 2.4, and 2.5 take a closer look at the similarity values for the NSS and SSIM measures. These figures segment the data into 3 groups from which the effect of task and quality loss separately (Figure 2.3) and combined (Figure 2.4) may be deduced. Figure 2.5 examines the effect of color on visual attention deployment. Tables 2.3 and 2.4 show the results of a statistical analysis on the similarity measures. Table 2.3 shows the results of paired-samples T-tests performed with the similarity values of the test versus reference data, on the one side, and the UESL obtained from the corresponding reference data, on the other side of the pair. Table 2.4 shows the results of a one-way ANOVA test followed by a Tukey Post-Hoc comparison for significant differences among the different test datasets within each image database (i.e., LC, LS, and RS).

2.5.2.1. Scoring task effect on viewing behavior

To look at the effect of task on saliency, we first focus on Test-Data-8 and Test-Data-9. Saliency maps in these datasets were collected for a scoring task, and then compared to the corresponding saliency maps collected for the same image material, but under a free looking task (Reference-Data-4 and Reference-Data-3, respectively). Thus, in this comparison, the factor under investigation is only the task (images are unchanged). Figure 2.3 shows that both for SSIM and NSS the test-reference similarity is significantly lower than the UESL, thereby indicating an effect of task on the viewing behavior, i.e., spatial attention deployment is different when observers are scoring or just freely looking at an image. Table 2.3

indicates that the difference in attention deployment as a consequence of task is significant for both NSS and SSIM. The effect of task on saliency can also be deduced by directly comparing Test-Data-2 (scoring of distorted LC images) with Test-Data-3 (free looking of distorted LC images). As shown in Table 2.4, Test-Data-2 maps are significantly more different from Reference-Data-1 than Test-Data-3 maps (at least in terms of SSIM). Since the difference between Test-Data-2 and Test-Data-3 is in the task and not in the images, we can compute the variation in viewing behavior to the scoring task. These results show that, across different experimental settings and using different image databases, a change in task repeatedly resulted in a significant difference in viewing behavior indicated by at least one similarity measure. In particular, when scoring images, observers attend image locations that they would not attend in regular, free looking of images. This finding agrees with findings available in literature [26], and can be explained by the fact that when scoring quality, observers may inspect more thoroughly peripheral locations in the image in order to evaluate the annoyance of artifacts appearing across the whole content [3].

2.5.2.2. Effect of visual quality losses on viewing behavior

When it comes to changes in image quality level, our results are less clear. We first look at the effect of visual quality losses on free looking visual attention. We therefore examine the results of Test-Data-3 and Test-Data-7, collected for free looking of impaired images, while compared to their respective reference data, collected for free looking of the corresponding high quality images. As visible from Figure 2.3, SSIM detects a significant decrease in similarity with respect to the UESL for Test-Data-3, but not for Test-Data-7 (see also Table 2.3). NSS data seem to indicate (almost) no effect of quality on the saliency maps of Test-Data-3 (Table 2.3 also indicates that the difference is not significant). Interestingly, the NSS data suggest that for Test-Data-7 the variability induced by the loss in quality is less pronounced than the UESL. As a result, it is not possible to conclude that quality has a well-defined effect when observers freely look at images.

Looking at the UESL (green line) in Figure 2.3, and especially at data points corresponding to Reference-Data-3 and Reference-Data-4 (two

rightmost points in the graphs) one can see that they fall within each other's confidence interval. These data-points represent self-similarity of free-looking saliency data. They are generated from the same original image content but just slightly compromised in quality for Reference-Data-3 and heavily impaired for Reference-Data-4 (see Table 2.2). The lack of a significant difference between these two values indicates that, independent on the quality level of the images assessed, observers are similarly consistent in visually inspecting them. As a result, in terms of the RS dataset, it tends to show that inter-observer consistency in viewing behavior is not sensitive to visual quality losses, under the free looking condition.

Finally, the similarity values for Test-Data-8 and Test-Data-9 (rightmost blue diamonds in Figure 2.3) could also provide useful insight on the impact of visual quality losses in attention deployment. Each of these data points indicates the similarity of saliency maps corresponding to the viewing of the same image, but under different task. The difference between the two points is given by the quality level of the images: low for Test-Data-8 and high for Test-Data-9. SSIM shows that when the quality of the images is high, scoring saliency maps are closer to free-looking saliency maps. The trend however is not repeated in the NSS data.

In general, saliency of an image does not change consistently significantly with changes of visual quality. There are small changes in saliency for some images and some metrics, but they may depend on the image content, the type of distortion, and the level of degradation. In terms of the application of saliency, such as investigating whether the small difference in saliency due to the change of quality is sufficient to yield a consistent difference when using saliency in image quality assessment algorithms, the results are detailed in [9], though for a subset of the images evaluated here.

2.5.2.3. How a combination of factors affects viewing behavior

Looking at the combined effect of changing both task and quality, as provided by the results of Test-Data-2, Test-Data-4, Test-Data-5 and Test-Data-6 in Table 2.3 and Figure 2.4, we found a significant difference both with the SSIM and NSS similarity measures. It is likely that each of the

two factors (i.e., both task and quality) somewhat influenced the viewing behavior. Once the two effects are accumulated, the difference in saliency becomes easily detectable with both similarity measures. We can also examine in more detail the effect of changing the quality level in scoring tasks, by comparing the datasets Test-Data-4, Test-Data-5 and Test-Data-6, as given in Table 2.4 (see also Figure 2.4). Only one comparison yields a significant difference, and only using SSIM. In particular, this is the case for the Low Quality versus the High Quality LS images, where the saliency maps of the latter are significantly more similar to the free looking maps than the saliency maps for the former. This may be due to the fact that the presence of artifacts, in combination with a quality scoring task does indeed distract attention to background areas (see also Section 2.6 for a more detailed analysis), and possibly does so most pronounced for images in which artifacts are most evident. In general, though, as already observed by [28], it is still difficult to precisely quantify the effect of distortions on visual attention.

2.5.2.4. How color affects viewing behavior

Test-Data-1 compares free-looking saliency of gray-scale images to free-looking saliency of the same colored images. To our own surprise Table 2.3 gives a significant difference when using the SSIM similarity measure (as also visible from Figure 2.5). Color might also have an impact when looking at a combination effect of saturation, quality, and task, i.e., by comparing Test-Data-1 with Test-Data-2 (see first row of Table 2.4). However, since in that comparison the quality and task are also changed, it is not possible to draw conclusions on the effect of color saturation separately. Nonetheless, the quantitative comparison of color and grayscale saliency has implications for the application of saliency in image quality assessment algorithms, where most of the existing metrics are based on only the luminance component of the image material. Modeling visual attention based on luminance only might be used to simplify the attention model, which could be plausibly added to image quality metrics [43]. It is worth investigating whether the observed difference between color and grayscale saliency is sufficiently large to actually affect the performance gain when adding both types of saliency to image quality metrics, which, however, is outside the scope of this paper.

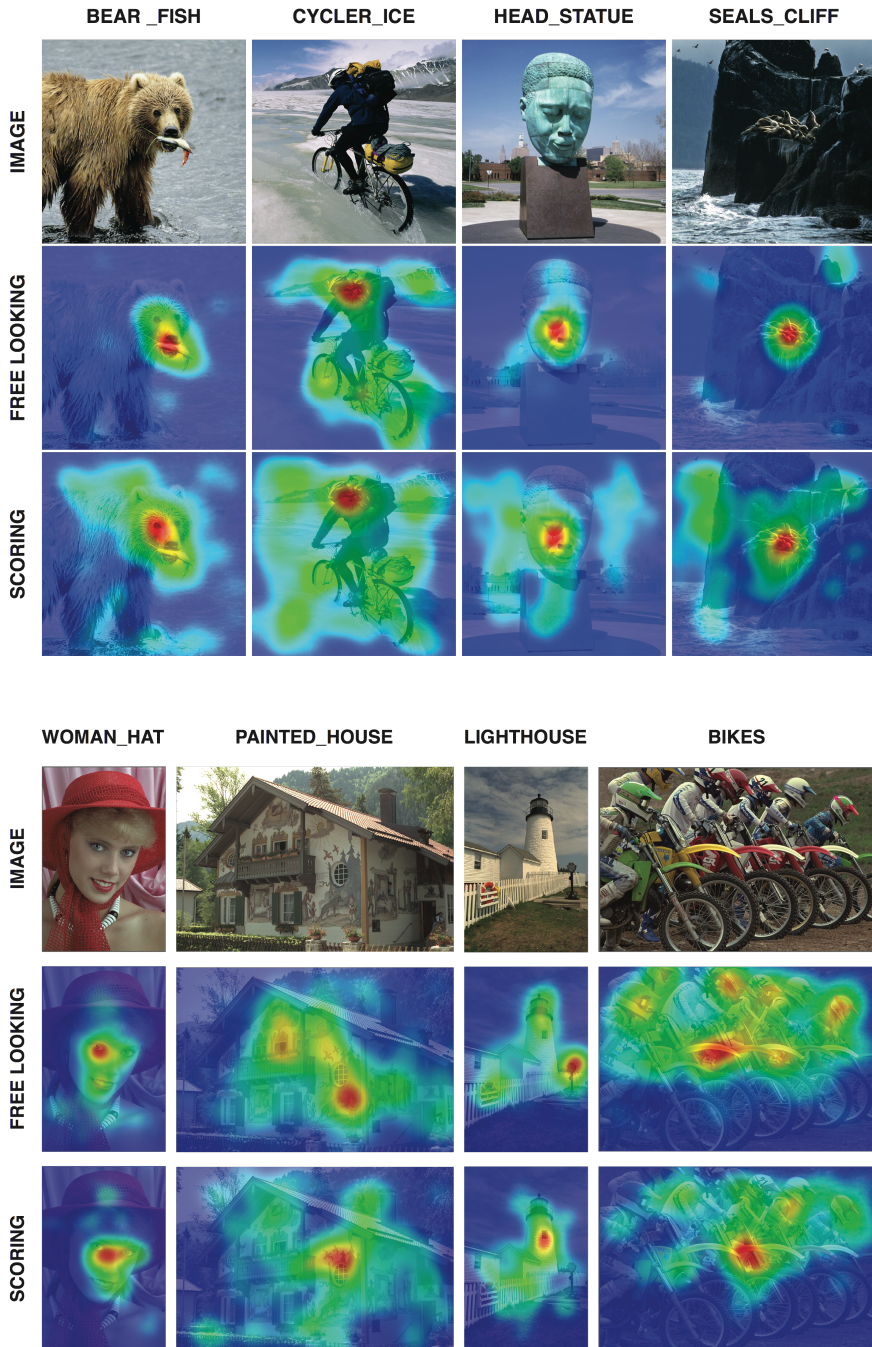


Figure 2.6. Difference in saliency maps between free looking and scoring tasks in the RS dataset (top) and the LS dataset (bottom).

2.5.2.5. Comparing similarity measures

The results in Section 2.5.1 show that the performance of similarity measures varies greatly from one metric to the next. Dependency on content seemed to be more prominent in LCC, KLD, and NSS than for SSIM. The capability of spotting significant differences between the tested data and the reference data was highest with SSIM and KLD. However, both LCC and KLD showed in some cases that the tested data was more similar to the reference data than the reference data to itself, which is somewhat unlikely. Therefore, similarity results given by the LCC and KLD are both considered unreliable, at least when examining the data used here. Judging by how many significant differences are detected and how possible it was to explain similarity trends in the data, the SSIM and NSS seemed to give the best results out of the used similarity measures.

Looking at the above analysis in Section 2.5.2, Figure 2.3 shows that the viewing behavior is significantly less similar to the Reference-Data when

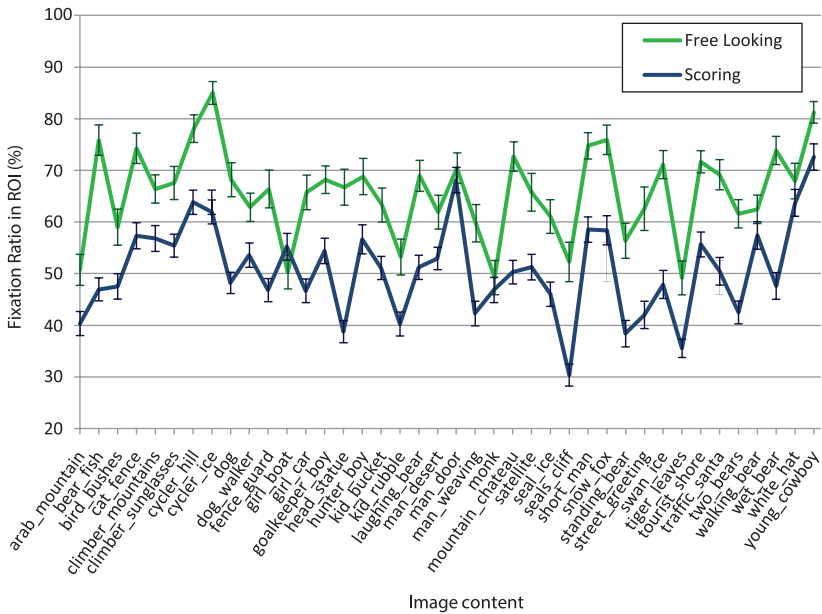


Figure 2.7. The percentage of fixations in the ROI for the images in the RS dataset for both the scoring and free viewing tasks. The error bars indicate the 95% confidence interval.

both task and color-saturation are changed compared to when only the color-saturation is changed. This effect is logical and only visible with SSIM. The other example is in Figure 2.4 which shows the viewing behavior becoming more similar to the Reference data as the quality of the images get higher (and thereby closer to the reference images). This is again only the case when SSIM is used. It is therefore difficult to argue that the SSIM is detecting significant differences where none actually exist.

2.6. Saliency changes with scoring task

In Section 2.5.1 we identified a clear impact of task on visual saliency. Nevertheless, whereas a change has been detected, its nature has not been explored. We are interested now in understanding whether there are systematic changes in the saliency distribution when a quality scoring task is in place. We examine here images from the RS dataset and from the LS set more in detail. The images from the first set have a clear ROI,

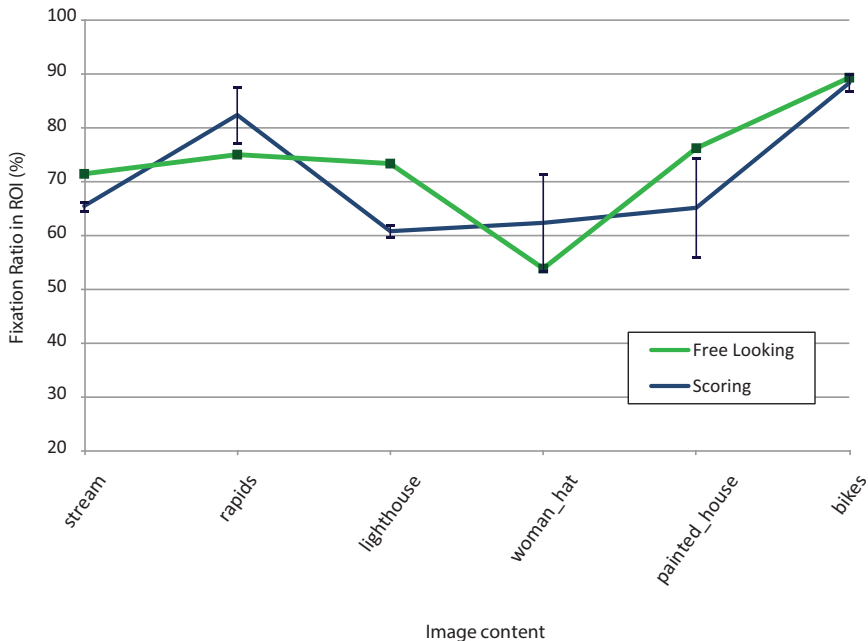


Figure 2.8. The percentage of fixations in the ROI for the images in the LS dataset for both the scoring and free viewing tasks. The error bars indicate the 95% confidence interval.

whereas those in the second set don't necessarily have a clear attentive focus. By examining both, we aim at accounting for content differences in our analysis.

We first start by visualizing the saliency maps in heat maps, where the areas of the map with the lower saliency values are colored in blue, while higher saliency gradually becomes green and yellow, and finally ends with red for the areas with the highest saliency values. Figure 2.6 shows eight example images with their corresponding heat maps. The four images on the top are from the RS dataset, and the four images on the bottom from the LS dataset. The second row in the figure shows heat maps from viewers looking freely at the images, while the third row does the same but for maps collected under the scoring task.

Looking at the example images of the RS dataset, it is clear from the figure that viewers looking freely at the images have their attention concentrated more on the ROI of the images. The size of the ROI for each of the images is different, which shows that the content has a great effect on the viewing behavior. When we compare the heat maps in rows 2 and 3, we can start to understand the difference in the viewing behavior. Even though the majority of the visual attention is still dedicated to the ROI when scoring, more attention is given to other parts of the image when scoring as compared to when freely looking. The viewer is clearly scanning the image for clues to determine its quality level. This behavior is more difficult to see on the examples from the LS dataset (Figure 2.6 bottom). Since there is no clear ROI, the visual attention is already more dispersed around the image for the free looking task. Therefore it is difficult to see a clear difference in behavior with the scoring task.

To quantify how much the ROI captured the viewer's attention, we start by identifying the ROI for each image. As the saliency maps have values between 0 (no attention) and 1 (maximum attention), a value of 0.2 was chosen as an attention threshold, indicating that at least one fifth of the observers attended the location. As a consequence, all pixels in the saliency maps with a value above 0.2 were considered to belong to the ROI. Subsequently, we defined a percentage of fixations in the ROI as the sum of the duration of the fixations inside the ROI divided by the total

duration of all fixations (multiplied by 100) [44]. For example, if the percentage of fixations of a given participant was 70 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.

Figure 2.7 shows the percentage of fixations in the ROI for all the images in the RS dataset, for the free looking and scoring tasks separately. The figure clearly shows that, for almost all images, the percentage of fixations in the ROI is higher when freely looking than when scoring the images. This indicates that the tendency we observed when examining the heat maps in Figure 2.6 (top) applies to most images in the RS dataset. The viewer generally looks more at the ROI of the image when looking freely. When they are asked to score the images, the attention deviates from the ROI to other regions of the image, possibly to evaluate the presence and annoyance of artifacts in less obvious regions of the image.

Figure 2.8 presents similarly calculated values, but for the images in the LS dataset. The first noticeable difference is that the fixation ratios here are much higher than for the images in the RS dataset. This may be due to the fact that images of this set were more cluttered, with multiple locations competing for attention. As a result, fixations may have been more spread to begin with, generating therefore larger ROIs (including in turn most of the fixations). One can also see that there is less of a clear difference in behavior of the percentage of fixations in the ROI between the free looking and scoring tasks. This also reflects the tendency observed in Figure 2.6, and confirms that when the ROI of an image is less defined, the difference in viewing behavior becomes less discernible.

2.7. Conclusions

In order to understand whether it matters which type of VA information to incorporate in objective quality metrics, we evaluated whether there are differences in VA depending on the viewing task under which the data are collected and depending on the quality level of the images observed. To this end, this paper examines a corpus of eye-tracking data collected over 4 years of work by the TU Delft IQ-Lab group [39]. All data used in

this article have been made available online at the TU Delft IQ-Lab repository for use by other researchers in the field.

We analyzed visual attention data in the saliency domain, using four different measures of saliency similarity [23]. Differences in saliency distribution were analyzed among VA data collected under different experimental conditions, i.e. when viewing task (scoring or free looking), quality level (presence or absence of impairments) and color saturation level (greyscale or color) were changed. Throughout the analysis, we evaluated different similarity measures, reported earlier in literature. We based our selection of most appropriate similarity measure on criteria as reliability with respect to self-similarity, sensitivity to image content variability and ability to detect differences in spatial distribution of saliency. The similarity measures performing well on these criteria are the Structural SIMilarity index (SSIM, [37]) and the Normalized Scanpath Saliency measure (NSS [34]).

When looking at the effect of task on saliency, we found a significant effect on saliency distribution across the three datasets. The analysis provides strong evidence that asking the viewers to score the image quality significantly changes their viewing behavior. Completely muting the color saturation also showed a significant change in saliency when analyzed with SSIM. This conclusion, however, is based only on data from a single image set, and therefore warrants further investigation. Quality losses were not found to consistently modify visual attention deployment, neither under free looking, nor during scoring. In some cases we found an effect using the SSIM measure, but this effect was not consistent over all data sets and was generally not confirmed with the NSS measure.

Examining in more detail how the change in task affected the viewing behavior shows that when looking freely at images viewers give most of their attention to the most prominent region of interest. When viewers are asked to score images, their attention deviates to other regions of the image scanning it for clues to the image quality level. This change in attention is stronger when the image has a clearly defined region of interest.

These conclusions show that it is not fair to compare the effect of adding saliency in objective metrics without specifying how the saliency was measured. Additionally, the differences in saliency reported here provide insights for designing objective metrics as it seems important to consider which saliency information they should incorporate.

References

1. Winkler, S. (2005). *Digital video quality: vision models and metrics*. John Wiley & Sons.
2. Hemami, S. S., & Reibman, A. R. (2010). No-reference image and video quality estimation: Applications and human-motivated design. *Signal processing: Image communication*, 25(7), 469-481.
3. Redi, J., Liu, H., Zunino, R., & Heynderickx, I. (2011, February). Interactions of visual attention and quality perception. In *IS&T/SPIE Electronic Imaging* (pp. 78650S-78650S). International Society for Optics and Photonics.
4. Engelke, U., Kaprykowsky, H., Zepernick, H., & Ndjiki-Nya, P. (2011). Visual attention in quality assessment. *Signal Processing Magazine, IEEE*, 28(6), 50-59.
5. Ninassi, A., Le Meur, O., Le Callet, P., & Barbba, D. (2007, September). Does where you gaze on an image affect your perception of quality? Applying visual attention to image quality metric. In *Image Processing, 2007. ICIP 2007. IEEE International Conference on* (Vol. 2, pp. II-169). IEEE.
6. Sadaka, N. G., Karam, L. J., Ferzli, R., & Abousleman, G. P. (2008, October). A no-reference perceptual image sharpness metric based on saliency-weighted foveal pooling. In *Image Processing, 2008. ICIP 2008. 15th IEEE International Conference on* (pp. 369-372). IEEE.
7. Moorthy, A. K., & Bovik, A. C. (2009). Visual importance pooling for image quality assessment. *Selected Topics in Signal Processing, IEEE Journal of*, 3(2), 193-201.
8. Gkioulekas, I., Evangelopoulos, G., & Maragos, P. (2010, September). Spatial Bayesian surprise for image saliency and quality assessment. In *Image Processing (ICIP), 2010 17th IEEE International Conference on* (pp. 1081-1084). IEEE.
9. Liu, H., & Heynderickx, I. (2011). Visual attention in objective image quality assessment: based on eye-tracking data. *Circuits and Systems for Video Technology, IEEE Transactions on*, 21(7), 971-982.
10. Larson, E. C., Vu, C., & Chandler, D. M. (2008, October). Can visual fixation patterns improve image fidelity assessment?. In *Image Processing, 2008. ICIP 2008. 15th IEEE International Conference on* (pp. 2572-2575). IEEE.
11. You, J., Perkis, A., Hannuksela, M. M., & Gabbouj, M. (2009,

- October). Perceptual quality assessment based on visual attention analysis. In *Proceedings of the 17th ACM international conference on Multimedia* (pp. 561-564). ACM.
12. Engelke, U., Barkowsky, M., Le Callet, P., & Zepernick, H. (2010, June). Modelling saliency awareness for objective video quality assessment. In *Quality of Multimedia Experience (QoMEX), 2010 Second International Workshop on* (pp. 212-217). IEEE.
 13. Le Meur, O., Ninassi, A., Le Callet, P., & Barba, D. (2010). Overt visual attention for free-viewing and quality assessment tasks: Impact of the regions of interest on a video quality metric. *Signal Processing: Image Communication*, 25(7), 547-558.
 14. Gao, X., Liu, N., Lu, W., Tao, D., & Li, X. (2010, October). Spatio-temporal salience based video quality assessment. In *Systems Man and Cybernetics (SMC), 2010 IEEE International Conference on* (pp. 1501-1505). IEEE.
 15. Culibrk, D., Mirkovic, M., Zlokolica, V., Pokric, M., Crnojevic, V., & Kukolj, D. (2011). Salient motion features for video quality assessment. *Image Processing, IEEE Transactions on*, 20(4), 948-958.
 16. Itti, L., Koch, C., & Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on pattern analysis and machine intelligence*, 20(11), 1254-1259.
 17. Le Meur, O., Le Callet, P., Barba, D., & Thoreau, D. (2006). A coherent computational approach to model bottom-up visual attention. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 28(5), 802-817.
 18. Rajashekar, U., Van Der Linde, I., Bovik, A. C., & Cormack, L. K. (2008). GAFFE: A gaze-attentive fixation finding engine. *Image Processing, IEEE Transactions on*, 17(4), 564-573.
 19. Wolfe, J. (2000). Visual attention. *Seeing*, 2, 335-386.
 20. Burr, D. C., Morrone, M. C., & Ross, J. (1994). Selective suppression of the magnocellular visual pathway during saccadic eye movements. *Nature*, 371(6497), 511-513.
 21. Yarus, A. L. (1967). *Eye movements and vision* (Vol. 2, No. 5.10). L. A. Riggs (Ed.). New York: Plenum press.
 22. Itti, L., Koch, C., & Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on pattern analysis and machine intelligence*, 20(11), 1254-1259.
 23. Redi, J. A., & Heynderickx, I. (2011, September). Image quality and visual attention interactions: towards a more reliable analysis in the saliency space. In *Quality of Multimedia Experience (QoMEX), 2011 Third International Workshop on* (pp. 201-206). IEEE.
 24. Buswell, G. T. (1935). How people look at pictures: a study of the psychology and perception in art.
 25. Noyes, C. E. (1907). *The Gate of Appreciation: Studies in the Relation*

- of Art to Life*. Houghton, Mifflin.
26. Ninassi, A., Le Meur, O., Le Callet, P., Barba, D., & Tirel, A. (2006). Task impact on the visual attention in subjective image quality assessment. In *Proceedings of European Signal Processing Conference*.
 27. Vuori, T., Olkkonen, M., Pölönen, M., Siren, A., & Häkkinen, J. (2004, October). Can eye movements be quantitatively applied to image quality studies?. In *Proceedings of the third Nordic conference on Human-computer interaction* (pp. 335-338). ACM.
 28. Le Meur, O. (2011, September). Robustness and repeatability of saliency models subjected to visual degradations. In *Image Processing (ICIP), 2011 18th IEEE International Conference on* (pp. 3285-3288). IEEE.
 29. Kim, C., & Milanfar, P. (2013). Visual saliency in noisy images. *Journal of vision*, 13(4), 5.
 30. Mantel, C., Guyader, N., Ladret, P., Ionescu, G., & Kunlin, T. (2012). Characterizing eye movements during temporal and global quality assessment of h. 264 compressed video sequences. In *IS&T/SPIE Electronic Imaging* (pp. 82910Y-82910Y). International Society for Optics and Photonics.
 31. Redi, J., Heynderickx, I., Macchiavello, B., & Farias, M. (2013). On the impact of packet-loss impairments on visual attention mechanisms. In *Circuits and Systems (ISCAS), 2013 IEEE International Symposium on* (pp. 1107-1110). IEEE.
 32. Le Meur, O., Ninassi, A., Le Callet, P., & Barba, D. (2010). Do video coding impairments disturb the visual attention deployment?. *Signal Processing: Image Communication*, 25(8), 597-609.
 33. Alers, H., Redi, J., Liu, H., & Heynderickx, I. (2013). Studying the effect of optimizing image quality in salient regions at the expense of background content. *Journal of Electronic Imaging*, 22(4), 043012-043012.
 34. Peters, R. J., Iyer, A., Itti, L., & Koch, C. (2005). Components of bottom-up gaze allocation in natural images. *Vision research*, 45(18), 2397-2416.
 35. Riche, N., Duvinage, M., Mancas, M., Gosselin, B., & Dutoit, T. (2013, December). Saliency and Human Fixations: State-of-the-art and Study of Comparison Metrics. In *Computer Vision (ICCV), 2013 IEEE International Conference on* (pp. 1153-1160). IEEE.
 36. Le Meur, O., & Baccino, T. (2013). Methods for comparing scanpaths and saliency maps: strengths and weaknesses. *Behavior research methods*, 45(1), 251-266.
 37. Wang, Z., Bovik, A. C., Sheikh, H. R., & Simoncelli, E. P. (2004). Image quality assessment: from error visibility to structural similarity. *Image Processing, IEEE Transactions on*, 13(4), 600-612.
 38. Borji, A., Sihite, D. N., & Itti, L. (2013). Quantitative analysis of human-

- model agreement in visual saliency modeling: A comparative study. *Image Processing, IEEE Transactions on*, 22(1), 55-69.
39. The TU Delft Image Quality Lab repository: <http://mmi.tudelft.nl/iqlab/>
40. IREcommendation, I. T. U. R. B. T. (2002). 500-11, Methodology for the subjective assessment of the quality of television pictures. *International Telecommunication Union, Geneva, Switzerland*, 4, 2.
41. Sheikh, H. R., Wang, Z., Cormack, L., & Bovik, A. C. (2005). LIVE image quality assessment database release 2., <http://live.ece.utexas.edu/research/quality>.
42. Stankiewicz, B. J., Anderson, N. J., & Moore, R. J. (2011). Using performance efficiency for testing and optimization of visual attention models. In *IS&T/SPIE Electronic Imaging* (pp. 78670Y-78670Y). International Society for Optics and Photonics.
43. Liu, H., & Heynderickx, I. (2010). Visual Attention Modeled with Luminance Only: from Eye-Tracking Data to Computational Models. Fifth International Workshop on Video Processing and Quality Metrics for Consumer Electronics - VPQM 2010, 2010.
44. Alers, H., Bos, L., & Heynderickx, I. (2011, September). How the task of evaluating image quality influences viewing behavior. In *Quality of Multimedia Experience (QoMEX), 2011 Third International Workshop on* (pp. 167-172). IEEE.

