THE EFFECT OF TASK ON VISUAL ATTENTION AND ITS APPLICATION TO IMAGE QUALITY ASSESSMENT METRICS

Lennart Bos

Supervisors:

Prof.dr. Ingrid Heynderickx

Ir. Hani Alers

MSc. Hantao Liu

Technical University of Delft

Department of Man-Machine Interaction (MMI)

Faculty Electrical Engineering, Mathematics

and Computer Science (EEMCS)

Mekelweg 4,

2628 CD Delft,

The Netherlands

February 2010

TU Delft

Table of Contents

Table of Figures

Abbreviations

ANOVA	Analysis of Variance. A statistical procedure to test whether two groups of variables originate from different populations.
CSF	Contrast Sensitivity Function. The sensitivity of the **HVS** to contrast as a function of the spatial frequencies.
DMOS	Differential **MOS**. The difference between the **MOS** of a test image and the **MOS** of its corresponding reference.
FR	Full-Reference. **IQA** metrics categorized as **FR** evaluate the quality of an image by comparing it with another full image, which is assumed to be a perfect quality.
GAFFE	Gaze-Attentive Fixation Finding Engine. Computational visual attention model that is able to find the fixation locations in a given picture.
GBIM	Generalized Block-edge Impairment Metric. A **NR IQA** metric specialized in the detection of blocking artefacts in digital images and video.
HVS	Human Visual System. A scientific model that is used in many research fields involved in digital image and video processing to simplify the complex behaviour of our perception.
IOR	Inhibition of Return. A property of the **HVS** that discourages attention from focusing on a previously attended region.
IQA	Image Quality Assessment. A research field that aims to objectively determine the perceptual quality of images.
JPEG	Joint Photographic Experts Group. One of the most commonly used digital image compression techniques, named after its inventors. **JPEG** compressed images can be identified by their `.jpg` file extensions.
JPEG2000	Joint Photographic Experts Group 2000. A compression standard designed by the Joint Photographic Experts Group in the year 2000, with the intention of superseding their original **JPEG** standard.
LIVE	Laboratory for Image & Video Engineering. One of the main contributors to **IQA** research, located in the University of Texas.
MOS	Mean Opinion Score. The average subjective score determined by a group of human test subjects.
MSE	Mean Squared Error. An objective measure to quantify the difference between two variables.

NR	No-Reference. **IQA** metrics categorized as **NR** are able to determine the quality of a given image without the need for any information that could serve as a reference.
PCC	Pearson's product-moment Correlation Coefficient. A widely used scientific formula to measure the linear dependence between two variables.
PSNR	Peak Signal-to-Noise Ratio. A measure similar to the **MSE** that calculates the ratio between the maximum possible power of a given signal and the power of the noise distorting that signal.
ROI	Region of Interest. The region in an image that draws the most attention from its observers.
RR	Reduced-Reference. **IQA** metrics categorized as **RR** are a compromise between **FF** and **NR** metrics: they only require a set of features extracted from the reference image to determine the quality of a test image.
SPSS	Statistical Package for the Social Sciences. A software application capable of many complicated statistical analysis procedures.
SROCC	Spearman's Ranked Ordered Correlation Coefficient. A widely used scientific formula to measure the non-linear dependence between two variables.
SSIM	Structural Similarity. The **SSIM** index is a popular **FR IQA** metric able to predict the perceptual quality of a given image.
VIF	Visual Information Fidelity. A **FR IQA** metric that is argued to be superior to the **SSIM**.

1. Introduction

Over the last few decades, technology has become an integral part of modern life. In every field of technology new achievements are made at an extraordinary rate. The field of digital photography is no exception. Today, most mobile phones are equipped with a small digital camera and more advanced high resolution cameras are becoming affordable to the average consumer. The ever-growing consumer demand for high quality digital photography is bounded only by our technological limitation. The increasing photo quality is too much for the storage and transmission capacity to handle. Hence, besides trying to raise the size and speed of the storage and network devices, reducing the size of digital images is essential. The field of digital image compression deals with this particular problem by storing images more efficiently.

The number of bytes needed to describe an image can be reduced by describing it in a different and more efficient code while preserving its quality, which is called *lossless* compression. In contrast, *lossy* compression is accompanied by a reduction in image quality. For this reason, lossy compression is theoretically able to obtain higher compression ratios than lossless compression. Moreover, the reduction of image quality is not necessarily perceivable by the human eye. Due to the limitations of our visual system, the image quality can be greatly reduced without any major visible effects.

To test the performance of an image compression algorithm, or any image processing algorithm, it is essential to be able to measure the perceptual quality of the image produced by the algorithm. Since humans are usually the ultimate receiver of images, the most common-sense approach to determine the quality of an image is to simply let people rate it. However, opinions differ greatly per individual, so in order to obtain an accurate and unbiased score, a large group of people is required, preferably from a variety of different ages, genders, ethnicities, and so forth. The average of all scores for a given image is usually referred to as the *Mean Opinion Score* (MOS). Although this method has been used for many years, it is too cumbersome, expensive, and time-consuming to be practical in most applications. The research field of *image quality assessment* (IQA) tries to circumvent this problem by developing automatic algorithms or *metrics*, which are able to objectively calculate a quality score for an image, without the need for any human effort.

IQA metrics have many applications, e.g. they can be used to optimize all sorts of image processing, restoration, enhancing, and compression algorithms. Such algorithms often contain parameters and settings that can be modified to produce different results. An IQA metric can be used to automatically determine the combination of parameters and settings that yield the best possible result. Secondly, IQA metrics can be used to evaluate and benchmark such algorithms, enabling designers of an image processing system to choose the algorithm that best suits their purpose. And thirdly, IQA metrics can be used to monitor the image quality in certain digital systems, such as video network streaming systems, where they can make adjustments to control the image quality depending on the available bandwidth.

One of the earliest IQA measures is the *mean squared error* (MSE), which is simply a measure for the difference in pixel value between the original image and its processed version. Closely related is the *peak signal-to-noise ratio* (PSNR). Although both metrics have been widely used for many years, they do not represent the human quality preference very well and have received much criticism [1, 2]. Over the last few decennia several attempts have been made to develop a superior metric that is able to predict the quality of an image as seen by a human observer. Some of these metrics are able to quantify all kinds of distortion [3, 4, 5], yet some are specialized in one particular type of distortion, e.g. blockiness [6, 7] or blur [8].

The internal mechanisms of the aforementioned metrics are entirely different, yet they all follow the same basic scheme. A diagram of this scheme is depicted in Figure 1.1. The inputs of the metric consist of a lossily compressed, or otherwise distorted image, and its corresponding original version, which is assumed to have a perfect quality. The metric compares the original image with the distorted image to generate a distortion map. The output value, the quality score, is simply the average of the distortion map. Some metrics are able to produce a distortion map without the need for a reference image, named *No-Reference* (NF) metrics. Other metrics, known as *Reduced-Reference* (RR) metrics, manage with merely a set of features extracted from the reference image. The most commonly used metrics, however, require the full original image as reference and are therefore known as *Full-Reference* (FR) metrics. The final quality score corresponding to the distorted image is the average of the whole distortion map. Consequently, no spatial information is taken into consideration, i.e. every pixel in the image is weighted equally. This does not necessarily represent human perception. Different regions in the image might have a different influence on its quality score, e.g. compression artefacts in the foreground could be more annoying than the ones in the background. This has lead to the idea of incorporating human visual attention into IQA metrics to weight the distortion map in a heterogeneous manner.

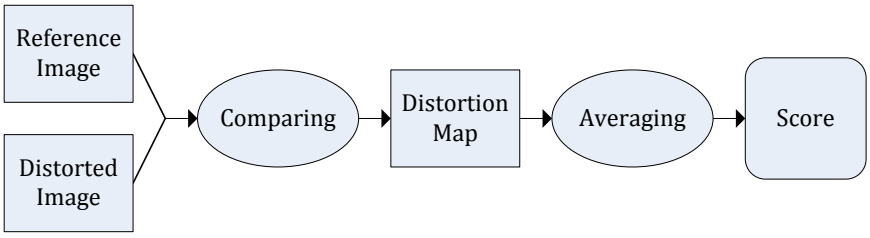


Figure 1.1: General full-reference image quality assessment metric scheme.

Visual attention has already been a research topic for decades, yet its application to IQA metrics is still relatively new. Several attempts to incorporate visual attention into a metric have already been made: some with promising results [9, 10, 11], yet others found no clear improvement [12]. The *Laboratory for Image and Video Engineering* (LIVE) has been a major contributor to the ongoing research in this field. The general approach is to obtain a *saliency map* for a given image, which depicts where a human observer looks at when viewing the image. The saliency map can be obtained in several ways, some more accurate than the other, e.g. by recording eye movements of observers with an eye tracker, by asking observers to express where they looked at, or by using an automatic computational model. The saliency map can be combined with the distortion map of the metric. The

average of this combined map is thus a weighted quality score for the corresponding image. A diagram of this scheme is depicted in Figure 1.2, where the visual attention related steps are highlighted in yellow.

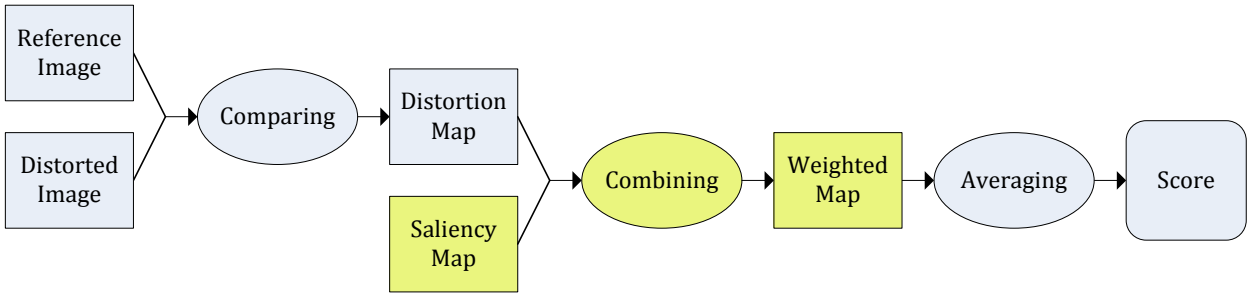


Figure 1.2: General FR metric scheme that incorporates visual attention.

The scheme of Figure 1.2 has two complications: first, how to obtain the saliency map, and second, how to combine it with the distortion map. The most accurate way of obtaining saliency is to objectively measure it. This can be done by recording the eye movements of observers with the help of an automated eye tracker. However, human ocular behaviour is not fixed and is subject to outside influences. One such influence is the goal or task of the observer. For example, an observer with the task of counting certain objects will most likely show different ocular behaviour than someone without a task. Of particular interest is the quality assessment task given to observers, since they are the people who determine the MOS of an image. It has already been suggested that there is a difference in visual behaviour between observers looking with and without a task [13, 14, 15, 16]. Different visual behaviour will result in a different saliency map, which in turn will affect the score of the metric. Therefore, we must first get a better understanding of the effects that a task can have on visual attention, which leads us to the first research question:

"What is the difference in visual attention between looking freely at an image and scoring the quality of an image?"

To answer this question an experiment will have to be done where the eye movements of test subjects are recorded while they are looking freely at an image and while they are assessing the quality of an image. From the data gathered in the experiment the saliency maps can be constructed. Hereafter, the maps can be analyzed for differences between the quality assessment task and the task-free condition. In the small experiment of C. T. Vu et al. [15], vague differences between looking freely and judging quality were found for a none-uniformly distributed distortion, e.g. JPEG compression. Our hypothesis is that with a bigger and more controlled experiment we will find a more apparent difference in visual attention between the two tasks, namely: when people are judging the image quality, they will tend to pay more attention to compression artefacts across the whole image, while when looking freely, they will focus more on the semantically informative region, also known as the *region of interest* (ROI).

After the difference in visual attention between looking freely and scoring has been analyzed, its effect on IQA metrics will be investigated. The saliency maps obtained through the experiment will be applied to a variety of different metrics in order to look for any consistent effects. Furthermore, the saliency maps from both sessions, the free looking session and the scoring session, will be applied separately to test for any differences between the two tasks. Hence, the second research question is:

"Can the performance of existing image quality assessment metrics be improved by using visual attention information?"

This question can be answered with the eye movement data of the first experiment. Additionally, LIVE also has a database available online, consisting of distorted images with corresponding saliency maps [17]. Therefore, the IQA metrics can be tested on both databases to ensure the effects are consistent across different sets of images. Our hypothesis is that applying saliency to metrics will improve their image quality prediction accuracy. Mainly because compression artefacts in the ROI are likely to be considered more annoying than artefacts in the background [12, 18]. For that reason, weighting the distortion map more heavily in the salient regions than in the rest of the image should lead to better results.

In section 2 an overview of the available literature will be given with more general background knowledge about visual attention and image quality assessment. Afterwards, the eye tracking experiment will be fully described in section 3, together with the results and important findings of a statistical analysis. In section 4 the data from the experiment will be used to try and improve existing IQA metrics. Finally, in section 5 a conclusion will be drawn with recommendations for possible future research.

2. Literature Review

In this section an overview is given of the basic background knowledge related to human perception, visual attention, and image quality assessment to familiarize the reader with the available literature.

2.1. Human Perception

Perception of humans and other animals is an extremely complex process and is currently still not entirely understood. Incoming light is processed in several steps before an image appears in our consciousness. The first steps consist of light entering the eye and converting it to electric nerve signals. This process is well-known and clearly defined. However, how the process continues in the brain after this step is less well understood and is still subject to much research.

Scientists are working on a model of human perception according to the present knowledge, known as the *Human Visual System* (HVS). This model tries to explicate the complex processes of sight and make them more practically usable [1, 19]. The model is improved whenever our understanding of human vision increases. The HVS models certain properties of vision that are important in many different application areas, e.g. display technology, photo printing, computer vision, and image/video processing. The HVS is used in these areas to create the best visual experience for the customers by optimally matching the technology with the biology, i.e. computers should only display what can actually be perceived by humans.

2.1.1. The Eye

Light is partially reflected by objects in the environment and eventually falls onto our eyes. The two human eyes are positioned inside sockets and are both rotatable with the use of six extraocular muscles [20]. Figure 2.1 shows a diagram of a cross section of the human eye. The outer layer of the eye consists of two tough parts, the transparent *cornea* that covers the anterior and the white *sclera* that covers the rest. Below the cornea lays the *iris*, with a central opening called the *pupil.* The iris can expand and contract the pupil from approximately 2 to 8 mm in order to control the amount of light entering the eye [21]. The size of the pupil only controls the intensity of the light and does not affect the total field of view [19]. Light coming through the pupil falls on the *retina* at the back of the eye, which consists of several layers of nerve cells, described in more detail in the next section. At the centre of the retina lays a highly pigmented spot, the *macula*, which is responsible for high acuity central vision.

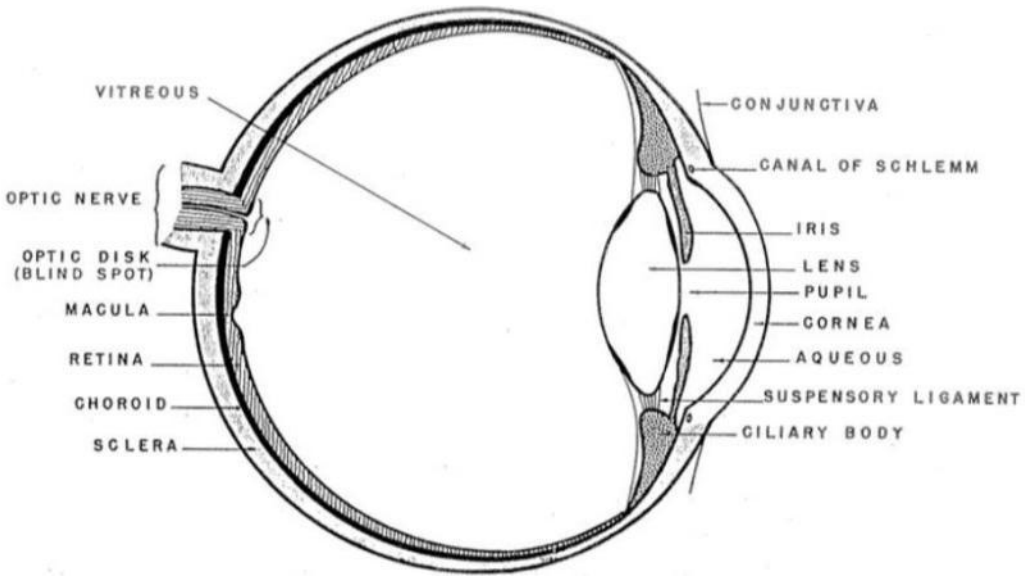


Figure 2.1: Cross section of the human eye.

The concentration of photoreceptor cells is highest in a small portion at the centre of the macula, the *fovea centralis*, a shallow circular shaped indentation with a diameter of about 0.25 mm [22]. This is much smaller than the pupil, so the incoming light has to be refracted by the *crystalline lens* to focus the light beams at the fovea. The *ciliary muscles* can increase the curvature of the lens to focus on nearby objects and decrease its curvature to focus on distant objects, schematically illustrated in Figure 2.2. This process, called *accommodation*, is very fast and can focus from a minimum distance of about 7 cm to an infinite distance in a fraction of a second. However, the ability to accommodate dramatically decreases with age [19].

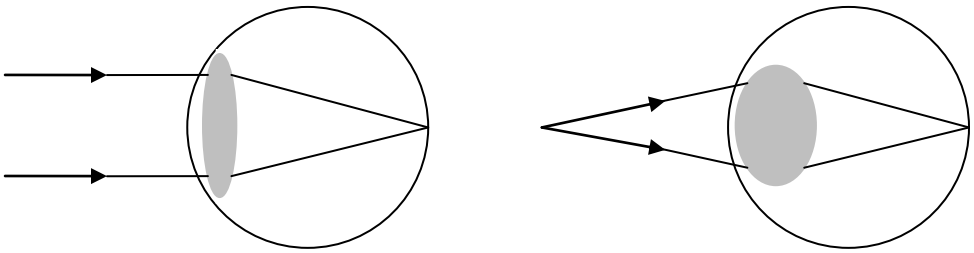


Figure 2.2: Focusing on a distant and near point by adapting the curvature of the lens.

The area where the optic nerves come together and go to the brain is called the *optic disk*. There are no photoreceptors on this disk, causing a *blind spot* in our vision. Normally, we are not consciously aware of this spot, because the brain fills in the missing information by surrounding patterns and information from the other eye [23, 24]. The actual sightlessness of the blind spot can easily be experienced by covering one eye and looking at two adjacent spots [24].

2.1.2. The Retina

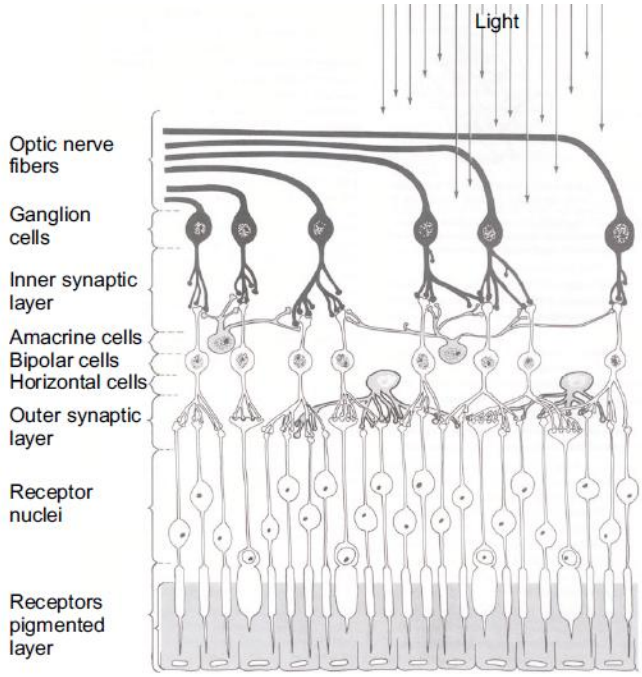


Figure 2.3: Schematic model of the different layers of cells in the retina.

The retina is the innermost layer of the eye that translates the incoming light into neural impulses with the use of several layers of nerves, shown in Figure 2.3. The incoming light passes through all but the last layer, the *pigment epithelium*, which reflects the light back onto a layer containing photoreceptors [24]. These receptors actually face away from the direction of the incoming light and catch the light after it is reflected backwards. There are two different types of photoreceptors, *rods* and *cones*, which have several different properties. The eye contains about 6 to 7 million cones which are highly concentrated at the fovea [21]. The number of rods is much larger, 100 to 150 million, and they mainly cover the periphery of the retina. Figure 2.4 shows the distribution of rods and cones for a cross section of the eye containing the fovea and the blind spot.

