Image Quality Evaluation in Lossy Compressed Images

Allen, E.

IMAGE QUALITY EVALUATION IN LOSSY COMPRESSED IMAGES

Elizabeth Allen, BSc(Hons), MSc

A thesis submitted in partial fulfilment of the requirements of the University of Westminster for the Degree of Doctor of Philosophy

This research programme was completed within the Imaging Technology Research Group at the University of Westminster

February 2017

Abstract

This research focuses on the quantification of image quality in lossy compressed images, exploring the impact of digital artefacts and scene characteristics upon image quality evaluation.

A subjective paired comparison test was implemented to assess perceived quality of JPEG 2000 against baseline JPEG over a range of different scene types. Interval scales were generated for both algorithms, which indicated a subjective preference for JPEG 2000, particularly at low bit rates, and these were confirmed by an objective distortion measure. The subjective results did not follow this trend for some scenes however, and both algorithms were found to be scene dependent as a result of the artefacts produced at high compression rates. The scene dependencies were explored from the interval scale results, which allowed scenes to be grouped according to their susceptibilities to each of the algorithms. Groupings were correlated with scene measures applied in a linked study.

A pilot study was undertaken to explore perceptibility thresholds of JPEG 2000 of the same set of images. This work was developed with a further experiment to investigate the thresholds of perceptibility and acceptability of higher resolution JPEG 2000 compressed images. A set of images was captured using a professional level full-frame Digital Single Lens Reflex camera, using a raw workflow and carefully controlled image-processing pipeline. The scenes were quantified using a set of simple scene metrics to classify them according to whether they were average, higher than, or lower than average, for a number of scene properties known to affect image compression and perceived image quality; these were used to make a final selection of test images. Image fidelity was investigated using the method of constant stimuli to quantify perceptibility thresholds and just noticeable differences (JNDs) of perceptibility. Thresholds and JNDs of acceptability were also quantified to explore suprathreshold quality evaluation. The relationships between the two thresholds were examined and correlated with the results from the scene

measures, to identify more or less susceptible scenes. It was found that the level and differences between the two thresholds was an indicator of scene dependency and could be predicted by certain types of scene characteristics.

A third study implemented the soft copy quality ruler as an alternative psychophysical method, by matching the quality of compressed images to a set of images varying in a single attribute, separated by known JND increments of quality. The imaging chain and image processing workflow were evaluated using objective measures of tone reproduction and spatial frequency response. An alternative approach to the creation of ruler images was implemented and tested, and the resulting quality rulers were used to evaluate a subset of the images from the previous study. The quality ruler was found to be successful in identifying scene susceptibilities and observer sensitivity.

The fourth investigation explored the implementation of four different image quality metrics. These were the Modular Image Difference Metric, the Structural Similarity Metric, The Multi-scale Structural Similarity Metric and the Weighted Structural Similarity Metric. The metrics were tested against the subjective results and all were found to have linear correlation in terms of predictability of image quality.

Acknowledgements

I would like to extend sincere thanks to my supervision team for their support, encouragement, and understanding and inspiration, during the time of my research. Dr Sophie Triantaphillidou has been my friend, colleague and supervisor, and has believed in, cajoled and encouraged me throughout. Professor Ralph Jacobson has been my teacher, colleague, and mentor. Both have helped to nurture my ideas and academic development for many years and have provided their huge experience and expertise openly and generously. I would also like to thank Dr Aleka Psarrou for her support, encouragement and interest in the work of our research group.

I would further like to thank my colleagues, Dr Efthimia Bilissi and John Smith, for the many hours of discussion and for their support in my research and their contributions to the academic work and research of our team. There are many other people within our department who have provided useful discussions, assistance and guidance in technical issues, including Jae Young-Park who deserves a special mention for all of his good-natured patience and help; Gaurav Gupta for giving up laboratory space and assisting with MATLAB implementation; Ana Tsifouti for discussion about methodology; Danny Garside and James Pickett for technical support; the photographic technical team and all the observers who have participated in my tests.

I also thank the many members of the graduate school for facilitating my ongoing research, and to the Faculty Dean, Professor Kerstin Mey, for her interest and support in the completion of my PhD and in the research of our group.

Finally I want to thank my family: My parents, who have always supported me in every way possible and in more ways than I could have appreciated; with encouragement, but without expectation; to Charlotte and Ben for always being there. Lastly I dedicate this to Jason, Anja and Freddie who provide their love and appreciation no matter what, and who mean more to me than anything.

Contents

List of Figures

List of Tables

Author’s Declaration

I declare that the work submitted in this thesis is my own. Any other sources of information, visual material or data produced by other researchers used directly or indirectly have been referenced appropriately.

Elizabeth Allen, 2017

1 Introduction

Digital images are, by now, so sophisticated, ubiquitous and embedded in modern culture that it seems somewhat incredible that consumer digital cameras were widely available only in the mid-1990s. Since the development of the first digital still camera in 1986 [HYPERLINK \l "Nak06" 1], which used a charge-coupled device (CCD) with only 250,000 pixels; pixel size, sensor technology, digital cameras, and imaging applications have continued to evolve in complexity, interactivity and interoperability with multiple systems and devices. Currently, digital cameras are a standard component of many types of mobile device, including phones, tablets, laptops and lately, smart watches. Consumer digital imaging has represented the biggest growth market in imaging for a number of years.

Many factors affect the required file size and the pixel resolution of the images produced, depending upon the image quality requirements of a given application, but also the output and on-going workflow of the images, and the way in which they are disseminated and archived. Data storage and processing capabilities, as well as transmission bandwidths, have progressed in tandem with, and driven by, the enormous expansion in imaging applications, and the need to compress data still remains. Data visualisation, new imaging modes and methodologies, and increasing file size ensure that image compression remains an important area of research.

1.1 Image Compression

The majority of images used in this research were captured using a professional level digital SLR, the Canon EOS 5D Mark II, with a full frame sensor resolution of approximately 21.1 megapixels and effective pixels 5616 x 3744 2]. A fully rendered, uncompressed 8-bit RGB image from this camera has a file size of 60.2 megabytes (Mb). At the time of writing there are a number of smartphones available with 16 megapixel image sensors. Imaging applications on smart phones and mobile devices have developed exponentially in the last few years, bringing relatively sophisticated image processing, High Dynamic Range (HDR) imaging and panoramic imaging (to name but a few) to the consumer, and are used to disseminate and share images to an extent which could not have been imagined when digital cameras were first developed.

A digital image is a discrete representation of either original scene intensities, or of the intensities of an analogue image, both of which are continuous intensity[1] functions. The input function is sampled and quantised at a required level, to ensure that discontinuities are not apparent to a human observer under given viewing conditions, as a result of the digitisation process.

Digitisation is achieved by sampling and quantisation of the original and may be represented by:

$$f(x, y, t, \lambda) = f'_n(i, j)$$

(1.1) [HYPERLINK \l "Tri11" 3][2]

Where *f* is the intensity level in the original scene or image at spatial location [*x, y*], integrated across spectral band *λ* for time *t,* and *f'* is the digital level assigned at the equivalent position in the output image denoted by discrete integers [*i,j*]. *n* denotes the nth colour channel.

