Creation & Assessment of a Video Quality Ruler

Maya Shah

Abstract

Image quality has always been a matter of great importance for anyone working with or consuming images and videos. Images and videos are, in general, for human consumption. Thus, image quality is essentially a subjective attribute of an image. Though there exist both objective and subjective approaches to assessing image quality, objective results may not correlate well with subjective results. On the other hand, pre-existing subjective methods for evaluating video quality can be time-consuming, resource-consuming, or unreliable.

The aim of this thesis was to create and analyse a method for subjectively determining video quality. Based on a previously established method for assessing still image quality prescribed in ISO 20462, this method, the Video Quality Ruler, establishes an absolute scale of perceptive video quality. The Video Quality Ruler consists of a series of 31 ordered video clips, varying solely in sharpness. Each of the 31 video clips are one perceptual unit apart. Users are able to determine the overall level of quality of a video through comparison against the Ruler. The method provides an easy, universal analysis of video quality that correlates well to human opinions, and is not as time- nor resource- consuming as many other subjective methods.

The creation of the Video Quality Ruler was, on the whole, successful. The calibration of the ruler videos differs slightly from the calibration of the Image Quality Ruler from ISO 20462, particularly for images on the blurrier end of the Ruler. Assessment of the Video Quality Ruler determined that a few levels of the Ruler may need to be adjusted, and that the software and laboratory used for the experiment should be improved slightly, to eliminate bias and reduce the variance of results.

Acknowledgements

I would like to thank my supervisors at Axis Communications, Henrik Eliasson, Xing Danielsson Fan, and Fredrik Pihl, and my supervisor at Lund University, Johan Lindstrom, for supporting me throughout this project.
I would also like to thank the 66 participants who took part in my (often long and tiresome) experiments.
Finally, I would like to thank Axis Communications for providing an open, motivating and creative environment to work in.

Contents

Abbreviations

HVS	Human Visual System
IQR	Image Quality Ruler
JND	Just Noticeable Difference
LSF	Line Spread Function
ML	Maximum Likelihood
MLE	Maximum Likelihood Estimation
MTF	Modulation Transfer Function
PSF	Point Spread Function
SQS	Standard Quality Scale
VQR	Video Quality Ruler

1 Introduction

Swedish-based company Axis Communications is the market leader in network video, having invented the world's first network camera in 1996.[1] This would not be the case without ensuring that the quality of their products is continuously improving. Though many factors contribute to the success of a product, a primary concern with providing excellent network video cameras, is the quality of the videos produced. This in itself can depend on a variety of factors, one example being the compression methods being applied to the video. Compression allows videos, which would ordinarily use a large amount of storage, to be stored at lower bandwidths. Many compression standards, such as H.264[2], already exist. Axis Communications aim to develop and improve such standards by introducing new compression algorithms, such as the relatively new Zipstream technology, which is based on the H.264 standard.[1] A new method of compression can only be deemed successful if the resulting videos are of good quality.

The determination of image and video quality is therefore a vital part of the network camera production process. Methods of assessing video quality can be divided into two distinct groups: objective and subjective methods. The former focus on evaluating video quality largely excluding human input. The latter solely take human opinions into account. As videos are, in general, for human consumption, subjective methods can provide the most informative results. However, subjective methods can be time- and resource- consuming. In contrast, objective methods are quick and easy to implement through the use of computer algorithms, but may not always correlate well with human opinions.

There are a variety of both objective and subjective methods of assessing video and image quality. A popular objective method is the SSIM Index, which appears to produce more accurate results than other objective methods.[3] An example of a subjective method is one in which participants are asked to watch video streams and rate the stream as 'bad', 'fair', or 'excellent', employed in a study of video quality analysis on mobile devices.[4] The rating scale can be changed to an alternative format, such as a numerical scale, as used in the MOS (Mean Opinion Score) Test. [5] As a contrast, pairs of videos may be compared directly to determine which is the better of two. This method may take approximately twice as long as the aforementioned subjective methods, but can detect very small differences between videos.[5] The ITU-R recommendation for the methodology for subjective assessment of the quality of television pictures further suggests a range of subjective methods with different purposes, under controlled conditions to avoid bias.[6] A number of articles analyse and support the use of objective and subjective methods of assessing video quality.[7],[8]
A subjective image quality method not introducted in the previous articles is the Image Quality Ruler, which is a set of images along an absolute scale of perceptual quality. Each of the 31 images within the ruler are one perceptual unit apart. The quality of an image is determined by comparison against the ruler, obtaining a standard quality scale value. It is faster to implement than a paired comparison experiment, and lacks the ambiguity of a non-calibrated rating scale.[9] The method has been successfully implemented in [10], [13], [11] The method has also been adapted for video, by both comparing videos against

the Image Quality Ruler [14], and through the creation of a Motion Quality Ruler [12], with success.

Though subjective methods are more expensive and take longer to implement, they provide vital insight into how humans perceive video quality. Subjective methods can not only enable development of the imaging world directly, through the improvement of compression algorithms and other related factors, but the information obtained from such methods can assist in the creation, analysis, and improvement of objective methods, which are much easier and cheaper to implement.

This thesis aims to develop the Image Quality Ruler by extending it for video purposes, and in particular, surveillance video purposes. Thus, 'quality', within this thesis, is defined from a surveillance perspective. The Image Quality Ruler for this use can be known as a Video Quality Ruler. The Video Quality Ruler will be created, with the help of existing research, and assessed for its accuracy and implementation procedures. The work builds on the ISO 20462 Standard [9], and previous studies on video quality rulers [14],[12]. The work within this thesis will differ from previous studies in a number of ways. Freitas et al. were not so rigorous when implementing the Image Quality Ruler.[14]. The Motion Quality Ruler was implemented in a cinema, whereas this thesis will detail the creation of a Video Quality Ruler for use in a laboratory on a computer display.[12]

2 Human Visual System (HVS)

In sufficiently understanding the complexities of determining image quality, an attribute heavily dependent on human opinions, some basic knowledge about the Human Visual System (HVS) must first be introduced.
The HVS consists of two parts, the eye and the brain. Figure 1 shows the cross-section of a human eye. Images are collected by the HVS through light waves emitted or reflected from a scene.

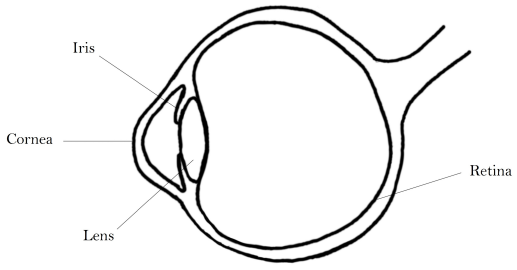


Figure 1: Cross-section of human eye

The cornea is a transparent protective layer that bends the light waves through the pupil, a small gap in the iris. Depending on the size of the pupil, more or less light is allowed into the eye. This is useful in different lighting conditions: the

iris will expand and contract in bright and dark conditions, respectively, allowing the HVS adapt in order to see as much as possible. The light waves then pass through the lens which varies in shape to focus the image onto the retina. For light with enough energy, in the correct spectral range (approximately 300 to 700 nm for the average human), the retina will send an electric signal to the brain through a nerve known as the optic nerve.[15]
The retina detects and converts light waves into electrical signals with the use of rods and cones, two types of photoreceptors. Rods are photoreceptors responsible for vision at low light levels, and can only perceive shades of grey, and cones, of which there are typically three types, are photoreceptors in charge of colour perception at higher light levels.[15] The three types of cones are often denoted ρ, γ & β, each of which detect light of different wavelengths.[16] There are some exceptions to this case, where some people's HVS diverge away from the average HVS. Some people can have fewer than 3 cones, and are partially or completely colour blind. Colour blindness is exhibited in the human being unable to distinguish between two or more colours, with the most common being the inability to discriminate red and green.[16]
Some other forms of HVS inconsistencies are short-sightedness, long-sightedness, and astigmatism. If the eyeball is of incorrect length, or exhibits incorrect curvature of the cornea or lens, this can result in an image not being focused onto the retina properly, resulting in blurry vision. However, these conditions can be (temporarily) corrected with glasses or contact lenses.[17]

3 Digital Imaging

Throughout this thesis, the term 'image' refers to both still and moving images (videos). Additionally, the focus will solely be on digital (as opposed to analogous) images, as this is currently the most commonly used format of images.

3.1 Mechanism & Operation of a Camera

Digital images, i.e., digital reproductions of scenes, can be captured by use of digital cameras. The mechanism of a camera is similar to that of the HVS. Light passes through a curved lens or lenses, which focus(es) the light through a small hole known as the aperture. At the back of a digital camera is a sensor, which collects the information travelling in the light. In general, cameras differ from the HVS by the use of a shutter, which is placed in front of the sensor, and controls when and how much light is collected. When the camera is clicked, the shutter is removed for a brief moment, light is collected by the sensor, and an image of the captured scene is recorded. This is the very basic structure of a digital camera.[18]

Various settings within a camera can be controlled in order to adjust the reproduced image. Some of these settings will be explored here. The camera lens is comprised of several lens elements that concentrate the light waves as accurately as possible onto the sensor. The type of lens can heavily affect the quality of the reproduced image.[19]

The exposure (brightness) in a reproduced image is mainly dependent on three things within the camera: Aperture, Shutter Speed, and ISO Speed. These

three elements are also responsible for other aspects of the reproduced image.

The aperture controls the size of the hole through which light can enter and be stored in the sensor. Instinctively, a smaller hole indicates that a darker image will be produced. The aperture size is denoted by the f-number, where a smaller f-number denotes a larger aperture, and vice versa. The shutter speed controls the duration that the sensor will be exposed to light. A slower shutter speed indicates that more light will be able to enter the camera, thus ensuring a more illuminated image. The ISO Speed controls the sensitivity of the camera's sensor to a given amount of light.[20]

Other than controlling the brightness of an image, the aperture can control the depth of field of the scene. The depth of field is defined as the range of distance over which objects appear in sharp focus. A smaller aperture (or larger f-number) will theoretically mean that objects in a larger range of distance will be in focus. However, with smaller apertures, diffraction becomes more prominent.[21] Diffraction is the bending of light when it passes through a small gap, and will cause the reproduced image to be slightly blurry. This will be discussed further in Section 3.6. Therefore, in reality, a small aperture may cause a loss of sharpness over the entire image. When capturing an image where it is desirable to have a large depth of field, it is useful to find the 'sweet spot' aperture size, where the aperture is as small as possible, while still exhibiting as little diffraction as possible.

As mentioned, the shutter speed can control the light exposure to the sensor. While a slow shutter speed may be desirable in areas with low light, or in conjunction with a small aperture, it could cause motion blur to be present in the reproduced image. This can be particularly noticeable when capturing a scene with moving objects. For example, when capturing a scene with fast moving water. A water droplet will move a certain distance while the shutter is being opened, and if the shutter speed is slow, the reproduced image will display the water droplet over the entire distance it moved in that time, resulting in motion blur and a loss of texture.

ISO Speed can increase the level of exposure in a reproduced image, however, there is a subjective highly undesirable side effect in the form of noise, which is an artifact (distortion) described in Section 3.6. High levels of ISO Speed can cause high levels of noise to be present in the image. Cameras in general hold ISO Speed values of $100, 200, 400, 800, ...$[20]

The colour temperature of a light source (that is illuminating an object), expressed in Kelvin, [16] describes the wavelength dependent properties of the light source. Though the HVS is able to see past colour temperatures to visualise the real colour of the object, a digital camera cannot realistically capture a scene without compensating for the colour temperature of the light source that the scene is exposed to. 3000 K and 9000 K light sources will place more importance on orange and blue wavelengths, respectively, thus producing images with measurably different colours.[22] In a digital camera, it is possible to compensate for the light source with a function known as 'White Balance'. It is known as White Balance as it ensures that whites appear white in the reproduced image. In scenarios containing multiple light sources with different colour temperatures, it may be difficult to apply white balance effectively.[22]

When capturing moving images, it is possible to specify the frame rate of the produced video clip. In general, the highest possible frame rate and the slowest possible shutter speed will interact with each other. Should the desired frame rate be a certain speed, for example, 25 frames per second, the slowest possible shutter speed for this frame rate will have to be faster than 1/25 fps.

3.2 Digital Representation of Images

Digital images are essentially matrices of values, representing a grid of pixels. Building upon this basic idea, there are different methods of encoding (i.e. digitally representing) images.

Colour Spaces

There are a number of methods to storing the colour information of images. One type, is an 'RGB' colour space, where information about the image is stored in a 5 dimensional matrix. The first two dimensions make up the pixel resolution of the image, and the remaining dimensions contain data about each of the three colour channels (red, green, blue) respectively. Recall that the HVS detects three types of colours. The 3 channels of the RGB colour space closely resemble the 3 colours detected by the HVS.[23] Though the colours in an RGB model can be implemented in different ways, a common implementation is the 24-bit model, where 8-bits, or 256 discrete levels are given to each colour channel. Therefore, any colour space based on such a model is limited to a range of $256 \cdot 256 \cdot 256 = 16{,}777{,}216$ colours.

Standardisation of the RGB colour space is necessary to ensure that colours are displayed correctly. One such standardisation is the 'sRGB' colour space. sRGB, shortened from 'standard RGB', is a colour space created in 1996 by HP and Microsoft, and has since become a de facto standard colour space used to represent digital images.[24]

'YCbCr' is a family of colour spaces in which RGB images are encoded in a different format. The 'Y' represents the luminance component, 'Cb' represents the difference between the blue component and a reference value, and 'Cr' represents the difference between the red component and a reference value. 'YCbCr' reduces the bandwidth of an image by acknowledging the fact that humans are more sensitive to luminance (i.e. black and white information) than to chrominance (colour information). This colour space increases the overall proportion of information that explains luminance information in comparison to chrominance information.[25] The proportion of the luminance information in a YCbCr image is determined by the type of RGB to YCbCr conversion used. An example of a YCbCr colour space is 'YUV 4:2:0'. In this format, the size of the luminance information is the same size as the pixel resolution, in that there is the same number of Y values as there are pixels. Conversely, there is half the amount of chrominance information as there are pixels, with the Cb and Cr components making up a quarter of the pixel resolution each. This reduces the bandwith of the image from 3 x 'pixel resolution' for an RGB' image, to $\frac{3}{2}$ x 'pixel resolution' for a YUV4:2:0 image. Both of the chrominance channels represent sets of 2x2 blocks of pixels in the image, meaning each Cb and Cr entry applies to 4 pixels.[26]

File Formats

Given the colour information of a digital image, the image can be compressed further and stored in a number of different formats. Compression is often necessary due to the data size of an uncompressed digital image. One compression format is the 'JPEG' format, which is commonly used to store images produced by digital photography. It is regarded to be a lossy compression, such that in order to reduce the data size, the information contained within the image is degraded, dependent on the degree of compression applied.[27] Similarly, 'MJPEG' is a format for compressing and storing videos, in which each frame is a JPEG image.[28] A file format which supports a method of lossless compression, is one in which the quality of the image is not reduced when stored. The 'Portable Network Graphics' (PNG) file format is the most commonly used lossless data storage on the internet.[29] Though the quality of a PNG image may be higher than the quality of the same image stored in JPEG format, the storage size of the PNG image will far exceed the storage size of the JPEG image.[27]

3.3 Image Processing Pipeline

In order to reproduce an image, the data recorded about the scene by the sensor must be sent through an image processing pipeline. Though the exact details of an image processing pipeline can vary from system to system, the stages implemented in digital colour cameras mostly follow a basic generalised pattern.

After the light information passes through the lens, it is collected and stored by the sensor. Instead of holding 3 sensors for the red, green and blue wavelengths, respectively, many digital cameras place a colour filter array over a single sensor. The most common color filter array is the 'Bayer' array, which consists of a red-green-blue checkered pattern that allows the sensor to collect different colour information for different pixels.[30] The Bayer pattern is shown in Figure 2. There are twice as many green pixels as there are red pixels and blue pixels each. Bryce Bayer created the Bayer pattern in this way in order to mimic the physiology of the HVS, which is most sensitive to green light.[31] After the first stage of the pipeline, the image appears as a checkered reproduction of the original scene. This image is known as the 'raw image'.

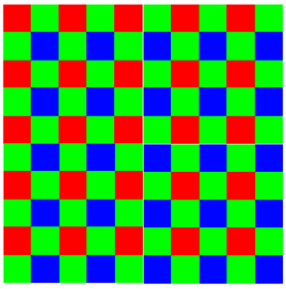


Figure 2: Bayer pattern

At the next stage, a number of algorithms are applied as part of the 'pre-

processing' stage. These often include missing/defective pixel correction and lens flare compensation.[30]

After pre-processing has been performed, the camera system applies the white balance.[30]

The next stage in the image processing pipeline is a step known as 'demosaicing', in which algorithms are applied to estimate the missing colour information due to the Bayer colour filter array. The algorithms that are applied vary depending on the camera system. After this is complete, the image loses the checkered effect and looks more like a natural image. Naturally, estimation of pixel colours will lead to artifacts being introduced into the image. After transforming the colour space of the image to sRGB, the remainder of the pipeline mainly focuses on correcting these artifacts and any others acquired during the pipeline so far. This final stage is part of the post-processing step, during which sharpening and noise reduction filters may also be applied.[30]

3.4 Camera Characterization

A digital imaging system, or in particular, a camera, can be characterised in a number of ways, for instance to determine the general quality of the images that are produced. Two such methods determine the sharpness of a digital image and possible image enhancing algorithms that may have been implemented in the image processing pipeline, respectively.

3.4.1 Modulation Transfer Function

The Modulation Transfer Function (MTF) is a key element in the quantification of sharpness (which will be defined in Section 3.6) in an image. Image filtering has its roots in signal processing (see Section 3.5), where an input signal is transformed to an output signal through an imaging system. If the input signal is an infinitely small point source, the response of the system is known as a Point Spread Function (PSF). The integral of the PSF is the 2 dimensional Line Spread Function (LSF). Further, a function known as the Optical Transfer Function (OTF) is obtained by taking the Fourier transform of the LSF.[32]
A perfect optical system would be able to distinguish the point source perfectly, however no optical system is perfect. The ability of an optical system to distinguish details is quantfied by the MTF, the modulus of the OTF in one dimension, or the modulus of the Fourier transform of the LSF. Thus, it is possible to characterise the sharpness captured by an optical system by examining its unique MTF. The MTF of optical systems can be measured in a number of ways, depending on the type of optical system. For example, the MTF of cameras has widely been measured using test charts containing slanted edges, [33] such as the test chart seen in Figure 3, used at Axis Communications.

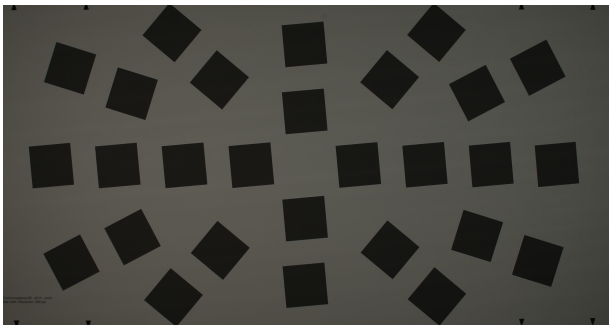


Figure 3: MTF Test Chart

An example MTF curve of an imaging system is pictured in Figure 4. The units of the x-axis can be, for example, cycles per pixel (CPP), cycles per mm, or cycles per degree (CPD), the last of which is a unit of frequency that is independent of viewing distance.

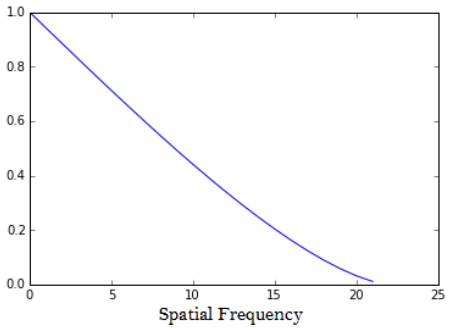


Figure 4: Example MTF Curve

When multiple optical systems with differing MTFs are used together to produce an image, the overall MTF of the image is easy to compute through an important property of the MTF. The cascading property of the MTF states that the MTF of the overall system is the point-by-point product of the individual subsystem MTFs.[32]

3.4.2 Opto-Electronic Conversion Function

The opto-electronic conversion function (OECF) describes the relationship between the optical input and digital output of a digital image. Algorithms that may be applied during the image processing pipeline affect the OECF of the digital image system, and thus measurement of the OECF can provide some insight into the general transformations made to the optical input. Additionally, the OECF of a camera can be used to correct the data of other camera characteristics, such as the MTF.[34]
The OECF of the camera can be calculated by capturing a test chart as seen in Figure 5. This particular test chart is used at Axis Communications, however similar styles are widely used to measure the OECF of digital cameras.[34]

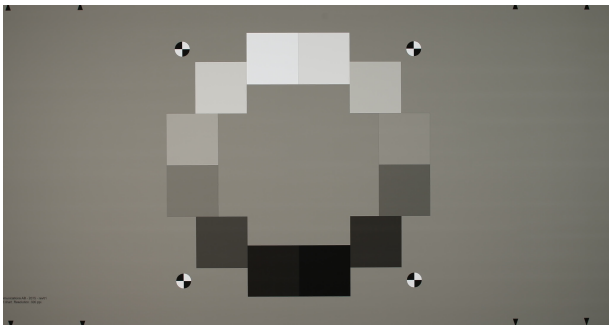


Figure 5: OECF Test Chart

3.5 Digital Image Processing

The pre- and post- processing algorithms touched upon in Section 3.3 largely consist of filtering the image in order to subjectively improve certain aspects of the image. Filtering can be used not only to enhance or restore an image, but also to degrade it.