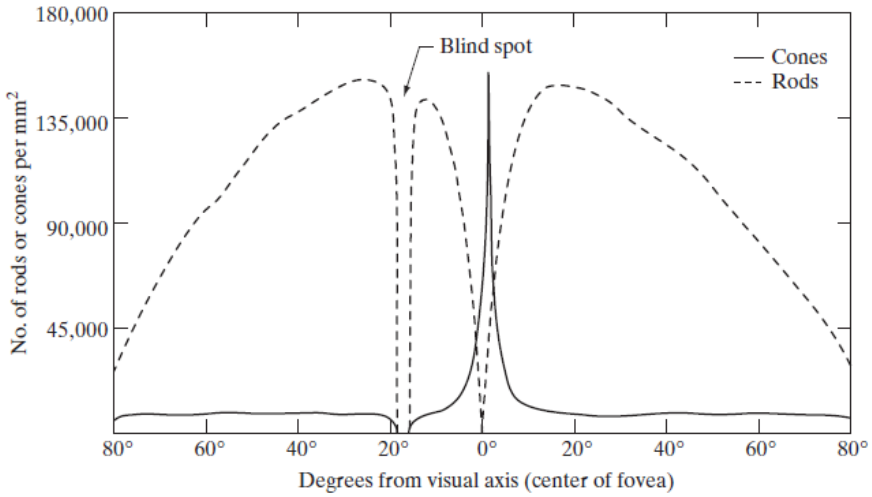


Figure 2.4: Distribution of rods and cones across the retina.

Cones and rods are both sensitive to different light intensities and wavelengths. Rods are activated by *scotopic* (dark) conditions and are most sensitive to light with a wavelength of about 500 nm [1, 24]. Cones, on the other hand, are activated by *photopic* (bright) conditions. They have one out of three different levels of pigment making them more sensitive to a certain wavelength range of light: the so-called S-cones to a range peaking at 430 nm (blue), the so-called M-cones to a range peaking at 540 nm (green), and the so-called L-cones to a range peaking at 570 nm (red), as depicted in Figure 2.5.

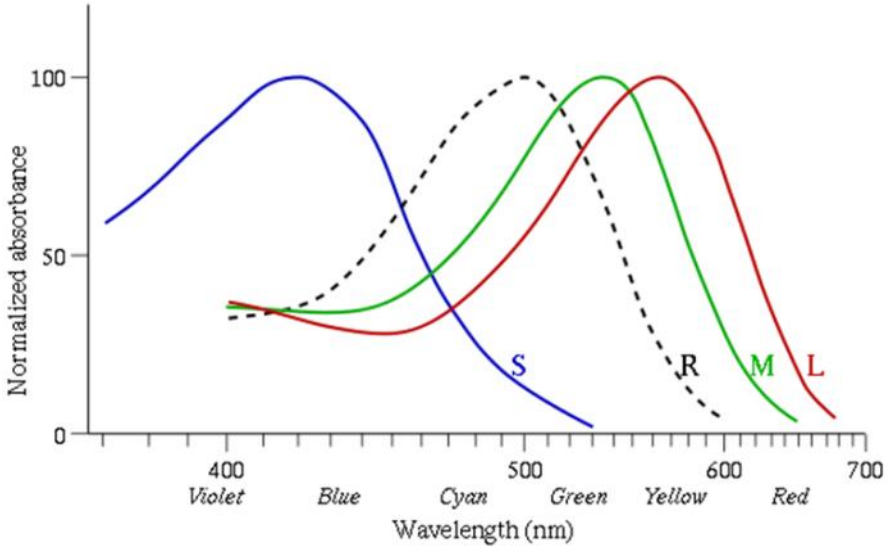


Figure 2.5: Normalized spectral sensitivities of the rods (R) and of the three different types of cones with a short (S), medium (M), or long (L) wavelength sensitivity.

The cones that are most sensitive to light with long (red) wavelengths cover about 65% of the total number of cones, followed by the medium (green) and short (blue) wavelength type cones, which cover about 30% and 5% respectively [22]. The relative neural signal strength of the cones encodes the colour information of the perceived light. Since all rods have the same spectral sensitivity, their impulses cannot encode colour information, but only light intensity information. That is why we perceive less colour in darker lighting conditions. Furthermore, rods have a much faster temporal response than cones, which makes us more sensitive to motion and flicker in peripheral vision than in foveal vision.

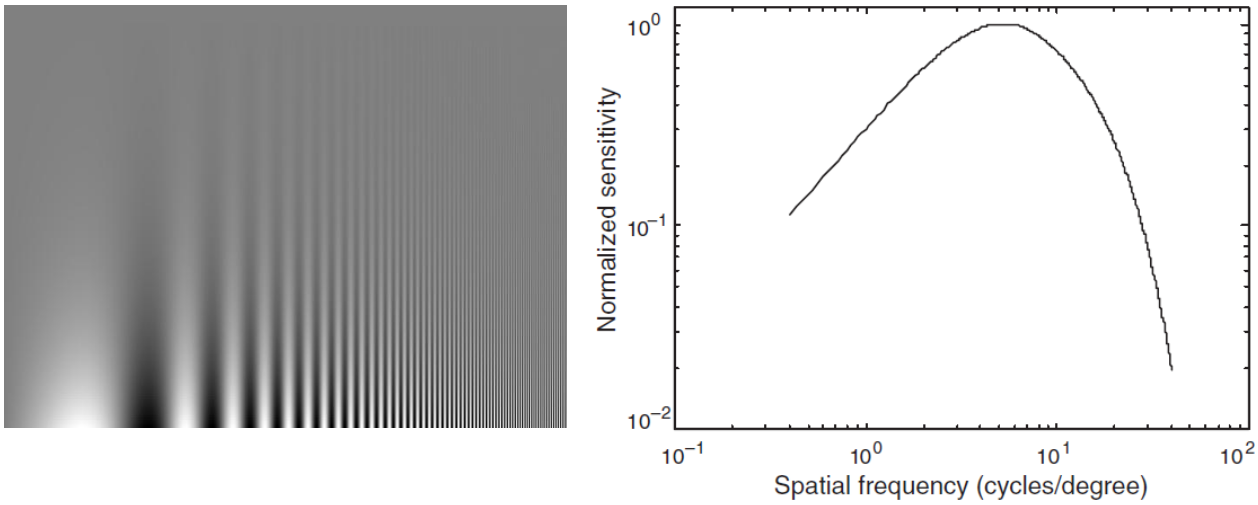
The neural impulses from the cones and the rods continue through a chain of other cells in the retina before they are collected by the *ganglion cells* in the top layer (see Figure 2.3). The distribution of these ganglion cells across the retina is comparable with that of the cones: highly concentrated at the fovea and rapidly decreasing in density as the distance from the fovea increases. The effect of this high spatial resolution at the fovea is that we can perceive great detail in the centre of our field of vision, while our peripheral vision is considerably more blurred, even though we are normally not consciously aware of it.

2.2. Perception

The previous section described how the incoming light is collected and converted into neural impulses. This following section will describe how these impulses influence our perception, especially the effects related to the perception of images. Knowledge about perception and the HVS is imperative in the field of image processing, since clever usage of certain limitations of the HVS can yield significant performance improvements in various image processing systems.

2.2.1. Contrast Sensitivity

The difference in visual properties that distinguish objects from each other, *contrast*, is an important aspect of perception. Contrast is determined by the colour and brightness of an object compared to another object or the background. Higher contrast increases our ability to distinguish objects and observe detail. Our sensitivity to contrast differentiates with the spatial frequency of the details. To demonstrate this, F. W. Campell and J. G. Robson [25] produced an artificial image (Figure 2.6a), which consists of a sinusoidal black-white pattern that increases in frequency along the horizontal axis, while it decreases in contrast along the vertical axis. Even though the decline in contrast is equal for all values on the horizontal axis, it appears as if the alternating black and white bars are highest around the middle of Figure 2.6a and lowest on its far left and right side. However, this observed effect is merely an illusion and not a property of the image. Under average lighting conditions, contrast sensitivity peaks at a spatial frequency of about 4 cycles per degree of visual angle and drops off rapidly at either side of the peak [25], as depicted in Figure 2.6b.



(a) (b)

Figure 2.6: The contrast sensitivity function. (a) The Campbell-Robson chart and (b) its resulting normalized contrast sensitivity as a function of the spatial frequency [1].

2.2.2. Texture Masking

The fact that the HVS is not sensitive to changes in contrast at a high frequency has important implications for our perception of images. One such effect is called *contrast masking* or *texture masking*. Masking in general refers to the situation where the presence of one image component obscures the presence of another. In this case, alternating contrast changes with a high spatial frequency act as the mask. As described above, we are not very

sensitive to such contrast variations. Therefore, areas in an image that contain such alternating patterns can mask distortion effects. A texture masking example is shown in Figure 2.7, where noise is added to a picture of a woman wearing a striped cloth. The noise is uniformly distributed across the whole image, yet it is more clearly visible in the smooth areas of the face than in the areas of the striped cloth.

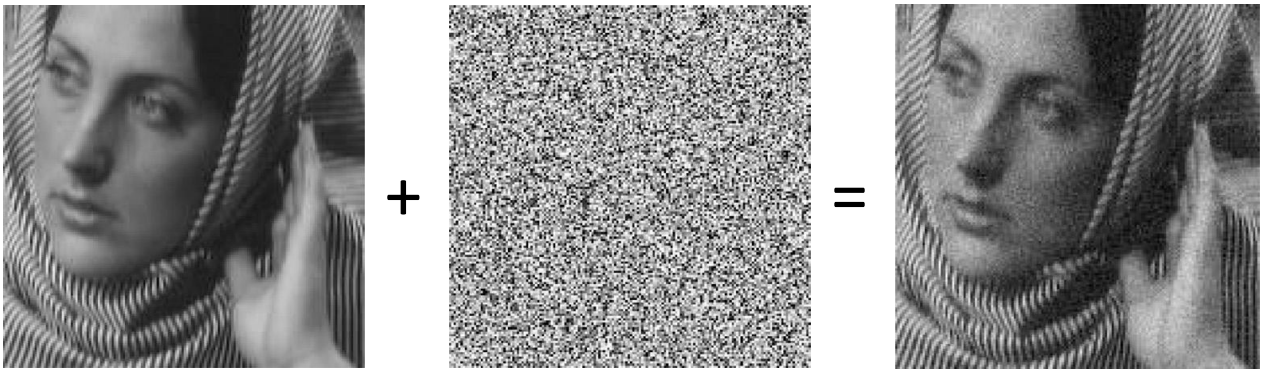


Figure 2.7: A texture masking example. The original image on the left is distorted with the uniformly distributed noise of the middle image. The resulting image on the right shows that the noise is being masked by the striped pattern and is less clearly visible in those areas than in smooth areas [1].

In the field of image compression, limitations of the HVS can be exploited to save bandwidth and decrease the number of bytes needed to store images. The fact that we do not perceive distortions as clearly in areas with a frequently alternating contrast can be used to the advantage of image compression. A clever algorithm can apply extra compression to high frequency areas and allocate the gained bits to low frequency areas. This means that the compression artefacts will be relocated towards regions of the image where we cannot perceive them as clearly.

2.2.3. Luminance Adaptation

Another important feature of the HVS is *luminance adaptation*. The HVS can process a wide range of light intensities, spanning from dark nights to bright days. However, it can only do so for a small band of the entire range at some moment in time, i.e. it cannot process both very low and very high light intensities simultaneously. The HVS copes with this problem by quickly adapting its dynamic range to the different lighting conditions. The amount of incoming light can be controlled by increasing or decreasing the diameter of the pupil. Moreover, the HVS adapts to luminance variations by adjusting the sensitivity of the retinal cells. *Dark adaptation,* when going from a well-lit environment to a poorly illuminated one, can take the HVS up to 15 minutes to fully adapt. *Light adaptation* on the other hand, when going from a dark environment to a bright one, only takes a few minutes.

As with the aforementioned limitations of the HVS, the fact that we are only sensitive to a narrow range of luminance intensities simultaneously can be exploited for the purpose of image compression. The HVS adapts its sensitivity to the average luminance of the field of vision. Hence, contrast between extreme intensity values that significantly deviate from the average intensity of an image will be as good as unperceivable. Therefore, compression artefacts can be relocated towards very bright or very dark regions to increase the overall perceptual image quality. This technique is often referred to as *light masking* or *luminance masking* [1, 26].

2.2.4. Foveal and Peripheral Vision

Fovea is the Latin word for *pit*. In our case, fovea refers to a small pit-shaped area in the centre of the retina with an extremely high density of photoreceptor cells, as described in the section 2.1.1. The high concentration of cells at the fovea allows us to perceive the light that falls onto it in great detail. However, because of its tiny size, the fovea only corresponds to about two degrees of our total field of view. That is about the size of a thumbnail held at arm's length. The fovea is merely 1% of the size of the retina, yet it takes up over 50% of the visual cortex to process its data [27]. Although we are normally not consciously aware of it, the remainder of our field of vision, *peripheral vision*, has a significantly lower spatial resolution. Sight becomes progressively blurred the farther away from the fovea. Figure 2.8 illustrates how an image would look if we were able to only fixate at one location on the image. The two images are almost indistinguishable when looking at the person in the lower part of the image from a proper viewing distance [1].

(a) (b)

Figure 2.8: An example of foveal and peripheral vision. (a) Original image. (b) Its foveated version as seen when focusing only on the man in the foreground [1].

2.3. Eye Movement

The previous section explained that we are only able to perceive a tiny portion of our field of view in high detail. This implies that the eye must be able to reposition itself, such that the light coming from any part of our surrounding is able to reach the fovea. Each eye is controlled by six extraocular muscles which almost purely rotates (and slightly translates) the eye in its socket [20]. These muscles allow for quick and accurate aiming of the eyes to almost anywhere in the field of view. Combined with head and body movements we are able to observe our entire environment. Our eyes are constantly moving, when we are just looking around and even when we fixate our gaze on a target [28, 29]. These motions can be roughly categorized into voluntary and involuntary movements, described separately in further detail below.

2.3.1. Voluntary

Voluntary eye movements are the ones we can consciously control in order to direct our vision towards regions or objects of interest. The process of voluntary movement mainly consist of rapid jumps from one location to another, so-called *saccades,* followed by a closer inspection of that location, a *fixation* [30]. During a saccade the mind selectively blocks visual processing in order to suppress the effects of motion blur. This process is known as *saccadic suppression*. Normally we are unaware of this gap in perception, but we can easily notice it ourselves by looking from one eye to the other in a mirror: it is impossible to see our own eyes in motion, while it is clear for an external observer [31].

There are, however, a few exceptions to saccadic suppression where we can move our eyes without temporal blindness. First of all, the *vestibulo-ocular reflex* stabilizes our vision by rotating the eyes in the opposite direction of the head. This allows for steady fixations, uninterruptable by concurrent head movements. It is possible to perceive this yourself by looking at your eyes in a mirror and rotating your head. Another type of eye movement is *smooth pursuit*, which is done by following a moving target with our eyes [32]. Closely related to it, is the *optokinetic reflex,* which is basically a repeated sequence of a short pursuit followed by a saccade back to the starting point. This is usually done when following targets with our eyes until they move out of our field of vision. For example, when the observer is in motion (e.g. looking at lantern posts at the side of the road while driving) or when the targets are in motion (e.g. looking at the wagons of a freight train passing by).

Another type of voluntary eye movement are the *vergence* movements [33], where both eyes move in opposite direction. These movements happen every time we focus our gaze on a different target, similar to the accommodation of the lens. Both eyes are rotated inward around their vertical axis to direct their vision to the same point. The direction of both eyes diverges to focus on distant objects and converges for near objects. When looking far into the infinite distance, the eyes diverge until they are nearly parallel.

2.3.2. Involuntary

Although we are normally unaware of it, our eyes continuously make small movements. Even when we fix our eyes on an stationary object, they still make *fixational eye movements* [28]. There are three main types of these movements, namely high-frequency *tremor*, slow *drifts*, and *microsaccades*:

- *Tremor* is a noisy random motion of the eyes with an amplitude of about the size of one cone and with a frequency of about 90 Hz [28]. It is difficult to measure these movements accurately and their contribution to vision is still unclear.
- *Drifts* are relatively slow movements that occur simultaneously with tremor and in-between microsaccades. They are much larger than tremor and can carry the object of fixation across several dozens of photoreceptors, as depicted in Figure 2.9.
- *Microsaccades* are fast movements that can carry the object of fixation across a wide range of several dozen, or even up to several hundred, of photoreceptors [28]. Microsaccades are believed to correct for drifts by moving in the opposite direction and bringing the fixational object back onto the centre of the fovea (Figure 2.9).

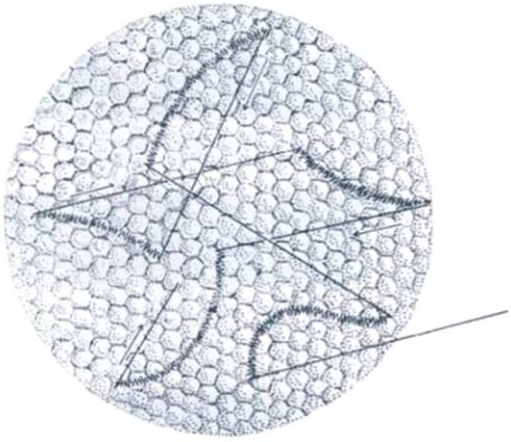


Figure 2.9: The path of fixational eye movements across the fovea. Slow drifts (including tremor) and microsaccades represented by the curved and the straight lines respectively.

The purpose of involuntary fixational eye movements has been the subject of much debate for many years and their exact role in vision is still uncertain. According to S. Martinez-Conde [28, 34] microsaccades serve to correct for the intersaccadic drifts and return the eye to the fixational target. It is also believed that fixational eye movements act to counter the effect of neural *adaptation* in the retina. Adaptation is the effect that all neurons *adapt* to their current impulse, such that steady impulses produce weak neural responses, whereas abrupt changes generate strong responses [28]. For vision this means that, due to adaptation, a stationary scene will simply fade from our sight after some time without any eye movement. This effect can be experienced in the periphery, because the perceptive fields there are considerably larger than near the fovea. Hence, fixational eye movements are not large enough to counteract fading there [28, 34], as demonstrated in Figure 2.10.

Figure 2.10: Demonstration of visual fading in the periphery, known as the *Troxler's effect*, first discovered by I.P.V. Troxler in 1804. After careful fixation on the dot in the centre for a few seconds, the annulus will vanish until further eye movement.

Under normal circumstances, we are unaware of visual fading and fixational eye movements, even though their magnitudes are substantial enough to be visible [28, 35]. This suggests that the brain compensates for such small eye movements and subtracts them from our actual awareness. According to I. Murakami [29, 35, 36] three sources of information are used to compensate for visual jitter and create a steady image: the motor command signals from the brain to move the eyes, proprioceptive feedback from the extraocular muscles, and retinal image motion. The last has been confirmed by a visual illusion that makes jitter visible through the after-effect of adaptation [29].

2.4. Visual Attention

The previous section was about how the eyes move, this section will focus on what guides the eyes to move to certain regions or objects of interest. Since we can only perceive a small portion of our entire field of view in detail, we constantly have to move our eyes across a scene in order to obtain any useful visual input. The scanning pattern of a scene is not random, but guided by certain properties of the scene and the goals or interests of the observer [37]. Eye movements can be measured easily with modern equipment, but attention is much more difficult to measure objectively and at present still little is known about its exact working. The general belief is that during each fixation the system, *visual attention*, rapidly detects potentially relevant parts of a scene as a target for the next fixation. Therefore, each saccade is preceded by a shift of attention to that same location [33]. This applies to all saccades, whether they are intentional or mere reflexes. However, this does not entail that this relation is also always true the other way around, i.e. that a shift of attention is always followed by a saccade. For it is possible to attend to a peripheral target without moving the eyes to its location [33].