The maximum values of *i* and *j* are *M* and *N,* the number of samples horizontally and vertically; the maximum of *n* is *C,* the number of spectral bands, and the maximum value of *f'* is *L-1* from:

$$L = 2^b$$

(1.2)

[1] Intensity is a generic term for input values, which might be brightness, luminance, illuminance, reflectance, transmittance, density

[2] Adapted from [HYPERLINK \l "Tri11" 3] using the convention of i, and j for rows and columns in the output image

Where L is the number of discrete quantised levels per channel and b is the bit depth per channel. The file size (S) of a digital image in bits is therefore given by:

$$S = M \times N \times C \times L \qquad (1.3)$$

The uncompressed file size is calculated assuming that the full number of bits are allocated per pixel, however this is rather inefficient, as there are numerous sources of redundancy within digital images of natural scenes [4] [5], which may be exploited to reduce file size. Redundancies may exist in the data (the bit stream used to encode the image) conveying the image information, or in the image information itself (the scene content).

Compression methods are either *lossless* or *lossy*. Lossless coding aims to reduce the average bit rate (bit allocation per pixel) without the loss of any information. Bovik [6] says:

"In lossless coding the image data should be identical both quantitatively (numerically) and qualitatively (visually) to the original encoded image"

Lossless compression is a requirement of certain imaging applications where image fidelity is paramount, for example in some forensic or medical imaging applications; it relies on the removal or reduction of correlation within images, and the reorganisation and encoding of the data in the most efficient manner possible. This is most successfully achieved by pre-processing the image to highlight or exploit the correlations, followed by a reduction in the amount of data using both interpixel and coding redundancy. However, the achievable compression rate is limited and is inadequate for many applications.

Lossy methods achieve much greater compression by reducing the image information as well as the data conveying it [4] [5] [6] [7]. Information beyond the limits of, or less relevant to, the human visual system, is removed or reduced, with a potential expense to image quality as a result of distortion being introduced. At lower compression rates, some lossy compression methods are considered *perceptually lossless*, as described by Bovik [6]:

“These methods require that the encoded and decoded images be only visually, and not necessarily numerically, identical”

In fact, the degree to which the decoded images are required to appear visually identical to their uncompressed originals is both context and image dependent. In reality, the images are unlikely to be viewed next to their original for comparison outside an imaging lab, therefore the degree to which distortions introduced by compression (and other sources of distortion in the imaging chain) are tolerated, depends upon the purpose of the image, the viewer and viewing environment and the visibility of the distortions.

Two such methods have been developed by the Joint Photographic Experts Committee, a “joint working group of the International Standardization Organization (ISO) and the International Electrotechnical Commission (IEC)” [8] set up for the specification of compression standards for continuous tone images.

The JPEG baseline standard [9] [10]**,** introduced in 1991, is a discrete cosine transform-based (DCT) algorithm originally developed for images displayed on screen, at a time when no such standard existed. JPEG has a number of characteristic distortions and its development occurred at a time when digital imaging was much less evolved and complex, meaning that it has some limitations when used in the modern imaging environment. For example, its native YCbCr colour space has a gamut developed for images to be viewed upon Cathode Ray Tube (CRT) displays-, and it is limited to 8-bit greyscale or 24-bit RGB images [10]. The nature of the quantisation means that it is rather poor at compressing images containing text. Nevertheless, baseline JPEG is probably the most widely adopted image compression method in existence at the time of writing; it is a standard rendered file format in the majority of digital cameras and has supplanted the majority of other image file formats in use in images across the Internet.

A newer approach to transform based compression uses the Discrete Wavelet Transform (DWT), as in the more recently specified JPEG 2000 standard [11]. Part 1 of the standard uses the DWT with bit-plane encoding. As well as providing improved compression efficiency and error resilience, it allows multiple resolution representation and progressive coding of image data. It has been developed to provide a number of other additional functionalities, including [12]: Lossy to lossless layer progressive encoding, image tiling, region of interest coding, random access and processing, improved colour space and ICC profile support and the use of alpha channels to meet the future needs of graphics and internet applications. Lossy JPEG 2000 has its own set of distinctive artefacts, which affect images in a different way to baseline JPEG.

As JPEG 2000 is the main algorithm investigated in this research it is of interest to note that the lossy version has been much less widely adopted than its predecessor. Nevertheless, JPEG 2000 is in use in a number of specialist applications, such as forensic imaging, where research into its use is ongoing [13]. Baseline JPEG has now been available for over two decades. It may be that consumer tolerance has habituated to the characteristic JPEG artefacts more than those of JPEG 2000. Otherwise this may be a reflection upon consumer preference in terms of visibility of distortions. It could simply be that JPEG is fit-for-purpose and there is not the motivation for manufacturers or consumers to invest in an alternative. The question of one type of artefact being more acceptable than another is an interesting area to explore in image quality.

1.2 Evaluating Compression

Because lossless coding methods result in images that are identical to their originals, their evaluation is generally focused upon the amount of compression, compression time and flexibility or usefulness.

Compression rate is calculated as a simple ratio of the output file size to input:

$$C_r = \frac{n_1}{n_2} \qquad (1.4)\ [4]$$

Where C_r is compression ratio, and n_1 and n_2 are the original and the compressed data sets respectively. The rate of compression may instead be expressed as a compression percentage, or more commonly as a bit rate, which defines the average number of bits required to represent a pixel in the compressed image. This is meaningful only when compared with the number of bits per pixel allocated in the original uncompressed image [5]. In lossy compression, compression rate is not enough as a performance measure, because of the distortion introduced, which must be somehow quantified both in terms of amount and relevance to the viewing system within the given context.

The simplest approaches to the assessment of the distortion evaluate the numerical difference between the original and the compressed image, which may then be used to form a rate-distortion curve, a graph of compression rate against distortion [5]. These so called *distortion metrics* are simple to calculate and quantify the total difference between the two images the average difference per pixel (Mean Absolute Error or Mean Squared Error), or the average difference in relation to the signal (Signal-to-Noise Ratio or Peak Signal-to-noise Ratio). Based upon first order statistics, these measures give no information about the spatial effects or location of the error [14] and are therefore poorly correlated with the perceived image quality of the compressed images.

1.3 Aims and Overview of the Project

This research explores image quality and compression with a particular focus upon scene dependency, that is, the influence of scene characteristics upon compression performance and their impact on human visual assessments of image quality.

The concept of image quality is difficult to define. It has different meanings depending upon context, application and purpose of the images. It varies widely across observers, images, systems, and scenes. The methodologies and approaches used in image quality evaluation are varied, originating from many different disciplines including imaging science, computer science and computer vision and psychology, to name a few. Objective image quality methods at their simplest quantify amounts of distortion and at their most complex attempt to predict image quality by modelling the combined effects of the scene, system components and visuo-cognitive processes of the human visual system. Subjective methods use *psychophysics,* which is the scientific study of the relationship between physical stimuli and human perception and involves the use of observers to evaluate images.