Though there exists both linear and non-linear filtering, this thesis focuses purely on the former, based on the theory of signal processing. Specifically, the filtering applied can be defined as a linear shift (spatial) invariant transformation. A linear transformation T is one in which the following property holds:

$$T(a_1p_1(x,y) + a_2p_2(x,y)) = a_1T(p_1(x,y)) + a_2T(p_2(x,y))$$

In the case of image processing, p_1 and p_2 are two sections of a digital image. In a shift invariant system, if an output signal $q(x,y)$ is the response of the system to an input signal $p(x,y)$, then $q(x-k_x, y-k_y)$ is the response of the system to $p(x-k_x, y-k_y)$.[35]
Based on a linear shift invariant system, there are two main ways to effectively apply a filtering algorithm to an image. In both methods, the data for a pixel is replaced by new pixel data, directly dependent on neighbouring pixel data. The only difference in the methods is the domain in which the transformation is applied.[36]

The first of these methods is conducted within the spatial domain, i.e. pixel by pixel. In this method, filters are matrices with a certain size, shape, and weights. A filter is placed over a pixel in a given image, and a weighted average of the neighbourhood of the pixel is computed as the response of the filter at that position. This is repeated for each pixel within the image, until a filtered image is obtained.[37] Formally, this filter is known as an impulse response, as it describes the reaction of the system as a function of spatial coordinates. Essentially, the output signal of the system is obtained as a convolution between the input signal and the impulse response.[35] Applying a filter H, to an image I, the filtered image is, [37]

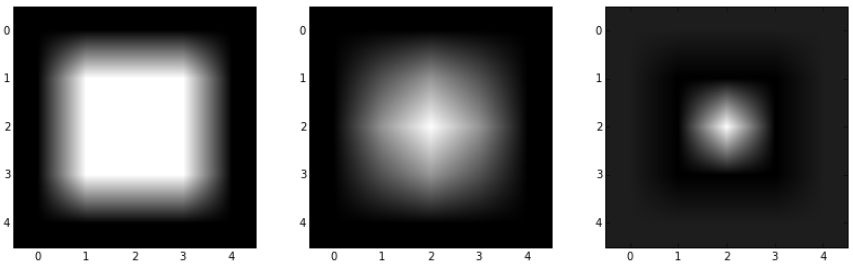
$$I' = I * H$$

Depending on the impulse response H, different transformations can be applied to any digital image. Among these transformations are blurring, noise reduction,

and sharpening. Though there are infinitely many filters that can perform transformations, three are shown below.[36]

$$\frac{1}{9}\begin{pmatrix}1 & 1 & 1\\ 1 & 1 & 1\\ 1 & 1 & 1\end{pmatrix}, \qquad \frac{1}{16}\begin{pmatrix}1 & 2 & 1\\ 2 & 4 & 2\\ 1 & 2 & 1\end{pmatrix}, \qquad \begin{pmatrix}-1 & -1 & -1\\ -1 & 9 & -1\\ -1 & -1 & -1\end{pmatrix}$$

The sum of all elements of a filter should equate to 1, and pixels at a distance larger than that covered by the filter are not considered when transforming the pixel at the centre of the filter. The first two filters are known as smoothing filters, or more formally, low-pass filters. The first filter is known as a Box filter, and the second filter is an example of a Gaussian filter, due to the weights in the elements of the matrices.[37] The Box blur places no importance on the position of the neighbouring pixels, whereas the Gaussian blur gives precedence to pixels at a closer distance to the centre pixel. The third filter is an example of a high-pass filter, and performs sharpening. Sharpening filters essentially subtract a smoothed version of the image from the original image, enhancing the edges of the image.[37] A 2-D visualisation of the impulse responses are shown in Figures 6a, 6b, and 6c.



(a) Box Blur (b) Gaussian Blur (c) Sharpening

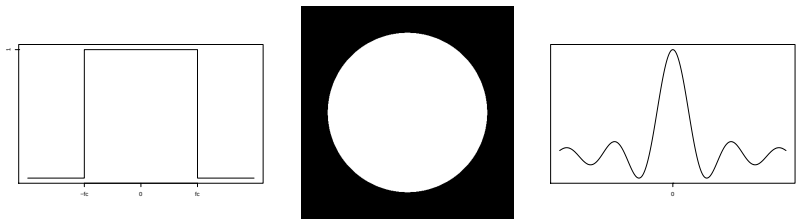
Figure 6: Spatial Filters

The disadvantage of filtering within the spatial domain is the use of convolution, which can be a very slow operation. This leads to the second method of linear filtering, performed in the frequency domain of the image. Where the spatial domain of an image is the image itself, digitally represented by a linear colour space, the frequency domain of the same image is the 2-dimensional Fourier transform of the image. The Fourier transform of an image characterises the image by its frequency content. It is usually impossible to make direct associations between specific components of an image and its Fourier transform. However, some general statements can be made about the relationship between the frequency and spatial domain of an image. For example, patterns of intensity variations in the spatial domain are associated with patterns in the frequency domain.[36]
The relation between filtering in the spatial and frequency domains can be characterised by the Convolution Theorem, which states that the Fourier transform of the convolution of two functions is equivalent to the product of the Fourier transform of each of the functions. That is,

$$\mathcal{F}(I * H) = \mathcal{F}(I) \cdot \mathcal{F}(H)$$

For more information on convolution, the Fourier transform, and the Convolution Theorem, see [38].
Thus, by taking the inverse Fourier transform of the product of the Fourier transforms of the image and the impulse response, it is possible to obtain the filtered image.[36] The filter applied in this domain, the Fourier transform of the impulse response, is known as a frequency response. Though image processing within the frequency domain can be conceptually more difficult to understand than image processing within the spatial domain, it can be performed more rapidly by use of the Fast Fourier Transform (FFT).

The spatial filters previously examined can be applied within the frequency domain. The term 'low-pass' for a smoothing filter is derived from its methodology within the frequency domain. A low-pass filter is a filter that allows signals with frequencies lower than a certain cutoff to pass, and removes any frequencies higher than the cutoff frequency.[37] In image processing terms, low-pass filters remove fine detail from images, essentially smoothing them. Conversely, high-pass filters allow signals with higher frequencies to pass, and remove lower frequencies.
The simplest low-pass filter is one that cuts off all frequency components above a specified frequency f_c, and passes all frequency components below f_c. This filter is known as the Ideal Low-Pass filter, and is visualised in Figure 7.[36]



(a) Frequency Response (b) 2D (c) Impulse Response

Figure 7: Ideal Low-Pass Filter

The graph for the frequency response perfectly describes the definition of the ideal low-pass filter. The impulse response for the ideal low-pass filter, computed through the inverse Fourier transform of the filter in the frequency domain, is a sinc filter. The sinc function, $\text{sinc}(x) = \frac{\sin(\pi x)}{\pi x}$, is infinitely long, and thus the ideal low-pass filter can only be applied effectively to images that are infinitely large. Therefore, implementation of the ideal low-pass filter introduces many artifacts in practice.
There exist filters which resemble the ideal low-pass filter, but do not quite match the signal processing quality achieved by the ideal low-pass filter. These filters offer a more practical approach to low-pass filtering.
One such filter is a Lanczos filter, which is known as the best compromise for image resampling in terms of the reduction of various artifacts.[39] The impulse response of the Lanczos filter is defined as

$$L(x) = \begin{cases} \text{sinc}(x)\text{sinc}(x/a), & -a < x < a \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

where $a \in \mathbb{R}_+$ is the order of the filter. A higher order Lanczos filter more closely resembles the ideal low-pass filter, however, higher orders may produce artifacts within the filtered image. Typically, values of $a = 2, 3$ are chosen.[39],[40]

3.6 Aspects of Image Quality

Having explored certain elements that go into creating an image and ensuring that said image is of 'good quality', the term 'good quality' will now be defined.

Image quality is defined as the perception of an image being 'good'. For the purposes of this study, an image of 'good' quality is said to be as close to reality as possible, that is, as close to reality as possible in terms of human perception. A 'good' quality image within this study is defined from the perspective of surveillance, as opposed to the perspective of, perhaps, an artist or consumer photographer. Cameras do not always capture exactly what we see. An example of this, is an area with both dark and illuminated parts. A camera is not always capable of capturing both the dark and light parts together well in situations where a human observer would have no difficulty, with some extreme exceptions. An image or video of good quality will tell us as much information about the scene as possible, without compromising information in any other way. As images and videos are, the majority of the time, for human consumption, 'image quality' is essentially a subjective matter. Further, humans with different preferences and experiences can also judge image quality in different ways, for example, professional film makers and editors versus film consumers, people with normal versus impaired vision.

There are various things that can affect or contribute to image and video quality. In the following, some of these factors will briefly be examined. The affecting attributes can be varied by the camera performance and settings applied to the camera, the shooting conditions, the methods involved in processing and compressing the images, and a number of other factors implemented in the image reproduction pipeline.

Texture Rendition

Texture is an attribute of image quality that can often be reproduced badly in an image. Texture can be described by the surface properties of objects, and can enhance object recognition. With changes in texture, rough objects reproduced in an image can appear smooth, and vice versa, therefore degrading the image quality. Another example of the use of texture in an image, is in identifying the direction of wind across a body of water. Often, as texture in an image decreases, objects in an image begin to look like an oil painting, losing their sense of reality.[41]

Motion

Motion is the attribute that separates videos (moving images) from still images. Videos are produced by capturing multiple images, or 'frames', which are then viewed successively at a rapid rate to fool the viewer into seeing a moving picture, by a phenomenon known as 'Persistence of Vision'.[42] Persistence of Vision is an example of an optical illusion, whereby the HVS blends multiple

images into a single moving image. Should the discrete images be similar enough, and the speed at which the images change is fast enough, the change can be unnoticeable to the human eye. However, if the change is noticeable, the viewer of the video can experience a decline in quality in comparison to the real life scene. As explained in Section 3.1, it is possible to set the number of frames per unit time, and naturally, a higher frame rate will in general produce higher quality videos, should the other factors remain constant.

Colour

The appearance of colour in an image can highly change the subjective quality rating. For instance, a grayscale image, or even an image with a blue tint, can cause loss of information from the image, therefore resulting in a loss of quality, particularly when looking at an image from which we want to gather accurate information about the subject. For artistic purposes, colour distortion may be desirable, however, for image quality as defined in this thesis, it is an attribute for which a minimal amount of departure from reality is required.
Another, less intuitive example of colour perception can be described by the Hunt effect, which states that as the level of illumination is decreased, colours lose their 'colourfulness'.[16]
There are 3 basic attributes that may be assigned to colour: brightness, hue, and saturation. Hue is the attribute of visual perception according to which an area appears to be similar to one or proportions of two of the perceived colours red, yellow, green and blue. Saturation defines the amount of hue that is exhibited in an area.[16] Brightness will be defined in the next section.

Lightness

Brightness is an attribute of visual perception in which a source appears to be radiating or reflecting light.[16] The brightness of an image may be related to the luminance of the scene, and is an important subjective attribute when defining image quality. If an image is very dark or very bright, and unable to portray objects, however true to life the image may be, it may subjectively be deemed to be of bad quality. Conversely, if an image can display both light and dark details well, and is not quite identical to the image that is seen by the HVS, it may subjectively be deemed to be of bad quality. In this way, it may be difficult to quantify image quality as a function of luminance. In this thesis, a good quality image as a function of luminance will be an image created in good lighting conditions that, upon capture, allow the least amount of perceived deviance from the real life scene.

Contrast

Contrast can be defined as the difference in colour and brightness between objects within the same field of view. The maximum contrast of an image is known as the dynamic range of the image. The HVS is more sensitive to contrast than it is to absolute luminance.[43]

Pixel Resolution

All digital images are made up of matrices of pixels. 'Pixel' is the name given to small dots containing information about that small area of the image. The pixel resolution of an image is the number of pixels per said image or area. For example, if an image has a pixel resolution of 4096x2160, it will contain $4096 \cdot 2160 = 8{,}847{,}360$ pixels. In the most general case, a higher pixel resolution can mean higher quality, provided all other elements of the image remain constant. This is because it will be more difficult to discern between pixels and therefore see a seamless image, if there are more pixels in a given physical area. Problems can also arise when an image captured at a certain pixel resolution is viewed on a display with a different pixel resolution, due to over- or under- sampling. This will be examined briefly later in this section.

Spatial Resolution

Spatial resolution is related to the amount of detail an image holds. In particular, it is defined as the smallest discernible detail in an image.[36] The resolution of a camera can be limited by various factors, amongst which one of them is diffraction. In the case of cameras, diffraction can occur as the light passes through the lens of the camera, causing the light to bend slightly, reducing the resolution of the image.[44] Another factor that can limit resolution is aberrations, whereby light from a single point of an object does not converge into a single point in the reproduced image.[45]
As a practical measure, when viewing alternating black and white bars, spatial resolution is the narrowest and closest bars that can be resolved by the camera.

Sharpness

Sharpness and spatial resolution are in some ways very similar aspects of image quality, in that they both describe the amount of detail of an image. However, where spatial resolution is an objective attribute, sharpness is a subjective attribute, describing the *perceived* detail of an image, and is not always correlated with spatial resolution. For example, sharpness levels will change with viewing distance. An image may appear to be sharp from a certain distance away, but upon reducing the distance between image and eye, the image may appear to be less sharp.[46]

Other Artifacts

As previously mentioned, artifacts are distortions that can be introduced into a digital image during any stage of the image processing pipeline. Some examples which will be briefly defined include chromatic aberration, noise and spatial distortion. Chromatic aberration occurs when wavelengths of colour are focused by the lens at different positions on the sensor. This results in blurred or coloured edges around objects, particularly in high-contrast situations.[45] Noise is defined as random variation of brightness or colour information in images and videos, and normally arises in the image reproduction pipeline during image acquisition and/or transmission.[36] Spatial distortion is heavily dependent on the lens of the camera, and is exhibited in an image by curved lines, that would ordinarily be straight in the real scene.[47]

In the image processing pipeline, it may be required to resample the image, i.e. rotate or resize the image. An artifact that can arise during the resampling of images is aliasing. An example of aliasing is a wavy looking pattern known as a moiré pattern. As frequency components above half the sampling rate (the Nyquist frequency) will cause aliasing, an anti-aliasing filter can be applied to the image prior to resampling, to remove components above the Nyquist frequency.[48] Such a filter, as defined in Section 3.5, is known as a low-pass filter. Aliasing problems can also arise when an image captured at a certain pixel resolution is viewed on a display with a different pixel resolution. Another artifact that can appear as a result of resampling is known as ringing. Ringing is exhibited in digital images as bands or 'ghosts' near edges.[37]

4 Image Quality Assessments

Existing image quality assessments can be divided into two groups of methods, objective and subjective methods. Both groups of methods and some examples of each will be explored within this section.

4.1 Objective Video Quality Assessments

There exists a variety of objective methods to determine image quality. The methods can be categorized into 3 types: Full Reference (FR), No Reference (NR), Reduced Reference (RR) measures. These methods are dependent on the amount of information we have about the reproduced image and the undistorted image. The undistorted image is considered to be of perfect quality, and an ideal comparison to the reproduced image in objective quality investigations (FR). However, if there is no raw image, or if there is incomplete information about the raw image, NR and RR methods must be implemented, respectively.[49]
Further, objective evaluations can be separated into bottom-up and top-down approaches. The former is where algorithms are developed to model relevant parts of the HVS, and integrated into the assessments, in order to obtain methods that behave in a similar way to the HVS. Conversely, top-down approaches treat the HVS as a mystery and attempts to predict what an average human observer will experience, resulting in a method that may or may not operate in a different way to the HVS. Naturally, there are drawbacks to both approaches. The HVS is an extremely complex system, and is therefore difficult to model. In comparison, top-down approaches are simpler but can still perform at the same level as bottom-down approaches. However, there is some debate over the validity of the hypotheses that the top-down approaches are based on.[49]

An example of a FR top-down objective method is one in which you can examine the MSE (mean-squared error) of the pixels of the image. It is possible to determine the mean-squared error and a related measure, the peak signal-to-noise-ratio (PSNR) of the image in order to determine the level of quality in an image.[49]

$$\mathrm{MSE} = \frac{1}{N}\sum_{i=1}^{N}(x_i - y_i)^2,$$

$$\mathrm{PSNR} = 10\log_{10}\frac{L^2}{\mathrm{MSE}},$$

where $\mathbf{x}$ and $\mathbf{y}$ are two images, and L is the largest allowable pixel value.
A benefit of the MSE is that it is easy to use. However, it has a major drawback, in that images with the same MSE level can have varying levels of subjective quality.

In an attempt to build an improved FR top-down objective method, the Structural Similarity (SSIM) Index was developed. The SSIM Index algorithm separates the image information into luminance, contrast, and structure. It is then applied to an image locally, rather than globally.[3]
The SSIM Index of two windows within images $\mathbf{x}$ and $\mathbf{y}$ is

$$\text{SSIM}(\mathbf{x},\mathbf{y}) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)}$$

where μ_x and μ_y are the averages of $\mathbf{x}$ and $\mathbf{y}$, respectively. Similarly, σ_x^2, σ_y^2 and σ_{xy} are the variances and covariance of $\mathbf{x}$ and $\mathbf{y}$. C_1 and C_2 are constants determined by the dynamic range of the pixel values, in relation to the luminance and contrast comparisons on the images, respectively.
Results show that the SSIM Index is an improvement on the problems of MSE, however, as it is a FR metric, it is not always useful.[3]
Objective methods are fast and easy to implement through the use of computers. However, they may not always correlate well with human opinions of image quality.

4.2 Psychophysics

Psychophysics, an important concept used in subjective experiments, is the scientific study of the relationships between sensations (ψ) in the psychological domain, and stimuli (ϕ) in the physical domain. One such relationship can be examined through the 'absolute threshold' or 'stimulus threshold', which is the smallest amount of stimulus energy necessary to produce a sensation. The 'difference threshold' can be interpreted as the amount of change in a stimulus required to produce a 'just noticeable difference' (JND) in the sensation. 1 JND within this thesis is defined as a 75:25 ratio of agreement and disagreement of observers in reference to intensity of stimulus.[50] This can also be interpreted as the following. If 75% of observer opinions are correct, and 25% are incorrect, it is possible to assume that the 25% guessed their answer, and guessed incorrectly. Presumably, an equal number of observers guessed correctly, suggesting that 50% of the observers did not guess at all. Therefore 1 JND defines a situation where the stimulus intensity difference is so small that half of the observers will guess, and half will not.[51] 0 JNDs define the case in which all of the observers will guess, resulting in 50% agreement.

4.2.1 Weber's Law

The German physiologist E. H. Weber discovered that two heavy weights must differ by a greater amount than two light weights to be perceived as heavier than the other. More precisely, the size of the difference threshold is a linear function of stimulus intensity. This applies to sensations applied to all of the human senses. This relationship is known as Weber's Law, and in scientific

terminology is [50]

$$\frac{\Delta\phi}{\phi} = c$$

Later studies of Weber's Law found that as intensity of stimulus increased, $\Delta\phi/\phi$ decreased and then became approximately constant. Therefore, a modification was applied to Weber's Law in order for it to more closely correspond what the empirical data stated.[50]

$$\frac{\Delta\phi}{\phi + a} = c,$$

where a is a constant with relatively small value. The exact significance of the constant has not been determined, but may represent sensory noise when the intensity of the stimulus is very low.[50]

4.2.2 Weber-Fechner Law

German psychologist G. T. Fechner, with a strong background in physics and mathematics, aimed to derive a mathematical equation for the relationship between physical events and conscious experience. He proposed that sensation magnitude could be quantified indirectly by relating the values of the change in stimulus intensity on the physical scale to the corresponding values of the JND in sensation on the psychological scale. By assuming the validity of Weber's Law, he developed the Weber-Fechner Law, which proposes that the sensation magnitude increases proportionally to the logarithm of the stimulus in units above the absolute threshold.

$$\psi = k \log \phi,$$

where k is a constant dependent upon the particular sensory dimension and modality. This can also be known as Fechner's Law.[50]

4.3 Subjective Video Quality Assessments

In Section 4.1, objective methods of assessing image quality were examined. As an alternate method, subjective methods, which employ the use of statistical analyses, can be implemented to evaluate image quality. Though subjective methods can be time- and resource- consuming, the results from such experiments are often highly correlated with human opinions. In this section, a number of different types of subjective methods will be examined.
Psychometric scaling is the generation of scales of stimulus intensity by human measurement. The stimulus intensity in this case is the level of (moving) image quality. Psychophysicist S. S. Stevens put forward a reference of scaling types in 1946, as shown in the table below.[52]

Scale Type	Operations
Nominal	Determination of equality
Ordinal	Determination of greater or less than
Interval	Determination of equality of intervals or differences
Ratio	Determination of the equality of ratios

Though subjective analyses can be expensive and time-consuming, they provide direct insight into how humans perceive image quality. Based on these types of scalings, a number of subjective video quality assessments will be examined.

4.3.1 Rank Order Method

A Rank Order Method is an ordinal method, whereby observers are given a number of samples (varying in quality level) and asked to rank them in order of increasing or decreasing quality. The data in the form of rankings are collected from each observer, and each rank is given a value. The values are averaged, from which an ordered list of samples can be compiled.[52] The benefit of using this method is that a mathematically ordered list is obtained, however, there is no indication of what the size of the difference is between each sample in the list.

4.3.2 Categorical Sort Method

A Categorical Sort Method can be used to produce nominal scales, interval scales or ordinal scales, though the last is more common. The method can be implemented by presenting observers with a series of samples and asking them to assign the samples to categories, labelled with terms such as 'excellent', 'good', 'bad', etc.[52] Each category is approximately 6 JNDs apart.[51] Similar to the Rank Order Method, this will provide an ordered list from which it is possible to analyse attributes of quality. Making an improvement on the Rank Order Method, it is possible to see the difference in stimulus intensity between each sample. However, the Categorical Sort Method can provide unreliable results. One reason for this is the lack of baseline for the images. Without a reference image, it may be difficult to decide whether a given image is 'excellent' as opposed to 'good'. Additionally, different observers may have different interpretations of the category titles.

4.3.3 Magnitude Estimation Method

The Magnitude Estimation Method is used to generate ratio scales. In this method, observers are asked to compare each test sample against a fixed reference sample that has been assigned a numerical value. Depending on the magnitude of level of quality of the test sample in comparison to the magnitude of level of quality of the reference sample, the observer is asked to assign a value to the test sample. For example, should an observer deem the test sample to be twice as good in quality as the reference sample, they would assign a value two times as large as the value assigned to the reference sample. Here, compiling the data of multiple observers is done by taking the geometric mean, rather than the arithmetic mean. Though the Magnitude Estimation Method is an improvement on the Categorical Sort Method, a drawback to using ratio scales in general is that humans think more naturally in terms of intervals rather than ratios, and so it may be difficult for observers to confidently apply values to the test samples.[51]

4.3.4 Paired Comparison Method

The Paired Comparison Method is one in which observers are given pairs of samples and asked to select the better quality sample out of the two. If the difference between a pair of samples exceeds approximately 1.5 JNDs, the magnitude of the quality difference cannot be estimated reliably because the response saturates as the proportions approach unanimity. For n samples, the size of the test becomes $\frac{n(n-1)}{2}$, i.e. $\frac{n(n-1)}{2}$ pairs of samples must be compared.[9] For a large number of samples, the test becomes incredibly large, and the Paired Comparison Method becomes more time consuming and expensive than any of the other subjective tests explored so far. Conversely, the test provides easy comparison with little bias. The Paired Comparison method will be explored in depth later in this thesis.

4.3.5 Quality Ruler Method

A Quality Ruler is a univariate series of reference images depicting the same scene, with images ordered approximately 1 JND apart in quality. A Quality Ruler and a number of test samples are given to the observer, and the observer is asked to assign each test sample to a position on the ruler.[9] As opposed to the Rank Order Method, the Quality Ruler provides results wherein it is possible to see the difference in stimulus intensity. It is not as time consuming as the Paired Comparison Method, and avoids the errors that can arise through the implementation of the Categorical Sort Method and the Magnitude Estimation Method. Additionally, as it is configured on an absolute universal scale of quality, it allows easy and quantitative indirect comparison of stimuli.