Visual attention is controlled by certain basic features of a scene, such as: colour, orientation, motion, size, curvature, depth, and shape [38]. These simple visual characteristics can be quickly processed during a fixation in order to determine the location of the next fixation. Therefore, fixations are not spread randomly across a scene, but follow a distinct scan path based on visual features. People display highly replicable scan paths when they are shown the same image multiple times [33]. However, it is possible to alter the visual behaviour by giving observers different instructions, e.g. counting certain objects in an image. This would suggest that eye movements are not solely controlled by visual features in the image itself, but also by the governing intentions of the observer. According to J. Henderson et al. [39], fixations are clustered on both visually and semantically informative regions. Moreover, the eyes are initially not drawn to a region for its meaning, but they will remain in that region longer if it appears to be more semantically informative upon first encountering it. The decision on whether or not a region is semantically informative depends on the context and intentions of the observer. This means that eye movements are guided by both a covert or "hidden" attention system for quick processing as well as an overt system for intentional orientation [33].

J. Wolfe [38] found that it is impossible to guide attention to two different locations simultaneously. Therefore, in order to see an entire image, it has to be scanned one location at a time. To quickly scan an image, it is crucial to have an efficient scanning system that does not scan the same area twice in a row. A phenomenon called *inhibition of return* (IOR) does exactly that [40]. After having fixated to a specific target, the IOR mechanism remembers its

location and discourages attention from re-orientating back to that same target. This is especially helpful during visual search. IOR is both space and object oriented, so even when a previously attended object has moved to a new location, attention to that object will remain impaired [33].

Several promising attempts have been made to model the visual attention mechanism and develop an algorithm that can automatically identify the ROI in an image. Such algorithms often follow a bottom-up approach, i.e. they calculate individual properties of an image and combine them in order to determine the most important areas. For example, L. Itti et al. [41] use colour, intensity, and orientation features, while H.Y. Chen & J.J. Leou [42] also use object recognition to simulate visual attention. O. Meur et al. [30] utilize even higher-level features to model the HVS behaviour, such as the contrast sensitivity function, perceptual decomposition, visual masking, and centre-surround interaction. The model developed by U. Rajashekar et al. [43], entitled *GAFFE* (Gaze-Attentive Fixation Finding Engine), is able to predict the location of the first few fixations by using four low-level image features: luminance, contrast, luminance-bandpass, and contrast-bandpass. GAFFE is also used in [44] to improve the performance of IQA metrics. A different approach is taken by F.W.M. Stentiford [45], who uses evolutionary programming based on the dissimilarity between randomly generated pixel neighbourhoods. This method is used in [46] to separately encode the ROI and the background using JPEG2000 compression.

2.5. Image Quality Assessment

Image quality assessment (IQA) is a vital element in the development of image processing algorithms. Without careful assessment we would not be able to properly test the performance of such algorithms. Determining the image quality as perceived by humans is best done by letting people score various images. However, people are all different and have individual preferences. Therefore, to get an unbiased score, a great variety of people is needed to determine an accurate average score. This is unfortunately a very expensive and time-consuming procedure, thus not practical in most applications. To solve this problem, automatic IQA metrics have been developed that aim to predict the perceptual quality of an image without the need for any human judges.

2.5.1. Mean Squared Error

One of the earliest and most commonly used IQA metrics is the *Mean Squared Error* (MSE). It is simply defined as the squared difference between the pixel values of the distorted image and its reference:

$$MSE = \frac{1}{MN}\sum_{i=1}^{M}\sum_{j=1}^{N}(R(i,j) - D(i,j))^2$$

where D is the distorted image and R is its reference, both of M by N pixels. The resulting value indicates the difference between the original and distorted image, and thus is a measure for the degree of distortion. However, this is not necessarily the best measurement for image quality as perceived by humans, since the HVS does not operate similarly on the pixel domain [1]. For this reason the MSE is said to be inferior to other metrics when it comes to perceptual IQA [2]. To illustrate the shortcomings of the MSE, a photo of Albert Einstein is degraded

using six different types of distortion, shown in Figure 2.11. Notice that the six resulting images all have clearly distinguishable distortion artefacts, yet all six have a nearly identical MSE value.

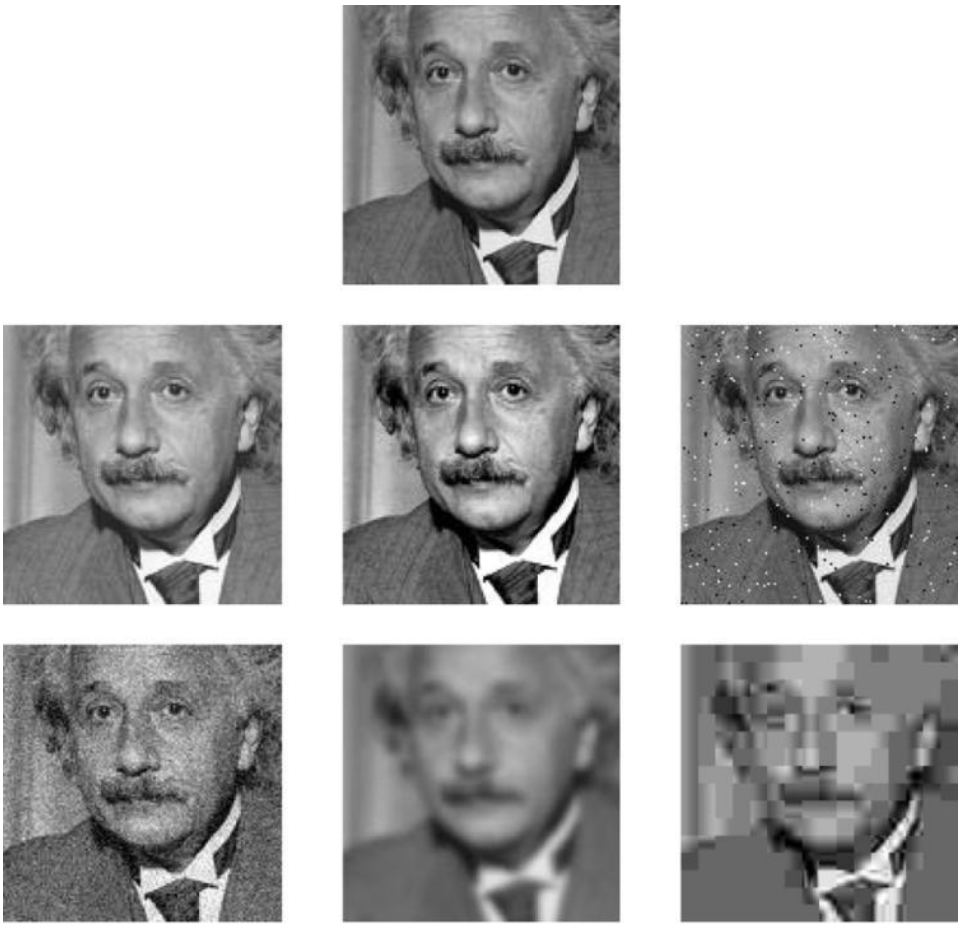


Figure 2.11: The original picture of Einstein (top) and six degraded version with different types of distortion: mean luminance shift, contrast stretch, impulsive noise, white Gaussian noise, blurring, and JPEG compression. All six images have approximately the same MSE value, yet an apparent different perceptual quality [1].

2.5.2. Full-Reference vs. Reduced-Reference vs. No-Reference Metrics

In the last few decades more advanced and complex IQA metrics have been developed that perform much better than the MSE. They can be categorized into three groups depending on what role the reference image plays in their design: No-Reference, Reduced-Reference, and Full-Reference metrics. They each are described separately below.

2.5.2.1. Full-Reference Metrics

The most commonly used metrics are the *Full-Reference* (FR) metrics. As their name suggests, they require the whole original image as a reference to calculate the quality of the distorted image. They use the difference between the distorted image and its perfect quality version to quantify the distortion. Because FR metrics need the whole reference image, they also consume the most bandwidth. In some applications this is not an important issue, yet in applications involving network transmissions it can make all the difference. For instance, in a network

application the distortion can be caused on purpose by compression to save bandwidth, or the transmission channel can be very noisy, so that the image will always get distorted to some extent.

An extensive evaluation of several state-of-the-art FR metrics can be found in [47]. The MSE, described in 2.5.1, and the closely related PSNR, are examples of two simple and well-known FR metrics. More advanced FR metrics have been developed that aim to exceed the quality prediction accuracy of the MSE and PSNR. One such metric is the *Picture Quality Scale* [48], which takes several properties of the HVS into account, such as texture masking, the CSF, and luminance sensitivity. Another advanced FR metric has been developed by Z. Wang et al. [3]. It measures the image quality by quantifying the structural similarity between the distorted and reference image. The structural similarity is defined as similarity in local luminance, contrast, and texture, further described in section 4.1.2. H.R. Sheikh et al. [5, 49] follow an information theoretic approach. They have designed a supposedly superior FR metric that quantifies the amount of information that is shared between the two images to evaluate the quality degradation. This metric is described in more detail in section 4.1.3. N. Damera-Venkata et al. [50] have developed a metric based on the assumption that the image quality degradation is caused by a combination of both linear frequency distortion and additive noise injection. Therefore, their algorithm consists of a separate measure for distortion and for noise to quantify the impact on the HVS.

2.5.2.2. No-Reference Metrics

In many applications the original image is simply not available as a reference. Therefore, developers have come up with metrics that use only the distorted image to calculate the quality, the so-called *Blind* or *No-Reference* (NF) metrics. Obviously, this is much more difficult than with a reference, since no information is available about the original perfect quality image. Instead of using the difference between the perfect and distorted image, the metric has to determine the quality from the distorted image alone. This is quite an easy task for a human being, because we possess prior knowledge about how objects look like in reality and how the image should look like if it were perfect quality, but it is much more difficult for a computer.

Since no information about the original image is available, NR metrics have to make assumptions about the reference of the distorted image; for instance, that the distorted image has been subject to a certain type of distortion. Therefore, NR metrics often specialize in the detection of artefacts caused by a specific distortion type. For example, blocking artefacts caused by block-based compression techniques, such as JPEG, can be objectively measured with the NR metrics developed by H.R. Wu & M. Yuen [6], R. Muijs & I. Kirenko [7], or H. Liu & I. Heynderickx [51]. Blur artefacts caused by wavelet-based compression techniques, such as JPEG2000, can be quantified by the metric of P. Marziliano et al. [8] or N.G. Sadaka et al. [18]. However, the performance of general-purpose NR metrics, which do not assume any specific distortion type and can be applied to any degraded image, is to date still insufficient to be of any practical use [1].

2.5.2.3. Reduced-Reference Metrics

Other metrics, known as *Reduced-Reference* (RR) metrics, are a compromise between FR and NR metrics. Instead of using the original image as reference, these metrics manage with merely a set of image features extracted from the reference. This feature set only costs little bandwidth and can be sent together with the image, or possibly on a

separate channel. Upon receiving the distorted image and the features of the original image, the RR metric can extract the same set of features from the distorted image. Afterwards, the metric compares the features of both images in order to quantify the distortion.

The metric developed by Z. Wang & E.P. Simoncelli [52] uses the Kullback-Leibler distance between the marginal probability distributions of wavelet coefficients of the reference and distorted images as a measure of image distortion. These distributions are approximated with a generalized Gaussian model to further reduce the required number of RR features. Different image features, namely the local harmonic amplitude information computed from an edge-detected image, are employed by the metric of I.P. Gunawan & M. Ghanbari [53]. By comparing the information from the reference with the information from a degraded image, the metric can identify different types of distortion, such as blockiness or blur. RR metrics are not only applicable to still images, but also to video data. For example, A.A. Webster et al. [54] have developed a RR metric that calculates several temporal features as well as spatial features to evaluate the quality of videos.

2.5.3. Bottom-Up vs. Top-Down Approach

IQA metrics usually follow one of two design approaches, either *top-down* or *bottom-up*. Metrics that follow the bottom-up approach aim to model each individual aspect of the HVS separately, and combine these aspects to yield a single quality index afterwards. Such HVS elements can include light/luminance masking, texture/contrast masking, and peripheral blur, which are described in section 2.2. The values from all these different elements have to be normalized before they can be combined in a meaningful manner. Afterwards, they can be pooled into a single value that represents the final perceptual quality of a given image.

The *Visible Difference Predictor* [55] models three important aspects of the HVS by using: (1) amplitude nonlinearity to model luminance adaptation, (2) a CSF filter to model the spatial frequency effects, and (3) a detection process that models the masking effects. A different approach is taken by J. Lubin & D. Fibush [56], who quantify the difference between a distorted image and a reference image in units of the modelled human *Just-Noticeable Difference*, the smallest difference that can be detected by a human observer. A similar approach is taken by the aforementioned FR metric of N. Damera-Venkata et al. [50]. However, bottom-up designed IQA metrics are usually rather computationally expensive and are generally not suited for real-time applications.

On the other hand, the top-down approach does not try to model the HVS, but takes a very different approach to the IQA problem. Instead, it focuses on the relationship between the input and the output and simply treats the whole HVS process as a black box [1]. The top-down designed metrics do not care how the HVS operates, as long as the input (the image) results in the correct output (the quality score). Such metrics are usually much simpler and efficient than bottom-up ones, while they are still able to achieve comparable results. The aforementioned FR metrics of Z. Wang et al. [3] and H.R. Sheikh et al. [5] are two examples of IQA metrics that use a top-down approach. H. Liu et al. [57] have implemented the black box design with the use of a neural network. A neural network can learn the relationship between the input (the image) and the output (the quality score), after which it can apply that knowledge to new images and calculate their quality score.

3. Effect of Task on Visual Attention

Image compression algorithms aim to reduce the number of bytes needed to store an image while keeping its perceptual quality as high as possible. The best way to measure this perceptual quality is to let a large number of people rate it to determine the m*ean opinion score* (MOS). However, this is an expensive and time-consuming process, which makes it unpractical for most applications. Therefore, *image quality assessment* (IQA) metrics have been developed that aim to quickly predict the quality of an image as perceived by a human observer, or in other words, to calculate the score of an image equal to its MOS, without the need for actual human observers. These IQA metrics are tested by comparing their calculated score with the MOS, so they are considered better the closer their calculated score is to the MOS. The current IQA metrics average their predicted score over the entire image, so they do not take into account that distortion in some areas of the image might be more annoying to a human observer than in other areas. To accommodate for this aspect of the HVS and possibly improve the IQA metrics, it is possible to include visual attention into the image quality prediction. According to Q. Ma et al. [11], visual attention or saliency can be used as a weighting function in existing IQA metrics in order to increase their performance. However, in their study they used a computational saliency model, which is not very accurate or representative for real human visual attention. Applying empirical saliency gathered from human observers could lead to even better results. A study that took this approach found that including empirical saliency improved the results of certain IQA metrics [9], yet another study found no clear improvement [12]. A possible explanation for the difference in these results might be the method used to acquire the saliency. In both studies participants were given a different task, and it is already known that task can have a profound influence on viewing behaviour [13, 14, 15, 16]. This raises the question, what type of saliency can best be used for improving IQA metrics: saliency based on the eye movements from observers that determined the MOS or from observers that were looking freely without a task? Does the task of scoring the quality of an image affect the way it is perceived? To answer these questions we have set up an experiment where we investigated the difference in visual attention between observers with a task and without a task, i.e. between *scoring* and *free looking*. The main research question of this experiment is as follows:

"What is the difference in visual attention between looking freely at an image and scoring the quality of an image?"

Our hypothesis is that there is indeed a difference between the two, namely, when people are scoring they will pay more attention to compression artefacts across the whole image, while when looking freely they will focus more on the semantically informative regions in the image, also known as the *regions of interest*. C.T. Vu et al. [15], found some evidence for this assumption, yet they were not able to make any firm conclusions, for they only used five test subjects in their experiment. In contrast, we used 60 test subjects for our experiment. In order to determine the influence of task on viewing behaviour, the experiment was divided into two separate sessions, each with their own task. In both sessions the test subjects were presented with stimuli, images in this case, while their eye movements were recorded with an eye tracker.

3.1. Experimental Setup

In the experiment we recorded the eye movements of the test subjects with the *iView X* eye tracker, produced by SensoMotoric Instruments GmbH [58]. It had a sampling rate of 50 Hz, a pupil tracking resolution of around 0.1°, a gaze position tracking accuracy between 0.5° and 1°, and an optimal operating distance (the distance between the test subject and the camera) of 0.4 to 0.8 meter. The eye tracker illuminated one eye of the observer with an infrared light and measured the position of the infrared reflection and the pupil in order to determine the gaze point. To get more accurate measurements the head of the test subjects was kept in place using a chinrest, which was positioned at a distance of 60 cm from the screen, as can be seen in Figure 3.1. The height of the chinrest and the chair were adjustable to suit the length of the test subject. The display screen was a 17-inch CRT monitor with a resolution of 1024x768. For eye tracking stability, brighter lighting conditions (70 lux) proved favourable over darker conditions, possibly caused by the decreased pupil diameter, which in turn decreased the likelihood of the pupil being partially covered by the eyelids. Test subjects were arbitrary recruited among the students or university employees. They mostly aged between 20 and 40 years, with no major prior experience in image quality assessment. They also did not see the stimuli before.

Figure 3.1: Experimental setup showing the participant (left) with his head on the chinrest. The experimenter (middle) can follow the experiment and control the eye tracker from the screen on the right. The eye tracker itself can be seen at the bottom-right of the participant's screen [59].

3.1.1. Stimuli

We expected the observers who had to score the images to look at the compression artefacts across the whole image, while people who were asked to look freely were expected to pay most attention to the region of interest. Therefore, we had chosen images with an apparent region of interest, usually in the foreground or centre of the image, and a non-empty background. We selected a set of 42 different coloured images, 40 from a National Geographic database and 2 from the LIVE database [60]. The images of the first database all had a resolution of 600x600 pixels and the two images of the LIVE database had a resolution of 640x512 pixels and 768x512 pixels respectively. Most images were natural photographs of a certain object, person, or animal. See Figure 3.2 for some examples.

To create compression artefacts all images were compressed using JPEG[1] compression. Each of the 42 images was compressed at four different compression levels, randomly selected between 10 and 100, which resulted in a total of $4 \cdot 42 = 168$ stimuli. The remainder of the display screen surrounding the stimuli was grey (R=50% G=50% B=50%) to keep the average contrast between the stimuli and the background as small as possible.

(a) (b) (c)

(d) (e) (f)

Figure 3.2: Some examples of the 42 images that were used to create the stimuli.

3.1.2. Sessions

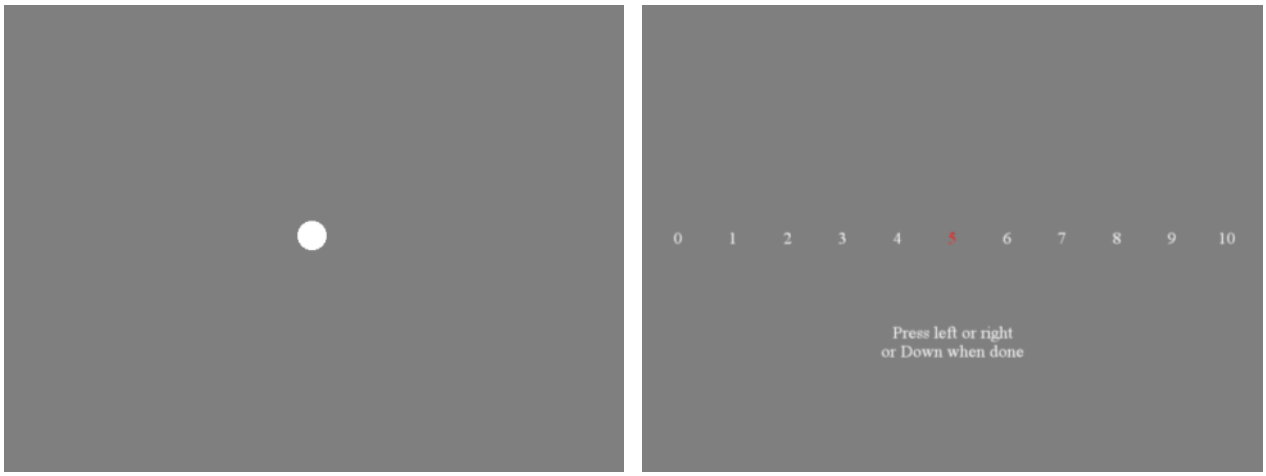
The experiment was divided in two sessions. In the first session the test subjects had no task; they were looking freely. The only instruction they had, was to look at the stimuli as if they would normally do, e.g. as through a photo album of a friend's holiday. In this session, every test subject only saw one of the four compression levels per original image, so 42 stimuli per test subject all with different image content. This was done because visual attention might be influenced by the number of times people look at a given image, i.e. one might look at the same image differently a second time, due to memories from having it seen the first time. For this first session we used 40 different participants who each saw 42 of the 168 stimuli. This resulted into $40 \cdot \frac{42}{168} = 10$ participants per stimulus. Each stimulus was shown for a fixed duration of 8 seconds.