Lossy compression is just one of the many sources of artefacts that are introduced into images in the digital imaging chain. Understanding the visual effects of these artefacts is of fundamental importance in the evaluation of lossy processes. The interactions between scene, imaging system and the human visual system will ultimately determine whether an application, method or device is successful and useful.

Chapter 2 introduces key theory underpinning image quality evaluation. Image quality is defined, and factors affecting image quality evaluation are discussed. Objective and subjective image quality evaluation approaches are introduced and summarised. Image quality attributes and their objective quantification are described, including an overview of image quality metrics. Human visual perception and visual phenomena affecting image quality judgements are described, in relation to the design of metrics. Psychophysical methods are also reviewed.

Chapter 3 introduces image compression in more detail. Redundancy in images is discussed as a basis for compression. Lossy transform based compression is defined and the architectures of JPEG and JPEG 2000 are described. Image artefacts and their effects upon image quality judgements, as well as their interactions with scene content are detailed. Scene dependency and its sources are summarised along with their potential impact upon image quality judgements. Approaches to the classification of scenes according to their properties and susceptibility to artefacts are discussed.

Chapter 4 describes a psychophysical experiment to evaluate JPEG and JPEG2000 in terms of their preferred image quality. A paired comparison experiment is implemented to evaluate observer preference between the two algorithms. An associated study classifying the scenes used in the experiment is also described.

Chapter 5 reports on an experiment to evaluate perceptibility and acceptability thresholds for JPEG 2000. The imaging workflow is described in detail. Tone reproduction characteristics at all stages of the image-processing pipeline are evaluated. Scene metrics are again used to describe the test scenes. Correlations are sought between perceptibility and acceptability thresholds and scenes are clustered. Similarities between scene characteristics within the clusters are identified.

Chapter 6 describes the implementation of ISO 20462-3 [15], the Soft-Copy Quality Ruler, a standardised psychophysical method that uses images of the same scene separated by known increments of quality JNDs (but varying only in a single attribute, in this case sharpness) as a set of standards against which test images may be matched. The images are presented in the form of a slider allowing the user to select the matched image interactively. The results are explored in terms of their application in identifying scene susceptibility and observer sensitivity when implementing a psychophysical test.

Chapter 7 describes experimental work to correlate a number of objective metrics with the subjective results. The selected metrics are, where applicable, adapted using the results from the experimental work from the earlier chapters.

Chapters 8 and 9 discuss the experimental work and the implications of the results, drawing final conclusions and proposing ideas for further research.

Chapter 10 summarises related work carried out during the process of this research.

2 Image Quality

2.1 Image Quality Definitions

The now well-established era of digital imaging has led to the convergence of many different disciplines into a relatively generic imaging industry [16], which encompasses numerous approaches to product design and performance assessment. Product cycle times continue to decrease, to keep up with the many developments in imaging applications and the burgeoning array of modes by which images are used and disseminated. These factors lead to an increasing need for fast and accurate methods of evaluating and predicting perceived image quality.

A number of definitions of image quality exist, including those by Jacobson [17]: "the subjective impression formed in the mind of the observer of the degree of excellence exhibited by an image" and Engeldrum [18]: "the integrated set of perceptions of the overall degree of excellence of the image". These descriptions imply that image quality cannot be separated from consideration of the observer, that it is fundamentally subjective in nature, and that it is the result of the combination of a number of different perceptual attributes.

Keelan [16] explains more specifically: "the quality of an image is defined to be an impression of its merit or excellence, as perceived by an observer neither associated with the act of photography, nor closely involved in its subject matter". As a result of this strict definition of the observer, the influences of what Keelan describes as 'personal attributes', which are those that affect the quality judgement of an image by someone who has been personally involved with the image, capture or with the subject matter, are eliminated. An example of such an attribute is 'the preservation of a cherished memory' which reduces the objectivity of an observer's judgement, making prediction of perceived image quality difficult.

These definitions indicate the complexity inherent in image quality investigations. It is not only difficult to describe something which is ultimately about a subjective impression in the mind of an observer, but image quality is multi-dimensional, the combined impact of variations in a range of attributes upon a number of different systems. The dimensions relating to single attributes and the various systems affected combine and interact in ways that are difficult to understand and predict.

2.2 Factors Affecting Image Quality Evaluation

According to Ford [19], a fundamental assumption in approaching the quantification of image quality must be that a relationship exists between measureable physical properties of the image and imaging systems and the subjective impression of quality that the image produces. He describes image quality as the interaction of three main systems with the original image properties to produce the overall perception of image quality for the observer. These systems are the display system, which includes viewing conditions, the 'visual' system, and the 'cognitive' system; clearly some aspects of these are more readily described and quantified than others. As shown in figure 2.1, such a model considers the image in different *states* as it passes through the different systems.

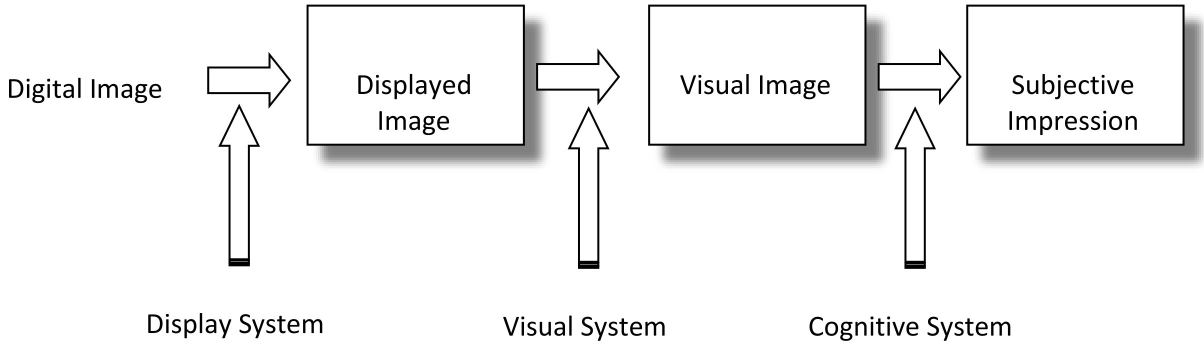


Figure 2.1 Image types and their location in a digital imaging system, according to Ford [19]

Some objective approaches to image assessment are based on the idea that the influence of the visual system on the perceived image quality is primarily as a result of lower level processing performed in the eye and early on in the visual pathway. These methods use models of various aspects of human visual processing, based upon knowledge of characteristics of the human visual response. Examples of such characteristics include the nonlinear visual response to changes in luminance and the contrast sensitivity function(s), describing the frequency response of the human visual system (HVS). Such methods assume that more complex processing such as feature extraction, pattern matching and the processing as a result of changes in attention are of secondary importance in terms of quality perception, to lower level visual processing [20]. Effectively they are measuring the 'visual image' from figure 2.1, without predicting anything about its interpretation.