3.

Studying the Effect of Optimizing Image Quality in Salient Regions

Abstract

Manufacturers of commercial display devices continuously try to improve the perceived image quality of their products. By applying post-processing techniques on the incoming signal, they aim to enhance the quality level perceived by the viewer. These post-processing techniques are usually applied globally over the whole image, but may cause side-effects, the visibility and annoyance of which differ with local content characteristics. To better understand and utilize this, a three-phase experiment was conducted where observers were asked to score images which had different levels of quality in their regions of interest and in the background areas. The results show that the region of interest has a greater effect on the overall quality of the image than the background. This effect increases with the increasing quality difference between the two regions. Based on the subjective data we propose a model to predict the overall quality of images with different quality levels in different regions. This model, which is constructed on empirical bases, can help craft weighted objective metrics which can better approximate subjective quality scores.

3.1. Introduction

In today's competitive market, commercial display manufacturers are striving to find new features to help them overtake competition. Since consumers find Image Quality (IQ) one of the deciding factors when choosing a display [1], research and development effort has been concentrated on improving image quality using various techniques. For some applications, however, the quality of the content is one of the bottlenecks. It has become quite common today to view video material on devices such as personal computers and mobile phones. Regardless of whether the video material is stored on the device itself or streamed from a remote server, the limitations that such devices have in storage capacity and data transfer bandwidth make it desirable to reduce the video data size as much as possible by means of data compression algorithms [2-4]. Unfortunately compression algorithms also introduce artifacts in the content.

It is possible to compensate for some of the artifacts caused by compression algorithms. For example, areas which have become blurred after compression can benefit from a sharpening filter [5, 6]. On the contrary, the impact of blocking artifacts may be reduced by applying a blur filter [7]. Since the visibility of each artifact may vary depending on the image content, one part of an image may be effected more by a specific artifact than others parts [8,9]. Therefore, applying an image enhancement filter may improve the perceived IQ in some areas of an image while making other areas worse. For example, applying a sharpening filter will enhance areas affected by blur, while it will make blocking artifacts more visible [10-16]. It is therefore important to know how the viewer evaluates the overall quality of the image if different regions of the image differ in their quality level. A more specific question is whether improving the quality of the Region Of Interest (ROI) results in a higher IQ rating for the entire image even if the quality of some background (BG) regions has become worse.

The work discussed in this article shows that subjectively measured MOS is different from an estimated score obtained by averaging the quality of all image regions. Nonetheless, the latter is what most quality estimation algorithms do; they locally estimate (based on pixel values) a quality

score, and then average these scores over the entire image [17-32]. As such, the calculated overall quality is an (area weighted) average of the local quality in different regions of the image. More advanced quality estimation algorithms include saliency weighting; i.e., the local quality values are weighted with the local saliency, as such giving more value to quality in the ROI than to quality in the BG [33-38]. So far, however, the weighting strategy for adding saliency has not been determined. Attempts to determine this weighting strategy have largely depended on trial and error [34-38]. As such, this chapter contributes in quantifying the optimal weighting strategy.

This paper describes a three-phase experiment that examines the significance of the ROI in determining the quality of the entire image. A database of images with a clear ROI was compromised in quality to different degrees using JPEG compression. The IQ levels of these images, as well as their natural ROI, were subjectively determined with the help of an eye-tracking system. The images were then manipulated to have different quality levels in the ROI and the BG regions. The overall IQ of the manipulated images was subjectively evaluated as well. These scores were then compared to the subjective IQ scores of the non-manipulated images to determine whether the ROI had a stronger effect on IQ than the rest of the image.

The methodology and the experimental protocol are discussed in Sections 3.2 and 3.3, respectively. Section 3.4 lists the results of the experiment, which are then discussed in Section 3.5. Finally Section 3.6 summarizes the conclusions and mentions some possibilities for future research.

3.2. Experimental set-up

3.2.1 Stimuli

The stimuli used in the experiment were created from 40 original images. All were natural images containing humans, animals, or structures. Considering the goal of the experiment, we only chose images which 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 (corresponding to a viewing angle of 20.2°) in order to have a standard size for all images.

Each image was further processed to produce 4 different versions, which resulted in a total of 160 stimuli used in the experiment. These versions were created with the JPEG compression function (`imwrite`), defined in MATLAB. The compression parameter for the (`imwrite`) function in the four compression levels used to process the images ranged between 10 (low quality) and 100 (high quality), and were different for the 40 different originals. Some example images are shown in Figure 3.1.

3.2.2 The eye tracker

To record the gaze location of the users viewing the images, an eye-tracking system (i.e., iView X system developed by SMI) was adopted. It tracks the eye movements of the users with an infrared camera, recording the reflections of a small infrared source at the eye's retina. Since infrared falls outside the spectral range of sensitivity of the human visual system, the viewers were not distracted by the infrared light emitted by the eye-tracker. The REDIII camera used by the eye-tracker had a sampling rate of 50 Hz and a tracking resolution of ± 0.1 deg. Viewers were asked to place their head on a head rest as recommended by the system's manual. The head rest restrained head movements and 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 eye-tracker was calibrated using a 13-points grid, and resulted in a gaze position tracking accuracy of ± 1 deg. 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.



Figure 3.1. Four examples of the images used in the experiment.

3.2.3 Facilities

The experiment was carried out in an isolated room. Only the experimenter and participant were present during the experiment. All stimuli were displayed on a CRT monitor with a resolution of 1024 by 768 pixels and an active screen area of 365x275mm. The experiment was controlled from a remote computer with its monitor positioned so that its content was not visible to the participant to avoid distractions (see Figure 3.2).

3.2.4 The participants

The experiment had a total of 75 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.

The participants were informed that they would carry-out an experiment on image quality research. They were told that their eye-movements



Figure 3.2. Participants place their head on a chin-rest positioned at a fixed distance from the display. The eye-tracker is positioned next to the display. The experimenter controls the eye tracker and runs the experiment using another monitor not visible to the participant.

would be recorded using an eye-tracking device. However, they were not informed about the goal of the experiment or how the data is going to be analyzed in order not to reduce the influence on their viewing behavior. After they gave their consent, a quick test was performed to check whether the eye-tracker locked on the participant's pupil. The latter was occasionally not possible due to reflections from eye glasses or to poor contrast between the pupil and the iris in the infrared spectrum. These participants had to be excluded from the experiment, and were replaced by new ones.

3.3. Experimental protocol

As mentioned before, the experiment included 3 separate phases. Phases 1 and 3 required people to examine images and to give each image a score based on the perceived quality. Participants in phase 2 were only asked to look at the images without a predefined task. The participants were divided such that we had 20 participants in phase 1, 40 in phase 2, and 15 in phase 3. A larger number of participants was needed in phase 2 to identify the natural ROI of the images (see Section 3.4.1). Phase 3 required less participants since the eye-tracking data was not going to be analyzed and the number of images was lower than that of phase 1. Each phase of the experiment adopted a within-subject design, where changes in the dependent variable (that is, IQ in phases 1 and 3 and VA deployment in phase 2) are analyzed as they appear across different images rather than across groups of different test subjects. However, different participants were recruited for each phase.

Participants who passed the check with the eye-tracker were asked to start the experiment. In order to insure consistency, the instructions for the experiment were given to the participants through the computer screen, together with examples of how to perform each step. After having read all instructions, the subjects were allowed to ask clarifying questions. Once they were ready to start, the experimenter started the eye-tracking calibration process, and then started showing them the stimuli. To avoid introducing a bias in the results, each participant saw the corresponding stimuli in a different random order.

3.3.1 Phase 1

Participants in phase 1 were shown all 160 stimuli of the experiment. The experiment was split in 4 sessions requiring the participants to evaluate 40 images in each session. Every session contained one version of each original image presented at a certain level of compression. The system chose the image at random insuring that at the end of the session, the participant saw one version of each of the 40 original images in the database. In the subsequent sessions, the participant was shown one of the remaining versions of each image. The order in which the images were shown in each session was 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 ITU39. The participants were shown a 50% gray screen (R,G, and B values set to 127) with a white dot in the center. They were 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 refine the eye-tracker's calibration (see also section 3.3.4). Subsequently, a randomly selected image was displayed on the screen centered on a 50% 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 to choose the required score. Then a click on the left mouse button saved the chosen score and took the participant again to the 50% gray screen with the white dot in the center. These steps were repeated until the end of the session, in which the participants had to score 40 different images. After a short break, the participants started the following session by first completing the 13-points calibration step described earlier, followed by another 40 images randomly chosen. This process was repeated in 2 more sessions taking each participant through the entire database of 160 stimuli.

3.3.2 Phase 2

In this phase 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 was later used to subjectively identify the natural ROI of the images. To avoid any deviation in the measured saliency due to a learning effect from viewing the same image content multiple times, participants only viewed one (compressed) version of each original image.

The second phase was performed concurrently with phase 1, 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. Before the experiment started, two sample images were shown to the participants. These images were separated by the 50% gray screen with the white dot in the center, similarly as in phase 1. 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 refine the eye-tracker's calibration.

After completing the training, the participants went through the 13-points calibration step as before and then started viewing the stimuli. Each stimulus was displayed on the screen for 8 seconds followed by the 50% gray screen. Basically, each participant saw a selection of all 160 stimuli as if he completed just one session of phase 1. As a result, every 4 participants saw the entire set of the 160 images presented at a random order. As a consequence, by the end of phase 2 we gathered the free looking gaze data of 10 participants for each compressed version of the 40 original images.

3.3.3 Phase 3

The last step of the experiment used stimuli generated from the same original content, but with a different level of quality for the ROI and background. Data collected from phase 2 of the experiment were used to identify the ROI of the images. In order to avoid that the size of the ROI region affected the results, only 20 of the original images with a similarly sized ROI, were used in phase 3. The size of the ROI ranged from

10-16% of the entire image area corresponding to a viewing angle of 2.0°-3.2° (see Figure 3.2).

From the original 160 images used in this experiment, only 80 were chosen for phase 3 (20 of the 40 original images) which had the most clear and uniquely identifiable ROI. Each stimulus in phase 3 contained data from two stimuli of phase 1. Basically, from every two versions of an image with different levels of quality in phase one, the contents of the ROI was swapped between the two images. This created two combined images with different levels of quality inside and outside the ROI (see Figure 3.3). The edge between the two regions was softened using a 3x3 pixel Gaussian function. In total, 80 stimuli were used in phase 3. They were shown to each participant in 4 sessions in a similar manner as used in phase 1. Figure 3.4 shows an example of the resulting combined images.

To ensure consistency, the experiment was conducted in the same lab and under the same conditions as the two previous phases. The same scoring protocol was used as the one described in phase 1. The eye-tracker was also used to ensure uniformity in the experimental conditions to make sure any change in the data is not caused by not using the eye tracker in phase 3. The data collected from the eye-tracker were not needed for this phase of the experiment.

3.3.4 The eye-tracking data

The eye-tracker collected the coordinates of the participant's gaze locations throughout each session. These data were then sorted into fixations and saccades by the eye-tracking system based on the gaze dispersion within a specified amount of time. For the experiment, the system was set to consider a gaze that remained within an area of 100 pixels (viewing angle of 3.4°) for 80 ms or longer as a fixation. Its location was calculated as the mean of the coordinates over the entire length of the fixation. If the eye dispersion exceeded 100 pixels, the tracker indicated the movement as a saccade. So all fixations had a duration of at least 80 ms, and all saccades spanned a distance of at least 100 pixels.

While testing the eye-tracker, we noticed that the recorded fixations were occasionally shifted from their correct location. This shift tended to be a constant displacement in horizontal and vertical direction of all fixations in a test session. To compensate for this error in the collected data, an additional calibration step was added to the experiment. Between each two images displayed on the screen, the system displayed a 50% gray screen with a white dot in the center. The participants were instructed to keep their eyes fixed on at the dot. As such, we aimed at having a uniform starting gaze location for each participant. Since the eye-tracker recorded where the participants were looking at, and we knew the coordinates of the dot that they were supposed to look at, it was possible to compensate for the possible shifts in fixations. The correction was performed in MATLAB by taking the mean coordinates of all fixation points collected on the gray screen for the entire session, and then applying an opposite shift to the rest of the fixation points recorded by the system.

The iView X eye-tracking system mainly collects fixations that need to be converted into saliency maps. These maps show a visual representation

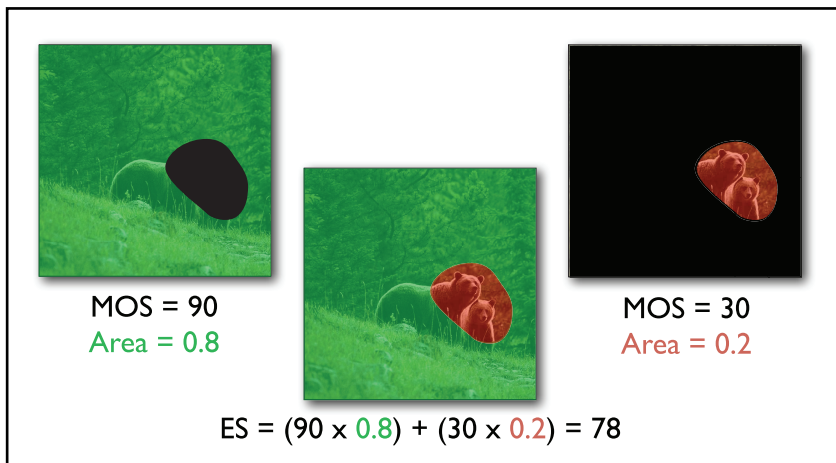


Figure.3.3. Illustrating how images were created for phase 3. Each image combined parts from two different images inside and outside the ROI. In this hypothetical example, the BG region is taken from the image on the left scored with MOS=90, and the ROI is taken from the image on the left with an MOS=30, giving the combined image in the center. The figure also shows how the ES is calculated for the resulting combined image.

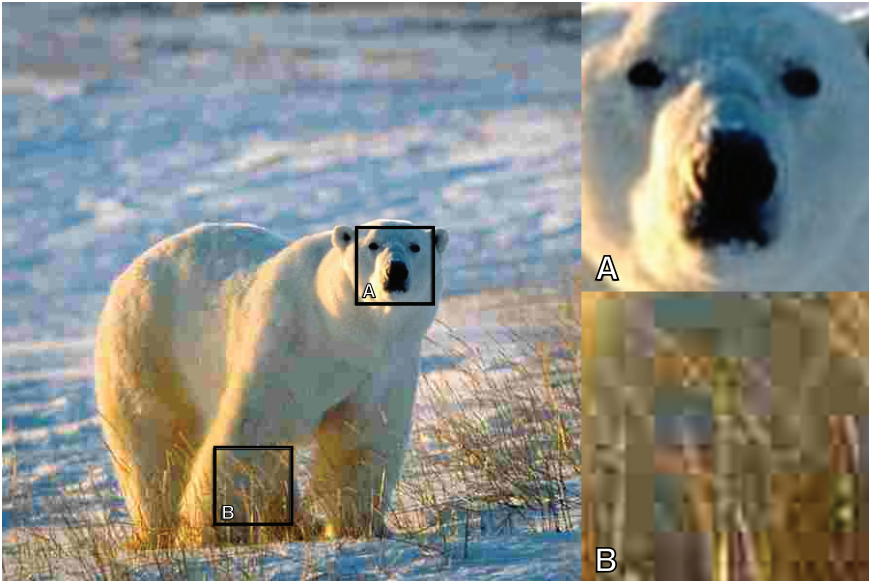


Figure 3.4. A sample of the combined stimuli with the blown up region A inside the ROI with quality of 77 and region B in the BG with quality of only 11.

of the probability that a location of the scene is attended by the average observer. To create the saliency map for a given image i , first a fixation map is created that includes all the fixation locations from all observers recorded for that image. These fixation maps are then converted to saliency maps by applying a Gaussian patch with a width σ to each fixation in the map. The width σ of the Gaussian patch is chosen to be 2° of visual angle, approximating the size of the fovea of the human eye and sufficiently accounting for inaccuracy in the measurement of the fixations. A mean saliency map that takes into account all fixations of all subjects is then calculated as follows:

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

where $S_i(k, l)$ indicates the saliency map for stimulus i , (k, l) refers to a pixel in the saliency map of size $M \times N$ pixels (i.e. $k \in [1, M]$ and $l \in [1, N]$), (x_j, y_j) indicates the spatial coordinates of the j th fixation ($j=1 \dots T$), 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].

3.4. Results

3.4.1 Natural Region Of Interest

As mentioned earlier, the images selected for this experiment were deliberately chosen to have a clearly identifiable ROI. It is expected that when observing the images without a specific task, the viewer's attention is mainly drawn towards the natural ROI of each image. For example, in a picture of a man standing on the beach, one would expect his head to be the ROI of the picture, while for a picture of a face features such as the eyes usually attract viewers' attention [40].

The ROI for each image was determined by the eye-tracking data collected in phase 2 of the experiment. For each of the 160 stimuli, the data of all 10 participants were averaged into one saliency map. Since the viewers saw compressed versions of the images in phase 2, we needed to determine whether the quality level effected their viewing behavior. To that effect, the highest and lowest quality version of each image were distributed into two separate sets. Then an independent sample t-test was performed to see whether the difference in quality resulted in a difference in the viewing behavior with the independent variable being the quality level (high or low) and the dependent variable being the similarity score that shows how similar the saliency maps are to each other [41]. The test showed no significant difference in the similarity of the saliency maps between the high and low quality images ($p = 0.48$, $F = 5.94$). This shows that when the observers were looking at the stimuli, they were not distracted by the compression artifacts and their viewing behavior did not change with the change in compression level. Therefore, the 4 saliency maps for each original content were averaged, giving us 40 saliency maps.

3.4.2 Defining the Region Of Interest

The ROI for each image is extracted from the saliency maps. Since these maps were normalized on a scale of 0-1 representing the level of

intensity of the saliency heat-maps. The ROI was identified as the area with the top 25% of the range (i.e. scoring 0.75-1 on the heat-map scale).

3.4.3 Significance of ROI on IQ

The IQ scores collected in phases 1 and 3 of the experiment were processed to calculate the Mean Opinion Scores (MOS). First the Z-scores were calculated for each image. Then the standard normal distribution of the resulting Z-scores was taken to give a score in the range (0-1). From the MOS scores obtained in phase 1, it is possible to estimate the Expected Score (ES) of a combined stimulus in phase 3, assuming that the observer will average out the overall quality of the image without giving more importance to the quality of a specific region. In that case, the ES is the weighted sum of the MOS scores of each stimulus, as obtained in phase 1, weighted only with the percentage of area of the ROI and background, respectively. Hence, under this assumption, we would calculate the ES as follows:

$$ES = MOS_{ROI} \cdot A_{ROI} + MOS_{BG} \cdot A_{BG} \quad (2)$$

With MOS_{ROI} and MOS_{BG} being the scores of the images used in the ROI and in the background regions respectively, and A_{ROI} and A_{BG} are the ratios of the area of the ROI and the background regions respectively to the entire image. The way the ES is calculated is also illustrated in Figure 3.3. By comparing the collected MOS values of all stimuli of phase 3 to their estimated ES, it is possible to extract the effect the ROI has on the overall quality of an image. Figure 3.5 presents these data, split up in two data groups: one with images which have a higher IQ in the ROI than in the BG (A), and one with images which have a lower IQ in the ROI than in the BG (B). Figure 3.5(A) clearly shows that the images with higher quality in the ROI have a tendency to get a higher MOS than what would be expected from the ES. Looking at the trend line, this effect seems to be stronger for images located in the central region of the quality range. The effect diminishes when the quality of the image is too high or too low. A similar tendency is seen in Figure 3.5(B); images with a lower quality in the ROI tend to get a lower MOS than what would be expected from the ES. The latter effect, however, seems to be weaker

than the one found for images with a higher quality in the ROI. Figure 3.5(B) shows that the effect is less prevalent for images that have the higher quality in the BG area of the image. Occasionally these images even gain a MOS that exceeds the ES.

It is also interesting to see whether the size of the quality difference between the ROI and the BG plays a role on the overall MOS of the combined stimulus. Figure 3.6 shows a scatter plot that attempts to illustrate this effect. In this figure, the horizontal axis represents the difference in quality between the ROI and the BG of the stimulus. All stimuli fall either in the negative half or the positive half of the graph, depending on whether the ROI region or the BG had a higher quality. The vertical axis represents the difference between the MOS collected from phase 3 and the ES.

If the effect of the ROI and BG on overall IQ was equal, all data points in Figure 3.6 would lay horizontally on the $Y=0$ axis, since the difference between the MOS and the ES would then be zero for all stimuli. It is clear that this is not the case. Instead, values tend to be negative when the ROI has a lower quality than the BG and positive when the situation is reversed. Moreover, this effect appears to be stronger as the difference in quality between the ROI and BG increases, the latter being especially the case for images with a higher quality in the BG than in the ROI (so, at the negative side of the X-axis in Figure 3.6). This trend is weaker when the

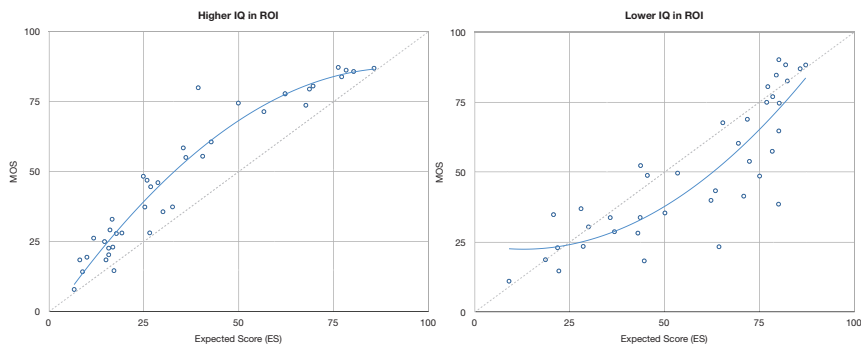


Figure 3.5 . Comparing the calculated ES to the subjectively collected MOS. Figure 3.5 (A) on the left represents images which had a higher IQ level in the ROI than in the BG, while those in Figure 3.5 (B) on the right had a lower quality in the ROI than in the BG.

quality of the background is strongly compromised in comparison to the quality of the ROI.

3.4.4 Modeling the influence of ROI on overall IQ

Using the data collected from the experiment, it is possible to estimate how much more important the ROI is in determining the overall quality of an image. To do that, we again look at equation (2) used to calculate the ES. Since we now know that there is a difference in how much each region affects the overall perceived quality, we calculated a more accurate Weighted Expected Score (WES) by introducing two weighting parameters to the equation, resulting in:

$$WES = MOS_{ROI} \cdot A_{ROI} \cdot w_{ROI} + MOS_{BG} \cdot A_{BG} \cdot w_{BG} \quad (3)$$

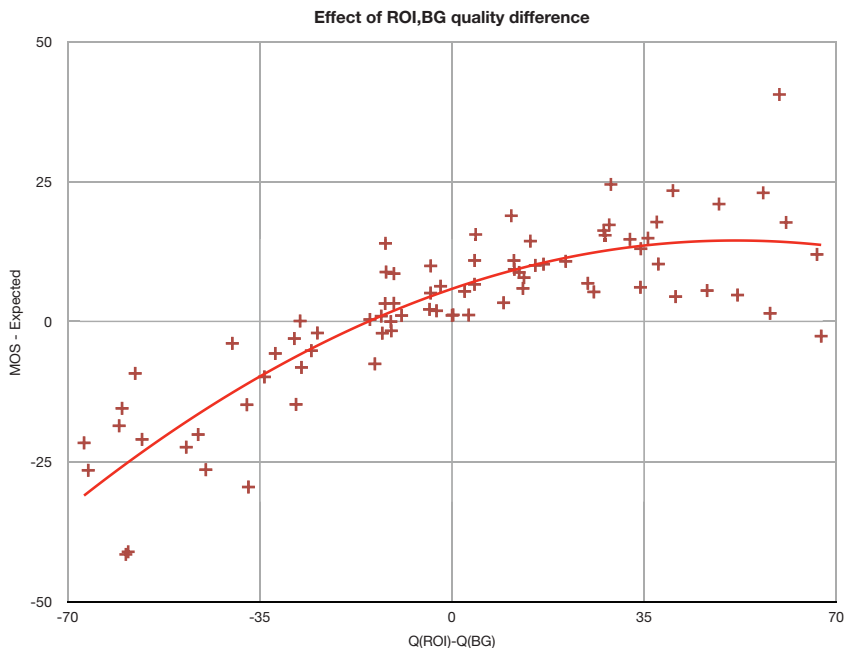


Figure 3.6 . The horizontal axis represents the difference in quality between the ROI and the BG regions, with the left side containing images with lower quality in the ROI than in the BG and right side images with higher quality in the ROI than in the BG. On the vertical axis, the difference between the MOS and the calculated ES is given. On the lower half, viewers scored the images lower than what would be expected from the ES, while in the upper half, viewers scored the images higher than what would be expected from the ES.

where w_{ROI} determines the weight of the ROI and w_{BG} the weight of the background region on the overall perceived quality. To calculate the values of these weights, a linear regression analysis was performed. The analysis used the MOS of the combined images as the dependent variable and the quality of each region multiplied by its corresponding area as the independent variables. The analysis returned the values $w_{ROI} = 3.80$ ($p < 0.001$, 95% confidence interval 3.34 to 4.27), and $w_{BG} = 0.65$ ($p < 0.001$, 95% confidence interval 0.58 to 0.71). The overall model fit had a $R^2 = 0.975$. The resulting relation is depicted in Figure 3.7, again for the two groups of data separately; i.e., in Figure 3.7(A) for the images with a higher IQ in the ROI than in the BG, and in Figure 3.7(B) for the images with a lower IQ in the ROI than in the BG.