[1] JPEG compression as specified by the `imwrite` function of MATLAB.

In the second session, the participants were instructed to score the quality of the stimuli. After having viewed a stimulus, the participants could press a key to continue to the scoring screen, where they could select a score from 0 to 10. As a result, the viewing duration of each stimulus was variable, instead of fixed like in the first session. For the second session we used 20 participants, all of whom were different from the participants of the first session. Consequently, the subjects of the second session had not seen the images before. Every participant was presented with all 168 stimuli. However, these 168 stimuli were spread across four sub-sessions. Each sub-session contained 42 stimuli, resulting from only one of the four compression levels per original image; thus similar as in the first session. In other words, the second session was essentially equal to four consecutive first sessions. As a result, we had 20 participants per stimulus for this session.

3.1.3. Protocol

To present the stimuli to the test subjects we used a software application called Presentation [61], which allowed us to easily display the stimuli in a randomized order. Before the experiment started, the participants received basic instructions about what was expected of them during the experiment. To familiarize the participants with the experimental procedure, they were trained by looking at example stimuli and practicing with scoring. Hereafter, the participants looked at 13 points across the screen in order to calibrate the eye tracker. To ensure there was no interdependency between the stimuli, we showed an empty screen with a white dot in the centre (Figure 3.3a) before every stimulus and instructed the test subjects to focus at this dot. This way all stimuli should have started with the same initial fixation to the centre. Furthermore, these fixations were later used to correct for possible drifts subsequent to the calibration, as will be explained in 3.2.1. In the scoring stage, the participants continued to the scoring screen (Figure 3.3b) after every stimulus, where they could easily select the score corresponding to the quality of the image on an 11-points scale.



(a) **(b)**

Figure 3.3: (a) A screenshot of the image shown in between two stimuli, and (b) a screenshot of the scoring scale on which participants could select their score

3.2. Data Processing

From the experiment we gathered the fixation locations and durations per stimulus for each test subject. These data can be visualised with circles, as shown in Figure 3.4a, where the size of each circle is determined by the duration of the fixation: the larger a circle the longer the duration of the fixation at that position. The green circles represent fixations onto the image, while the red circles represent the fixations onto the intermediate screen (Figure 3.3a) that was shown before the actual stimulus.

(a) (b)

Figure 3.4: (a) Original fixations for one example image, and (b) its corrected fixation locations.

3.2.1. Fixation Location Correction

From the visualization method of Figure 3.4 we noticed that in some cases the fixations were clustered right next to a presumed salient location. Additionally, the fixations to the intermediate screen deviated from the centre. This would indicate an offset in the measurements, possibly caused by head movements after the calibration. To compensate for this error, we used the locations of the fixations during the intermediate screen, and calculated the difference between the average of these locations and the centre of the screen. This difference, a constant vector, was then added to the location of all other fixations. The result of this correction can be seen in Figure 3.4b, where all fixations are translated a little to the lower right, such that the red circle is positioned in the centre of the image, indicated by the small white cross. Note that the clusters of fixations adjacent to the most salient regions, the face and ball, now lay on top of those regions, as one would expect.

It should be noted that this method is sensitive to whether the participant properly fixated at the dot during the intermediate screen. Unfortunately, this was not always the case: sometimes people did not neatly look at the dot and were just looking around the empty screen or were blinking. Using such incorrect fixations to correct the

fixations of the succeeding image would lead to an even worse result. Thus, instead of using only the fixations of the intermediate screen preceding a given image, the average of the fixations to all intermediate screens was used.

3.2.2. Saliency Maps

The visualization method of Figure 3.4 is well suited to represent the data of a single test subject, but for the combined data of all test subjects this method would result in a big blob of circles that does not entail any useful information. Therefore, another visualization technique was used: a height map, or landscape, where the height at a given coordinate indicates the total duration of the fixations of all test subjects to that coordinate. A fixation is, however, not limited to a single coordinate, but rather represent a small spot centred at that coordinate. This spot is not uniform, since detailed vision falls off rapidly with increasing distance from the centre of the fixation (see section 2.1.2). Therefore, a fixation is approximated with a bell-shaped curve using the following Gaussian function:

$$f(x) = t \cdot e^{-\frac{x^2}{2\sigma^2}}$$

where x is the distance from a given coordinate on the map to the centre of the fixation, t is the duration of the fixation, and σ is the width of the bell-shaped curve, which value is chosen to simulate foveal vision. In this way it is possible to create a saliency landscape map by adding a Gaussian function on top of the map for each fixation, with the centre of the Gaussian function at the fixation coordinate. After all fixations of all test subjects are added to the map, the intensity of the height map is normalized to a range from 0 to 1 by dividing all values by the highest value of the map. Figure 3.5a and Figure 3.5b show two examples of this method. In these examples, the height, or saliency, is represented with the use of colour, according to the scale shown in Figure 3.5c. These landscape maps clearly illustrate that the highest concentration of fixations is around the expected region of interest in the images: faces, people, animals, and other objects in the centre or foreground.

(a) (b) (c)

Figure 3.5: Example saliency maps of the combined fixations of all test subjects.

3.2.3. Region of Interest

The saliency maps described above can be used to objectively determine the region of interest (ROI) of an image. One way to do this is to consider all values in the saliency map above a certain threshold to belong to the ROI. Assuming a threshold of e.g. 0.2 would imply that for Figure 3.5 the ROI would be the area containing the green, yellow, and red colours. This area is then enhanced with a mathematical procedure called binary *opening*[2]. Simply explained this procedure entails that each area is first eroded by removing pixels from its outer boundary, after which the remaining areas are dilated by adding pixels to their outer boundary. The result of this operation is that small areas, such as noise, are removed by the erosion, yet the remaining areas recover their approximate original size by the subsequent dilation. The end result can be seen in Figure 3.6.

(a) (b)

Figure 3.6: The regions of interest (green) as determined from the saliency maps.

3.2.4. Saliency Metric

Before we can do a proper statistical analysis on the saliency data and test our hypothesis, we first have to define a metric that describes a whole saliency map in a single value. Since our hypothesis is based on whether test subjects look more to the ROI of a given image depending on the task, this value is in our case: the duration of the fixations inside the ROI divided by the total duration of all fixations. For example, if the saliency map of a given participant who looked at one stimulus gave a value of 0.7, this then would mean that this participant fixated 70% of the time to the ROI and 30% to the background of the image.

3.3. Results

The saliency metric described above, can be used to analyze the effect of different experimental factors, such as task, image content, and compression level. As explained above, the viewing time per stimulus was fixed at 8 seconds in the first session, yet the second session had variable viewing times, because the participants could continue to the scoring screen at any time. To make the two sessions comparable, the average viewing time of the second session was used as a limit to the viewing time of the first session. That is to say, the average viewing of the

[2] Binary opening as specified by the `imopen` function of MATLAB.

second session was about 5 seconds, and so for the first session only the data recorded during the first 5 seconds was used for further analysis, while the data of the remaining 3 seconds was discarded.

3.3.1. Task

The main effect that first had to be analyzed in order to answer the research question was the effect of task. Did the people who were scoring the image behave differently from the people who were looking freely, i.e. did task have a significant effect on visual attention? Our hypothesis stated that the percentage of fixations to the ROI would be higher without a task and lower for the scoring task, since we expected that people who were scoring would try to find compression artefacts, and hence would pay less attention to the ROI compared to people who were looking freely. An *analysis of variance* (ANOVA) was done with the help of the software program SPSS (version 16.0) [62] to statistically determine the effect of task. In this case a *between-subject* test was performed, since each test subject participated in only one of the two tasks and the difference between these two groups of subjects was tested. Therefore, only the main effect of task could be calculated; secondary effects, such as participant, image content, or compression level are analyzed in the next section for both tasks separately. The resulting ANOVA table (Figure 3.7) confirms our hypothesis: the between-subject test shows a significant different in the task ($p < 0.001$), which means that people who are asked to score the image quality indeed look differently to the ROI than people who look freely.

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Intercept	551.692	1	551.692	1562.19	0.001
Task	11.154	1	11.154	31.58	0.001
Error	17.658	58	0.353		

Figure 3.7: Test of the between-subject effect of task on the fixations to the ROI, performed with SPSS.

The significance value p alone does not tell us what the difference is, merely that it exists. Whether the percentage of fixations inside the ROI is higher for free looking or for scoring is another question. This question can be easily answered by plotting the values of our saliency metric per image for both tasks separately, as is done in the graph of **Error! Reference source not found.**. The plot shows that there is a lot of variation amongst the 42 different mages, but for almost all of them the percentage of fixations in the ROI is higher for the free looking task than for the scoring task. This means that people who are given the task to score the quality of an image tend to look significantly less to the ROI and more to the surrounding than people without a task. The only exception is the *girl_boat* image, shown in Figure 3.2f. Apparently, its exceptionally dark background masks the compression artefacts, compelling observers to pay more attention to the region of interest to assess the image quality.

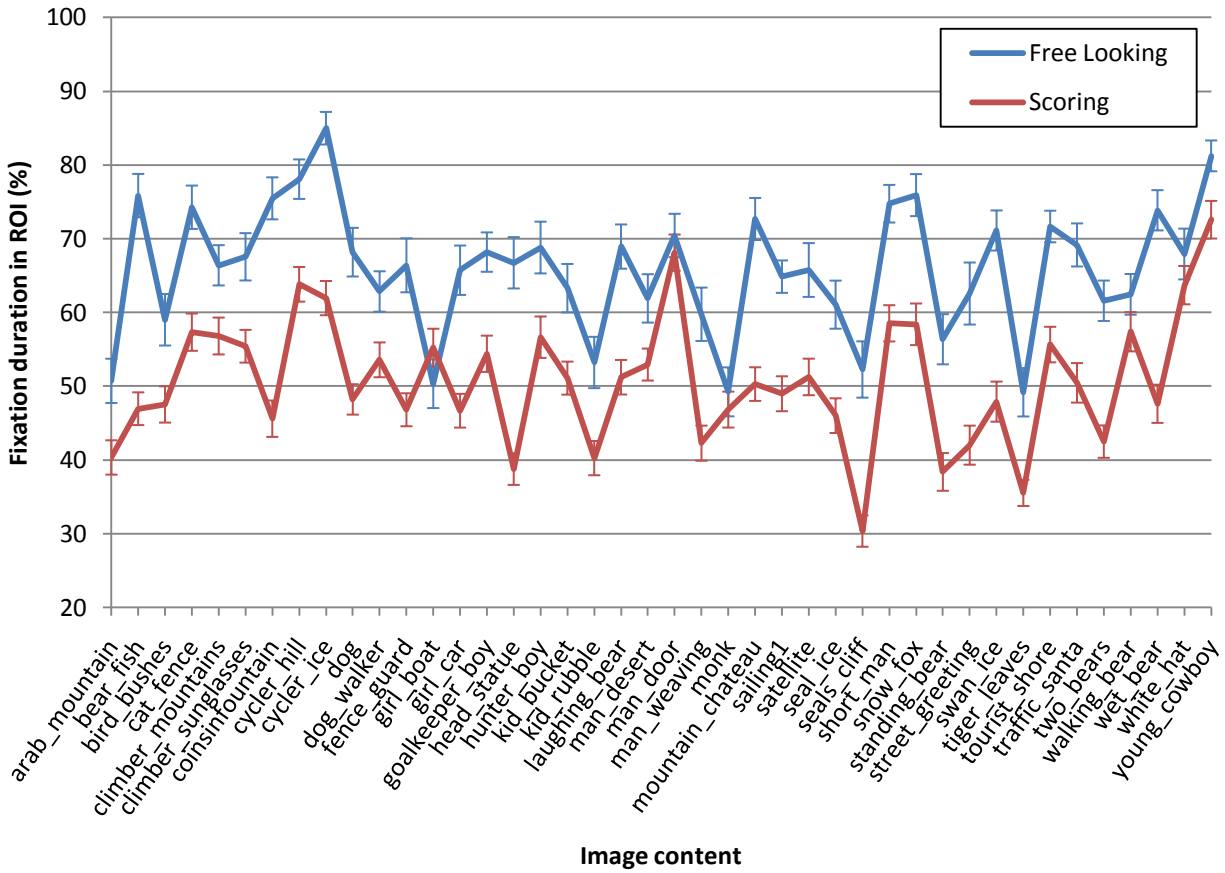


Figure 3.8: Percentage of fixations inside the region of interest per task for all 42 images. The error bars represent the standard error.

3.3.2. Other Effects

The data gathered from this experiment can also be used to investigate other effects on the percentage of fixations in the ROI. From Figure 3.7 it is already clear that the task, i.e. free looking or scoring, has a significant effect on the percentage of fixations in the ROI, so the additional effects are analyzed for each task separately. The ANOVA results are shown in Figure 3.9. As expected, participant and image content have a significant effect ($p < 0.001$) for both tasks. More interesting is the effect of the compression level: for the task of scoring this has a significant effect ($p < 0.001$), while for free looking it does not ($p = 0.28$). In the case of scoring there is a significant interaction effect ($p = 0.005$) between the image and its degree of compression. Furthermore, the four different groups, or sub-sessions, in the second session of the experiment do not have a significant effect ($p = 0.57$), which indicates that the number of times a participant has seen an image does not significantly influence the percentage of fixations to the ROI.

Task	Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Free Looking	**Intercept**	141.921	1	141.921	1646.198	.001
	Participant	5.254	39	.131	3.951	.001
	Image	2.884	41	.070	2.116	.001
	Compression level	.039	1	.039	1.170	.280
	Image * Compr. level	1.231	41	.030	.903	.646
	Participant * Compr. level	1.028	39	.026	.773	.846
Scoring	**Intercept**	161.678	1	161.678	558.781	.001
	Participant	9.487	19	.527	14.768	.001
	Image	8.135	41	.198	5.559	.001
	Compression level	.471	1	.471	13.199	.001
	Group	.072	3	.024	.674	.568
	Image * Compr. level	2.440	41	.060	1.667	.005
	Image * Group	4.185	123	.034	.953	.627
	Group * Compr. level	.081	3	.027	.757	.518
	Participant * Compr. level	.753	19	.042	1.173	.275

Figure 3.9: Tests of effects on the fixation to the region of interest for separate tasks.

3.3.3. Fixation Duration

In the previous sections we investigated spatial effects, but interesting observations can be made about temporal effects as well. The mean fixation duration in milliseconds for both the fixations inside and outside the region of interest are shown in the table of Figure 3.10 for both tasks separately. As can be seen from the table, the mean durations of the fixations outside the ROI are almost similar ($F = 0.53,\ df = 1,\ p = 0.47$) for both tasks, while they appear to be significantly different ($F = 80.92,\ df = 1,\ p < 0.001$) for the fixations inside the ROI. The fixations inside the ROI are much longer for the free looking task than for the scoring task, while the fixations outside the ROI last approximately just as long for both tasks. This suggests that not only the fixation location, but also the fixation duration shows a different behaviour depending on the task of the observer.

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

Figure 3.10: Mean duration (in ms) of the fixations inside and outside the region of interest per task.

3.3.4. Time

Analyzing visual behaviour over time revealed more interesting trends. Figure 3.11 shows the percentage of fixations in the ROI as a function of viewing time. It is noteworthy that for both tasks the percentage peaks after about half a second, after which it decreases and stabilizes, yet the percentage remains higher for the free looking task than for the scoring task.

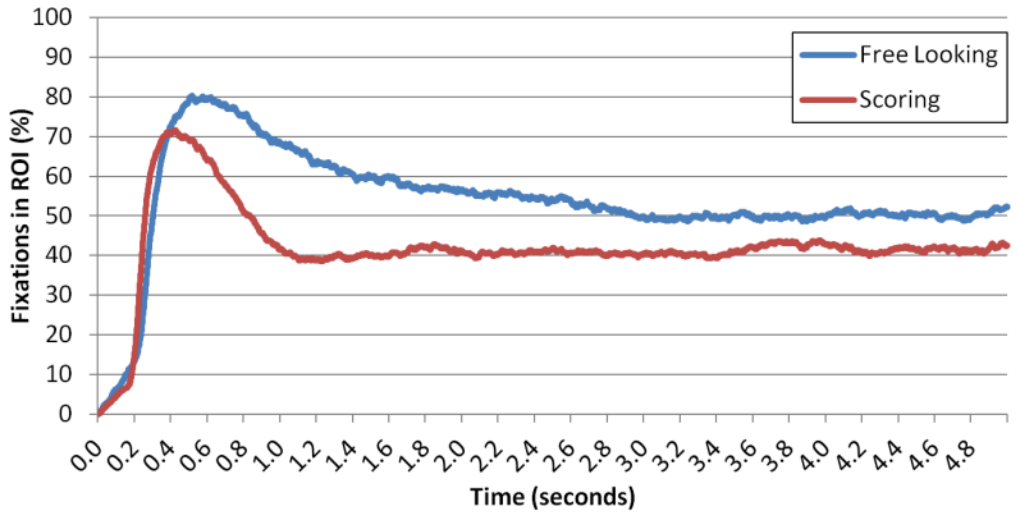


Figure 3.11: Fixations in region of interest plotted against the viewing time.

3.3.5. Scores

The subjective scores given by the test subjects in the second session of the experiment were first converted to standard scores, so-called *z-scores*, to compensate for the fact that every individual participant had a different quality judgment. Some participants scored all images worse than others, resulting in an overall lower mean. Additionally, some participants only used a small range of the scoring scale, resulting in an overall lower variation. Therefore, each score of every participant was first subtracted by his or her mean score and then divided by his or her standard deviation, according to the following formula:

$$z(s) = \frac{s - \mu}{\sigma}$$

where z is the standardized score of a given image, s is its original score, μ is the mean score of a given participant, and σ is the standard deviation of the scores of that participant. Each original score was converted from a scale of 0 to 100 into a z-score that denoted how many standard deviations it was away from the mean score. This resulted in a basically infinite range, so to make interpretation easier the scores were converted back into a range of 0 to 100, with the following formula:

$$\Phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz$$

where z is the aforementioned standardized score and Φ the resulting converted score. After the scores were converted, they could be compared to the actual quality of the images as defined through the JPEG compression level. As can be seen in Figure 3.12, there is no linear relation between the scores obtained in this experiment and the JPEG compression level. Differences in compression at a low quality level seem to affect the image quality scores much more than differences in compression at a high quality level. Between a compression level of 10 and 30 there are a lot of clearly visible compression artefacts, but above 30 these artefacts become harder to perceive by a human observer.

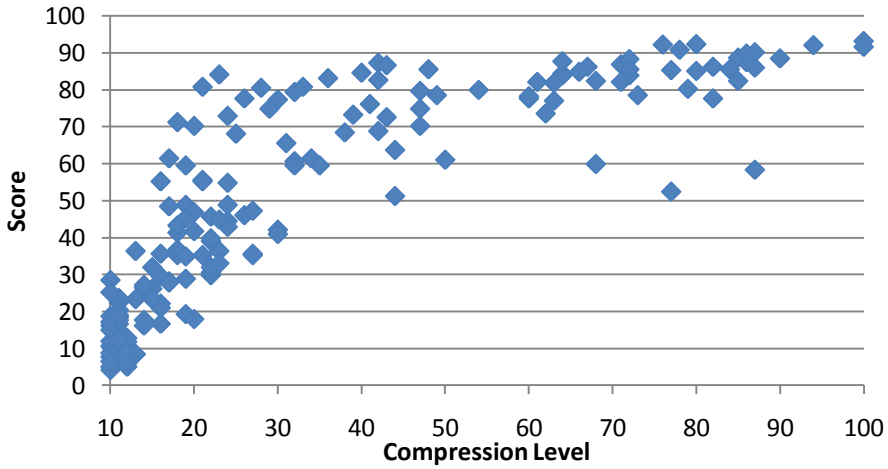


Figure 3.12: The quality scores of all stimuli plotted against their compression level.

The fact that the artefacts are harder to perceive by a human observer in images that have been compressed at a higher quality level can also cause the positive correlation ($\rho = 0.74$) between the quality score and the time needed by the test subjects to make their assessment. The plot in Figure 3.13 depicts this trend. Images that are considered to have a low quality (as indicated by the low quality score) are viewed for a much shorter time than images of a high perceptual quality.

