Subjective Image Quality Evaluation Using the Softcopy Quality Ruler Method

Maria Persson

June 23, 2014

Abstract

Image quality is in essence based on subjective experience, and the involvement of human subjects is therefore necessary in one way or the other in order to make reliable assessments of it. This topic is of interest for Axis Communications AB – the market leader in surveillance cameras.

In this thesis, the softcopy quality ruler method described in ISO 20462 is evaluated through a study of one specific image quality problem; human preference in the noise reduction – texture loss space. The method involves quality or attribute judgment by comparison of test images to a series of ordered, univariate (variation in one attribute only) reference images, called "ruler images". Sharpness is used as reference attribute for the ruler images.

For the purpose of the thesis project, a working environment for performing subjective studies was set up. A total of 47 persons (observers) were invited to judge a set of test images varying in noise level and amount of noise reduction as well as scene content.

From the data, confidence intervals were calculated using the non-parametric bootstrap method as well as using the normal distribution. The two methods gave very similar results, and thus it could be confirmed that the data is well described by a normal distribution.

The results of the study indicate that it is possible to determine an optimum level of noise reduction for a given noise level.

For increasing noise levels, the variances of the observer judgments increase, while they decrease for increasing noise reduction levels. This difference may be explained by making the observation that increasing noise reduction levels result in increasing texture blur, which appears more similar to sharpness loss in comparison with noise. It is reasonable to assume that the uncertainty in observer judgments should be lower when comparing images with similar degradations. Therefore, it might be argued that the variability of the judgments depends to a large extent on the perceived similarities between the ruler and test images, in terms of the attribute(s) considered. If the ruler images appear similar to the test images, the variances of the judgments will be lower than for less similar ruler images and test images, and also more strongly dependent on the number of observers. In order to reduce the number of observers when testing some particular image quality attribute(s), care should be taken to use ruler images varying in an attribute similar in appearance to the test images.

Also noticeable are discrepancies between experienced (having experience in judging or evaluating images) and unexperienced observers, as

concluded by a Welch t-test. When judging image quality, experienced observers at Axis, such as imaging engineers, camera designers etc., seem more tolerant to noise than unexperienced observers.

.

Preface

This thesis has come together thanks to many people.

First of all I would like to express my thanks and gratitude to my supervisor at Axis Comminucations AB, Henrik Eliasson, for support and help during this project.

I would also like to express my thanks to all of the 47 observers participating – of course for making it possible to perform subjective studies, but also for making it fun!

I would also like to thank the Core Technologies Imaging Department (where I performed this project), lead by Mats Thulin, for being friendly and including me in the team spirit of the group. I am grateful for the helpful discussions (and computer support(!)) provided when needed, but also for the great company and the discussions not involving this project.

I would also like to thank my supervisor at LTH, Soren Vang Andersen, for showing engagement during the whole process of the project.

At the end, I would like to thank Axis Communications AB for letting me preform from my thesis in a innovative and creative environment that is ”always open”.

Contents

1 Introduction

Axis Communications AB is the market leader in network cameras and in the category surveillance cameras [1]. To establish this position, an important factor has been the ability to have and maintain a high product quality. For a camera, one of the most important properties is the ability to deliver images of high quality. When developing and improving cameras, the ability to specify image quality in a measurable way is therefore of great importance [2].

Then, what is image quality? There is no de facto definition, but for the purpose of this thesis the following is used: "Image quality is the integrated set of perceptions of the overall degree of excellence of the image" [3].

There are several *attributes*, image characteristics, that contribute to image quality. Examples of such attributes are *resolution*, *contrast*, *noise*, etc. These attributes can be measured by physical measurement techniques, usually with some kind of instrument, but also by analyzing images captured of carefully specified test targets.

When approaching the topic of image quality and the ability to measure it, these separate measurements have to be combined in some way. The combination is not trivial. Furthermore, objective attribute measurements are not always coincident with the experience of image quality perceived by humans [4, 5]. Another obstacle, making image quality even harder to specify, is that of unmeasurable, but still obvious, properties that contribute to the overall perception of the image. Compression artifacts are an example of such a property.

With this being said, it should be obvious that creating an overall metric for image quality is a complex problem with no absolute solution. Why is it like that? The answer is related to the property of perception – image quality is not objective. Something that is objective is existing independently of an individual's perception [6]. Image quality, which arises through perception, is instead subjective. Something that is subjective is dependent on an individual's perception for its existence [7]. Just like beauty, happiness, coldness, warmth, etc., image quality stems from, or takes place in, peoples' minds rather than in the external world.

Is it impossible to somehow specify image quality in a measurable way then? The answer is no. What is impossible is to exclude subjectivity when dealing with image quality. Even if only using objective attribute measurements, the combination of them will be subjective and the unmeasurable properties are highly likely to be left out. If trying to somehow also involve the unmeasurable properties, subjectivity again comes in to the picture.

To manage this problem one can let many humans, *observers*, judge many images. Subjective methods can be described as a procedure of collecting data on images by letting humans somehow record their experience of perceived quality [2, 8, 9]. By utilizing statistical methods, the data can be used to quantitatively specify image quality.

To be able to use subjective methods when investigating image quality, one ground assumption has to be made. This postulate is that people tend to have a general idea of what a picture of good quality looks like, at least within some confidence interval. If that was not the case, there would be no point in investigating image quality in the subjective domain at all. As a matter of fact, a large body of research shows that there indeed seems to be a general opinion on perceived image quality [2, 4, 8, 9, 10, 11, 12, 13, 5]. Intuitively,

this assumption may seem evident, but is still worth mentioning. Nothing is certainly black or white when dealing with subjectivity.

2 Aim of the thesis

This thesis has three main purposes.

1. Set up a working environment at Axis Communications AB for performing subjective image quality studies – a "Visual lab"
2. Make a survey of existing subjective methods and decide which is best suited for Axis' needs
3. Perform one subjective study addressing one specific image quality problem. The specific problem is human preference in the noise reduction - texture loss space

A more detailed evaluation of the results of the study should be performed, involving statistical methods.

The Visual lab, and the first study conducted in it, will hopefully be useful as a starting point for later studies to be conducted at Axis.

3 Background

3.1 The determination of image quality

One possible way to describe image quality is to create a model that visualizes it. An example of such a model is The Image Quality Circle [2], made by Engeldrum. This model will now be examined, as an introduction to the topic of image quality.

3.1.1 The Image Quality Circle

The Image Quality Circle is a model or a framework arranging the elements that contribute to image quality, visualized in figure 1. The first purpose of the model is to make the subject of image quality more understandable and complete [2]. The Image Quality Circle also serves as a process model in imaging product development projects. In this thesis, the Image Quality Circle serves the first purpose, not the second.

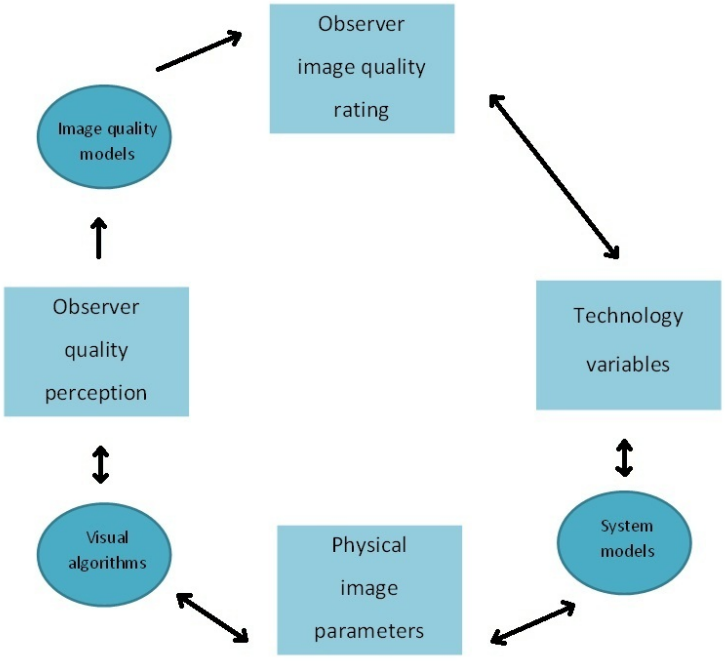


Figure 1: The Image Quality Circle

Walking around the circle, *technology variables* are connected to *physical image parameters* by *system models*. After that follows *observer quality perceptions*, connected to *physical image parameters* by *visual algorithms*. *Observer quality perceptions* are connected to *observer quality perceptions* by *image quality models*.

The Image quality circle is incomplete without the involvement of human observers. About half of the Circle, from visual algorithms to image quality rating, requires human judgments. *Psychometric scaling* can be described as "mind measuring" [2]. Photometric scaling methods are psychological measurement methods based upon human judgments. In Engeldrum's model, humans are called customers, but for the purpose of this thesis they will be called observers.

An *attribute* is a characteristic of an image describing it in some sense. A *perception* is a sensation received by the human mind through human senses. A *perceptual attribute* is a an attribute that a human can perceive by the sense vision; a visual image characteristic. Sharpness, colorfulness and lightness are examples of perceptual attributes.

The four elements of the Image Quality Circle

The Image Quality Circle consists of four elements, shaped as boxes, see figure 1.

- The *observer quality rating* represents the judgment of the overall image quality made by the observer. The observer quality rating is the image quality "value".

- The *observer quality perceptions* are the perceptual attributes that form the basis of the judgment of the overall image quality made by the observer. The observer quality perceptions are "what is seen".
- The *technology variables* are controllable variables describing the hardware of an imaging system (the camera, the imaging material etc.). Examples of such camera properties are properties of the lens and the sensor, such as angle of view, the maximum aperture, pixel per inch, pixel size, etc. Examples of image material properties are paper thickness, waterfastness etc. The technology variables are "what is controlled".
- The *physical image parameters* are quantitative measurements of an image. The parameters are usually obtained by objective measurement techniques that may involve physical instruments. Examples are resolution, contrast, noise etc. More complex examples are functions of spatial frequency, like the Modulation Transfer Function for example. The physical image parameters are "what is measured".

The units of Observer quality ratings and observer quality perception values are based on human perception. The units of physical image parameters or technology variables are instead of physical nature.

The connecting links of the Image Quality Circle
The four elements described above are connected by three connecting links, "recipes"; models, formulas, computer code, etc., translating some kind of value(s) to other kind(s) of value(s), shaped as ellipses, see figure 1.

- *System models* predict the physical image parameters from the technology variables or vice versa. In the simplest case, a system model can be one single physical measurement.
- *Visual algorithms* predict observer quality perception values from physical image parameters. In the simplest case, a visual algorithm computes one observer quality perception value from one measured physical image parameter.
- *Image quality models* link observer quality perceptions and observer quality ratings. Model inputs are values of observer quality perceptions and the output are observer quality rating values. In the simplest case, an image quality model translates one observer quality perception value to an observer quality rating value.

3.2 Image quality attributes

In the Image Quality Circle, image quality attributes are labeled observer quality perceptions and/or physical image parameters. As mentioned, an image attribute is a characteristic of an image describing it in some sense. There are several such attributes that contribute to image quality. Some of them, important for this thesis, will now shortly be presented.

3.2.1 Resolution

The detail an image holds is related to the resolution of the image [14]. In conventional terms, image resolution quantifies how close lines can be to each other and still be visibly resolved. Image resolution can be seen as minimum resolvable distance.

One way to quantify resolution is through the Rayleigh criterion [15]. Two point sources emitting light are said to be resolved when the main diffraction maximum of the image of the first coincides with the first minimum of the image of the other, see figures 2, 3 and 4. In other words, two points in an image are resolved if the distance between them is greater or equal to the distance between the main diffraction maximum of one and the first minimum of the other. The smallest spot size in the image is given by that distance.

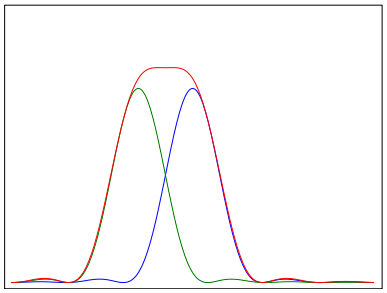


Figure 2: Unresolved

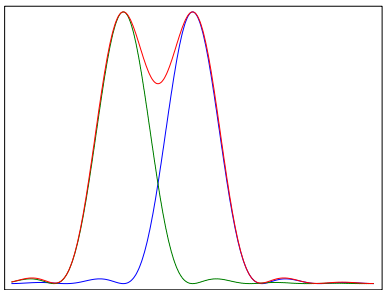


Figure 3: Resolved limit

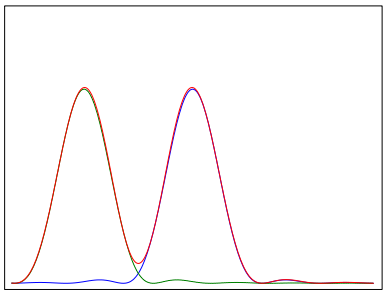


Figure 4: Fully resolved

Another way to quantify resolution is to use a resolution bar [15]. A resolution bar is a series of light and dark bars decreasing in width and distance. The spatial frequency of the resolution bar is the inverse distance between two bars, in units of line pairs per millimeters (lp/mm) or cycles per millimeters (cy/mm). The finest line structure that can be resolved by a human observer under some specific viewing conditions can be seen as a measure of the spatial resolution of the optical system.

Pixels are units of digital images. Pixel resolution is associated with the capacity of the sensor of an optical system [14]. The resolution in digital images depends to some degree on the number of pixels. The larger the amount of pixels, the higher the resolution in the image.

3.2.2 Contrast

The luminance and/or color difference in an image that makes objects distinguishable is corresponding to the contrast of the image [16].

The level of contrast can be quantified in terms of *modulation*. Modulation is the ratio of the difference in luminance of two objects and their average luminance, see A.1.4, Appendix A. This definition will be used when discussing the *Modulation Transfer Function (MTF)*.

3.2.3 Sharpness

Sharpness is a subjective quality attribute of an image [17]. Sharpness indicates the perceived quality of details of an image.

The sharpness impression can be described in terms of modulation and spatial frequencies using the MTF. The MTF of an optical system is the ratio of the image modulation to the object modulation at all spatial frequencies. It is a measure of the modification of contrast from object to image over the spatial frequency spectrum. The MTF is the magnitude of a one-dimensional slice of the optical transfer function (OTF) or the magnitude of the Fourier transform of the line spread function (LSF). See sections A.1.1, A.1.2, A.1.3, A.1.5, in appendix A, for definitions of the PSF, the LSF, the OTF and the MTF. The MTF has become a widely used means of specifying the performance of many sorts of systems. Lenses, the atmosphere, displays, camera films to mention a few. If the MTFs for the independent components in a system are known, the total MTF of the system is then often their product. This follows from the convolution theorem, see B.2, Appendix B. This is an important property

of the MTF. The understanding of the MTF and especially this multiplication property will be important later in this thesis.

3.2.4 Color

Color is a highly subjective attribute. It is dependent on the physics of light, the chemistry of matter, the geometric properties of the object and human visual perception [18]. When light meets the human eye, a series of complex events leads to the sensation of color [8]. Some more information on color can be found in section 3.3.

3.2.5 Noise

Noise in an image is usually defined as random fluctuation within the image [19]. It is a deviation from the captured object. Noise is normally particularly visible in uniform areas.

Noise can be fluctuations in color (generating chromatic noise), fluctuations in luminance (generating luminance noise) or a combination of the two [20].

The attribute is usually characterized in terms of standard deviation [19]. Standard deviation represents the dispersion of gray levels around the mean luminance gray level. The signal to noise ratio (SNR) is then the ratio of the mean value and the standard deviation. Another unit, which can be derived from the SNR is the dynamic range, which may be defined as the ratio between the highest and lowest gray luminance, where the lowest gray luminance level is usually taken as the value where the SNR is equal to 1.

3.2.6 Texture rendition

Texture helps in identifying objects in images [10]. Texture thereby provides contextual information to the image. Many objects contain important texture elements. These elements enhance the recognition of the objects. Examples are hair, skin, textile, fabrics and foliage etc. When texture decreases, objects can begin to appear smooth, shiny, waxy, melted and/or blurry.

It should be noted that even if the texture decreases in a image, it is not necessary that the sharpness becomes affected. For instance, modern noise reduction algorithms have a tendency to blur out texture while still maintaining the appearance of sharpness in the image by keeping edges in the image intact.

3.3 The Human Visual System

In Engeldrum's Image Quality Circle, the *Human visual system (HVS)* and its subsystems correspond to visual algorithms, since they predict human perception values from physical image parameters. In order to understand the importance of device calibration when performing subjective image quality studies, some knowledge of the HVS is necessary.

The eye consists of several subcomponents [18]. The *cornea* is a transparent layer at the outer surface or the eye. The *lens* is located behind the cornea, posterior to the *iris* and the *pupil*. The pupil is in control of how much light that enters the eye. The iris consists of two muscles controlling the size of the pupil.