To test the stability of this fit, the 80 stimuli of experimental phase 3 were split into two subgroups of 40 stimuli each. Both subgroups spanned the entire range of the quality scale. The two counterparts of each combined picture (i.e., one with a higher IQ in the ROI and the other with a higher IQ in the background) were joined in the same subgroup in order to avoid having the same image content repeated in both subgroups and thereby influencing the analysis. We then conducted a linear regression analysis in the same manner as described above on one of the two sub-groups. This analysis yielded the values $w_{sg-ROI} = 3.54$ ($p < 0.001$, 95% confidence interval 2.85 to 4.25), and $w_{sg-BG} = 0.68$ ($p < 0.001$, 95% confidence interval 0.58 to 0.78). The overall model fit had a $R^2 = 0.974$.

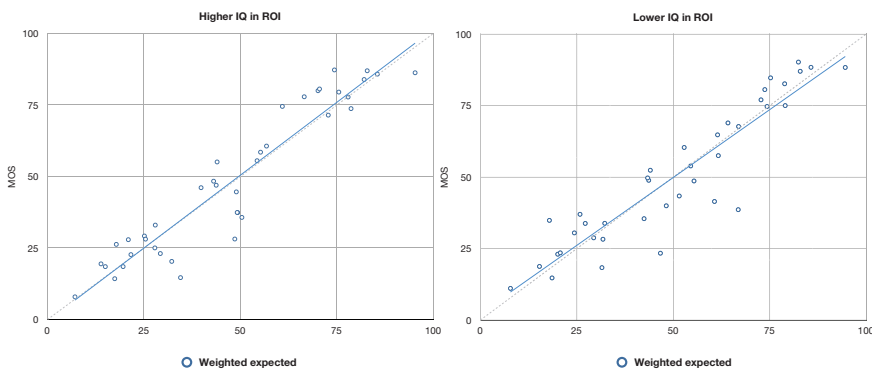


Figure 3.7. Comparing the calculated WES to the subjectively collected MOS. Figure 3.7 (A) on the left represents images which had a higher IQ level in the ROI than in the BG, while Figure 3.7 (B) on the right represents images which had a lower quality in the ROI than in the BG.

Both weighting factors were close in value to the ones found for the whole ensemble of stimuli, indicating that the result of the fit was not very sensitive to the particular selection of stimuli used. Subsequently, the new values of w_{sg-ROI} and w_{sg-BG} were used in equation (3) to calculate the WES of the second subgroup of stimuli (see Figure 3.8).

Finally, a similar plot as the one shown in Figure 3.6 is generated using the WES scores from the second subgroup of stimuli and the result is shown in Figure 3.9. The data points are now scattered around the $Y=0$ axis, indicating that with the proper weighting factors for the quality of the ROI and the BG, the overall quality of an image, locally varying in quality, can be predicted. By examining the values of w_{sg-ROI} and w_{sg-BG} , we can conclude that the quality of the ROI is about 5 times as important than the quality of the BG.

3.5. Discussion

The results of the experiment clearly show that when people assess image quality, they give greater significance to some regions of the image over others. It is not possible to obtain the overall image quality by simply averaging the quality of the different regions of the image. The subjectively measured MOS is clearly different from the estimated score obtained by averaging the quality of all image regions, even when taking

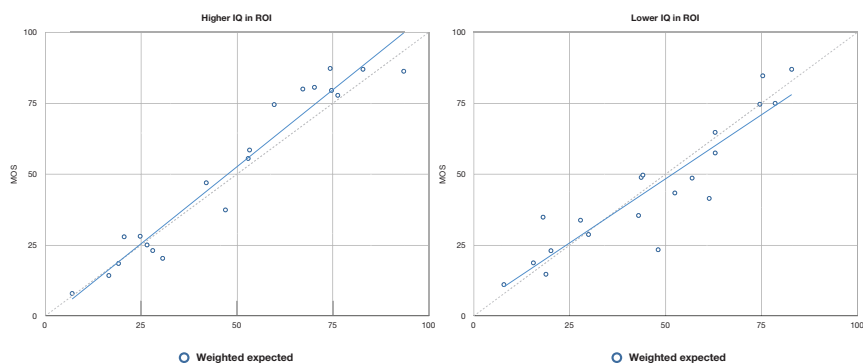


Figure 3.8. Comparing the calculated WES to the subjectively collected MOS for the stimuli belonging to the second subgroup. Figure 3.8 (A) on the left represents images which had a higher IQ level in the ROI than in the BG, while Figure 3.8 (B) on the right represents images which had a lower quality in the ROI than in the BG.

into account their relative area in the image. Results of phase 3 of the experiment demonstrated that there is a relation between the ROI quality and the MOS given to a combined stimulus. Stimuli which had a higher quality in the ROI mostly scored higher than expected. Stimuli with a lower IQ in the ROI scored lower than expected, though this effect was less clear. The extent to which the subjectively determined MOS differed from the ES was affected by the amount of quality difference between the two image regions, as was shown in Figure 3.6. Looking at the lower left corner of the figure, one can see that as the quality of the ROI gets more degraded, the MOS shifts further away below the ES. In the center of the figure, where the quality level in both image regions is very close, the MOS and the ES are very close as well. One can notice, however, that the MOS is slightly higher than the ES, where equality is expected. The shift may be attributed to differences in scoring between the first and third phases of the experiment, for example as a consequence of the different groups of participants used. As such, this shift may be considered as an estimation of the reproducibility of the quality scores over the whole experiment. Since this shift is considerably smaller than the difference between the MOS and ES measured at both ends of the quality (difference) range, we are convinced that the impact of the quality of the ROI on the overall quality is not an artifact of the limited reproducibility of the quality scores. As the quality of the ROI continues to increase (towards the right side of the graph), the difference between the MOS and the ES stops growing and even seems to diminish at the extreme end of the graph. This seems to suggest that even if the degradation is only present at the BG region, at a certain point the degradation becomes so bad that it plays a bigger role in determining the MOS for the entire image.

We also quantified the effect of the quality of the ROI on the overall quality using linear regression. The resulting values, i.e., $w_{ROI} = 3.803$ and $w_{BG} = 0.648$, suggest that the ROI region is more than 5 times more significant in determining the overall quality of an image than the BG region. This is even more impressive when one takes into account that the ROI in the used images occupied only 10-16% of the entire image area. Subsequent analysis also proved that this simple linear regression model already resulted in a considerable improvement in predicting the MOS value of stimuli with a different quality level in different areas of the image.

At this point, one can wonder whether changing how the ROI is defined will influence how much more significant it will be in determining the overall image quality [42]. Since we used in our experiment only images with a clear ROI which was always occupied by a human or an animal, the ROI was well defined. One should keep in mind that during the experiment, we used a white dot in the intermediate screen between stimuli and asked the participants to focus on this dot. This procedure helped us to refine the calibration of the eye-tracker, but may have introduced a center-bias in the saliency map as well. The effect of a fixed starting point on a center-bias in a saliency map may have occurred in the first fixations, but is expected to rapidly disappear over time (as extensively discussed in the literature [43]). Indeed, the example given in Figure 3.3 clearly illustrates that the saliency map calculated over the full presentation time of 8 seconds results in the expected ROI, away from the image center. Hence, we are convinced that we were able to

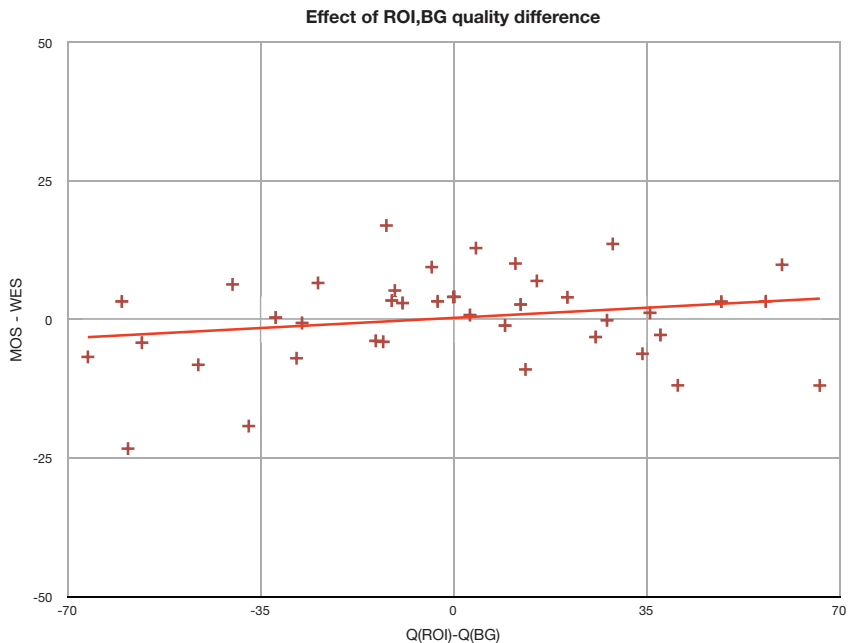


Figure 3.9. Illustration of how well the WES estimated from equation (3) predicts the subjectively obtained MOS for the second subgroup of combined stimuli used in phase 3 of the experiment. The horizontal axis represents the difference in quality between the ROI and the BG, while the vertical axis gives the difference between the MOS and the calculated WES.

accurately find the ROI in all images of our data set. In real practice, however, not all images have a clear ROI and the ROI has to be determined with an algorithm. Both aspects will reduce the accuracy with which the ROI is defined, and most probably the resulting ROI will be more scattered than what we used to establish our model. As a consequence, the importance of the quality of the ROI may be overestimated in real applications. Nonetheless, most images taken for entertainment purposes contain a main subject which constitutes an ROI, and the performance of ROI estimation algorithms is expected to further improve. Thus, the quality of the ROI being 5 times more important than the quality of the BG is a ratio that may be used for a wide variety of images uploaded online today. One can even consider an alternative approach in which the importance factor is decreased when the area of the ROI becomes bigger, but this approach would need further research to establish the relationship.

A limitation of our study is that all images in the database were degraded using JPEG compression. JPEG is still one of the de facto formats used to save digital imagery, but its specific nature may have affected the low importance of the quality of the BG to the overall quality score. Since the human eye is not good in detecting details in the periphery, it is possible that the observers are not capable of detecting the lower (or higher) quality as a consequence of blockiness or blur in the BG region. It is, however, good to realize that most post-processing manipulations address spatially detailed information, and so, the relative importance of the quality of ROI and BG is expected to hold for a broad range of signal processing algorithms. An exception may be artifacts with a temporal nature, since our peripheral vision is more sensitive to temporal artifacts than our foveal vision. Hence, temporal artifacts in the BG may be more easily detected or may be more annoying than temporal artifacts in the ROI, and so, these artifacts may change the relative importance in overall quality of ROI and BG.

Finally, it is good to realize that there are already several algorithms and image formats available that can encode different regions of an image at different quality levels [44-47]. There are also mechanisms available which can objectively estimate the ROI [48, 52]. It is therefore already

possible to implement the functionality that would optimize the encryption of different regions of images while predicting how it will affect the overall quality. This can be practical for saving content on mobile devices where memory-space is limited, or when making content for the web to save bandwidth.

3.6. Conclusions

Our results have proven that it is important to take the ROI of an image into consideration when trying to apply any manipulation to original image content with the aim of improving its overall IQ. If the manipulation lowers the quality of the ROI, then the perceived IQ of the entire image will be lower even if the majority (84% to 90%) of the image area has benefited from the manipulation. It is therefore risky to apply naive image enhancement algorithms which do not take the ROI into consideration.

When the quality of the ROI is higher than that of the BG, the viewers tend to give the image a higher quality score than its average quality level. We propose a simple model to estimate the overall perceived quality from the different quality levels of ROI and BG regions. This model illustrates that the quality of the ROI is about 5 times more important for the overall quality judgment than the quality of the BG.

It would be interesting to extend this study to video content. Since the dynamic nature of video lends greater significance to the ROI, we would expect the results to be more pronounced. On the other hand, video is expected to be more prone to temporal artifacts which may be more annoying in the BG (our peripheral vision) than in the ROI (our foveal vision). Further research is needed to establish which of the two video related aspects dominate the relative importance of quality in the ROI and BG.

References

1. Engeldrum, P. G. (2000). *Psychometric scaling: a toolkit for imaging systems development*. Imcotek Press.
2. Ghanbari, M. (2003). *Standard codecs: Image compression to advanced video coding* (No. 49). Iet.

3. Richardson, I. E. (2004). *H. 264 and MPEG-4 video compression: video coding for next-generation multimedia*. John Wiley & Sons.
4. Yuen, Michael, and H. R. Wu. "A survey of hybrid MC/DPCM/DCT video coding distortions." *Signal processing* 70.3 (1998): 247-278.
5. Farias, M. C., Moore, M. S., Foley, J. M., & Mitra, S. K. (2004). Perceptual contributions of blocky, blurry, and fuzzy impairments to overall annoyance. In *Electronic Imaging 2004* (pp. 109-120). International Society for Optics and Photonics.
6. Koh, C. C., Mitra, S. K., Foley, J. M., & Heynderickx, I. E. (2005). Annoyance of individual artifacts in MPEG-2 compressed video and their relation to overall annoyance. In *Electronic Imaging 2005* (pp. 595-606). International Society for Optics and Photonics.
7. Reeve III, H. C., & Lim, J. S. (1984). Reduction of blocking effects in image coding. *Optical Engineering*, 23(1), 230134-230134.
8. Girod, B. (1989). The information theoretical significance of spatial and temporal masking in video signals. In *OE/LASE'89, 15-20 Jan., Los Angeles. CA* (pp. 178-189). International Society for Optics and Photonics.
9. Liu, H., & Heynderickx, I. (2007). A simplified human vision model applied to a blocking artifact metric. In *Computer Analysis of Images and Patterns* (pp. 334-341). Springer Berlin Heidelberg.
10. Xia, J., Shi, Y., Teunissen, K., & Heynderickx, I. (2009). Perceivable artifacts in compressed video and their relation to video quality. *Signal Processing: Image Communication*, 24(7), 548-556.
11. Shen, M. Y., & Kuo, C. C. J. (1998). Review of postprocessing techniques for compression artifact removal. *Journal of Visual Communication and Image Representation*, 9(1), 2-14.
12. Luo, J., Chen, C. W., Parker, K. J., & Huang, T. S. (1996). Artifact reduction in low bit rate DCT-based image compression. *Image Processing, IEEE Transactions on*, 5(9), 1363-1368.
13. Van Zon, K., & Ali, W. (2001). Automated video chain optimization. *Consumer Electronics, IEEE Transactions on*, 47(3), 593-603.
14. Le Meur, O., Ninassi, A., Le Callet, P., & Barba, D. (2010). Do video coding impairments disturb the visual attention deployment?. *Signal Processing: Image Communication*, 25(8), 597-609.
15. Ninassi, A., Le Meur, O., Le Callet, P., & Barba, D. (2007, September). Does where you gaze on an image affect your perception of quality? Applying visual attention to image quality metric. In *Image Processing, 2007. ICIP 2007. IEEE International Conference on* (Vol. 2, pp. II-169). IEEE.
16. Liu, H., Engelke, U., Wang, J., Le Callet, P., & Heynderickx, I. (2013). How does image content affect the added value of visual attention in objective image quality assessment?. *IEEE Signal Processing Letters*, 20(4).
17. Wang, Z., & Bovik, A. C. (2009). Mean squared error: love it or leave

- it? A new look at signal fidelity measures. *Signal Processing Magazine, IEEE*, 26(1), 98-117.
18. Daly, S. J. (1992, August). Visible differences predictor: an algorithm for the assessment of image fidelity. In *SPIE/IS&T 1992 Symposium on Electronic Imaging: Science and Technology* (pp. 2-15). International Society for Optics and Photonics.
 19. Lubin, J. (1993, October). The use of psychophysical data and models in the analysis of display system performance. In *Digital images and human vision* (pp. 163-178). MIT Press.
 20. Safranek, R. J., & Johnston, J. D. (1989, May). A perceptually tuned sub-band image coder with image dependent quantization and post-quantization data compression. In *Acoustics, Speech, and Signal Processing, 1989. ICASSP-89., 1989 International Conference on* (pp. 1945-1948). IEEE.
 21. Watson, A. B., Hu, J., & MCGOWAN, J. F. (2001). Digital video quality metric based on human vision. *Journal of Electronic imaging*, 10(1), 20-29.
 22. Wu, H. R., & Yuen, M. (1997). A generalized block-edge impairment metric for video coding. *Signal Processing Letters, IEEE*, 4(11), 317-320.
 23. Wang, Z., Bovik, A. C., & Evan, B. L. (2000). Blind measurement of blocking artifacts in images. In *Image Processing, 2000. Proceedings. 2000 International Conference on* (Vol. 3, pp. 981-984). Ieee.
 24. Liu, H., & Heynderickx, I. (2009). A perceptually relevant no-reference blockiness metric based on local image characteristics. *EURASIP Journal on Advances in Signal Processing*, 2009, 2.
 25. Marziliano, P., Dufaux, F., Winkler, S., & Ebrahimi, T. (2002). A no-reference perceptual blur metric. In *Image Processing. 2002. Proceedings. 2002 International Conference on* (Vol. 3, pp. III-57). IEEE.
 26. Ferzli, R., & Karam, L. J. (2009). A no-reference objective image sharpness metric based on the notion of just noticeable blur (JNB). *Image Processing, IEEE Transactions on*, 18(4), 717-728.
 27. Liu, H., Klomp, N., & Heynderickx, I. (2010). A no-reference metric for perceived ringing artifacts in images. *Circuits and Systems for Video Technology, IEEE Transactions on*, 20(4), 529-539.
 28. Wang, Z., Bovik, A. C., Sheikh, H. R., & Simoncelli, E. P. (2004). Image quality assessment: from error visibility to structural similarity. *Image Processing, IEEE Transactions on*, 13(4), 600-612.
 29. Sheikh, H. R., & Bovik, A. C. (2006). Image information and visual quality. *Image Processing, IEEE Transactions on*, 15(2), 430-444.
 30. Sheikh, H. R., Bovik, A. C., & Cormack, L. (2005). No-reference quality assessment using natural scene statistics: JPEG2000. *Image Processing, IEEE Transactions on*, 14(11), 1918-1927.
 31. Venkatesh Babu, R., Suresh, S., & Perkiş, A. (2007). No-reference

- JPEG-image quality assessment using GAP-RBF. *Signal Processing*, 87(6), 1493-1503.
32. Redi, J. A., Gastaldo, P., Heynderickx, I., & Zunino, R. (2010). Color distribution information for the reduced-reference assessment of perceived image quality. *Circuits and Systems for Video Technology, IEEE Transactions on*, 20(12), 1757-1769.
 33. Liu, H., & Heynderickx, I. (2011). Visual attention in objective image quality assessment: based on eye-tracking data. *Circuits and Systems for Video Technology, IEEE Transactions on*, 21(7), 971-982.
 34. Barland, R., & Saadane, A. (2006, October). Blind quality metric using a perceptual importance map for jpeg-20000 compressed images. In *Image Processing, 2006 IEEE International Conference on* (pp. 2941-2944). IEEE.
 35. Rao, D. V., Sudhakar, N., Babu, I. R., & Reddy, L. P. (2007, March). Image quality assessment complemented with visual regions of interest. In *Proceedings of the International Conference on Computing: Theory and Applications* (pp. 681-687). IEEE Computer Society.
 36. Ma, Q., & Zhang, L. (2008, December). Image quality assessment with visual attention. In *Pattern Recognition, 2008. ICPR 2008. 19th International Conference on* (pp. 1-4). IEEE.
 37. Sadaka, N. G., Karam, L. J., Ferzli, R., & Abousleman, G. P. (2008, October). A no-reference perceptual image sharpness metric based on saliency-weighted foveal pooling. In *Image Processing, 2008. ICIP 2008. 15th IEEE International Conference on* (pp. 369-372). IEEE.
 38. Moorthy, A. K., & Bovik, A. C. (2009). Visual importance pooling for image quality assessment. *Selected Topics in Signal Processing, IEEE Journal of*, 3(2), 193-201.
 39. Recommendation, I. T. U. R. B. T. (2002). 500-11, Methodology for the subjective assessment of the quality of television pictures. *International Telecommunication Union, Geneva, Switzerland*, 4, 2.
 40. Buswell, G. T. (1935). *How people look at pictures* (pp. 142-144). Chicago: University of Chicago Press.
 41. Redi, J. A., & Heynderickx, I. (2011, September). Image quality and visual attention interactions: towards a more reliable analysis in the saliency space. In *Quality of Multimedia Experience (QoMEX), 2011 Third International Workshop on* (pp. 201-206). IEEE.
 42. Wang, J., Chandler, D. M., & Le Callet, P. (2010, February). Quantifying the relationship between visual salience and visual importance. In *IS&T/SPIE Electronic Imaging* (pp. 75270K-75270K). International Society for Optics and Photonics.
 43. Tseng, P. H., Carmi, R., Cameron, I. G., Munoz, D. P., & Itti, L. (2009). Quantifying center bias of observers in free viewing of dynamic natural scenes. *Journal of vision*, 9(7), 4.
 44. Said, A., & Pearlman, W. A. (1996). A new, fast, and efficient image

- codec based on set partitioning in hierarchical trees. *Circuits and Systems for Video Technology, IEEE Transactions on*, 6(3), 243-250.
45. Taubman, D. (2000). High performance scalable image compression with EBCOT. *Image Processing, IEEE transactions on*, 9(7), 1158-1170.
 46. Taubman, D. S., Marcellin, M. W., & Rabbani, M. (2002). JPEG2000: Image compression fundamentals, standards and practice. *Journal of Electronic Imaging*, 11(2), 286-287.
 47. You, J., Korhonen, J., Perkis, A., & Ebrahimi, T. (2011). Balancing attended and global stimuli in perceived video quality assessment. *Multimedia, IEEE Transactions on*, 13(6), 1269-1285.
 48. You, J., Perkis, A., & Gabbouj, M. (2010). Improving image quality assessment with modeling visual attention. In *Visual Information Processing (EUVIP), 2010 2nd European Workshop on* (pp. 177-182). IEEE.
 49. You, J., Korhonen, J., & Perkis, A. (2010, July). Attention modeling for video quality assessment: Balancing global quality and local quality. In *Multimedia and Expo (ICME), 2010 IEEE International Conference on* (pp. 914-919). IEEE.
 50. Long, M., & Tai, H. M. (2002, August). Region of interest coding for image compression. In *Circuits and Systems, 2002. MWSCAS-2002. The 2002 45th Midwest Symposium on* (Vol. 2, pp. 11-172). IEEE.
 51. Itti, L., Koch, C., & Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on pattern analysis and machine intelligence*, 20(11), 1254-1259.
 52. Rajashekar, U., Van Der Linde, I., Bovik, A. C., & Cormack, L. K. (2008). GAFFE: A gaze-attentive fixation finding engine. *Image Processing, IEEE Transactions on*, 17(4), 564-573.

4.

Examining the Effect of Task on Viewing Behavior in Videos Using Saliency Maps

Abstract

Research has shown that when viewing still images, people will look at these images in a different manner if instructed to evaluate their quality. They will tend to focus less on the main features of the image and, instead, scan the entire image area looking for clues for its level of quality. It is questionable, however, whether this finding can be extended to engulf videos considering their dynamic nature. One can argue that when watching a video the viewer will always focus on the dynamically changing features of the video regardless of the given task. To test whether this is true, an experiment was conducted where half the participants viewed videos with the task of quality evaluation while the other half were simply told to watch the videos as if they were watching a movie on TV or a video downloaded from the internet. The videos contained content which was degraded with compression artifacts across a wide range of quality. An eye tracking device was used to record the viewing behavior in both conditions. By comparing the behavior during each task, it was possible to observe a systematic difference in the viewing behavior which seemed to correlated to the video quality.

4.1. Introduction

Researchers have been studying visual attention deployment for many decades now [1,2]. This knowledge has been shown to be useful in a number of applications (e.g. [3,4]), especially when extracting image saliency information [5] through attention prediction models (e.g. [6,7,8]). One example is its implication in visual quality perception, which has been largely studied in images [9,10,11]. However, when it comes to videos there have been few efforts in trying to understand the relation between task, quality, and viewing behavior [12].

This research expands on earlier work performed on still images, focused on the effect that a task given to the observers can have on their viewing behavior [13]. In that work it was shown that the task does impact several spatial and temporal characteristics of the viewing behavior. On the other hand, similar work conducted on video material [12] has suggested that the viewing behavior is not affected by the given viewing task. Therefore, here we extend that study [13] to video material and focus on the task of scoring the quality of videos.

By following a similar methodology as the one used on still images, it is interesting to see whether the results are duplicated or whether the viewing task indeed has no effect on the viewing behavior. Furthermore, this study also looks at the interaction effects between the video encoding quality and the viewing behavior. The goal is to better understand how humans watch videos for entertainment. This information can then be used to optimize the video encoding algorithms to produce the best viewing experience while consuming less resources. For analyzing the viewing behavior, the methodology used here again extends on previous work done on still images [13] and, therefore, employ the use of saliency maps. The paper explores different algorithms for studying saliency maps [14] in an effort to determine which are more suitable for use in videos.

We designed a psychometric experiment to investigate the effects of task and visual quality on humans viewing behavior. We recorded the eye movements of a panel of observers while they were watching a set of distorted videos. These videos were degraded with compression artifacts across a wide range of video qualities. Half of the viewers was instructed

to evaluate the visual quality of each video. The other half was asked to watch the videos (as much as possible) as if they were freely watching them in a home setting. In the following we analyze the resulting eye-tracking data across different tasks and visual quality levels. To do so, we first convert them into saliency information, producing saliency maps averaged across all participants for each video and under each viewing condition. From (dis)similarities in saliency, we are able to detect analogies and differences in viewing behavior depending on task and visual quality level. The experimental methodology and the protocol are discussed in Section 4.2. The methodology followed to analyze the data is described in Section 4.3. Section 4.4 reports on the analysis of the eye-tracking data, which is then discussed in Section 4.5. Conclusions and future research are prospected in Section 4.6.

4.2. Methodology

4.2.1 Stimuli

A video database was created which consisted of 25 video segments with a duration of 20 seconds each. Since the main purpose of the experiment is to detect if there was a difference in viewing behavior, we wanted to use stimuli that have clearly identifiable natural saliency regions in order to make it easier to detect any differences that may result from changing the viewing task. It was assumed that highly dynamic scenes can seize the focus of the viewers under natural viewing conditions. Therefore, the video segments were extracted from action based movies (some sample frames are shown in Figure 4.1). From each video, two distorted versions were produced using an H.264 video encoder. These two versions were degraded to two different levels of quality. The x264 encoder was used as provided by the ffmpegX software [16]. The resulting videos had a resolution of 1280x720 pixels, and a frame rate of 25 frames per second. The coding parameters were not uniform across the videos, to allow a variety of quality levels to be judged by the observers. Eventually, the database included 50 distorted videos spanning a wide range of quality.