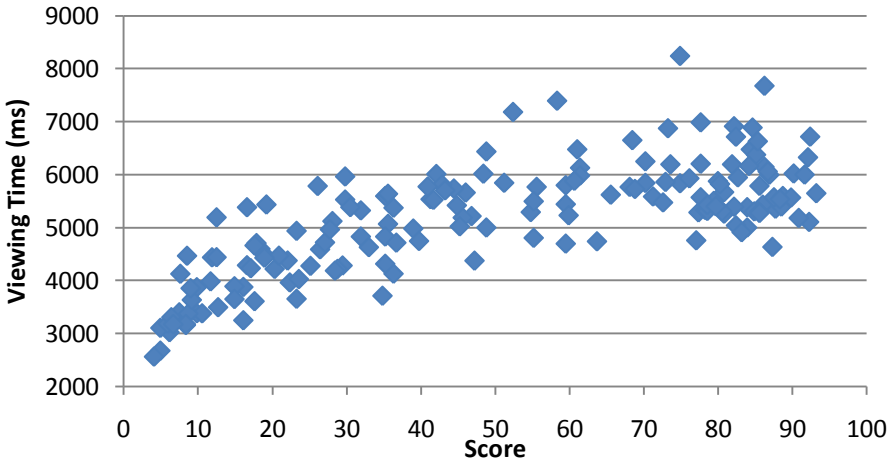


Figure 3.13: Viewing time (i.e. time needed by the test subjects to make the quality assessment) of all stimuli plotted against their quality score.

3.4. Discussion

In this section we have investigated the influence of a quality assessment task on visual attention. For this purpose, we conducted an experiment where one group of participants looked at a set of images freely, while the other group rated their quality. The eye movement data from the experiment were transformed into saliency maps, from which we could objectively derive the ROI. With the percentage of fixations in the ROI we could perform a statistical analysis to examine the differences between the two groups of participants. The choices and problems encountered during each step of this procedure are discussed separately below.

3.4.1. Experimental Procedure

- **Presence of the eye tracker**

  During the experiment the eye movements of all participants were recorded with a remote eye tracker positioned next to the screen. Even though the eye tracker does not physically interfere with the participants, the mere presence of an eye tracker might already influence their behaviour, since it could make them more conscious of their own actions. However, recording the eye movements without the participants knowing it is technically impossible (and morally arguable), so for the experiment we considered the eye tracker as the best option to determine saliency as reliably as possible.

- **Additional experimental session**

  The eye tracking experiment covered in this report consisted of two sessions: in the first session the test subjects looked freely to a set of images, and in the second session they were given a quality assessment task. An additional third session was performed separately by H. Alers et al. [59], in which the participants scored image quality, similarly as to our second session. However, the stimuli in this third session had a separate compression level for the ROI and for the background. This way it was possible to investigate if improving the quality of the ROI at the expense of the quality of the background would result in an overall higher quality score. They concluded that it is important to take the ROI into consideration during image processing, since the quality of the ROI greatly determines the perceived quality of the entire image, though the quality of the background cannot fully be neglected.

3.4.2. Post-Processing

- **Determining the ROI**

  After the experiment, the fixations were approximated to a bell-shaped curve and summed up to form a saliency map. There were several ways to objectively determine the ROI with the help of these saliency maps. One possibility was to define the ROI as a circle, with a fixed size of e.g. 25% of the image size, and with its centre at the point of highest saliency. Consequently, all stimuli would have a ROI with the same size and shape, which would be favourable for statistical consistency. However, the salient area in most images was hardly circular and some images even had multiple salient areas. We therefore chose to simply threshold the saliency maps, i.e. to consider the area above a saliency value of 0.2 as the ROI. The value of the threshold determines the size of the ROI (the higher the threshold the smaller the ROI). Deviations of about 0.1 from the used value only have a minor impact on the final results. A very different approach was taken by A. Santella &

D. DeCarlo [63], who determined the ROI with the use of clustering techniques. Their method might be a good alternative to find the ROI in the fixation data of a single test subject, but in our case it was almost impossible to find any sensible clusters in the combined data of all test subjects.

- **Misleading saliency maps**

  The saliency maps clearly showed the area in the image with the highest fixation density around the expected ROI (Figure 3.5). Although, this does not necessarily imply that people looked more to the ROI than to the rest of the image. For example, it is possible that they looked in equal amounts to the ROI and to the background, but that the fixations to the background were more spread out than the fixations to the ROI. In other words, the saliency maps merely depict the density, but not the total coverage. Therefore, we only used the saliency maps to objectively determine the ROI. To further statistically analyze the data, we came up with a new saliency measure, namely the ratio between the duration of the fixations inside the ROI and the total duration of all fixations (section 3.2.4).

3.4.3. Analysis

- **Influence of task**

  The main finding of our statistical analysis was that the task had a significant effect on visual attention, i.e. that scoring the quality of an image influenced the way it was observed, which is in agreement with the research of [15] and [16]. People tend to pay less attention to the ROI when they are asked to score than when they are looking freely, which confirmed our hypothesis.

- **Image content and compression level**

  The image content had a significant effect on the saliency metric for both tasks, yet the compression level only had a significant effect for the quality assessment task. This suggests that people tend to pay more attention to the compression artefacts when they are scoring than when they are looking freely. C. Vu et al. [15] also found that a none-uniformly distributed distortion, e.g. JPEG compression, influences the visual attention during quality assessment. Furthermore, we found that during quality assessment the image content had a significant interaction effect with the compression level. This means that the effect of the compression level on the saliency metric is not the same for all images. This interaction can possibly be explained by masking effects of the HVS, such as contrast, texture, or luminance masking (as described in section 2.2.1 through 2.2.3), which are more pronounced in some images compared to others. In other words, JPEG compression artefacts can be masked by rough textures or intensity extremes in some image content.

- **Fixation duration**

  The difference between the tasks was not only spatial, but temporal as well: when scoring, the attention strayed from the ROI earlier than when looking freely. Furthermore, the mean duration of the fixations inside the ROI was longer for the free looking task than for the scoring task, while the mean duration of the fixations outside the ROI did not differ. A possible explanation for this different behaviour might be that when scoring, people only briefly look at the ROI to recognize the image and then quickly divert their attention to the task

they are given, namely searching for artefacts across the whole image. Our findings are in contradiction to the results of A. Ninassi et al. [16], who found that the fixation duration was longer for scoring than for free looking. However, in their case original undistorted images without a clear ROI were used and they did not separately investigate the fixations inside and outside the ROI.

- **Variable viewing time**

  The time needed by the participants to score a given stimulus shows a remarkable relation with the perceptual image quality, namely it is shorter for low quality images and becomes longer for higher quality images (Figure 3.13). The task of scoring could be responsible for this positive correlation: the participants are searching for compression artefacts, which are more apparent and more quickly detected in lower quality images than in higher quality ones. Hence, they will be able to conclude their score faster for images with a low quality.

- **Memory**

  The participants in the second session of the experiment saw each image four times, every time with a different compression level. We expected that the number of times a participant saw a given image would influence the visual attention in some way, since people who saw an image for the second time would remember aspects of it from the first time. However, we did not find any evidence for this hypothesis during our analysis. This finding is in agreement to a study reported in [16], in which the authors also did not find any visual adaptation or task learning throughout their experiment. Unfortunately, we were unable to compare this effect between the two groups of participants and to investigate its task dependency, because in our experiment the participants without a task were only shown each image once.

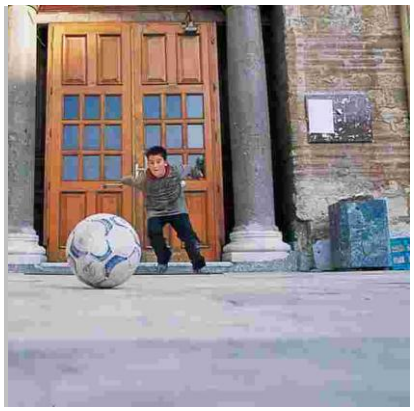
In the next section, we examine the influence of visual attention on existing IQA metrics, in particular the difference between both tasks. The current metrics do not take visual attention into account. Therefore, the saliency maps described in this section are used in the next section to investigate if they can be used to improve the performance of several popular IQA metrics.

4. Applying Visual Attention to IQA Metrics

The current image quality assessment (IQA) metrics operate on an image as a whole, i.e. they do not take into account that some areas in the image might be more salient than other areas. Distortion located in a salient area is likely to be considered as more annoying by an observer, thus should receive a lower perceptual quality score. Applying saliency to IQA metrics has already proved to be a promising research topic by several other studies [9, 11, 44, 64], yet they did not investigate the effect of applying saliency from different tasks. In the previous section we have seen that the task has a profound influence on the visual attention of observers, but how does this difference translate into IQA metrics? To investigate this we have used the eye movement data gathered in our previous experiment to answer the following research question:

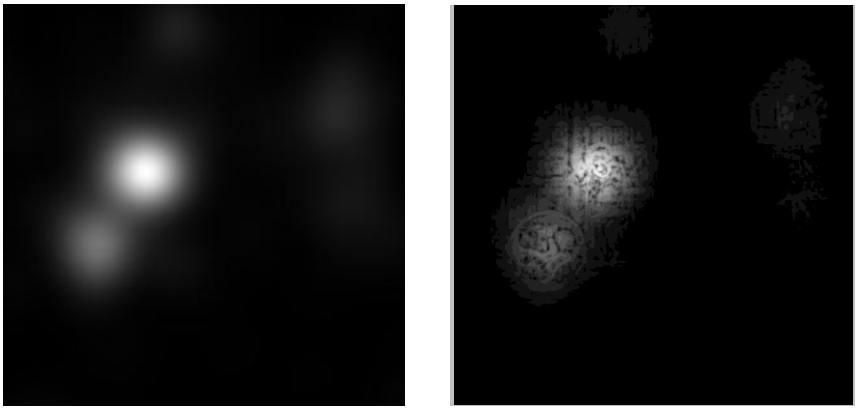
"Can the performance of existing image quality assessment metrics be improved by using visual attention information?"

The people at the Laboratory for Image and Video Engineering (LIVE) have also performed an extensive experiment, in which they obtained the free looking saliency of 29 original images [47]. These images were then processed with different distortion types, e.g. JPEG and JPEG2000 and the resulting stimuli were scored on quality. The data set and the subjective quality scores are freely available for download at [17]. We compared the findings on including saliency in IQA metrics for their and our database to ensure the conclusions were consistent across the two databases and not simply the result of poorly selected stimuli. To get a good overview of the effect of incorporating saliency, some well-known metrics, both full-reference and no-reference, were tested. These metrics are described in further detail in the next section. The general approach for adding saliency to an IQA metric is to weight the saliency map with a certain distortion map generated by the metric and then average the result. Figure 4.1 illustrates this procedure for the SSIM metric: Figure 4.1b is the distortion map that represents the distortion of the stimulus in Figure 4.1a. This map is multiplied with the saliency map of Figure 4.1c, resulting in the weighted map of Figure 4.1d. In the next section a brief description will be given of the metrics that were tested and how they were modified to incorporate saliency.



(a)

(b)



(c) (d)

Figure 4.1: An example of an SSIM map weighted with saliency: (a) the JPEG compressed image, (b) the SSIM map, (c) the saliency map, and (d) the SSIM map weighted with the saliency map.

4.1. Full-Reference Metrics

Full-reference (FR) IQA metrics cannot rate the quality of a distorted image without the undistorted image as a reference, which is assumed to be of a perfect quality. A FR metric can quantify the distortion by calculating the difference between the reference and the distorted image. This is in contrast to No-Reference (NR) metrics that can function even when no reference image is available. The NR metrics will be described in the section 4.2, whereas the following popular FR metrics are described here: the PSNR, the SSIM index, and the VIF criterion.

4.1.1. Peak Signal-to-Noise Ratio

The *Peak Signal-to-Noise Ratio* (PSNR) is a commonly used measure for the quality difference between a distorted image and its reference. It is most easily calculated via the *Mean Squared Error* (MSE), which is defined as the average of the squared difference between the distorted and reference image:

$$MSE = \frac{1}{MN}\sum_{i=1}^{M}\sum_{j=1}^{N}(R(i,j) - D(i,j))^2$$

where D is the distorted image and R is the original reference of M by N pixels. The PSNR can now be calculated with the following formula:

$$PSNR = 10\,log_{10}\left(\frac{MAX^2}{MSE}\right)$$

where MAX is the maximum possible pixel value in the image. In our case the pixel intensity values are all in the interval [0, 1], so $MAX = 1$. To incorporate the saliency into this metric the above formulas have to be modified such that the difference of each pixel value is weighted with the salience value:

$$WMSE = \frac{1}{\sum_{i=1}^{M}\sum_{j=1}^{N}S(i,j)}\sum_{i=1}^{M}\sum_{j=1}^{N}(R(i,j) - D(i,j))^2 \cdot M_S(i,j)$$

where M_S is the saliency map corresponding to the image. The final weighted PSNR is then simply calculated similar to the normal PSNR:

$$WPSNR = 10\ log_{10}\left(\frac{MAX^2}{WMSE}\right)$$

4.1.2. Structural Similarity Index

The *Structural Similarity* (SSIM) index, developed by Z. Wang et al. [3], is a more complex metric than the PSNR and is supposed to yield better results. Over the years it has become a widely accepted IQA metric that is used in diverse areas of image processing. The SSIM compares the reference image with the distorted image in three components: luminance, contrast, and structure:

$$\begin{aligned} SSIM(R,D) &= luminace\ similarity \cdot contrast\ similarity \cdot structural\ similarity \\ &= \left(\frac{2\mu_R\mu_D + C_1}{\mu_R^2 + \mu_D^2 + C_1}\right) \cdot \left(\frac{2\sigma_R\sigma_D + C_2}{\sigma_R^2 + \sigma_D^2 + C_2}\right) \cdot \left(\frac{Cov(R,D) + C_3}{\sigma_R\sigma_D + C_3}\right) \end{aligned}$$

where R is the reference image with a mean intensity μ_R and a standard deviation in intensity σ_R, D is the distorted image with a mean intensity μ_D and a standard deviation in intensity of σ_D, $Cov(R,D)$ is the covariance of R and D, and C_1 to C_3 are small constants that are included to avoid unstable results when the denominators are very close to zero. The algorithm calculates a similarity value at each pixel of in the image using an 11×11 circularly symmetric Gaussian kernel with a standard deviation of 1.5 samples. This produces a so-called SSIM map, which depicts the local distortion between the original and distorted image at every pixel. The SSIM map is averaged into a single value, sometimes referred to as the Mean SSIM (MSSIM), which represents the overall quality of the image:

$$MSSIM(R,D) = \frac{1}{MN}\sum_{i=i}^{M}\sum_{j=1}^{N} M_{RD}(i,j)$$

where M_{RD} is the SSIM map that represents the distortion between images R and D as calculated by the metric. An implementation of the SSIM metric for MATLAB is available for download at the LIVE homepage [65]. To include visual attention in this metric, we multiply the SSIM map M_{RD} with our saliency map M_S, after which it is averaged by dividing it with the sum of the saliency map:

$$WSSIM = \frac{1}{\sum_{i=1}^{M}\sum_{j=1}^{N} M_S(i,j)}\sum_{i=i}^{M}\sum_{j=1}^{N} M_{RD}(i,j) \cdot M_S(i,j)$$

4.1.3. Visual Information Fidelity Criterion

The *Visual Information Fidelity* (VIF) criterion [5] is argued to be even more advanced than the SSIM. An implementation of the algorithm is available in MATLAB for both the pixel domain version [66] and a supposedly superior wavelet domain version [67]. However, the pixel domain version is more suitable for our purpose, because our saliency maps are also in the pixel domain, and therefore, are more easily incorporated into the metric.

Instead of modelling the HVS and trying to predict the visual quality by comparing a distorted image against a reference image, the VIF aims to quantify the Shannon information that is shared between the two images. The VIF is simply defined as the fraction of information from the reference image that can ideally be extracted by the brain from the distorted image:

$$VIF = \frac{Distorted\ Image\ Information}{Reference\ Image\ Information}$$

In the metric proposal [5] the wavelet domain version of the VIF is described, which defines the image information of the two images with the following formula:

$$Distorted\ Image\ Information = \frac{1}{2}\sum_{i=1}^{M}\sum_{j=1}^{N} log_2\left(1 + \frac{g^2 s_i^2 \lambda_j}{\sigma_v^2 + \sigma_n^2}\right)$$

$$Reference\ Image\ Information = \frac{1}{2}\sum_{i=1}^{M}\sum_{j=1}^{N} log_2\left(1 + \frac{s_i^2 \lambda_j}{\sigma_n^2}\right)$$

where σ_n^2 is a constant HVS model parameter for the variance of the internal neuron noise (which we left at the default value of 2), and:

$$\begin{aligned} s_i^2 \lambda_j &= \widehat{\mathrm{Cov}}(R,R) \\ g &= \widehat{Cov}(R,D) \cdot \widehat{Cov}(R,R)^{-1} \\ \sigma_v^2 &= \widehat{Cov}(D,D) - \widehat{Cov}(R,D) \cdot \widehat{Cov}(R,R)^{-1} \cdot \widehat{Cov}(R,D) \end{aligned}$$

where the covariances are approximated by sample estimates using sample points from the corresponding blocks in the distorted image D and reference image R. To incorporate saliency into the metric, the three covariance matrixes $\widehat{Cov}(D,D)$, $\widehat{Cov}(R,R)$, and $\widehat{Cov}(R,D)$ are simply multiplied with our saliency maps and then reinserted into the remainder of the algorithm.

4.2. No-Reference Metrics

No-reference (NR) metrics are able to predict the quality of images when no perfect quality image is available as a reference, as already described in 2.5.2. Instead of comparing the distorted image with a reference, they try to detect artefacts in the distorted image alone. A significant artefact in compressed images is blockiness. Two of the NR metrics described below predict the perceived quality as a consequence of the occurrence of blockiness. The

third metric measures blur, which is an artefacts that is often found in JPEG2000 compressed images. Therefore, we are only unable to test this metric with the data of our experiment, but can only evaluate it with the JPEG2000 compressed images of the LIVE database.

4.2.1. Generalized Block-edge Impairment Metric

The *Generalized Block-edge Impairment Metric* (GBIM), developed H.R. Wu and M. Yuen [6], is a NR metric that is specialized in the detection of blocking artefacts in image and video data. It measures the blockiness separately in horizontal and vertical direction, after which the two directions are combined into a single quality value. The GBIM assumes that the artefacts occur on a grid of blocks of 8×8 pixels, which is common for most compression standards. The horizontal and vertical part of the metric are essentially the same, so it is sufficient to describe only the horizontal part. Given an image f with width N_c consisting of $\{f_{c1}, f_{c2} \dots f_{cN_c}\}$ columns, then the interpixel differences between each of the horizontal block boundaries in the image are defined as:

$$D_c f = \begin{bmatrix} f_{c8} - f_{c9} \\ f_{c16} - f_{c17} \\ \vdots \\ f_{c(N_c-8)} - f_{c(N_c-7)} \end{bmatrix}$$

Now the metric for the horizontal blockiness can be defined as:

$$\begin{aligned} M_h &= \|W D_c f\| \\ &= \left[\sum_{k=1}^{N_c/8-1} \left\| w_k [f_{c(8\times k)} - f_{c(8\times k+1)}] \right\|^2 \right]^{1/2} \end{aligned}$$

where $\|\dots\|$ is the l_2-norm and $W = \{w_1, w_2 \dots w_k\}$ is a diagonal weighting matrix which takes into account the local spatial characteristics of the image. The horizontal map M_h is multiplied with our saliency map M_S to incorporate visual attention. The result is then normalized by the average interpixel difference E_h between pixels which are not at block boundaries:

$$M_{h,NORM} = \frac{M_h \cdot M_S}{E_h}$$

This process is repeated in the vertical direction, resulting in a vertical map M_v, which is also multiplied with our saliency map M_S, and then normalized by the average interpixel difference E_v. Finally, the horizontal and vertical components are combined into a single quality value:

$$GBIM = \frac{M_{h,NORM} + M_{v,NORM}}{2}$$

4.2.2. Philips Blockiness Metric

The blocking artefact metric, developed by R. Muijs & I. Kirenko [7], does not assume a fixed block size of 8×8 pixels like the aforementioned GBIM, but is able to detect the position of the block grid. As with the GBIM, it processes the horizontal and vertical direction separately and combines them in a single quality value afterwards. Given an image I of $M \times N$ pixels, the metric determines the visibility of blockiness at every pixel $I(i,j)$ using the local gradient. This horizontal local gradient is normalized using the average gradient calculated over the K adjacent pixels to the left and right:

$$M_{h,NORM}(i,j) = \frac{|I(i+1,j) - I(i,j)|}{\frac{1}{2K}\sum_{k=-K\ldots K, k\neq 0}|I(i+k+1,j) - I(i+k,j)|}$$

The resulting horizontal gradient map is multiplied with our saliency map M_S and then summed over all N lines in the image:

$$D_h(i) = \sum_{j=1}^{N} M_{h,NORM}(i,j) \cdot M_S(i,j)$$

The blocking strength BS_h of the horizontal blocking artefacts can be determined by comparing the average of D_h at the block edges $\overline{D}_h(block)$ with the average at the intermediate positions $\overline{D}_h(non-block)$:

$$BS_h = \frac{\overline{D}_h(block)}{\overline{D}_h(non-block)}$$

The same process is repeated in the vertical direction and both components together make up the final blocking artefact quality score:

$$Blockiness = \frac{BS_h + BS_v}{2}$$

4.2.3. Blur Metric for JPEG2000

P. Marziliano et al. [8] devised an IQA metric that measures the degree of blur in JPEG2000 compressed images. The database of our experiment only contained JPEG compressed images, hence we only tested this metric on the JPEG2000 compressed images of the LIVE database. The metric first uses an edge detection algorithm[3] to find all strong vertical edges $e_1, e_2 \ldots e_N$ in a given image. The width of each edge $w_1, w_2 \ldots w_N$ is calculated as a measure for the local blur. Finally, the sum of all edge widths is average by the total number of edges N:

[3] Edge detection as specified by the `edge` function of MATLAB.

$$Blur = \frac{1}{N}\sum_{k=1}^{N} w_k$$

To incorporate saliency into this metric, we multiply the width of a given edge with the saliency at the location of that edge. The sum of these weighted edge widths is then averaged by the total saliency at the location of all edges:

$$Blur = \frac{1}{\sum_k^N M_S(k)}\sum_{k=1}^{N} w_k \cdot M_S(k)$$

where $M_S(k)$ is the saliency at the location of edge e_k that has a width of w_k.