Image Quality Ruler

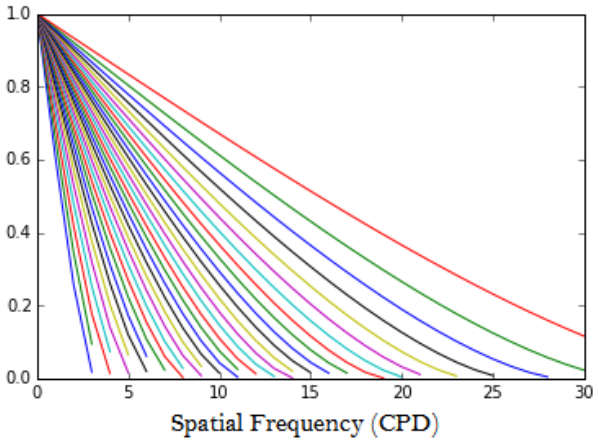


Figure 8: Aim MTFs of the images in the IQR

Based on the Quality Ruler Method, an Image Quality Ruler (IQR), as defined in ISO 20462 [9], has been constructed by ordering a series of 31 images, approximately 1 JND apart. The images were created by adjusting the sharpness of the baseline image according to the MTF of the complete system (camera,

image processing, display). Sharpness is chosen as a useful attribute to adjust, as it is easily manipulated through image processing, it is correlated with MTF, which is readily determinable, it has low scene and observer variability, and it exerts a strong influence on quality in practical imaging systems.[9] One minor disadvantage of varying sharpness is that the scale in JNDs are only valid at a single viewing distance.[51] The aim MTFs for the baseline image are shown in Figure 8, and closely conform to the shape of the MTF of a diffraction-limited lens.[9] The diffraction MTF is based on the diffraction effect, and therefore, blur, that occurs when the aperture of a camera is too small. The utilisation of this form of MTF is to make the blur as realistic and natural as possible.[32] The MTFs, $m(\nu)$, are given by

$$m(\nu) = \begin{cases} \frac{2}{\pi}\left(\arccos(k\nu) - k\nu\sqrt{1-(k\nu)^2}\right), & k\nu \leq 1 \\ 0, & k\nu > 1 \end{cases} \tag{2}$$

where k is a constant and ν is the spatial frequency in cycles per degree (CPD) at the eye of the observer over the range 0 to 30 CPD. The MTFs of the images should match the aim MTFs over each of the frequency bands 0 to 5, 5 to 10, ..., 25 to 30 CPD to within 0.05 CPD. The values for the constant k should be computed from

$$\text{SQS} = \frac{17249 + 203792k - 114950k^2 - 3571075k^3}{578 - 1304k + 357372k^2} \qquad (1 \leq 100k \leq 26) \tag{3}$$

where SQS stands for the Standard Quality Scale. The SQS is a fixed numerical scale in which one unit increase in scale value corresponds to an improvement of one JND of quality. Properties of the SQS include that the numerical scale is anchored against physical standards, and a value of zero corresponds to an image having so little information content that the nature of the subject of the image is difficult to identify.[9] Therefore, an image with SQS level 31 is the image with highest perceived quality within the ruler, and is 1 JND apart from the image with SQS level 30.
Images are created by adjusting the sharpness of the baseline image according to the MTF of the complete system, which is the camera, image processing, and display. Upon measuring the MTF of the camera and the chosen medium of display, and given the aim MTF, it is possible to determine the blur to apply to the baseline image through the cascading property of MTFs.

The viewing conditions of the IQR should be as neutral as possible to reduce observer fatigue. This can be achieved by painting the viewing room in a neutral grey colour. The use of a chin rest is encouraged, in order to ensure the viewing distance is kept constant at all times,[53] which should be at least 2500 times the 'pixel pitch' of the display.[9] The pixel pitch of a display is the distance in physical units (inches, cm, etc.) from the centre of one pixel to the centre of the next pixel. The IQR can be implemented in both hard-copy format and soft-copy format, though the use of a soft-copy ruler makes more sense, due to the large number of images that will be viewed. The soft-copy ruler should be implemented on a large display panel with a sufficiently high pixel resolution in order to allow side-by-side presentation of images. Using two displays (i.e. one image per display) may reduce the reliability of the results as it can be challenging to match any two displays exactly in colour and tone. In the soft-copy ruler,

the test image is displayed on one side, and the ruler image is displayed on the other side, with a slider beneath to move between positions on the ruler. The wall behind the display should be uniformly illuminated by flourescent tubes emulating D65 (standard daylight illuminant) so as to match the luminance of the average pictorial screen rendered on the display. This should be the only light source within the room.[53]

20 IQRs containing different scenes were previously provided by Aptina (though they are no longer available from this source). Each IQR contains 31 images. However, the images are not calibrated to SQS values 1, 2, ..., 31. Files are provided for different viewing distances, detailing the SQS value of each image in an IQR. The SQS values have a precision of 2 decimal places. The scenes depicted in the rulers cover a variety of different subjects, such as text, people, and buildings.

5 Statistical Concepts

The subjective methods explored in Section 4.3 are examples of statistical experiments. An experiment can be defined as a test in which purposeful changes are made to the input variables of a system so that observations can be made about the possible changes to the output response. Experiments can confirm or disprove theories or hypotheses about a system that have previously been established or predicted. Such experiments should be planned and conducted according to the strict statistical rules that ensure that appropriate data is collected and analysed, resulting in valid and objective conclusions. [54]
There are two parts to any experimental problem, which are closely linked: the design of the experiment and the statistical analysis of the data. [54]
A successful experiment with any number of input variables may: [54]

- determine which variables (if any) are most influential on the output response,
- determine how to set the variables so that the output response is as close to a desired value as possible,
- minimise the variance in the output response, or,
- determine how to set influential variables so that effects of uncontrollable variables are minimised.

There are three basic principles to experimental design: randomisation, replication, blocking. The most important of these, randomisation, minimises the prevalence of systematic error, resulting in errors that should theoretically be independent and identically distributed random variables. Replication is defined as an independent repeat run of each factor combination. Through replication, an estimate of the experimental error can be made. Additionally, averages can be taken over replications, obtaining a more precise result. The third basic principle to experimental design, blocking, is a design technique in which experimental units are grouped together dependent on certain similarities. This can reduce the variance of the output response due to factors that may influence the output but in which there is no relevance to the objectives of the experiment.[54]
Based on these three principles, two examples of experimental designs are: [54]

- One-factor-at-a-time (OFAT) experiments in which a single factor is varied over a range while the other factors are kept at a constant predefined baseline level;
- Factorial experiments, which build upon OFAT experiments by also varying factors together to examine the interactions between factors.

There are variations of both types of experimental designs listed above, and alternatives to both of these designs. As well as fulfilling the three basic principles, an experiment should be constructed to best fit the aims that the experiment wishes to achieve.[54]
Upon completion of the experiment and statistical analysis, further testing and experimentation should be performed to validate the conclusions of the experiment.[54]

5.1 Hypothesis Testing

In general, an experiment should be designed with a theory or hypothesis in mind. Through experimental procedures, the proposed hypothesis will be rejected or accepted. This forms a hypothesis test in which there is a null hypothesis H_0 and a proposed (alternate) hypothesis H_1. Should there be enough evidence, a null hypothesis is rejected in favour of the alternate hypothesis. Hypothesis tests can be two-sided or one-sided.
Two types of errors can be produced when conducting hypothesis tests. A Type I error is the case in which a hypothesis test finds that there is sufficient evidence to reject H_0, even though H_0 is true. A Type II error is the case in which a hypothesis test deduces that H_0 should not be rejected, even though H_0 is false. The probabilities of producing Type I and II errors are α and β, respectively. That is,

$$\alpha = P(\text{Type I error}) = P(\text{reject } H_0 | H_0 \text{ true})$$

$$\beta = P(\text{Type II error}) = P(\text{accept } H_0 | H_0 \text{ false})$$

α can also be known as the significance of the test, and $1 - \beta$ as the power of a test. In an experimental problem, the significance of the hypothesis must be set by the experimenter. A typical value of α is 0.05. Smaller values of α will produce more accurate results, at the cost of producing results with less precision. This will be explored further in Section 5.1.2. Further, the experiment must be designed based on the significance level of the test and a desirable power to achieve. This is known as power analysis, and will be explored further in Section 5.1.1.
The general methodology of a hypothesis test is as follows:

1. Establish the null and alternate hypotheses
2. Compute a test statistic based on the data and the underlying distribution of the test
3. Compare the test statistic to the quantile(s) of the underlying distribution of the test
4. Accept or reject the null hypothesis in favour of the alternate hypothesis

The significance of the test is utilised in the third point, in calculating the quantiles. If the hypothesis test is two-sided, the $100 \cdot \frac{\alpha}{2}\%$ and $100 \cdot (1 - \frac{\alpha}{2})\%$ quantiles are computed. If the hypothesis test is one-sided, the $100 \cdot \alpha\%$ quantile is computed. The test statistic computed in the second point is computed dependent on the hypothesis test implemented, and the underlying distribution of the test. Some examples of hypothesis tests with different functions will be explored later in this section.
When conducting multiple hypothesis tests in parallel, it is wise to apply the Bonferroni correction, whereby the significance level of the overall test is divided amongst the individual hypothesis tests equally.[55]
An alternative method to conducting hypothesis tests is through the use of p-values, which is the probability of achieving a more extreme test statistic in the direction of the alternate hypothesis, given that the null hypothesis is true. If the p-value is smaller than or equal to the significance level of the test, there is sufficient evidence to reject H_0.

t Test

A hypothesis test for the mean value of a set of normally distributed data which unknown variance can be conducted by use of Student's t distribution. Aptly named the 't Test', the hypothesis test can be conducted based on an estimate s^2 of the unknown variance σ^2 of the normally distributed population.t tests can further be divided into two categories, the t test for comparison of means of two populations, and the t test for the mean value. The former is based on an additional assumption of equal (but unknown) variances, and the test statistic in this case is

$$t = \frac{(\bar{x_1} - \bar{x_2}) - (\mu_1 - \mu_2)}{s\left(\frac{1}{n_1} + \frac{1}{n_2}\right)^{\frac{1}{2}}},$$

where $\bar{x_i}$ is the average of all n_i observations, m_i is the hypothesised mean of a group i, and s is the pooled sample standard deviation, i.e.

$$s^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

The test statistic should be compared to Student's t distribution with $n_1 + n_2 - 1$ degrees of freedom.
The test statistic of a t test for the mean value of one population is

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}},$$

where $\bar{x}$ is the average of all n observations, μ_0 is the hypothesised mean of the population, and s is the sample standard deviation, i.e. $s^2 = \frac{\sum(x-\bar{x})^2}{n-1}$. The test statistic should be compared to Student's t distribution with $n - 1$ degrees of freedom.[56]

Z Test

Another hypothesis test for the mean value of a set of normally distributed data can be conducted by use of the Normal distribution. However, this test is

more powerful and relies on a known variance σ^2. Like the t test, there are two notable Z tests, a one sample test of the mean value, and a two sample test of the difference between the mean values. The test statistic of the former is

$$Z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}},$$

where $\bar{x}$ is the mean over n samples and σ^2 is the population variance. The test statistic is compared to the Normal distribution.
Likewise, the test statistic of the two sample test,

$$Z = \frac{(\bar{x_1} - \bar{x_2}) - \mu_0}{\left(\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)^{\frac{1}{2}}},$$

where $\bar{x_i}$ is the mean over n_i samples and σ_i^2 is the population variance, is compared to the Normal distribution.[56]

Chi-Squared Test of Variance

The Chi-squared distribution is a positively skewed distribution that can aid the analysis of variance. In such a case in which it is desirable to test an unknown variance of a sample of normally distributed data, a hypothesis test known as the Chi-squared test can be useful. The test statistic

$$\chi^2 = \frac{(n-1)s^2}{\sigma_0^2},$$

where s^2 is the sample variance of a sample with size n, and σ_0 is the hypothesised population variance, is compared to the Chi-squared distribution with $n-1$ degrees of freedom.[56]

Chi-Squared Goodness of Fit Test

The Chi-squared test can also refer to a 'Goodness of Fit' test to analyse whether sampled data fits an assumed (known) distribution. The test statistic is

$$\chi^2 = \sum_i^k \frac{(O_i - E_i)^2}{E_i},$$

where O_i is the frequency of the ith group, and E_i is the expected frequency of the ith group. The term 'group' could denote an interval containing samples of data. The test statistic should be compared to the Chi-squared distribution with $k-1$ degrees of freedom.[56]

Normality Tests

Most of the hypothesis tests mentioned thus far rely on an assumption of normality, although, the tests are reasonably robust to deviations from this assumption. In this section, a number of ways to assess normality will briefly be explored.
Normality can be assessed graphically by the use of quantile-quantile (Q-Q) plots. In Q-Q diagrams, the sample data are ordered and plotted against expected quantiles of a Normal distribution. Thus, for samples that are normally

distributed, Q-Q plots would form approximately straight lines that go through the origin.[57] There exist various other graphical methods of assessing normality, such as probability-probability (P-P) plots, histograms, and boxplots.[58] There also exist analytical methods of assessing normality, which should be used in conjunction with graphical methods. Tests include the Kolmogorov-Smirnov test and the Shapiro-Wilk test, the latter of which is a more powerful test. As per research conducted by Royston in 1995, the Shapiro-Wilk test may be used for sample sizes within a range of 3 and 5000. For sample sizes on the lower end of the spectrum (≤ 30), the power of the Shapiro-Wilk test is still relatively low.[59]

5.1.1 Power Analysis

As previously mentioned, there can exist two types of errors in conducting a hypothesis test. The first, a Type I error occurs when the null hypothesis is falsely rejected. A Type II error defines when a false null hypothesis is not rejected. The probability of a Type II error is known as β, and is commonly desired to be at most 0.2.[54],[60] The power of a test is

$$Power = 1 - P(\text{Type II error}),$$

and is thus set to be ≥ 0.8 to ensure a statistically powerful test. The power of a test is closely linked with the significance level and sample size of a test, and can provide information on the smallest sample size required to achieve a power of a desired value. By setting the smallest deviance from the null hypothesis that the test should detect, the effect size, it is possible to determine the minimum sample size required to achieve a level of power.[60]

5.1.2 Confidence Intervals

Given the significance level of a hypothesis test, it is possible to compute confidence intervals around an obtained result that describe a region of values that the result covers. As the size of α decreases, the interval that contains the result becomes wider, and less precise. Naturally, setting $\alpha = 0$ and removing the probability of a Type I error occuring would produce confidence regions that cover an inifite range of values.
A confidence interval surrounds the obtained result based on the underlying distribution of the hypothesis test. For example, the confidence region in a two-tailed Z test with significance level 0.05 is $\bar{x} \pm z_{0.025}\frac{\sigma}{\sqrt{n}}$.

5.2 Maximum Likelihood Estimation

The Maximum Likelihood Estimation (MLE) method is a method for estimating the parameters of the underlying distribution of a sample of data. Though popularised by R. A. Fisher between the years of 1912 and 1922, the method had existed and been used before this time by other mathematicians.[61]
The methodology of computing a maximum likelihood (ML) estimator of a parameter is as follows:

1. Specify the joint probability density function of the observations.

2. The likelihood function is the joint probability density function when the observations are fixed and the parameters are allowed to vary; compute the likelihood function for the observations.

3. Compute the log-likelihood function, which is the logarithm of the likelihood function. As the observations may often be independent and identically distributed, the joint probability density function in point 1 will then be the product of probability density functions. Taking the logarithm of the likelihood function would return a summation as opposed to a product, simplifying later computations.

4. Differentiate the log-likelihood function with respect to the parameter of interest.

5. Find the maximum value of the parameter by equating the differentiated function to 0. As the logarithm is a monotonically increasing function, the maximum of the likelihood function and the maximum of the log-likelihood function is the same. The maximum value is the ML estimator of the parameter.

For more complex problems, numerical optimisation methods [62] must be performed on the likelihood function. However, the numerical optimisation methods for multi-parameter maximum likelihood estimation are sensitive to the initial parameter(s). Thus, good starting estimates for the parameters are required.[63] Numerical optimisation methods include Nelder-Mead and quasi-Newton.
The ML estimator is asymptotically unbiased, and tends to the true value as the sample size increases. Additionally, as the sample size increases, the variance of the estimator decreases and the distribution of the estimator approaches normality.[63]

5.3 Bootstrapping

Initially proposed by B. Efron in 1979, bootstrapping is a resampling technique used to make estimations on a samples of data.[64] The method works by drawing from the sample of data, with replacement, and obtaining estimates for parameters and their uncertainty based on the resampled data. The reasoning behind the use of this method is based on the Law of Large Numbers [65], which describes that with a sufficient sample size, the empirical distribution is a good approximation of the true distribution.
There are two main methods of bootstrapping, namely parametric and non-parametric bootstrapping. The latter is a method carried out when the underlying distribution of the sample data is unknown, and the data is uniformly resampled without any assumptions made to the data. The former method operates by fitting a parametric model to the data and drawing instead from the fitted model.
The bootstrap method is, in general, asymptotically consistent, and can be employed for a wide range of purposes. Examples include:

- Non-parametric bootstrapping when the underlying distribution of the sample data is unknown, in order to estimate parameters such as the mean and variance of the data

- Non-parametric bootstrapping when the sample size is small, to test for the normality of the data (or test for any other distribution)
- Non-parametric or parametric bootstrapping to compute confidence intervals of parameters, such as the mean and variance of the data

The methodology of the parametric version of the final example is as follows:

1. Given a sample of data of size n, fit a parametric model to the data, perhaps by the maximum likelihood estimation method.
2. Specify a large number of bootstraps m.
3. Generate m sets of n samples from the fitted model.
4. Estimate the desired parameter(s) for each of the m bootstraps.
5. Compute the differences between the bootstrapped parameter(s) and the parameter(s) fitted in point 1.
6. Order the differences.
7. Set the significance level α of the confidence interval.
8. Select the $\frac{\alpha}{2}\%$ and $(1-\frac{\alpha}{2})\%$ quantiles of the ordered differences. The fitted parameter plus the quantiles for the differences form the confidence interval(s) for the parameter(s).

For more information on the popular resampling method, see [66].

6 Method

Through the exploration of already established subjective image quality evaluation methods, it is clear that the Image Quality Ruler provides the most accurate and reliable results, with minimal time consumption. Additionally, the IQR operates with a universal absolute scale of image quality. In extending this idea for the usage of moving image (video) quality evaluations, this thesis focuses on the creation of a Video Quality Ruler (VQR).
The requirements of the Quality Ruler were defined in the previous section. In the interests of building an IQR with new SQS values, a minimum of 3 scenes (and preferably 6) should be averaged to determine a reported SQS value for each experimental treatment.[9] However, this would increase the size of the test drastically. For this thesis, to establish and test methodology, only one scene is used. Considering this, there were two main options when it comes to the creation of the VQR.

1. Create a series of video clips, varying only in sharpness. Find the JNDs between each clip through Paired Comparison tests and create a VQR with video clips ordered 1 JND apart.
2. Create a series of video clips, varying only in sharpness. Compare the video clips against an Image Quality Ruler whose reference images hold similar attributes, in order to determine the SQS values of each of the video clips. Order the video clips and create a VQR approximately 1 JND apart.

Both options are now explored in detail in order to assess which is the preferred method to use in this thesis.

Option 1

The method for Option 1 is as follows:

(a) Create a large number of video clips varying only in sharpness, spanning a pre-specified range of sharpness. The video clips can be calibrated at this stage either by following the specifications provided in ISO 20462 or by formulating a new set of calibration values. As a first step, it is more efficient to use the ISO 20462 calibration levels and modifying after initial studies if required.

(b) Pairs that will inevitably be extremely far apart in quality, for example, the least blurred image against the most blurred image, will not be presented to observers. Therefore, the size of the test can be reduced from $m(m-1)/2$, where m is the total number of video clips.

(c) The time taken for observers to make a decision about each pair will be determined through small pilot studies. Pilot studies can also determine whether or not the levels of blur defined in ISO 20462 for the IQR are appropriate for a VQR.

(d) Elements of the pilot study may be modified dependent on the results obtained, and a larger study is conducted. For example, if the calibration levels specified in ISO 20462 do not apply well to video clips, the calibration may need to be altered. Other factors that may need to be altered depend on the time taken or the physical environment of the test.

(e) Upon completion of the larger study, select the pairs for which there is approximately 75% agreement amongst observers.

(f) Order the selected pairs, thus creating a ruler of video clips.

(g) Perform a validation experiment of the VQR, whereby video clips from the ruler are tested against the ruler.

Option 2

The method for Option 2 is as follows:

(a) Create more than 31 video clips varying only in sharpness, spanning the same range of sharpness as the Image Quality Ruler.

(b) Ask observers to compare the video clips to the Image Quality Ruler, resulting in SQS values for each of the video clips.

(c) Select video clips for each SQS value, whose SQS distributions have the smallest standard deviation, i.e. the video clips for which there was the most agreement for their SQS values. Discard the remaining video clips. This will result in 31 video clips 1 JND apart, creating a VQR.

(d) Perform a validation experiment of the VQR, whereby video clips directly extracted from the VQR are compared in a Paired Comparison test. The results should prove that successive video clips are 1 JND apart, i.e. there should be approximately 75% agreement amongst observers for level of quality of video clips.

It is possible to compare video clips to the Image Quality ruler, and has been implemented successfully in the past.[14] However, in this method, the viewing distance wasn't regulated. Test video clips were displayed at the correct viewing distance (though without the use of a chin rest), but ruler images were displayed on a tablet held in the observers hand. This may have produced unreliable results.

Comparison of the two options

- It is easy to see that Option 1 will take longer to implement and will require more participants than Option 2. This is because Paired Comparison tests make up the larger part of Option 1, and is far more time- and resource-consuming. Whilst Paired Comparison tests are implemented in Option 2, they are utilised during the validation stage, during which far fewer participants are required, and far fewer pairs need to be compared.
- Option 2 makes use of 2 tests, namely the Image Quality Ruler test and the Paired Comparison test that was used to build the Image Quality Ruler. In comparison, Option 1 makes use of 1 test only, the Paired Comparison test. This means that Option 2 will theoretically produce more errors than Option 1.
- Since Option 2 depends heavily on the rules based upon the IQR, should video clips provide a different subjective effect to images, the rules provided in ISO 20462 will essentially be void. Thus, it will be impossible to create a complete ruler.
- Due to the nature of Option 2, it will be possible to produce a ruler consisting of a maximum of only 31 video clips, as opposed to Option 1, where it will be possible to produce a ruler of any size.
- In relation to the previous point, Option 1 will allow for ruler creation that spans any range of quality, whereas Option 2 will be limited to the range set by the Image Quality Ruler as defined in ISO 20462.

Comparing the advantages and disadvantages for each of the two options, it is decided to implement Option 1. Though this option takes longer to implement, the results are more accurate and reliable, and a better basis for future VQR creation. It should be noted that most of the variables that can arise in the creation of a VQR should be kept as close to the settings of the IQR in order to allow for easy comparison of the two methods.

6.1 Video Clip Capture

The initial step in creating the VQR is to capture a video clip that will be the basis for the ruler. This video clip will later be calibrated to produce multiple video clips varying in sharpness.