The implication in such an approach is that the visual image, *prior* to cognitive processing, may thus be simply described as the result of the physical properties of the original stimuli (scene or image), the influence of the viewing conditions, and various low level linear and non-linear processing performed by the HVS. There is a question, however, as to whether it is possible to so clearly define a separation between the visual and cognitive systems, whether such a 'visual image' exists, considering the continuously changing nature of attention and selection performed as an observer looks at the scene in front of them. Where methods based on psychophysical aspects of the HVS are fairly successful at predicting thresholds of detection in images (as determined in *image fidelity* studies, see section 2.4), they are not so effective for suprathreshold quality estimation [21]. This seems to suggest that the emphasis between high-level and low-level visual processing and cognitive processing changes, depending upon the visual task and image context.

Jacobson and Triantaphillidou [22] summarise some of the additional factors that influence observer image quality judgements (as well as the particular imaging context) in the term 'quality consciousness'. Quality consciousness includes the combined impact of memory, association, experience in judging images, scene content, emotions, environmental conditions and many other factors difficult to define or quantify. Ahumada and Null [23] suggest, for example, that when various types of image artefacts are suprathreshold, the variations in observers' experiences of the artefacts, which affect their quality consciousness, will lead to differential weightings of the artefact dimensions.

Engeldrum notes that the criteria for judging quality vary significantly depending upon the imaging context [24]. For example, in medical or forensic imaging, the perceived image quality is often dependent on an expert observer being able to detect and recognise image features and to interpret them in the context of a particular image class. This idea of 'fitness of purpose' of images in quality evaluations is more formally explored in the measure of 'usefulness' as a component of image quality, described in the next section. This can be contrasted against a more generalised view of quality judgement, which he describes as: "a 'beauty contest' selection from images produced by competing products". Indeed, the motivation for much image quality research comes from the manufacturers of imaging products; in many cases, they are interested in the judgements and perceptions of non-expert observers. The visual tasks in these two broad classes of image quality judgement are different, or rather are prioritised differently.

Consideration of the image in context is clearly extremely important as this has a primary effect on the expectations of the observer. Janssen and Blommaert [25] coined the terms *usefulness* and *naturalness* of images to describe requirements of image quality relating to context and expectation. The former relates to the suitability of the image for a task or application; the latter to the relationship between the image characteristics and the observer's 'internal references', which may be thought of as a combination of memory, association and expectations (equivalent to quality consciousness described above).

Yendrikhovskij [26] notes that researchers in colour science have a long history of investigating the relationship between memory and preference. A number of experiments [27] [28] have identified inconsistencies between the measured colour of objects, so-called 'memory colours' and preferred colours. It has been suggested by Newhall et al [29] that these discrepancies are caused by the influence of *memory prototypes*, which are examples of typical colours in an object category (the average of the category), stored in the memory and used for comparison with actual object colours.

Contemporary theories of visual perception suggest that our understanding of scenes is based upon a hierarchical structure of perceived attributes, which may be broadly classified into high-level and low-level attributes, and which contribute to two different visuo-cognitive mechanisms [21]. Hochstein and Ahissar [30] define the two levels in the hierarchy, as "vision at a glance", which produces a general categorisation of scene content, a 'broad brush' impression of the scene, and "vision with scrutiny", where attention is focused on details. They propose a new view of the hierarchy, contrary to the classical view, called 'Reverse Hierarchy Theory' for the order in which these two levels operate. The classical view of visual hierarchy suggests that the outputs from neurons from low-level cortical areas, e.g. those dealing with visual inputs to represent simple attributes such as edges and lines, are gradually combined with those from other low level attributes, at subsequent cortical levels, to build up an overall understanding of global features, from the bottom-up. Reverse Hierarchy Theory suggests that a scene is instead perceived using a top-down hierarchy, beginning with high-level processing in the cortex, to produce a generalised impression of the scene, with in-depth scrutiny of the scene following later, to fill in the details.

Work by Leisti et al [21] suggests that subjective quality evaluations are based upon what they term *Image Quality Experience*, in which observers also use a hierarchy of high and low-level image quality attributes. According to descriptions by observers of their image quality experience, subjective attributes can be 'concrete' and related in a straightforward manner to physical properties of an image, or more general and 'abstract' (although in some cases such attributes can be demonstrated to relate directly to physical attributes, but the context of the relationship is less clear). Examples of the former type include 'sharpness' and 'graininess', while the latter are less easily defined, such as 'naturalness' or 'clarity'. The two classes of attributes are interrelated, and interdependent, with a hierarchy which may be related to the hierarchical nature of human vision as described above. This could in part explain the difficulty in correlating traditional objective approaches for quality prediction, with results from subjective experiments. The relationships between physical image properties and low level attributes are well defined, as are suitable objective measurements for them, many of which are described in section 2.6. Defining types of abstract high level attributes, as well as their relationships to physical image properties is a more complex task.

Complementing the different perspectives on visual perception and subjective quality evaluation, theories on image quality also consider the problem both from top-down and bottom-up:

As described by Yendrikhovskij [26], a bottom-up approach to quality explores the physical parameters underlying image quality, such as properties of the imaging system, the display and viewing conditions, and the properties of the human visual system. This classical 'signal processing' perspective is typical of traditional approaches to image quality metrics. The signal is processed by the physical parameters of system and HVS to produce perceptual attributes such as brightness, colourfulness etc. Individual perceptual attributes are assumed to combine to form higher-level attributes such as image quality. More recently, image quality has been considered from an 'information processing' perspective, which is a top-down approach; visual information is processed, reconstructed and interpreted in the context of memory representations, such as the prototype memory colours discussed earlier. Yendrikhovskij notes that the requirements of both signal processing and information processing approaches may be used to develop a general model of image quality. He describes [26] a model containing the three attribute dimensions: *fidelity, usefulness* and *naturalness* (FUN), in which the overall quality of an image can be modelled as a weighted sum of the three attributes. The FUN model is illustrated in figure 2.2.

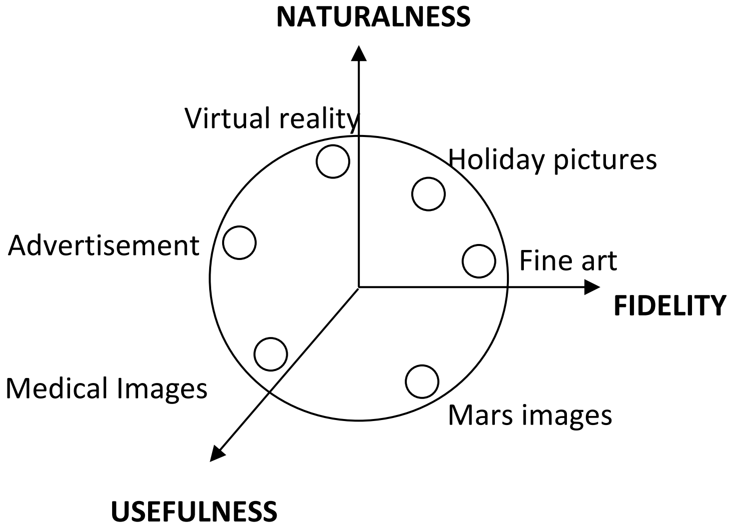


Figure 2.2 Different image types illustrated in the FUN model of image quality, from Yendrikhovskij [26]

The *Fidelity* attribute is the degree of correspondence between the reproduced image with the external reference (the original). In classical image quality approaches, fidelity metrics are used to identify thresholds of perceptibility. Fidelity is described in more detail in section 2.4. A high degree of fidelity is important in applications requiring image matching.