The *retina* is a neurosensory layer at the inner surface of the eye. The retina converts light into signals and further transmits these signals to the brain. The light that reaches the retina has passed through the cornea, the iris, the pupil, the lens and some additional layers. The *photoreceptors* (detecting light) at the very back of the retina exist in two varieties; *rods* and *cones*. The rods function in lowlight conditions and transmit the perception of contrast, brightness and motion. The cones function in bright light conditions and transmit color vision. The light level range where both rods and cones are active is when the lowlight condition and the bright light condition overlap.

Ganglion cells constitute the connection between the retina and the rest of the brain. The primary location for processing signals is the part of the brain called the *visual cortex*. The exact process of the visual cortex is yet to be revealed. Conclusions on its functions are dependent on the measurement technique used.

In psychophysics, section 3.5, the HVS is treated as a black box. By measuring task response, perception, some of its functionality can be understood.

3.3.1 Color and color spaces

The color processing in the HVS is complex. There are three types of cones [18], all sensitive to their own range of light wavelength, with different peak sensitives. Therefore, the human color vision is said to be *trichromatic*. A signal transmitted by a photoreceptor is proportional to the mix of wavelengths that are detected by it, weighted by the specific sensitivity curve of the receptor. The perception of color is based upon the comparison of signals produced by the three cones.

Color can be described with a vector with three elements, due to the trichromatic nature of the HVS. Through psychological experiments, the CIE XYZ color space has been defined to represent color vision on a basic level.

3.3.2 Contrast sensitivity function

The human eye does not interpret all spatial frequencies equally, in fact, it also treats detail information in color different from luminance. This is manifested in the human *Contrast Sensitivity Function (CSF)*. The ganglion cells are more sensitive to spatially varying patterns for certain spatial frequencies of the pattern. The response of ganglion cells can be described by using a visual stimulus consisting of a combination of many sine waves.

A threshold curve can be obtained, by recording the response to sine waves with different spatial frequencies. This curve is the Contrast Sensitivity Function, the CSF. The luminance CSF has a peak at a specific spatial frequency, which corresponds to the frequency where the ganglion cell is most sensitive. Contrast sensitivity applies to ganglion cells, as well as human psychological perception. Humans can not sense visual stimuli not detected by the ganglion cells.

3.4 Rendering

A characteristic display has one red, one green and one blue *primary*. The primaries can differ a lot from display to display, witch means that each display

has a different range of colors that it can display. RGB color spaces are thereby device-dependent. With two displays with different primaries, the same RGB values will be displayed differently. The CIE XYZ color space is based on a set of imaginary primaries, that are well defined and device-independent.

If wanting to compare displays, and thereby calibrating their colors, the display specific RGB values can be transformed to the CIE XYZ color space. The three primaries and the white point of a display are needed in order to construct a transformation matrix (3×3). The white point of a display is the color obtained by the maximum of all three primaries. It can be described as the color coordinates defining the color "white". Usually, the white point of displays is D65. The name D65 is a shorthand for "daylight 6500 Kelvin", and suggests that the color temperature should be 6500 Kelvin, but in reality it is closer to 6504 Kelvin.

The sRGB is a standardized color space. If two displays work in sRGB, they match each other and there should be no need for conversion between different device-dependent RGB color spaces. In the definition, the sRGB standard also uses three primaries and the D65 white point.

3.5 The history of Psychophysics

Psychophysics, "mind measuring", is the scientific study of the relation between *stimuli* (ϕ) in the physical domain and *sensations* (ψ) in the psychological domain [8]. In other words, it is a study of mental sensation, also called *sensory experience*, caused by some physical stimuli or stimulus. Subjective methods of measuring image quality, or anything else, are methods of psychophysics. The history of psychophysics will be presented in the following.

Thinkers have, for centuries, recognized the understanding of sensation as something important. Sensation is fundamental in understanding the human mind. Today, psychophysics is a central part of experimental psychology and subjective methods are important tools.

In 1860, the German philosopher and psychologist Gustav Fechner (1801-1887) published the book "Elements of Psychophysics" [8, 21]. His work contained theories and methods for the measurement of sensation. Psychophysics was founded. The German physician Ernst Heinrich Weber (1795-1878) had earlier, in the 1830's, obtained experimental results on sensation, mainly with weight as stimulus. Weber's results were of great inspiration for Fechner.

In 1879, the German philosopher and physician Wilhelm Wundt (1832-1920) created the fist lab in history designed only for experimental work on psychological processes. Experimental psychology was established as an independent science.

Weber, Fechner and Wundt were all influenced by the British empiricists. Empiricism states that knowledge comes from sensory experience. In the formation of ideas, sensory experience plays the central role. In science, empiricism emphasizes the role of evidence, especially revealed through experiments. Theories and hypotheses must be tested against observations of the natural world rather than by only relying on reasoning or intuition; empirical evidence is acquired by observation or experimentation [22]. Accordingly, the empiricist view of experience and evidence, the meaning of the concepts and their role in science, were picked up by the founders of psychophysics and experimental psychology.

3.5.1 Absolute and difference thresholds

In 1824 the German philosopher and psychologist Johann Friedrich Herbart (1776-1841) had apprehended the fact that mental sensations had to be stronger than some critical amount to be experienced or noticed. The term *sensory threshold* was introduced and later became a central theoretical concept in psychophysics [8]. Measurements were not a part of Herbart's description of the sensory threshold, but other scientists were going to pick up on the concept. That would be Weber and Fechner. They performed measurements of sensitivity limits of human senses.

Weber and Fechner experimentally measured sensory thresholds. Measurement techniques from physics were combined with human judgment by trained observers. The goal was to specify the weakest detectable sensation in relation to the stimulus intensity required to reproduce the sensation. The measurements succeeded and resulted in the definition of the *absolute threshold*. The absolute threshold was defined as the minimum amount of stimulus needed to produce a sensation. Both Weber and Fechner performed a large number of experiments and averaged the results to obtain accurate estimates.

The quantitative experiments of the sensitivity limits of human senses also resulted in another important type of sensory threshold. Weber and Fechner realized that when the stimulus intensity exceeded the absolute threshold, it had to be increased some critical amount to produce a change in the sensation. The *difference threshold* was defined as the amount of change in stimulus needed to produce a *just noticeable difference* in the sensation. The term just noticeable difference, *JND*, will be well examined later on, and the concept is of great importance of this thesis.

3.5.2 Weber's law

When performing his experiments on subjectivity in the 1830's, Weber discovered a relationship between the level of stimulus intensity and the difference threshold [8]. He mainly worked with weight as physical stimulus and noticed that a pair of two relatively heavy weights had to differ more in weight than a pair of two relatively light weights for a human to experience a difference when lifting them. Weber concluded that lighter weights were associated with a smaller difference threshold than heavier weights. More generally put; as the stimulus intensity (amount of weight in this case) increased, it took a greater change in stimulus intensity to change the sensation magnitude (the perceived heaviness) by some critical amount. After many experiments Weber concluded a relationship; the difference threshold was discovered to be a linear function of the intensity of the stimulus.

Theorem 3.1 Weber's law: *The change in stimulus intensity that can just be discriminated* ($\Delta\phi$) *is a constant fraction* (c) *of the starting intensity of the stimulus* (ϕ)*:*

$$\Delta\phi/\phi = c \tag{1}$$

or

$$\Delta\phi = c \times \phi \tag{2}$$

Weber's fraction, $\Delta\phi/\phi$, is expected to be constant at all levels of stimulus intensity. This is a confirmed result for a wide range of intensity levels [8]. At very low intensities however, Weber's fraction tends to increase substantially. More recent studies similar to the ones Weber performed were made by Köning and Boodhun in 1889 and Miller in 1947 [8]. Both showed results that $\Delta\phi/\phi$ first decreased as a function of ϕ and then became approximately constant. The original version of Webers's law did not hold at intensity values near the absolute threshold. A modification of Weber's law, more accurately corresponding to empirical data at all levels of stimulus intensity, is:

$$\Delta\phi = c \times (a + \phi) \tag{3}$$

In a modern interpretation, the value of a represents the amount of sensory noise (spontaneous activity in the nervous system) present in the neural system of a human when the value of the starting intensity of the stimulus is zero [8]. The value of a depends on the type of stimulus.

Introducing the concept of sensory noise in neural activity enables a better understanding of the thresholds. The absolute threshold can be seen as the value of the stimulus intensity needed to increase the neural activity level above the sensory noise level by some critical amount. The difference threshold can be regarded as the change in the stimulus intensity needed to produce a critical difference in neural activity levels.

Weber's law (with modification) provides a good explanation of most stimulus intensity data. Notable exceptions to mention are experiments made on discrimination of pure tones (by auditory sensation) and vibration (by tactile sensation) [8].

3.5.3 Fechner's law

Fechner, with background in physics and mathematics, approached the problem of measuring senses quite differently from scientists before him [8].

One of his early proposals was that an arithmetic series of sensation intensity values might correspond to a geometric series of physical stimulus intensities. Fechner later realized that this proposal was corresponding to what Weber had come up with through his experiments.

Fechner suggested that the sensation magnitude could be quantified indirectly by relating the values of the change in stimulus intensity to the corresponding values of the difference threshold, in sensation. He proposed that the psychological scale (quantifying sensation) could be related to the physical scale (quantifying stimuli) somehow. He assumed that difference thresholds, or just noticeable differences, were subjectively equal.

Fechner derived a general formula from Webers' law by integrating over the stimulus intensity.

Theorem 3.2 Fechners law: *The sensation magnitude* (ψ) *increases proportionally to the logarithm of the stimulus* (ϕ) *in units above the absolute threshold:*

$$\psi = k \times \log \phi \tag{4}$$

where k is a proportionality constant.

In absolute units, the law is $\psi = k \times \log \phi \ + c$, where c is a constant. Fechners law is often called Weber-Fechner's law. It is valid when Weber's law holds, i.e., at levels of stimulus intensity not too close to the absolute threshold.

3.6 Subjective image quality measurement methods

In Engeldrum's Image Quality Circle, subjective methods of measuring image quality ranges over observer quality perceptions, image quality models and observer quality ratings. There are two distinct ways of measuring image quality; *objective* measurements and *subjective* measurements [2]. Objective measurements involve physical measurements of defined image quality attributes. Objective measurements are however not always coincident with the experience of image quality perceived by humans. Subjective measurements involve human judgment of various aspects of image quality. These methods assign numerical values to perceptual attributes or to the perceived overall image quality. There are numerous ways to subjectively investigate image quality and to build subjective metrics. Later on, subjective measurement scales are going to be defined and a number of subjective methods are going to be presented. But first, perception and a commonly used unit in subjective measuring are going to be examined.

3.6.1 The statistical nature of perception

Perception or judgment can be seen as a random variable with a distribution described by all possible perceptions or judgments of the individuals in a population.

As shown by the Central Limit Theorem [32], the normal distribution describes many phenomena that depend on the sum of many independent events. Therefore, the normal distribution is a good candidate to use as a basis for quantifying perception or judgment. In image quality evaluation, the independent events are the human judgments of one image. The judgments of many humans added together can then be seen as normally distributed.

3.6.2 Just Noticeable Difference, JND

The term *Just Noticeable Difference (JND)*, has earlier been described historically in section 3.5.1.

Today, when building subjective metrics, the definition is based upon the outcome of a comparison of two stimuli. A 50 % JND is defined as the difference between two slightly different stimuli when 50 % of the observers perceive the difference and the other 50 % do not. The observers have to answer, so the 50 % who do not perceive the difference are guessing whether they perceive a difference or not. This results in, assuming that 50 % of the guessing observers say that they perceive the difference, 75 % positive responses. In this thesis the term JND will refer to this commonly used 50 % JND. The definition of a 50 % JND is:

Definition 3.1 *A 50 % JND is the stimulus difference that leads to a 75:25 proportion of positive and negative, respectively, responses in a paired comparison.*

When dealing with image quality, there are two kinds of JND units; *Attribute JNDs* and *Quality JNDs*. An Attribute JND is a measure of the perceived difference in the appearance of some particular attribute. A Quality JND is a measure of perceived difference in overall image quality.

A *JND increment* is associated with objective measurements:

Definition 3.2 *A* JND increment *is the number of units of an objective metric required to separate two stimuli by one JND.*

3.6.3 Cross-modal psychophysics

Comparisons and/or matchings are useful approaches when investigating image quality, making it possible to somehow rank or relate images to other images. Is it possible to compare and/or match different image quality attributes?

Comparing and matching seem to bee possible even for different senses. Cross-modal correlation between vision and audition has been well described, for example in [23], [24]. The relationship between vision and audition is extensively documented. In [25], subjective experiments made suggest the existence of robust correspondences between vision and olfaction as well. These studies, among others, state that there is a correlation between human perception of different senses.

In other words, it seems possible to somehow compare and/or match the magnitude of different senses. Thereby, the comparing and/or matching of different attributes (within the same sense) should be even "easier" for humans. According to [9], the success of observers matching image quality attributes (by vision) is remarkably good. This matching of different image quality attributes is of great importance for this thesis project.

3.6.4 Subjective measurement scales

In Engeldrum's Image Quality Circle, subjective measurement scales are referred to as observer quality ratings. The scale types [2] organizing the results of subjective measurements are explained in table 1:

Table 1: Classification of scale types

Scale type	Operations	Description	Examples
Nominal	Determination of *equality*	Labels or names assigned to organize individual elements	Names of colors, Numbers on soccer player shirts
Ordinal	Determination of *greater* or *less than*	Labels or numbers to represent *order* of individual elements	Rank order of favorite candy
Interval	Determination of equality of intervals or difference	Adds the property of *distance* to the Ordinal scale	Fahrenheit and Celsius temperature scales
Ratio	Determination of equality of ratios	Adds an *origin* to the distance property of the interval scale	Kelvin temperature scale (with absolute zero)

It is now time to approach the subjective methods. Different subjective methods have different ways of getting there, but they all have one common goal – to obtain quality or attribute judgments from observers. The judgments can then be represented in some type of scale.

3.6.5 Rank order methods

Rank order methods are having the observers ranking the images in order, from best to worst. If there are n images to be ranked, the rank goes from 1 to n. Rank order methods yield ordinal scales.

3.6.6 Category scaling methods

Category scaling methods are having the observers placing images in categories. The category names can be ”Bad”, ”Poor”, ”Fair”, ”Good”, ”Excellent”, or ”Imperceptible”, ”Perceptible but annoying”, ”Slightly annoying”, ”Annoying”, ”Very annoying” or ”Acceptable”, ”Not Acceptable” for example. There should be a minimum of two categories. A large number of categories allows observers to express themselves more, but too many categories can however be confusing. Category scaling usually requires a defined scenario to establish a context for the observer, since the categories can have different meanings in different contexts. Category scaling methods yield nominal or ordinal scales.

3.6.7 Paired-comparison methods

Paired-Comparison methods are having the observers comparing images in pairs to judge which of each image is preferred. If there are n images to be judged, this leads to a total of $n(n-1)/2$ comparisons – the total number of all possible combinations of n images shown two at a time. Results of the paired-comparison method are generally converted to an interval scale. This is usually done via Thurstone’s law of comparative judgment [2].

Theorem 3.3 Thurstone's law of comparative judgment:

$$\mathrm{S_A} - \mathrm{S_B} = \mathrm{z_{A-B}}\sqrt{\sigma^2_{A-B}} \tag{5}$$

with

$$\sigma^2_{A-B} = \sigma^2_A + \sigma^2_B - 2\rho\sigma_A\sigma_B$$

S_A and S_B are psychological scale values for stimuli A and B
$\mathrm{z_{A-B}}$ is the proportion of times that stimulus A is judged greater than stimulus B
σ^2_A and σ^2_B are the variances of the observers' responses of A and B
ρ is the correlation between the variances of the observers' responses of A and B

There is no known general solution to the equation, some simplifying assumptions about the unknown parameters have to be made. These assumptions are organized into five cases. Some cases are practically identical and therefore three case categories are presented here. Case category one requires knowledge on all parameters. Case category two makes the assumption that the correlation between the observers' responses are zero, $\rho = 0$. Case category three makes the same assumption as case category two, $\rho = 0$, and also that the variances of the observers' responses are equal, $\sigma^2_A = \sigma^2_B$. Case category three is the most applied one, since it enables the practical application.

3.6.8 Anchored scaling methods

Anchored scaling methods are types of the category scaling method with reference images, *anchors*, used as categories. With anchor images that have been calibrated, the method can yield both ratio and interval scales (depending on the kind of calibration). With anchor images that have not been calibrated, the methods yield nominal or ordinal scales. An example of an anchored scaling method is the *Softcopy Quality Ruler method* described in the third part of the ISO standard 20462 [27]. The softcopy quality ruler method will now be examined more in detail.