The baseline video clip for the VQR should follow certain rules. Excluding the effect of the attribute varied within the Quality Ruler, the reference stimulus should have high image quality, with pleasing colour and tone reproduction, and an absence of significant degradation from artefacts under the existing viewing conditions.[9]

To ensure the best possible lighting, the scene is chosen to be outdoors, in sunny weather. In a scene containing only one type of lighting, it is simple for the camera to compensate for this and adjust the white balance so that colours are reproduced as authentically as possible. Though this lighting provides a large amount of illumination for the scene, it can also create large and dramatic shadows, resulting in high dynamic range. As previously examined, high dynamic range may be difficult reproduce with the camera. Therefore, scenes that contain as few dark shadows and bright highlights as possible should be chosen.

The requirements on the chosen scene were as follows: participants should have many objects that they can focus on and use to detect the difference in quality between two video clips. Conversely, the scene should not be too crowded. The chosen scene should contain at least one face, clear and legible text, and some sort of detailed pattern. Face(s) and text are included within the scene as they are objects that humans can easily recognise and interact with on a daily basis. The inclusion of a detailed pattern within the scene is linked to the idea behind the MTF: as sharpness decreases, details above certain frequencies will not be discernible.

The camera used to capture the scene is a Canon EOS-1D C, with a Canon TS-E 45mm f/2.8 tilt-shift lens. This camera and lens combination is chosen as it is one of the best quality cameras for video use on the market.[67],[68] The lens has no zoom function, and must be focused on the subject manually (as opposed to auto-focus performed by the camera).

The white balance of the camera was set to approximately 5200K, to account for the sunlight in the scene. It is desirable to keep the ISO Speed as low as possible to ensure that noise in the video clip is at a minimum. As ISO Speed, shutter speed, and aperture all interact with each other to control the brightness level in the reproduced image, shutter speed and f-number (aperture size) are determined by trial and error. It is desirable to keep the shutter speed as fast as possible, in order to reduce the effects of motion blur. The aperture should be small enough so that most of the scene is in focus, but large enough so that diffraction blur does not occur. Both the shutter speed and aperture size should be controlled so that the correct luminance is exhibited. The shutter speed is chosen as 1/125, i.e. a single frame is captured in $\frac{1}{125}$th of a second. The aperture is set to f/11 (where the smallest possible aperture is f/22, and the largest is f/2.8). The lens is focused using the optical zoom function on the viewfinder, i.e. zooming in on the scene on the viewfinder before the images are captured, and adjusting the focus on the lens accordingly. The focus is applied to the stand-alone bus stop in the centre of the scene, but as the aperture is relatively small, a good proportion of the scene is in focus. The video clip is recorded at a pixel resolution of 4096 x 2160, and the maximum frame rate for this resolution, 25 fps, is selected. In a bid to obtain the rawest possible set of images, noise reduction and highlight tone priority algorithms are turned off. However, the

camera settings do not allow for the removal of sharpening filters. The images are recorded in the sRGB colour space. It is not possible to retrieve the raw images for a video clip, so the images are saved as a Motion JPEG (MJPEG) video.
A roughly 12 second segment (295 frames) of the scene is selected. A few frames are shown below.

Figure 9: Frames 7, 94, and 260 of the captured scene

The scene is static, with the exception of the moving bus, the moving person, and a few pedestrians and cyclists in the background. As can be seen, there is text in a few areas within the scene. These can be found in the bus stop shelter, in the stand-alone bus sign, next to the shelter, and on the bus itself, moving away from the camera. This ensures that there is text on both still and moving parts of the scene. There is a person walking towards the camera,

allowing a clear face to be seen for approximately half of the scene. The person is wearing a scarf with a detailed pattern, fulfilling the third requirement of the scene. The person and the bus appear in the scene at different times to allow participants to focus on one thing at a time. There are many other objects within the scene, giving participants of the test a wide selection to choose from when making a decision upon viewing the clip. This includes bushes and trees, cyclists, and buildings in the background. There are few extremely dark shadows and extremely bright highlights, keeping the dynamic range at an acceptable level.
The first frame contains all the static parts of the scene and the bus, and the last frame contains all the static parts of the scene and the person walking towards the camera. When the video clip will be played in the test, it will be played in a loop. The large difference between the last and first frame will give an indication to the participants that the video is restarting. This should hopefully reduce boredom.

6.2 Calibration

The selected scene now undergoes calibration to create multiple video clips with different levels of sharpness. As mentioned in Section 3.5, low-pass filters reduce the frequency content (and thus the sharpness) of an image. The calibration process follows a number of steps to create and apply the low-pass filter. A script is written in Python to carry out the majority of the process described in this section, unless otherwise stated.
The general methodology applied in the calibration process is as follows:

1. Measure the camera OECF in controlled conditions.
2. Change the file format of the video clip from JPEG to YUV420 (i.e. obtain a more raw form of the video).
3. Separate the video clip into PNG images, where each image is a single frame.
4. Separate each frame into the three colour channels.
5. Dependent on the camera OECF, linearise the colour space of each colour channel of each frame.
6. Decide the factor by which the pixel resolution of the frames will be reduced.
7. Recall that image resampling can cause artifacts to appear within the resampled image. This can be solved by the use of low-pass filters. Create a low-pass filter(s) to apply before decimation of the frames.
8. Resample the pixels of each colour space of each frame to reduce the pixel resolution.
9. Create blur filters to apply to each colour channel of each frame, and apply.
10. Non-linearise the colour space of each colour channel of each frame by making use of the OECF of the camera.

11. Combine the 3 colour channels of each frame.
12. Crop each frame slightly to remove the edges.
13. Combine the PNG frames to create multiple video clips of varying sharpness in YUV420 format.

The preparatory steps to create the blur filter mentioned in point 9 are as follows:

1. Measure the camera MTF in controlled conditions.
2. Create a low-pass filter(s) of the same type as used in point 7 above. Note that the filter(s) created here are slightly different to the filter(s) created in point 7.
3. Determine the MTF of the screen that will be used to display the video clips in the experiment.
4. Determine an appropriate distance at which to view the video clips in the experiment.
5. Compute the values of k from Equation (3) that will be used as the parameter in the diffraction MTF, Equation (2).
6. Transform the constants k dependent on the chosen viewing distance for the experiment.
7. Depending on the new k values, create filters describing the aim diffraction MTFs.
8. Compute the blur (low-pass) filters that should be applied to the frames.

The steps briefly described will now be explained in detail. The camera MTF is measured by capturing a video clip of the test chart in Figure 3. The settings applied to the camera mimic the settings applied to the camera whilst filiming the selected scene, with one exception. The test chart is placed in an Image Lab which is completely dark and free of external lights. Two tungsten lights are placed in the Image Lab to shine light onto the test chart at an angle of 45 degrees. White umbrellas used to reduce the intensity of light are placed over the bulbs. Thus, the white balance applied during video capture correct for tungsten lights, and not for sunlight as in Section 6.1. Additionally, a luminance measurer is used to check that the luminance is roughly uniform over the whole test chart. As the lens used is a manual focus lens, a detailed image is temporarily placed over the test chart, and optical zoom on the viewfinder is used to assess whether the lens is correctly focused. A Python script created by Axis Communications computes the MTFs from the frames of the video clip. The MTF of the frames are corrected by inversion of the OECF, which is measured in the same conditions by capturing the test chart in Figure 5. Another Python script created by Axis Communications is used to quantize the measurements made. The MTFs are averaged over a number of frames within the middle of the video clip. The camera MTF is pictured in Figure 10.

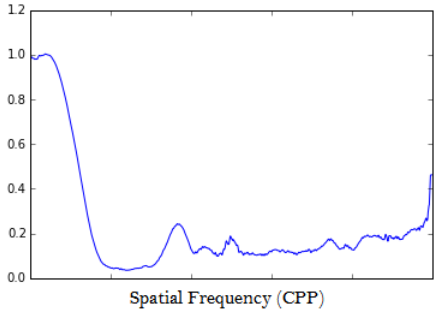


Figure 10: Camera MTF

The pixel resolution of the images provided in the IQR are roughly close to one quarter of the current pixel resolution of the video clip captured for this thesis. Thus, a decimation factor of 4 is chosen to make the methodology used in this thesis as close to the methodology used in ISO 20462. There is no indication of what low-pass filter was used to decimate the images in the IQR.[11] Thus, a Lanczos filter of order 3 is used, as it appears to be the best filter for resampling images, as mentioned in Section 3.5. Additionally, it is decided that decimation should occur in two steps. In both steps, the pixel resolution should be reduced by half, resulting in an overall reduction by a factor of 4. Note that the Lanczos filter in the frequency domain is the 2 dimensional Fourier transform of the function in Equation 1.

As mentioned in Section 3.4.1, the MTF is the modulus of a frequency response. Thus, it is a function described within the frequency domain of an image. Applying frequency filters to an image would naturally alter the MTF of said image. Recallling the cascading properties of the MTF, it is easy to compute the altered MTF of the image, if the frequency filter and original MTF of the image are known. Reversing this logic, it is possible to compute the frequency filter to apply if the aim MTF and original MTF are known. This knowledge is applied to the aims of this thesis. Figure 11 demonstrates the image alterations applied within the frequency domain of the frames with pixel resolution 4096x2160. In order to compute the blur filters that need to be applied to the video clips, the MTFs and frequency filters of the entire system must be computed first.

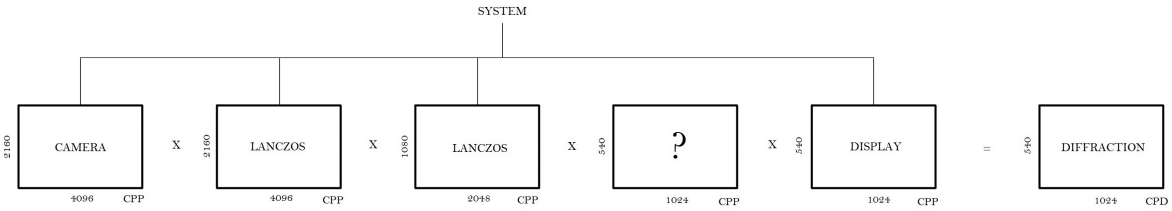


Figure 11: Calibration Process

The frequency filters are created as matrices. It should be noted that the multiplication and division between the matrices is element-wise multiplication and division. The camera MTF found previously is a set of values as opposed to a

function, so the values are interpolated to create a 2-dimensional matrix of the same size as the pixel resolution of the image.

As seen in Figure 11, different sized filters must be applied to the frames at different stages. This makes calculating the system MTF (camera MTF · Lanczos MTF · Lanczos MTF · display MTF) impossible. Thus, each matrix that makes up the system matrix is created to the size of the pixel resolution of the frames by oversampling the MTFs. The blur matrix will later be cropped to the size of the image it will be applied to, removing the oversampling. In creating 2-dimensional Lanczos filters, one is oversampled by a factor of 2 and the other remains the same.

The MTF of a display is derived by assuming that the pixels of the screen are 'perfect', i.e. each pixel forms a perfect square and there is no space between one pixel and the next. The latter condition indicates that the pixel pitch of the display is equal to the width and height of each pixel. Let the pixel pitch of the display be denoted by h. The MTF of the display is [10]:

$$m(\nu) = |\text{sinc}(\nu h)|$$

In units of CPP (cycles per pixel), the MTF is

$$m(\nu) = |\text{sinc}(\nu)|$$

The matrix describing the display MTF is created by mapping the MTF over spatial frequencies from 0 to 1, multiplied by a factor of 4 to oversample the function.

The screen that will be used to display the video clips is an EIZO ColorEdge CG276 monitor. The pixel pitch of the display is 0.2331mm. The diffraction MTFs formulated in ISO 20462 are described in units of CPD (cycles per degree). As seen in Figure 11, the units of the other elements within the calibration process are CPP. Thus, the diffraction MTFs must be transformed so that the units match the remaining filters and MTFs. The conversion between CPD and CPP is

$$\nu_{\text{CPD}} = \frac{\pi D}{180h}\nu_{\text{CPP}}, \tag{4}$$

where D is the distance between the display and the viewer, h is the pixel pitch and ν_i the spatial frequency.
The images provided in the IQR are calibrated to distances of 25 inches, 34 inches, and 43 inches. It would be wise to choose one of these distances so that it is possible to compare results at a later stage. A distance of 34 inches between display and observer is selected. However, the pixel pitch of the display used to create the IQRs is 0.250mm [11]. Therefore, the equivalent viewing distance for the EIZO display is $34 \cdot \frac{0.2331}{0.250} = 31.7$ inches. In order to oversample the aim MTF, the spatial frequency is also multiplied by the decimation factor (in this case, 4). With a pixel pitch of 0.2331mm = 0.009177 inches a viewing distance of 31.7 inches, and a decimation factor of 4, Equation (2) becomes

$$m(\nu_{\text{CPP}}) = \frac{2}{\pi}\left(\arccos(k\frac{4\pi D}{180h}\nu_{\text{CPP}}) - k\frac{4\pi D}{180h}\nu_{\text{CPP}}\sqrt{1-(k\frac{4\pi D}{180h}\nu_{\text{CPP}})^2}\right) \tag{5}$$

The values of the constants k are obtained from Equation (3) and stored to 4 significant figures. With these k values, matrices mapping the diffraction MTFs in 2 dimensions are created. Similar to the display MTF, the diffraction MTFs are created over spatial frequency values covering a unit circle multiplied by the constant $\frac{4\pi D}{180h}$.

The blur filters are now calculated by dividing the aim diffractions MTFs by the system MTF. As the blur filters are applied to the frames after their pixel resolutions have been decreased, the blur matrices are cropped to select the middle 1024x540 part of the matrices. Cropping to select the centre of the filter also removes the oversampled parts of filters that were created to account for the change in matrix size.

The video clip itself is converted from JPEG to YUV 420 format using a program known as FFmpeg, as this is the format that is accepted by the software that will be used during the testing stages. The conversion to YUV 420 format unpacks the JPEG image, thus obtaining a more raw form of the video clip. Using the same program, the video clip is separated into single frames and stored in a PNG format, a lossless compression format. The frames are saved as PNG files as opposed to YUV files as image filtering must be completed on linear colour channels.

(a) Blurriest (b) Sharpest

Figure 12: Frame at both ends of the sharpness spectrum

The frames are then entered into Python where they run through a pipeline in which the colour channels are separated then linearised by inversion of the OECF previously found. The pipeline then moves to apply the first Lanczos filter before resampling the frame by selecting every second pixel in both the x and y directions. This is repeated with a slightly different filter due to the current pixel resolution of the frame at this stage within the pipeline. The blur filters are then applied. Following this, the colour channels are reverted back to the non-linear format by use of the OECF, and combined. The resulting image has slightly distorted edges due to the method by which image filtering is conducted. Recall that image filtering within the spatial domain is conducted using neighbouring pixel data. Pixels close to the edges will therefore have a lack of available data, resulting in distortions. The resulting frames are all cropped to the same, previously determined, size. The size was determined simply by viewing the blurriest frames and selecting a sufficient crop area instinctively. The frame data are now converted back to 8-bit integer format, and saved in PNG format. The frames are combined using FFmpeg and saved in YUV 420 format. Figure 12 shows a frame at both ends of the sharpness spectrum.

6.3 Theoretical Preparation for the Paired Comparison Test

The results of a Paired Comparison Test can be modelled in a number of ways, which will be explored here. The model for the Paired Comparison test should also provide a suitable model for the validation test of the VQR. Upon determining the best general set of models, more practical preparations will later be made to prepare for the study.

6.3.1 Binomial Distribution

A first intuitive guess in analysing the results of a Paired Comparison test, is modelling the results of each pair with a Binomial distribution. For a difference of 1 JND, the probability of success (.e. correct selection of the best image) for a pair would be 0.75. Similarly, a difference of 0 JND would be based upon a Binomial distribution with probability of success 0.5.

$$X_0 \sim B(n, 0.5),$$

$$X_1 \sim B(n, 0.75),$$

where X_0, X_1 are the variables representing the number of correct results on a pair of video clips that are 0 and 1 JND apart, respectively, and n is the sample size. The Binomial distribution is useful in this instance as it computes the discrete number of successes in a group of n independent binary experiments, each of which is successful with probability $p = 0.5, 0.75$.

If the difference between two clips can be modelled with an arbitrary distribution, the psychological scale values of each individual stimulus should be modelled dependent on the arbitrary distribution of the differences. In this case, it is difficult to obtain the distribution of the SQS values of each individual video clip. However, if the sample size n is large enough, by the Central Limit Theorem [69], it is possible to obtain a Normal approximation to the Binomial distributions of the JND between images. That is, for a random variable X from a Binomial distribution with a sufficiently large sample size n and probability of success p, the Central Limit Theorem implies that

$$\frac{X - np}{\sqrt{np(1-p)}} \xrightarrow{d} N(0, 1)$$

Thus,

$$X_0 \sim N(0.5n, 0.5n(1 - 0.5)) = N(0.5n, 0.25n),$$

$$X_1 \sim N(0.75n, 0.75n(1 - 0.75)) = N(0.75n, 0.375n)$$

The difference of two Normal distributions is still a Normal distribution, therefore a Normal approximation to the Binomial distribution for the JND values between each pair allows for a Normal assumption of the SQS values of each video clip.

6.3.2 Bradley-Terry Model

R. A. Bradley and M. E. Terry introduced a model in 1952 for Paired Comparison tests known as the Bradley-Terry Model.[70]
For a pair of stimuli A and B, the probability of perceiving stimulus A with a higher psychological scale value than stimulus B is

$$P(S_A > S_B) = \frac{\pi_A}{\pi_A + \pi_B},$$

where S_i is the psychological scale value of stimulus i, and π_i is the number of times a stimulus i is chosen. The Bradley-Terry model suggests substituting π_i with an exponential function, such that

$$\pi_i = \exp(\gamma_i)$$

Hence,

$$\begin{aligned} P(A > B) &= \frac{\exp(\gamma_A)}{\exp(\gamma_A) + \exp(\gamma_B)} \\ &= \frac{1}{2} + \frac{1}{2}\tanh\left(\frac{\gamma_A - \gamma_B}{2}\right) \end{aligned} \tag{6}$$

Since $P(S_A > S_B) = P(S_A - S_B > 0)$, Equation (6) becomes the cumulative distribution function of the difference in psychological scale values, $S_A - S_B$. If γ_i is substituted with μ_i/s, the function in Equation (6) matches the cumulative distribution function for the Logistic distribution with mean as the mean psychological difference between stimuli A and B, $\mu_A - \mu_B$, and scale parameter s.[71] In this way, the Bradley-Terry model essentially performs logistic regression on pairs of stimuli.[70]
The Bradley-Terry model can be adapted to use within this thesis. Here, 1 JND is defined as 0.75 proportion agreement amongst observers. Thus, $P(S_A > S_B) = 0.75$ for a mean quality difference of 1. Equation (6) becomes

$$0.75 = \frac{1}{2} + \frac{1}{2}\tanh\left(\frac{1}{2s}\right)$$

Rearranging this obtains a scale parameter of approximately 0.9102. Therefore, the quality difference between two video clips follows a Logistic distribution with scale parameter 0.9102. Following this model, the estimator for the perceived difference in quality, dependent on the proportion agreement p, is

$$\text{JND} = 1.8204 \operatorname{arctanh}(2p - 1)$$

As an example, a difference of 2 JND requires participant agreement of approximately 0.9000.

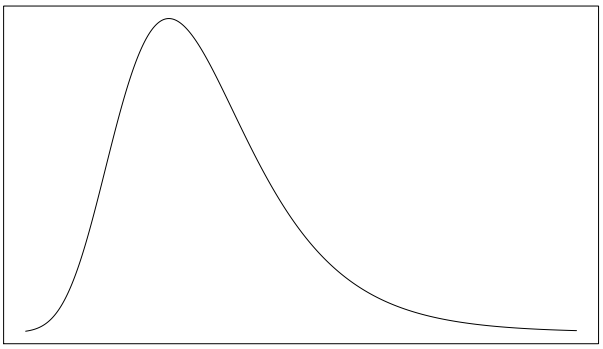


Figure 13: Gumbel PDF

For $S_A - S_B$ to have a Logistic distribution, the random variables S_A and S_B must have Gumbel distributions.[71] As pictured in Figure 13, the probability density function of the Gumbel distribution is skewed. Application of the Bradley-Terry model on a video quality test would suggest that the quality scores of any video clip would be perceived according to a skewed distribution. Assuming this model, most participants should perceive any given video clip as the true quality level. Out of the remaining participants, the spread of opinions should be uneven, with a larger variance amongst participants who should perceive the video clip as better than true quality than those who should perceive the video clip as worse than true quality. In the absence of bias, this seems like an unrealistic model.

6.3.3 Thurstone's Law of Comparative Judgement

American psychologist L. L. Thurstone proposed in 1927 a Law of Comparative Judgement, wherein one can model measurements obtained from a paired comparison experiment.[72] In particular, it was established to provide insight in psychophysical experiments. Like Fechner, Thurstone proposed that sensation could be measured only indirectly through measurement of stimulus discrimination.[50]
Thurstone's model assumes that any sensation is perceived normally, i.e. can be modelled with a Gaussian distribution.[50] Let there be two stimuli A and B, and their corresponding psychological scale values S_A, S_B, respectively. Then,

$$S_A \sim \mathrm{N}(\mu_A, \sigma_A^2), \qquad S_B \sim \mathrm{N}(\mu_B, \sigma_B^2)$$

The perceived difference between the two stimuli, i.e. the difference between their psychological scale values, is $S_A - S_B$. The sum of two normal random variables is still a normal random variable. Thus,

$$S_A - S_B \sim \mathrm{N}(\mu_A - \mu_B, \sigma_A^2 + \sigma_B^2 - 2\rho_{AB}\sigma_A\sigma_B) = \mathrm{N}(\mu_{AB}, \sigma_{AB}^2),$$

where ρ_{AB} is the correlation between S_A and S_B. In this way, the perceived difference in stimuli can be modelled normally. The probability of perceiving A

with a higher psychological scale value than B is $\mathrm{P}(S_A > S_B)$.

$$\begin{aligned}
\mathrm{P}(S_A > S_B) &= \mathrm{P}(S_A - S_B > 0) \\
&= \int_0^\infty \frac{1}{\sqrt{2\pi\sigma_{AB}^2}} \exp\left(\frac{-(x-\mu_{AB})^2}{2\sigma_{AB}^2}\right) dx \\
&= \int_{-\mu_{AB}}^\infty \frac{1}{\sqrt{2\pi\sigma_{AB}^2}} \exp\left(\frac{-x^2}{2\sigma_{AB}^2}\right) dx \\
&= \int_{-\infty}^{\mu_{AB}} \frac{1}{\sqrt{2\pi\sigma_{AB}^2}} \exp\left(\frac{-x^2}{2\sigma_{AB}^2}\right) dx \\
&= \int_{-\infty}^{\mu_{AB}} \frac{1}{\sigma_{AB}} f\left(\frac{x}{\sigma_{AB}}\right) dx \\
&= \int_{-\infty}^{\frac{\mu_{AB}}{\sigma_{AB}}} f(x) dx \\
&= F\left(\frac{\mu_{AB}}{\sigma_{AB}}\right), \qquad (7)
\end{aligned}$$

where $F(z)$ and $f(z)$ are the standard normal cumulative distribution function (CDF) and probability density function (PDF), respectively. The probability of perceiving A with a higher psychological scale value than B, $\mathrm{P}(S_A > S_B)$ can be estimated by the empirical proportion of a sample of people, n. Then, from (7), given a known variance σ_{AB}^2, it is possible to obtain an estimator for the perceived difference in stimuli μ_{AB},

$$\hat{\mu_{AB}} = \sigma_{AB} F^{-1}\left(\frac{C_{A,B}}{n}\right), \qquad (8)$$

where $C_{A,B}$ is the number of people perceiving A with a higher psychological scale value than B. This estimate is known as Thurstone's Law of Comparative Judgement.[71] Since the parameter σ_{AB}^2 may be difficult to estimate, Thurstone developed five different cases of his law, which aim to simplify the law by making assumptions on this parameter. Case V, the simplest and most popular case, is explained further here.