Usefulness, as described earlier by Janssen and Blommaert [25] considers the image in context and quantifies its suitability for a particular task. Usefulness is important in imaging applications requiring detection or discrimination of objects or details.

Naturalness is the degree of correspondence between the reproduced image and internal reference; the knowledge of reality as stored in memory i.e. prototypes. Yendrikhovskij in [31], states that 'a basic assumption underlying the naturalness constraint is that images of good quality should at least be perceived as natural'. This is not necessarily the case however, when all three attributes are considered to be components of image quality.

The definitions above indicate that the three attributes may be conflicting. Janssen and Blommaert note, particularly, that the requirements of usefulness and naturalness may not coincide. For example, maximum usefulness may be achieved by enhancing an image to improve object discrimination (for example by contrast enhancement or edge detection), thereby reducing its naturalness (and its fidelity).

2.3 Objective and Subjective Image Quality Assessment

Objective image quality *measures* involve the physical measurement of imaging systems, images and image attributes which contribute to overall quality. Engeldrum calls them - in the context of his *Image Quality Circle,* [18], [32], - 'Physical Image Parameters'. Research into the measurement of image structure, tone and colour reproduction has a long and well-documented history, leading to numerous standardised practices. Many have been devised and are used individually as performance measures for imaging devices, processes and systems, for comparison, or to drive product development. Various adaptations of these standard methods have been developed for the assessment of digital imaging systems. Some of the attributes and their measurement are described later in this chapter. However, as noted by Engeldrum, although an important aim in developing robust objective methods of image evaluation is to relate image attributes to image quality, individual measures alone are typically unsuccessful, as they quantify the image properties but not the visual system.

Nevertheless, performance measures are important as components in objective image quality metrics (IQMs). Visual image quality metrics (VIQMs) aim to combine physical measurements from images with psychophysical characteristics of the human visual system to predict perceived image quality (or fidelity – see below), often producing single numbers or figures of merit. Jacobson [22] notes, however, that one of the most significant questions asked by researchers in relation to IQMs is whether the single-number approach is a valid one in the assessment of image quality. Nevertheless, a wide variety of metrics have been developed, differing in the physical image properties quantified, as well as the models and parameters of the human eye used.

For objective measures to be useful to image quality assessment, they must correlate with subjective impressions of image quality and ideally should be both standardised and independent of the imaging systems or processes involved. This last requirement is difficult to fulfil, meaning that many solutions are very application specific and difficult to translate for the comparison of different types of systems. In particular, if a metric has been developed for a digital imaging process which gives rise to one particular type of artefact, it can produce very different results from a process producing different kinds of artefacts [33]. As the visibility and severity of such artefacts are typically dependent upon scene content, this may be viewed as a form of *scene dependency*. The problem of this type of scene dependency is being addressed in current research [34] [35]. In general, metrics that are applicable to different types of systems tend to be more complex and computationally intensive than those that are application specific [36].

Subjective image evaluation approaches involve the collection and analysis of judgements of aspects of image quality by human observers. The field of psychophysics, a branch of psychology concerned with the quantification of perception, has its roots in the work of Weber and Fechner in the nineteenth century. Their research related discriminable differences in the perception of different sensations, to measured changes in physical properties (for example, the relationship between the perceived taste of 'saltiness' and the concentration of sodium chloride in solution). Psychometrics is an area of psychophysics that describes experiments designed to quantify relationships between the perceptions of stimuli to variations of more than one objectively measureable attribute; it is used extensively in subjective image quality assessments. Presenting sample stimuli to relatively large groups of observers may produce various measurement scales, correlated to aspects of image quality. Some psychometric experimental methods derive scales by the rating of attributes of single images, in experiments to determine suprathreshold magnitudes of image quality. Others, using paired comparison experiments, are based on Thurstone's Law of Comparative Judgement [37] (1927), and are more commonly used to evaluate small differences in image quality, useful for identifying thresholds of perceptibility of stimuli changes, or Just Noticeable Differences (JNDs) [16] [18].

There are some challenges in implementing subjective methods of quality assessment. The data analysis involved is often demanding and in practice, data collection can be time-consuming. Environmental conditions must be carefully controlled and large numbers of observers are often required to produce meaningful results. Factors such as the experience of the observer group in the judgement of images, and the level of fatigue of observers can have a significant effect on the results obtained. Therefore the observer group must be carefully selected and the length of time taken for observations must be controlled, which limits the numbers of images that may be evaluated. The results are also very dependent upon the type and range of sample stimuli and careful consideration must be given to image selection. This makes it difficult to provide comparisons between derived scales, unless they are calibrated in some way to a common standard. Keelan's *Handbook of Image Quality* [1] and further work on an extension of ISO standard 20462, the Softcopy Quality Ruler [38] [39], are examples of approaches which aim to provide solutions to this problem of calibration and comparison between scales resulting from different psychometric experiments.

The necessary time and complexity of psychophysical experiments mean that such methods are not always practical for the assessment of systems which have a short product cycle time; this is a primary motivation in the development of suitable objective image quality measures and metrics. Nevertheless, subjective methods remain an important aspect of image quality investigation; they are an important means of evaluating and benchmarking objective metrics.

2.4 Distortion, Fidelity and Quality

Various aspects of image quality may be quantified, and these are differentiated by three terms, *image distortion*, *image fidelity* and *image quality*. Although all are related to the evaluation of images and imaging systems, they are described somewhat inconsistently in the literature under the general umbrella of image quality [19] [40], but have distinct meanings.

Image distortion is assessed objectively and is concerned with the quantification of errors introduced by an imaging system or process, without reference to their visual impact. The simplest methods measure physical differences between digital images. The example in figure 2.3 illustrates a form of distortion measurement, where a compressed image has been subtracted from the original and the results scaled to produce a third image which is a difference map between the two images. The resulting image can be quantified to produce a measure of the magnitude of error incurred by the compression process. *Distortion metrics* such as root mean squared error (RMSE) [4]and Peak Signal-to-Noise Ratio (PSNR) [7], which evaluate the average error magnitude (per pixel in most cases), and the amount of error relative to the peak value of the signal, respectively, are two widely used examples.