3.6.9 The softcopy quality ruler method

The ISO 20462 standard is titled *"Psychophysical experimental methods to estimate image quality"* [26]. The third (and last) part of it, titled *"The softcopy quality ruler method"*, describes a method that involves quality or attribute judgment by comparison of digital images to a series of ordered, univariate (in sharpness) digital reference images. The word "softcopy" refers to the fact that the images are digital and shown on a screen. The images to be judged are called the *test images* and the reference images (used to judge the test images with) are called *ruler images*. When assessing test images by matching them with ruler images, observers are performing an appearance matching. According to [9], the success of observers matching appearances is remarkably good.

General description
A *softcopy quality ruler* is a series of digital ruler images depicting the same scene, but varying in one single attribute. The quality differences between the ruler images are defined and known; they differ by known numbers of JNDs.

The ISO 20462 standard provides 21 softcopy quality rulers depicting different scenes, each with 31 ruler images varying in sharpness. The ruler images of one softcopy quality ruler are spaced with approximately one JND of quality. Sharpness is an appropriate reference attribute because:

1. It has a strong effect on image quality
2. It is easily varied by image processing
3. It exhibits relatively low variability between different scenes and observers
4. It is correlated with the MTF

Since the JND values of the ruler images are known, the softcopy quality ruler allows a numerical value to be associated with the test image immediately upon judgment by the observer. The test images can be scored by their quality directly in JND units.

The ISO 20462 standard defines a numerical scale; *The Standard Quality Scale, SQS*. The SQS is an absolute numerical scale of quality anchored against physical standards. It has one unit corresponding to one JND. A value of zero corresponds to a very bad image, so that the content is difficult to identify. The ruler images supplied, when displayed as specified in the standard, are calibrated against the SQS.

The ISO 20462 standard specifies and describes how to create softcopy quality rulers varying in other attributes than sharpness, if desired. Only one attribute shall change within a given ruler.

System MTF
In order for the ruler images and test images to be displayed on the screen in a predictable manner, the artifacts introduced by the capturing device (camera) and the screen must be discounted. This is achieved by performing a calibration of the imaging chain. The two most important factors to take into account are sharpness and color. In the case of sharpness, a characterization of the system MTF must be done.

The system MTF [28] associated with the softcopy quality ruler method is:

$$MTF_{System} = MTF_{Camera} \times MTF_{Prefilter} \times MTF_{Display} \tag{6}$$

The system MTF, the total MTF, is the product of the MTFs for the independent components. Here is an explanation of the components:

- In the creation of the Ruler images in the ISO 20462 standard, two different cameras were used; a Canon camera and a Kodak camera. The different cameras possess different camera MTFs, in units of cycles per millimeter(cy/mm).
- Prefilter means the processing of the image, in this case down sampling. The prefiltering was performed to prevent aliasing artifacts that might

be produced when decimating images. Different filters were used for the Canon and the Kodak camera, with different prefilter MTFs, in units of cycles per millimeter (cy/mm).

- The display used in the standard was an Apple 30 inch Cinema display. The display MTF is in units of cycles per millimeter (cy/mm) as well.

The system MTF is associated with a specific viewing distance (distance between the display and the eye of the observer). It is defined in units of cycles per degree (cy/degree) at the eye of the observer.

Calibration files
The ruler images supplied in the ISO 20462 standard are calibrated against the SQS. There are calibration files in the standard describing the translation between ruler levels (corresponding to ruler images, indexed 0 to 30) and SQS values. With the calibration files, the test images can be scored in a standardized way. The calibration files are valid when the ruler images are displayed as specified in the standard. They have been created based on the system MTF.

There are calibration files for three viewing distances (distance between the display and the eye of the observer): 25, 34, and 43 inches for both the Canon camera and the the Kodak camera.

Creation of ruler images
The ruler images supplied in the ISO 20462 standard are created using the MTF of the complete imaging system (capture, image processing and display). The system MTF shall be characterized by measurement of a resolution bar, see 3.2.1 and/or based upon linear system theory, see A.1, appendix A, and B.1, appendix B. An aim MTF (that mimics the response of a monochromatic, on-axis diffraction-limited lens) is defined as:

$$MTF(v) = \frac{2}{\pi}\left((\cos(kv))^{-1} - kv\sqrt{1-(kv)^2}\right) \tag{7}$$

v is spatial frequency (in cycle per degree) at the eye of the observer
k is constant

The creation of the ruler images is made in three steps [28]:

1. A spatial kernel is created to modify the system MTF to approximate the aim MTF. After the spatial filtering, the system MTF should conform adequately to the aim MTF.

2. The value of the constant k (that fits the aim MTF) is used to compute relative JNDs of quality. The JND values represent SQS values for a high quality image with great color and tone reproduction, average scene content and absence of significant artifacts and/or noise.

3. The spatial kernel is applied to an image captured by a specific camera. A series of ruler images are created for a series of k values.

4. The spatial kernel is applied to an image captured by a specific camera. A series of ruler images are created for a series of k values.

3.6.10 Experienced and unexperienced observers

Observers evaluating image quality can be divided into two groups; experienced and unexperienced observers[2]. Experienced observers have experience in judging or evaluating images and can distinguish among categories of specific attributes to a much greater degree than unexperienced observers.

When judging overall image quality, experienced observers seem to weight various attributes differently, and thereby focus on different aspects, as compared to unexperienced observers [5]. Experienced observers seem to be more influenced by perceived sharpness; they are more annoyed by sharpness loss, and thereby base their judgment more on overall image quality on sharpness than unexperienced observes. Unexperienced observers seem to be more influenced by perceived colorfulness; they are correlating their scores on overall image quality much more with colorfulness than experienced observers do. This difference between experienced and unexperienced observers was also shown by [37]. The study was based on the analysis of eye movements during subjective image quality evaluation. The study showed that both experienced and unexperienced observers first looked to the colorfulness of the object in the image. Second, unexperienced observers looked at high-brightness areas in the images, whereas the experienced observers rather looked at regions containing details and contour.

3.7 Evaluation of data - frequentist statistics

In frequentist statistics, conclusions based on the frequency of the data can be drawn from the data. The data y is seen as an observation of a random variable Y. The random variable Y has some distribution, denoted $\mathbb{P}$. Accordingly, an estimate $t(y)$ is an observation of the random variable $t(Y)$. If a study is repeated, the data y, and thereby the estimate $t(y)$, may well obtain different values.

3.7.1 Uncertainty of estimators

If $\tau = \tau(\mathbb{P})$ is some property, called the estimand, of the distribution $\mathbb{P}$, the error associated with an estimate $\Delta(y) = t(y) - \tau$ is an observation of the random variable $\Delta(Y) = t(Y) - \tau$. In order to evaluate the uncertainty of the estimator, the distribution function of the error $\Delta(Y)$ can be analyzed. The error distribution is denoted $\mathbb{F}$.

Assuming that the error distribution $\mathbb{F}$ is known, the confidence interval, $(L(y), U(y))$ on confidence level α for the estimand τ, see section C.1, appendix B , is:

$$I = \big(t(y) - \mathbb{F}^{-1}(1-\alpha/2), t(y) - \mathbb{F}^{-1}(\alpha/2)\big) \tag{8}$$

For example, if $y = (y_1, \ldots, y_n)$ are observations of n independent variables distributed as $\mathcal{N}(\mu, \sigma)$, the normal confidence interval on confidence level α is:

$$I = \left(t(y) - \lambda_{\alpha/2}\frac{\sigma}{n}, t(y) + \lambda_{\alpha/2}\frac{\sigma}{n}\right) \tag{9}$$

where the value of the quantile $\lambda_{\alpha/2}$ can be found in a tabular. As mentioned in section 3.6.1, the normal distribution is a good candidate to use as a basis for quantifying perception or judgment.

3.7.2 Bootstrapping

Bootstrapping is a method of estimating properties of estimates $t(y)$ [30]. The procedure is to draw samples from an approximated distribution. If the unknown distribution of the estimand τ is $\mathbb{P}$, the approximated distribution is denoted $\hat{\mathbb{P}}$.

The non-parametric bootstrap method
In the non-parametric bootstrap method, no assumptions are made on $\mathbb{P}$, apart from the samples being independent and identically distributed random variables. $\mathbb{P}$ is not assumed to belong to a certain parametric family. The approximated distribution, $\hat{\mathbb{P}}$, is the empirical distribution function.

The empirical distribution gives equal weight, $\frac{1}{n}$ to each of the data points in the data set. Therefore, if X is a random variable with the distribution $\hat{\mathbb{P}}$, X takes the value of one certain data point with the probability $\frac{1}{n}$, where n is the number of data points. Therefore, the simulation of $\hat{\mathbb{P}}$ is carried out by drawing, with replacement, from the data set. These are the steps of the algorithm:

1. Simulate the empirical distribution $\hat{\mathbb{P}}$ from the data set.
2. Simulate a large number of new data sets, estimators, by drawing, with replacement, n times among the data points.
3. For all simulated data sets, compute the desired property.
4. For all simulated data sets, take the difference of the desired property of the original data set and the desired property of the estimator. The values obtained are then approximately distributed according to the error distribution. These values can be used for evaluation of uncertainty.

These are some of the cases when the bootstrap method can be useful [31]:

- If the distribution of a statistic is complicated or unknown.
- If the sample size is small and it is desired to draw conclusions from the obtained data. Even if the distribution is known, distortions can appear if the sample is not fully representative of the whole data set.

4 Method

4.1 Background aspects

The goal of the thesis project is to set up a working environment for performing subjective image quality studies, and to carry out one larger study covering one particular image quality problem. To accomplish this, a number of initial decisions are made. This is available:

1. A room reserved for subjective testing
2. A high quality color monitor (EIZO ColorEdge CG276)
3. A software tool, a GUI (Graphic User Interface), for displaying and judging images as well as recording results

Questions that need answers are:

(a) What image quality problem is to be investigated in the study?

(b) How is the study going to be performed?

(c) How is the Visual lab going to be designed?

(d) Who is going to participate in the study?

(e) When is the study going to be performed?

(f) How is the evaluation of the performance of the Visual lab and the results going to be performed?

The first step is to explore the literature on subjective methods for evaluating image quality; what methods there are, how they work, how they are used, how effective they are and how their results turns out, etc. Simultaneously, the subject of image quality is investigated through the literature as well. All the following decisions are made in consideration of the present and future needs of Axis Communications AB.

4.2 The image quality problem

The image quality problem to be investigated is how noise and texture loss in combination affected the perceived image quality. When reducing noise in images by image processing, there is a trade off between the amount of noise reduction and texture loss. The question that is interesting in this case is if there exists an optimal combination of parameters from which image quality can be maximized for different situations.

4.3 The softcopy quality ruler method

The subjective method chosen for this study is the softcopy quality ruler method. The reasons for this choice are:

- The method is time effective (and thereby cost effective)
- The method provides real-time results (which is one of the reasons why it is time effective)
- The method is beneficial when judging images not closely spanned in quality (thanks to the wide quality range of the ruler images)
- The results of the method are in units of SQS values, JNDs (the results are standardized and thereby comparable to standardized results from other studies)

4.4 The Visual lab

The room reserved for the subjective testing is a normal sized office space. The ISO 20462 standard [27] provides a description of the appropriate setup of the physical apparatus for performing softcopy quality ruler tests. A paper describing the implementation on the softcopy quality ruler [28] give directions on how to apply the standard to the setup of the Visual lab.

4.4.1 Lab environment

The room is painted in neutral, approximately 18 % gray, color in order to reduce observer fatigue and to avoid possible distractions. Grey curtains, as close as possible in color to the walls, are hung up to cover the windows and the glass door of the room for the same reasons as well as to eliminate incoming light.

According to ISO 20462, the following physical apparatus shall be included in the lab in order to perform a Softcopy Quality Ruler test:

1. A display with necessary hardware to display images and a keypad or mouse for data entry.
2. A lighting system for controlling the surrounding illumination.
3. A headrest or other device to constrain the viewing distance (from the observer's eye to the display)

An EIZO ColorEdge CG276 display is bought for the Visual lab. The hardware is a standard computer having both a keypad and a mouse. The display is placed on a neutral, approximately 18 %, gray desk. In the standard, the creation of the ruler images was made on an Apple 30 inch Cinema display. To account for the different display of the Visual lab, a modification of the viewing distance is made. More on this in section 4.4.3.

To meet the second criterion of the standard, four flourescent tubes with a color temperature of approximately 6500 Kelvin ("Lysrör Fullfärg de Luxe 36 W/965 Dagsljus Special"[1]) from the manufacturer Philips are mounted standing in pairs in armatures on the gray desk. Four variable density tube guards ("American PLAS-100100"[2]) from the manufacturer American Made Plastics, are placed around the tubes. All this is in accordance with the implementation paper [28]. The luminance from the tubes can be controlled by the variable density tube guards so that the wall behind the display can be fairly uniformly illuminated. Adjusting the tube guards properly, the luminance on the wall behind the display can be similar to that of an average pictorial scene rendered on the display. The ceiling lamp should be turned off when performing tests in the Visual lab; with no light source directly illuminating the display the dynamic range of the images displayed can be higher. The light sources, in combination with the gray surround, can reduce observer fatigue compared to the case of viewing images with a dark surround [28]. The wish to minimize structural artifacts and the ability to calibrate the results of the study are the reasons for the headrest criterion [28]. The calibration of the results will be examined in section 4.4.3.

Instead of using a headrest, an additional desk is put in front of the desk with the display on it. The position of the observer heads can then be adjusted so that their eyes will be on the same distance from the display as the edge of the additional desk. Having something to refer to makes it easier for the observers to hold the position. A head rest would be more precise, but likely less comfortable for the observers. In figures 5 and 6 the Visual lab is shown.

[1] http://ljusbutiken.nu/produkt.php?art=8136965
[2] http://www.1000bulbs.com/product/247/PLAS-100100.html?tid=pacc

Figure 5: The visual lab: view of the display from the observers position

Figure 6: The visual lab: side view from the lab setup

4.4.2 The software tool

The software tool, the GUI (Graphic User Interface), for displaying and judging images and for recording judgments was prepared by a summer worker in 2013. It was implemented in Python. Python is a frequently used language, which is why it was chosen. Some minor modifications are made in order to meet the needs of the study. Figure 7 shows a screen dump of the main window of the program.

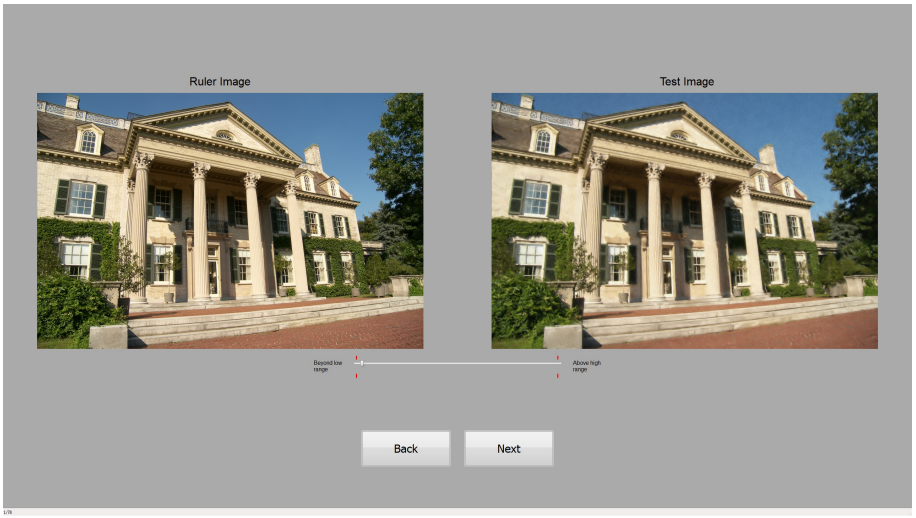


Figure 7: The GUI of the Visual lab

The ruler image is varied by moving the slider bar. There are a total number of 33 ruler positions. The first one (to the left) is labeled ”Beyond low range”. If an observer judges the test image to be of higher quality than all ruler images, i.e., out of scale of the quality ruler, that is where the slider bar is placed. The following 31 positions correspond to the 31 ruler images of the scene, ranged from best to worst. The last position (to the right) is labeled ”Above high range”. If an observer judges the test image to be of lower quality than all ruler images, that is where the slider bar is placed. When a test image has been judged (by a slider bar position), the ”Next”-button is pressed and the next test image in order is presented. If wanting to re-evaluate a test image, the ”Back”-button can be pressed.

4.4.3 Calibration of the display

A calibration of the EIZO display is performed, to enable standardized scoring with SQS values. In the following, the procedure is described:

Distance calibration
In order for the calibration files provided in the ISO 20462 standard to be valid, the ruler images have to be displayed as specified in the standard. An Apple 30 inch Cinema display is used in the ISO 20462 standard when creating the ruler images and the calibration files. The display of the Visual lab is an EIZO ColorEdge CG276. The difference in display MTF (and thereby system MTF) has to be compensated for.