In the Case V model, Thurstone implemented assumptions that each psychological scale value have equal variance and zero correlation.[71]

$$\sigma_A^2 = \sigma_B^2$$

$$\rho_{AB} = 0$$

The variance of the perceived difference in the Case V model is then [71]:

$$\sigma_{AB}^2 = 2\sigma_A^2 = 2\sigma_B^2 = 2\sigma^2,$$

and the estimator for the perceived difference in stimuli using the Case V model is [71]:

$$\hat{\mu}_{AB} = \sqrt{2}\sigma F^{-1}\left(\frac{C_{A,B}}{n}\right) \qquad (9)$$

Now Thurstone's Case V Law of Comparative Judgement can be applied to a hypothetical Paired Comparison test of video quality, in which a pair of video

clips, A and B, can be known as 'just noticably different' if 75% of the participants agree on the preference of image. That is, a test in which a result of 1 JND would be obtained by 75% agreement amongst participants. The proportion of participants choosing clip A over clip B, $\frac{C_{A,B}}{n}$ is set to 0.75 and the mean perceived difference in quality for this proportion is set to 1. Thus, the estimated variance for the perceived difference in quality in JNDs is

$$\begin{aligned}\hat{\sigma}^2_{AB} &= 2\hat{\sigma}^2 \\ &= \left(\frac{\mu_{AB}}{F^{-1}(C_{A,B}/n)}\right)^2 \\ &= \left(\frac{1}{F^{-1}(0.75)}\right)^2 \\ &= 2.198109\end{aligned}$$

Thus, the estimated variance of the quality levels of videos A and B is 1.099055. Therefore, videos of all quality levels will be perceived according to a normal distribution with variance approximately equal to 1. With the estimated variance of perceived difference in quality, it is now possible to estimate the proportion of participants required to obtain other JND values, by inverting (9). Recall that 0 JND is equivalent to 50% agreement amongst participants. The distributions and proportions of agreement of a few perceived quality differences are tabulated below.

JND	μ	$\hat{\sigma}^2_{AB}$	Proportion Agreement
0	0	2.198109	0.5
0.5	0.5	2.198109	0.6320339
1	1	2.198109	0.75
2	2	2.198109	0.9113283

In this way, Case V of Thurstone's Law of Comparative Judgement is adapted for application of the JND unit.
Thurstone's assumption of perceiving sensations as normally distributed is not quite accurate in practise. Systematic discrepancies are observed at larger stimulus differences, which means that larger stimulus differences are needed to drive participants to unanimity than would be anticipated. Empirical results and theoretical assumptions begin diverging away from each other at around 1.5 JND.[9]

6.3.4 Angular Distribution

With the Gaussian distribution comes an imprecision in the tails as the distribution approaches infinity. For 100% agreement amongst observers, the JND between a pair of clips according to the Normal distribution would be infinity, which is highly inaccurate. Keelan proposed replacing the Normal distribution with a distribution known as the Angular distribution. The Angular distribution closely resembles the Normal distribution but does not approach infinity due to truncation of the distribution at $\pm\sqrt{\frac{\pi^3}{8}}$.[51]

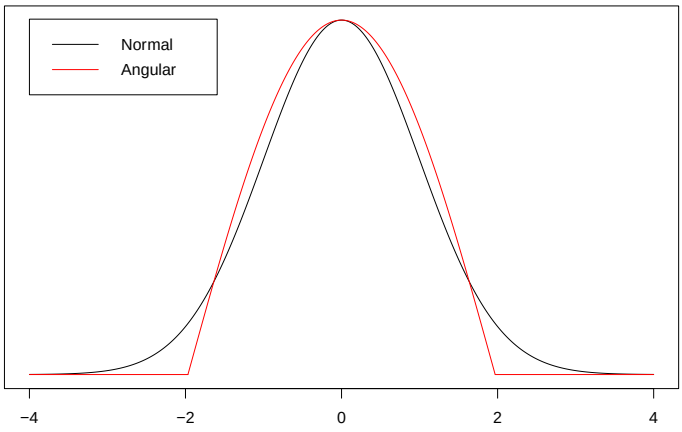


Figure 14: Normal & Angular Density Functions

As seen in Figure 14, the Angular distribution solves the issue of approaching infinity with larger probabilities, however, with an Angular assumption, smaller stimulus differences are predicted to be required to approach unanimity amongst participants. This prediction deviates from the Normal distribution prediction close to one standard deviation of the Normal and Angular distributions. Therefore, JND analysis must be limited to within one standard deviation.[51]
The cumulative distribution function for the Angular distribution is defined as [51]

$$F(z_a) = \begin{cases} 0, & z_a < -\sqrt{\frac{\pi^3}{8}} \\ \left(\sin\left(\frac{z_a}{\sqrt{2\pi}} + \frac{\pi}{4}\right)\right)^2, & |z_a| \leq \sqrt{\frac{\pi^3}{8}} \\ 1, & z_a > \sqrt{\frac{\pi^3}{8}} \end{cases}$$

The probability density function for the Angular distribution is [51]

$$f(z_a) = \begin{cases} \frac{1}{\sqrt{2\pi}} \sin\left(\sqrt{\frac{2}{\pi}} z_a + \frac{\pi}{2}\right), & |z_a| \leq \sqrt{\frac{\pi^3}{8}} \\ 0, & \text{otherwise} \end{cases}$$

The angular deviate for obtaining 75% agreement amongst participants, i.e. for obtaining 1 JND, can be calculated by inverting the cumulative distribution function for $F(z_a) = 0.75$. The angular deviate is 0.6562337 (to 7 significant figures), and thus the normalizing unit for JNDs based upon the Angular distribution is approximately 0.656. Hence, the proportion of agreement required for JND values other than 0 and 1 can be computed from

$$JND = \frac{F^{-1}(p)}{0.6562337},$$

where p is the proportion agreement amongst observers and $F^{-1}(\cdot)$ desrcibes the inverse cumulative distribution function of the Angular distribution. The proportion agreement required for different JND values are tabulated below.

JND	Proportion Agreement
0	0.5
0.5	0.6294095
1	0.75
1.5	0.8535534
2	0.9330126

The standard deviation for the distribution in this case is

$$\frac{1}{0.6562337} = 1.523847$$

Recall that JND analysis must be limited to within one standard deviation. Thus, the Angular distribution has approximately the same limitation of 1.5 JND as the Normal distribution.
It is unclear what the underlying distribution of the SQS values of the stimuli are, based on Angular distributions being applied to the difference in SQS values. In order to compute the probability distribution of the SQS values, deconvolution of the difference distribution must be performed.

6.3.5 Discussion

All of the models explored in this section hold advantages and disadvantages.

The Bradley-Terry model assumes a skewed distribution for the individual quality scale values, which seems highly unlikely if the experiment is free of bias or systematic error. This model should only be used if the underlying distributions are arbitrary or skewed.

The Binomial model is simple, yet the Normal approximation that needs to be applied imitates Thurstone's Law of Comparative Judgement. In making a Normal approximation of the Binomial distribution as opposed to initially assuming a Normal model for the perceived difference in quality seems to be counter-inuitive. Additionally, approximating will cause further errors to be added to the experiment.

The Thurstonian and Angular models are both inaccurate above approximately 1.5 JND, limiting analysis to within a region of (0, 1.5) JND. Within this region, the Angular and Thurstonian models are very similar.

The benefits of choosing the Angular model over the Gaussian model, are firstly in the ability to analytically calculate a deviate given the JND, and secondly, in estimating the JND for a pair of video clips where there is 100% agreement of preference. As this is outside of the region of accuracy, the latter is an unnecessary benefit. The former benefit eases analysis and computation. However, deconvolution is required to determine the distribution of the SQS values of individual video clips.

Thus, the Thurstonian model appears to be the best model for the analyses within this thesis. Upon completion of the Paired Comparison study, the proportion agreement for each pair of video clips should be analysed to determine whether or not the pair are 1 JND apart or not. This forms a two-tailed hypothesis test of means, in which the test statistic is computed from Equation

(8). Analyses should be limited to within approximately 1.5 JND.

It is important to note that the IQR as specified in ISO 20462 is based upon an Angular assumption. It is unclear what distribution the SQS value of each individual image is assumed to be.[9]

6.4 Practical Preparation for the Paired Comparison Test

6.4.1 Participants

ISO 20462 for psychophysical experimental methods for estimating image quality specifies certain criteria for the participants involved in the test. Firstly, participants should be free of any personal involvement of the experiment and the generation of or subject matter depicted by the test stimuli. The participants should have normal colour vision and good visual acuity at approximately the viewing distance employed in the experiment.[9] However, previous studies have found that participants with and without colour blindness do not produce significantly different results.[5] Thus, the results for participants with partial or complete colour blindness will be noted and removed if necessary.
ISO 20462 states that at least 10 observers (and preferrably 20) should contribute data to the analysis.[9] However, no discussion is provided on why this should be the requirement. As discussed in the previous Section, Paired Comparison tests with the purpose of evaluating relative image quality can be modelled with the use of Thurstone's Law of Comparative Judgement, and this model will be utilised in the analysis performed in this thesis. Power analysis can be performed on this model to obtain an estimate for the smallest sample size required to achieve reliable results. As the variance of the population is known (= 2.198), the Z test as described in Section 5.1 may be used to assess the results of the Paired Comparison test. In order to achieve a power of 0.8, 70 participants are required in order to be able to differentiate between a result of 1 JND and 1.5 or 0.5 JND. Due to time restrictions, 70 is deemed as too large of a sample size. If the effect size of the test is increased so that it is possible to differentiate between integers of JNDs, each non-integer JND will be rounded to the closest integer. In order to differentiate between 0 and 1 JND, a sample size of 18 achieves the same power. As this sample size is relatively low, it is possible to increase the power to 0.9, resulting in a minimum sample size of 24. Thus, a minimum of 24 participants should take part in the Paired Comparison test.
Potential observers within Axis Communications are emailed with invitations to participate. The technical imaging experience of the participants is recorded. Participants are split into two types, those who view images frequently with the aim of analysing said images, 'experienced observers', and those who have little to no technical interaction with images, 'naïve observers'. Previous studies have shown that naïve observers judge still image quality more accurately than experienced observers.[10] A roughly equal number of both types of observers take part.

An interesting side note relates to the images provided in the IQR. As mentioned in Section 4.3.5, the images are all calibrated to JND levels which have a precision of 2 decimal places. The documentation provided with the IQR suggests the use of the Angular distribution in configuring the JND levels. Noting

the similarities between the Angular distribution and the Normal distribution, it is possible to conduct power analysis by use of the t test, which assumes that the data are normally distributed. With a confidence level of $\alpha = 0.05$ and a desired power of 0.8, over 180,000 participants are required to ensure detection of differentiation between 1 and 1.01 JND, i.e. obtain a precision of 0.01 JND. For a confidence level of $\alpha = 0.10$, over 140,000 participants are required for the same purpose. Higher confidence levels and lower powers would naturally reduce the required sample size.If these requirements were ignored and fewer observers were tested, it is possible that the JND levels were obtained through interpolation. In this case, the precision to 2 decimal places is unnecessary and invalid. These results are not completely accurate as there are differences between the Angular distribution and Normal distribution, but they provide a rough idea of the samples sizes required for results with a precision to 2 decimal places.

6.4.2 Video Clips

A select set of pairs will be presented to observers, in order to remove any pairs that are so far apart in quality that they will result in 100% agreement amongst participants. Thus, the size of the test is reduced dramatically. The number of pairs to be tested is computed as follows:

Let there be m video clips, ordered in sharpness. It has been determined that each video clip should be compared only to n video clips on either side. Instinctively, a video clip far from either end of the sharpness range must be compared with $2n$ video clips. However, video clips near or at the end of the range have special cases. Starting from the least sharp video clip, this clip must be compared with n video clips. Moving on to the next video clip, this must be compared with $(n+1)$ video clips (the n clips sharper than the current clip, and the 1 clip that is more blurry than the current clip). However, one of these pairs has already been compared, therefore only n video clips must be compared at this stage. Successively, n video clips must be compared at each stage of sharpness, until the $(m-n+1)$th video clip is reached. From here, each video clip is compared with 1 fewer video clip than the previous video clip. The final and sharpest video clip has now been compared with all video clips within the n range, and does not need to be compared with any. The last n clips must be compared with $(n-1), (n-2), ..., 0$ video clips respectively. Mathematically, this translates to a total of

$$\begin{aligned} n(m-n) + \sum_{i=0}^{n-1} i &= n(m-n) + \frac{1}{2}(n-1)n \\ &= n(m - \frac{1}{2}n - \frac{1}{2}) \end{aligned}$$

pairs that need to be compared.

The pairs are presented in a randomised order. A randomisation script is written in Python, in which the list of videos that will be presented are randomised within each pair, and within the entirety of the list. The randomised list of videos is implemented in the software, which will be discussed in the following section. A null pair, consisting of two video clips that are close to either end of

the quality spectrum, is added into the middle of the test in order to determine whether the observer is paying attention or giving random responses. If an observer gives an incorrect response to the null pair, the data for that observer is later discarded. The number of pairs to be tested is now

$$n(m - \frac{1}{2}n - \frac{1}{2}) + 1 \tag{10}$$

It is possible to mirror the video clips horizontally, as discussed in [12]. When viewing images, it can be difficult to focus on and compare moving objects. Mirroring the clips makes it easier to notice differences close to the line of symmetry. Upon completion of the VQR, video clips which may not contain the same scene as a reference clip may be used against the VQR. This could be the case for testing different video compression algorithms, for which there is no raw image data. In cases like these, the clips to be tested would be compared against a VQR containing a scene with similar attributes to the test clips. Mirroring video clips in such instances where the test image does not match the VQR image, would not provide any benefit. The VQR should be created to be applied as generally as possible. Additionally, at this stage, the VQR calibration levels are based on those prescribed in ISO 20462 for the IQR, in which the images are presented non-symmetrically. Thus the pairs of video clips are presented in the software without any symmetry.

6.4.3 Laboratory Set Up & Software

A room (Visual Lab) has previously been prepared for the use of conducting psychophysical tests prescribed in ISO 20462. The walls have been painted a neutral grey, and curtains covering any windows closely match the colours of the walls. Pairs of video clips are presented on one computer display, in order to reduce any variation due to inconsistencies between displays. The wall behind the display is moderately uniformly illuminated by fluorescent lights emulating D65 to match the average luminance of the display. This is the only light source within the room, other than the display itself.[53] The display is an EIZO ColorEdge CG276 monitor, whose colours are adjusted to compensate for external luminance and other factors, with the use of a colour calibration device (X-Rite i1Display Pro). In addition to the previously mentioned equipment, the Visual Lab also contains a chin rest, in order to regulate the distance between the observer and the monitor. The monitor and chin rest are double-sided-taped to grey tables, with 31.7 inches between the eyes of the observers and the display surface.

Figure 15: Laboratory Set Up & Software

Software has previously been built by Axis Communications to use for the paired comparison test. The program starts by requesting the participant enter their name and click 'Start'. Some brief instructions appear before the participant can click a button to commence the test. During the test, pairs of video clips appear side by side on the screen. The video clips play in sync and in a loop. The clips have a space between them so as to make it easy for users to differentiate between clips. The space between clips is small so as to reduce the distance that the observers' eyes are required to move when viewing clips. Beneath the video clips are two buttons titled 'Vote Left' and 'Vote Right' respectively, indicating which button is to be clicked for the preferred video clip. Beneath these are a numerical indication of progress within the test.
The data recorded for each observer provides information about the observer's response and time taken for each pair.

6.4.4 Outline for the Experiment

The outline for the psychophysical experiment is as follows:

1. The observer arrives at the Visual Lab.

2. The observer completes vision tests to confirm normal vision. The observer is allowed to wear any normal visual aids (such as glasses, contact lenses) that they regularly use.

   A visual acuity test is performed by using a full size A4 print out of an eye test chart, taken from [73]. The observer stands 16 inches away from the test card and reads the line for the viewing distance required. A maximum of one mistake is allowed to be made. This corresponds to slightly lower relative acuity than 20/20 vision.[53] In this case, the video clips are calibrated for a viewing distance of 31.7 inches, therefore the line labelled '32 inches' should be read.

   Before performing the colour blindness test, observers are informed about the possible results of the test, should they be unaware of their colour vision abilities. The colour blindness test consists of viewing three images taken from a book containing a set of testing charts that implement the Ishihara test.[74] The three images chosen, images containing characters

'92, 83', '823', 'GREEN' respectively, test for red-green weakness, red-green blindness, and total colour blindness. Upon viewing the images, observers with normal colour vision are able to distinguish the contents of the image from the background, whereas observers with a complete or partial lack of colour vision are unable to see some or all of the contents.

Observers who fail the visual acuity test do not participate in the psychophysical experiment. Observers who pass the visual acuity test but fail the colour vision tests are welcome to participate in the psychophysical experiment, but the data obtained from such observers is noted for possible removal.

3. The observer reads a hard copy of instructions about the completion of the experiment. The instructions are presented as follows:

> Thank you very much for participating in this study!
>
> In this experiment, you will be evaluating overall quality of video clips through the comparison of pairs of clips, as part of a test known as the Paired Comparison Test. Please remember, there are no right or wrong answers, as the experiment wishes to determine the level of **perceived** image quality of the video clips.
> Before starting the test, place your chin in the chin rest and get as comfortable as possible.
> After entering your name in the field provided, the experiment will begin. During the test, please do not move your face away from the chin rest, as this will result in biased answers.
> A pair of video clips varying only in sharpness will appear on the screen. You must decide which video clip appears to be of **better overall quality. Select the option for the video that appears to be of better quality.**
> Throughout the test, you will be able to see your progress.
>
> Please let me know if you have any further questions!

4. The observer has the opportunity to ask the assessor any questions or clarifications about the instructions.

5. The observer sits down and places their chin in the chin rest, adjusting the chair and/or chin rest heights in order to get comfortable.

6. After the assessor leaves, the observer completes a trial experiment, consisting of 5 pairs or video clips with varying levels of sharpness. The video clips used in the trial differ from those used in the recorded experiment. Before the trial experiment is commenced, condensed instructions appear on the screen to reiterate the main points of the instructions given previously.

7. The observer has the opportunity to ask the assessor any questions and make any adjustments to the chair and/or chin rest.

8. The observer completes the real experiment, after viewing the condensed instructions on the screen.

9. Upon completion of the experiment, the assessor asks the observer a series of questions about their experience during the experiment:

   (a) What proportion of the time were you guessing?
   Less than $\frac{1}{2}$ */* $\frac{1}{2}$ */ More than* $\frac{1}{2}$

   (b) Out of those times, did you ever make a choice because you didn't know?

   (c) Out of those times, did you ever make a choice due to fatigue?

   (d) Any other reason?

   (e) How did you judge whether or not the video was of good quality – what objects did you look at?

   (f) Did you feel there were enough things for you to observe?

   (g) What did you think about the length of the video?
   Too short / Fine / Too long

   (h) Did anything about the physical elements make you uncomfortable?

   (i) Did anything about the software make you uncomfortable?

   (j) Did anything else make you uncomfortable?

   (k) Do you have any other comments?

   Question (a) is asked to determine the general perceived level of difficulty of the test. Question (b) - (d) attempt to pinpoint the reasoning behind any 'guessed' answers. In particular, question (c) can determine whether the test is too long. Question (d) can determine if there are any other variables that can affect a 'guessed' answer. Questions (f) and (g) provide some insight whether the scene is acceptable for this purpose. Question (e) can also have this function, but it can also be used to explore the way different types of people approach image quality. The answers to questions (h) - (j) can be used to alter the experimental process, if required.

6.5 Pilot Study

Before the study is implemented and participants are tested, a pilot study must be completed. Pilot studies are useful in order to practise the methodology and conditions for the larger study, and find any problems with the experiment before it begins. Additionally, it can loosely be used to determine whether the SQS values specified in ISO 20462 for the IQR will also apply to a VQR. After analysis on the pilot study data, it may appear that the levels of blur applied to static images may not be appropriate to apply to video clips. Should this be the case, the calibration should be adjusted to produce video clips with different blur levels. Finally, a pilot study can also estimate the length of a typical experiment, providing a basis for future adjustments of the experiment. The videos are calibrated to the sharpness levels prescribed in ISO 20462, following

the procedure in Section 6.2. Instead of performing a full paired comparison test, each video clip is chosen to be compared to 5 others on each side (i.e. any video clip is compared to a maximum of 10 in total). Since results that exhibit JNDs larger than 1.5 are disregarded, it is unnecessary to compare videos which are theoretically over 2 JNDs apart. However, as the JND assumptions for the pilot study are only an initial estimation, it is wise to examine pairs within a slightly larger range than 2 JNDs. Thus, each video clip is compared to 5 on either side. Using equation (10), the size of the test is now $5(31 - \frac{1}{2}5 - \frac{1}{2}) + 1 = 141$.
10 participants take part in the pilot study, of which data is not recorded for 1 due to a fault in the software. The error occurs when a participant enters a character not included in ASCII, for example, the character 'ä', in the field designated for the participant's name. The data obtained from the test is automatically saved to a Microsoft Excel '.csv' file, and it is not possible to save such characters with this software. Out of the remaining 9 participants, 4 are categorised as 'experienced' observers, and 5 as 'naïve' observers. All 9 participants have acceptable vision acuity. All but 1 participant have perfect colour vision, with the remainder exhibiting red-green colour weakness.

6.5.1 Relevant Results

The results for all 9 participants are collected and examined. All 9 participants pass the 'null-image' test, and the results for the partially colour blind participant are compared to the other participants, and exhibits no irregularities. Thus, all 9 results are included within the analysis.

One participant commented that they made a few errors in which they accidentally clicked a choice immediately, without viewing the videos. The participant was unable to change the decision as there was no option within the software to go back one or more steps. The errors (2 in total) were identified in the data as decision times of length 0, and removed.
Figure 16 shows a set of boxplots of the average times in seconds for each theoretical JND. For example, the average times for pairs (1, 2), (2, 3), ..., (30, 31) are grouped together. The distributions of the average times for each theoretical JND appear to be roughly normally distributed. The median time to make a decision decreases as the theoretical quality range between a pair of video clips increases.