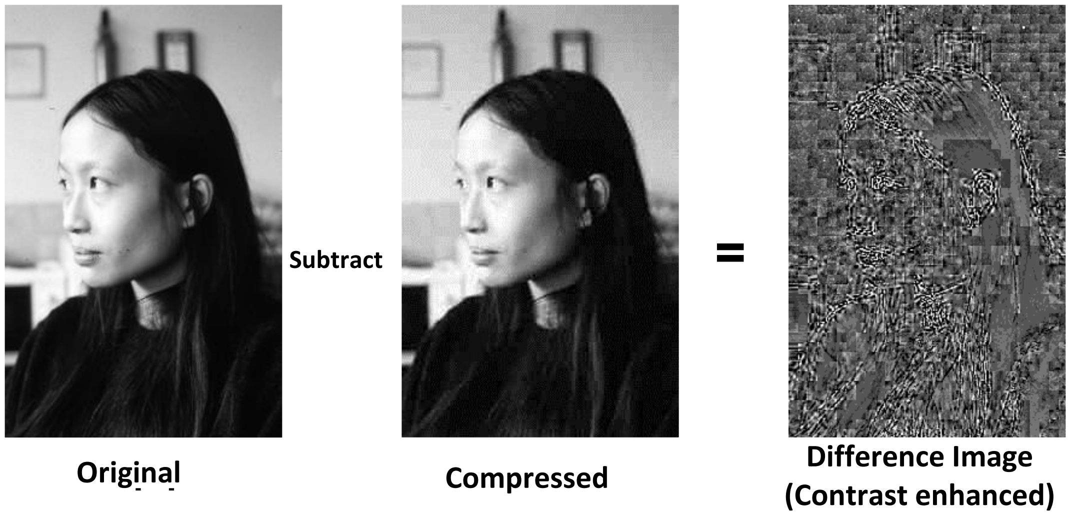


Figure 2.3: An example of simple distortion measurement (Image© S. Triantaphillidou)

Such measures are often employed to quantify the effects of imaging processes that introduce loss of information, such as image compression; however, their poor correlation with the subjective perception of quality [20] means that they have limited usefulness as visual quality measures. There are various reasons for a lack of correlation. Because simple distortion metrics are based on first order statistics, they take no account of the spatial structure of the image and therefore the location or visibility of the errors (as discussed later, certain spatial image characteristics may provide a masking effect for some types of artefacts, which more sophisticated objective approaches attempt to model). Additionally, without any model of the HVS incorporated, distortion measures cannot predict how problematic the errors will be, as their visual significance cannot be weighted. Indeed, the low level introduction of certain types of artefacts can represent an improvement in perceived quality for some types of scene content [33] [34], resulting in a negative correlation between such a distortion measure and image quality. Furthermore, certain global variations to images, such as a small value change in brightness [19](by the addition or subtraction of a small constant to every pixel), or a translation of all pixels by the same amount, will result in large distortion values with no perceived effect on the image. Nevertheless, distortion measures have a place in assessing the level of information loss introduced by some processes. Additionally, certain types of metrics rely upon *error pooling* as a final stage, after taking into account the properties of display and visual systems.

Silverstein and Farrell [41] define image fidelity as "the ability to discriminate between two images" and referring to "the ability of a process to render an image accurately without any visible distortion or information loss". Like distortion metrics, image fidelity measures employ image comparison of the same scene to evaluate differences; however they are concerned with identifying and quantifying threshold levels of detection of those differences and relating those to physical properties. As they provide a measure of the point at which distortion is perceived by the visual system, they effectively quantify the combined effects of the physical image properties, the display system, and aspects of visual processing. Because fidelity measures deal with thresholds, the differences between the compared images tend to be small (although a range of distortions will be evaluated). The differences may be as a result of a change in a single physical attribute, for example by contrast or colour modification, or can be the combined effects of multiple attributes and artefacts, as is typical of the distortion caused by lossy compression.

Fidelity may be assessed subjectively in paired comparison experiments, where observers provide a yes/no answer to the question of whether they can detect a difference between two stimuli, one distorted and the other a reference. Alternatively, observers may change a physical parameter on a single image to identify the point at which a difference is detectable: the Just Noticeable Difference (JND). Once the JND is established, a scale of magnitudes of fidelity loss may be calculated using multiples of the JND.

Objective measures attempt to predict the threshold point by combining physical image properties with models of the reproduction system and models of the HVS. The perceptual characteristics upon which these models are based are discussed in more detail in section 2.7. The visual models deal with lower order processing; typically they incorporate functions for luminance adaptation, contrast sensitivity and masking effects, where the effects of a signal can be masked by the presence of another signal, for example the presence of noise of a particular frequency can be masked by the presence of that frequency in the signal at the same spatial location [42]]. Pappas, Safranek and Chen [36]] term these types of measures: 'Perceptual Metrics'. These are classified as 'full reference' quality measures [24], in that a test image is compared to a reference image. Both images are passed through a number of different processing stages, which simulate the viewing and display conditions, followed by selected characteristics of the HVS. The result in each case is a theoretical 'visual image' [19]. The final stage is error pooling, the quantification of the differences between the two images, producing either a single figure which equates to quality, or a map of distortions (e.g. colour differences for each pixel).

It seems logical to assume that when an observer is comparing two images with relatively small differences between them, then they will tend to scrutinise the images with care; this is particularly true in the case of subjective fidelity studies using expert observers. From the earlier discussion regarding current theories of visual perception, this implies that hierarchical visual processing in fidelity judgements is likely to be operating from the bottom-up, (i.e. Hochtein and Ahissar's 'vision with scrutiny' [30]) where consideration of lower order attributes, takes precedence. This would explain why objective fidelity metrics, which use low level image attributes and models of lower level visual processing appear to correlate relatively well with subjective fidelity. However, as already discussed, suprathreshold quality judgement involves more high level visuo-cognitive processing. As Ford [19]] points out, without additional cognitive inputs, appropriate scaling to predict quality is difficult with perceptual metrics.

Image quality in the 'true' sense is concerned both with thresholds of perceptibility (to establish the just noticeable difference) and with suprathreshold magnitudes, [19] (which may usefully be calibrated in a scale of JND values). This involves the overall perception of the 'goodness' of the image, and combines the effects of all aspects of visual and cognitive processing with the further factors influencing observer quality preferences (such as quality consciousness and imaging context). Figure 2.4 illustrates the different factors involved in distortion, fidelity and quality measurements, and the points at which they are carried out.