4.3. Results

The modified IQA metrics described above were compared to their original version to test if integrating saliency in the metrics improved their quality prediction. The performance of the metrics was tested by calculating the correlation between their predicted score and the subjective score given by human observers, the MOS. The LIVE data does not use the absolute value of the MOS, but the so-called *Differential Mean Opinion Score* (DMOS). The DMOS is simply the difference between the score of a distorted image and the score of the original reference image. This is especially useful when the quality of the reference is not considered to be close to perfect and already receives a relatively low score.

The performance of the metrics was measured by means of both the linear *Pearson's product-moment Correlation Coefficient* (PCC) and the *Spearman's Rank Order Correlation Coefficient* (SROCC), which are measures for the prediction accuracy and monotonicity respectively. The Pearson's correlation r between two variables is defined as the covariance of the two variables divided by the product of their standard deviations:

$$r_{X,Y} = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}$$

where σ_X and σ_Y are the standard deviations of X and Y respectively, and $Cov(X,Y)$ is their covariance. In MATLAB this can easily be achieved with the `corrcoef` function. The Spearman's correlation ρ is a little more complex:

$$\rho_{X,Y} = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)}$$

where d_i is the difference between the ranks of the corresponding values X_i and Y_i and n is the total number of values in the data set. First, the metrics were tested on our own empirically gathered data. Secondly, they were

tested on the data of the LIVE database [17]. By doing so we can investigate whether the results are consistent across different databases, which would make our conclusions much stronger.

4.3.1. Own Data

The data gathered from our own experiment consist of the saliency maps of 42 original images and of the MOS of all stimuli generated by JPEG compression of the original images at four different compression levels. As described in 3.1.2, the saliency information was acquired in two sessions: one session for the free looking task and one for the scoring task. This resulted in two saliency maps per image, one for each session. These two saliency maps are integrated into to the aforementioned metrics separately in order to investigate the influence of the saliency map on the metrics' performance. First we examine the three FR metrics, followed by two of the three NR metrics. The third NR metric measures blur artefacts in JPEG2000 compressed images, while our own database only contained JPEG compressed images. Hence, that metric was tested with the LIVE database only, as described in the next section. The Pearson's Correlation Coefficient (PCC) and Spearman's Rank Order Correlation Coefficient (SROCC) between the score predicted by the FR metrics and the MOS are shown in the table and corresponding bar graph of Figure 4.2.

		Original	Scoring	Freelooking
PSNR	PCC	0.502	0.536	0.560
	SROCC	0.495	0.526	0.551
SSIM	PCC	0.739	0.768	0.788
	SROCC	0.753	0.788	0.807
VIF	PCC	0.791	0.802	0.806
	SROCC	0.826	0.841	0.844

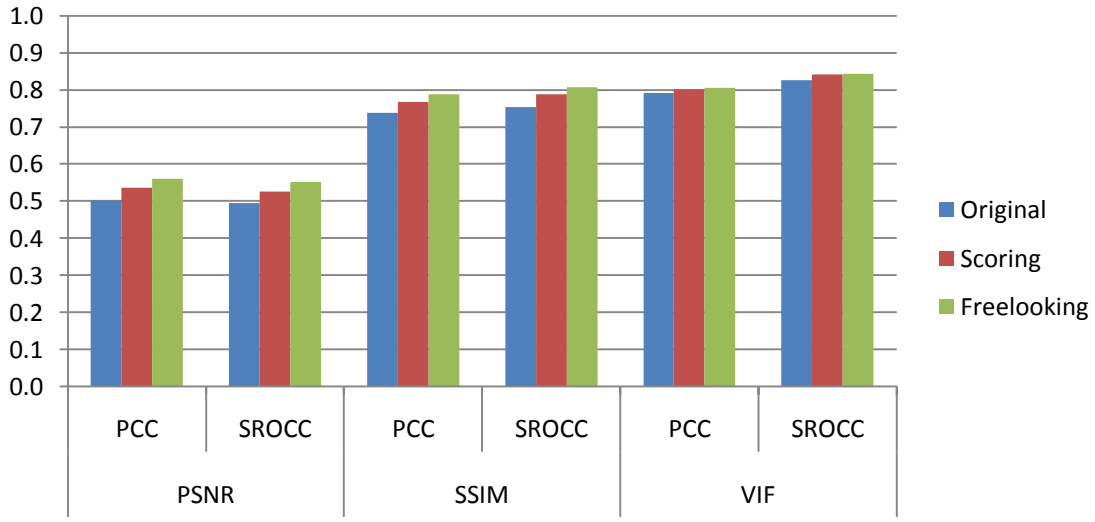


Figure 4.2: Correlation coefficients between the MOS and the scores predicted by PSNR, SSIM, and VIF respectively, once weighted with the saliency of the scoring task and once with the saliency of the free looking task.

As shown in Figure 4.2, there are clear differences between the three metrics: the SSIM correlates a lot better with the MOS than the PSNR, and the VIF even slightly exceeds the performance of the SSIM. More interestingly, however, is that the weighted versions of the metrics perform better than their original version. Weighting with the saliency of the scoring task clearly improves all the results, yet weighting with the saliency of the free looking task improves the metrics' performance even more. This trend applies to all three metrics and both correlation coefficients.

The same correlation coefficients are calculated for the two NR blockiness metrics and the results are depicted in the table and bar graph of Figure 4.3. The results here are less conclusive than the results for the FR metrics: saliency improves the Pearson's correlation of the GBIM, but it lowers its Spearman's correlation. The performance of the Philips blockiness metric is decreased when including the saliency of the free looking task, and is decreased even more when including the saliency of the scoring task. In summary, adding saliency to a NR metric seems to have no, or at most less, added value when compared to adding saliency to a FR metric.

		Original	Scoring	Freelooking
GBIM	PCC	0.788	0.822	0.825
	SROCC	0.913	0.892	0.890
BLOCK	PCC	0.656	0.623	0.644
	SROCC	0.866	0.806	0.856

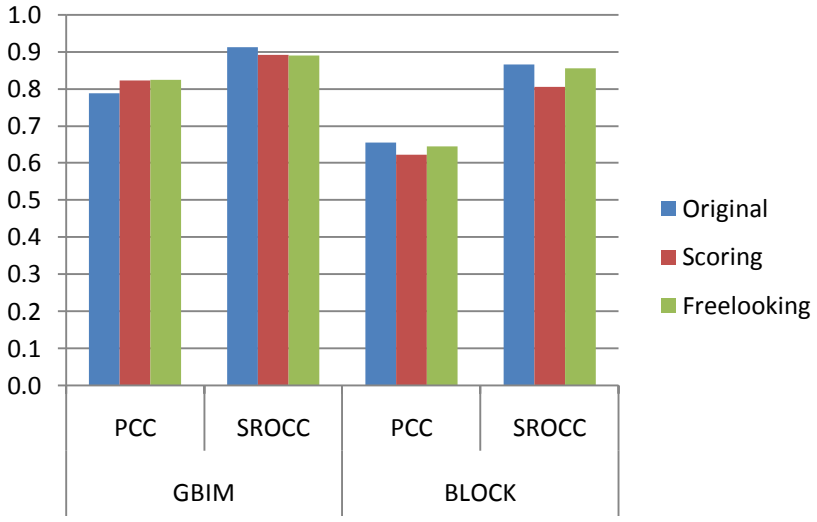


Figure 4.3: Correlation coefficients between the MOS and the scores predicted by the GBIM and Philips blockiness metric respectively, once weighted with the saliency of the scoring task and one with the saliency of the free looking task.

4.3.2. LIVE Data

The LIVE database [17] consists of 29 natural images, compressed using both JPEG and JPEG2000 at several compression levels between 1 and 100, resulting in 233 and 227 stimuli respectively. A Differential Mean Opinion Score (DMOS) is available for all stimuli and a free looking saliency map for all 29 original images. The DMOS values are again compared to the predicted quality scores of the three FR metrics by means of the PCC and the SROCC. The results of the three FR metrics are presented in the table and graph of Figure 4.4.

		Original	Weighted
PSNR	PCC	0.842	0.859
	SROCC	0.841	0.859
SSIM	PCC	0.850	0.863
	SROCC	0.903	0.908
VIF	PCC	0.905	0.905
	SROCC	0.905	0.905

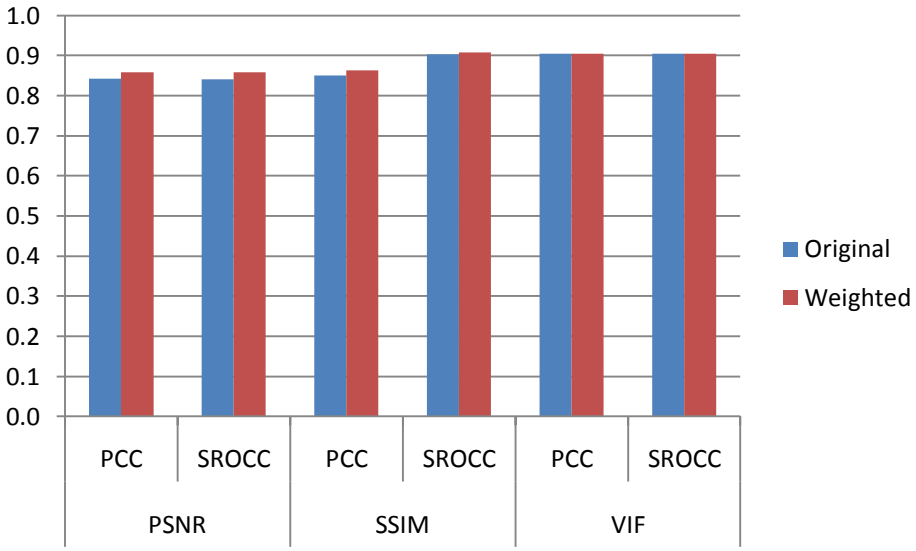


Figure 4.4: Correlation coefficients between the DMOS and the scores predicted by the PSNR, SSIM, and VIF: the original metric vs. the version weighted with free looking saliency.

Figure 4.4 confirms the trends we found for our own data (see section 4.3.1). The PSNR and SSIM clearly perform better when saliency is included than in their original version without saliency, according to both the Pearson's and the Spearman's correlation coefficients. The VIF shows no clear difference between the original and the saliency weighted version. The PSNR performs much better on this database than our own database, while the performance of the SSIM and the VIF are more or less comparable across the two databases. The performance difference of the PSNR might be due to the different compression levels used to create the stimuli in the two databases; the LIVE database contains more low quality stimuli than our database.

The results of the three NR metrics can be found in the table and corresponding graph of Figure 4.5. The performance of the GBIM is lower than with our own database, yet saliency improves both correlations, while saliency only improved the Pearson's correlation on our own database. As with our database, the Philips blockiness metric yields inconclusive results for the LIVE database, since the Pearson's correlation coefficient decreases and the Spearman's correlation coefficient slightly increases. The results of the blur metric are the other way around: the Pearson's correlation coefficient slightly increases, while the Spearman's correlation coefficient decreases.

		Original	Weighted
GBIM	PCC	0.593	0.602
	SROCC	0.873	0.900
BLOCK	PCC	0.544	0.504
	SROCC	0.875	0.888
BLUR	PCC	0.545	0.560
	SROCC	0.664	0.650

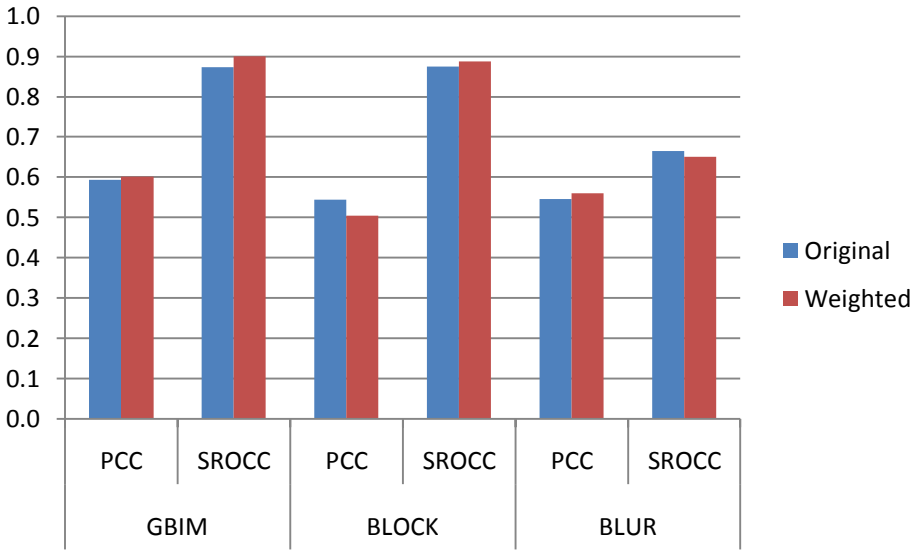


Figure 4.5: Correlation coefficients between the DMOS and the scores predicted by the GBIM, the Philips blockiness metric, and the blur metric: the original metric vs. the version weighted with free looking saliency.

4.3.3. Alternative Weighting Methods

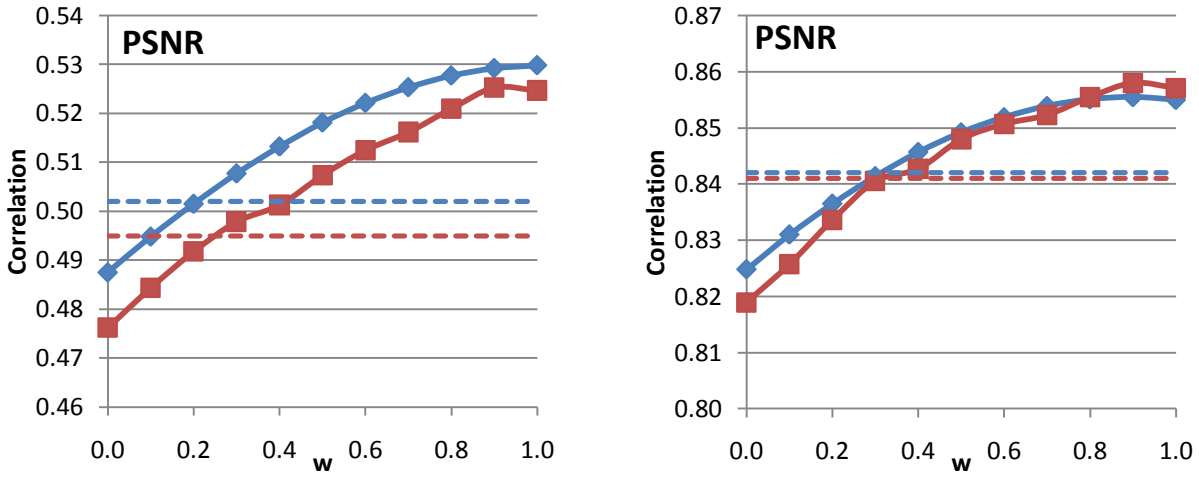
The results above are obtained by simply multiplying the distortion map of the metric with the saliency map. To further improve the performance gain, we have evaluated different methods of applying this saliency information obtained from freely looking to the images. This subsection will describe the steps in the research process of finding a better weighting technique than multiplication with the saliency map.

Region of Interest Weighting

Assuming that compression artefacts in the ROI are more annoying to a human observer than artefacts in the background, IQA metrics should weight the distortion in the ROI more heavily. Similar to the method used by U. Engelke & H.J. Zepernick [9] we calculated the distortion in the ROI and the background separately. A weighting value w is applied to both components in order to define their relative importance. We have already determined the ROI in section 3.2.3, so we used that to calculate a new image quality score by adding the quality in the ROI to the quality in the background:

$$IQ = w \cdot IQ_{ROI} + (1 - w) \cdot IQ_{BG}$$

where IQ_{ROI} and IQ_{BG} are the image quality values of the ROI and background respectively, and $w \in [0,1]$ is the weighting value. To find the optimal value of w for our own database and the LIVE database, we calculated the correlations with the MOS for the PSNR and the SSIM with the weighting value ranging from 0 to 1. The results are depicted in Figure 4.6. The plots show that when $w > 0.3$ both metrics already surpass their original versions, indicated by the horizontal dashed lines. Moreover, they perform best near $w = 1.0$, which means that only the ROI is taken into account and the background information is entirely discarded. Nevertheless, they do not perform as good as the metrics that are weighted with the entire saliency map.



(a) (b)

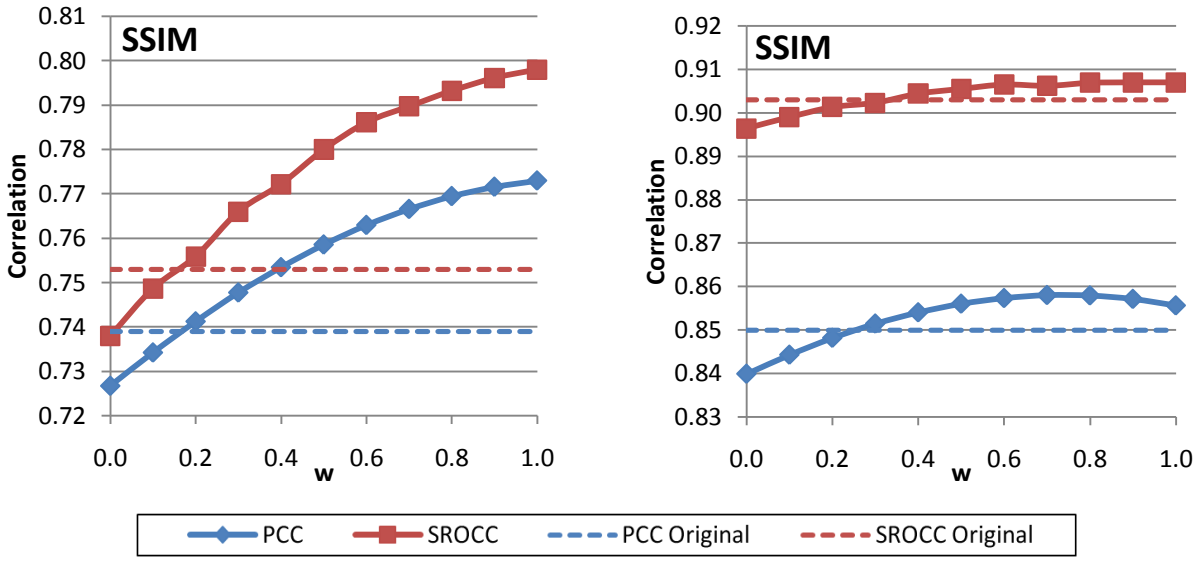


Figure 4.6: The Pearson's (PCC) and Spearman's (SROCC) correlation for different values of w, being the relative contribution of the quality of the ROI to the quality of the background: (a) the PSNR for our database, (b) the PSNR for the LIVE database, (c) the SSIM for our database, and (d) the SSIM for the LIVE database.