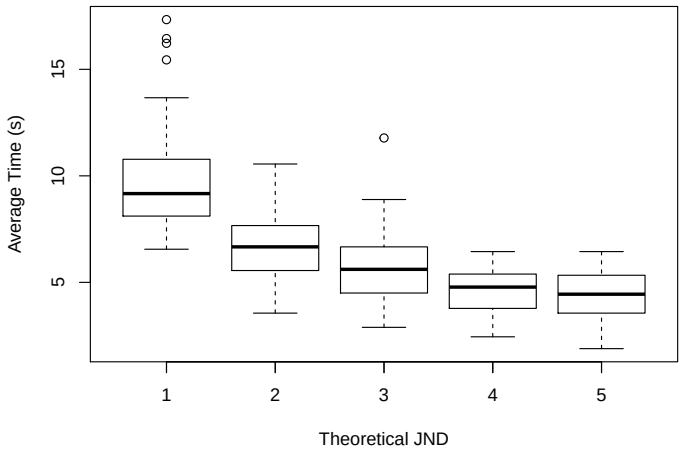


Figure 16: Average Time vs. Theoretical JND between pairs of images

The scatterplot in Figure 17 explains the average time taken to make a decision for the theoretical average SQS value of a pair. The blue line over the scatterplot is the line of best fit of the averaged data, showing a general increasing trend. As the video clips become sharper, the average time to make a decision generally increases. It should be noted that the highest average times to make decisions occur for the lowest and highest average SQS values, equivalent to the pairs (1, 2) and (30, 31), respectively. Comparatively, taking an average SQS value of 15.5, the pairs that contribute to the average time data are (13, 18), (14, 17), and (15, 16). Each of the three pairs have a different JND between them. As seen in Figure 16, observers make faster decisions on pairs that are further apart in quality. Thus, for cases such as those for the average SQS of 15.5, the high and low decision times roughly balance out. The average SQS end-points are only based only upon pairs that are theoretically 1 JND apart in quality, explaining the seemingly high average decision times.

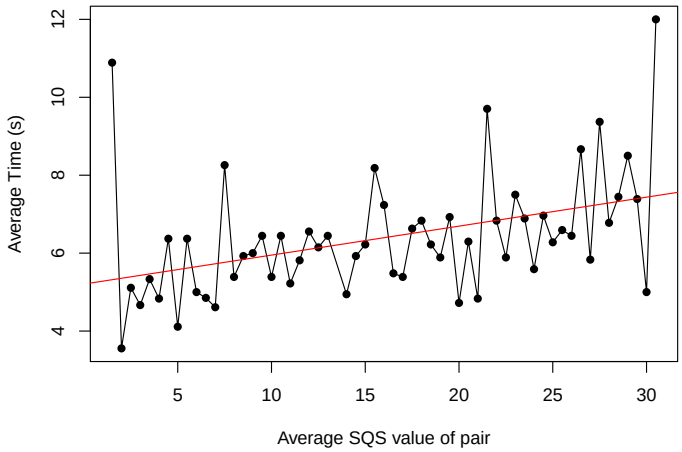


Figure 17: Average Time vs. Theoretical Average SQS Value of pairs of images

The average total time of the Paired Comparison test alone was approximately

15 minutes.
One participant commented that at the beginning they were not aware that the videos would play in a loop, so would rush to make a decision before the video ended. Therefore, the results about the decision times contain some random errors (due to the randomisation of the pairs).

Recall that the minimum sample size required to participate in the Paired Comparison test is 24. In achieving a power of 80% with a sample size of 9, the smallest JND that can be noticed is approximately 1.5. The JND between each pair is computed by substituting the known variance 2.198 and the proportion agreement between the pair into Equation (8). Using the Z-test, confidence intervals of the empirical JNDs are computed. Figure 18 contains a plot showing the proportion agreement for each pair, and denotes whether or not the confidence interval of the JND between each pair covers the theoretical JND between the pair. The x-axis non-linearly increases by average theoretical SQS value. It is also secondarily ordered by increasing theoretical JND between the pair. Thus, the x-axis is ordered, in theoretical SQS pairs: (1, 2), (1, 3), (2, 3), (1, 4), ..., (29, 30), (28, 31), (29, 31), (30, 31). A result of 100% agreement results in a mean JND of infinity according to the model. Thus, confidence intervals for such proportions are centered around infinity and cannot contain the theoretical JND between the pair.

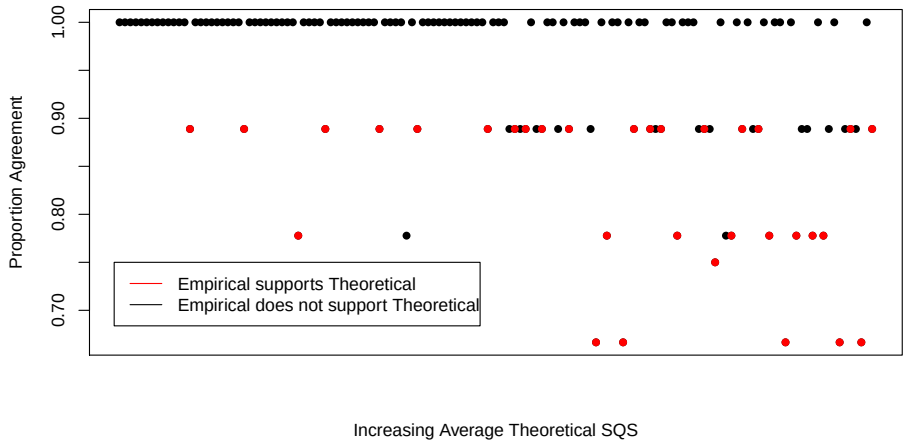


Figure 18: Proportion Agreement between Pairs vs. Average Theoretical SQS of Pairs ordered (1, 2), (1, 3), (2, 3), (1, 4), ..., (29, 30), (28, 31), (29, 31), (30, 31)

It can be seen that for blurrier videos, the majority of the pairs obtained 100% agreement. As the videos become sharper, the pairs with complete agreement grow sparse. Approximately 23% of the confidence intervals contain the theorised value.
Recall that JND analysis should be restricted to within 1.5 JND. According to Equation (9), 1.5 JND is equivalent to approximately 84% agreement. If this analysis is to be restricted to below this threshold, approximately 11% of the results can be used.
The following is qualitative analysis dependent on the questionnaire executed

after the tests. As questions can be very subjective and open to errors, the results are not as reliable as the quantitative results above.
The majority of participants felt that the length and content of the video were acceptable. One participant, a naïve observer, would have preferred the scene to contain more of the person, in relation to both the duration and size of the face. However, the observer was content with the other objects within the scene despite the lack of person.
Reasons for 'guessed' responses, other than the inability to make a decision, were boredom and the aforementioned accidental immediate click of button. Only one participant made choices out of boredom. Another participant commented that the choice made as a 'guessed' response was dependent on the location of the mouse at that time. For example, if the previous choice had been the left button, if the participant was not able to make a choice between the current pair, they would select the left video again.
Of the 9 participants, 4 concentrated on a mixture of static and moving objects within the scene, whilst the remaining 5 concentrated on static objects. 75% of the experienced observers and 60% of the naïve observers focused only on the static objects within the scene.

6.6 Large Study

Following the Pilot study, small parts of the methodology need to be modified. Firstly, it seems that the sharpness levels prescribed within ISO 20462 may not be accurate, particularly for the blurrier video clips. This can be seen by the shorter decision times for blurrier videos and a large number of complete participant agreement for blurrier videos. It is possible that some of the clips that are theoretically 1 JND apart are empirically further than 1 JND apart. Thus, a further 30 video clips are created with sharpness levels halfway between each existing sharpness level. An additional calibration level with a theoretical SQS of 0.5 is also applied to the video clip, resulting in a total of 62 video clips spanning roughly the original range of sharpness. All of the k values of Equation 3 that are used in the large study are plotted against the theoretical SQS values according to ISO 20462 in Figure 19.

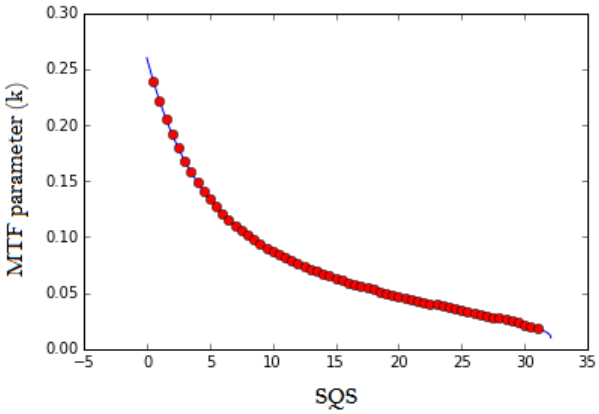


Figure 19: k values

Within the pilot study, each video clip was compared with 5 others on either side, spanning a range of ± 5 theoretical JNDs. Only 11% of the pilot results are within the acceptable analysis range of 1.5 JNDs, thus testing over a range of 5 theoretical JNDs seems unecessary. This is particularly true for the blurrier end of the spectrum, where a majority of the results were of 100% agreement. If the spread of results larger than 1.5 JND was uniform across all the results, the JND testing range should be restricted to approximately $0.11 \cdot 5 = 0.55$ JND. Due to the non-uniformity of the results, and the small sample size of the pilot test, other methods of restricting the range should be employed. According to Figure 16, average decision times for theoretical JNDs larger than or equal to 4 drop below 5 seconds. A good limitation to set would be comparisons within 3 theoretical JNDs. With the new calibrated levels, this would result in 352 comparisons that would have to be made. As one participant already made 'guessed' decisions out of boredom, this seems too large. The video clips are now chosen to be compared with 3 on either side, i.e. each video clip is compared with images within a theoretical range of ±1.5 JND. This results in 180 pairs and one null pair that will be compared. Participants of the large study are booked in for one hour sessions in the Visual Lab. This time includes testing the participants' vision, reading instructions, running the trial assessment, the test itself, and completion of the questionnaire following the test. Should the hour come close to completion, the participant will be asked to stop and invited to return another time to complete the test.
The hard copy of instructions given to participants is slightly edited, to include the statement, "Please note that the video clips will play in a loop so you don't need to rush." Participants are additionally instructed to enter only ASCII characters in the software.
The software of the Paired Comparison test is edited to include a 'back' button, to enable users to go back and edit a decision they have already made.
Finally, the questionnaire is edited to remove questions about the length and content of the video (questions (f) and (g)).

As noted in Section 6.4.1, a minimum of 24 observers are required to participate within the Paired Comparison test. 41 persons within Axis Communications take part. 20 are classified as 'experienced' observers, and the remaining 21 as 'naïve' observers. 1 'experienced' observer and 1 'naïve' observer exhibit red-green blindness.

7 Results

2 of the 'naïve' observers do not complete the Paired Comparison test within the alloted time. They are invited to return at a later stage to complete the test, but due to time constraints, the test is not completed for these participants. The incomplete results are discarded and not included within these results.
Of the remaining 39 observers, the percentage correctness for each participant is computed. Examining box plots of the percentage correctness, the 2 partially colour blind observers do not appear to be outliers, and thus are included within the study.

7.1 Z test

Two sets of Z tests are performed on each pair to determine the JND between the videos within each pair of video clips. The first hypothesis tests whether or not the JND between the pair is equal to or larger than 0. If it is found that the JND between the pair is larger than 0, the second hypothesis tests whether or not the JND between the pair is equal to or larger than 1. Both tests are conducted at the 2.5% level, in order to ensure that the overall significance level of the test of each pair is 5%, as per Bonferroni's Correction. The test statistic utilised in the Z tests are computed from Equation (8), where $C_{A,B}/n$ is the proportion of participants that selected the sharper video over the blurrier video. Pairs that fail both tests are determined to be more than 1 JND apart, pairs that fail the first test but not the second test are 1 JND apart, and pairs that fail neither test are 0 JND apart.
Figure 20 shows the JND between the video clips, which are labelled with their theoretical SQS values. The majority of the blurrier videos are approximately more than 1 JND apart. There are very few pairs that are more than 1 JND apart on the sharper end of the spectrum, with most clips being approximately 1 JND apart. It is around the theoretical SQS range of [8.5, 12] that the results appear to change and deviate away from the theorised JNDs.

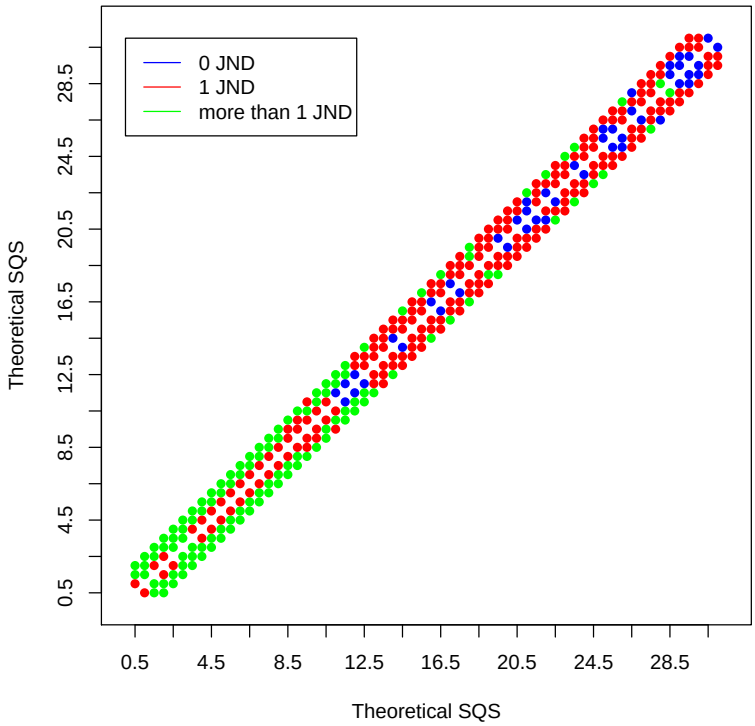


Figure 20: JND between pairs

In addition to testing the JND between each pair, confidence intervals of the JND between the clips are computed and plotted, see Figure 21, in order of ascending average theoretical SQS value. As in the pilot study, video pairs with

lower average sharpness appear to be further apart in subjective quality than previously hypothesised.

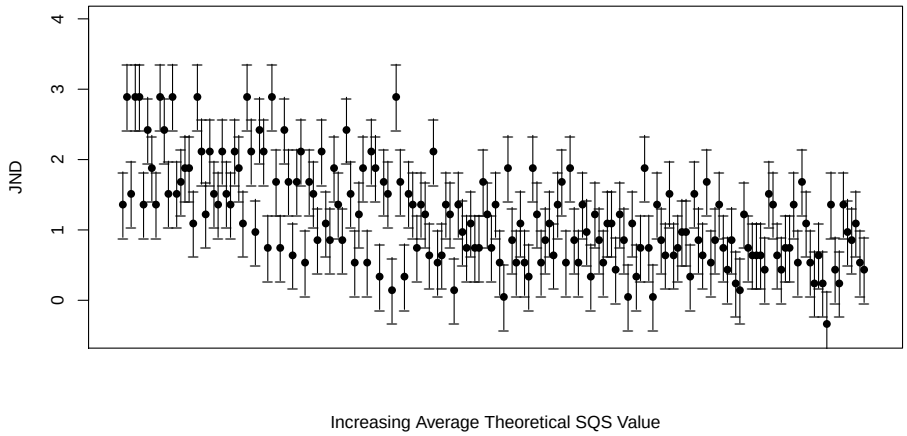


Figure 21: JND between pairs

There are three possible methods to creating a ruler in which each level is 1 JND apart:

1. Find all possible permutations of the video clips, in which each video clip is 1 JND apart according to the Z tests. Select the permutation for which the sum of the p-values for the Z test of $\mu = 1$ of the difference between the videos is highest.

2. Find all possible permutations of the video clips, in which each video clip is 1 JND apart according to the Z tests. Take the average of all permuations.

3. Find all possible permutations of the video clips, in which each video clip is 1 JND apart according to the Z tests. Take the weighted average of all permuations, depending on the overall sum of p-values of the distances between each video clip.

A script is written to compute all possible permutations of the video clips, according to a few conditions as follows:

1. The distance between one step of the permutation and the next step is 1 JND.

2. The distance between one step of the permutation and the step after the next step is more than 1 JND.

3. If the distance between one step of the permutation and the step after the next step is not known, this is not included.

Due to the second and third conditions, any permutations that are computed sit only within the blurrier end of the spectrum. For many pairs $(i, i+1)$ which are 1 JND apart, the pairs $(i, i+2)$ and $(i+1, i+2)$ are also 1 JND apart. Recall that JND analysis becomes inaccurate as the proportion agreement of

participants increases. Thus, the 2^{nd} condition is edited, so that if there exists a clip(s) that has been found to be more than 1 JND away, is it used as the 2^{nd} step away. If there is no such clip, the clip with the furthest theoretical JND that has been tested is used as the 2^{nd} step away. Additionally, the third condition is removed completely.
According to these conditions, there are over 3000 possible permutations of varying lengths, covering the theoretical SQS range [3.5,31]. Below the theoretical SQS of 3.5, there are few video clips that have been found to be 1 or less JND apart.
The sum of the p-values of each of the videos within the permutations are computed. There are many permutations for which the sums are close together in value. Thus, the first method of creating a ruler through permutations is eliminated. The unweighted and weighted averages of the permutations are computed. The permutations are averaged depending on their length, resulting in 2 sets of 6 permutations of lengths 29, 30, 31, 32, 33, 34. Each permutation is checked and adjusted to ensure that the distance between each step of the permutation is approximately 1 JND, satisfying the conditions of the permutations set previously. The permutations of length 31 are selected to enable easy comparison between the ISO 20462 standard and the new results obtained within this thesis. The weighted permutation is selected due to its higher overall p-value. The selected pairs, their renamed SQS values, and their 1 JND p-values are shown in Table 1.

Old SQS	Old SQS	New SQS	New SQS	P-Value
3.5	4	1	2	0.699
4	4.5	2	3	0.352
4.5	5	3	4	0.129
5	5.5	4	5	0.129
5.5	6	5	6	0.699
6	6.5	6	7	0.900
6.5	7	7	8	0.283
7	7.5	8	9	0.283
7.5	8	9	10	0.128
8	**8.5**	**10**	**11**	**0.0504**
8.5	9.5	11	12	0.699
9.5	11	12	13	0.352
11	**12**	**13**	**14**	**0.0504**
12	13	14	15	0.129
13	14.5	15	16	0.129
14.5	15	16	17	0.283
15	16.5	17	18	0.129
16.5	**17**	**18**	**19**	**0.0504**
17	18.5	19	20	0.542
18.5	19.5	20	21	0.542
19.5	20.5	21	22	0.542
20.5	22	22	23	0.699
22	23	23	24	0.542
23	24	24	25	0.900
24	**25**	**25**	**26**	**0.0504**
25	**26**	**26**	**27**	**0.0504**
26	27	27	28	0.128
27	28	28	29	0.283
28	29.5	29	30	0.128
29.5	30.5	30	31	0.900

Table 1: Successful Video Pairs

The bolded entries in Table 1 highlight the pairs for which the p-value is relatively small. The results for videos within these pairs may not be validated during implementation of the VQR.
The confidence intervals of the new SQS values according to the confidence intervals previously found are computed and plotted in Figure 22. The assigned SQS values are along the red line.

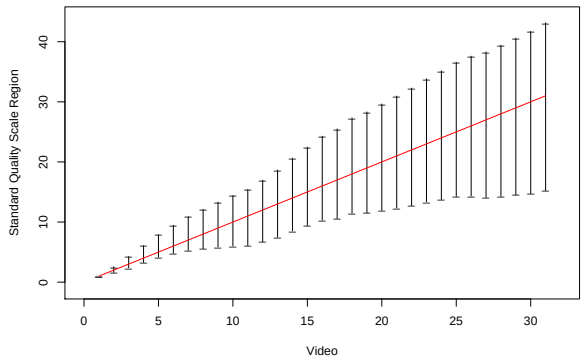


Figure 22: Standard Quality Scale

7.2 Weber-Fechner Law

In order to confirm the validity of the Weber-Fechner Law, the natural logarithm of the k parameters created in Section 6.6 are computed. The new SQS values are plotted against the corresponding $\log(k)$, shown in Figure 23. The relationship obtained through linear regression is additionally plotted in the graph, to show any possible deviations away from the Weber-Fechner Law. The majority of the SQS values appear to follow the Weber-Fechner Law. However, at the tails, the empirical curve begins to deviate away from the straight line, albeit by a small amount.

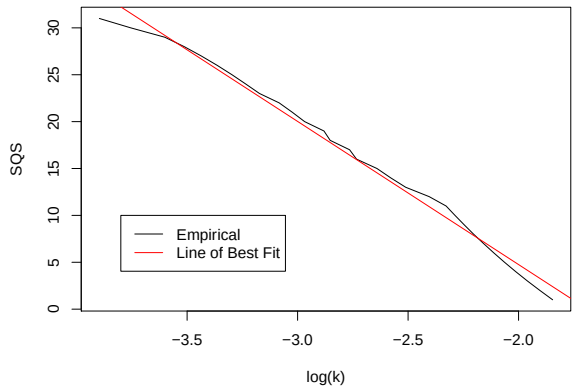


Figure 23: Weber-Fechner Law

7.3 Comparison to the ISO 20462 Standard

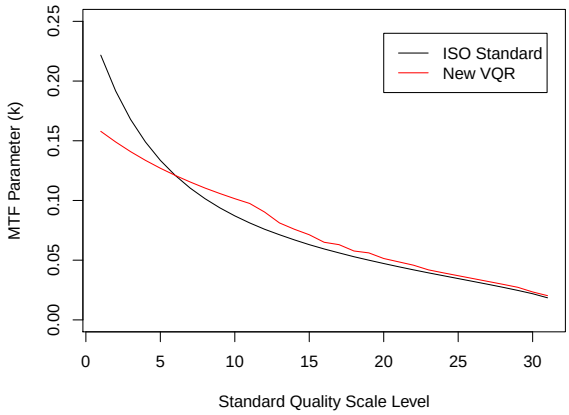


Figure 24: Comparison of the parameter

In order to compare the calibration of the old (ISO 20462 standard) and new quality rulers, the k parameter is plotted for both sets of SQS values. It can be seen that sharper videos follow a similar shape to the ISO 20462 standard ruler. However, the curve of the new ruler begins to deviate away from the old ruler at a SQS value of around 13. The gradient of the curve of the new ruler is more constant than that of the old ruler. Additionally, the new ruler does not span the same range of quality as the old ruler, as video clips worse than the now lowest quality video had insufficient results for extending the ruler further.

7.4 Difference between Experienced and Naïve Observers

In order to compare the results of experienced and naïve observers, a Z test for the difference between two means is conducted. The proportion agreement for each pair for both groups is computed, from which the JND between each pair for both groups is computed. The JNDs for each pair for both groups are compared at the 95% level, excluding any pairs for which one or both of the groups completely agreed, i.e. pairs that produced a theoretical JND of infinity. Out of the remaining 166 pairs, 33 cause the two groups to obtain different results. For 4 of the 33 pairs, the naïve observers produce larger JNDs than the experienced observers. For the remaining 29 pairs, the experienced observers produce larger JNDs than the naïve observers.