The display MTF of the Apple display has been quantitatively measured [28]. A single display pixel was lit in the center of the display with a black background. A calibrated camera was used to capture an image of the illuminated pixel. The image was magnified such that about 870 sensor pixels of the camera imaged the single display pixel, giving very good resolution of structure. A Fourier transformation was performed on the point spread function of the image to

obtain both horizontal and vertical MTFs. The display MTF was derived by taking the ratio of the measured MTF to the known camera MTF. The measured horizontal and vertical MTFs are displayed in a graph in [28].

The values, see table 2 are read in the graph and inserted in Matlab for future comparisons. By averaging the horizontal and vertical MTF of the display, a single display MTF is obtained.

Table 2: Modulation transfer values for the Vertical MTF and the Horizontal MTF

Spatial frequency	Vertical modulation	Horizontal modulation
0	1	1
0.5	0.98	0.98
1	0.92	0.91
1.5	0.79	0.79
2	0.63	0.55
2.5	0.45	0.4
3	0.33	0.25
3.5	0.22	0.15
4	0.09	0.07
4.5	0.02	0.05
5	0.05	0.08
5.5	0.09	0.1
6	0.11	0.09
6.5	0.12	0.07
7	0.1	0.05
7.5	0.09	0.02
8	0.07	0.04
8.5	0.06	0.06
9	0.03	0.08
9.5	0.03	0.09
10	0.02	0.08

There is no information about the MTF of the EIZO display. The procedure described above can be applied on the EIZO display as well. However, it is rather time consuming and requires special equipment not available for this study. The manufacturer may have some knowledge about the MTF, but for different reasons it is often difficult to obtain the MTF directly from them. In the absence of information about the MTF of the EIZO display, the MTF can be derived theoretically by making an assumption about its pixels. The theoretical MTF of the Apple display can be derived as well, in which case the sanity of this approach can be checked by comparing it to its measured MTF. The best assumption is that the displays can be described with the same MTF by assuming that both displays have ”perfect” pixels.

A perfect pixel is perfectly square. The pixel pitch of a display is the distance from the center of one pixel to the center to the next pixel. The pixel pitch of the Apple display is 0.25 mm in both directions. The pixel pitch of the EIZO

display is 0.2331 mm. Perfect pixels have no space between them, meaning that the pixel pitch corresponds to the width (and height) of a pixel.

Now, the MTF of a display with perfect pixels is going to be derived. Assume there is a perfect pixel. Denote the pixel pitch by L. The area of the perfect pixel is then $L \times L$. The PSF of the perfect pixel is a "box" described by:

$$PSF(x,y) = \begin{cases} 0, & (|x|,|y|) > L/2 \\ 1/L^2, & (|x|,|y|) < L/2; \end{cases} \tag{10}$$

The PSF is normalized, i.e the integral of it is 1. The LSF of the perfect pixel is the integral in one direction (horizontal or vertical) of the PSF:

$$LSF(x) = \begin{cases} 0, & |x| > L/2 \\ 1/L, & |x| < L/2; \end{cases} \tag{11}$$

The MTF is the the magnitude of the Fourier transform of the LSF. The MTF of the perfect pixel is:

$$MTF(k) = \left| \int_{-L/2}^{L/2} \frac{1}{L} e^{-2\pi i x k} dx \right| = \left| \frac{\sin(\pi k L)}{(\pi k L)} \right| = |\operatorname{sinc}(kL)| \tag{12}$$

For the Apple display, the pixel pitch is 0.25 mm

$$MTF_{Apple}(k) = |\operatorname{sinc}(0.25 \times k)| \tag{13}$$

The first zero will be at spatial frequency = 1/0.25 mm = 4 cy/mm (cycles per millimeter), the second one at k = 2/0.25 mm = 8 cy/mm, etc.

For the EIZO display, the pixel pitch is 0.2331 mm

$$MTF_{EIZO}(k) = |\operatorname{sinc}(0.2331 \times k)| \tag{14}$$

The first zero will be at spatial frequency k = 1/0.2331 mm = 4.29 cy/mm, the second one at k = 2/0.2331 mm = 8.58 cy/mm etc.

With the assumption that both displays have perfect pixels, the Display MTFs scale with the size of the pixel pitch:

$$MTF_{Eizo}(k) = MTF_{Apple}(k \times \frac{0.2331}{0.25}) = \left| \operatorname{sinc}(0.25 \times k \times \frac{0.2331}{0.25}) \right| = |\operatorname{sinc}(0.2331 \times k)| \tag{15}$$

Comparing the values of the measured MTF of the Apple display with its approximated MTF, the conclusion is that the assumption made on its pixels is appropriate. See figure 8. The measured MTF and the theoretical MTF are quite similar.

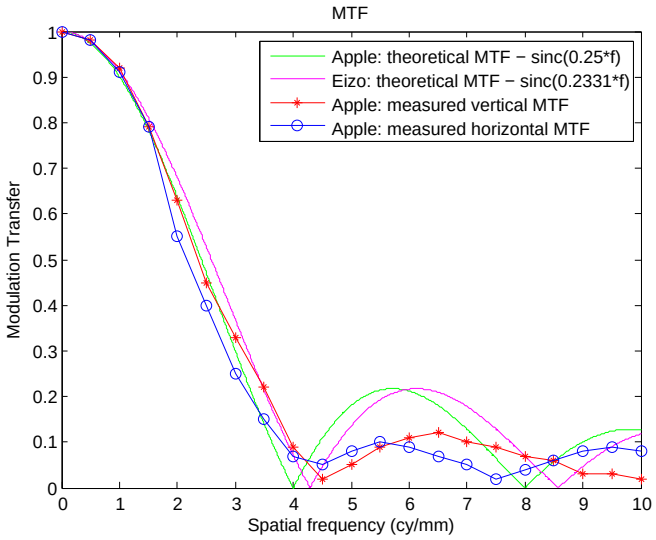


Figure 8: Display MTFs; theoretical MTFs for Apple and EIZO and measured horizontal and vertical MTF for Apple

By a geometric reasoning, the size of a pixel at the eye of an observer is proportional to the viewing distance (from the observers eye to the display). With known pixel pitches of the two displays and with the assumption that both displays can be described with the same MTF, the calibration files from the ISO 20462 standard can be used for the EIZO display, but for different viewing distance. The viewing distances scale with the sizes of the pixel pitches. With the EIZO display the calibration file for 25 inches should be used at a distance of $25 \times \frac{0.2331}{0.25} = 23.3$ inch and the file for 34 inches should be used at $34 \times \frac{0.2331}{0.25} = 31.7$ inch etc.

Color calibration
A color calibration of the EIZO display is performed, in order to match the gray color on the screen with the illuminated gray wall behind the screen. The white point of the display is set at 6500 Kelvin, which gives a good visual match with the back wall illuminated by the fluorescent tubes.

4.4.4 Ruler images

Three softcopy quality rulers provided in the ISO 20462 standard are used as reference scenes. One softcopy quality ruler depicts one scene. The goal is to choose three scenes that are as different as possible in color, object, etc. Three is considered a good number, explained in section 4.5. The test images are created from the ruler images, as described in section 4.5.

For a study with a different purpose, the reasoning on the ruler images would possibly be different. If investigating the difference in perception between different scenes for example, a choice of more scenes would be more reasonable.

If investigating test images not generated from the quality rulers, there would probably be a need to create new quality rulers, associated with the test images. The three scenes chosen are shown in figures 9, 10 and 11.

Figure 9: George Eastman house scene

Figure 10: Grass people scene

Figure 11: Snow scene

4.5 Test images

The test images are created from the ruler images. According to [34], it is advantageous to use ruler images as the starting point in the creation of test images, so that the image content of the test images and ruler images is matched, making the evaluation easier.

Number of test images
Since two attributes are going to be investigated, the number of test images per scene has to be relatively high. The task of judging images with the softcopy quality ruler method may be experienced as rather tedious. In order to avoid observer fatigue resulting in less reliable judgments, the total number of test images should not be too high. The following facts motivates the chosen total number of test images:

- Two attributes, noise and texture, are going to be investigated in the study
- According to ISO 20462, a minimum of 3 ruler scenes should be used
- The duration an average observer spends on each judgment is approximately 15 seconds [28].

25 test images per quality ruler scene is considered enough to cover the noise - texture loss space. The minimum number of 3 scenes is chosen, resulting in a total number of 75 test images. The importance of the alertness and focus of the observers is the higher priority than a larger number of test images.

According to [28], with 25 test images per scene, one scene would be judged in approximately $25 \times 15/60 = 6.25$ minutes. With 3 scenes, this would result in 18.75 effective minutes per observer spent on judging. That amount of time seems reasonable.

Creation of test images
The test images of one scene are created in two steps. First noise is added and second noise is reduced with a software tool, resulting in texture loss. The same procedure is repeated for all three scenes.

Test images should not be too similar or differ too much; they should neither be judged equal (or close to equal) or judged to be out of the Quality Ruler scale. Too similar test images (within the scale) will give little information on preference. Too different test images will give little information as well. If judged out of the quality ruler scale, no SQS values are assigned to the test images.

The first step is now going to be examined in detail. Five levels of Gaussian noise is added to ruler image number three. According to [34], it is recommended to use the second or third highest ruler image as a starting point when creating test images. Gaussian noise, generated in Matlab, is added in equal amount to each of the three color channels. Different levels of noise, with zero mean, correspond to different standard deviations. The Gaussian noise is chosen because:

- It is controllable, by the size of the mean (set to zero) and the standard deviation; the noise level steps can be equally spaced.
- It is easy to generate (Matlab was used)
- It is well used in theory. It may not be a very realistic noise in images and image processing, but the purpose of the study was to investigate the effect of noise on image quality from a theoretical point of view.

The magnitudes of the standard deviations are chosen by an iterative process. The first standard deviation value is chosen simply by guessing a value, looking at the noisy image, adding noise with a different value, looking at the new image etc. until a satisfactory value of the first standard deviation value is found. In the same manner, the step size of the standard deviation value is initially guessed, and the iterative process of looking and changing values is repeated. Five test images with equally spaced noise levels are generated, all with quality within the range the 31 ruler images, spanning a satisfactory range of quality themselves. The chosen standard deviations are {10 20 30 40 50{ in units of digital levels. These levels are labeled {1 2 3 4 5}.

The second step is now going to be examined in detail. The added noise of the five test images is reduced in four steps, using the software tool Neat Image [38]. Unfortunately, there is no public information on how Neat Image operates on images. The algorithm is confidential. The reasons Neat Image is chosen as an appropriate noise reduction program for the purpose of the study are:

- Noise can be reduced in percent and thereby the step of noise reduction amount can be equally spaced in a controlled way.
- It is a modern, well known noise reduction tool with properties comparable to that of other such tools with respect to how texture detail is traded off with noise level.
- It is a well known and widely used tool, open for everyone to buy.

The iterative process described above was repeated when choosing the noise reduction levels. The %-values of noise reduction chosen were {20 40 60 80{. This resulted in $5 \times 4 = 20$ additional test images with different amounts of texture loss. The more noise that is reduced, the more the texture of the image is lost.

The total 25 test images represent 25 different degrees of quality loss. Combining the noise level vector with the noise reduction vector, the quality loss matrix can be described as:

$$\begin{pmatrix} \{1,80\} & \{2,80\} & \{3,80\} & \{4,80\} & \{5,80\} \\ \{1,60\} & \{2,60\} & \{3,60\} & \{4,60\} & \{5,60\} \\ \{1,40\} & \{2,40\} & \{3,40\} & \{4,40\} & \{5,40\} \\ \{1,20\} & \{2,20\} & \{3,20\} & \{4,20\} & \{5,20\} \\ \{1,00\} & \{2,00\} & \{3,00\} & \{4,00\} & \{5,00\} \end{pmatrix} \tag{16}$$

The same two step process was repeated for all three scenes and $25 \times 3 = 75$ different test images were obtained.

(a) Test image "{1,80}" (b) Test image "{3,80}" (c) Test image "{5,80}"

(d) Test image "{1,40}" (e) Test image "{3,40}" (f) Test image "{5,40}"

(g) Test image "{1,0}" (h) Test image "{3,0}" (i) Test image "{5,0}"

Figure 12: 9 test images (out of 25) of the George Eastman scene

A closer look at some of the test images is displayed in figure 13.

(a) Test image "{1,80}"

(b) Test image "{5,80}"

(c) Test image "{3,40}"

(d) Test image "{5,0}"

Figure 13: 4 test images (out of 25) of the George Eastman scene

4.6 Observers

4.6.1 Number of observers

According to the ISO 20462 standard, a minimum of 10 observers shall be used, but preferably 20. The number of observers chosen for the pre-studies and the large study will be motivated later on.

Visual acuity test
Observers have to have normal visual acuity at the specified viewing distance [33] 31.7 inches. Before participating, the observers are tested to make sure that this is the case. The eye test card is taken from [35], see figure 14. The requirement chosen for the study is that the observer has to be able to read the row marked "32 inches" with maximum one error at a viewing distance of 29 inches. The observers are allowed to wear whatever visual aids they normally wear [33].

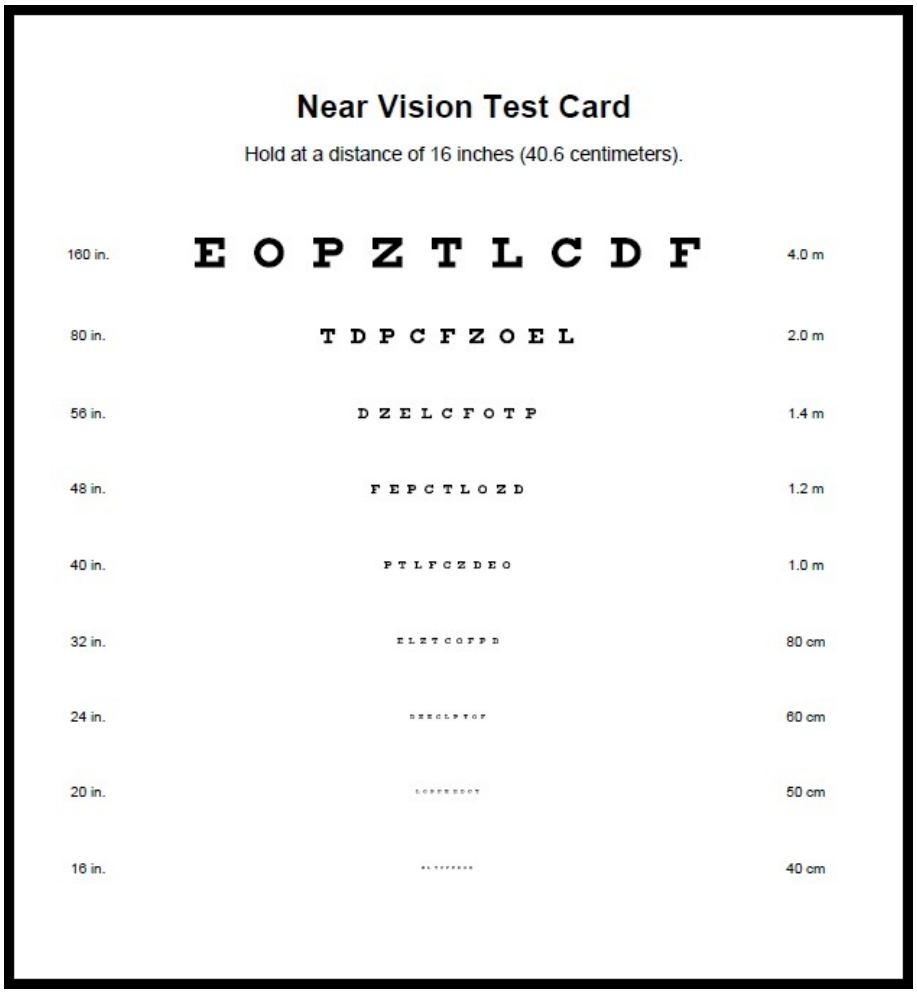


Figure 14: The near vision test card [35]

Color vision test

For any color related study, the observers have to be tested for color vision as well [33]. For the noise reduction - texture loss study, it would not have been mandatory. For future studies to be performed in the Visual lab it however might be. Therefore, for consistency, a color vision test is performed. A test chart book for color deficiency is used for the color vision test [36]. This book consists of several Ishihara test patterns designed to evaluate various aspects of human color vision. Since the purpose in this case is to find out if a person could be considered to have normal color vision or not, a smaller subset of the images in the book is chosen. The requirement chosen is that the observer has to be able to distinguish three objects in the book (a kangaroo, an elephant and a machine gun (AK 47)). Those are objects that a person with normal color vision should be able to see, but that would have been difficult to identify for a person with any of the most typical color deficiencies.

4.6.2 Experienced vs unexperienced observers

Observers can, offhandedly, be divided into two groups; experienced and unexperienced observers, see section 3.6.10. Both experienced and unexperienced observers are chosen to participate in the study.