Two collections of stimuli were generated (collections I and II), each including only one of the two distorted versions generated for each video. The assignment of videos to collections was made randomly, therefore the videos in each collection spanned roughly the same range of quality.

Table 4.1. Between-subjects design of experiment to determine the impact of task on viewing behavior

		Task	
		Scoring	Free looking
Collection	I	Group 1	Group 3
	II	Group 2	Group 4

To give an idea of the wealth of this quality range, Figure 4.1 shows a sample frame on the left of one of the videos encoded with the highest bitrate (1237 bit/s), and another on the right which is among those encoded with the lowest bitrate (209 bit/s).

4.2.2 The experiment setup

Given that the focus of the experiment was on the viewing task, a between subjects design was chosen, where half the participants looked freely at the videos and the other half was asked to evaluate their visual quality. This meant that every viewer saw each video segment only once, and ensured that there was no memory effect influencing the viewing behavior. With 2 viewing conditions and 2 viewing tasks, the experiment took the form of a 2x2 design requiring 4 groups of observers, as shown in Table 4.1. Each group counted 12 participants, for a total of 48. Participants included master, graduate, and postgraduate students of the Electrical Engineering, Mathematics and Computer Science (EEMCS) faculty building at the Delft University of Technology (TU Delft).



Figure 4.1. Video segments used for the experiment were taken from action movies and were chosen to have highly dynamic sequences. Varying the bitrate used to compress the videos from higher bitrates (left 1237 bit/s) to lower ones (right 209 bit/s) gave a wide range of quality for the generated videos.

The videos were displayed using a late 2008 MacBook on a 17-inch CRT monitor with a resolution of 1280x960 pixels. The experiment was controlled from a remote computer with its monitor positioned so that it would not interfere with the participant's task. In order to avoid outside elements interfering with the results, the experiment was carried out in a controlled environment inside the Delft Experience Lab located in the EEMCS faculty building at the TU Delft. Only the experimenter and the viewer were present while performing the experiment. The illumination level was kept constant. Eye movements were recorded binocularly at 250 Hz with a video-based infrared eye tracker (SR-Research, EyeLink-II). The eye-tracker data was saved to disk for off-line analysis. The experiment setup is shown in Figure 4.2.

4.2.3 The experiment protocol

Of the four groups of participants (Table 4.1), the first two groups (1 and 2) as well as the last two (3 and 4) went through identical protocols but watching a different collection of videos. In all cases, participants were given a printed description of the experiment and a list of the instructions they needed to follow. They were then seated so that their viewing distance measured 60 [cm] from the display plain. After being fitted with the head mounted eye tracker, the experimenter ran a 9-point calibration for the gaze location.



Figure 4.2. Experimental setup. The viewer watched the videos on a CRT monitor while wearing a head mounted eye tracking device. The experimenter ran the experiment from a non-intrusive position.

Groups 1 and 2 watched the first and second collection of videos respectively. They performed a scoring task which used a single stimulus numerical scaling setup [15]. After the calibration, observers were shown 4 training videos (Which were not a part of the collections I and II), which were representative of the entire range of quality used in the video collections. The training videos helped the participants in getting acquainted with the user interface of the experiment and gave them an idea of the range of quality they could use for scoring the videos. For every video to be evaluated, during both the training and actual experiment, the following steps were performed. The participants first saw a drift correction screen, which helped the head mounted eye tracker compensate for any shifts in its position. To do that, they simply had to fixate their gaze at a red dot in the center of the screen and press the space bar. They were then shown a 20 second video segment. This was followed by a scoring window with a continuous slider going from 0 to 10 with the labels 'poor' at the lower end and 'excellent' at the other. Once a score was chosen, the process was repeated with the drift-correction screen followed by another video. After the first four videos, a dialog window was shown indicating that the training session was over and the process then continued with the videos from one of the collections. Participants from each group always saw the same 25 video segments from the assigned collection, but the order in which the videos were shown was randomized in order to avoid any bias (i.e., learning or fatigue effects) in the results.

Participants in groups 3 and 4 were again shown collections I and II respectively. They followed the same protocol described above except that they were not given the scoring window after each video. Of course, the experiment instructions they had were also adjusted so that they were told to only watch the videos as if they were watching TV or a video downloaded from the Internet.

4.3. Analyzing the data

The eye tracker collected fixation and saccade information. Smooth pursuit eye behavior exhibited when the viewer followed the movements of objects on the screen was registered by the eye tracker as fixations. In order to perform a precise spatial analysis of attention deployment, we

decided to transform fixation data into saliency information, adapting the procedure proposed in [14]. First, per each video sampling point we grouped in a single fixation map all the fixation locations of all observers. This resulted into as many fixation maps as sampling points per each video, giving a too fine granularity for the purposes of our analysis. Thus, videos were divided in coarser slots of 1 second each, and fixation maps were averaged over these time slots, resulting in 20 fixation maps per video. These fixation maps were finally converted to saliency maps to better reflect the characteristics of human vision. In particular, a Gaussian patch with a width σ approximating the size of the fovea (about 2° visual angle) was applied to each fixation. A mean saliency map that takes into account all fixations of all subjects was calculated as follows:

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

where $S_{t,i}(k, l)$ indicates the saliency map for stimulus l_i for the time slot $t \in [1, 20]$ of size $M \times N$ pixels (i.e. $k \in [1, M]$ and $l \in [1, N]$), (x_j, y_j) indicates the spatial coordinates of the j th fixation ($j=1 \dots T$), T is the total number of all fixations over all subjects in that time period, and σ indicates the standard deviation of the Gaussian. The intensity of the resulting saliency map is linearly normalized to the range $[0, 1]$. Each of these maps specifies the saliency distribution over a specific time slot of a specific video. In other words, it is a representation of the probability, pixel by pixel, that at a given time slot, for a given video and task, the average observer will fixate on a specific pixel. This process was repeated for each 1 second time slot for each video sequence. Hence, we obtained in total 20 (slots) x 25 (videos) x 2 (collections) x 2 (tasks) = 2000 saliency maps.

4.4. Results

The scores collected from groups 1 and 2 for each encoded video sequence were processed to calculate one mean-opinion-score (MOS) [17] representing the subjective quality level of that video segment. Graphs shown in Figure 4.3 illustrate the range of subjective quality used in the videos. The top graph shows the MOS values for the two degraded

versions of each video segment included in collections I and II. The graph clearly shows that the videos in both groups were well distributed across the used range of quality. It also conveys that the difference in quality between the two encoded versions of each video segment varies across the generated collections. The graph on the bottom of Figure 4.3 plots the MOS as a function of the video bitrate. Despite the wide spread that can be observed in the middle of the scale, a linear relation can be observed between the two values approximated by the stapled line plotted in the graph.

As previously stated, we are interested not only in the impact of task on visual attention but also on that of the quality level. Thus, it is useful for the analysis to sort the collected data into groups depending on the MOS quality levels. We redistributed the videos from the two collections into a High Quality (HQ) and Low Quality (LQ) groups. For each video, the version that received the lower MOS is collected in the LQ group and the one with the higher quality is assigned to the HQ group. By taking the 2 tasks into consideration, we end up with 4 sets of data as shown in Table 4.2.

In order to see whether the quality scoring task (the independent variable) affected the viewing behavior, the saliency maps collected under each task are compared to measure the level of similarity among them. Many approaches for analyzing similarities between saliency maps have been proposed in similar research studying attention data in still images [14]. In the sake of being thorough, 4 of the measures proposed in the literature [14] are computed here for the saliency maps in order to find the most appropriate measures capable of highlighting differences in the viewing behavior in videos. These approaches are: Linear Correlation Coefficient (LCC), Kullback-Leibler divergence (KLD), Normalized Scanpath Saliency (NSS), and the Structure Similarity Index (SSIM) [18]. A value of $LCC = 1$ indicates identical maps, while $LCC = 0$ indicates uncorrelated maps. This is also the case for SSIM with a range of [0-1] and higher scores indicating more similarity. The NSS will return a value greater than zero if there is a greater correspondence between the two saliency maps than expected by chance. The NSS value of zero would mean there is no such correspondence and a value of less than zero would mean there is anti-correspondence between the saliency maps.

Table 4.2. The redistributed data sets used in the analysis of the results

		Task	
		Scoring	Free looking
Quality	Higher	S-HQ	F-HQ
	Lower	S-LQ	F-LQ

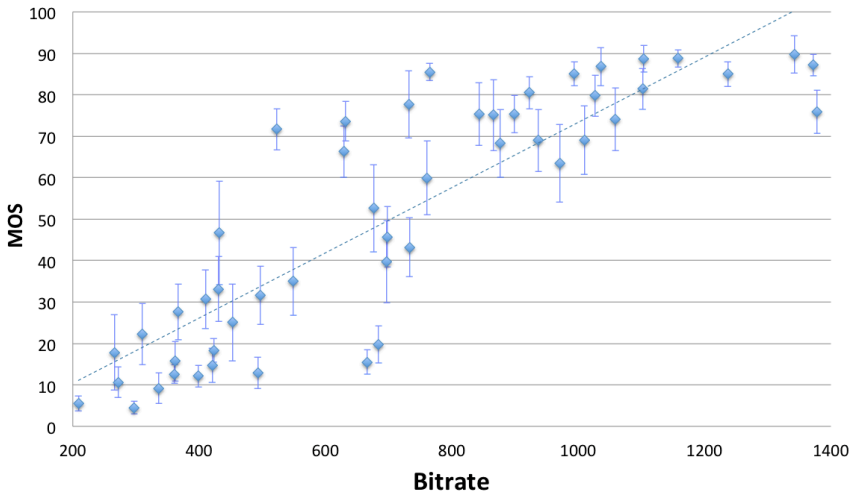
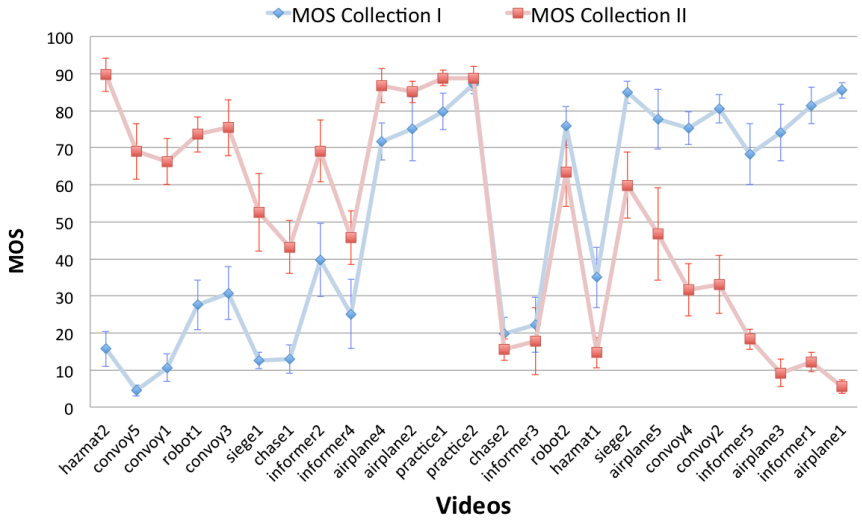


Figure 4.3. Two graphs illustrating the range of quality in the used videos. On the top are the MOS values for the two versions of videos in each collection sorted by the difference in MOS. On the bottom the MOS for all 50 videos is plotted against the bitrate used to encode the videos.

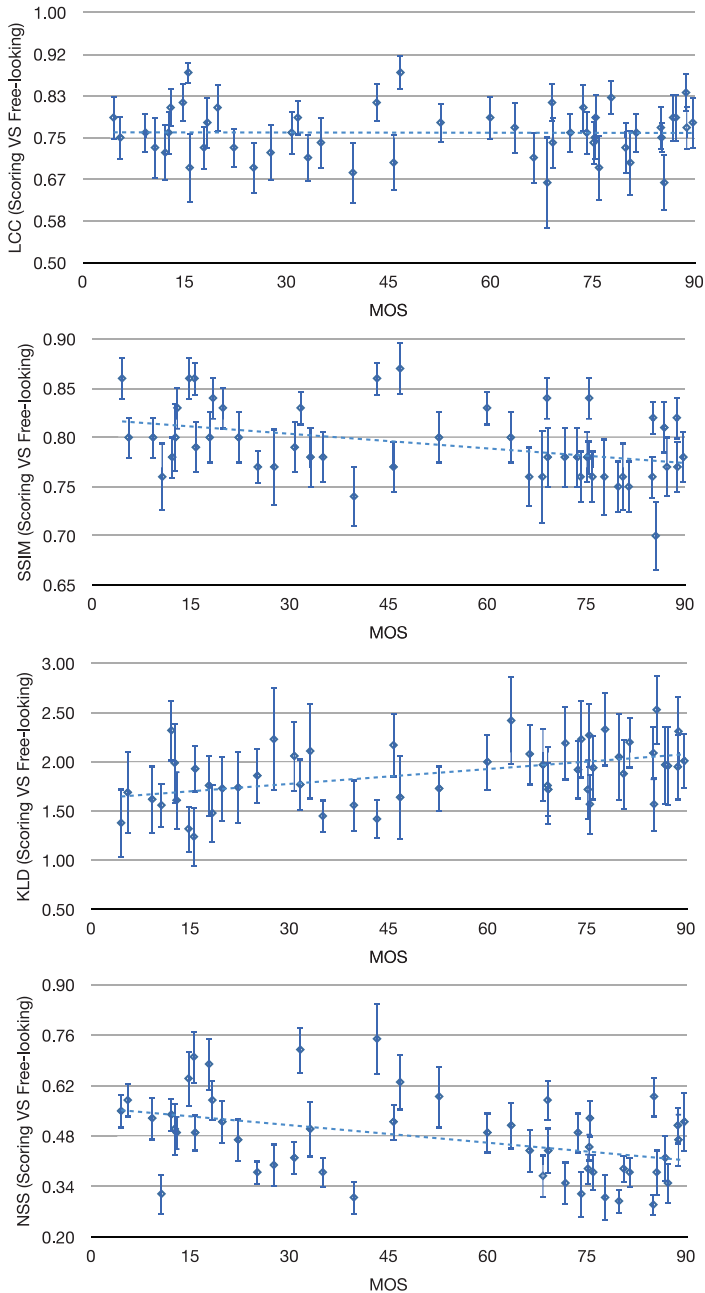


Figure 4.4. Four different similarity measures are applied on 50 videos. They compared the saliency maps collected while scoring and free-looking, plotted against the MOS values. LCC and SSIM have the range $[0-1]$ with higher values indicating more similarity. With NSS, a value of 0 represents no similarity with higher values representing more similarity. KLD is the opposite starting at 0 for perfect similarity, with higher values meaning less similarity.

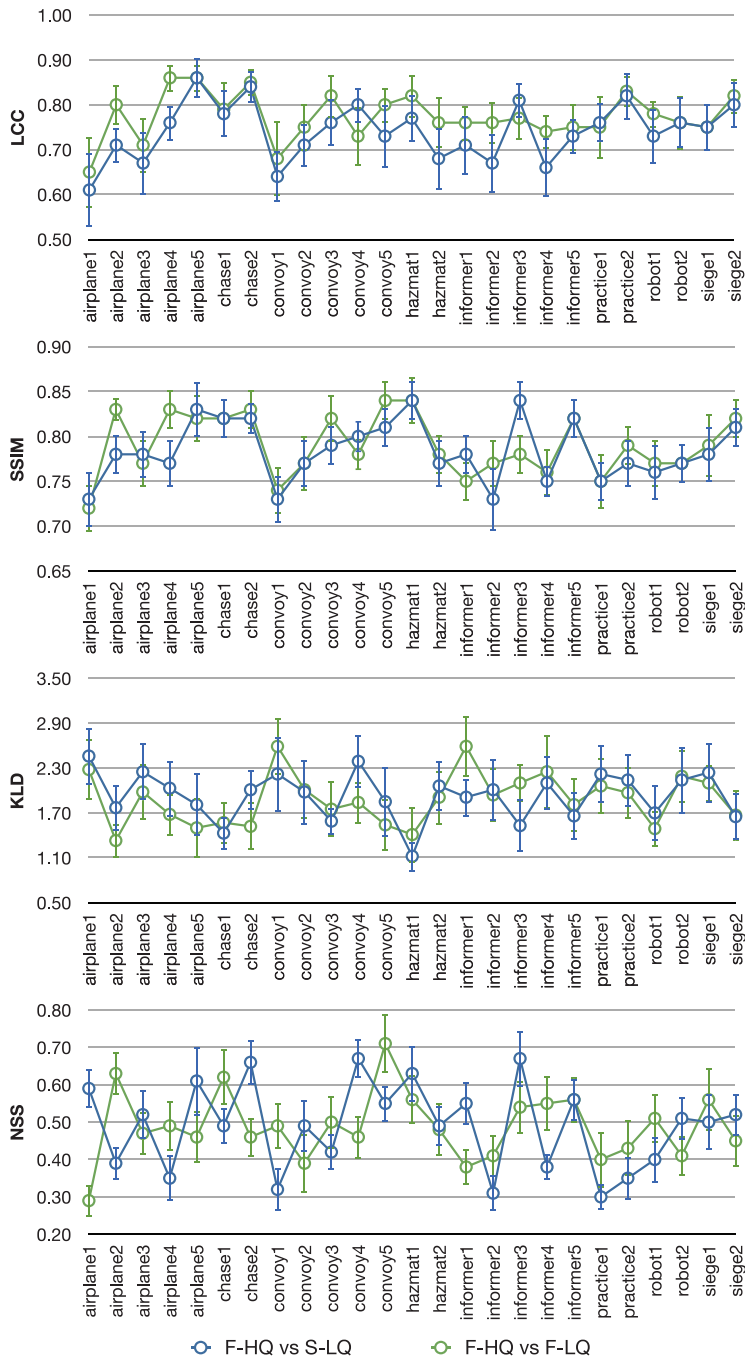


Figure 4.5. By taking the free looking high quality as a reference. The similarity with the low quality videos is measured for both the free looking and the scoring conditions

Finally, the KLD is a positive quantity that increases with the dissimilarity of the maps, and $KLD = 0$ only in the case of identical maps. More details on what each of these method measures and how they are calculated can be found in the literature [14].

First the LCC was computed for each pair of saliency maps corresponding to the same video and time slots for each of the two tasks. In other words, the LCC was calculated for each video for each saliency map in S-HQ and the corresponding saliency map in the F-HQ dataset. The same process is then repeated for the S-LQ vs F-LQ, S-HQ vs S-LQ, and F-HQ vs F-LQ data sets. In this way, we had, per each second and per each encoded video, an indication of the similarity of viewing behavior under different tasks and under different quality levels. This process was then repeated using the other three similarity measures (SSIM, KLD, and NSS). Figure 4.4 shows the similarity values for the task effect for all 4 similarity measures. Each data point represents the average value over 20 time slots for each of the 50 encoded video. The error bars represent the 95% confidence interval for the 20 time slots for each video segment. The similarity scores are plotted against the subjective MOS quality and a trend line is plotted representing the best linear function fitting the data.

Since the KLD scale is reversed (*higher* values indicate *less* similarity), the trend lines of the SSIM, KLD, and NSS seen in Figure 4.4 indicate that the viewing behavior becomes less similar as the quality level increases. The trend line for the LCC values is virtually flat and neither support nor oppose this observation. When it comes to explaining the effect of the task however, it may be helpful to have a reference measure to compare the data against.

We next take the free-looking data collected while viewing the high quality segments (F-HQ) as a reference. With no assigned task and the relatively higher level of quality, it represents the closest saliency data to the natural saliency of the original 25 video segments. We use this reference data to examine the similarity of the viewing behavior for the lower quality versions of the same video segments under free-looking (F-LQ) and scoring (S-LQ) conditions. Figure 4.5 shows this comparison using the 4 similarity measures.

As all observers are human, It is possible that there are differences between observers and how they would view the same stimulus in a different manner. This difference is known in the literature as inter-observer variability [14]. In order to determine whether the observed differences in viewer behavior were a result of inter observer variability or whether they are the results of a systematic shift in viewing behavior, saliency maps collected under the same viewing conditions are analyzed against each other for similarity to establish a so called upper empirical similarity limit. (UESL) [14].

Since the free looking high quality data is used as a reference for the analysis, this data is also chosen to be used for calculating the UESL. The data collected from the 12 participants is split in two subgroups. Two groups of saliency maps are constructed from each subgroup of viewing data. The similarity scores between the saliency maps are then calculated. This process is repeated 10 times and the average similarity score is calculated to establish the UESL. In order to keep the analysis fair, the rest of the similarity scores shown in Figure 4.5 are recalculated with half of the number of participants as well. The results of this analysis for the four different similarity measures is shown in Figure 4.6. Statistical analysis of the S-LQ and F-LQ revealed no significant difference viewing behavior between the two conditions.

4.5. Discussion

Looking at the graphs shown in Figure 4.4 one can safely say that all similarity measures have a similar level of performance considering the spread of the data points and the size of the resulting confidence intervals. It is therefore difficult to single out any of them as the best or worst. Hence, the analysis of the results looks at all the four similarity measures.

Figure 4.4 also shows that there is a trend of lower similarity in behavior as the quality of the videos increases. In other words, when looking at a better quality video, the viewing task has a greater influence on behavior. One possible explanation is that higher quality videos contain less artifacts, and therefore requires the viewers to actively search for clues of quality by ignoring the natural salient regions and scanning the entire

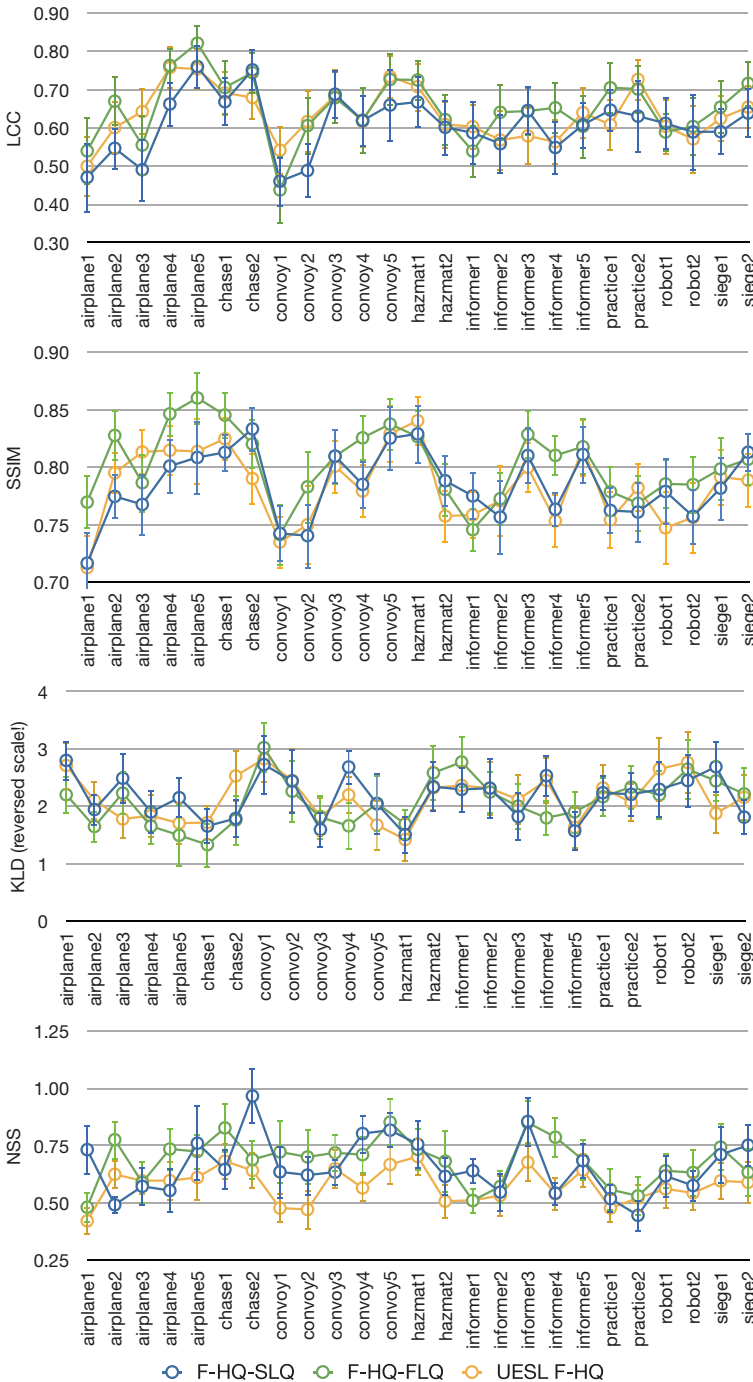


Figure 4.6. Similar analysis to that in Figure 5, but now with the UESL analysis. By taking the free looking high quality as a reference, the similarity with the low quality videos is measured for both the free looking and the scoring conditions

video area. This trend, however, is not visible when analyzing the data using LCC (represented by the flat trend line in Figure 4.4 top), which may be an indication that it is not a very strong trend.

As mentioned before, freely viewing the higher quality versions of the video segments (F-HQ) gives the closest viewing characteristics to natural saliency. For that reason it is used as the reference value for the analysis shown in Figure 4.5. To see whether the task has an effect on the viewing behavior, we measure the similarity of the reference data (F-HQ) to that collected with F-LQ. This is then repeated for the reference data (F-HQ) and S-LQ to see if changing the task gives a different result. Since the difference of quality is equal in both cases, we assume that it does not play a significant role in this comparison.

The similarity analysis is then repeated for all data taking the UESL into account. The results of that analysis are shown in Figure 4.6. In the figure, it is possible to see that the UESL does not fall above all other data points as expected. This is the case for all four similarity measures and is more apparent with NSS where the UESL values mainly fall below the other values of the analysis. This result does not fall in line with what is expected from a UESL analysis. It simply seems that the data collected within the parameters of this experiment are not suitable for a UESL analysis.

A statistical analysis of the results showed no significant difference between the free looking and scoring tasks. It was therefore not possible to replicate the results found in earlier research performed on still images [13] This may be due to the highly dynamic character of the video segments used in this experiment which makes it difficult to deviate attention from the natural scene saliency. It may also be the result of the different method used for analyzing the data, since even within this data set not every used similarity measure was able to detect the difference in viewing behavior.

4.6. Conclusions

In this paper we examined the effect of the given task on the viewing behavior when watching videos. By tracking the eye movements of the observers under both conditions it was possible to generate saliency

maps representing the viewing characteristics under each condition. A set of different possible measures for saliency similarity [14] was used to analyze the data. From these measures, the SSIM was the only one sensitive enough to detect all observed effects.

Using these similarity measures, it was possible to see a trend of the task having a stronger effect on viewing behavior if the video has a higher level of encoding quality. Viewers seem to focus more on searching for clues of the image quality if no clear artifacts are present. However, it was not possible to detect a systematic difference in the viewing behavior when viewers were given the task of scoring the quality of the videos. The UESL analysis did not yield any solid conclusions other than that the analysis is not suitable for analyzing the data of this type of experiments.

With regards to future work, we are looking deeper into the video segments using the calculated similarity measures to find specific scenes that exhibit higher sensitivity to the given task and try to identify common characteristics in these scenes. Additionally, it may be interesting to look into applying other image saliency analysis techniques to see how they perform on this data. The saliency data generated in this experiment has also been made available on the Internet for other researchers in the field [19] to offer them the chance to use it in related research.