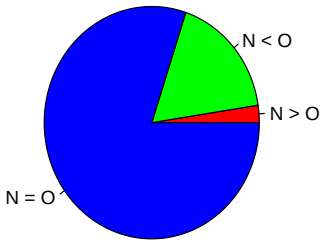


Figure 25: Proportion of pairs obtaining different responses between groups

For the majority of the pairs, the two groups judge video quality to roughly the same level. The majority of the remaining pairs are judged with more accuracy by the experienced observers. There does not appear to be any obvious pattern for which pairs are judged more accurately by one group than the other.

7.5 Qualitative Results

The responses to the question regarding the objects concentrated on within the video clip for both the pilot and larger study are collated. Roughly 57% of all participants said that they concentrated on a mix of static and moving objects within the clips. 35% focused solely on static objects, and the remaining 8% solely on moving objects.
Upon asking the final question of the questionnaire, 6 participants, approximately 15%, commented that they felt that they consistently preferred the colours of the video clips on the right side. The colour and illumination of the display were tested by use of a colorimeter (Konica Minolta CS-100). The results showed that there was a difference of 3 Δ_E in colour between the centre and right corner of the display, which confirmed the perceived the difference in 'colourfulness'. This is a clearly noticeable difference for uniform coloured patches, but less so with a detailed background. As the videos within each pair are randomised between the left and right side, the results may have only been compromised by the possible increase in variance. The results of all participants are grouped, and the number of 'left' responses are compared to the number of 'right' responses. 48% of the responses were 'right', and 52% were 'left'.
Approximately 44% of the participants commented that they felt the test was too long, and felt tired.

8 Validation

8.1 Method

The results found in the Section 7 must be validated to confirm or disprove the findings. Validation of the results may be carried out by implementing the VQR with the 31 calibrated levels. In this case, each of the 31 video clips of the VQR must be viewed by observers and assigned an SQS value by comparison

against the VQR itself. Recall the use of the Thurstonian model for analyses within this thesis. Thus, the results obtained in Section 7 are validated if the perceived SQS values for each video clip are normally distributed about the true SQS value, with variance 1.099. More concisely, the methodology in this section aims to confirm that:

- The responses, i.e. the perceived SQS value, for each video clip are normally distributed.
- The mean of the responses for each video clip are the true SQS values of each video clip.
- The variance of the responses of each video clip are 1.099.

If the aims of this method are not fulfilled, the results found in Section 7 may be inaccurate.

8.1.1 Video Clips

The scene used in the Paired Comparison test is used in the VQR. The 31 sharpness levels found in Section 7.1 are used for the VQR.

8.1.2 Participants

Based on a Normal model assumption, power analysis is performed in order to determine the minimum number of participants required to achieve reliable results. The mean and variance of the responses will be analysed by use of the Z test and Chi-squared test for variances. In order to achieve a power of 0.8, a minimum of 12 participants are required for both tests. The former test will differentiate between 2 integer SQS values, and the latter test will differentiate between a standard deviation of 1.05 and 2. The Shapiro-Wilk test for normality requires a minimum of 30 participants, which is not possible in this case due to time constraints. However, as the Shapiro-Wilk (and other analytical normality tests) are only additions to graphical methods, it is not required to perform this test. 15 observers participate in the VQR experiment, of which 7 are experienced and 8 are naïve.

8.1.3 Laboratory Step Up & Software

The Visual Lab used for the Paired Comparison test is used for the VQR test.

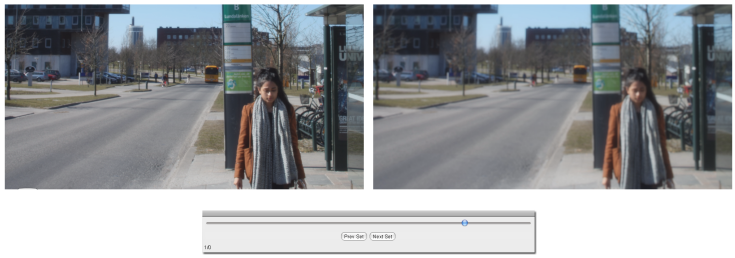


Figure 26: Video Quality Ruler Software

Software has previously been created by Axis Communications to implement the VQR method. After the observer enters their name in the designated field, a brief set of instructions appear on the screen. The user then clicks a button to begin the test. During the test, the test clips appear on the right side, and the ruler clips appear on the left side. The software is set up as shown in Figure 26. The ruler clip is always initially presented as the lowest quality video. It is not possible to randomise the ruler and test clip sides, suggesting that if the difference in colour across the display becomes noticeable by all participants, the means of the responses will be shifted to the right of the true value. For example, the mean perception of the video with SQS value 10 may be an SQS value of 11 or more. However, if it goes unnoticed by some and noticed by others, it may change the shape of the distribution of the responses, possibly skewing the responses.
The video clips play in sync and on a loop. The clips have the same amount of space between them as in the Paired Comparison test. Beneath the video clips is a slider which controls the SQS level of the ruler. The slider can be adjusted using the mouse or the arrow keys. Beneath the slider are a pair of buttons which allow the user to go back to the previous test video, or progress to the next test video. Beneath the buttons are a numerical indication of the progress within the test.
The data recorded for each observer includes: the SQS value of the ruler video clip chosen by the user, the SQS value of the test video clip, and the time taken to make a decision for each test video clip.

8.1.4 Outline for the Experiment

The outline for the psychophysical experiment is as follows:

1. The observer arrives at the Visual Lab.

2. The observer completes the same vision tests as in the Paired Comparison test to confirm normal vision. The observer is allowed to wear any normal visual aids (such as glasses, contact lenses) that they regularly use.

   Observers who fail the visual acuity test do not participate in the psychophysical experiment. Observers who pass the visual acuity test but fail the colour vision tests are welcome to participate in the psychophysical experiment, but the data obtained from such observers is noted for possible removal.

3. The observer reads a hard copy of instructions about the completion of the experiment. The instructions are presented as follows:

Thank you very much for participating in this study!

In this experiment, you will be evaluating overall quality of video clips. Please remember, there are no right or wrong answers, as the experiment wishes to determine the level of **perceived** image quality of the video clips.

After entering your name in the field provided, the experiment will begin. During the test, please do not move your face away from the chin rest, as this will result in biased answers.

A pair of video clips varying only in sharpness will appear on the screen. Below the video clips will be a slider controlling the video clip on the left hand side. Adjust the slider to change the sharpness of the video clip, until you believe the left hand side matches the right hand side in **overall quality**. The slider can be adjusted by the mouse or by using the left and right arrow keys. When you are happy with your decision, select 'Next'.

Please note that the video clips will play in a loop so you don't need to rush. If you make an incorrect decision, you are welcome to go back and change your choice.

Please let me know if you have any further questions!

4. The observer has the opportunity to ask the assessor any questions or clarifications about the instructions.

5. The observer sits down and places their chin in the chin rest, adjusting the chair and/or chin rest heights in order to get comfortable.

6. After the assessor leaves, the observer completes a trial experiment, consisting of 4 test video clips of varying sharpness being compared to a 'ruler' made up of the same 4 video clips. The video clips used in the trial differ from those used in the recorded experiment. Before the trial experiment is commenced, condensed instructions appear on the screen to reiterate the main points of the instructions given previously.

7. The assessor returns and the observer has the opportunity to ask the assessor any questions and make any adjustments to the chair and/or chin rest.

8. The assessor leaves again and the observer completes the real experiment, after viewing the condensed instructions on the screen.

9. Upon completion of the experiment, the assessor asks the observer a series of questions about their experience during the experiment:

   (a) How did you judge whether or not the video was of good quality – what objects did you look at?

   (b) Did anything make you uncomfortable?

   (c) Other than sharpness, did you notice anything different between the videos?

   (d) Do you have any other comments?

Question (c) is asked in relation to the colour non-uniformity of the display.

8.2 Results

Upon examining boxplots of the responses for each SQS level, one naïve participant frequently produced large outliers of 5 - 6 difference in SQS in their results. The results of this participant were removed for analysis, resulting in 7 naïve participants and 7 experienced participants. The spread of the remaining results are plotted in Figure 27. The true SQS values are along the red line. It appears that the majority of the results are centred around the true SQS values with differing variances. As the SQS values tend to the endpoints [0, 31], the results become skewed due to the inability to select videos outside of the given range.

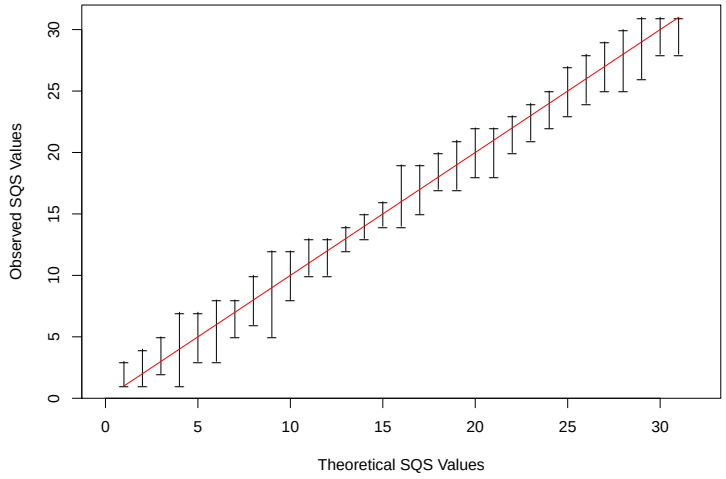


Figure 27: Spread of Results

8.2.1 Chi-Squared Goodness of Fit

The results of the 14 participants are tested by the Chi-Squared Goodness of Fit test, to assess the validity of the three conditions listed in Section 8.1. The expected distribution of the results of each video clip is a Normal distribution with variance 1.099 and mean as the true SQS value of the video clip.
For each of the video clips, the responses are separated into bins that cover the range of SQS $\pm$ 4. The range is chosen as $\pm$ 4 for two reasons, the first being that over all responses, 4 is the largest deviation away from the true SQS value. The second reason depends on the expected number of participants out of 14 who perceive a video clip as 5 units away from the true SQS value, 0.000123. This is a small enough number to not greatly affect the results.

The 9 bins for SQS values [5, 27] are:

$$[\text{SQS} - 4 - 0.5, \text{SQS} - 4 + 0.5)$$
$$[\text{SQS} - 3 - 0.5, \text{SQS} - 3 + 0.5)$$
$$[\text{SQS} - 2 - 0.5, \text{SQS} - 2 + 0.5)$$
$$[\text{SQS} - 1 - 0.5, \text{SQS} - 1 + 0.5)$$
$$[\text{SQS} - 0.5, \text{SQS} + 0.5)$$
$$[\text{SQS} + 1 - 0.5, \text{SQS} + 1 + 0.5)$$
$$[\text{SQS} + 2 - 0.5, \text{SQS} + 2 + 0.5)$$
$$[\text{SQS} + 3 - 0.5, \text{SQS} + 3 + 0.5)$$
$$[\text{SQS} + 4 - 0.5, \text{SQS} + 4 + 0.5)$$

The remaining SQS values of 1, 2, 3, 4, 28, 29, 30, 31 are expected to have truncated Normal distributions. The distributions of the blurrier videos are truncated on the left side at SQS = 1 - 0.5. The distributions of the sharper videos are truncated on the right side at SQS = 31 + 0.5. In this way, the number of bins for each of the truncated distributions decreases as the mean value of the distribution becomes more extreme.
The responses for each of the video clips are assigned a bin, thus obtaining the observed frequency for each bin. The Chi-Squared Goodness of Fit statistic is computed from the observed and expected frequencies. The p-value for each video clip is shown in the table below. P-values smaller than 0.05 are bolded to demonstrate the video clips that do not validate the results found in Section 7.

SQS Value	Chi-Squared P-Value
1	0.945
2	0.997
3	0.899
4	**0.014**
5	0.732
6	0.211
7	0.983
8	0.733
9	**0.00**
10	1.00
11	0.802
12	0.983
13	0.448
14	0.769
15	0.818
16	0.201
17	0.314
18	0.499
19	0.691
20	0.940
21	**0.028**
22	0.563
23	0.223
24	0.735
25	0.097
26	0.727
27	0.478
28	0.418
29	0.187
30	0.939
31	0.141

Videos 4, 9, and 21 break some or all of the conditions required to validate the results. Video 9 has a particularly small p-value. Histograms of the responses for these video clips are shown in Figure 28. The responses for videos 4 and 9 are roughly symmetrical, however, not necessarily normally distributed. The responses for video 21 are negatively skewed. The responses for videos 4 and 9 have a large spread around the mean.

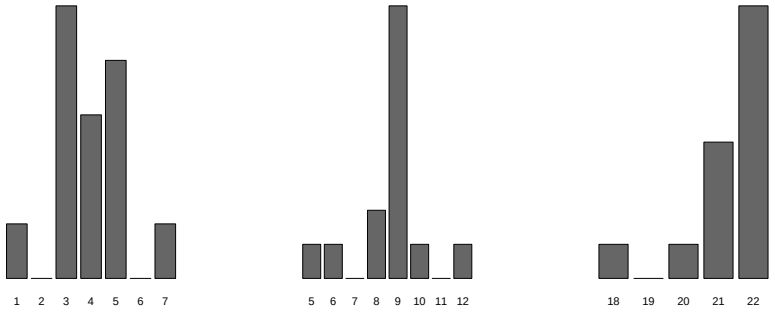


Figure 28: Histograms for videos 4, 9, and 21

8.2.2 t Test & χ^2 Test of Variance

The benefit of the Chi-Squared Goodness of Fit test is that it is fast and easy to assess the fit of a complete model to the data. However, it is not possible to determine which part of the model assumptions are inaccurate, be it the shape of the model, the mean, or the variance. The t test and Chi-Squared test of variance assume that the data are normally distributed, and assess the equality of the mean and variance to values, respectively.

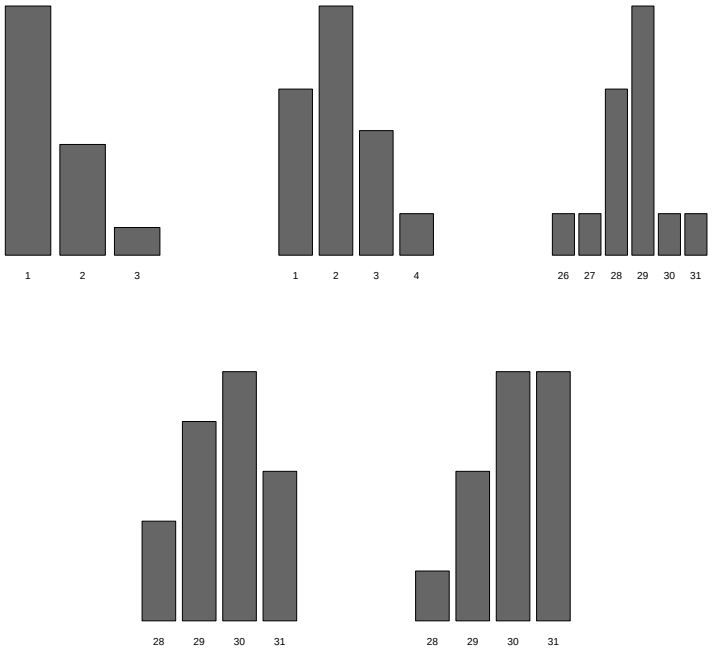


Figure 29: Histograms for videos 1, 2, 29, 30, and 31

The Chi-Squared Goodness of Fit test found that there is insufficient evidence to assume that the majority of the video clips are not normally distributed with mean as the true SQS values of the video clips and variance 1.099. The arithmetic means of the responses for each of the video clips (including videos 4, 9 and 21) are tested at the 95% level for equality to the true SQS values. The standard sample variances of the responses for each video clip are tested at the 95% level for equality to 1.099. According to the histograms in Figure 29, the responses for videos 1, 2, 29, 30, and 31, may have truncated normal distributions. The results of the tests for these videos should largely be ignored, as the tests used in this section do not account for truncated Normal distributions. Additionally, the results of the tests for videos 4, 9 and 21 should be ignored. As before, the videos for which the p-values are less than 0.05 are in bold.

SQS Value	Mean	t P-Value	Variance	Chi-Squared P-Value
1	**1.43**	**0.0275**	**0.418**	**0.0470**
2	2.07	0.775	0.841	0.603
3	3.21	0.426	0.951	0.819
4	3.93	0.856	2.07	0.0536
5	5.21	0.459	1.10	0.886
6	6.21	0.533	1.57	0.277
7	6.86	0.547	0.747	0.430
8	7.93	0.807	1.15	0.808
9	8.64	0.431	**2.71**	**0.00474**
10	9.93	0.807	1.15	0.808
11	10.9	0.752	0.687	0.329
12	11.9	0.547	0.747	0.430
13	13.0	1.000	**0.308**	**0.0111**
14	14.3	0.165	0.527	0.126
15	14.9	0.435	0.440	0.0589
16	16.6	0.104	1.49	0.340
17	16.9	0.699	1.82	0.124
18	18.6	0.0715	1.19	0.742
19	18.6	0.212	1.49	0.340
20	20	1	1.69	0.190
21	21.3	0.365	1.30	0.573
22	21.6	0.189	1.34	0.514
23	22.8	0.459	1.10	0.886
24	23.9	0.500	0.593	0.198
25	24.8	0.583	2.03	0.0626
26	25.8	0.533	1.57	0.277
27	26.7	0.435	1.76	0.154
28	27.8	0.551	1.72	0.174
29	28.6	0.212	1.49	0.340
30	29.6	0.208	1.02	0.949
31	**30.0**	**0.00185**	0.923	0.765

The means of the responses of all of the videos appear to be the true SQS values of the videos. A notable video is 18, which has a small (albeit insignificant within this thesis) p-value. It should also be noted that the majority (approximately 70%) of the mean values for each video are insignificantly lower that the true

SQS value, seemingly negating the possibility of the colour non-uniformity of the display having an effect on responses.
It can be seen that the variance for video 13 appears to be significantly less than 1.099. Other notable videos are video 15 and 25. A relatively low variance for video 14 indicates that the calibration levels determined for videos 13, 14 and 15 may possibly be invalid.

8.2.3 Maximum Likelihood Estimation Method

Instead of computing the arithmetic mean and sample variance of the responses, the maximum likelihood estimates of the mean and variance of the responses of each video clip are computed. The underlying distributions of the video clip responses are assumed to be Normal, though this may not be the case for video clips 4, 9, and 21, as found in Section 8.2.1. Continuing the use of bins, the likelihood of a bin $(i-0.5, i+0.5)$ with m observations is

$$\prod_{j=1}^{m}(F(i+0.5)-F(i-0.5))$$

As each of the m observations is independent and identically distributed, the log likelihood is then

$$m\log(F(i+0.5)-F(i-0.5))$$

Then, the log likelihood for each video clip is the sum of the log likelihoods of the bins of the video clip. The parameters μ and σ^2 can be estimated by maximising this function through optimisation using the Nelder-Mead method. The optimisation has starting estimates as the arithmetic mean and the sample variance, and is limited to the positive numbers. For the majority of the videos, the Normal distribution with 9 bins (as in Section 8.2.1) is assumed. The remaining videos are assumed to have truncated Normal distributions.
The ML estimators for the mean and variance of each of the videos is computed. Further, the responses of each video are bootstrapped paramterically according to the ML estimators 1000 times. $1000 \cdot 31 \cdot 2$ ML estimators are computed from the $1000 \cdot 31 \cdot 14$ bootstrapped responses, using the same ML method. The 2.5% and 97.5% quantiles of the estimators are found, obtaining the confidence intervals for each estimator. The ML estimators for the means and their confidence intervals are plotted in Figure 30. The same for the ML variances are plotted in Figure 31. Both figures also include the confidence intervals for the means and variances found in Section 8.2.2 for easy comparison.

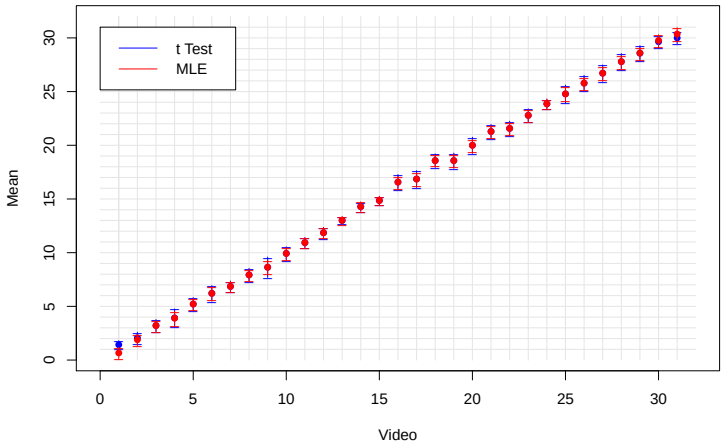


Figure 30: t and ML/bootstrapped confidence intervals of the mean

The majority of the confidence intervals for the means of the responses of the video clips cover the true SQS value of the respective video clips, with the exception of video clips 16, 18 and 31. The mean of video clips 16 and 18 are greater than 16 and 18, respectively, and the mean of video clip 31 is less than 31. The confidence intervals for the mean values obtained from the ML method are mostly slightly narrower than the confidence intervals for the mean values obtained from the t test.

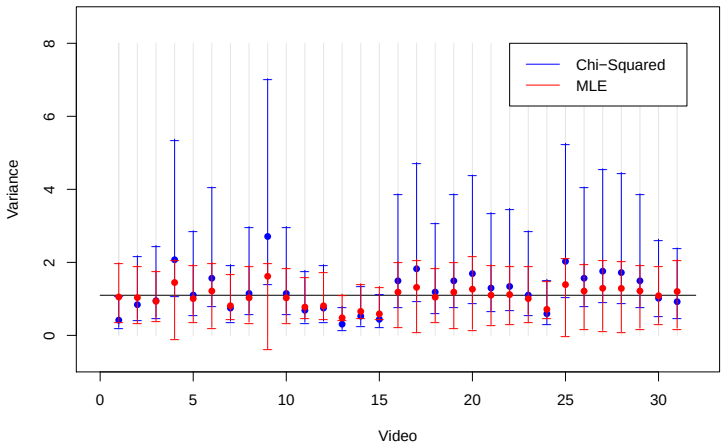


Figure 31: χ^2 and ML/bootstrapped confidence intervals of the variance

In general, the majority of the confidence intervals for the variances obtained through the ML method cover a narrower range than the confidence intervals for the variances obtained through the Chi-Squared test for variance. According to

the ML method, the responses to all videos have the desried variance. However, responses to video 13 only marginally have a variance of 1.099. Additionally, adjacent videos 14 and 15 have low variances.

8.2.4 All Tests

The results of all tests for each video clip are presented in the table below. Cells of the table contain letters signifying whether the relevant test was passed or failed by the video clip. 'P' denotes a pass, and 'F' denotes a fail.