4.7 The outline for participation in the study

Taking into account the time consumption of the visual acuity test, the color vision test, the reading of the instructions and the ”test session” added to the approximately 18.75 minutes of effective judging, a supported guess is that the large study to be performed will take somewhere between 30 and 45 minutes per observer. There is no time limit set, observers are themselves to choose how long time they spend on each judgment. The outline for one observer participating is the following:

1. The observer arrives in the Visual lab at the scheduled time and gets welcomed by the instructor.
2. The observer takes the visual acuity test as well as the color vision test. (The tests should be passed in order to participate.)
3. The instructor makes sure the observer finds a comfortable position in front of the screen with eyes at the correct viewing distance. The instructor points out the importance of holding the position through the whole study.
4. The observer reads the instruction text presented on the screen and the instructor answers any questions that may come up.
5. The observer takes a ”trial session”, performing a few trial judgments. In the trial, no judgments are recorded. The instructor makes sure that the observer has understood the task and feels comfortable with the method.
6. The observer starts the actual study where the judgments are recorded and the instructor leaves the Visual lab. One scene is judged at the time.
7. When the observer has judged all Tests images, the study is over and a short text thanking the observer is presented on the screen. The observer can leave the Visual lab.

The instructions to the observers presented on the screen are:

> Welcome to the Visual Lab!
>
> You will now participate in a test evaluating image quality. It is important that you judge the OVERALL QUALITY of the WHOLE image - not some particular attribute or a particular area of the image.
>
> A pair of images will be presented on the monitor. The image on the left is labeled "Ruler image" and the image on the right is labelled "Test image". The sharpness of the ruler image is varied by moving the slider bar. For each test image on the right, you are supposed to adjust the ruler image on the left so that there is a match in the overall quality between the two images.
>
> There are 78 test images for you to judge, varying in noise and/or texture detail. Your response will be recorded when you press the "Next" button. If you want to re-evaluate a previous test image that is not currently on the screen, press the "Back" button.
>
> Before the actual test, there will be a training session. Play with the slider to get used to how it works. When you are ready to start the actual test, press "Exit Trial" and the test starts.
>
> Please remember that there are no right or wrong answers - image quality is defined by observer perception, and the purpose of this test is to find out what YOUR perception is of these test images. Good luck! :)

4.8 Preparatory pre-studies

Two pre-studies are carried out before the large study. This is done in order to "warm up" the Visual lab, to verify that it is functioning properly, and to make adjustments and/or changes in the procedure if necessary.

4.8.1 Initial study: Weber-Fechner's law

When the Visual lab is set up, a small initial test is performed in order to "warm up" the lab and to verify that it is appropriately running. This is done by investigating Weber-Fechner's law. The law states that the sensation magnitude (ψ) increases proportionally to the logarithm of the stimulus (ϕ) in units above the absolute threshold $\psi = k \times \log \phi$. In absolute units, the law becomes $\psi = k \times \log \phi \ + c$, where k and c are constants.

The stimulus used is noise only and the sensation magnitudes are the judgments, SQS values recorded with the softcopy quality ruler method.

Preparation
Twenty test images are created solely for the initial study. Twenty levels of Gaussian noise is added to Ruler image number 3 of one chosen scene, see figure 15, according to the same procedure described in section 4.5. The chosen values of the standard deviations are {3 6 9 12 15 18 21 24 27 30 33 36 39 42 45 48 51 54 57 60}.

Figure 15: Farm stand scene

Observers
Four observers are participating in the initial study. Four is not a sufficient number of observers for the softcopy quality ruler method 4.6.1, but the initial study is only an informal sanity check of the performance of the Visual lab. The intention is to avoid using plenty of time and/or resources.

Performance
The study outline described above is almost followed, with the exception being that the instruction text was not written at that time. The instructor instead describes the task to the observers. In order to ensure that the observer-screen distance is hold, the observers are monitored, discreetly, during the test.

4.8.2 Pilot study

Before performing the large study, another pre-study simulating the large study, called the pilot study, is run for several reasons:

- To find out if the test images were appropriate, neither too similar nor too different from each other and within the quality range of the Softcopy Quality Rulers.
- To find out if the observers find the number of test images reasonable
- To informally verify if it seems if it is suitable to perform the study without a head rest for controlling the viewing distance

- To be prepared on what kinds of questions that the observers might have
- For the instructor to become comfortable with everything concerning the Visual lab and the procedure of the study
- To spot any inconsistencies or errors in the experimental setup.

Preparation
The test images are created according to section 4.5.

Observers
The observers of the pilot study are people from the department where this thesis project is performed; the Core Technologies Imaging department (CTI) at Axis Communications AB. The group is chosen for convenience; the people are familiar with the study and the Visual lab is located at the department floor. The goal is to involve at least 10 observers 4.6.1. Invitations are sent out to the employees (totally 25 persons) and everyone who wants to is welcome to participate, resulting in a number of 17 observers.

Performance
All observers passed the visual acuity test and everyone but one passed the color vision test. A note is taken on the observer who did not pass the color vision test and then the observer is allowed to participate. The study outline described above was followed.

4.9 The large study

Having performed the pilot study, it is time to carry out the large study.

Preparation
The same test images as for the pilot study are used. The only difference from the pilot study is that three ruler images, "null images" , are added to the test images, see explanation in section 6.3. Ruler image number three is chosen as null image for each scene.

Observers
30 observers are invited. A relatively large number (with the minimum number being 10 4.6.1), but preferable since the softcopy quality ruler method is going to be evaluated. When inviting observers, the goal is to attain diversity in gender, age, imaging experience etc. Since the Visual lab is located at Axis, the observers need access to the company premises. Visitors have to be registered in the reception and picked up by an employee in order to enter the house, which is a time demanding process. Therefore, it is natural to almost only invite employees. 28 Axis employees and 2 visitors are participating in the study.

Performance
All observers passed the visual acuity test and the color vision test. The study outline described above is followed, the only difference being that there are 78 test images to judge instead of 75.

4.10 Data analysis

The observer judgments of the test images are modeled as random variables Y_i, where $i = 1, \ldots, 75$. The judgments of one test image are assumed to be independent since the observers are assumed not to influence each other. The n judgments for test image number one, for example, are thereby observations $y_{1,1}, \ldots, y_{1,n}$) of the random variable Y_1.

The main purpose of the data analysis is to investigate the estimators $t(Y_i)$, where $i = 1, \ldots, 75$ being the means of the judgments. In order to assess the uncertainty of the estimators, the errors $\Delta(Y_i)$ are be analyzed. As mentioned in section 3.6.1, a good candidate for quantifying perception or judgment is the normal distribution. The average of the judgments can be approximated with a normal distribution. This is investigated by comparing bootstrap and normal confidence intervals for the judgments.

1. Bootstrap confidence intervals were calculated, with the only assumption that the judgments were independent and identically distributed.

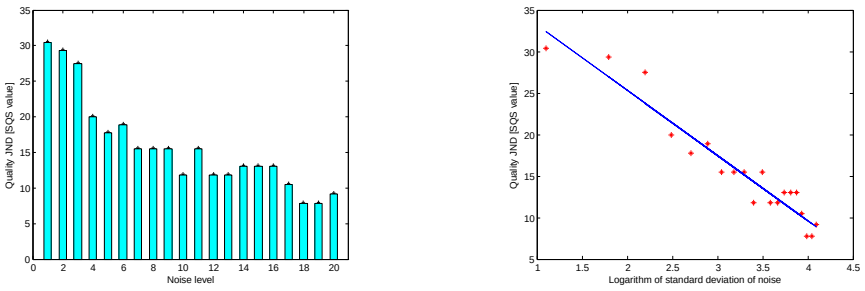
2. Assuming independent normally distributed judgments, normal confidence intervals were calculated as well; $Y_1, \ldots, Y_{75}$ were assumed to be distributed as $\mathcal{N}(\mu_1, \sigma_1), \ldots, \mathcal{N}(\mu_1, \sigma_{75})$

The confidence intervals are calculated for all test images, for every scene as well as for the mean of the scenes and for every study with a confidence level of 95 % , i.e., $\alpha = 0.05$. The normal approximation is thereby tested using the non-parametric bootstrap method.

5 Results

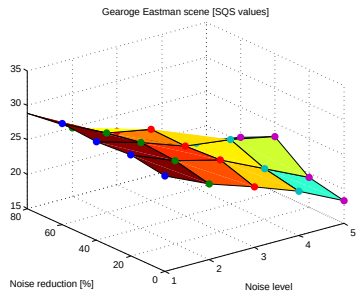
5.1 Initial study: Weber-Fechner's law

In figure 16, the results of the initial study are shown. The averaged judgments of the four observers are plotted against the standard deviation of the noise in figure 16 (a). The averaged judgments of the four observers are plotted against the logarithm of the standard deviation of the noise as well as a line fitted to the data in figure 16 (b).

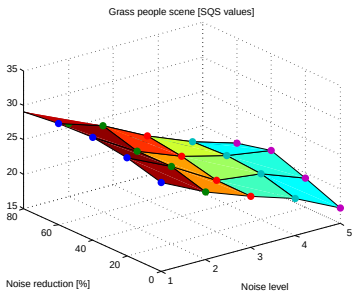


(a) Perceived quality (b) Weber-Fechner's law

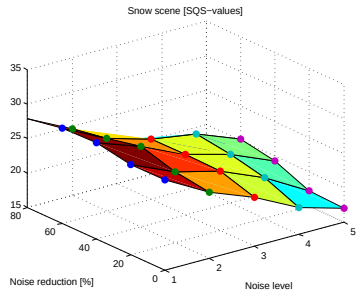
Figure 16: Initial study: Weber-Fechner's law



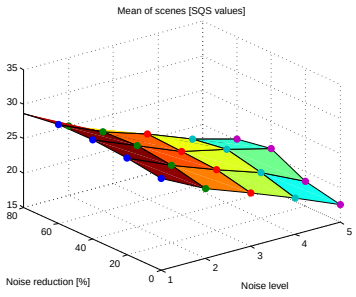
(a) George Eastman scene



(b) Grass people scene



(c) Snow scene



(d) Mean of scenes

Figure 17: The pilot study: the three scenes and the mean of the scenes

The polynomial of degree one was fitted to the data with the Matlab commands POLYFIT and POLYVAL, optimizing the fit in a least squares sense. The polynomial is: $\psi = -7.86 \times \log \phi + 41.0$. A goodness of fit of the polynomial was made. The coefficient of determination, the R^2-value [3], was 0.926.

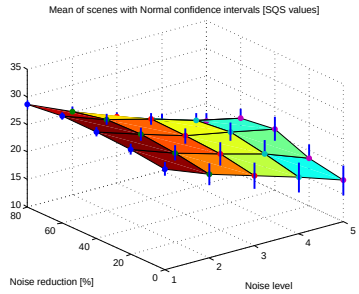
5.2 The pilot study

In figure 17, the results of the pilot study are shown. The averaged judgments of the 17 observers are plotted for each scene as well as for the mean of the scenes.

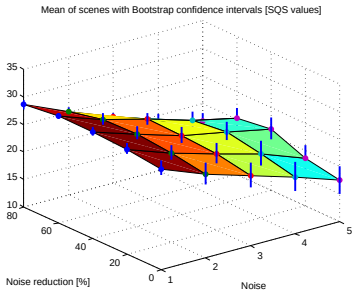
In figure 18, the size of the two sided normal and bootstrap confidence intervals (the difference between the upper and lower bound) are shown for the averaged judgments (of the 17 observers) for the mean of the scenes. The confidence level was set at 95 % , i.e., $\alpha = 0.05$.

The mean value and maximum value of the normal confidence intervals are 3.13 JNDs and 5.49 JNDs, respectively. The mean value and maximum value of the bootstrap confidence intervals are 2.99 JNDs and 5.70 JNDs, respectively. The difference between the estimates of confidence intervals are small; assuming the normal approximation therefore seem justifiable.

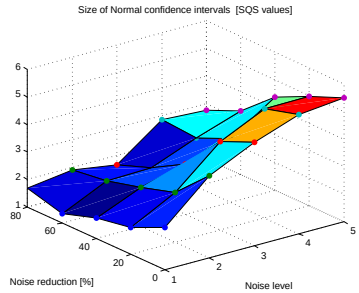
[3] The R^2-value, the *Coefficient of determination* [29], indicates how well data fits a statistical model. It ranges from 0 to 1, where a value of 1 indicates that the model fits the data perfectly and a value of 0 implies the opposite.



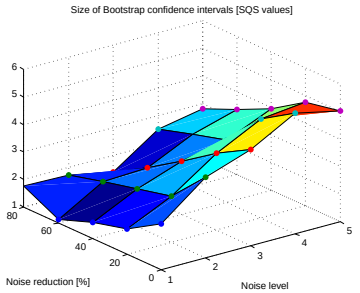
(a) Mean of scenes and normal confidence intervals



(b) Mean of scenes and bootstrap confidence intervals



(c) Size normal confidence intervals



(d) Size bootstrap confidence intervals

Figure 18: The pilot study: size of confidence intervals for the mean of the scenes

The observers were discretely observed to verify if they could hold the viewing distance without a headrest, and that seemed to be the case.

5.3 The large study

In table 3, the judgments of the null images are shown. The third ruler image is the correct response for all three scenes. The table shows the deviation from the correct response. Judging a null image as ruler image 2 gives a deviation of 1 JND (ruler images of one softcopy quality ruler are spaced with approximately one JND of quality), judging a null image as ruler image 6 gives a deviation of 3 JND etc. The standard deviation is calculated as:

$$\sqrt{\frac{1}{3-1}\sum_{i=1}^{3}(3-x_i)^2} \tag{17}$$

where x_i is the observed value for one of the three scenes. The standard deviation is calculated in order to be consistent with the results from [12], see section 6.4. Data from three observers out of the 30 were removed in the large study, see section 6.4, resulting in a total number of 27 observers.

In figure 19, the results of the large study are shown. The averaged judgments of the 27 observers are plotted for each scene as well as for the mean of the scenes.
In figure 20, the size of the two sided normal and bootstrap confidence intervals (the difference between the upper and lower bound) are shown for the averaged judgments (of the 27 observers) for the mean of the scenes. The confidence level was set at 95 % , i.e., $\alpha = 0.05$.
The mean value and maximum value of the normal confidence intervals are 2.25 JNDs and 4.31 JNDs, respectively. The mean value and maximum value of the bootstrap confidence intervals are 2.13 JNDs and 4.13 JNDs, respectively. Also here, the normal approximation gives a good estimate of the confidence intervals.

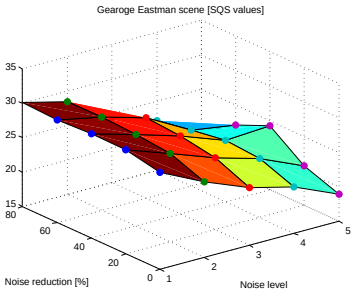
5.4 The total study

The results of the pilot study can be added to the results of the large study, since the two studies were identical except from the null images. That difference was considered negligible considering the observer performances. Therefore, the results from a total number of 47 observers could be obtained. Together, the two studies are called the total study. Since three observers were removed from the data, a total number of 44 observers were included in the total study.