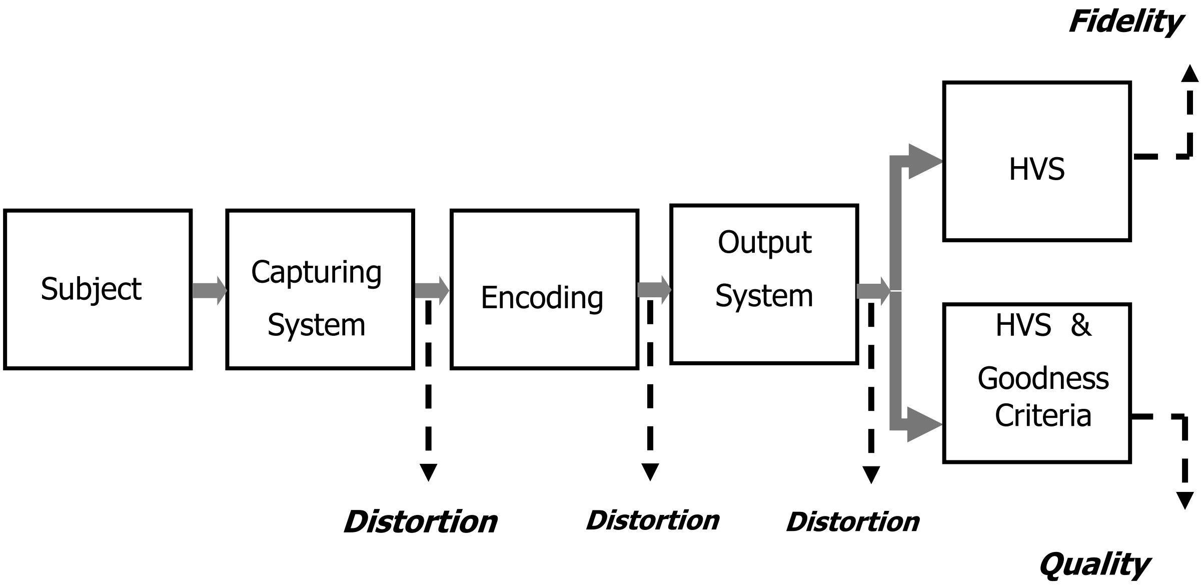


Figure 2.4: Factors involved in the measurement of image distortion, image fidelity and image quality and points where their measurements are performed (from Triantaphillidou [40]

Distinguishing between fidelity and quality measures can be confusing, especially as similar methodology may be used, particularly in subjective assessment. The fundamental difference is in the task for the observer; in fidelity, it is one of discrimination (establishing *perceptibility*), whereas in quality the observer is also asked for a preference, or a judgement (for example, establishing *acceptability*). Because fidelity measurements assess the degree of visible distortion, they may also correlate negatively with perceived quality [41], as for distortion measures, for the reasons described previously. Part of this research, detailed in chapter 5, explores subjective assessment of perceptibility and acceptability within the same experimental study, and the relationship between them, particularly with reference to scene content.

Quality may be assessed subjectively, using either comparison between images to indicate preferences (in paired comparisons, or by ranking groups of images), or on single images (so-called *no-reference* methods). Without comparison, Ford notes [19] that the impact of the observer's quality consciousness becomes more dominant. Psychometric evaluations produce various different types of quality scales, as detailed in the later section on subjective evaluation methods. Image quality may be assessed objectively using various performance measures, to evaluate systems, or by modelling *image quality attributes* (described below) either singly or in combination in Image Quality Metrics (IQMs) [[19] [22] [40] [14]. A fundamental assumption in the development of many IQMs is that image quality is multi-dimensional, but that the attribute dimensions can be individually scaled and then combined to obtain an overall figure of merit. As described in section 2.6, five main attributes have been traditionally considered to be the main contributors to image quality in analogue imaging systems, and derived metrics often use combinations of these attributes with HVS characteristics. Types of metrics are introduced in the later section on objective measurement.

2.5 Physical and Perceptual Image Quality Attributes

Keelan, in providing a working definition of image quality [16], classifies a range of image quality attributes into the following different types, according to their nature and amenability to objective description:

Personal attributes, as described at the beginning of this chapter, e.g. to do with capturing the 'essence' of the subject, or providing a flattering image of someone.

Aesthetic attributes, such as lighting and composition.

Artefactual attributes, which are not always evident in an image, but are defects introduced by the imaging system, and commonly (but not always) lead to degradation in image quality. Examples include blur, noise, and digital artefacts.

Preferential attributes, these are always evident in an image and can be identified by a preferred point, dependent upon observer, scene content and imaging context. Examples include tone and colour reproduction.

Of these, the first two types of attributes are highly subjective in nature, and very variable from image to image, therefore are not very amenable to objective quantification. The effects of these types of attributes are limited by careful definition of the observer, purposeful scene selection and quality evaluation across many different images. The latter two types of attributes are the ones of interest in image quality investigation, having two important characteristics: they are both amenable to objective description, as described below, and they are strongly influenced by the properties of the imaging system.

In analogue imaging, five basic physical attributes have been traditionally considered to influence image appearance [19], [22] [40]: tone, colour, resolution, sharpness and noise. These are sometimes referred to as the main *dimensions* of image quality. There are numerous well-defined physical measures relating to each attribute, as defined in table 2.1, and these are used individually or in combination in image quality metrics to characterise the physical properties of the image.

Where physical attributes are considered to be objective measures, they are related to *perceptual attributes,* which are often termed 'nesses' [18] [32].

These are the terms typically used by observers to describe the way in which the image characteristics are perceived; some of those relating to the physical attributes described above are illustrated in table 2.2.

IMAGE ATTRIBUTE	**MEASURES**
Tone	Characteristic curve, density differences, transfer function and OECF, contrast, gamma, histogram, dynamic range
Color	Spectral power distribution, CIE tristimulus values, colour appearance values, CIE colour differences
Resolution	Resolving power, imaging cell, limiting resolution
Sharpness	Acutance, ESF, PSF, LSF, MTF
Noise	Granularity, noise power spectrum, autocorrelation function, total variance (σ^2_{TOTAL})
Image content and efficiency	Information capacity, entropy, detective quantum efficiency.

Table 2.1:Physical Image attributes and objective imaging performance measures (from Triantaphillidou [40])

PHYSICAL ATTRIBUTE	**VISUAL DESCRIPTION**	**RELATED PERCEPTUAL ATTRIBUTES**
Tone	Macroscopic contrast; reproduction of intensity	Brightness, Lightness, Contrast
Colour	The reproduction of/differences in lightness, chrominance and saturation	Lightness, Brightness, Chrominance, Saturation, Colourfulness, Hue
Resolution	The ability to perceive/reproduce fine detail	Fineness (detail)
Sharpness	The reproduction of edges	Sharpness, Fineness
Noise	The presence of random and non-random spurious information	Noisiness

Table 2.2: Physical attributes, their descriptions and relationships with perceptual attributes (adapted from Ford [14]).

Note that sharpness is considered both a physical attribute, quantified by the objective measurement of the frequency content in a reproduced edge, and a perceptual attribute, describing the subjective impression formed from the reproduction of an edge. In some cases there is a direct relationship between a single *ness* and a physical attribute, but in others the perceptual attribute is either dependent on more than one physical attribute, or the physical attribute may be partially described by several different *nesses*. For example, the subjective impression of sharpness is to do with both edge frequency content and edge contrast. Equally, five different perceptual attributes may be used to describe aspects of colour reproduction.

The perceptual attributes described in table 2.2 are fairly well defined, but as described in section 2.2, recent research identifies further more complex perceptual attributes that have an influence on image quality. Examples include *usefulness* and *naturalness*, which are dependent not only on the physical attributes of the image, and psychophysical characteristics of the visual system, but also on higher level visuo-cognitive processing, which references the user's quality criteria in relation to the imaging application and their expectations and preferences based upon their internal references for the particular type of image or scene.

2.6 Objective measures of physical attributes

The following sections describe some of the well-defined measures of imaging performance relating to the five basic image quality attributes described above.