References

1. Buswell, G. T. (1935). *How people look at pictures* (pp. 142-144). Chicago: University of Chicago Press.
2. Yarbus, A. L. (1967). *Eye movements and vision* (Vol. 2, No. 5.10). L. A. Riggs (Ed.). New York: Plenum press.
3. Liu, H., & Heynderickx, I. (2011). Visual attention in objective image quality assessment: based on eye-tracking data. *Circuits and Systems for Video Technology, IEEE Transactions on*, 21(7), 971-982.
4. Adams, M. D. (2001). The JPEG-2000 still image compression standard. *ISO/IEC JTC, 1*, 2001-09.
5. Koch, C., & Ullman, S. (1987). Shifts in selective visual attention: towards the underlying neural circuitry. In *Matters of Intelligence* (pp. 115-141). Springer Netherlands.
6. Itti, L., Koch, C., & Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on pattern analysis and machine intelligence*, 20(11), 1254-1259.

7. Le Meur, O., Le Callet, P., & Barba, D. (2007). Predicting visual fixations on video based on low-level visual features. *Vision research*, 47(19), 2483-2498.
8. Bruce, N. D., & Tsotsos, J. K. (2009). Saliency, attention, and visual search: An information theoretic approach. *Journal of vision*, 9(3), 5.
9. Vuori, T., Olkkonen, M., Pölonen, M., Siren, A., & Häkkinen, J. (2004, October). Can eye movements be quantitatively applied to image quality studies?. In *Proceedings of the third Nordic conference on Human-computer interaction* (pp. 335-338). ACM.
10. Ninassi, A., Le Meur, O., Le Callet, P., Barba, D., & Tirel, A. (2006, September). Task impact on the visual attention in subjective image quality assessment. In *Proceedings of European Signal Processing Conference*.
11. Redi, J., Liu, H., Zunino, R., & Heynderickx, I. (2011, February). Interactions of visual attention and quality perception. In *IS&T/SPIE Electronic Imaging* (pp. 78650S-78650S). International Society for Optics and Photonics.
12. Le Meur, O., Ninassi, A., Le Callet, P., & Barba, D. (2010). Overt visual attention for free-viewing and quality assessment tasks: Impact of the regions of interest on a video quality metric. *Signal Processing: Image Communication*, 25(7), 547-558.
13. Alers, H., Bos, L., & Heynderickx, I. (2011, September). How the task of evaluating image quality influences viewing behavior. In *Quality of Multimedia Experience (QoMEX), 2011 Third International Workshop on* (pp. 167-172). IEEE.
14. Redi, J. A., & Heynderickx, I. (2011, September). Image quality and visual attention interactions: towards a more reliable analysis in the saliency space. In *Quality of Multimedia Experience (QoMEX), 2011 Third International Workshop on* (pp. 201-206). IEEE.
15. Recommendation, I. T. U. R. B. T. (2002). 500-11, Methodology for the subjective assessment of the quality of television pictures. *International Telecommunication Union, Geneva, Switzerland*, 4, 2..
16. <http://www.ffmpeg.com>
17. Sheikh, H. R., Sabir, M. F., & Bovik, A. C. (2006). A statistical evaluation of recent full reference image quality assessment algorithms. *Image Processing, IEEE Transactions on*, 15(11), 3440-3451.
18. Wang, Z., Bovik, A. C., Sheikh, H. R., & Simoncelli, E. P. (2004). Image quality assessment: from error visibility to structural similarity. *Image Processing, IEEE Transactions on*, 13(4), 600-612.
19. Alers, H., Liu, H., Redi, J., & Heynderickx, I., "TUD Video Quality Database: Eye-Tracking Release 2", http://mmi.tudelft.nl/iqlab/video_task_eye_tracking_1.html

5.

Quantifying the Importance of Preserving Video Quality in Visually Important Regions

Abstract

Advances in digital technology have allowed us to embed significant processing power in everyday video consumption devices. At the same time, we have placed high demands on the video content itself by continuing to increase spatial resolution while trying to limit the allocated file size and bandwidth as much as possible. The result is typically a trade-off between perceptual quality and fulfillment of technological limitations. To bring this trade-off to its optimum, it is necessary to understand better how people perceive video quality. In this work, we particularly focus on understanding how the spatial location of compression artifacts impacts visual quality perception, and specifically in relation with visual attention. In particular we investigate how changing the quality of the region of interest of a video affects its overall perceived quality, and we quantify the importance of the visual quality of the region of interest to the overall quality judgment. A three stage experiment was conducted where viewers were shown videos with different quality levels in different parts of the scene. By asking them to score the overall quality we found that the quality of the region of interest has 10 times more impact than the quality of the rest of the scene. These results are in line with similar effects observed in still images, yet in videos the relevance of the visual quality of the region of interest is twice as high than in images. The latter finding is directly relevant for the design of more accurate objective quality metrics for videos, that are based on the estimation of local distortion visibility.

5.1. Introduction

With the booming of digital video [1], and especially of its distribution via IP networks, compression algorithms that can reduce the video size have become crucial for video distribution and delivery to the user. Video coding is indeed essential to allow storing and transmitting, within existing bandwidth and storage constraints, the huge amount of video material available in online communities and repositories. Despite the remarkable progress that video coding has made in the past few years (i.e., first MPEG-2 [2], then H.264/AVC [3] and the most recent HEVC codec [4] for ultra-high definition video), most commercially used compression schemes are lossy, invariably causing a decrease in the perceptual quality of the eventually delivered video. Post-processing algorithms, such as sharpening or deblocking filters [5,6], can to some extent compensate for the quality loss. However, though generally beneficial, these filters also may cause visible artifacts, e.g., by introducing blur when reducing the visibility of blocking artifacts. In fact, these filters are usually designed to be applied uniformly to entire video frames. So, even if local video texture masks the visibility of compression artifacts, still post-processing filters are used, possibly hampering the video quality in regions where it was still acceptable.

Hence, for quality control and enhancement to be effective it is necessary to have some mechanisms in place that allow us to estimate as accurately as possible the visibility and the annoyance of artifacts, on a local basis. A wide amount of research is currently dedicated to the automated estimation of artifact annoyance and related perceived quality of video, by means of so-called Objective Quality Metrics. These typically estimate artifact visibility and/or annoyance depending on image/video characteristics and the related perceptual processes in the human visual system [7,8]. Lately, there have been attempts at designing objective quality metrics that also include information on the visual importance of the position of the artifacts in the video. It has been commonly assumed that artifacts in visually important areas of the image/video may be more noticeable and weigh more in the overall quality judgment. Similarly, artifacts in background regions may be less visible and thus neglected by users [9].

The notion of visual importance is tightly coupled to that of visual attention. Research has shown that humans have limited resources to allocate to the perception of visual stimuli [10]. In order to cope with the high level of complexity that visual scenes can contain, the human visual system directs the focus of vision to conspicuous regions in the scene, deemed to carry the most relevant visual information [11,12]. Two types of visual attention mechanisms are distinguished: driven by scene saliency (bottom-up mechanisms) or visual importance (top-down mechanisms). Bottom-up saliency refers to the fact that (local) contrast in scene features such as luminance, color and orientation can attract people's attention, typically unconsciously. Top-down mechanisms are instead triggered by a more conscious process and may require understanding of the scene. They regulate the phenomenon according to which scene elements with a prominent semantic connotation (such as a human face or a single dog in a picture of an empty field) receive more attention than other regions of that scene [13]. Visually important regions, also known as Regions Of Interest (ROI), can be considered as the scene elements carrying the most relevant visual information, while the remaining regions contain simply background information that is there to support the ROI or provide context.

The link between visual importance and artifact annoyance has been thoroughly investigated and proved for images [14-17]. In multiple cases, the integration of visual importance information into objective quality metrics for images has been shown to be beneficial in terms of improving the quality prediction accuracy. These results are hardly generalizable to video, though. When attending video material, the viewing behavior is different from that deployed when viewing images. When looking at still images, humans rapidly fixate on a specific region in the image, scan it, and then move on to other regions. In videos, fixations on moving objects are enabled through smooth pursuit eye movements [18]. This allows for longer fixation durations, which may also be needed to follow the gradual change of the content in the ROI. As a result, the ROI tends to capture most of the visual attention [9], and that leaves less time to scan the background regions of the scene. However, since video artifacts are also dynamic in nature, one can also assume that they will be more distracting than artifacts in still images, thereby becoming salient. One would

Table 1. Between-subjects design of experiment to determine the impact of task on viewing behavior

		Task			
		Scoring		Free looking	
Video Collection	I	Phase 1	Group 1	Phase 2	Group 3
	II		Group 2		Group 4
	Comp I	Phase 3	Group 5	—	
	Comp II		Group 6	—	

therefore wonder whether the quality of the ROI in videos has in fact a larger influence on overall perceived quality, or whether the appearance of distracting artifacts in background regions is instead most disruptive for overall quality perception [17].

Research in [19] has actually shown that packet-loss artifacts, which are strongly localized in space and time, are more annoying when appearing in the ROI than when appearing in the background of the same video. The extent to which distortions in the ROI would be more annoying than those in the background was, however, not quantified, nor was the difference in annoyance proven to be true for diffused artifacts such as blockiness or blur. Evidence of a reduced distraction power of background artifacts in videos with respect to images was also found. Salient video areas were found to be mostly unaltered when compression artifacts would appear outside the ROI [20]. To the best of the authors' knowledge, however, it is still unclear whether this reduced distracting power has an influence on the overall quality assessment, and if it does, to what extent background artifacts contribute to the formulation of the overall quality judgment in the user. Unclear results in this sense have also been found when attempting at incorporating visual importance information in objective quality metrics for videos. This has been typically done by weighing estimated blockiness strength depending on the visual importance of its location. This practice has led to marginal if any improvement [9,21], leaving unclear the role played by visual importance in video quality perception, when dealing with distributed, compression-generated artifacts.

Therefore, the current study aims at shedding some light on the relationship between the annoyance of blocking artifacts and the visual importance of their spatial location. In particular, we are interested in determining: (1) whether artifacts located in the ROI of the video are significantly more annoying than those in the background and, if so (2) to what extent their annoyance contributes to the final quality evaluation, compared to the annoyance of artifacts located in background areas. We investigate these two questions by using a similar methodology as applied to images in [17]. We first determine the visual importance of different spatial regions in a video by means of eye-tracking. We then create a video database containing videos with a different compression level in the ROI and in the background. Based on subjective evaluations of these videos, we establish a model for the annoyance of artifact visibility in regions with prominent visual importance. The resulting model illustrates how, for a ROI as small as 10% of the video size, the artifacts located in it have a 10 fold higher importance in the determination of the final video quality than those located in the background.

Section 5.2 of this chapter explains how the videos were created and which experimental methodology was used. We also explain how the eye tracker data were used to identify the ROI of the shown videos. Section 5.3 lists the results, which are then discussed in Section 5.4, followed by the conclusions and further thoughts in Section 5.5.

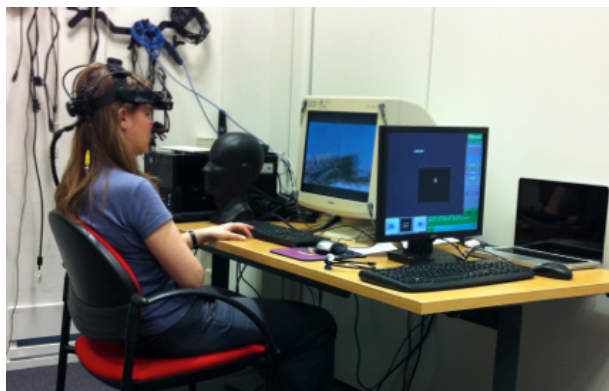


Figure 5.1. Illustration of the experimental setup, showing on the left-hand side a participant watching the videos on a CRT monitor while wearing a head mounted eye tracking device. The experimenter ran the experiment from a non-intrusive position, and assisted the experimental task without moving or interfering with it, in order to prevent causing distractions to the participant.

5.2. Methodology

Visual attention mechanisms can be studied via the analysis of eye movements. Eye-trackers are typically used to this purpose; they track fixations and saccades during scene observation. Fixations are known to be driven by bottom-up saliency. Recent literature has shown that for most images there is also a relation between visual fixation patterns and ROI or visual importance, with the early fixations being more predictive for the ROI than later fixations [22,23]. Hence, these findings justify the use of fixation density maps to determine the ROI, as is exploited in this paper. In the following, we provide the required details for all the three phases of the experiment.

5.2.1 Experimental setup

Our study was conducted based on a three-phase experiment using a between-subject design (see Table 1). Participants in Phase 1 and Phase 3 were asked to evaluate the visual quality of the videos while those in Phase 2 were simply instructed to look freely as if they were watching the videos on a TV or in a movie theater. All participants who took part in Phase 2 were seeing the videos for the first time to ensure that no memory effect influenced their viewing behavior. In all phases, the eye movements of the participants were recorded binocularly at 250 Hz with a video-based infrared eye tracker (SR-Research, EyeLink-II). The eye-tracking data were saved to disk for off-line analysis.

In total, 6 groups of participants were divided over the assigned tasks and collections of video stimuli as shown in Table 1. Groups 1, 3 and 5 performed exactly the same task as Groups 2, 4 and 6, respectively, albeit on a different set of stimuli, as will become clear in Section 5.2.2. The reason for splitting up videos over two groups of people was not only to keep the experimental sessions short, but also (mainly) to take care that all participants saw each video only once (i.e., either for scoring or free-looking, and with only one quality level). Groups 1 to 4 included 12 participants each, while Groups 5 and 6 included 9 participants each. Since no analysis on the eye-tracking data collected in Phase 3 (Groups 5 and 6) was planned, a lower number of participants was sufficient to quantify the perceived quality of the video stimuli. In total, 66 participants

took part in the experiment. They included master, graduate, and postgraduate students of the Electrical Engineering, Mathematics and Computer Science (EEMCS) faculty building at the Delft University of Technology (TU Delft).

The videos were displayed using a late 2008 MacBook connected to a 17-inch CRT monitor with a resolution of 1280x960 pixels. The experiment was controlled from a remote computer with its monitor positioned so that it would not interfere with the participant's task. In order to avoid outside elements interfering with the results, the experiment was carried out in a controlled environment inside the Delft Experience Lab located in the EEMCS faculty building at the TU Delft. Only the experimenter and the viewer were present while performing the experiment. The experimental setup is shown in Figure 5.1.

5.2.2 Creating the video stimuli

5.2.2.1 Uniformly Degraded Video Collections

A video database was created which consisted of 25 video segments with a duration of 20 seconds each. The goal of the study was to detect whether the quality of the ROI had an impact on the overall perceived quality; thus, we wanted to use stimuli that had clearly identifiable regions of interest, in order to make it easier to detect any differences that may result from changing the video quality in a spatially localized way. This peculiarity is not always retrievable in video quality databases (e.g., LIVE [24]); therefore, we opted for creating a dataset specific to our purpose.



Figure 5.2. Video segments used for the experiment were taken from action movies and were chosen to have highly dynamic sequences. Varying the bitrate used to compress the videos from higher bitrates (left 1237 bit/s) to lower ones (right 209 bit/s) gave a wide range of quality for the generated videos.

We assumed that highly dynamic scenes can seize the focus of the viewers under natural viewing conditions. Therefore, the video segments were extracted from action based movies (some sample frames are shown in Figure 5.2, the full dataset can be found at [25]). We opted for 25 different contents to make sure that, although related to videos with a clear region of interest, our results would be independent on specific content semantics. From each video, two distorted versions were produced using an H.264 video encoder (specifically, the x264 encoder provided by the ffmpegX software [26]). The resulting videos had a resolution of 1280x720 pixels, and a frame rate of 25 frames per second. From each video, two versions of it were created. Each was degraded to a different level of quality. We deemed it sufficient to analyze effects of only two quality levels in this experiment, as we were interested in proving that the quality of the ROI was more relevant to that of the background, and for this proof of concept, the comparison of two quality levels was enough. Furthermore, the coding parameters were constant over the whole length of one video, but were not uniform across the 25 different videos, to allow a variety of quality levels to be judged by the participants. Eventually, the database included 50 distorted videos spanning a wide range of quality. To give an idea of the breadth of this quality range, Figure 5.2 shows a sample frame on the left of one of the videos encoded with the highest bitrate (1237 bit/s), and another on the right which is among those encoded with the lowest bitrate (209 bit/s). Bitrate values are also displayed in Figure 5.3.

Two collections of stimuli were generated (Collections I and II, as indicated in Table 1). Each collection included only one of the two distorted versions generated for each of the 25 videos. The assignment of the distorted videos to the two collections was made randomly, so that the videos in each collection spanned roughly the same range of quality. Hence, the participants in Group 1 and Group 2 saw the same 25 original videos, but at a different quality level. Exactly the same holds for the participants in Group 3 and Group 4. All videos included in Collections I and II, along with their corresponding saliency maps, are available for further use and can be retrieved from the Delft IQLab repository [25,27].

5.2.2.2 Identifying the Region Of Interest

As mentioned earlier, the videos selected for this experiment were deliberately chosen to have a clearly identifiable ROI. It is expected that when observing the videos without a specific task (as in our experiment participants belonging to Groups 3 and 4 did), the viewer's attention is mainly drawn towards the ROI of each scene shown, as suggested by the research of Engelke et al. [22] and Wang et al. [23]. Smooth pursuit eye behavior exhibited when the viewer followed the movements of objects on the screen was registered by the eye tracker as fixations [18]. In order to perform a precise spatial analysis of attention deployment, fixation data were transformed into saliency information [28]. First, for each video frame we grouped in a single fixation map all the fixation locations of all participants. This resulted into as many fixation maps as frames for each video, giving a too fine granularity for the purpose of our analysis. Thus, videos were divided in coarser timeslots of 1 second each, and fixation maps were averaged over these timeslots, resulting in 20 fixation maps per video. These fixation maps were finally converted to saliency maps to better reflect the characteristics of human vision. In particular, a Gaussian patch with a width σ approximating the size of the fovea (about 2° visual angle) was applied to each fixation. A mean saliency map that takes into account all fixations of all subjects was calculated as follows:

$$S^{t,i}(k,l) = \sum_f^F \exp \left[\frac{(x_f^{t,i} - k)^2 + (y_f^{t,i} - l)^2}{\sigma^2} \right] \quad (1)$$

where $S^{t,i}(k, l)$ indicates the saliency map for video V_i , $i = 1, \dots, 50$, at timeslot $t \in [1, 20]$, given that V_i has a spatial resolution of $M \times N$ pixels (i.e. $k \in [1, M]$ and $l \in [1, N]$); $(x_f^{t,i}, y_f^{t,i})$ indicates the spatial coordinates of the f -th fixation ($f=1 \dots F$) performed on video V_i at timeslot t , where F is the total number of all fixations recorded for V_i from all subjects during timeslot t . The intensity of the resulting saliency map is linearly normalized to the range $[0, 1]$. Each of these maps specifies the saliency distribution over a specific timeslot of a specific video. In other words, it is a representation of the probability, pixel by pixel, that at a given timeslot, for a given video and task, the average participant fixates on a specific pixel. This process

was repeated for each 1 second timeslot for each video sequence. Hence, we obtained in total 20 (timeslots) x 25 (videos) x 2 (collections) x 2 (tasks) = 2000 saliency maps. Eventually, we decided to define the ROI as the region including all areas with a saliency value of 0.2 or more on the normalized range of [0, 1]. This threshold was chosen to ensure that the ROI occupied less than 10% of the entire video area (see Section 5.2.2.3)

5.2.2.3 Composite Video Creation

With the process described above we obtained for each timeslot of each video 2 ROIs from the free-looking data (i.e., from Groups 3 and 4). Each ROI was extracted from a video with a different quality level. To construct composite videos we first averaged these two ROIs by averaging the saliency maps and extracting a new ROI. Once the averaged ROI was identified, we created composite versions of the videos where the ROI area (which turned out to be 6 to 9% of the video size) and the background area (the remaining 91 to 94% of the video size) corresponded to a different version of the original video. As a result, within the same composite video, the encoding bitrate of the ROI was different from the bitrate used in the background area.

To better explain how we built our composite videos, we describe here an example. Let us consider the original video V_o , from which two distorted version V_{o1} and V_{o2} (included in Collection 1 and Collection 2 respectively) were created. Let us also assume that V_{o1} was compressed at a higher bitrate than V_{o2} . By analyzing the eye-tracking data with the procedure explained in Section 5.2.2.2, we were able to identify an averaged ROI for every timeslot of V_{o1} and V_{o2} , which we indicate with $ROI(V_o, t)$. From V_{o1} and V_{o2} we created two new composite videos, namely V_{c1} and V_{c2} , with the following characteristics:

$$\begin{aligned}
 V_{c1}(x, y, t) &= \begin{cases} V_{o1}(x, y, t) & \text{if } (x, y) \in ROI(V_o, t) \\ V_{o2}(x, y, t) & \text{otherwise} \end{cases} \\
 V_{c2}(x, y, t) &= \begin{cases} V_{o2}(x, y, t) & \text{if } (x, y) \in ROI(V_o, t) \\ V_{o1}(x, y, t) & \text{otherwise} \end{cases} \quad (2) \\
 t &= 1, \dots, 20
 \end{aligned}$$

with (x,y) representing a pixel in the video at timeslot t . Thus, this process resulted in two new composite versions; one version (V_{c1}) contained the ROI region of the video with the higher bitrate (V_{o1}) and the background region of the video with the lower bitrate (V_{o2}), while the second version (V_{c2}) contained the ROI region of V_{o2} and the background region of V_{o1} . As such, V_{c1} had a higher level of quality in the ROI than in the background, while V_{c2} had a lower level of quality in the ROI than in the background. A 3x3 pixel Gaussian smoothing filter was applied only to the pixel boundary between ROI and background in these composite videos to avoid a jarring transition in quality, which could distract the viewer or provoke annoyance. After applying this filter the presence of the boundary between both regions was hardly noticeable for most videos. Nonetheless, 5 of the resulting videos had to be discarded because the clash between the composite ROI and background regions was too intrusive on the video content. Therefore only 40 videos were included in the rest of the experiment and analysis. These 40 composite videos of the types (V_{c1}) and (V_{c2}) were distributed between the new Collections Comp I and Comp II, where each group contained one randomly selected version of the video (see Table 1).

5.2.3 Experimental protocol

Of the six groups of participants (again see Table 1), Groups 1, 2, 5, and 6 went through identical protocols, but watching a different collection of videos. Groups 3 and 4 followed a similar protocol except they were not asked to score videos in quality. In all cases, participants were given a printed description of the experiment and a list of the instructions they needed to follow. They were then seated so that their viewing distance measured 60 cm from the display screen. After being fitted with the head mounted eye tracker, the experimenter ran a 9-point calibration for the gaze location.

Groups 1 and 2 watched the first and second collection of videos respectively. They were requested to score each video, after it ended, in a single stimulus set-up with a continuous scale, ranging from 0 to 10, including the labels 'poor' at the lower end of the scale and 'excellent' at the higher end [29]. The choice of a continuous scale rather than a discrete, categorical one (often used in literature [29]), was dictated by the need of obtaining clear interval values for the quality scores, to be directly employed in the follow-up analysis. Categorical scales such as

the widely used 5-point Absolute Category Rating (ACR, [29]) are known for returning evaluations that cannot be immediately translated into interval values, due to the fact that the width of the categories may be variable and subject to the participants' interpretation (e.g. the 'poor' category may refer to a larger quality range than the 'fair' category) [30,31]. Continuous interval scales allow instead the user to express judgments that map linearly to the quality scale [30].

After the calibration of the eye-tracker, participants were shown 4 training videos (that were not part of Collections I and II), which were representative for the entire range of quality used in the video collections. The purpose of this separated training session was twofold: (1) it would help the participants in getting acquainted with the user interface of the experiment, and (2) gave them an idea of the range of quality used in the experiment, as such limiting contextual effects [32], as shown in previous work [33]. For every video to be evaluated, during both the training and actual experiment, the following steps were performed. Participants first

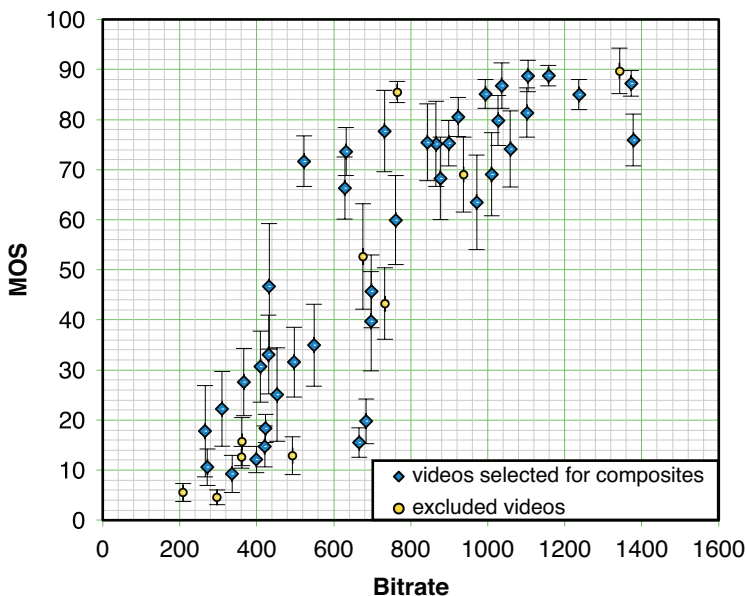


Figure 5.3. The relation between the H.264 bitrate and the MOS given by the viewers in Phase 1. Yellow, round markers identify the MOS scores for both versions of the five videos excluded from the video composite creation (see Section 5.2.2.3). The figure also visualizes the 95% confidence intervals for each MOS score using error bars.

saw a drift-correction screen, which helped the head mounted eye tracker to compensate for any shift in its position. They were then shown a 20-seconds video segment. This was followed by a scoring window presenting a slider to be moved on the continuous rating scale. Once a score was given, the process was repeated with the drift-correction screen followed by another video. After the first four videos, a dialog window was shown indicating that the training session was over and the process then continued with the videos from one of the collections. Participants from each group always saw the same 25 video segments from the assigned collection, but the order in which the videos were shown was randomized in order to avoid any bias (i.e., learning or fatigue effects) in the results.

Participants in Groups 3 and 4 were also shown Collections I and II, respectively. They followed the same protocol described above except that they were not given the scoring window after each video. Of course, the experiment instructions they had were also adjusted so that they were told to only watch the videos as if they were watching TV or a video downloaded from the Internet.

Finally, participants in Groups 5 and 6 also followed the exact same protocol explained above. The only difference was that participants were shown videos from Collections Comp I and Comp II instead. The training videos were also replaced by composited videos created with the same method as the composited videos in the collections. It is worthwhile mentioning that the eye movements of participants in Groups 5 and 6 were also tracked during their quality evaluation. Although these data were not directly analyzed in this study, we considered it necessary to collect them anyway, to keep experimental conditions as similar as possible throughout the three experimental phases.

5.3. Results

5.3.1 Scoring experiments

The video quality scores collected in Phases 1 and 3 of the experiment were processed to extract one Mean Opinion Score (MOS) for each video [29,34]. First the Z-score of each video was calculated using the mean

and standard deviation over all videos (so, both including Phases 1 and 3). Then the standard normal distribution of the resulting Z-scores was scaled to the range [0, 100]. For the collected MOS values in Phase 1, the relation between the level of H.264 bitrate that the videos had and the MOS is shown as an error bar plot in Figure 5.3.