4.3.3.1. Adaptive Weighting

After further examining the cause behind the trend in Figure 4.6, we found that the importance of the background distortion increased for images with gradient colours in the background, e.g. the sky. These areas were extra sensitive to blocking artefacts caused by JPEG compression. Therefore, fully discarding the background quality was generally not the best option. To effectively incorporate the background quality, we examined an adaptive weighting method, where the weighting value w was based on the difference in distortion characteristics of the ROI and the background. For this purpose, we tested two distance measures based on the distortion histogram, namely the Euclidean distance d_E and the intersection distance d_I:

$$d_E(h_{ROI}, h_{BG}) = \sum_{m=0}^{M} (h_{ROI}[m] - h_{BG}[m])^2 \qquad d_I(h_{ROI}, h_{BG}) = 1 - \sum_{m=0}^{M} \min(h_{ROI}[m], h_{BG}[m])$$

where h_{ROI} and h_{BG} are the histograms of the distortions in the ROI and background respectively, consisting of M bins. Based on the histogram distance measure we determined the weighting value w: if the distance was above the average distance, we weighted the ROI more ($w = 0.8$), and if the distance was below average, we weighted the background more ($w = 0.2$). The results of this method are presented in the table and corresponding graph of Figure 4.7.

		Our Database			LIVE Database		
		Original	Euclidean	Intersection	Original	Euclidean	Intersection
PSNR	PCC	0.502	0.530	0.525	0.842	0.854	0.853
	SROCC	0.495	0.525	0.524	0.841	0.856	0.854
SSIM	PCC	0.739	0.775	0.768	0.850	0.858	0.854
	SROCC	0.753	0.796	0.795	0.903	0.907	0.905

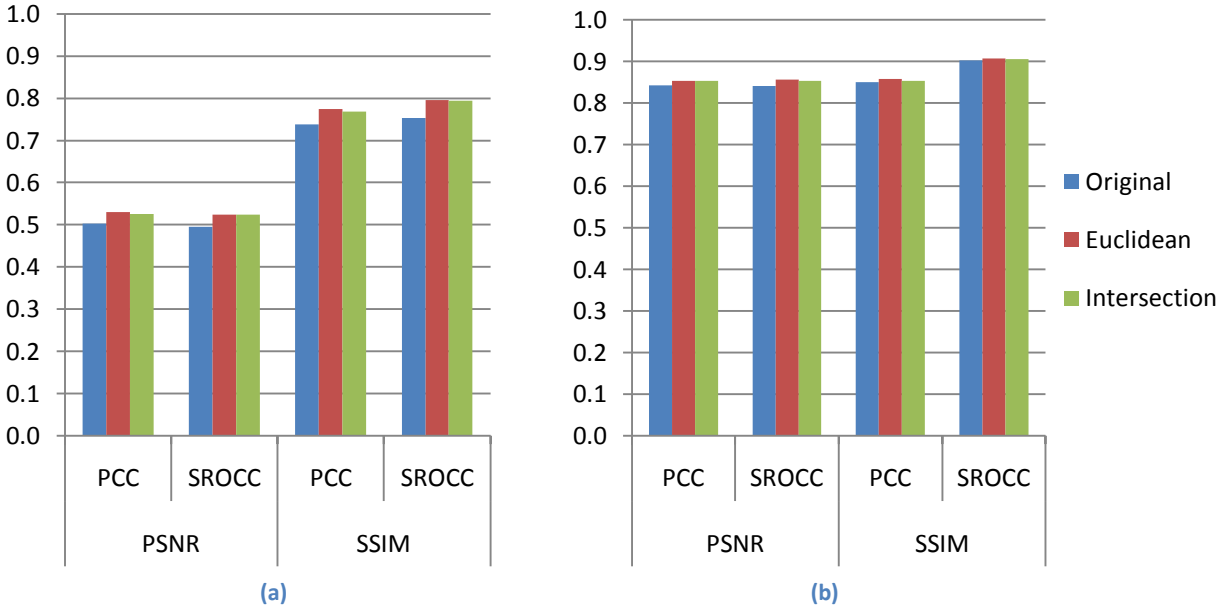


Figure 4.7: Correlation coefficients between the MOS and PSNR and SSIM applied in three versions, namely the original metric, the metric weighted with the Euclidean distance between the distortions in ROI and background, and the metric weighted with the intersection distance: (a) for the results of our database, and (b) for the results of the LIVE database.

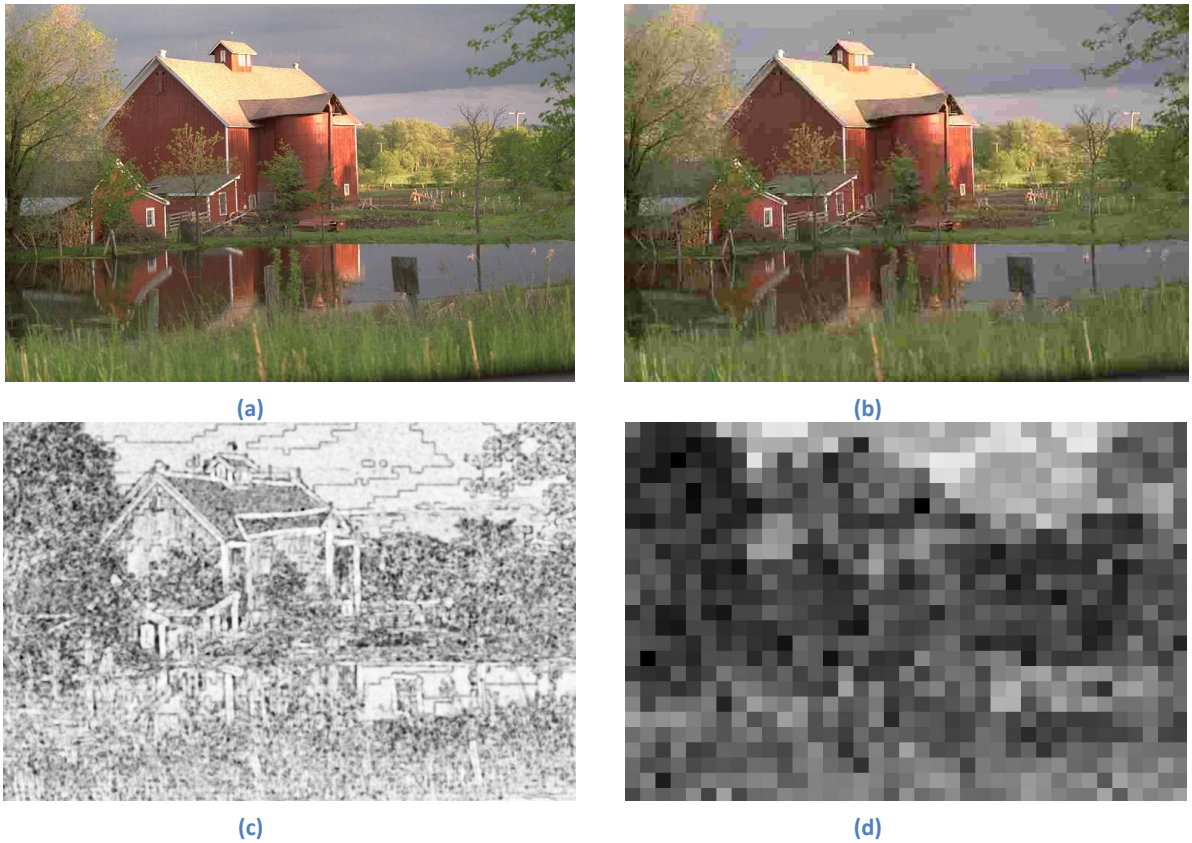
These results show that the adaptive weighting method improves the performance of the original metrics. Weighting the quality with the Euclidean distance between distortions in the ROI and background is slightly better than weighting with the intersection distance. However, the performance gain does not exceed the much simpler ROI weighting method described in the previous section, i.e. a fixed weight of $w = 1.0$ outperforms the adaptive weighting method.

4.3.3.2. Variance-Based Weighting

From the previous two weighting methods we learned that the distortion in the ROI had a dominant contribution to the overall perceived quality, while the contribution of the distortions in the background should be neglected. In part, this can be explained by the fact that observers were more annoyed by compression artefacts in the ROI, and thus weighted these artefacts more heavily in their image quality judgement. However, even when annoying blocking artefacts appeared in the background, the predicted quality score of the metrics still performed best when these artefacts were ignored and only the distortion in the ROI was taken into account. This might suggest

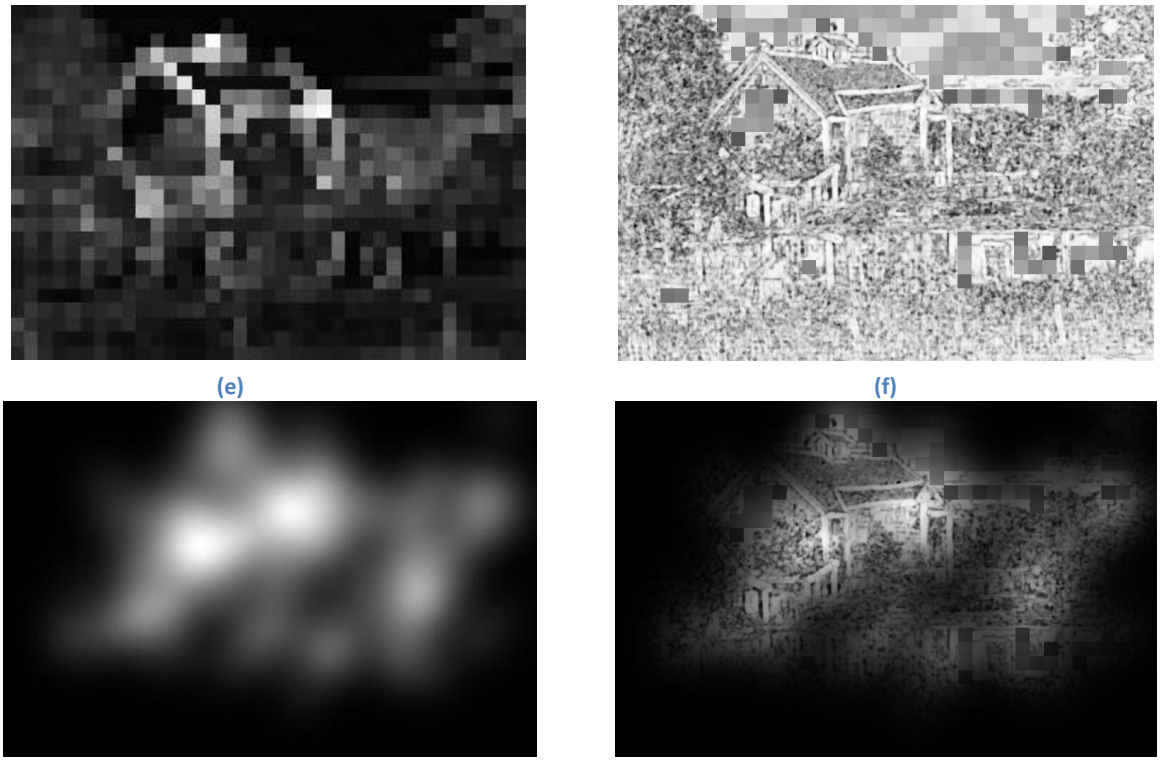
that the metrics were just not able to predict the background quality properly. With that in mind, we took a different approach to improve the performance of the metrics: instead of developing a more effective weighting method, we focused on correcting the metrics' erroneous quality prediction in the background.

Examining the distortion map of the stimuli reveals that blockiness is not well represented by the metrics. To illustrate this we use the example image in Figure 4.8a. After compressing this image, noticeable blocking artefacts appear in the sky, as can be seen in Figure 4.8b. However, the SSIM map (Figure 4.8c) represents the sky as a bright colour, indicating high quality, with merely a few darker lines around the block edges. Averaging this area would result in an overall high quality score, while a person would probably rate the area fairly low in quality. To circumvent this issue, we first divide the image into small blocks of 20×20 pixels, and then use the minimum SSIM value per block, instead of the average (Figure 4.8d). By doing so, the dark lines at the block edges represent the quality of the whole block. However, high resolution detail would be lost with this approach. For that reason, we only apply this process to the smooth areas in an image, which are defined based on the variance in intensity per block of the original image. The dark areas in the variance map (Figure 4.8e) clearly cohere with the erroneous bright areas in the SSIM map. The variance map is normalized to a range of 0 to 1 by dividing it with the maximum variance. Finally, the blocks with a variance below 0.005 are filled with their minimum SSIM value, and the blocks with a higher variance retain their detailed SSIM values, resulting in the new SSIM map of Figure 4.8f.



(a) (b)

(c) (d)



(e) (f)

(g) (h)

Figure 4.8: Illustration of variance based weighting for an example image: (a) the original image, (b) its JPEG compressed version, (c) the SSIM map calculated between (a) and (b), (d) the minimum SSIM value per block of 20×20 pixels, (e) the variance per block of the original image, (f) the combined SSIM map, where the variance per block determines whether the content of (c) or (d) is used, (g) the saliency map obtained from the free looking task, and (h) the new SSIM map weighted with the saliency map.

The new distortion map is then multiplied with the saliency map obtained for freely looking to the image (Figure 4.8g). This results in a saliency weighted distortion map, shown in Figure 4.8h. We tested this new approach for both the PSNR and SSIM and evaluated the outcome with the PCC and SROCC. The resulting correlation coefficients are shown in the table and graph of Figure 4.9. As a comparison, we also included the earlier results obtained with the original metrics and with the metrics that were weighted with the saliency maps for freely looking at the images (section 4.3.1 and 4.3.2).

It is clear from both correlation measures that the new variance-based method (green bars in Figure 4.9) outperforms the original metrics (blue bars). Even without adding saliency information the new method (green bars) performs better than the original metrics weighted with the saliency maps obtained with a free looking task (red bars in Figure 4.9). However, whereas the original metrics are clearly improved by saliency, the variance-based versions do not show consistent improvement when saliency information is included (purple bars in Figure 4.9): in the case of the SSIM adding saliency to the new approach improves the prediction performance for our own database, yet not for the LIVE database. In the case of the PSNR it is the other way around: adding saliency to the new approach slightly improves the prediction performance for the LIVE database, yet decreases its performance for our own database. In other words, the new method clearly improves both metrics, but the additional benefit of integrating saliency is inconclusive.

		Our Database				LIVE Database			
		Original	Saliency	Variance	Var. + Sal.	Original	Saliency	Variance	Var. + Sal.
PSNR	PCC	0.502	0.560	0.616	0.599	0.842	0.859	0.863	0.871
	SROCC	0.495	0.551	0.612	0.592	0.841	0.859	0.857	0.869
SSIM	PCC	0.739	0.788	0.829	0.832	0.850	0.863	0.887	0.877
	SROCC	0.753	0.807	0.836	0.842	0.903	0.908	0.914	0.912

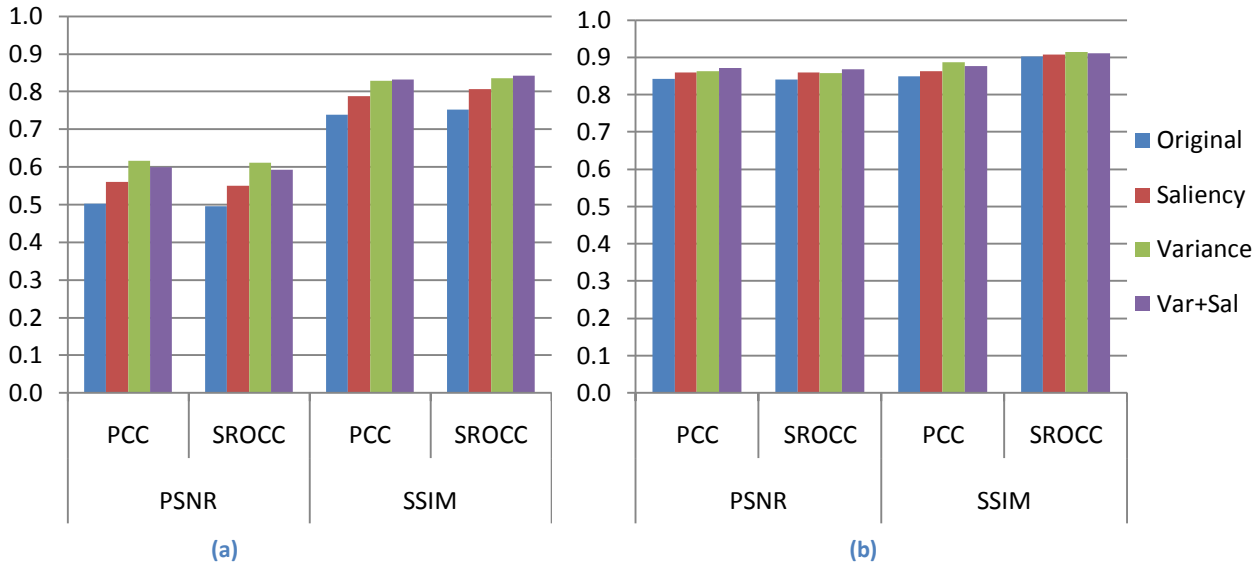


Figure 4.9: Correlation coefficients between the MOS and PSNR and SSIM based new metrics. The results for the original metrics ("Original") and the versions weighted with saliency of the free looking task ("Saliency") are the same as in Figure 4.2 and Figure 4.4. The new variance-based method is indicated with "Variance", and the new method weighted with the saliency of the free looking task is indicated with "Var. + Sal.". (a) represents the results for our own database, and (b) for the LIVE database.

4.4. Discussion

In this section we have investigated the possibility of enhancing the performance of existing IQA metrics by using visual attention. Current IQA metrics average the detected distortions across a whole image to obtain a single quality value. Hence, they do not take into account that distortions in salient regions of an image might be more annoying to a human observer than distortions in other regions. Therefore, adding visual attention to IQA metrics has been a promising research topic in the last decade. In addition to the ongoing research, we compared the two types of visual attention, i.e. obtained from observers with or without a task, more particularly, saliency while assessing image quality, or while looking freely. Several methods were used to incorporate this saliency information. Here we discuss our findings for each method separately.

4.4.1. Saliency Maps

- **Performance improvement**

  Our initial method of incorporating visual attention was to multiply the saliency map with the distortion map of the metric. For this purpose we used the saliency maps that we derived from the eye tracking experiment, described in section 3, and the saliency maps of the LIVE database. Based on the Pearson's and Spearman's

Rank Order correlations coefficients, we found that all three FR metrics (i.e. the PSNR, SSIM, and VIF) were clearly improved by including saliency, and this observation was consistent across both databases. This finding is in agreement with the research of [9], [11], and [44], who also found that saliency information enhanced the quality prediction of the PSNR, the SSIM, and/or the VIF. However, we found no clear performance gain for the NR metrics (i.e. the GBIM, the Philips blockiness metric, and blur metric).

- **Difference between both tasks**
  From the experimental data we generated two saliency maps, one for the free looking task and one for the quality assessment task. In section 3 we have already seen that there is a difference in visual attention between the two tasks. Therefore, we applied the saliency maps of both tasks to the metrics separately. We found that the saliency maps of the free looking task performed better than the maps of the scoring task. This was in contradiction to our expectations; our initial thought was that including the visual attention of people who determined the MOS would yield predicted quality scores closer to the MOS.

- **Original images as stimuli**
  Given that the saliency maps of participants without a task resulted in a better performance of quality prediction than the maps of participants with a quality assessment task, suggests that the visual attention of observers viewing original undistorted images might yield an even better quality prediction performance. Our saliency maps depict visual attention of observers who were looking at JPEG compressed images, and thus, compression artefacts could have distracted our participants and influenced their ocular behaviour. Nonetheless, we expect that using original images as stimuli for measuring saliency would only yield a minor improvement in the quality prediction performance, if any at all, since in section 3.3 we have shown that the effect of the compression level on the fixation duration in the ROI was not significant

4.4.2. ROI Maps

- **Fixed weight**
  Similar to the method used by Engelke & H.J. Zepernick [9], instead of using the full saliency map, we only used the ROI (e.g. Figure 3.6). This way we could separately weight the distortion in the ROI and the distortion in the background. We expected that the weight of the ROI would be higher than the weight of the background, since observers might consider the compression artefacts in the ROI to be more annoying than the ones in the background. However, we obtained results that exceeded our expectations: the metrics performed best when the background distortion was fully discarded and only the distortion in the ROI was taken into account. The resulting correlation coefficients between the perceived and predicted quality scores, however, were still quite low: they did not surpass the correlation of the metrics that are weighted with the saliency maps. We expected that these low correlations were due to the fact that human viewers did not fully discard the distortions in the background.