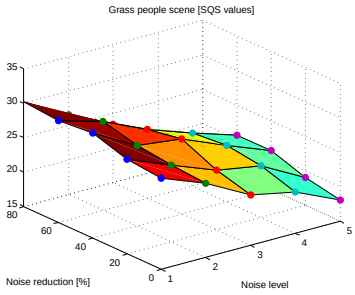
In figure 21, the results of the large study are shown. The averaged judgments of the 44 observers are plotted for each scene as well as for the mean of the scenes.
In figure 22, contour plots corresponding to figure 21 (d) are displayed.
In figure 20, the size of the two sided normal and bootstrap confidence intervals (the difference between the upper and lower bound) are shown for the averaged judgments (of the 44 observers) for the mean of the scenes. The confidence level was set as 95 % , i.e., $\alpha = 0.05$.
The mean value and maximum value of the normal confidence intervals are 1.83 JNDs and 3.37 JNDs, respectively. The mean value and maximum value of the

Table 3: Deviation from correct response (0 indicates a correct response) and standard deviation of all scenes

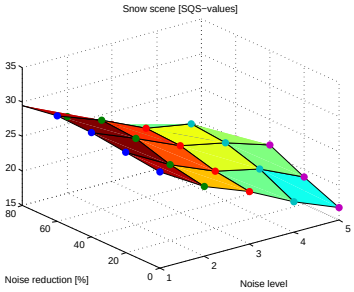
Observer	George Eastman	Grass people	Snow	Standard deviation
1	1	1	1	1.22
2	**3**	**1**	**3**	**3.08**
3	1	1	2	1.73
4	1	1	0	1
5	0	2	2	2
6	0	0	3	2.12
7	1	1	3	2.34
8	**4**	**1**	**4**	**4.06**
9	1	0	0	0.71
10	2	1	2	2.12
11	2	1	0	1.58
12	0	2	0	1.41
13	0	2	3	2.56
14	2	1	1	1.41
15	2	1	2	2.12
16	**2**	**2**	**8**	**6**
17	1	3	2	2.66
18	0	2	2	2
19	2	3	2	2.92
20	1	2	2	2.12
21	1	2	1	1.73
22	0	0	2	1.41
23	1	1	0	1.22
24	2	3	3	3.32
25	2	2	2	2.45
26	1	1	1	1.22
27	0	1	0	0.71
28	0	1	1	1
29	2	0	2	1.73
30	1	1	4	3



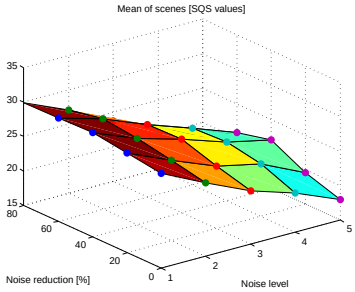
(a) George Eastman scene



(b) Grass people scene

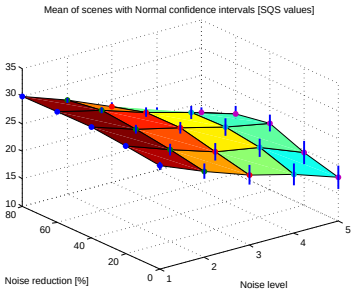


(c) Snow scene

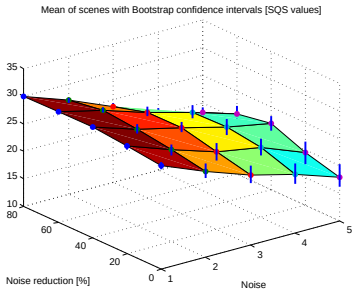


(d) Mean of scenes

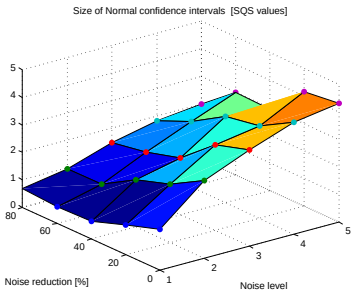
Figure 19: The large study: the three scenes and the mean of the scenes



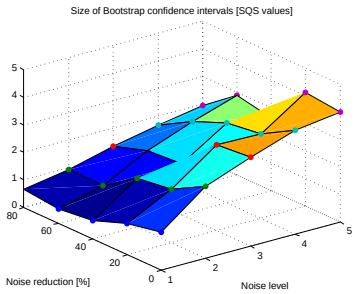
(a) Mean of scenes and normal confidence intervals



(b) Mean of scenes and bootstrap confidence intervals

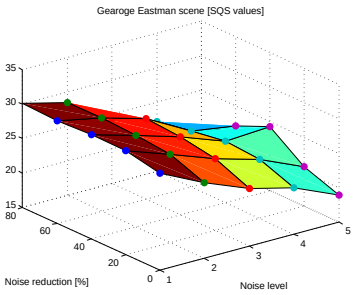


(c) Size of normal confidence intervals

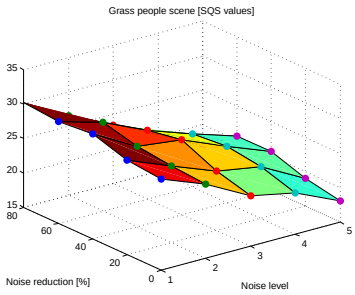


(d) Size of bootstrap confidence intervals

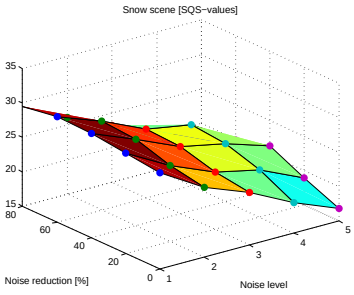
Figure 20: The large study: size of confidence intervals for the mean of the scenes



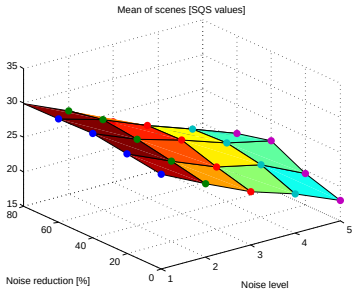
(a) George Eastman scene



(b) Grass people scene



(c) Snow scene



(d) Mean of scenes

Figure 21: The total study: the three scenes and the mean of the scenes

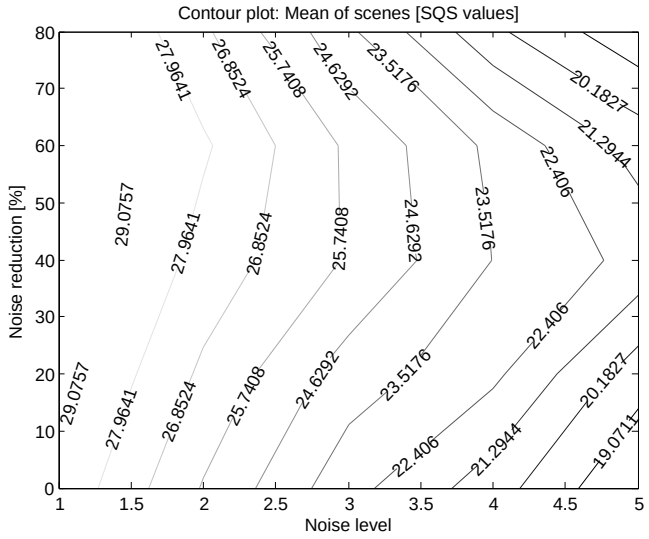
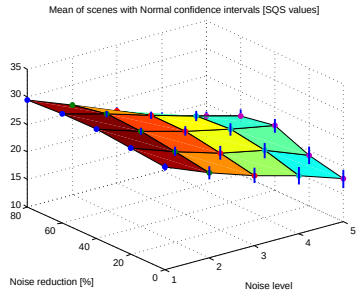
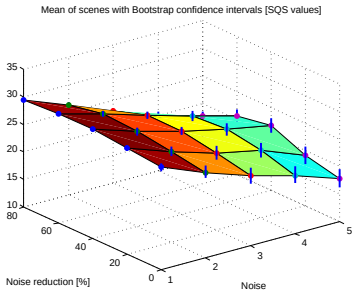


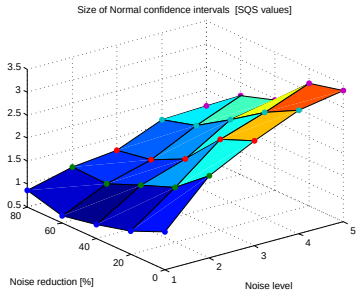
Figure 22: Mean of scenes



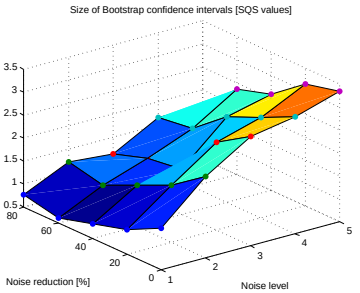
(a) Mean of scenes and normal confidence intervals



(b) Mean of scenes and bootstrap confidence intervals



(c) Size normal confidence intervals



(d) Size bootstrap confidence intervals

Figure 23: The total study: size of confidence intervals for the mean of the scenes

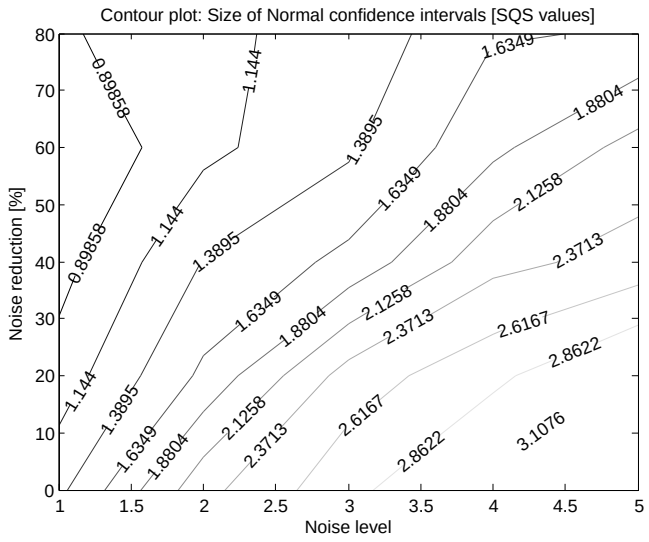


Figure 24: Size normal confidence intervals

bootstrap confidence intervals are 1.78 JNDs and 3.40 JNDs, respectively. The ranges of the normal and bootstrap confidence intervals are similar; the normal distribution (tested using the non-parametric bootstrap method) seems to well describe the data.
In figures 24 and 25, contour plots corresponding to figures 23 (c) and 23 (d) are displayed.
Recall the quality loss matrix 16 describing 25 different degrees of quality loss that represent the 25 test images:

$$\begin{pmatrix} \{1,80\} & \{2,80\} & \{3,80\} & \{4,80\} & \{5,80\} \\ \{1,60\} & \{2,60\} & \{3,60\} & \{4,60\} & \{5,60\} \\ \{1,40\} & \{2,40\} & \{3,40\} & \{4,40\} & \{5,40\} \\ \{1,20\} & \{2,20\} & \{3,20\} & \{4,20\} & \{5,20\} \\ \{1,00\} & \{2,00\} & \{3,00\} & \{4,00\} & \{5,00\} \end{pmatrix} \tag{18}$$

As mentioned, the normal distribution seems to describe the data well, and in figure 26, the size of the two sided normal confidence intervals for 9 test images are plotted as functions of the number of observers (the mean of 30 random permutations of choosing observers).
In figure 27, the size of the normal confidence intervals are plotted as a function of the averaged judgments of the observers for the mean of the scenes.
Since the background of the observers (in terms of position) were known, 11 experienced observers (imaging engineers, camera desig as well as 11 unexperienced observers (Financial controllers, Human resources employees etc.) were picked out and put in two groups. The difference between these two groups are shown in figures 28.
In figures 29 and 30, contour plots corresponding to figure 28 are displayed.

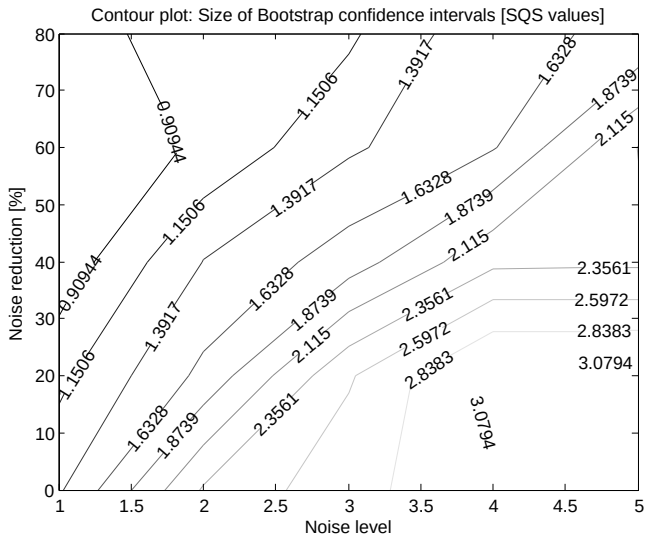


Figure 25: Size bootstrap confidence intervals

In order to investigate the significance of the difference between the two groups, a statistical hypothesis test, Welch's t-test, in which the test statistic follows a student's t-distribution if the null hypothesis is supported, was made for the mean of the scenes for every test image. As mentioned, the normal distribution seems to describe the data well, and therefore it is appropriate to perform this test assuming normal distributed data:

1. The data (per test image) from the experienced observers are 11 observations drawn from a normal distribution $\mathcal{N}(\mu_e, \sigma_e)$.

2. The data (per test image) from the unexperienced observers are 11 observations drawn from a normal distribution $\mathcal{N}(\mu_{ue}, \sigma_{ue})$.

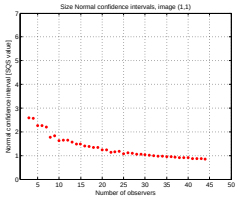
3. The two datasets are independent.

Therefore, $\hat{\mu}_e - \hat{\mu}_{ue}$ are observations drawn from $\mathcal{N}\left(\mu_e - \mu_{ue}, \sqrt{\frac{\sigma_e}{11} + \frac{\sigma_{ue}}{11}}\right)$. The hypothesis are:

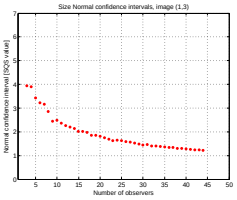
$$H_0 : \mu_e = \mu_{ue}$$
$$H_1 : \mu_e > \mu_{ue}$$

In words:

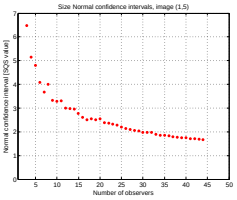
H_0 : "There is no difference between the means of the judgments of the two groups."
H_1 : "The mean of the judgments of the experienced observers is higher then the mean of the judgments of the unexperienced observers."



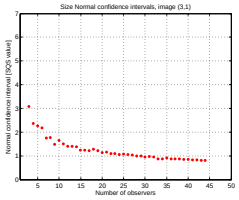
(a) Test image "{1,80}"



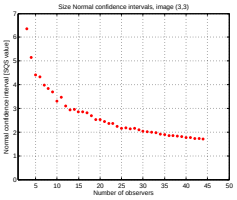
(b) Test image "{3,80}"



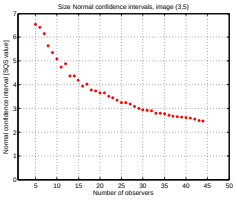
(c) Test image "{5,80}"



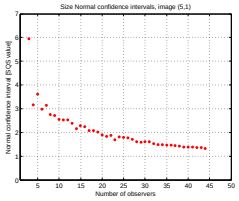
(d) Test image "{1,40}"



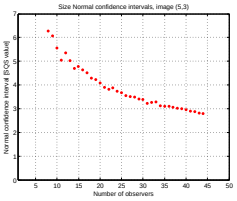
(e) Test image "{3,40}"



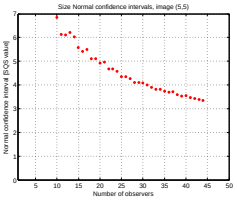
(f) Test image "{5,40}"



(g) Test image "{1,0}"



(h) Test image "{3,0}"



(i) Test image "{5,0}"

Figure 26: Size of normal confidence intervals as a function of number of observers for the mean of the scenes

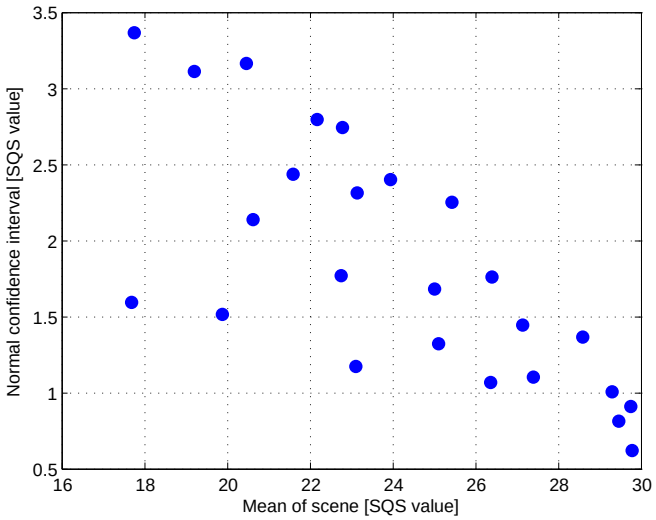
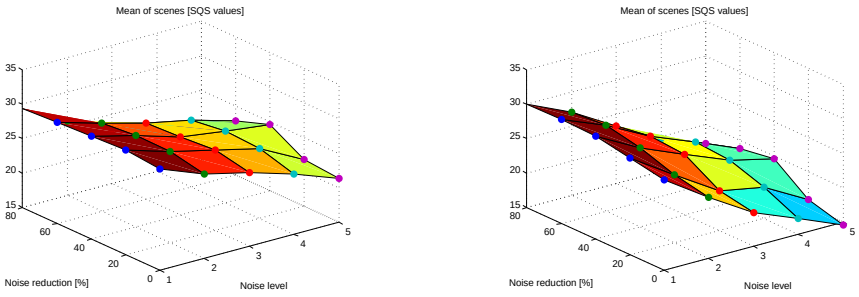


Figure 27: Normal confidence interval size dependency on overall judgment mean values for 47 observers