The plot shows good distribution of the data points across the bitrate and the quality ranges. Yellow, round markers indicate the MOS scores obtained by both versions of the five videos excluded for the creation of composite videos (Section 5.2.2.3). It can be observed that those five videos span a quality range overlapping with those of the remaining videos included in the experiment (blue diamonds). Thus, we can assume their exclusion from the third phase of the experiment did not bring any context effect [32] possibly distorting the mean opinion scores.

Figure 5.4 shows an overview of the scores obtained for all four categories of videos: (1) V_{o1} encoded with high bitrate, (2) V_{o2} encoded with lower bitrate than V_{o1} , (3) V_{c1} composite with higher bitrate in the ROI than in the background, and (4) V_{c2} composite with higher bitrate in the background than in the ROI. Scores for V_{o1} and V_{o2} are derived from the first experimental phase (i.e., participants of Groups 1 and 2), while scores for V_{c1} and V_{c2} are derived from the third experimental phase (i.e., participants of Groups 5 and 6). It can be seen that, as expected, scores for videos in the V_{o1} category are significantly higher than scores of videos in the V_{o2} category. Also as expected, scores for V_{c1} and V_{c2} are somewhere in between, i.e., significantly higher than V_{o2} and significantly lower than V_{o1} (as confirmed by an ANOVA with video category as independent variable and quality score as a dependent one: $df = 3$, $F = 43,935$, $p < 0.001$). A post-hoc Tukey test confirmed the significance of the difference between all pairs of categories, besides between the mean scores for categories V_{c1} and V_{c2} , as also visible in the graph. This hints that videos with high quality in a small (about 9% of the whole area) but visually important region (V_{c1}) have just as high quality as videos whose largest part is of high quality, but are highly degraded in the ROI (V_{c2}). If we compare the corresponding used file sizes we find that V_{c1} videos are only 5% larger than their original, low quality counter-part V_{o2} . Conversely, videos in category V_{c2} (with higher bitrate in the background) required a file size of about 60% larger than V_{o2} . Still both V_{c1} and V_{c2} have an increase in overall quality of about 30% (blue bars in Figure

5.4(B)). The remainder of this section further investigates and quantifies the impact of the ROI quality to the overall quality evaluation in more detail.

5.3.2 Significance of the visual quality of the ROI for the overall video quality

If we assume that the quality in the ROI has the same impact on the total video quality as that of the background region, it is possible to calculate an Expected Score (ES) indicating the quality level of each composite video used in Phase 3. This is possible since we have the MOS scores for the two videos used to create each composite video (from Phase 1), as well as the percentage of the area that each of these videos occupied in the combined video. The ES of the video is then calculated as the area weighted average of the MOS scores of the two videos as follows:

$$ES = (MOS_{ROI} \cdot A_{ROI}) + (MOS_{background} \cdot A_{background}) \quad (3)$$

where MOS_{ROI} is the MOS for the video from Phase 1 used in the ROI region of the composite video, $MOS_{background}$ is the MOS for the video used in the background region, A_{ROI} and $A_{background}$ are size percentages of the ROI and background region to the total area of the video respectively. In other words, we assume that people evaluate the quality of all regions of the video and then average their scores without giving a higher value to one region over another.

By comparing the collected MOS values to the calculated Expected Scores, it is possible to see the effect the ROI has on the overall quality of the video. In Figure 5.5 we have split the data in two groups: one containing videos which have higher quality in the ROI than in the background (A) and one with videos which have lower quality in the ROI than in the background (B). From Figure 5.5(A) it is clear that the videos with a higher quality in the ROI have a tendency to get a higher MOS than what the ES suggests. The figure shows a fairly consistent positive trend throughout the quality range. In addition, the size of the deviation is in most cases larger than the confidence intervals around the MOS. Figure 5.5(B) shows that the effect is more prevalent for videos that have the higher quality in the background area of the video. The majority of the MOS lies far below the ES calculated values. This is also highlighted by

the strongly skewed 2nd order fitting line in the graph. We only have a tentative explanation to this observation. Both in Figure 5.5(A) and 5.5(B) we observe a ‘halo effect’ [35], where overall video quality is rewarded or punished depending on the quality of the ROI. If the discrepancy between MOS and ES was only due to this ‘halo effect’ we would expect a symmetrical behavior in both figures. Instead, videos with severe artifacts in the ROI (Figure 5.5(B)) show a saturation towards the low part of the evaluation scale. It seems therefore that the presence of such artifacts in the ROI enhances the halo effect, whereas the absence of artifacts in the ROI (Figure 5.5(A)) does not affect it. Thus we can hypothesize that, not only the quality of the ROI plays a dominant role in the overall quality perception, but also that the presence of strong artifacts in the ROI makes this role even more prominent.

It is also interesting to evaluate whether the size of the quality difference between the ROI and the background plays a role on the overall MOS of the combined videos. Figure 5.6 shows a scatter plot that attempts to illustrate this effect. In this figure, the horizontal axis represents the difference in quality between the ROI and the background of the video as determined from the difference in MOS found in Phase 1 of the experiment for the two videos used in the composite video. The videos

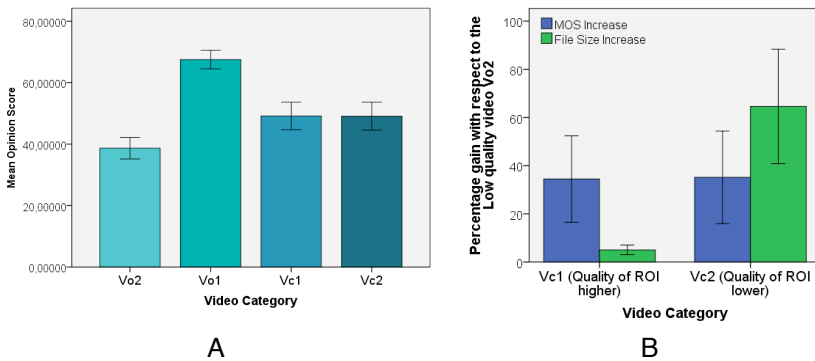


Figure 5.4. (A) Mean Quality scores for the videos in the four categories evaluated in Phase 1 and Phase 3 of the experiment: V_{o1} = encoded with high bitrate, V_{o2} = encoded with lower bitrate than V_{o1} , V_{c1} = composite with higher bitrate in the ROI than in the background, and V_{c2} = composite with higher bitrate in the background than in the ROI. (B) comparison of the difference in increase in video size (green columns) and corresponding increase in video quality (blue columns) for both categories of composite videos.

used in the experiment either fall in the negative half or the positive half of the graph depending on whether the ROI region or the background had a higher quality. The vertical axis represents the difference between the MOS collected in Phase 3 of the experiment (composite videos) and the ES. If there was no effect of the difference in quality between the ROI and background on the quality of the composite video, we would find a horizontal line in this graph. Moreover, if there was no difference between the effect the ROI and the background had on the overall quality, all data points would lay horizontally on the $Y=0$ axis, since the difference between the MOS and the ES would then be zero for all videos. It is clear that both are not the case. Instead, values tend to be negative when the ROI has a lower quality than the background and positive when the situation is reversed. Moreover, this effect appears to be stronger as the quality of the ROI (in relation to the quality of the background) becomes lower. This trend is weaker when the quality of the background is strongly compromised in comparison to the quality of the ROI.

5.3.3 Modeling the influence of ROI on overall VQ

Using the data collected from the experiment, it is possible to estimate how much more important the ROI is in determining the overall quality of a video. To do that, we again look at equation (3) used to calculate the ES. Since we now know that there is a difference in how much each region affects the overall perceived quality, we calculated a more

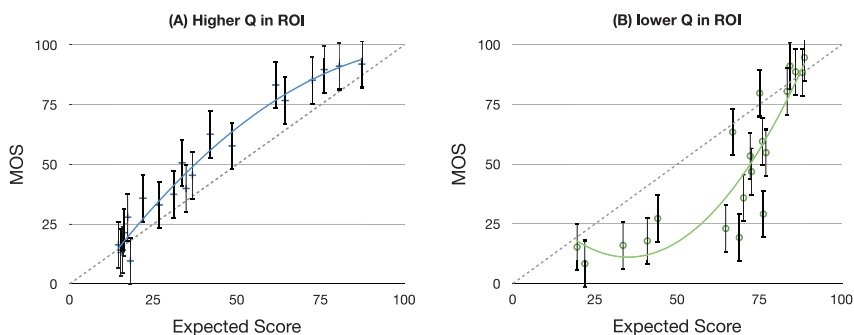


Figure 5.5. Comparing the calculated ES to the subjectively collected MOS. Figure 5.5(A) on the left represents videos with higher quality in the ROI than in the background, while videos in Figure 5.5(B) on the right have lower quality in the ROI.

accurate Weighted Expected Score (WES) by introducing two weighting parameters to the equation, resulting in:

$$WES = MOS_{ROI} \cdot A_{ROI} \cdot w_{ROI} + MOS_{background} \cdot A_{background} \cdot w_{background} \quad (4)$$

where w_{ROI} determines the weight of the ROI and $w_{background}$ the weight of the background region on the overall perceived quality. To calculate the values of these weights, a linear regression analysis was performed. The analysis used the MOS of the composite videos as the dependent variable and the averaged quality of each region multiplied by its corresponding area as the independent variables. The analysis returned the values $w_{ROI} = 6.17$ ($p < 0.001$, 95% confidence interval 4.65 to 7.69), and $w_{background} = 0.57$ ($p < 0.001$, 95% confidence interval 0.46 to 0.69). The overall model fit had a R^2 of 0.96. The resulting relation is

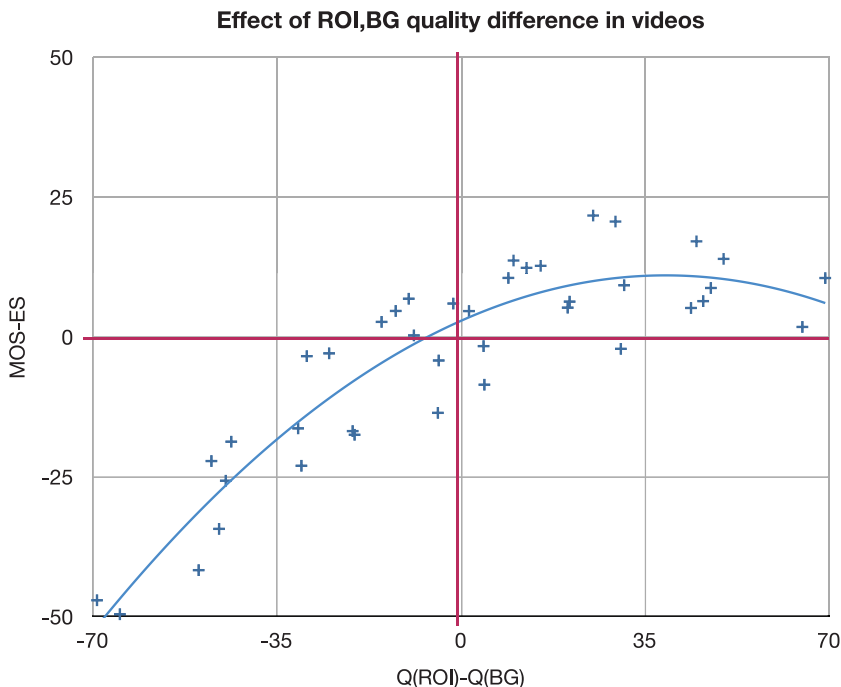


Figure 5.6. The horizontal axis represents the difference in MOS between the ROI and the background region, where the left side contains videos with lower quality in the ROI than in the background and vice versa. On the Y axis, the difference between the MOS and the calculated ES is represented. On the lower half, viewers scored the videos lower than what the ES expected and vice versa.

depicted in Figure 5.7, again for the two groups of data separately; i.e., in Figure 5.7(A) for the images with higher quality in the ROI than in the background, and in Figure 5.7(B) for the images with lower quality in the ROI than in the background.

To test the stability of this fit, the 40 videos of Phase 3 of the experiment were split into two subgroups of 20 videos each. Both subgroups spanned the entire range of the quality scale. The two counterparts of each composite video (i.e., one with higher quality in the ROI and the other with a higher quality in the background) were joined in the same subgroup in order to avoid having the same video content repeated in both subgroups and thereby influencing the analysis. We then conducted a linear regression analysis in the same manner as described above on one of the two subgroups to calculate a subgroup weight for the ROI (w_{sg-ROI}) and the background ($w_{sg-background}$) regions. This analysis yielded the values $w_{sg-ROI} = 5.96$ ($p < 0.001$, 95% confidence interval 3.64 to 8.27) and $w_{sg-background} = 0.60$ ($p < 0.001$, 95% confidence interval 0.42 to 0.79). The overall model fit had a R^2 of 0.96. Both weighting factors were close to the ones found for the whole ensemble of videos, indicating that the result of the fit was not very sensitive to the particular selection of videos used. Subsequently, the new values of w_{sg-ROI} and $w_{sg-background}$ were used in equation (4) to calculate the WES of the second subgroup of videos, the results of which are shown in Figure 5.8.

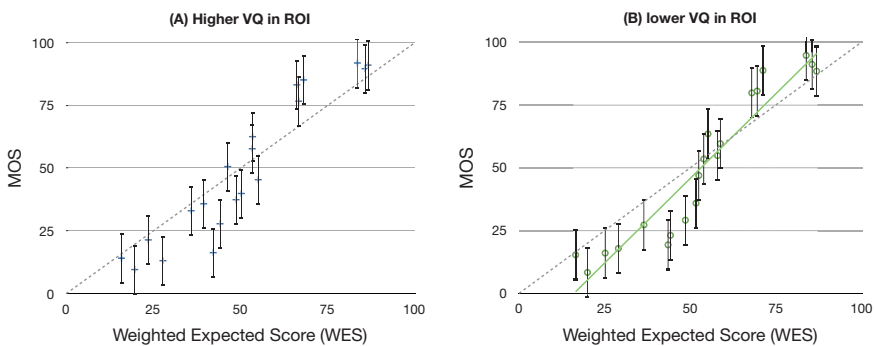


Figure 5.7 Comparing the calculated WES to the subjectively collected MOS. Figure 5.7(A) on the left represents images which had higher quality in the ROI than in the background, while Figure 5.7(B) on the right represents images which had lower quality in the ROI than in the background.

The latter figure illustrates that for most videos the calculated WES is a good prediction of the MOS (R^2 of the fit was 0.852 for Figure 5.8(A) and 0.859 for Figure 5.8(B)), showing that the quality prediction model can generalize well on unseen video content. There are some exceptions though; for some videos the predicted WES falls outside the confidence interval of the perceived MOS. To check whether these deviations were systematically related to the quality difference between ROI and background in the composite video, we constructed Figure 5.9.

Figure 5.9 is a similar plot as the one shown in Figure 5.6, but now generated using the WES scores from the second subgroup of videos. The data points are now scattered around the $Y=0$ axis, indicating that with the proper weighting factors for the quality of the ROI and the background, the overall quality of a video, locally varying in quality, can be reasonably well predicted. There are still videos for which the deviation of the predicted score from the perceived MOS is about 1/5 to 1/4 of the total scoring scale. For these videos a linear weighting of the perceived quality of the ROI with the perceived quality of the background may be too simple.

5.4. Discussion

Our results clearly show the importance of the contribution of the quality of the ROI to the overall quality judgment. This finding may have two

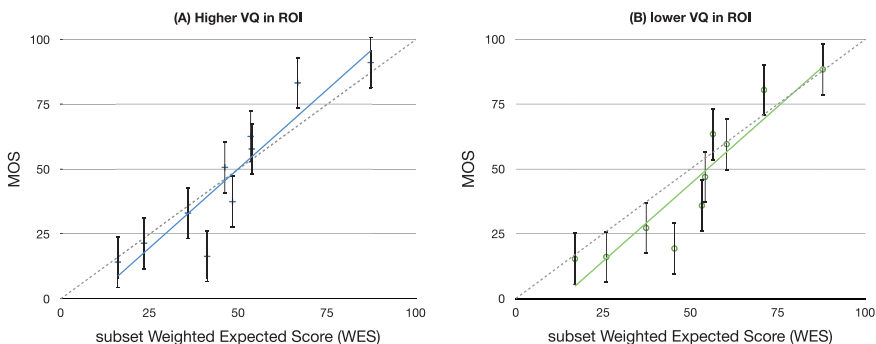


Figure 5.8. Comparing the calculated WES to the subjectively collected MOS for the stimuli belonging to the second subgroup. Figure 5.8(A) on the left represents images which had higher quality in the ROI than in the background (R^2 of the fit was 0.852), while Figure 5.8(B) on the right represents images which had lower quality in the ROI than in the background (R^2 of the fit was 0.859).

important spin-offs: (1) it may be used to further improve coding schemes, and (2) it may increase the accuracy of visual-importance-based objective quality measures. The latter application is clearly more straightforward than the former. Because most coding schemes are block-based, a pixel-wise weighting of saliency in the compression scheme may become impractical, and so, a different approach should be designed. In addition, videos with a mostly flat background (e.g., landscapes) already require less bits for compressing the background, and as such compromising the quality of the background at the expense of the quality of the ROI is not very practical. The latter though becomes relevant for videos with a highly textured background (e.g., woods, sports, architectural environments). More straightforward is the application of our findings in the design of objective quality metrics, and so, this application gets further attention in the rest of the discussion.

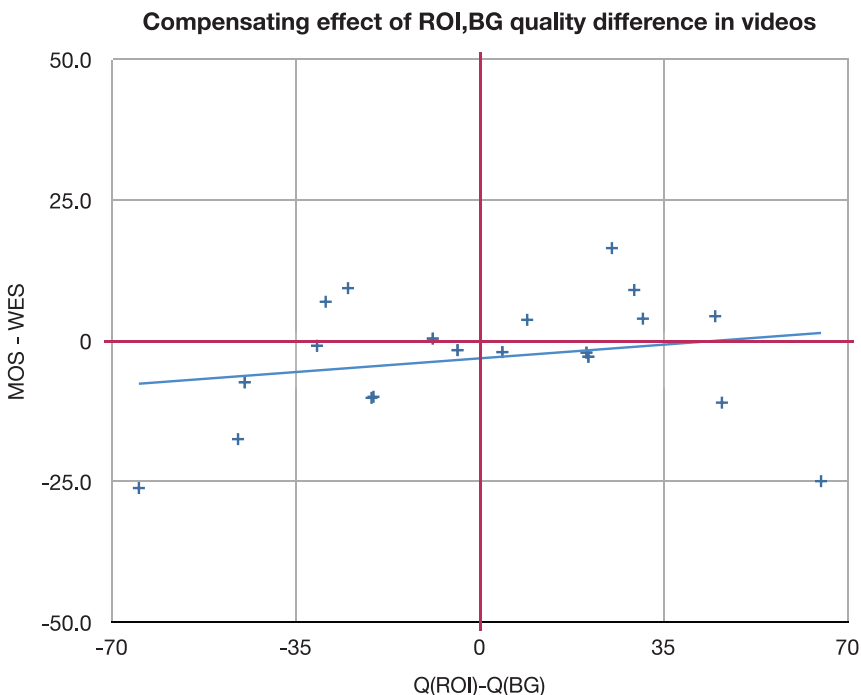


Figure 5.9. Illustration of how well the WES estimated from equation (4) predicts the subjectively obtained MOS for the second subgroup of composite videos used in Phase 3 of the experiment. The horizontal axis represents the difference in quality between the ROI and the background, while the vertical axis gives the difference between the MOS and the calculated WES.

5.4.1. Effect of ROI on perceived quality

From the results presented in Section 5.3, it seems possible to conclude that, when people assess video quality, they give greater significance to some regions of the video over others. Thus, it would be unwise to assess the overall video quality by simply averaging the quality of the different regions of that video. This is shown in Figure 5.6 where participants gave the video a score (MOS) different from the one we calculated (ES) by simply averaging the quality of all video regions. The latter observation explains why objective quality assessment algorithms can benefit from utilizing saliency information [36, 37]. Moreover, from our results, we would expect that the latter is even the case for quality prediction of videos degraded with diffused coding artifacts, contrary to what has been suggested in [9, 21, 38]. However, deciding on how exactly to use this saliency data has largely remained an arbitrary process. In that respect, this paper provides insight in how humans perceive visual quality in relation to the saliency of the content. Additionally, it makes a first attempt in creating an empirical based model for using saliency information in calculating overall perceived visual quality.

Indeed, the results of Phase 3 of the experiment have clarified the relation between the ROI quality and the overall perceived quality of a composite video as expressed with a MOS. Videos that have a higher quality in the ROI tend to be scored higher than expected. Reversely, videos that have lower quality in the ROI than in the background area are scored lower than expected. Figure 5.6 illustrates how the difference between the subjectively collected MOS and the calculated ES is affected by the amount of quality difference between the two video regions. Looking at the lower left corner of the figure, one can see that as the quality of the ROI gets more degraded, the MOS shifts further away below the expected score. In the center of the figure, where the quality level in both video regions is very close, the MOS and ES are very close as well. As the quality of the ROI continues to increase (i.e., towards the right side of the figure), the difference between the MOS and the ES stops growing and even tends to diminish again. The latter observation seems to suggest that even if the degradation is only present at the

background region, at a certain point the degradation becomes so bad that it plays an equally important role in determining the MOS for the entire video. Literature has indeed shown that excessive background distortions can be strong attractors of attention and are perceived as highly annoying [26]. In other words, the distortion artifacts become ROI themselves and thereby decrease the importance of the natural ROI of the videos.

5.4.2. An ROI sensitive video quality assessment model

In an attempt to quantify the impact of the quality of the ROI to the overall video quality we performed a linear regression analysis. The resulting values, i.e., $w_{ROI} = 6.17$ and $w_{background} = 0.57$, suggest that the ROI region is about 10 times more significant in determining the overall quality of a video than the background region. This is even more impressive when one takes into account that the size of the ROI in the used videos occupied only 6% to 9% of the entire video area. Subsequent analysis also proved that this simple linear regression model already results in a considerable improvement in predicting the perceived overall quality of a video with different quality levels in different areas. Still the model is not perfect; for some videos the predicted quality deviates 1/5 to 1/4 of the total scoring scale from the perceived quality. The latter is not so surprising if we realize that we simply linearly weighted the contributions of ROI and background to the overall predicted quality, whereas multiple figures in this paper show strong non-linear behavior. Figure 5.5, for example, shows that the difference between the perceived and expected quality scores depends in a non-linear way on the quality difference between the ROI and background region of the video. This observation is not taken into account in the linear weighting model. Despite this simplification in the model, we still believe that the current model may be applied in the spatial pooling step of objective quality metrics based on estimating local distortion visibility (e.g. [39]), to further improve their accuracy. In most metrics, spatial pooling is done by either averaging across the whole image or by weighting distortion visibility proportional to saliency. With our model, it would be possible to simply average distortion visibility values within and outside the ROI, and then take a weighted combination (based on w_{ROI} and $w_{background}$) as an estimate of the MOS.

5.4.3. Significance of ROI to overall quality for images and videos

Comparing our current findings to similar work previously reported on still images [17] reveals great similarities. The latter is particularly impressive if one keeps in mind that there were some differences in the experimental setup between both studies. One of these differences was the eye-tracker hardware used in both experiments. Additionally, the resolution of the content used changed from 600x600 pixels for the images to 1280x720 pixels for the videos. Finally, only 40 videos were used in the current study instead of the 80 images used in the previous study. Nonetheless, the methodology used to process the data and the statistical analysis was identical between both studies.

Figure 5.10(A) mirrors the analysis shown in Figure 5.6 but then on still images. The similarities between the two figures are undeniable. Both show that when the quality of the ROI is lower than that of the background, the subjectively measured MOS is lower than the Expected Score, calculated assuming an equal importance of the quality of the ROI and background. This trend is more pronounced in videos than in images as the curve in Figure 5.6 is steeper than the curve in Figure 5.10(A). Both figures approach zero when the quality of the ROI and the background get close. When the ROI quality is higher than the background quality, both figures show a rise in the deviation between the perceived quality and predicted quality that rapidly levels off and then tends to decrease as the quality of the background region becomes very low. This level of similarity in the results is quite surprising. Given the dynamic nature of video content and how that may affect the significance of the ROI and the annoyance level of compression artifacts, one would have expected more pronounced differences in the results.

It is also interesting to compare the ROI of the still images to the ROI of the videos. For the images, the ROI covered 10 - 16% of the entire image area. However, using the same methodology, the ROI for the videos occupied only 6 - 9% of the video area. This difference may probably be attributed to the timeslots over which the ROI was calculated. For the videos, frame related saliency maps were combined into saliency maps over 1 second of video (as explained in Section 5.2.2.2). For the images,

on the other hand, viewers saw each image for 8 seconds, and so, the saliency maps cover visual attention over a longer timeslot. With that in mind, it becomes even more impressive that Figure 5.6 shows the same trends as Figure 5.10(A). Moreover, given that the size of the identified ROI in the videos is smaller than the size of the ROI in images, one would expect that the relative contribution of the quality of the ROI to the overall quality in videos is also smaller than in images. However, the latter is not the case; the ROI in images is only 5 times more important than the background region, while for videos the ROI is 10 times more important than the background. Apparently, in videos the viewing behavior of people is much more concentrated in a smaller area of the video, which then also gets a higher impact on the overall quality judgment.

Finally, Figure 5.10(B) shows a graph equivalent to Figure 5.8, but again now for images instead of videos. So, half of the data from Figure 5.10(A) were used in a linear regression analysis to establish a Weighted Expected Score model, and this model was then applied to the other half of the data. Using Weighted Expected Scores, taking into account the relative importance to VQ of the ROI and the background, the curved trend shown in Figure 5.10(A) disappears. When comparing Figures 5.10(B) and 5.8, we can see that the use of Weighted Expected Scores delivers similar performance for both images and videos. The data points in both graphs are mostly contained within the -25 to 25 range of the y-axis.

The similarity between images and videos is also impressive in view of the fact that different content was used: the still images evaluated in [17] were not just stills from the videos used in the current study. What all content shared though was an obvious ROI. For content with a less pronounced ROI, the weighting coefficients will change, with the weighting coefficient of the background quality ultimately becoming zero when the ROI covers the whole image, but this obviously is not a relevant condition for the current research. The evaluation of the model in terms of generalization towards different content (as presented in Section 5.3.3) showed that the weighting coefficients are rather stable over variable size of the ROI within reasonable limits (i.e., for images up to 16% of the size).