SQS Value	χ^2 GoF	t	χ^2 Variance	ML Mean	ML Variance
1	P	**F**	**F**	P	P
2	P	P	P	P	P
3	P	P	P	P	P
4	**F**	P	P	P	P
5	P	P	P	P	P
6	P	P	P	P	P
7	P	P	P	P	P
8	P	P	P	P	P
9	**F**	P	**F**	P	P
10	P	P	P	P	P
11	P	P	P	P	P
12	P	P	P	P	P
13	P	P	**F**	P	P
14	P	P	P	P	P
15	P	P	F	P	P
16	P	P	P	**F**	P
17	P	P	P	P	P
18	P	P	P	**F**	P
19	P	P	P	P	P
20	P	P	P	P	P
21	**F**	P	P	P	P
22	P	P	P	P	P
23	P	P	P	P	P
24	P	P	P	P	P
25	P	P	P	P	P
26	P	P	P	P	P
27	P	P	P	P	P
28	P	P	P	P	P
29	P	P	P	P	P
30	P	P	P	P	P
31	P	**F**	P	**F**	P

The majority of the video clips do not fail the tests. Video clips 1 and 31 fail the t test of means due to the inability of the t test to account for the truncation of the Normal distribution at the end-points. Additionally, video clip 31 obtains a maximum likelihood estimate of the mean as < 31. The maximum likelihood estimate is for video clip 31 is based on an underlying distribution of a truncated Normal distribution.

Video clips 4, 9, and 21 fail the Chi squared Goodness of Fit test, and video clip 9 fails the Chi squared test of variance. Video clip 4 narrowly passes the Chi squared test of variance.
Video clip 13 fails the Chi squared test of variance, and only marginally achieves the desired variance through MLE.
Video clips 16 and 18 obtain ML estimates of means that are larger than 16 and 18 respectively, and narrowly pass the t test of means.

8.2.5 Tests for the Offset

The true SQS value for each video is subtracted from the responses of each video. The offset values for all videos are collated, and a histogram of the offset of the responses is shown in Figure 32. The line over the histogram demonstrates the ideal shape of the histogram according to the Normal distribution with variance 1.099. It appears that the histogram may be slightly negatively skewed, contrasting the expected postive skew due to the colour non-uniformity of the display.

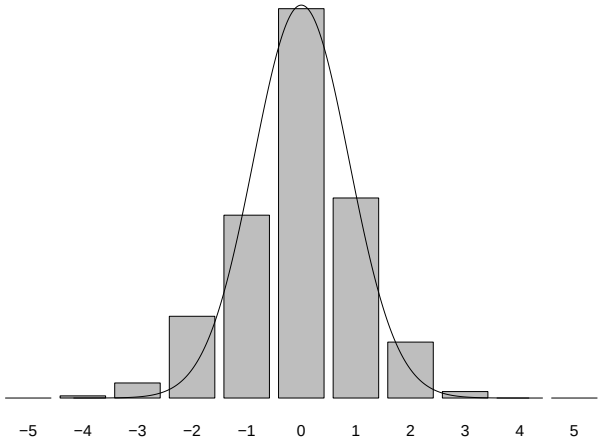


Figure 32: Histogram of the offset

A binned Chi-Squared Goodness of Fit test is performed. The p-value is 0.05145. Though this value is not less than 0.05, it is relatively small, suggesting that there may be some evidence to conclude that the N(0,1.099) model is incorrect. The ML estimators of the offset are $(\hat{\mu}, \hat{\sigma}^2) = (-0.074, 1.08)$. The bootstrapped 95% confidence intervals for the ML estimators are:

$$\mu \in [-0.182, 0.0305]$$
$$\sigma^2 \in [1.01, 1.31]$$

The confidence interval for the mean narrowly covers 0. The confidence interval for the variance covers the ideal variance of 1.099, suggesting that the results as a whole may validate the assumption of a variance equal to 1.099.

8.2.6 Qualitative Results

2 out of the 14 observers (14%) noticed a difference in colour between the video clips.

3 out of the 14 observers (21%) felt that it was more difficult to make a decision for the blurrier videos.

The responses regarding the objects focused on by participants from all three tests (pilot study, larger study, validation study) are collated. 59% of experienced participants and 63% of naïve participants focused on a mixture of static and moving objects. 38% of the experienced participants and 27% of the naïve participants focused solely on static objects. A larger proportion of naïve observers than experienced observers focused solely on moving objects.
The objects which made up a large proportion of participants' responses were, in decreasing order: static text, trees and/or bushes, the bus (which includes any text on the bus), the face of the person walking towards the camera, the clothing of the person walking towards the camera, buildings in the background (in particular, a tall tower), and the texture on the ground.

9 Conclusion

The captured scene was chosen as it contained a wide variety of objects to examine, both moving and static. This was decided with the intention of reducing participant boredom, and to enable an averaged result over all types of scenes. Different participants focused on different objects in order to make a decision on the quality of the video(s). A majority of the participants focused on a mixture of both static and moving objects. It is possible that participants would focus on objects, and when they were unable to come to a decision, they would move to a different object, carrying on until they could come to a decision. As the test progressed, and participants became tired, they may have focused only on the initial object, before making a 'guessed' decision. If this was the case, the decisions would become very object dependent as the test progressed, thus reducing the averaged scene effect of the results.

The new Video Quality Ruler was created with 31 levels, spanning a slightly narrower sharpness range than prescribed in ISO 20462. The results obtained from the larger Paired Comparison test showed that the calibrated levels prescribed in ISO 20462 were not quite accurate for the blurrier videos. The ISO standard suggested that for blurrier images that are further apart in sharpness, the perceptual difference between the images is smaller. The experiments conducted for this thesis dispute this suggestion for moving images. It is unclear whether this case is solely restricted to videos, or still images as well. Contrasting this finding, 21% of the participants of the Video Quality Ruler experiment commented that they felt the blurrier videos were more difficult to tell apart. This appears to support the suggestion in the ISO standard. However, the validation experiment confirmed that the variances of the blurrier videos were as they should be, and thus the change in calibration for the blurrier videos is correct.

The Weber-Fechner Law was tested on the calibration levels of the Video Qual-

ity Ruler. The majority of the levels abided by the law, with the exception of sharper videos, which slightly deviated away from the ideal result. Therefore, the Video Quality Ruler obeys a key law of psychophysics that underpins the studies performed in this thesis.

Contrasting previous studies on image quality which suggested that experienced observers perceive image quality less accurately than naïve observers, the Paired Comparison experiment conducted within this thesis found that the majority of the time, the two groups perceived sharpness with roughly equal accuracy. For a small portion of the tested pairs, the two groups produced differing results, with the experienced observers outperforming the naïve observers more often than not. As the Video Quality Ruler was created from the results of a roughly even number of naïve and experienced observers, it is applicable for use by observers with all levels of image quality experience.

Conduction of the larger Paired Comparison test found that the screen used to display the pairs of video clips is non-uniform in illumination. This was exhibited by a portion of the observers finding the video clips on the right side consistently more 'colourful'. Additionally, a larger number of observers found the experiment too long and felt tired after some time. Both of these effects may have influenced the results by increasing the variance. However, as it was not possible to quantify the change in variance, the variance was assumed to be the original theoretical value. Analyses on the larger Paired Comparison test were conducted with the assumption of the known theoretical variance.
The instructions given to participants informed them that the video clips vary only in sharpness, but to judge the overall quality of the images. If the participants were unaware of the true differences between the videos, it is possible they would focus less on the sharpness, and more on the overall quality, causing a greater proportion of participants to experience the colour difference. Thus, in this case, these particular instructions were beneficial to the experiment.

Implementation of the Video Quality Ruler in a controlled environment did not validate the results of the Paired Comparison test for calibration levels 4, 9, 13, 16, 18, 21 and 31. The remaining levels were found to prove the results obtained in the Paired Comparison test.
Examining the histogram for level 4, in Figure 28, it can be seen that the responses are centred around the true value, however, the shape of the responses does not follow that of the hypothesised Normal distribution. From the high frequencies of the levels surrounding level 4, it is possible to hypothesise that video 4 may be perceptually too close to video 3 and video 5. The p-values of the Paired Comparison test for video pairs (3,4) and (4,5) show that there is very little evidence to suggest that the videos are not 1 JND apart. With an empirical JND of 1.36 for both pairs, the results of the Paired Comparison experiment dispute the hypothesis that the video pairs are too close together in perceptual quality. As level 4 did not fail any of the other tests conducted as part of the validation experiment, it is unclear what has caused the irregular results for level 4.
Responses for level 9 exhibited a larger than expected spread according to the Chi squared test for variance. This is also made clear by examining the histogram for level 9 in Figure 28. Conversely, level 9 was found to have the correct variance by the maximum likelihood estimation method. A p-value of

0.128 from the Paired Comparison test for pair (9,10) shows that there is very little evidence to suggest that videos 9 and 10 are too close together in perceptual quality. Similarly, a p-value of 0.28 for pair (8,9) suggests that there is no evidence to dispute that videos 8 and 9 are 1 JND apart.
Level 13 also failed the Chi-squared test for variance. However, the variance of the responses for video 13 is less than the hypothesised variance, suggesting that videos (12,13) and/or (13,14) are too far apart in perceptual quality. The Paired Comparison test for pair (13,14) produced a much smaller p-value than the test for pair (12,13). Additionally, videos 14 and 15 had (insignificantly) small variances. It is possible that level (13,14) are further than 1 JND apart.
Levels 16 and 18 obtained larger than expected maximum likelihood estimates of the mean value of the responses. This may be attributed to the combination of the non-uniformity of the display and the fact that the ruler video was always placed on the left side in the validation experiment. In addition to this, the p-value of the Paired Comparison test for video pair (18,19) was small, suggesting that there may be some evidence to suggest that video pair (18, 19) are closer than 1 JND.
The histogram of level 21 in Figure 28 shows that the responses are skewed and most of the observers perceive video 21 as video 22. This may also be attributed to the non-uniformity of the display.
The responses for video 31 had a mean value of less than 31. The software created for the Video Quality Ruler specified that the ruler slider is always initialised at the blurriest level. It is possible that observers would start at the blurriest end of the spectrum and move the slider along until they reached what they thought was the correct level, and not go any further to assess the accuracy of their choice. If this was the case, the initial level of the ruler slider would have somewhat of an impact on results. It is possible that this is the case for level 31, the sharpest end of the spectrum.

Compilation of all of the results of the validation experiment suggested that the Video Quality Ruler created through the Paired Comparison test was calibrated correctly. However, there was some evidence to suggest that there is an offset in the mean of responses. There are two possible attributes that have already been explored, that can affect the mean of all responses. The first is the colour non-uniformity of the display, which exhibits itself by a mean value of responses as higher than the true value of the video. The second attribute is the initial positioning of the ruler slider. This would cause a negative offset, opposite to the former attribute. The offset exhibited in the validation experiment was negative, suggesting that the slider effect was more influential than the non-uniformity of the display.

The Video Quality Ruler is, in general, a successful creation. However, a few changes should be made to ensure a more accurate and reliable test procedure. These changes are outlined in Section 10.

10 Suggestions for Future Work

1. The validation experiment showed that some of the calibration levels may be inaccurate. In particular, videos (9,10) and (18,19) may be closer than 1 JND apart, and videos (13,14) may be further than 1 JND apart.

According to the results of the Paired Comparison test and the validation experiment, a new set of calibration levels, denoted by the SQS values of the current Video Quality Ruler, is proposed:

1, 2, 3, 4, 5, 6, 7, 8, 9, 10.5, 11.25, 12.17, 13.25, 14, 15, 16, 17, 18, 19.3, 20.5, 21.5, 22.3, 23.5, 24.5, 25.5, 26.5, 27.5, 28.5, 29.3, 30.5, 31.5

The validity of this hypothesis should be tested either by the Paired Comparison test or through implementation of the Video Quality Ruler.

2. Adjust the software of the Video Quality Ruler, to allow randomisation of the initial slider position.
3. Perform a validation experiment on a more uniform display.
4. Extend the range of quality of the ruler.
5. Adjust the software of the Video Quality Ruler, to allow users to say that a test video clip is outside (either greater than or less than) the quality range of the ruler.
6. Test the Video Quality Ruler whereby the test and ruler videos are displayed vertically (as opposed to horizontally) and/or are displayed in a mirrored format. Compare the results of the test to those without the alterations.
7. Perform a Quality Ruler test with both the Video Quality Ruler and Image Quality Ruler, as prescribed in ISO 20462, simultaneously, to compare the two methods.
8. Compare the Video Quality Ruler with other subjective methods for efficiency and correlation.
9. Perform a Paired Comparison test of images from the Image Quality Ruler to determine the validity of the calibrated levels prescribed in ISO 20462.
10. Develop rulers containing different scenes, by creation of scenes and implementation of validity experiments.

References

[1] `www.axis.com`

[2] Team, J.V., 2003. Advanced Video Coding for Generic Audiovisual Services. *ITU-T Rec. H.264*.

[3] Z. Wang, A. C. Bovik, H. R. Sheikh, E. P. Simoncelli, 2004. Image Quality Assessment: From Error Visibility to Structural Similarity. *Transactions On Image Processing, IEEE*, 13(4), pp.600-612.

[4] A. K. Moorthy, L. K. Choi, G. de Veciana, A. C. Bovik, 2012, January. Subjective Analysis of Video Quality on Mobile Devices In *Sixth International Workshop on Video Processing and Quality Metrics for Consumer Electronics (VPQM), Scottsdale, Arizona.*

[5] M. H. Pinson, L. Janowski, Z. Papir, 2015. Video Quality Assessment: Subjective Testing of Entertainment Scenes. *Signal Processing Magazine, IEEE*, 32(1), pp.101-114.

[6] *Methodology for the Subjective Assessment of the Quality of Television Pictures*. ITU-R BT.500-13 (01/2012).

[7] Y. Chen, K. Wu, Q. Zhang, 2015. From QoS to QoE: A Tutorial on Video Quality Assessment. *Communications Surveys & Tutorials, IEEE*, 17(2), pp.1126-1165.

[8] P. M. A. Kumar, S. Chandramathi, 2015. Video Quality Assessment Methods: A Bird's-Eye View. *World Academy of Science, Engineering and Technology, International Journal of Computer, Electrical, Automation, Contorl and Information Engineering*, 8(5), pp.734-740.

[9] 2005. *ISO-20462: Photography – Psychophysical Experimental Methods for Estimating Image Quality*

[10] M. Persson, 2014. Subjective Image Quality Evaluation Using the Softcopy Quality Ruler Method. *Master's Theses in Mathematical Sciences.*

[11] E. W. Jin, B. W. Keelan, J. Chen, J.B. Phillips, Y. Chen, 2009. Softcopy Quality Ruler Method: Implementation and Validation. In *IS&T/SPIE Electronic Imaging* (pp. 724206-724206). International Society for Optics and Photonics.

[12] R. Thiel, P. Clark, R. B. Wheeler, P. W. Jones, M. Riveccie, J. F. Dupont, 2007. Assessment of Image Quality in Digital Cinema Using the Motion Quality Ruler Method. *Motion Imaging Journal, SMPTE* 116(2-3), pp.61-73.

[13] P. D. Burns, J. B. Phillips, D. Williams, 2013. Adapting ISO 20462 Softcopy Quality Ruler Method for on-line Image Quality Studies. In *Proceedings of SPIE-IS&T Electronic Imaging Symposium* (pp. 86530E-1).

[14] P. G. Freitas, J. A. Redi, M. C. Q. Farias, A. F. Silva, 2015. Video Quality Ruler: A New Experimental Methodology for Assessing Video Quality. In *Quality of Multimedia Experience (QoMEX), 2015 Seventh International Workshop on* (pp. 1-6). IEEE.

[15] T. Maintz, 2005. Digital and medical image processing. *Universiteit Utrecht.*

[16] R. W. G. Hunt, M. R. Pointer, 2011. *Measuring Colour.* 4th Edition. Chichester. John Wiley & Sons, Ltd.

[17] `http://www.webmd.com/eye-health/understanding-vision-problems-symptoms`

[18] S. Sagers, R. Patterson (2010) *Mechanics of a Digital Camera*. Mechanics.

[19] `http://www.cambridgeincolour.com/tutorials/camera-lenses.htm`

[20] `http://www.cambridgeincolour.com/tutorials/camera-exposure.htm`

[21] `http://www.cambridgeincolour.com/tutorials/depth-of-field.htm`

[22] `http://www.cambridgeincolour.com/tutorials/white-balance.htm`

[23] M. Tkalcic, J.F. Tasic, 2003. Colour Space - Perceptual, Historical and Application Background. In *Eurocon*.

[24] International Electrotechnical Commission, 1999. *Multimedia systems and equipment-colour measurement and management-part 2-1: Colour management-default RGB colour space-sRGB*. IEC 61966-2-1.

[25] G. Sharma, 2003. *Digital Color Imaging Handbook*. Boca Raton, Florida. CRC Press LLC.

[26] `https://en.wikipedia.org/wiki/YUV`

[27] Miano, J., 1999. *Compressed Image File Formats: Jpeg, png, gif, xbm, bmp.* Reading, Massachussets. Addison-Wesley Professional.

[28] `https://en.wikipedia.org/wiki/Motion_JPEG`

[29] `https://w3techs.com/blog/entry/the_png_image_file_format_is_now_more_popular_than_gif`

[30] R. Ramanath, W. E. Snyder, Y. Yoo, M. S. Drew, 2005. Color Image Processing Pipeline. *Signal Processing Magazine, IEEE*, 22(1), pp.34-43.

[31] Bayer, B.E., Eastman Kodak Company, 1976. *Color imaging array.* U.S. Patent 3,971,065.

[32] G. D. Boreman, 2001. *Modulation Transfer Function in Optical and Electro-Optical Systems*. (Vol. 4). Bellingham, Washignton: SPIE Press.

[33] J. K. M. Roland, 2015. A Study of Slanted-Edge MTF Stability and Repeatability. In *IS&T/SPIE Electronic Imaging* (pp. 93960L-93960L). International Society for Optics and Photonics.

[34] W. F. Hsu, K. W. Chuang, Y. C. Hsu, 2000. Comparisons of the camera OECF, the ISO speed, and the SFR of digital still-picture cameras. In *Photonics Taiwan* (pp. 104-111). International Society for Optics and Photonics.

[35] B. Porat, 1997. *A Course in Digital Signal Processing.* New York, New York. John Wiley.

[36] R. C. Gonzalez, R. E. Woods, 2002. *Digital Image Processing.* 2nd Edition. New Jersey. Prentice Hall

[37] W. Burger, M. J. Burge, 2008. *Digital Image Processing: An Algorithmic Introduction using Java.* New York, New York. Springer Science & Business Media.

[38] A. Deitmar, 2005. *A First Course in Harmonic Analysis.* Second Edition. New York, New York. Springer-Verlag New York, Inc.

[39] K. Turkowski, 1990. Filters for Common Resampling Tasks. In *Graphics gems* (pp. 147-165). Academic Press Professional, Inc..

[40] `https://en.wikipedia.org/wiki/Lanczos_resampling`

[41] J. B. Phillips, S. M. Coppola, E. W. Jin, Y. Chen, J. H. Clark, T. A. Mauer, 2009. Correlating objective and subjective evaluation of texture appearance with applications to camera phone imaging. In *IS&T/SPIE Electronic Imaging* (pp. 724207-724207). International Society for Optics and Photonics.

[42] `http://www.explainthatstuff.com/camcorders.html`

[43] `https://en.wikipedia.org/wiki/Contrast_(vision)`

[44] `https://en.wikipedia.org/wiki/Diffraction-limited_system`

[45] J. Sasian, 2012. *Introduction to Aberratoins in Optical Imaging Systems.* Cambridge. Cambridge University Press.

[46] `http://www.photoreview.com.au/tips/shooting/sharpness,`
`-acutance-and-resolution`

[47] `https://photographylife.com/`

[48] A. B. Tucker, 2004. *Computer Science Handbook.* 2nd Edition. Boca Raton, Florida. CRC Press.

[49] Z. Wang, A. C. Bovik, 2006. *Modern Image Quality Assessment.* Morgan & Claypool.

[50] G. A. Gescheider, 2006. *Psychophysics: The Fundamentals.* 3rd Edition. Mahwah, New Jersey. Lawrence Erlbaum Associates, Inc..

[51] B. W. Keelan, 2002. *Handbook of Image Quality.* Boca Raton, Florida. Taylor & Francis Group, LLC.

[52] P. G. Engeldrum, 2000. *Psychometric Scaling.* Winchester, Massachusetts. Imcotek Press

[53] E. Jin. System Requirements for Implementing the Softcopy Quality Ruler Method.

[54] D. C. Montgomery, 2013. *Design and Analysis of Experiments*. Eighth Edition. Hoboken, New Jersey. John Wiley & Sons.

[55] R. G. Miller, Jr., 1991. *Simultaneous Statistical Inference*. New York, New York. Springer-Verlag.

[56] G. K. Kanji, 2006. *100 Statistical Tests*. 3rd Edition. London, UK. Sage Publications.

[57] H. C. Thode, Jr., 2002. *Testing for Normality*. New York, New York. Marcel Dekker, Inc..

[58] A. Ghasemi, S. Zahediasl, 2012. Normality Tests for Statistical Analysis: A Guide for Non-Statisticians. *International journal of endocrinology and metabolism*, 10(2), pp.486-489.

[59] N. M. Razali, Y. B. Wah, 2011. Power Comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling tests. *Journal of Statistical Modeling and Analytics*, 2(1), pp.21-33.

[60] P. D. Ellis, 2010. *The Essential Guide to Effect Sizes*. Cambridge. Cambridge University Press.

[61] J. Aldrich, 1997. R. A. Fisher and the Making of Maximum Likelihood 1912 - 1922. *Statistical Science*, 12(3), pp.162-176.

[62] J. Nocedal, S. J. Wright, 2006. *Numerical Optimization*. Second Edition. New York, New York. Springer Science + Business Media, LLC.

[63] H. Cramér, 1962. *Mathematical Methods of Statistics*. Bombay. Asia Publishing House.

[64] B. Efron, 1992. *Bootstrap Methods: Another Look at the Jackknife*. (pp. 569-593). Springer New York.

[65] E. Seneta, 2013. A Tricentenary history of the Law of Large Numbers. *Bernoulli*, 19(4), pp.1088-1121.

[66] B. Efron, R. J. Tibshirani, 1998. *An Introduction to the Bootstrap*. Boca Raton, Florida. CRC Press LLC.

[67] `https://www.ephotozine.com/article/canon-ts-e-45mm-f-2-8-lens-review-26462`

[68] `http://www.imaging-resource.com/PRODS/canon-1dc/canon-1dcA.HTM`

[69] H. Fischer, 2011. *A History of the Central Limit Theorem*. New York, New York. Springer Science + Business Media, LLC.

[70] R. A. Bradley, M. E. Terry, 1952. Rank analysis of incomplete block designs: I. The method of paired comparisons. *Biometrika*, 39(3/4), pp.324-345.

[71] K. Tsukida, M. R. Gupta, 2011. *How to Analyze Paired Comparison Data*. (No. UWEETR-2011-0004). WASHINGTON UNIV SEATTLE DEPT OF ELECTRICAL ENGINEERING.

[72] L. L. Thurstone 1927. A Law of Comparative Judgement. *Psychological review*, 34(4), p.273.

[73] `http://i-see.org/block_letter_eye_chart.pdf`

[74] C. Li, 1981. *New Color Vision Testing Chart*. Liaoning. Liaoning People's Press