- **Adaptive weight**

  To improve the results of the fixed weighting method, we adapted the weights of the ROI and background based on the characteristics of the distortion. Nonetheless, the performance gain of this new method did not surpass the results of the much simpler fixed weighting method. This suggested that the distortion in the background had to be neglected by the metric independent of the distortion characteristics in the background. Since the MOS suggested that viewers nonetheless were to some extent annoyed by artefacts in the background, we had to conclude that IQA metrics were unable to properly predict the perceived image quality in the background. To examine the likelihood of this hypothesis, we focused on the distortion maps.

4.4.3. Distortion Maps

- **Gradient colours**

  By carefully investigating individual stimuli we discovered that blocking artefacts often appear in smooth areas with gradient colours, such as the sky. However, the PSNR and SSIM metrics are not well suited to represent blockiness in such areas properly, as can be seen in Figure 4.8c. To compensate for these erroneous areas, we altered the metrics such that the minimum distortion value, instead of the average value, was used in smooth areas. This simple modification to the PSNR and SSIM resulted in a quality prediction performance that clearly surpassed the performance of the original metrics, even after integrating the original metrics with saliency information. When expressing the performance in a Pearson's correlation coefficient between the MOS and the predicted score, we found a gain in performance ranging from about 4% for the SSIM applied to the LIVE database, to about 23% for the PSNR applied to our own database. The gain in performance is even more pronounced when looking particularly to those images that have a pronounced area with a gradient colour. Considering only the half of the images with the lowest mean variance, the performance gain increases to about 8% for the SSIM applied to the LIVE database and to about 30% for the PSNR applied to our own database.

- **Including saliency**

  The modified distortion maps were multiplied with the saliency maps to evaluate the additional benefit of integrating visual attention. We did not find a consistent improvement across the two databases and metrics. This is in agreement with the research of A. Ninassi et al. [12]. They also concluded that the improvement in quality prediction by including saliency could be the effect of a spatial coherence of the saliency with the errors in quality prediction rather than the effect of the saliency information itself. In other words, by weighting with saliency, the areas in the background with an erroneous quality prediction are weighted less, which in turn improves the overall prediction accuracy. Therefore, if we first reduce the errors in quality prediction in the background, by modifying the distortion maps, the additional advantage of weighting the quality prediction with saliency is minimal. This would also explain why the saliency obtained from a free looking task performs better than the saliency obtained from a scoring task, since the latter covers more of the erroneous areas. Furthermore, it might have caused the absence of any improvement in the NR blockiness metrics when including saliency, since these metrics are better suited to correctly detect the blocking artefacts in the smooth areas of the background.

5. Conclusions and Recommendations

In this report we have investigated the visual attention of people who looked at images and the effect of giving them an image quality assessment task. For this purpose, we conducted an experiment where the ocular behaviour of two groups of test subjects was recorded with an automated eye tracker. The first group looked at a set of JPEG compressed images freely, while the second group was instructed to rate their quality. The gathered eye movement data were first converted into saliency maps, which depict the fixation density at an image, after which we could objectively determine the ROI. We statistically analyzed the duration of the fixations inside the ROI compared to the total duration of all fixations, in order to determine the differences between the two groups of subjects. From the analysis we made the following interesting observations:

- Participants without a task look more to the ROI and less to the background than observers with a quality assessment task.
- Participants in general first look at the ROI, but the attention strays away from the ROI faster during quality assessment.
- The average duration of the fixations to the ROI is longer for participants without a task, while the average duration of the fixations to the background does not differ between the two groups.
- Participants determine the score of low quality images faster than of high quality images.
- The degree of compression only affects the visual attention of participants with a scoring task.
- Viewing an image multiple times does not significantly alter the visual attention.

In the second part of this report we have investigated the application of visual attention to IQA metrics. Current IQA metrics do not take visual attention into account, hence we examined the possibility of improving their performance by incorporating saliency. For this purpose, we used the saliency maps derived from the aforementioned eye tracking experiment, together with the saliency maps available for the LIVE database. In this way we could test whether our results were consistent across different databases. Besides weighting the saliency maps with the distortion maps of the metrics, we also evaluated different weighting methods. Namely, we separately weighted the distortions in the ROI and background, using both a fixed weight and an adaptive weight. Finally, we used a method where we modified the distortion maps before weighting them with the saliency maps. From these tests we found the following results:

- Weighting the FR metrics (PSNR, SSIM and VIF) with the saliency maps clearly improves their image quality prediction.
- The NR metrics (GBIM, Philips blockiness, and blur) do not show a consistent improvement when weighted with the saliency maps.
- The saliency maps of the participants without a task yield a higher performance gain than the maps of the participants with a quality assessment task.
- The distortion in the ROI has a more positive influence to the overall prediction accuracy than the distortion in the background.

The performance improvement found for the FR metrics can be explained by two factors. First, observers are likely to consider distortion in the ROI to be more annoying than in the background. Secondly, the quality prediction in the background is likely to be more prone to errors than in the ROI, since the FR metrics are unable to properly predict the annoyance of blocking artefacts in smooth areas in the background. The background often has smooth areas with gradient colours (e.g. the sky), which are sensitive to blockiness caused by JPEG compression. Weighting the metrics with the saliency maps reduces the influence of the erroneous smooth regions in the background, thus increasing their performance. Since the participants without a task paid less attention to the background than the participants with a quality assessment task, the saliency maps of the former outperformed the maps of the latter. Furthermore, the inconclusive results of the NR metrics could be caused by their more sophisticated distortion detection and quality prediction algorithm. Nevertheless, further research is still required to assess the validity of this hypothesis and to fully understand the causality between incorporating saliency and the quality prediction improvement.

In this report we have used visual attention gathered from real observers. Despite the fact that this method is the most reliable, it is too cumbersome and time-consuming to be practical in most applications. Therefore, this research could be extended to artificial saliency/ROI-detection algorithms. Incorporating such algorithms into current IQA metrics would allow them to operate automatically on a given image, without the need for any further human intervention. Such attention models could be integrated into image compression systems, which would then be able to allocate more bits to the ROI, possibly yielding a higher overall perceptual quality. Another option to expand this research would involve video quality assessment. The methods used for still images cannot be simply applied to all frames in a video sequence, since each frame is only viewed for a fraction of a second and the element of motion is an important cue for visual attention. Hence, further research is necessary in order to apply the visual attention mechanism to video quality assessment metrics.

Appendix A. Poster Presented At ECVP 2009

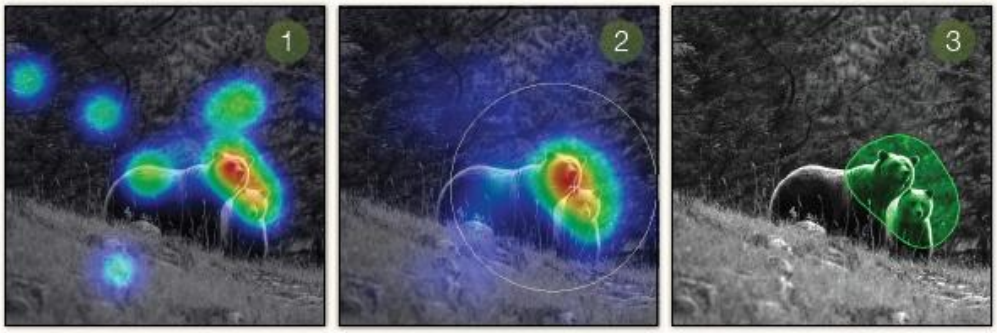
TASK AND VIEWING BEHAVIOR

Using Eye Tracking to Assess the Task Effect on Viewing Behavior

Most image quality experiments use some subjective quality scores. To collect such scores, observers are given a task to assess the quality of a set of images. This makes it important to have an understanding of how the given task influences the way people look at the image material. This work shows that the quality assessment task actually changes the behavior of the observers.

IDENTIFYING THE ROI

1. Combining fixation points with eye tracker
2. Averaging the gathered data from all participants
3. A proper threshold is chosen to select the ROI from saliency



	S.H.	S.L.	F.H.	F.L.
S.H.	x	Differ	Same	Same
S.L.	Differ	x	Differ	Differ
F.H.	Same	Differ	x	Same
F.L.	Same	Differ	Same	x

S : Scoring
F : Free-looking

H : High quality images
L : Low quality images

Differ : Significant difference
Same : No significant difference

ANALYZING SALIENCY

The relation between task and saliency is examined for different ranges of image quality

HANI ALERS, HANTAO LIU, LENNART BOS, PROF. INGRID HEYNDERICKX
Man-Machine Interaction Group
Delft University of Technology
Mekelweg 4, 2628CD Delft, The Netherlands

TUDelft PHILIPS

Bibliography

[1] Wang, Z. and A. C. Bovik (2006). Modern Image Quality Assessment, Morgan & Claypool.

[2] Wang, Z. and A. C. Bovik (2009). "Mean Squared Error: Love It or Leave It?" IEEE Signal Processing Magazine: 88-117.

[3] Wang, Z., A. C. Bovik, et al. (2004). "Image Quality Assessment: From Error Visibility to Structural Similarity." IEEE Transactions on Image Processing **13**: 600-612.

[4] Wang, Z. and A. C. Bovik (2002). "A Universal Image Quality Index." IEEE Signal Processing Letters **9**(3): 81-84.

[5] Sheikh, H. R. and A. C. Bovik (2004). "Image Information And Visual Quality." Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing.

[6] Wu, H. R. and M. Yuen (1997). "A Generalized Block-Edge Impairment Metric for Video Coding." IEEE Signal Processing Letters **4**(11): 317-320.

[7] Muijs, R. and I. Kirenko (2005). "A No-reference Blocking Artifact Measure for Adaptive Video Processing." Proceedings of 13th European Signal Processing Conference (EUSIPCO).

[8] Marziliano, P., F. Dufaux, et al. (2004). "Perceptual blur and ringing metrics: application to JPEG2000." Signal Processing: Image Communication **19**: 163-172.

[9] Engelke, U. and H.-J. Zepernick (2009). "Optimal Region-Of-Interest Based Visual Quality Assessment."

[10] Moorthy, A. K. and A. C. Bovik (2009). "Visual Importance Pooling for Image Quality Assessment." IEEE Journal of Selected Topics In Signal Processing **3**: 193-201.

[11] Ma, Q. and L. Zhang (2008). "Saliency-Based Image Quality Assessment Criterion." ICIC 2008: 1124–1133.

[12] Ninassi, A., O. L. Meur, et al. (2007). "Does Where You Gaze On An Image Affect Your Perception Of Quality? Applying Visual Attention to Image Quality Metric." ICIP07 **2**(169-172).

[13] Zhong, Y., I. Richardson, et al. (2004). Influence of Task and Scene Content on Subjective Video Quality. Image Analysis and Recognition.

[14] Sundstedt, V., A. Chalmers, et al. (2004). "Top-Down Visual Attention for Efficient Rendering of Task Related Scenes."

[15] Vu, C. T., E. C. Larson, et al. (2008). "Visual Fixation Patterns when Judging Image Quality: Effects of Distortion Type, Amount, and Subject Experience." SSIAI 2008.

[16] Ninassi, A., O. L. Meur, et al. (2006). "Task Impact On The Visual Attention In Subjective Image Quality Assessment." EUSIPCO-06.

[17] Sheikh, H. R., M. F. Sabir, et al. (2006). "Subjective Database Release 2." from http://live.ece.utexas.edu/research/quality/release2/databaserelease2.zip.

[18] Sadaka, N. G., L. J. Karam, et al. (2008). "A No-Reference Perceptual Image Sharpness Metric Based On Saliency-Weighted Foveal Pooling." IEEE International Conference on Image Processing: 369-372.

[19] Roorda, A. (2002). "Human Visual System - Image Formation." The Encyclopedia of Imaging Science and Technology: 539-557.

[20] Cadik, M. (2004). Human Perception and Computer Graphics.

[21] Gonzalez, R. C. and R. E. Woods (2002). Digital Image Processing. New Jersey, Prentice-Hall, Inc.

[22] Nave, C. R. (1997). "Hyper Physics." Retrieved October 16, 2008, from http://hyperphysics.phy-astr.gsu.edu/hbase/hframe.html.

[23] Molavi, D. W. (1997). "Eye and Retina." Retrieved October 16, 2008, from http://thalamus.wustl.edu/course/eyeret.html.

[24] Montgomery, T. M. (1998). "Anatomy, Physiology and Pathology of the Human Eye." Retrieved October 16, 2008, from http://www.tedmontgomery.com/the_eye/overview.html.

[25] Campbell, F. W. and J. G. Robson (1968). "Application Of Fourier Analysis To The Visibility Of Gratings." Journal of Physiology **197**: 551-566.

[26] Eckert, M. P. and A. P. Bradley (1998). "Perceptual Quality Metrics Applied to Still Image Compression." Signal Processing **70**: 177-200.

[27] Mason, C. and E. R. Kandel (1991). Principles of neural science. New York, Elsevier Science Publishing.

[28] Martinez-Conde, S., S. L.Macknik, et al. (2004). "The Role of Fixational Eye Movements in Visual Perception." Nature Reviews **5**.

[29] Murakami, I. (2002). "Visual Jitter." Retrieved October 27, 2008, from http://www.brl.ntt.co.jp/people/ikuya/demo/visualjitter/VisualJitter.html.

[30] Meur, O. L., P. L. Callet, et al. (2006). "A Coherent Computational Approach to Model Bottom-Up Visual Attention." IEEE Transactions On Pattern Analysis And Machine Intelligence **28**.

[31] Tatler, B. W. and T. Troscianko (2002). "A rare glimpse of the eye in motion." Perception **31**: 1403-1406.

[32] Krauzlis, R. J. (2003). "Recasting the Smooth Pursuit Eye Movement System." Neurophysiol **91**(591-603).

[33] Hoffman, J. E. (1995). Visual Attention and Eye Movement. Attention. H. Pashler**:** 119-153.

[34] Martinez-Conde, S., S. L. Macknik, et al. (2006). "Microsaccades Counteract Visual Fading during Fixation." Elsevier Inc. **Neuron 49**: 297-305.

[35] Murakami, I. and P. Cavanagh (1998). "A Jitter After-Effect Reveals Motion-Based Stabilization of Vision." Nature **395**: 789-801.

[36] Murakami, I. and P. Cavanagh (2001). "Visual Jitter: Evidence for Visual-Motion-Based Compensation of Retinal Slip Due to Small Eye Movements." Vision Research **41**: 173-186.

[37] Hoffman, J. E. and B. Subramaniam (1995). "The role of visual attention in saccadic eye movements." Perception & Psychophysics.

[38] Wolfe, J. M. (2000). "Visual Attention." De Valois KK: 335-386.

[39] Henderson, J. M. and A. Hollingworth (1999). "High-level Scene Perception." Annu. Rev. Psychol. **50**: 243-271.

[40] Klein, R. M. (2000). "Inhibition of Return." Trends in Cognitive Sciences **4**(4): 138-147.

[41] Itti, L., C. Koch, et al. (1998). "A Model of Saliency-Based Visual Attention for Rapid Scene Analysis." IEEE Transactions on Pattern Analysis and Machine Intelligence **20**(11).

[42] Chen, H. Y. and J. J. Leou (2008). "A New Visual Attention Model Using Texture and Object Features." IEEE International Conference on Computer and Information Technology Workshops.

[43] Rajashekar, U., I. v. d. Linde, et al. (2008). "GAFFE: A Gaze-Attentive Fixation Finding Engine." IEEE Transactions on Image Processing **17**: 564-573.

[44] Moorthy, A. K. and A. C. Bovik (2009). "Perceptually Significant Spatial Pooling Techniques for Image Quality Assessment."

[45] Stentiford, F. W. M. (2001). "An Evolutionary Programming Approach to the Simulation of Visual Attention." Congress on Evolutionary Computation: 851-858.

[46] Bradleya, A. P. and F. W. M. Stentiford (2003). "Visual Attention for Region of Interest Coding in JPEG 2000." Journal of Vision Communications & Image Representation **14**: 232-250.

[47] Sheikh, H. R., M. F. Sabir, et al. (2006). "A Statistical Evaluation of Recent Full Reference Image Quality Assessment Algorithms." IEEE Transactions on Image Processing **15**(11): 3440-3451.

[48] Miyahara, M., K. Kotani, et al. (1998). "Objective Picture Quality Scale (PQS) For Image Coding." IEEE Trans. Communications **46**(9): 1215-1225.

[49] Sheikh, H. R., A. C. Bovik, et al. (2005). "An Information Fidelity Criterion for Image Quality Assessment Using Natural Scene Statistics." IEEE Transactions on Image Processing **14**(12): 2117-2128.

[50] Damera-Venkata, N., T. D. Kite, et al. (2000). "Image Quality Assessment Based on a Degradation Model." IEEE Transactions on Image Processing **9**(4): 636-650.

[51] Liu, H. and I. Heynderickx (2009). "A Perceptually Relevant No-Reference BlockinessMetric Based on Local Image Characteristics." EURASIP Journal on Advances in Signal Processing **2009**.

[52] Wang, Z. and E. P. Simoncelli (2005). "Reduced-Reference Image Quality Assessment Using a Wavelet-Domain Natural Image Statistic Mode." Human Vision and Electronic Imaging X, Proc. SPIE **5666**.

[53] Gunawan, I. P. and M. Ghanbari (2003). "Reduced-Reference Picture Quality Estimation by Using Local Harmonic Amplitude Information." Proc. London Communications Symposium: 137-140.

[54] Webster, A. A., C. T. Jones, et al. (1993). "An Objective Video Quality Assessment System Based On Human Perception." Proc. SPIE **1913**: 15-26.

[55] Daly, S. (1993). The Visible Differences Predictor: An Algorithm for the Assessment of Image Fidelity. Digital Images and Human Vision (A.B. Watson, ed.). Cambridge, MA, The MIT Press**:** 179–206.

[56] Lubin, J. and D. Fibush (1997). "Sarnoff JND Vision Model." cont. to G-2.1.6 Compression and Processing Subcommittee.

[57] Liu, H., J. Redi, et al. (2010). "No-Reference Image Quality Assessment Based on Localized Gradient Statistics: Application to JPEG and JPEG2000." IS&T/SPIE Electronic Imaging 2010, Human Vision and Electronic Imaging XV.

[58] SMI. (2008). "SensoMotoric Instruments GmbH." from http://www.smivision.com/.

[59] Alers, H., H. Liu, et al. (2010). "Studying the Effect of Optimizing the Image Quality in Saliency Regions at the Expense of Background Content." IS&T/SPIE Electronic Imaging 2010, Image Quality and System Performance VII.

[60] Rajashekar, U. and A. Moorthy. (2008). "Laboratory for Image & Video Engineering." Retrieved December 22, 2008, from http://live.ece.utexas.edu/.

[61] NBS. (2008). "Presentation." from http://www.neurobs.com/presentation.

[62] (SPSS). "Statistical Package for the Social Sciences ", from http://www.spss.com/.

[63] Santella, A. and D. DeCarlo (2004). "Robust Clustering of Eye Movement Recordings for Quantification of Visual Interest." Proc. ETRA '04 Symposium: 27-34.

[64] Sadaka, N. G., L. J. Karam, et al. (2008). "A No-Reference Perceptual Image Sharpness Metric Based on Saliency-Weighted Foveal Pooling." Proc. IEEE Int. Conf. ICIP: 369-372.

[65] Wang, Z. (2003). "SSIM Index, Version 1.0." from http://www.ece.uwaterloo.ca/~z70wang/research/ssim/ssim_index.m.

[66] Sheikh, H. R. and A. C. Bovik. (2005). "VIF Multiscale Pixel Domain." from http://live.ece.utexas.edu/research/Quality/vifp_release.zip.

[67] Sheikh, H. R. and A. C. Bovik. (2005). "VIF Wavelet Domain." from http://live.ece.utexas.edu/research/Quality/vifvec_release.zip.