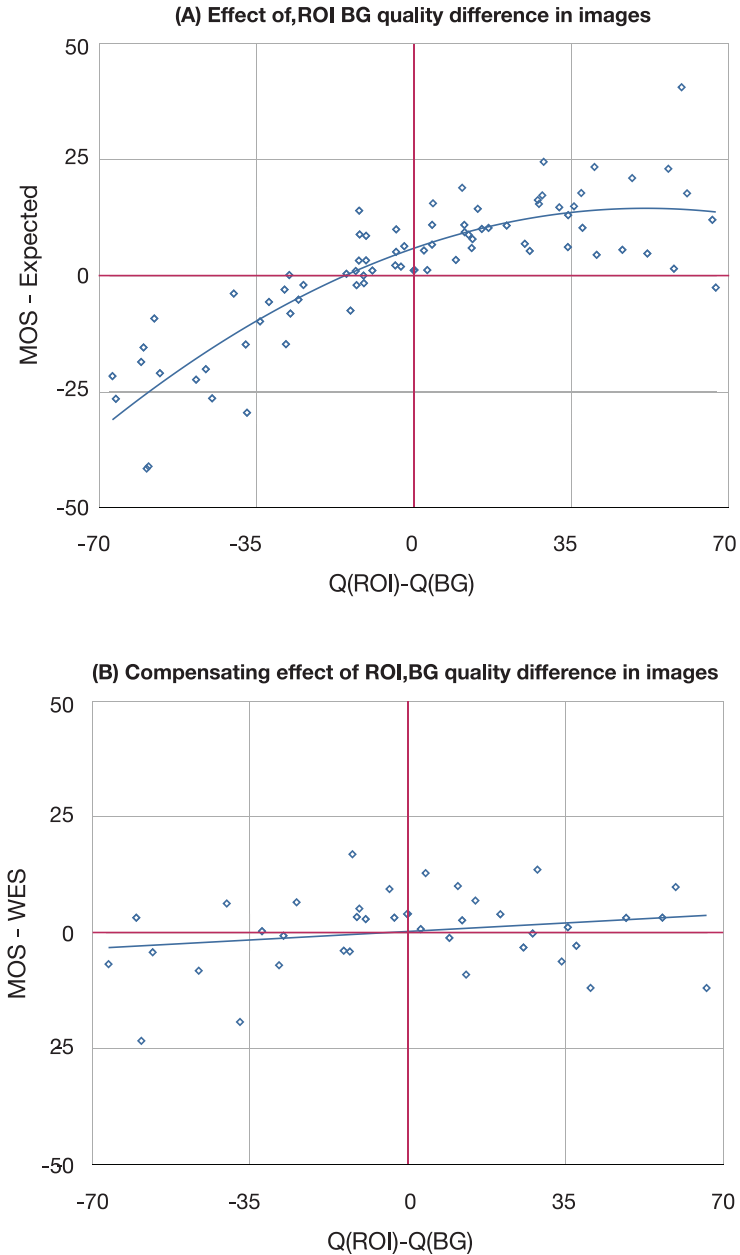


Figure 5.10. The results of similar work performed on still images [17]. Figure 5.10(A) (on the left) shows the relation between the expected scores and the subjective MOS in a similar fashion to Figure 5.5. While Figure 5.10(B) (on the right) represents how the calculated weighted expected score performs after performing a similar regression analysis as illustrated here in Figure 5.8.

5.5. Conclusions

In this article we examined whether the visual quality in the Region of Interest of a video has more impact on the overall video quality than that of the background region. We found that when the quality of the ROI (kept to an extension corresponding to 6-9% of the video area) is higher than that of the background, the viewers tend to give the video a higher quality score than its average quality level. The collected data was used to create a simple model to estimate the overall perceived video quality from the different quality levels of ROI and background region. This model illustrates that the quality of the ROI is about 10 times more important for the overall quality judgment than the quality of the background. This answers our first research question showing that artifacts located in the ROI of the video are indeed significantly more annoying than those in the background region.

The extent to which the annoyance of the artifacts in the ROI contributes to the final quality evaluation is dependent on a number of variables. One of those is the size of the ROI which in turn is dependent on the nature of the video content. For the highly dynamic videos used in our study, it was shown that the quality judgment was marginally affected by the background region occupying 91% to 94% of the scene. Viewers based their quality judgment mainly on the visual quality of the ROI, 10 times more so than on the quality of the background. Though these values will defer with different types of video content, it is encouraging evidence that objective video quality assessment metrics should incorporate information on visual importance of video regions. The results also show that it is risky to apply naive video enhancement algorithms which do not take the location of the ROI into consideration.

When comparing the significance of the ROI in videos to that in images, our data show remarkable similarities in the findings. Still when identifying the region of interest in videos, we found that it was much more focused than in images. The size of the calculated ROI in images ranged between 10 - 16% of the entire image area, while it only occupied 6 - 9% of the video area. Nevertheless, when modeling the significance of the ROI on overall quality, it came out that in images it is only 5 times more significant than the background region, while in videos it is 10 times more so.

Further work can still look at the effect of different types of artifacts (such as blur or desaturation), which are more difficult to perceive with human peripheral vision. This makes these artifacts less distracting while at the same time offering reduction in used bandwidth. This may make the trends observed in this article even stronger, especially in videos. It may also be interesting to see whether the same results can be replicated in video content with a less prominent ROI.

References

1. Cisco, I. (2012). Cisco visual networking index: Forecast and methodology, 2011--2016. CISCO White paper, 2011-2016.
2. Haskell, B. G. (1997). Digital Video: An Introduction to MPEG-2: An Introduction to MPEG-2. Springer.
3. Schwarz, H., Marpe, D., & Wiegand, T. (2007). Overview of the scalable video coding extension of the H. 264/AVC standard. *Circuits and Systems for Video Technology, IEEE Transactions on*, 17(9), 1103-1120.
4. Sullivan, G. J., Ohm, J., Han, W. J., & Wiegand, T. (2012). Overview of the high efficiency video coding (HEVC) standard. *Circuits and Systems for Video Technology, IEEE Transactions on*, 22(12), 1649-1668.
5. Shao, L. (2008). Enhancement of compressed video signals using a local blockiness metric. In *2008 IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 1397-1400).
6. Shao, L., Wang, J., Kirenko, I., & De Haan, G. (2011). Quality adaptive least squares trained filters for video compression artifacts removal using a no-reference block visibility metric. *Journal of Visual Communication and Image Representation*, 22(1), 23-32.
7. Lin, W., & Jay Kuo, C. C. (2011). Perceptual visual quality metrics: A survey. *Journal of Visual Communication and Image Representation*, 22(4), 297-312.
8. Hemami, S. S., & Reibman, A. R. (2010). No-reference image and video quality estimation: Applications and human-motivated design. *Signal processing: Image communication*, 25(7), 469-481.
9. Engelke, U., Kaprykowsky, H., Zepernick, H., & Ndjiki-Nya, P. (2011). Visual attention in quality assessment. *Signal Processing Magazine, IEEE*, 28(6), 50-59.
10. Simons, D. J., & Chabris, C. F. (1999). Gorillas in our midst: Sustained inattentive blindness for dynamic events. *Perception-London*, 28(9), 1059-1074.
11. Carrasco, M. (2011). Visual attention: The past 25 years. *Vision research*, 51(13), 1484-1525.
12. Koch, C., & Ullman, S. (1987). Shifts in selective visual attention:

- towards the underlying neural circuitry. In *Matters of Intelligence* (pp. 115-141). Springer Netherlands.
13. Smith, A. T., Singh, K. D., & Greenlee, M. W. (2000). Attentional suppression of activity in the human visual cortex. *Neuroreport*, 11(2), 271-278.
 14. Redi, J., Liu, H., Zunino, R., & Heynderickx, I. (2011). Interactions of visual attention and quality perception. In *IS&T/SPIE Electronic Imaging* (pp. 78650S-78650S). International Society for Optics and Photonics.
 15. Ninassi, A., Le Meur, O., Le Callet, P., Barba, D., & Tirel, A. (2006, September). Task impact on the visual attention in subjective image quality assessment. In *Proceedings of European Signal Processing Conference*.
 16. Vu, E. C. L., & Chandler, D. M. (2008). Visual fixation patterns when judging image quality: Effects of distortion type, amount, and subject experience. In *Image Analysis and Interpretation, 2008. SSIAI 2008. IEEE Southwest Symposium on* (pp. 73-76). IEEE.
 17. Alers, H., Liu, H., Redi, J., & Heynderickx, I. (2010). Studying the risks of optimizing the image quality in saliency regions at the expense of background content, *IS&T/SPIE Electronic Imaging, Image Quality and System Performance VII*.
 18. Kowler, E. (2011). Eye movements: The past 25years. *Vision research*, 51(13), 1457-1483.
 19. Engelke, U., Pepion, R., Le Callet, P., & Zepernick, H. J. (2010). Linking distortion perception and visual saliency in H. 264/AVC coded video containing packet loss. In *Visual Communications and Image Processing 2010* (pp. 774406-774406). International Society for Optics and Photonics.
 20. Le Meur, O., & Le Callet, P. (2009). What we see is most likely to be what matters: Visual attention and applications. In *Image Processing (ICIP), 2009 16th IEEE International Conference on* (pp. 3085-3088). IEEE.
 21. Le Meur, O., Ninassi, A., Le Callet, P., & Barba, D. (2010). Overt visual attention for free-viewing and quality assessment tasks: Impact of the regions of interest on a video quality metric. *Signal Processing: Image Communication*, 25(7), 547-558.
 22. Engelke, U., Zepernick, H., & Maeder, A. (2009). Visual attention modeling: region-of-interest versus fixation patterns. In *Picture Coding Symposium, 2009. PCS 2009* (pp. 1-4). IEEE.
 23. Wang, J., Chandler, D. M., & Le Callet, P. (2010). Quantifying the relationship between visual salience and visual importance. In *IS&T/SPIE Electronic Imaging* (pp. 75270K-75270K). International Society for Optics and Photonics.
 24. Seshadrinathan, K., Soundararajan, R., Bovik, A. C., & Cormack, L. K. (2010). A subjective study to evaluate video quality assessment

- algorithms. In IS&T/SPIE Electronic Imaging (pp. 75270H-75270H). International Society for Optics and Photonics.
25. Alers, H., Liu, H., Redi, J., & Heynderickx, I., TUD Video Quality Database: Eye-Tracking Release 2, http://mmi.tudelft.nl/iqlab/video_task_eye_tracking_1
 26. Redi, J., Heynderickx, I., Macchiavello, B., & Farias, M. (2013). On the impact of packet-loss impairments on visual attention mechanisms. In Circuits and Systems (ISCAS), 2013 IEEE International Symposium on (pp. 1107-1110). IEEE.
 27. Alers, H., Redi, J. A., & Heynderickx, I. (2012). Examining the effect of task on viewing behavior in videos using saliency maps. In IS&T/SPIE Electronic Imaging (pp. 82910X-82910X). International Society for Optics and Photonics.
 28. Redi, J. A., & Heynderickx, I. (2011). Image quality and visual attention interactions: towards a more reliable analysis in the saliency space. In Quality of Multimedia Experience (QoMEX), 2011 Third International Workshop on (pp. 201-206). IEEE.
 29. Recommendation, I. T. U. R. B. T. (2002). 500-11, Methodology for the subjective assessment of the quality of television pictures. International Telecommunication Union, Geneva, Switzerland, 4, 2.
 30. Keelan, B. (2002). Handbook of image quality: characterization and prediction. CRC Press.
 31. Engeldrum, P. G. (2000). Psychometric scaling: a toolkit for imaging systems development. Imcotek Press
 32. de Ridder, H. (2001). Cognitive issues in image quality measurement. *Journal of Electronic Imaging*, 10(1), 47-55.
 33. Redi, J., Liu, H., Alers, H., Zunino, R., & Heynderickx, I. (2010, January). Comparing subjective image quality measurement methods for the creation of public databases. In IS&T/SPIE Electronic Imaging (pp. 752903-752903). International Society for Optics and Photonics.
 34. Sheikh, H. R., Sabir, M. F., & Bovik, A. C. (2006). A statistical evaluation of recent full reference image quality assessment algorithms. *Image Processing, IEEE Transactions on*, 15(11), 3440-3451.
 35. Nisbett, R. E., & Wilson, T. D. (1977). The halo effect: Evidence for unconscious alteration of judgments. *Journal of personality and social psychology*, 35(4), 250.
 36. Lu, Z., Lin, W., Yang, X., Ong, E., & Yao, S. (2005). Modeling visual attention's modulatory aftereffects on visual sensitivity and quality evaluation. *Image Processing, IEEE Transactions on*, 14(11), 1928-1942.
 37. You, J., Korhonen, J., & Perkis, A. (2010). Attention modeling for video quality assessment: Balancing global quality and local quality. In Multimedia and Expo (ICME), 2010 IEEE International Conference on (pp. 914-919). IEEE.

38. Ninassi, A., Le Meur, O., Le Callet, P., & Barbba, D. (2007). Does where you gaze on an image affect your perception of quality? Applying visual attention to image quality metric. In *Image Processing, 2007. ICIP 2007. IEEE International Conference on* (Vol. 2, pp. II-169). IEEE.
39. Wang, Z., Bovik, A. C., Sheikh, H. R., & Simoncelli, E. P. (2004). Image quality assessment: from error visibility to structural similarity. *Image Processing, IEEE Transactions on*, 13(4), 600-612.

6

Effect of Image Quality on Disaster Response Applications

Abstract

There has been a significant amount of research investigating how image quality is evaluated in a home setting for entertainment purposes. It is, however, still unclear how different tasks can impact the perception of image quality. In turn it is interesting to understand whether image quality can affect the performance for such tasks. In this paper we use the setting of a disaster situation to study the relation between image quality and performance. An experiment was conducted where participants viewed slideshows of disaster situations using different levels of image quality. By measuring how well they were able to reconstruct the events they saw, we show that the reduced image quality did not have an effect on their performance.

6.1. Introduction

When using lossy compression algorithms to reduce the size of image files, a part of the original image information is permanently lost. As a result, only a distorted version of the original image can be reconstructed for the viewer. People perceive such distorted images as being of lower image quality (IQ) than the originals. This effect has been studied extensively to try and understand how much such distortions affect the perceived IQ [1-3]. However, this work mainly focuses on a viewing task for the purposes of entertainment in a home setting.

When taking other tasks into consideration, research has shown that, images which have been greatly distorted (and therefore have a significantly lower IQ score) can still be considered of high quality with respect to the desired application [4]. Images used as means to transfer information, for example, can withstand a considerable level of degradation with its content still being recognizable [5]. For some applications where data transmission bandwidth is limited and the main focus is the exchange of information through images, highly compressed images are desired for their small files sizes. It is however still unclear whether the degradation of the IQ has a hindering effect on the exchange of information. We are interested in studying whether a significant loss in image quality can cause a drop in performance even if the required task is not directly impacted by the image distortions.



Figure. 6.1. One of the water disaster scene images showing a freight boat colliding with a bridge causing it to collapse and sending a car into the water channel.

This paper focuses on using images in the emergency response application domain, where resources are extremely limited but ,at the same time, having a clear and undisturbed transfer of information is of vital importance [6]. We constructed an experiment where a set of images was used to identify the events that took place in a disaster situation. By compromising the quality of the displayed images and measuring the resulting effect on the performance in the required task we saw whether the reduction in IQ had an adverse effect on performance. The paper starts in Section 6.2 by explaining how the disaster scenes were simulated in order to create the image slideshows. Section 6.3 goes through the steps of the experiment showing the task the test participants needed to perform using these slideshows. The paper then shows how the data was analyzed in Section 6.4. The generated results are discussed in Section 6.5 and the main conclusions of the paper are listed in Section 6.6.

6.2. Methodology

6.2.1. The stimuli

The participants were meant to play the role of witnesses of a disaster simulation. They would see the events of this disaster through a series of static images. To create these images, a miniaturized disaster situation model was created using Playmobil toy sets. Two models were built that depicted disaster scenes involving fire and water accidents. Photos of

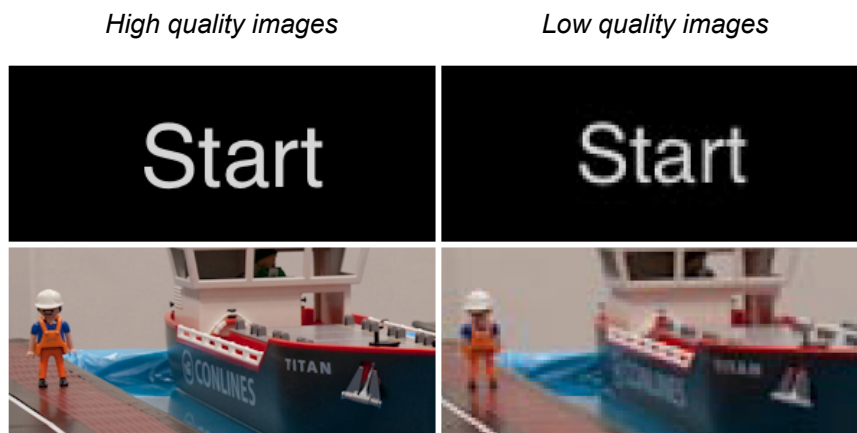


Figure 6.2. Two examples showing the difference between the low and high quality images shown in the slideshows.

Table 1. The number of images each slideshow included for each view point of the created scenes.

Scene	View point	Images
Fire	A	15
	B	9
Water	A	11
	B	11

these worlds were taken using 2 Canon cameras (Models 30D, and 40D), with the first using a 50mm fixed focal length lens and the other using a 17-85mm zoom lens. Wide angle photos were taken to help the viewers see their surroundings and orient their location, while long focal lengths were used for photos zooming in on specific locations to simulate the viewer focusing on specific events. All images were taken from two specific locations representing the viewpoints of two different observers looking at the disaster from different angles. The disaster models were adjusted while shooting the pictures to represent the development of the accident situation. Some of the images were later manipulated using Adobe Photoshop to add fire effects. Figure 6.1 shows one of the wide angle images from the water disaster scene.

For each of the two scenes, two slideshows (from different view points) were created from the photo shoots, giving us 4 different slideshows. As shown in Table 1, the number of images that were needed to show the events of the disasters scene sometimes differed between the two view points.

The slideshow that showed the events of the disaster unfolding. Each image was displayed on the screen for a duration of 5 seconds, and each slideshow was shown only once.

The slideshows were created using the Adobe Photoshop Lightroom (version 2.5) application. Two versions of the slideshows were created that contained the same images but under different quality levels. The high quality (HQ) version contained images in their original quality with a resolution of 1200x800 pixels. The low quality (LQ) Slideshow contained

images which were scaled down to a resolution of 480x320 pixels. This resolution was chosen since it is the native resolution of mobile device currently being evaluated for use in the field of disaster response [7]. The compression quality of the generated images was also compromised to reflect low bandwidth data transfer in real-life networks for mobile devices. The low resolution images were compressed with the JPEG codec at a quality level of 50 (on a scale of 0 to 100, implemented via the Lightroom software). This level of quality produced clearly visible compression artifacts in the images (see Figure 6.2), but nonetheless ensured that all the information which the participants needed to remember during the experiment was still clearly visible. Since originally 4 slideshows were constructed for the two scenarios (fire and water) from two view points, having two levels of quality from each slideshow meant that the experiment involved 8 slideshows in total.

6.2.2. The experimental setup

All experimental sessions were held in the Pi-lab located in the Electrical Engineering, Mathematics and Computer Science (EEMCS) faculty



Figure. 6.3. A test participant indicating the events of the disaster scenario using the magnetic board.

Table 2. The complete scheme of the experiment showing which viewpoint (V), session, and image quality (IQ) was shown to each participant

Participant #	V	Scenario 1		Scenario 2	
		IQ	Session	IQ	Session
1, 3, 17, 20	A	L	Fire	H	Water
5, 7, 22, 24	A	L	Water	H	Fire
9, 11, 26, 28	A	H	Fire	L	Water
13, 15, 30, 32	A	H	Water	L	Fire
2, 4, 18, 19	B	L	Fire	H	Water
6, 8, 21, 23	B	L	Water	H	Fire
10, 12, 25, 27	B	H	Fire	L	Water
14, 16, 29, 31	B	H	Water	L	Fire

building at the Delft University of Technology. The lighting conditions were controlled and kept at a lighting setting of typical office conditions. The images were displayed on the screen of an Apple MacBook computer with a 13.3" widescreen and a native resolution of 1280x800 pixels. The viewing distance was kept to 60 cm with the screen directly facing the viewer.

The experiment had a total of 32 participants. They were collected from the faculty of Computer Science at the Delft University of Technology, and were either students or staff members. 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.

6.3. The experimental protocol

The experiment started by giving the participant written instructions explaining the steps of the experiment. The participant was then shown one of the slideshows showing one of the accident situations (fire or water accident scenario). The order of which scenario was shown first was alternated to avoid any systematic effect on the results. Table 2

shows the complete plan of the experiment sessions.

After viewing the slideshow, the test participants were presented with a magnetic board containing a bird's-eye-view map of the disaster area (see Figure 6.3). They were also provided with magnetic icons representing objects and characters from the disaster. The task given to the participants was to construct a situation map reflecting the events that took place in the slideshow. The participants were given an unlimited amount of time to adjust the created map until they were convinced that they could not make any further improvements. A photo of the created map was then taken for later evaluation as shown in Figure 6.4.

Consequently, the second slideshow was displayed to the test subject showing images of the second disaster situation. As shown in Table 2, this slideshow was shown in a different level of quality than the first one. In other words, if the first slideshow was shown in the HQ version then the second one was shown in the LQ version, and vice-versa.

Finally, the experiment was concluded by asking the participant to fill in a Likert-scale questionnaire. Since the participants were not informed beforehand that there was a difference in image quality between the slideshows, we asked them whether they noticed any difference in the

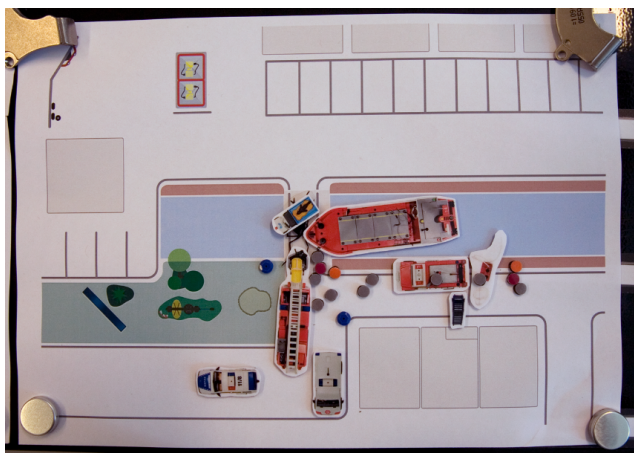


Figure. 6.4. An example of a situation map created by a test participant representing the events of the disaster scenario

image quality and whether it affected their ability to complete the required task. This was followed by an interview discussing general impressions of the experiments and the motivation behind the answers given in the questionnaires.

6.4. Results

Test participants were asked to create situation maps that represented the main events of the disaster situation they saw. In order to evaluate their performance, key-maps were created for each slideshow representing the ideal recreation of the events shown in the displayed images. The maps created by the users were then compared to those key-maps, and points were subtracted based on whether the actors, objects, and events were included in the map in the correct way. The location indicated on the map was also taken into consideration, subtracting points if it deviated from the positions on the key-maps. After taking all these aspects into account, the performance resulted in a ratio score that ranges from 0 (map is completely wrong) to 1 (an exact match of the key-map).

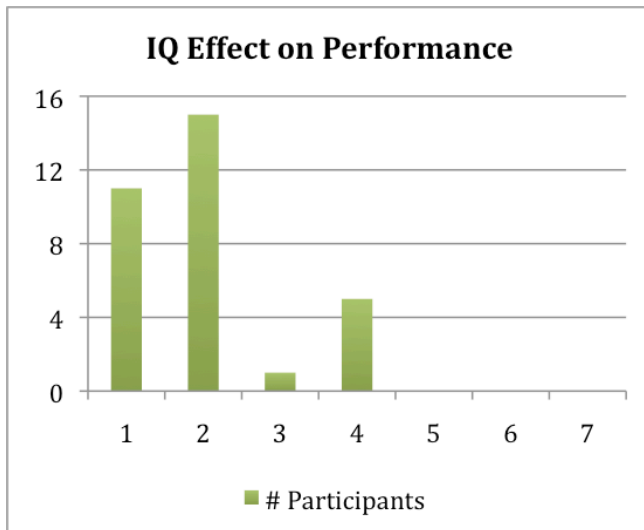


Figure. 6.5. Participant response to the question: "Do you think the picture quality affected your ability to perform the required task?", with 1=Very Uninfluential, and 7= Very Influential.

By comparing the performance with respect to the image quality in the slideshows we saw that the mean performance was higher for the HQ images ($M=0.65$, $SD=0.18$) than for the LQ images ($M=0.61$, $SD=0.20$). However, when we applied a one-way ANOVA test, the difference in the means was not found to be statistically significant ($F= 0.67$, $df=1$, $p=0.41$). This indicates that no significant effect of image quality on performance was detected using a sample size of 32 participants.

The subjective responses given to the questionnaire confirmed that test participants indeed did not think that the quality of the images affected their performance (see Figure 6.5).

A more surprising result was that the majority of the participants even indicated that they did not notice the difference in the image quality as shown in Figure 6.6.

6.5. Discussion

Looking at the sample mean performance of remembering details of a

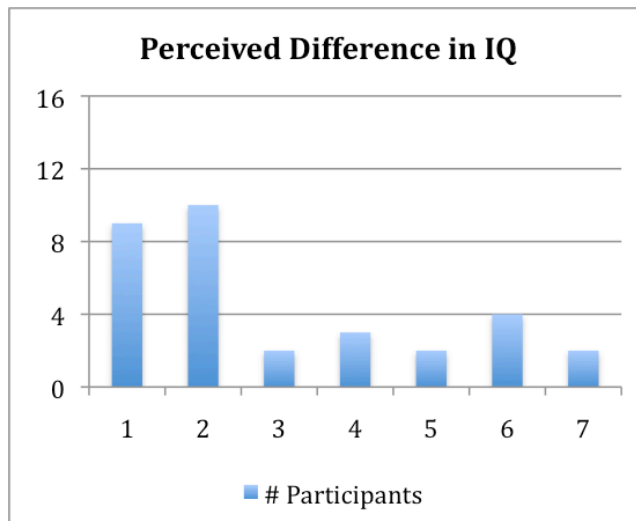


Figure 6.6. Participant response to the question: "Did you notice a difference in the image quality between the two scenarios?", with 1= No Difference and 7=Strong Difference.

disaster situation with respect to quality of images representing the disaster situation, there is indeed a slight increase in the accuracy of the generated maps when higher quality images are used. However, the relatively low statistical significance for a group of 32 participants suggests that if any effect is present, it cannot be considered as highly relevant. We therefore assume that the image quality did not hinder the performance of these non-expert participants in remembering the disaster situation. This assumption is supported by the subjective data collected at the end of the experiment, where most participants stated that the image quality did not influence their ability to complete the task.

A more surprising result was that the majority of the participants who completed the experiment did not notice the difference in quality between the sideshows in the two scenarios. As shown in Figure 6.2 above, the difference in quality was quite significant. In addition, the slides that contained text had sharp high contrast edges, which significantly deteriorated in the JPEG compressed images. Despite this huge difference in quality, people did not notice it, mainly because they were fully concentrated on the task. Information gathered in the interviews indeed confirmed that the experiment had put participants under high stress. This was caused by the difficulty of the task, and the relatively short time the images were displayed on the screen with respect to the amount of information participants needed to notice and remember. With their attention focused solely on the given task, they were mostly oblivious to the quality of the displayed images.

6.6. Conclusions

This paper shows that the performance in reconstructing disaster situations does not suffer from the sharp reduction in the quality of images representing the situation. As long as the necessary information is visible, low quality images are sufficient to give the viewer an understanding of the situation map of the disaster location. This can allow administrators to save resources (such as transmission bandwidth and data storage space) which are always in limited supply during rescue operations.

Another important conclusion is that the task performed by the viewers

can mask obvious flaws in image quality. This can have implications for media generated for specific uses (such as instructional videos or educational lectures). In such applications, high image quality is not only unnecessary, but also unnoticeable by the viewers. This implies that while generating such content, it is possible to ignore the image quality aspect and concentrate solely on the content.