(a) Experienced observers

(b) Unexperienced observers

Figure 28: Experienced vs unexperienced observers

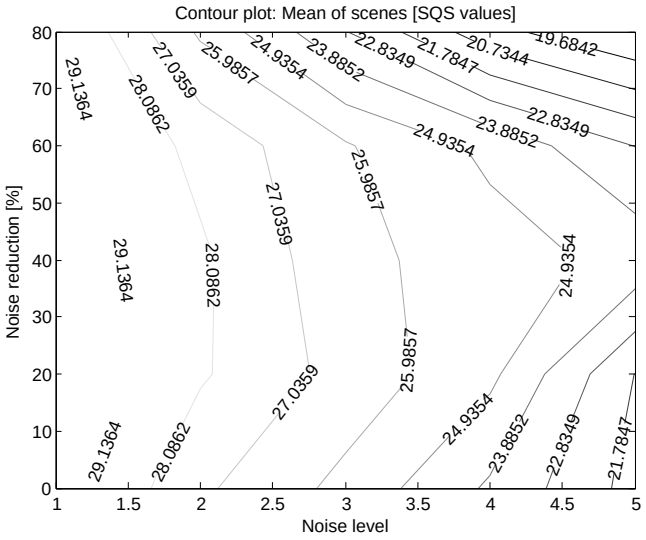


Figure 29: Experienced observers

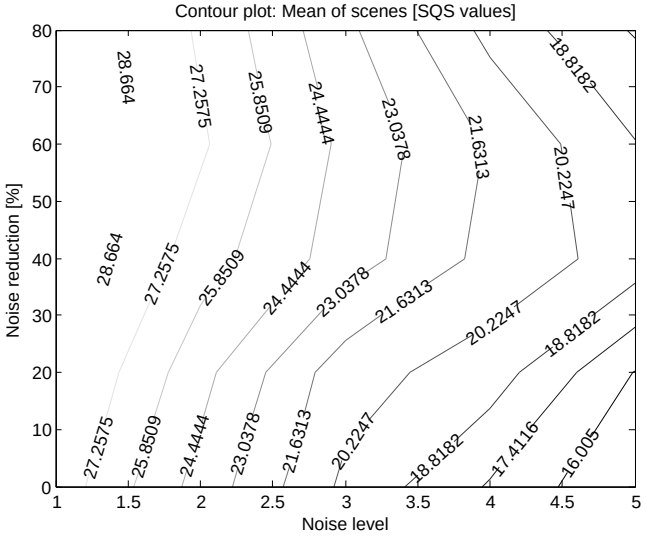


Figure 30: Unexperienced observers

The test statistic is:

$$T = \frac{\hat{\mu}_e - \hat{\mu}_{ue} - 0}{\sqrt{\frac{\hat{\sigma}_e}{11} + \frac{\hat{\sigma}_{ue}}{11}}} \tag{19}$$

H_0 is rejected when $T > t_\alpha(f)$, with $\alpha = 0.05$ and $f = 20$ (degrees of freedom). The 25 T-values, for the mean of the scenes for every test image (the elements of the matrix represents the test images named according to 16) are:

$$\begin{pmatrix} -1.31 & -1.71 & -0.523 & 0.190 & 1.24 \\ -0.936 & 0.404 & 2.33 & 2.42 & 2.61 \\ -0.0856 & 2.12 & 2.06 & 2.63 & 2.92 \\ 1.67 & 3.04 & 3.4270 & 2.91 & 2.53 \\ 1.96 & 2.10 & 2.9715 & 2.83 & 2.72 \end{pmatrix} \tag{20}$$

The quantile is $t_{0,05}(20) = 1.72$ (from student's t-distribution). Therefore, with a hypothesis test on level $\alpha = 0.05$, the null hypothesis can be rejected for images marked "Y" below:

$$\begin{pmatrix} -N & -N & -N & N & N \\ -N & N & Y & Y & Y \\ -N & Y & Y & Y & Y \\ N & Y & Y & Y & Y \\ Y & Y & Y & Y & Y \end{pmatrix} \tag{21}$$

I.e., the images marked "Y" are judged as significantly better by the experienced observers. The images marked "N" are not significantly judged as better by the experienced observers. The images marked "– N" are judged as better by the unexperienced observers, however not with significance.

6 Conclusions

As mentioned, the data from the pilot study and the large study were added to one study, called the total study. Now, the three separate studies; the initial study in section 6.2, the pilot study in section 6.3 and the large study in section 6.4 will be discussed separately in that given order. Then, the total study will be discussed in section 6.5. That is where the main discussion is held and it is also there the main conclusions are drawn. Finally, a general discussion on the visual lab and the softcopy quality ruler method will be held. But first of all, the distribution of the judgments will be discussed.

6.1 Comparing bootstrap and Normal confidence intervals

The normal approximation was tested using the non-parametric bootstrap method, and the conclusion is that the data can be well described by the normal distribution. As seen in figures 18, 20, and 23, normal and the bootstrap confidence intervals are similar in size, for all studies. The more observers that are taken into account, the more similar the bootstrap and normal confidence intervals become. This is not a surprising result, considering the central limit theorem. Therefore, the conclusion drawn is:

Conclusion: *The judgments of the softcopy quality ruler can be said to be normally distributed, also for a smaller amount of observers*

6.2 The initial study

The data from the initial study could be well described by Weber-Fechner's law, $\psi = k \times \log \phi \ + c$. The fitted linear equation $\psi = -7.86 \times \log \phi + 41.0$ gives a R^2-value of 0.926. This comparably high value indicated that the Visual lab was set up in a proper way and that the methodology was reasonable. Therefore, it was possible to proceed with the pilot study.

6.3 The pilot study

The following conclusions can be drawn from the pilot study:

- The test images were appropriately chosen, since the judgments spanned a wide range of quality
- The observers found the number of test images reasonable, but the general opinion was that more had been too many
- The observers were of the opinion that they had been able to hold the viewing distance through the test. They had been discreetly observed by the instructor as well, which indicated the same thing.
- The overall response of the observers was positive. Many had enjoyed the study, and expressed that they learned something about their personal opinion on noise and texture loss.

The one observer who did not pass the color vision test did not show notably different judging behavior when looking at the data; the plot of SQS values of the different scenes as well as the mean of the scenes did not appear different from the averaged values of all observer. Also, the averaged values did not change when excluding the observer's data, and would not have changed the conclusions made in the study. Because of that, no data were removed. No further conclusions were drawn on the color vision test topic.

When the pilot study was carried out, an idea on detecting the observers attentiveness came up, thanks to a paper [12] that was read at that point. Adding additional test images corresponding to ruler images, "null images", would give an indicator on observer performance. The observer responses of the null images, the "null responses", would indicate how consistently observers can match ruler images to ruler images. Other than adding null images, no other modifications were considered needed. A great advantage, resulting in a possession of data from 17 observers to be added to the large study.

6.4 The large study

The null responses indicated that most of the observers made "consistent" judgments, see table 3. In [12], data from two observers out of twenty (10 %) were removed from the analysis of a softcopy quality ruler experiment "because the standard deviation of their null responses was significantly larger than 2.5

JNDs, noticeably separating their task performance from that of the other participants."

In the large study, six observers out of thirty (20 %) gave null responses with standard deviations larger than 2.5 JNDs; 3.08, 4.06, 2.56, 2.66, 2.92 and 3.32 JNDs. Only three of these observer null responses were considered significantly larger than 2.5 JNDs. 3.08, 4.06 and 3.32 JNDs. Therefore, data from the three observers out of the thirty (10 %) were removed; observer number 2, 8 and 16 (boldfaced in the tabular).

Since the autos of [12] does not define how much larger (than 2.5 JND) "significantly larger", no futher conclusions could be drawn on how our study performed to their. We removed the same number of observers (10 %), but we had a total number of six observers with null responses larger than 2.5 JNDs.

Three null images might have been too few. Images from the middle as well as the end of the soft copy quality ruler should have been added as well, in order to make further conclusions. As mentioned, the idea on adding null images came up close to the starting day of the large study, so there was unfortunately not much time spent on them. However, including them and analyzing the responses of them was a good lesson for future work, an initial start on the ability to detect less alert observers in the future.

6.5 The total study

6.5.1 The mean of the judgments of different test images

Seen in figures 17, 19 and 21, the means of the judgments were showing a regularity, indicating that the study and the method used were consistent. By looking in the noise level direction, it can be seen that the five curves were rather straight lines with negative slopes. The higher the noise level, the lower the SQS value. By looking in the noise reduction direction, it can be seen that the five curves were rather shaped as second order polynomials with maxima around the 40 % level of noise reduction. The higher the noise level, the more noise reduction is performed by Neat Image. The slopes of the curves increased for higher noise levels. The question if there exists an optimal combination of parameters from which image quality can be maximized for different situations, can be answered by considering the shape of the curves discussed as well as the contour plot in figure 22 in the following conclusion:

Conclusion: *In the study performed in this thesis project, the optimal noise reduction level seems to be 40 % for all five noise levels, even if the noise reduction was rather unnecessary for the one or two first noise levels. Even so, the noise reduction algorithm does not seem to unnecessarily damage the image at these low noise levels. The other noise reduction levels; 0 %, 20 % and 80 %, resulted in lower SQS values than for the 40 % level.*

6.5.2 The variance of the judgments of different test images

Seen in figures 18, 20, 21 and 23, the variances of judgments were largest in images with a high amount of noise and less noise reduction. This can be seen in the contour plots 24 and 25 as well. As mentioned, noise reduction algorithms tend to blur out texture when reducing noise. By looking at the test images,

the impaired texture rendition appears more similar to sharpness loss than the added noise does. The fact that sharpness was the reference attribute in the study is probably therefore the explanation of why the confidence intervals were largest in the images with most noise. It seems to be easier to match images that differ less in type of quality loss.

As seen in figure 27, the sizes of the normal confidence intervals can not be said to inherit any obvious dependency on the overall mean values of the judgments. However, the figure shows that large mean values give small confidence intervals and that the spread in the confidence interval sizes becomes larger when the mean is smaller. In this study, the test images are generated from a high quality ruler image (number three, with SQS value around 30), which probably is the explanation of that behavior of the plot. The less processed the test image, the more similar to the (high quality) ruler images and thereby easier to judge, resulting in large mean values and more consistent judgements.

In figure 26 one can see how the confidence intervals decrease differently for different test images. Both the "offsets" and the shapes of the curves are dependent on the test image, whereby it is hard to draw any further conclusions on a general curve describing the variance in terms of number of observers. However, the figure does reinforce the conclusion that the variances of the judgments depend on the appearance of the test images (in relation to the quality ruler). By looking in the noise level direction (from left to right), the offset of the confidence intervals increases. By looking in the noise reduction direction (from bottom to top), the offset of the confidence intervals decreases. Also noticed is the difference in observer dependence. By looking in the noise level direction (from left to right), the curves show a weaker dependence on the number of observers. By looking in the noise reduction direction (from bottom to top), the curves show a stronger dependence on the number of observers. This discussion can be summarized in the following conclusion:

Conclusion: *When using the softcopy quality ruler method for evaluating image quality, the variances of the judgments in this study seem to depend considerably on the perceived similarity of the ruler images and the test images. If the ruler images appear similar to the test images, the variances of the judgments will be lower than for less similar ruler images and test images. Therefore, if it is important to minimize the number of observers, the choice of attribute used for the ruler images should be carefully considered.*

Matching the attribute of the quality ruler with the attribute to be investigated in a test has been pointed out in a previous study [4]. In this case it was noted that it could be difficult to use the softcopy quality ruler (varying in sharpness) of the ISO standard for validating other attributes (texture in this case). The paper does, however, not provide any experiments or investigations to corroborate that assumption, and no other study discussing this issue has been found. This thesis project does indicate that the "closer" in apperance the attribute to be judged is to sharpness, the less the variance of the judgements seem to become.

If wanting to investigate the attribute sharpness, or one or several attributes that appear similar to sharpness, trivially sharpness is a good reference attribute and the softcopy quality rulers of the ISO standard should be used. Sharpness

is also a good reference attribute if wanting to evaluate overall image quality in the same study, explained in 3.6.9. It should not be forgotten that it is perfectly possible to use sharpness as reference attribute when judging different attributes as well. In the studies performed in this thesis project, both noise and texture loss could well be evaluated, only with a variation in variances of the judgments.

Acceptable variance
According to the ISO 20462 standard, a minimum of 10 observers shall be used, but preferably 20. In this study, in the "worst case" – figure (i) in 26 – 10 observers would give a confidence interval of size 7 JNDs, [μ - 3.5, μ + 3.5], and 20 observers would give a confidence interval of size 5 JNDs, [μ - 2.5, μ + 2.5].

Considering that 1 JND is just about discernible, 3.5 JNDs could be considered reasonable in most cases, but for more critical work 2.5 JNDs might be the tolerable limit. Therefore, this study can agree with the ISO 20462 standard on the number of observers that shall be used in a softcopy quality ruler experiment.

6.5.3 Experienced vs unexperienced observers

In figures 28, 29 and 30, the difference in judgments between the two groups can be seen. Intuitively, it seems that the experienced observers (imaging engineers, camera desig are judging the test images as "better" than the unexperienced observers do. In matrix 21, it can be seen that the differences in judgments are significant, on a 95 % confidence, for 9 out of the 25 test images. All five images with only noise added were significantly judged as better by the experienced observers, while the five images with the maximum amount of noise reduction (80 %) were not significantly judged as different. The following conclusion can be drawn:

Conclusion: *When judging overall image quality, experienced observers working with cameras (imaging engineers, camera desig are more tolerant to noise than unexperienced observers*

According to the discussion in section 3.6.10, experienced observers are more negatively influenced by sharpness than unexperienced observers. This conclusion can be reinforced by this study.

In the figures and the matrix, it can also be seen that the experienced observers were relatively less tolerant when judging the test images with more texture rendition than the ones with more added noise, as compared to the unexperienced observers who, in relation, judged texture rendition and added noise similarly.

One point to make here is that there are different types of experience. In this case, they were all working with cameras. The results may turn out different if the experienced observers were working with a different aspect or purpose.

7 Suggestions for future Work

7.1 Evaluation of video streams

Axis cameras typically deliver video streams. This project has been focusing on still images. However, the methodology should be possible to use for video quality evaluation as well.

7.2 Camera evaluation and benchmarking

The visual lab provides a controlled environment for looking at and/or judging images and videos. Therefore, it is a good resource in the evaluation, and thereby the development, of the cameras of the company.

The visual lab could also be used for benchmarking. In this way it is possible to quantify the difference between cameras in a subjective way.

7.3 Reducing variance of the judgments

This thesis project does indicate that the "closer" in apperance the attribute to be judged is to the reference attribute sharpness, the less the variance of the judgements seems to become. This "simulatiry impact on variance" becould be investigated further in future studies.

Two set of ruler images
One suggestion is to perform studies with two sets of ruler images. Every test image could be judged two times (with the different rulers) and the variances of the judgments could then be investigated.

7.4 Ideas for future modifications of the Visual lab

Headrest
As for the setup for the Visual lab, a headrest for controlling the viewing distance should be bought. For the purpose of the studies performed in this thesis project, it was concluded not to be needed, but for future use of the lab it will be; if wanting to calibrate own ruler images, a headrest is very important.

Null images
As mentioned, three null images were used in the large study. There were no specific conclusions drawn from the null responses. This topic should be investigated more deeply in the future, in order to detect observers that might should be removed in the analysis. More null images should then be used, from the whole range of the softcopy quality rulers.

The authors of [12] removes observers with null responses "significanlty larger" than 2.5 JNDs. The limit number should be investigated more accurately, in order to be confident in the possible removal of observer data.

Judging video
The Visual lab set up is appropriately for judging video as well, no modifications are considered needed for that purpose.

Level of experience of observers
As stated, the difference in observer experience affects the judgments. In the future, when performing subjective studies, one suggestion is to have the observers answer a short form on their background of and experience of image quality.

8 Summary

In this thesis project, a working environment for performing subjective studies was set up at Axis Communications AB. The ISO 20462 subjective softcopy quality ruler method was evaluated through a study of one specific image quality problem; human preference in the noise reduction - texture loss space. In the study, a total of 47 observers judged images varying in scene content and level of noise and amount of noise reduction. Here follows the three main conclusions drawn. The first conclusion is based on the comparison of confidence intervals calculated with the normal approximation and the non-parametric bootstrap method, and concerns distribution of the judgment data:

Conclusion: *The judgments of the softcopy quality ruler can be said to be normally distributed, also for a smaller amount of observers*

The second conclusion concerns the optimal trade-off between noise and texture loss:

Conclusion: *In the study performed in this thesis project, the optimal noise reduction level seems to be 40 % for all five noise levels, even if the noise reduction was rather unnecessary for the one or two first noise levels. Even so, the noise reduction algorithm does not seem to unnecessarily damage the image at these low noise levels. The other noise reduction levels; 0 %, 20 % and 80 %, resulted in lower SQS values than for the 40 % level.*

The third conclusion concerns the variance of the judgments:

Conclusion: *When using the softcopy quality ruler method for evaluating image quality, the variances of the judgments in this study seem to depend considerably on the perceived similarity of the distortion attribute in the ruler images and the test images. If the ruler images appear similar to the test images, the variances of the judgments will be lower than for less similar ruler images and test images. Therefore, if it is important to minimize the number of observers, the choice of attribute used for the ruler images should be carefully considered.*

The fourth conclusion concerns the difference between experienced and unexperienced observers:

Conclusion: *When judging overall image quality, experienced observers working with cameras (imaging engineers, camera designers etc.) are more tolerant to noise than unexperienced observers*

A
Appendix A

A.1 The formation of images described by Fourier Optics

Fourier optics provide a beautiful way to treat optical systems in terms of spatial frequencies [15]. The concept of Linear systems, see Appendix B, is a key point in the analysis of optical systems. Linear systems theory can be applied to optical systems provided that [15]:

(a) The optical system can be adequately modeled as a as a linear system.