References

1. Eckert, M. P., & Bradley, A. P. (1998). Perceptual quality metrics applied to still image compression. *Signal processing*, 70(3), 177-200.
2. Sheikh, H. R., Sabir, M. F., & Bovik, A. C. (2006). A statistical evaluation of recent full reference image quality assessment algorithms. *Image Processing, IEEE Transactions on*, 15(11), 3440-3451.
3. Liu, H., Redi, J., Alers, H., Zunino, R., & Heynderickx, I. (2010). No-reference image quality assessment based on localized gradient statistics: application to JPEG and JPEG2000. In *IS&T/SPIE Electronic Imaging* (pp. 75271F-75271F). International Society for Optics and Photonics.
4. Rouse, D. M., Pepion, R., Hemami, S. S., & Le Callet, P. (2009, February). Image utility assessment and a relationship with image quality assessment. In *IS&T/SPIE Electronic Imaging* (pp. 724010-724010). International Society for Optics and Photonics.
5. Rouse, D. M., & Hemami, S. S. (2008). Analyzing the role of visual structure in the recognition of natural image content with multi-scale SSIM. In *Electronic Imaging 2008* (pp. 680615-680615). International Society for Optics and Photonics.
6. Gunawan, L. T., Oomes, A. H., Neerincx, M., Brinkman, W. P., & Alers, H. (2009, September). Collaborative situational mapping during emergency response. In *European Conference on Cognitive Ergonomics: Designing beyond the Product--Understanding Activity and User Experience in Ubiquitous Environments* (p. 6). VTT Technical Research Centre of Finland.
7. Gunawan, L. T., Oomes, A. H., & Yang, Z. (2009). Navigation Support for the Walking Wounded. In *Universal Access in Human-Computer Interaction. Applications and Services* (pp. 197-206). Springer Berlin Heidelberg.

7.

Thesis discussion

In this research we explored the way humans generate a subjective opinion about the quality of image content, taking into account how they divide attention over the content. Considering the complicated interplay of the human visual system (HVS) and the subsequent perceptual and cognitive processes, image and video quality assessment is by far not fully understood, and including how attention affects quality assessment is just at the beginning of being explored. More particularly, our research focused on three research questions. First, we wanted to know how the task given to the observer and the quality level of the stimulus affected their viewing behavior, and how that would be different between images and videos. Second, we were interested in understanding how the observer would evaluate the overall quality of a stimulus if different parts of the scene conveyed a different level of quality, and whether that would differ between images and videos. Finally, we also wanted to know whether the task given to the observer could mask the perception of artifacts in the scene.

We determined from previous work in the field that a good way to build a better understanding of consumers' viewing behavior and subsequent quality assessment is through the concept of visual attention (VA). This research has therefore centered around subjective experiments conducted with the aid of an eye tracking system. In order to have comparable data between image and video content, the research tried, as much as possible, to follow the same approach for the experiments investigating still images and the experiments investigating video.

The impact of task on viewing behavior was evaluated for two tasks often occurring in this type of research: freely looking at the images or scoring the quality of the images. The quality of the stimuli varied by compressing the content at different quality levels. The resulting sets of visual stimuli were shown to the participants, while collecting data about their viewing behavior with the help of the eye tracking equipment. This same setup

was used for both still images and video content. The outcome of this first part of the research is discussed in more detail below.

In order to determine the role different parts of the visual scene played in determining the overall perceived quality, special content was created that combined visual stimuli with different levels of visual degradation in different regions. These regions were carefully selected based on eye tracking data. These data were used to identify the region of interest (ROI) in the scene and the remaining background (BG) region of the scene. Subsequently, for the combined stimuli the ROI was given either a higher or lower quality level than the BG region. From the scores given to the overall quality of the scene, it was possible to determine how much each region affected the quality judgment. Again, this approach was repeated for both image and video content, and also these results are discussed in more detail below.

Finally, we wanted to know whether a task inflicting a high cognitive load on the viewer affected how this viewer would perceive visual quality. We therefore simulated a disaster situation in a miniaturized world and created a scenario where the viewer would play the role of a witness assisting the rescue services. By showing the viewer a slideshow of images for a limited time and with a highly demanding cognitive task, we were interested in knowing whether that viewer spotted the difference in visual quality between different of these images and how that would affect how well he or she performed the required task. Also, these results are shortly discussed below.

7.1. Using eye tracking for visual analysis

Using eye-tracking to analyze viewing behavior is not a novel approach. It has already been shown to be an effective way to understand how we look at images under different circumstances [1]. Here we expanded on this basic premise by using the eye tracker under controlled conditions in different experimental settings. These experimental settings spanned differences in task, image properties, and dynamics (i.e., by including videos). Analyzing viewing behavior was accomplished by first constructing saliency maps representing the viewer's attention to different regions in the image content, and by then studying their mutual similarity.

This approach proved useful in providing insights in how visual attention differed with task and image quality level for static images. The visual attention was concentrated more in the ROI when observers were looking freely at the images. For dynamic videos, however, we didn't find an effect of task or quality level on ROI, implying that in videos people more constantly look to the natural ROI, independent of the task or the quality level.

When studying the effect of task and content integrity on the viewing behavior, we tried to keep the approach as consistent as possible between still images and videos. Yet, it is still difficult to say why the conclusions differ between still images and videos. One thing is clear though, the dynamic nature of the video content introduces extra levels of complexity in the analysis of viewing behavior via saliency maps. The content of the scene changes 24 times a second providing only 0.04 seconds of saliency information per frame. At this speed, an eye tracker running at 50 Hz only has two samples, a hardly large enough sample for proper analysis. Therefore averaging saliency from several frames is necessary, but the question is how to do that. One may argue that frames with similar visual properties should be grouped together, meaning that the duration of each time stamp should depend on the changes in characteristics of the video content. Clearly, scene cuts may be used for this purpose, but scenes may also change considerably in characteristics between scene cuts. Hence, such grouping of scenes is not uniquely defined, and so, carries the risk of influencing the results by the subjective selection of segmentation. Additionally, if various time stamps have a different duration, different parts of the video get saliency maps with a different level of accuracy, which also may influence subsequent analysis. The other approach is to split the data into constant time segments (i.e., the approach we followed in this book). The question then is how big should this segment be? Keep in mind that the scene is changing and that stimuli in the previous scene may still influence the viewer. Also, human reaction causes delay, which results in a corresponding delay in the gathered saliency data. What would be an appropriate way to compensate for such a delay? These questions are still unanswered, and the particular choices made in this research may have affected the observed difference in the effect of task and image integrity on saliency between images and video.

An additional difference is the lower number of people that participated in the experiment with video than in the experiment with still images. The latter was simply due to the longer duration of the experiment with videos and the logistics of securing test participants to take part in such a time consuming experiment. As such, the saliency maps obtained with video may still suffer from more inter-observer variability.

One way to address inter-observer variability is by using the Upper Empirical Similarity Limit (UESL) [2], an analysis used in different parts of this thesis. This analysis seems to be sound in theory, and has become the de-facto approach to examine saliency data. The upper limit is the similarity in saliency between data collected under the same conditions from different participants, and as such is a measure for the difference in saliency caused by inter-observer variability. Therefore, as the name implies, it should fall above the similarity in saliency between two datasets collected under different conditions. In this work, we applied the UESL approach on saliency data collected for still images and videos in chapters 2 and 4, respectively. In both chapters, the UESL analysis was performed for four different similarity measures used to compare the level of similarity between two saliency maps. These similarity measures are: Linear Correlation Coefficient (LCC), Kullback-Leibler divergence (KLD), Normalized Scanpath Saliency (NSS), and the Structural Similarity Index (SSIM). When utilized to our data, the UESL did not function as expected. When analyzing videos, the UESL repeatedly was smaller than the similarity calculated between saliency maps measured under different conditions, i.e. task and video integrity, for each of the 4 similarity measures applied. So far, we have no explanation for this unexpected result.

One advantage of having a large corpus of data like the one used in Chapter 2 is that it allows us to examine the analysis methodology across a large sample of data collected and processed in the same manner. Interestingly, the UESL did not work properly with the LCC and KLD, again giving values falling below those comparing saliency maps collected under different conditions. It therefore seems from this work that UESL is, at best, not always reliable to analyze visual attention data. At least not with the used methodology.

7.2. Task effect on viewing behavior

It has long been known that changing the task of the observer can fundamentally change his or her viewing behavior [3]. Here we focus on the task of scoring visual quality. This task is important for the field of image quality perception since it is needed to collect subjective scores for test images and videos. These scores are then collected in reference databases and considered to be the “ground truth” which objective models should aspire to match.

Our results show that the scoring task indeed affects the viewing behavior significantly. The region of interest of the image is no longer the face or the animal shown in the picture. Instead the task has introduced new interesting features that the viewer is attentive to, namely the noise and artifacts. The natural region of interest is still important and initially captures the attention of the observer, but that attention quickly deviates to the background region.

The above finding is clearly observed when asking people to view still images, but what happens in the case of videos? First, one should keep in mind that videos are different from images since they are dynamic. As explained above, this means that the way we analyze how viewers look at the region of interest over time is different in nature from the approach used for images. Having said that, we still have to state that the type of analysis that we were able to perform did not yield conclusive results.

One way to interpret the lack of finding a change in the region of interest with task for videos is to say that the scoring task did distract the viewer less when watching video content than when watching still images. The continuously changing region of interest in videos is difficult to ignore, and therefore, the viewer does not have a chance to start scanning the scene for artifacts. Indeed, the video segments we chose for the experiment were of action scenes selected specifically to have a clear and active ROI. Since we already saw in still images that the natural ROI initially captured the attention of the viewer regardless of the task, we can argue that a continuously changing ROI continues to capture the attention of the observer.

On the other hand, we should remember that the results obtained with videos exhibited peculiar outcomes with respect to the comparison of similarity in saliency maps with the UESL. For all four similarity measures used in the analysis, the UESL occasionally was lower than the similarity between saliency maps measured under different conditions, as explained above. This may be an indication that the results are not reliable and that modifications are needed in the analysis or in the experimental setup for videos.

Since this research was originally initiated to improve the visual experience of watching TV, it is interesting to point out a design dilemma that the results indicate. When people are shopping for a new TV set, they are usually interested in evaluating the image quality of different sets in order to choose the one with the best picture quality. However, this is probably the only time that they examine the image quality of the TV set this closely. Since this work shows that there are differences in viewing behavior caused by viewing tasks, TV manufacturers are left with a critical choice. They can either optimize their TV sets for the viewing behavior during quality assessment, meaning that they will outperform other TV sets in the showroom and sell more units. Or, they can optimize for the free looking viewing behavior, and thereby giving their clients a better home viewing experience. An ideal solution is to have a showroom mode that optimizes the TV settings for the former condition and a normal mode for the latter condition. Such a showroom mode currently is included in high-end TVs, but basically focuses on brightness and color rendering. Taking care that image integrity is high over the whole TV-screen in the showroom mode, while mainly focusing all post-processing capacity on the natural ROI in the normal-use mode may be considered as an option to further improve a TV's overall perceived quality.

7.3. Importance of ROI

The region of interest in visual stimuli is a recurring point of focus in this thesis. Moreover, the results from most of the experiments show that the ROI is worth our attention. We saw that the ROI is the first thing that the viewers' gaze gravitates to. It is the first part of the scene examined by the viewers, thereby becoming the first element that shapes the viewing experience. Moreover, even when the scene is a static image and the

viewers are given the freedom to look at it for as long as they want, we see that they keep coming back to the ROI and examine it throughout their visual experience. This is a clear indication that the ROI is indeed the most important part of the scene.

When it comes to visual quality assessment, the results from this research leave little room for doubt. When assessing the visual quality of a scene, the quality judgment of the observers is mainly based on the visual quality of the ROI. No matter whether it has a higher or lower quality than the rest of the image, the quality judgment tends to follow the quality level of the ROI. This effect is found to be twice as strong in videos as in images, and so supports the above findings that the viewer is more drawn to the natural saliency of the scene when viewing videos than when viewing images.

An interesting point of discussion raised by these findings is whether there is a direct relation between viewing behavior and quality judgment. We have seen before that the viewing behavior changes when the observer is given a scoring task. Yet, here we see that the quality score given to an image mainly follows the quality of the ROI. Keep in mind that these ROI areas were identified using natural scene saliency. However, we showed that when scoring compromised images, the viewers' attention shifts from the natural scene saliency in search for image artifacts, making such artifacts a new ROI. Nevertheless, when giving a quality score, viewers seem to base their scoring on the quality of the natural ROI.

These findings also allow for a great potential to apply lossy compression intelligently. Since the quality judgment of the viewer is mainly based on the ROI, it means that it is possible to compromise the quality of the rest of the scene without losing much of the “perceived” quality. One challenge that immediately comes to mind is how to obtain the natural saliency information, so that compression can be applied to the background region. A possible solution is to gather this saliency information, when the content is created. There are many points during the creation of image and video content that allow for collecting saliency information. For example, it can be collected while the content is being recorded from the camera operator, or while the images or videos are

being culled and edited. There are also opportunities to collect saliency data before the content is prepared for consumption. The collected information on the ROI can then be added as metadata to the video stream. With this additional information, there is a great potential for compression optimization with little loss in perceived quality.

7.4. Quality masking by task

One interesting conclusion of this work is the realization that there is a limit to how much image integrity is required depending on the viewing conditions. The amount of time the viewer is allowed to look at images, and the task the viewer is requested to perform have the ability to suppress their quality perception. Even if the region of interest is degraded in quality, the viewer may not bother by it as long as he or she is not impaired in performing a task, the latter being proven for at least tasks with a high cognitive load. So, high cognitive loads and time limitations may mask quality perception of image content, making an otherwise disturbing loss of image quality completely unperceivable. This uncovers yet another dimension in the way we look at images. It is clearly a complicated process that we are just starting to understand.

References

1. Buswell, G. T. (1935). How people look at pictures: a study of the psychology and perception in art.
2. Stankiewicz, B. J., Anderson, N. J., & Moore, R. J. (2011). Using performance efficiency for testing and optimization of visual attention models. In *IS&T/SPIE Electronic Imaging* (pp. 78670Y-78670Y). International Society for Optics and Photonics.
3. Yarbus, A. L. (1967). *Eye movements and vision* (Vol. 2, No. 5.10). L. A. Riggs (Ed.). New York: Plenum press.

8.

Conclusions and Recommendations

8.1 Thesis conclusions

The different experiments described in the separate chapters of this thesis contribute to a better understanding of how humans look at images and videos. We know that scenes have a naturally salient region. The attention of the viewers is initially gravitated towards this region, no matter what task the viewers are given. Moreover, they keep examining this salient area repeatedly even after starting to explore the rest of the scene. However, how the rest of the scene is subsequently explored is affected by the task the viewers are given. If they are given the task of scoring the quality of the scene, their attention shifts to areas with visible artifacts. However, even then the salient part of the image content remains important, since that is what they focus on to make sense of the scene and try to comprehend and understand it. The background area of the scene is of little importance, and mainly gets attention only for scoring an image that is visibly degraded in the background area.

In addition to the above, we have also shown that the viewer tends to score the visual quality of images based on the quality of its most important regions. In a similar manner, the viewer tends to score the visual quality of videos based on the quality of its most important regions. Hence, understanding the region of interest of an image or video may allow more optimal data compression by keeping the quality of the important regions high, while degrading the background more. Finally, we have seen that putting a high cognitive load on the viewer can mask visible artifacts in images, and hence, in these applications having efficient information transfer is most important.

8.2 Thesis recommendations

Despite the interesting findings related to the research presented in this thesis, there are still a number of open endings that might require

additional research. These open endings relate to the way we analyzed our results, but also to further fine-tune some of the relations we found.

Time duration in video analysis.

In order to analyze the eye tracking data for the video content, we chose to average each one second of eye tracking data together in one saliency map. This choice of a fixed time interval is by no means optimal. It does not correspond to the changes in video content, and therefore a saliency map can combine saliency information of multiple scenes not fitting logically together (e.g., split by a scene cut). We suggest that similar research in the future generates video clips that remain consistent for equal time intervals, which can then be used to segment the eye tracking data. Filming video material specifically for the experiment will also allow for having fewer variables in the scene that can affect the outcome of the experiment (e.g. camera movements, activity in the background of the scene). On the other hand, creating such controlled content may skew the results in a way that using realistic content, as we did in this work, would not.

The Upper Empirical Similarity Limit Approach

To compensate for the inter-observer variability, the research community has adopted the Upper Empirical Similarity Limit as an approach. This approach examines two sets of data collected under the same conditions to determine what the upper limit of similarity between viewers can be. The assumption is that comparing the similarity of data collected under different conditions will always fall below that limit.

In this thesis, we have applied the UESL approach to a huge set of data, which included images and video, and while using various similarity measures. In some cases we saw that similarity between saliency maps obtained under different conditions was larger than the UESL similarity between saliency maps obtained for the same content and task. We should remember that both similarities are statistical measures. Therefore, it may be possible that one indeed is higher than the other. The question then is how accurately should the UESL be measured to

assure that it would be a more reliable upper limit.

We recommend further research to investigate the UESL approach in order to define how to apply it reliably. So far, the approach is not standardized in terms of how many participants to use for having an accurate UESL and which similarity measure to use. It would be helpful to apply the UESL analysis on large sets of data using the same approach in order to define when stability in UESL is used and to determine which similarity measure gives the most reliable information. The approach used in Chapter 2 of this book can serve as a good example for such a study.

Stimuli with less apparent ROI

In the presented studies, the chosen images and videos contained a clear ROI that captures the viewer's attention. Images contained something like a face of a human or an animal, while the videos showed fast action sequence as a ROI. Such stimuli was chosen to be a proof of concept in order to explore if there is any merit to this line of work. As a result, the conclusions we reached and the models we constructed all apply for stimuli with similar visual properties. It is now important to expand the work to include other type of stimuli such as images with uniform textures or a news broadcast where a person remains central in the scene. It will be interesting to see how much the results will differ when no clear ROI is present.

The threshold of cognitive load and quality perception

In Chapter 6 we saw that the cognitive load of the observer can mask bad quality of the shown images. The experiment, however, was limited to one specific level of cognitive load and image quality. We expect a relation between cognitive load of the image material and image degradation that is affordable. We recommend that further studies investigate a larger range of quality levels and cognitive load levels to establish a reference range. It will also be interesting to see how the task load affects quality perception in video content.

Acknowledgments

First and foremost I would like to thank my promotor Ingrid Heynderickx who have given me ample amount of help and support throughout my time working on this project and beyond. She has always been available for discussions whether the topic is related to the research or not. Always accommodating and understanding, it is difficult to imagine how any other supervisor could have done this job better. The successful fruition of this work is as much a result of her efforts as it is of mine.

Second mention goes to my companion in this journey, Nike Gunawan. As colleagues and friends, our activities included us partnering up in performing research, authoring scientific articles, pursuing our passion for photography, crafting countless projects, and starting our own company, just to name a few. I have always found her seemingly limitless energy to be a great motivator, and working with her made no obstacle seem too difficult. I want to thank her for the unforgettable times together and wish for many more to come as we are connected at the soul.

In addition, I am grateful to Hantao and Judith for their valuable support in my research. A special mention goes to Willem-Paul, my always helpful advisor on statistical challenges. My research mentors Zhenke and Ramon who gave me the benefit of their years of experience as PhD students, as well as Pascal and Joost who served as my role models for what a sound researcher should be like. My group comrades Tim, Kristina, Maaike, Chris, Alina, and Iris for the countless adventures in and outside the university, and in and outside the country. Yun, Chao, Yangyang, Vanessa, Iulia, Siska, Junchao, and Ni for being wonderful colleagues and dear friends as well. Anela for her help in translating the propositions of this thesis to Dutch. Our group's tireless guardians Bart and Ruud who always welcomed me with a smile and helped me solve any technical problems I came across. Toss, Helen and Anita, who made sure that all arrangements were taken care off while I was working at the group.

Of course, I am immensely grateful to my parents without whom none of this would have been possible. Their looking over me and supporting me

through my life and managing to make wise choices for us through good and challenging times is the reason I have reached where I am today. Knowing that I can rely on their unconditional love and support gives me the motivation to keep going, and their faith in my abilities is the source of my confidence. Also my two brothers and partners in my life journey. They are the people most similar to me in this world even though we are each so different. I want to thank them both for being there for me. Especially my brother Zaid who has been by my side longer than any other person on earth. He is more than my brother as he is also my best friend. My admiration to him is something that cannot be put in words. Last but not least, I would like to thank God for watching over me, because those familiar with all details of my PhD journey know that it would have never been completed without one or two (minor) miracles.

List of Publications

Journal

1. Alers, H., Redi, J., Liu, H., & Heynderickx, I. (2015). Effects of Task and Image Properties on Visual Attention Deployment in Image Quality Assessment. *Journal of Electronic Imaging*.
2. Alers, H., Redi, J., & Heynderickx, I. (2014). Quantifying the Importance of Preserving Video Quality in Visually Important Regions. *Journal of Signal Processing: Image Communication*
3. Alers, H., Redi, J., Liu, H., & Heynderickx, I. (2013). Studying the effect of optimizing image quality in salient regions at the expense of background content. *Journal of Electronic Imaging*, 22(4), 043012-043012.
4. Liu, H., Zunino, R., Heynderickx, I., Redi, J., & Alers, H. (2011). Efficient neural-network-based no-reference approach to an overall quality metric for JPEG and JPEG2000 compressed images. *Journal of Electronic Imaging*, 20(4), 043007-043007
5. Gunawan, L. T., Alers, H., Brinkman, W. P., & Neerincx, M. A. (2011). Distributed collaborative situation-map making for disaster response. *Interacting with Computers*, 23(4), 308-316.

Conference

6. Fitrianie, S., Hultgren, A., Alers, H., & Guldmond, N. A. (2013). A SmartTV Platform for Wellbeing, Care and Social Support for Elderly at Home. In *Inclusive Society: Health and Wellbeing in the Community, and Care at Home* (pp. 94-101). Springer Berlin Heidelberg.
7. Hultgren, A., Detweiler, C., Alers, H., Fitrianie, S., & Guldmond, N. A. (2013). Towards Community-Based Co-creation. In *Human Factors in Computing and Informatics* (pp. 585-592). Springer Berlin Heidelberg.
8. Alers, H., Redi, J. A., & Heynderickx, I. (2012). Examining the effect of task on viewing behavior in videos using saliency maps. In *IS&T/SPIE Electronic Imaging* (pp. 82910X-82910X). International Society for Optics and Photonics.
9. Pommeranz, A., Fitrianie, S., Alers, H., & Guldmond, N. (2012). Care@ Home: An integrated approach to care and social inclusion of elderly. In *AAL Forum, Eindhoven*.
10. Alers, H., Bos, L., & Heynderickx, I. (2011). How the task of evaluating image quality influences viewing behavior. In *Quality of Multimedia Experience (QoMEX), 2011 Third International Workshop on* (pp. 167-172). IEEE.

11. Gunawan, L., Alers, H., Brinkman, W. P., & Neerincx, M. (2010). Effect of map sharing and confidence information in situation-map making. In *Proceedings of the 28th Annual European Conference on Cognitive Ergonomics* (pp. 41-48). ACM.
12. Alers, H., Gunawan, L. T., Brinkman, W. P., & Heynderickx, I. (2010). Effect of image quality on disaster response applications. In *Quality of Multimedia Experience (QoMEX), 2010 Second International Workshop on* (pp. 41-45). IEEE.
13. Liu, H., Redi, J., Alers, H., Zunino, R., & Heynderickx, I. (2010). No-reference image quality assessment based on localized gradient statistics: application to JPEG and JPEG2000. In *IS&T/SPIE Electronic Imaging* (pp. 75271F-75271F). International Society for Optics and Photonics.
14. Redi, J., Liu, H., Alers, H., Zunino, R., & Heynderickx, I. (2010). Comparing subjective image quality measurement methods for the creation of public databases. In *IS&T/SPIE Electronic Imaging* (pp. 752903-752903). International Society for Optics and Photonics.
15. Gunawan, L. T., Oomes, A. H., Neerincx, M., Brinkman, W. P., & Alers, H. (2009). Collaborative situational mapping during emergency response. In *European Conference on Cognitive Ergonomics: Designing beyond the Product---Understanding Activity and User Experience in Ubiquitous Environments* (p. 6). VTT Technical Research Centre of Finland.
16. Gunawan, L. T., Oomes, A. H., Neerincx, M., Brinkman, W. P., & Alers, H. (2009). Collaborative situational mapping during emergency response. In *European Conference on Cognitive Ergonomics: Designing beyond the Product---Understanding Activity and User Experience in Ubiquitous Environments* (p. 6). VTT Technical Research Centre of Finland.
17. Alers, H., Liu, H., Bos, L., & Heynderickx, I. (2009). Using eye tracking to assess the task effect on viewing behaviour. In *PERCEPTION* (Vol. 38, pp. 23-23). 207 BRONDES BURY PARK, LONDON NW2 5JN, ENGLAND: PION LTD.
18. Mast C. A. P. G., Alers H. (2008). Comparing Web-Based Services in HCI Teaching, In *Proceedings of HCI2008 workshop - HCI for technology enhanced learning* (pp.29-35) isbn: 9789-0-813811-2-3.
19. Alers, H., Mast, C. A. P. G. (2008). A Case Study of Using Web-Based Services in Higher Education. *Euromedia'2008*.

Thesis Propositions

English list of propositions

1. Asking to evaluate the visual integrity of images or videos significantly changes how an observer looks at this content.
2. A viewer's judgment of visual integrity is mainly based on a region of interest in images or videos.
3. High cognitive load masks degradation in visual integrity in images.
4. Convincing people that "Trickle Down Economics" is a sound economic principle is not an indication of economical skills.
5. The extensive use of SPSS by usability experts has not resulted in improved usability of SPSS.
6. Science dissemination is just as important as science generation.
7. Changes of our own values from one decade to the other do not result in more tolerance for the values of other cultures.
8. The fact that the educational system has hardly changed in the past few decades is an indication of woeful deficiencies.
9. Discussing the existence of God will never reach a definitive conclusion.
10. It is ironic that researchers keep complaining about how the scientific community recognizes their achievements, even though the community is run by the researchers.

Dutch list of propositions (proefschrift stellingen)

1. Gevraagd worden om de kwaliteit van beelden te beoordelen, beïnvloedt aanzienlijk de wijze waarop we die beelden waarnemen.
2. Ons oordeel over beeldkwaliteit wordt hoofdzakelijk bepaald door de plek op de foto of video waar we ons op focussen.
3. Hoge cognitieve belasting maskeert de verslechterde kwaliteit van beelden.
4. In staat zijn om mensen te overtuigen dat “Trickle Down Economics” een gezond economisch beginsel is, impliceert geen economische vaardigheden.
5. Het is ironisch dat veelvuldig gebruik van SPSS door usability experts niet heeft geresulteerd in het verbeteren van gebruiksvriendelijkheid van het programma.
6. Wetenschap overbrengen is even belangrijk als wetenschap creëren.
7. Veranderingen van onze eigen waarden door de tijden heen hebben niet geresulteerd in meer tolerantie voor waarden van andere culturen.
8. Het feit dat het onderwijssysteem nauwelijks is veranderd in de afgelopen tientallen jaren impliceert verschrikkelijke tekortkomingen ervan.
9. Discussies over het bestaan van God zullen nooit leiden tot een eenduidige conclusie.
10. Het is ironisch dat onderzoekers klagen over de wijze waarop zij erkend worden in de wetenschap terwijl zijzelf de wetenschap aansturen.