(b) The object to be imaged is the input to the linear system.

(c) The image is the output of the linear system.

A.1.1 Point Spread Function, PSF

Assume there is an shift invariant linear system $g(x) = \mathcal{H}[f(x)]$. Using the properties of the δ -function and equation 37 in Appendix B, the system can be written:

$$g(x) = \mathcal{H}\left[\int_{-\infty}^{\infty} f(x')\delta(x'-x)dx'\right] = \int_{-\infty}^{\infty} f(x')\mathcal{H}[\delta(x'-x)]dx' \tag{22}$$

$\mathcal{H}[\delta(x'-x)]$ is denoted $h(x'-x)$. Since the system is shift invariant $h(x'-x) = h(x)$. $h(x)$ is called the *Impulse Response Function (IRF)* of the system.. Equation 22 can therefore be written as:

$$g(x) = \int_{-\infty}^{\infty} f(x')h(x'-x)dx' \tag{23}$$

This representation is recognized as the convolution of $f(x)$ and $h(x)$. More compactly written:

$$g(x) = f(x) * h(x) \tag{24}$$

The IRF describes the response of the linear system to a delta function (an impulse) located at x. When extending this to two dimensions the following is obtained:

$$g(x,y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x',y')h(x'-x,y'-y)dx'dy' \tag{25}$$

$h(x,y)$ is called the *Point spread function (PSF)* of the linear system. The PSF describes the response of a linear system to a point source located at (x,y). Applying this to the formation of images, the PSF describes the response of an optical system to a point source located at (x,y) at the object to be imaged. Equation 25 is the two-dimensional convolution of $f(x,y)$ and $h(x,y)$:

$$g(x,y) = f(x,y) * h(x,y) \tag{26}$$

A.1.2 Line Spread Function, LSF

The one-dimensional counterpart to the PSF is the *Line Spread Function (LSF)*, denoted $l(x)$. In the formation of images, the *LSF* describes the response of an optical system to a line source with infinitesimal width at an object.

$$l(x) = \int_{-\infty}^{\infty} h(x, y)dy \tag{27}$$

A.1.3 Optical Transfer Function, OTF

The two-dimensional Fourier transform of the PSF is called the *Optical Transfer Function (OTF)* [15].

$$OTF(\eta, \xi) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} h(x, y)e^{-2\pi i(x\eta+y\xi)}dxdy \tag{28}$$

The OTF describes how features at varying spatial frequency are transferred from object to image.

A.1.4 The Modulation of a system

A useful parameter in the evaluation of the performance of a system is the *modulation* or *contrast* of the system [15]. The modulation is defined as the ratio of the amplitude of a function and the average function value:

$$\text{Modulation}(f(x)) \equiv \frac{(f_{max} - f_{min})/2}{(f_{max} + f_{min})/2} = \frac{f_{max} - f_{min}}{f_{max} + f_{min}} \tag{29}$$

The modulation of a function corresponds to the amount the function varies about its mean value divided by the mean value.

A.1.5 Modulation Transfer Function, MTF

The *Modulation Transfer Function (MTF)* of a system is defined as the ratio of the output modulation to the input modulation at all spatial frequencies:

$$MTF(\eta) \equiv \frac{\text{Output modulation}}{\text{Input modulation}} \tag{30}$$

It is therefore a measure of reduction in contrast from object to image over the spatial frequency spectrum.

To explain this further, assume that the following signal is input to a linear system:

$$f(x) = a + b\sin(2\pi x) \tag{31}$$

The output of the system will be given by:

$$\begin{aligned} g(x, y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (a + b\sin(2\pi x')h(x - x', y - y')dx'dy' = \\ = \int_{-\infty}^{\infty} (a + b\sin(2\pi x')l(x - x')dx' = \\ = a + b\left(\int_{-\infty}^{\infty} \sin(2\pi x')l(x - x')dx'\right) = \\ = \dots \end{aligned} \tag{32}$$

After some calculations, the following is obtained:

$$g(x) = a + b \times MTF(\eta) \sin(2\pi x + \Delta(x)) \tag{33}$$

where $\Delta(x)$ is a phase shift, called the *Phase Transfer Function (PTF)* and $MTF(\eta) = \left| \int_{-\infty}^{\infty} l(x) e^{-2\pi i x \eta} dx \right|$. The MTF is therefore given by the absolute valuer of the Fourier transform of the LSF. The MTF is also the magnitude of a one-dimensional slice of the OTF:

$$MTF(\eta) = |OTF(\eta, 0)| \tag{34}$$

B
Appendix B

B.1 Linear Systems

A linear system is a mathematical systems model based on the use of a linear operator describing input and output relations.

Definition B.1 Linear system: *A system consists of an input $f(x)$, passing through an operator $\mathcal{H}$, resulting in an output $g(x)$. This input-output relation is written as $g(x) = \mathcal{H}[f(x)]$. The system is linear if the properties of superposition hold [40].*

Homogeneity property of superposition:

$$a \times g(x) = a \times \mathcal{H}[f(x)] = \mathcal{H}[a \times f(x)] \tag{35}$$

Additivity property of superposition:

$$\begin{aligned} g_1(x) + g_2(x) + ... + g_n(x) = \\ = \mathcal{H}[f_1(x)] + \mathcal{H}[f_2(x)] + ... + \mathcal{H}[f_n(x)] = \\ = \mathcal{H}[f_1(x) + f_2(x) + ... + f_n(x)] \end{aligned} \tag{36}$$

Combining the homogeneity and additivity properties, a linear system satisfies

$$\begin{aligned} a_1 \times g_1(x) + a_2 \times g_2(x) + ... + a_n \times g_n(x) = \\ = a_1 \times \mathcal{H}f_1(x) + a_2 \times \mathcal{H}f_2(x) + ... + a_n \times \mathcal{H}f_n(x) = \\ = \mathcal{H}[a_1 \times f_1(x) + a_2 \times f_2(x) + ... + a_n \times f_n(x)] \end{aligned} \tag{37}$$

for all permitted inputs and all constants $a_1, a_2, ...a_n$.

Definition B.2 *A linear system is* space invariant *or* shift invariant *[40] if the changing of input position only changes the position of the output, not the functional form of it. In a space invariant system the output is only dependent on the shift in input position. If $g(x)$ is the output of the system to an input $f(x)$, the output becomes $g(x - x')$ when the input is $f(x - x')$.*

B.2 The Convolution Theorem

Theorem B.1 The convolution theorem: *Assume there are two functions $f(x)$ and $h(x)$ with Fourier transforms $\mathcal{F}\{f(x)\} = f(\eta)$ and $\mathcal{F}\{h(x)\} = h(\eta)$ respectively. Then if $g(x) = f(x) * h(x)$:*

$$\mathcal{F}\{g(x)\} = \mathcal{F}\{f(x) * h(x)\} = \mathcal{F}\{f(x)\} \times \mathcal{F}\{h(x)\} \tag{38}$$

or

$$G(\eta) = F(\eta) \times H(\eta) \tag{39}$$

where (x) represents the spatial domain and (η) represents the spatial frequency domain [15].

C
Appendix C

C.1 Frequentist statistics and confidence intervals

Data y is seen as an observation of a random variable Y. Y has the distribution $\mathbb{P}$. An estimate $t(y)$ (t is a function of the data) is an observation of the random variable $t(Y)$. Accordingly, an estimate $t(y)$ is an observation of the random variable $t(Y)$.

If $\tau = \tau(\mathbb{P})$ is some property, called the estimand, of the distribution $\mathbb{P}$, the error associated with an estimate $\Delta(y) = t(y) - \tau$ is an observation of the random variable $\Delta(Y) = t(Y) - \tau$. $\Delta(Y)$ has the distribution $\mathbb{F}$.

Assuming that the error distribution $\mathbb{F}$ is known, the confidence interval, $(L(y), U(y))$ on confidence level α for the estimand τ is:

$$\begin{aligned} 1-\alpha &= \\ &= \mathbb{P}(L(Y) \leq \tau \leq U(Y)) \\ &= \mathbb{P}(t(Y) - L(Y) \geq t(Y) - \tau \geq t(Y) - U(Y)) \\ &= \mathbb{P}(t(Y) - L(Y) \geq \Delta(Y) \geq t(Y) - U(Y)) \end{aligned} \tag{40}$$

The confidence interval should satisfy:

$$t(Y) - L(Y) = U(Y) - t(Y) = -(t(Y) - U(Y))$$

Therefore:

$$\begin{aligned} \mathbb{F}(t(Y) - L(Y)) - (1 - \mathbb{F}((t(Y) - L(Y))) = \\ = 2\mathbb{F}(t(Y) - L(Y)) - 1 = \\ 1 - \alpha \end{aligned} \tag{41}$$

And:

$$\begin{aligned} 1 - \mathbb{F}((t(Y) - U(Y)) - \mathbb{F}((t(Y) - U(Y)) = \\ = 1 - 2\mathbb{F}(t(Y) - U(Y)) = \\ 1 - \alpha \end{aligned} \tag{42}$$

The following is obtained:

$$\begin{aligned} \mathbb{F}(t(Y) - L(Y)) = 1 - \alpha/2 \quad &\Leftrightarrow \quad L(Y) = t(Y) - \mathbb{F}^{-1}(1 - \alpha/2) \\ \mathbb{F}(t(Y) - U(Y)) = \alpha/2 \quad &\Leftrightarrow \quad U(Y) = t(Y) - \mathbb{F}^{-1}(\alpha/2) \end{aligned} \tag{43}$$

The confidence interval on level α is thereby :

$$I = (t(y) - \mathbb{F}^{-1}(1 - \alpha/2), t(y) - \mathbb{F}^{-1}(\alpha/2)) \tag{44}$$

If $y = (y_1, \ldots, y_n)$ are observations of n independent variables distributed as $\mathcal{N}(\mu, \sigma)$. Then, $\Delta(Y) = t(Y) - \tau$ are distributed as $\mathcal{N}(0, \sigma/n)$, and:

$$\begin{aligned} \mathbb{F}^{-1}(1 - \alpha/2) &= \lambda_{\alpha/2}\frac{\sigma}{n} \\ \mathbb{F}^{-1}(\alpha/2) &= -\lambda_{\alpha/2}\frac{\sigma}{n} \end{aligned} \tag{45}$$

where the value of $\lambda_{\alpha/2}$ can be found in a tabular. The normal confidence interval on confidence level α is:

$$I = (t(y) - \lambda_{\alpha/2}\frac{\sigma}{n}, t(y) + \lambda_{\alpha/2}\frac{\sigma}{n}) \tag{46}$$

References

[1] About Axis. *Axis Communications AB.* `http://www.axis.com/corporate/index.htm` (April 9, 2014)

[2] Engeldrum, P. G. (2000) *Psychometric Scaling: A Toolkit for Imaging Systems Development.* Winchester: Imcotek Press.

[3] ibid, page 1.

[4] Phillips, J. B. and Christoffel, D. (2010) *Validating a texture metric for camera phone images using a texture-based softcopy ruler attribute.* Proc. SPIE, Volume 7529, 75290.

[5] Heynderickx, I., Bech, S. (2002) *Image quality assessment by expert and non-expert viewers.* Proc. SPIE, Volume 4662, pp. 129–137.

[6] Definition of objective. *Oxford Dictionaries.* `http://www.oxforddictionaries.com/definition/english/objective?q=objective` (April 9, 2014)

[7] Definition of subjective. *Oxford Dictionaries.* `http://www.oxforddictionaries.com/definition/english/subjective?q=subjective` (April 9, 2014)

[8] Gescheider, G. A. (1997) *Psychophysics: The fundamentals.* 3rd edition. Mahwah: Lawrence Erlbaum Associates, Inc.

[9] Keelan , B. W. (2002) *Handbook of Image Quality.* Boca Raton: CRC Press.

[10] Phillips, J. B., Coppola S. M., Jin, E. W. Chen Y., Clark J. H. and Mauer T. A. (2009) *Correlating objective and subjective evaluation of texture appearance with applications to camera phone imaging.* Proc. SPIE, Volume 7242, 724207.

[11] Johnson, G. M. and Fairchild, M. D. (2004) *The Effect of Opponent Noise on Image Quality.* Proc. SPIE, Volume 5668, pp. 82–89.

[12] Keelan, B. W., Jin, E. W. and Prokushkin, S. (2011) *Development of perceptually calibrated objective metric of noise.* Proc. SPIE, Volume 7867, 786707.

[13] Mantiuk K. M., Tomaszewska, A., and Mantiuk, R. (2012) *Comparison of four Subjective Methods for Image Quality Assessment.* Computer Graphics Forum, Volume 31, pp. 2478–2491.

[14] Image Resolution. *Wikipedia.* `http://en.wikipedia.org/wiki/Image_resolution` (March 19, 2014)

[15] Hecht, Eugene. (1987) *Optics.* 2nd edition. Boston: Addison-Wesley Publishing Company, Inc.

[16] Contrast (Vision). *Wikipedia.* `http://en.wikipedia.org/wiki/Contrast_(vision)` (March 19, 2014)

[17] Modulation Transfer Function (MTF). *DxOMark*. `http://www.dxomark.com/About/In-depth-measurements/Measurements/Sharpness` (March 19, 2014)

[18] Reinhard, E., Khan, E. A., Akyuz, A. O. and Johnson, G. M. (2008) *Color Imaging*. Wellesley: A K Peters, Ltd.

[19] What is noise?. *DxOMark*. `http://www.dxomark.com/About/In-depth-measurements/Measurements/Noise` (March 19, 2014)

[20] Digital camera image noise. *Cambridge in color*. `http://www.cambridgeincolour.com/tutorials/image-noise-2.htm` (March 19, 2014)

[21] Fechner, G. T. (1860) *Element der Psychophysik*. Leipzig: Breitkopf & Härterl.

[22] Hume, D. (1748) *Inquiry Concerning Human Understanding*.

[23] Marks, L. E. (1975) *On colored-hearing synesthesia: Cross-modal translations of sensory dimensions*. Psychological Bulletin, Volume 82, pp. 303-331.

[24] Marks, L. E. (1987) *On cross-modal similarity: Auditory-visual interactions in speeded discrimination*. Journal of Experimental Psychology: Human Perception and Performance, Volume 13, pp. 384-394.

[25] Gilbert, A. N., Martin, R., Kemp S. E. (1996) *Cross-modal correspondence between vision and olfaction: The color of smells*. Am. J. of Psych., Volume 109, pp. 335-351.

[26] International standard ISO 20462. (2012) *Photography − Psychophysical experimental methods for estimating image quality*.

[27] International standard ISO 20462–3. (2012) *Photography − Psychophysical experimental methods for estimating image quality − Part 3: Quality ruler method*.

[28] Jin, E. W., et al. (2009) *Softcopy Quality Ruler method: Implementation and validation*. Proc. SPIE, Volume 7242, 724206.

[29] Coefficient of determination. *Wikipedia*. `http://en.wikipedia.org/wiki/Coefficient_of_determination` (April 19, 2014)

[30] Efron, B., Tibshirani, R. J. (1994) *An Introduction to the Bootstrap*. New York: Chapman & Hall.

[31] Adèr, H. J., Mellenbergh G. J., Hand, D. J. (2008) *Advising on research methods: A consultant's companion*.

[32] Blom, G., et al. (2005) *Sannolikhetsteori och statistikteori med tillämpningar*. Lund: Studentlitteratur AB, ISBN:91-44-02442-8.

[33] Jin, E.W. *System Requirements for Implementing the Softcopy Quality Ruler method.*
`http://www.aptina.com/ImArch/` (November 10, 2014)

[34] Jin, E.W. and Keelan, B. W. *Aptina Softcopy Quality Rulers User's manual.*
`http://www.aptina.com/ImArch/` (November 10, 2014)

[35] Eye test chart.
`http://i-see.org/block_letter_eye_chart.pdf` (April 7, 2014)

[36] Li, C. (1981) *New color vision testing chart.* Liaoning: Liaoning People's Press.

[37] Deffner, G. et al. (1994) *Evaluation of display-image quality: experts versus non-experts.* SID94 Digest, pp. 475.

[38] Neat Image.
`http://www.neatimage.com/` (April 9, 2014)

[39] Lubeck, j. (2008) *Föreläsningsanteckningar i statistikteori* Lund: KFS i Lund AB.

[40] Spanne, S. (1982) *Lineära System.* 3rd edition. Lund: KFS i Lund AB.