2.6.1 Tone Reproduction

Tone reproduction is concerned with the relationship between the scene luminances and their reproduction in images. Although a subset of colour reproduction, it is considered separately because of its fundamental importance to image quality, as a result of the amount of information carried by the achromatic visual channel [40]. Objective tone reproduction for an imaging system or device is described by the relationship of input to output intensities in one or several transfer functions. Subjective tone reproduction is dependent on this objective relationship from input to output, but also takes into consideration viewing conditions (luminance level, surround luminance, flare, etc), which influence psychophysical characteristics (the non-linear perception of luminance, *amplitude non-linearity*) and therefore the perception of tones and tonal differences.

The transfer functions of imaging materials or devices are evaluated using sensitometric measurements, by inputting a scale of known intensity values of usually equal visual steps and measuring the output. The transfer function for silver halide materials, known as the photographic *characteristic curve*, or the H and D curve, after F.Hurter and V.C. Driffield [43] [44], is a sigmoid shaped curve obtained by plotting output densities against log relative exposures.

An important measure, which may be derived from device transfer functions is termed gamma (γ), which is correlated with image contrast. In the photographic characteristic curve, it is known as *sensitometric contrast* [45] and is defined as the gradient evaluated from any two points lying on the straight line portion of the curve:

$$\gamma = \frac{D_2 - D_1}{\log H_2 - \log H_1} \tag{2.1}$$

Transfer functions for digital systems are obtained in a similar fashion. They are more typically plotted on linear-linear axes and in many cases exhibit a non-linear relationship between input and output, often completely specified by some form of a power function. Interestingly, these transfer functions in some cases are not intrinsic to the devices but are applied by some form of internal correction to conform to transfer functions which have been standardised for image interchange and video transmission [46] [47] [48]. A well-known example of a power transfer function (which is native to the device) is that of a Cathode Ray Tube (CRT) display system, which has a non-linear relationship between input voltages and the produced luminances on the display surface, often described by the *gamma model* [47]below:

$$L = o + gV^{\gamma D} \tag{2.2}$$

Where L = normalised luminance
o = system offset (from display brightness setting)
g = system gain (from display contrast setting)
V is the normalised voltage (or normalised pixel values)
γ_D is the non-linearity of the contrast in the displayed image

On a correctly adjusted display, the offset and gain are set to values of 0 and 1. Many CRTs, when correctly set up have an exponent close to a value of 2.5.

Typically, in a digital imaging chain, the gamma of successive components is different, and *gamma correction* is applied, to compensate for the various individual gamma values, modifying image pixel values to obtain the required overall tone reproduction. The system gamma is calculated by cascading individual gamma values [47]:

$$\gamma_{\mathrm{sys}} = \gamma_{\mathrm{O}} \mathrm{x}\, \gamma_{\mathrm{I}}\, \mathrm{x}\, \gamma_{\mathrm{D}}\, \mathrm{x}\, \gamma_{\mathrm{C}} \tag{2.3}$$

Where the subscripts O,I,D and C represent output, input, display and correction respectively.

The expression above implies that gamma correction will be applied only once in the imaging chain, but in reality, it may be applied individually at several different stages. The importance of the CRT display gamma models is that many devices (and colour encodings) incorporate gamma correction to ensure that an image will be displayed on a typical CRT correctly. The term gamma correction originates from the television industry and is defined by Poynton [48] as: “The process by which a quantity proportional to intensity, such as CIE luminance...is transformed into a signal by a power function with an exponent in the range 0.4 to 0.5”. This describes the conversion of a linear signal to one that will be correctly displayed on a CRT with a non-linear response curve.

This form of gamma correction may be generalised by the expression:

$$V' = V^{1/\gamma} \tag{2.4}$$

Where V' is the gamma corrected signal, V is the original signal, and γ is the gamma value from the power function modelling the response of the intended display device.

As mentioned earlier, because of the initial prevalence of CRT display technology in digital imaging chains, many other devices are gamma corrected to a theoretical CRT display. Liquid Crystal Displays (LCD), for example, have various different native transfer functions depending on their operating modes, but are essentially linear devices. Their signals are often gamma corrected to ensure that they are suitable for standardised image interchange, often resulting in a gamma corrected signal, which mimics that of a CRT [46].

Digital Still Cameras (DSCs) and scanners have sensors with approximately linear luminance responses. However their transfer functions are usually described using the Opto-Electronic Conversion Function (OECF), which is a system transfer function produced from the combination of sensor, firmware and software. Gamma correction is typically applied as part of the signal-processing pipeline to ensure that the images are correctly displayed [47]. If RAW data are to be output, this happens when the image is previewed during RAW conversion (usually using one of a selection of standard RGB colour spaces). If a fully rendered file is produced, in the majority of DSCs it will be rendered using a standard RGB colour image encoding, such as sRGB or Adobe RGB 1998. Both colour encodings have transfer functions that approximate power functions. That of sRGB is not an exact power function; it has a linear bottom part and despite the power in the encoding being 2.4, it has a nominal gamma value of 0.45 (which corresponds to a gamma correction of 1/2.2,) because it assumes that the encoded image will be output to a standard CRT with a gamma of 2.5. Adobe RGB 98 is a pure power function with an exponent of 0.45.

While the goal of objective tone reproduction is an overall gamma of unity (where all component gammas compensate for each other), it is subjective tone reproduction that is important in terms of image quality. Ford [14] and Triantaphillidou [46] note that subjective tone reproduction is dependent on scene brightness relative to white, and that perceived contrast changes as a result of viewing conditions and surround. Fundamentally this means that perceived contrast cannot be defined without consideration of environmental and viewing conditions.

2.6.2 Colour Reproduction

2.6.2.1 Colour Specification

It is quite common in everyday language to ascribe certain colours to objects or light sources: for example 'grass is green'. However, this is an erroneous description, for colour is not an object attribute but rather an attribute of visual sensation; it cannot exist without an observer.

Whatever form the 'observer' takes (it need not be human) encompasses a means of both detecting and of interpreting electromagnetic radiation. In the human observer, the retinal photoreceptors detect the physical stimulus (electromagnetic radiation in the visible region of the spectrum) and then the neural connections in the visual pathway, and the cognitive system, process and interpret the signals produced by the stimulus [49]. The observer is the final component of what is termed the *triangle of colour* [50]. The first component is a source of visible electromagnetic energy, which will have its own spectral signature. The second component is an object, the chemical and physical properties of which modulate the energy from the source. To fully specify and describe colour, all three components of the triangle of colour require quantification.

The complexity of colour is reflected in the methods and models of colour description, which are numerous and multi-dimensional in nature. There is not a single perceptual attribute for example, from table 2.2, which can alone describe the appearance of a colour. It is generally accepted that three main perceptual attributes are required to fully express the observer's response to a colour [51]. Briefly, the *hue* of an area describes its relationship or apparent similarity to the perceived colours, red, green, yellow or blue (or combinations of two of them); its *brightness* is the degree to which an area appears to emit more or less light; and its *colourfulness* is the amount by which an area appears to exhibit more or less of its hue [52]. *Lightness* and *chroma* are correlates of brightness and colourfulness respectively where the area is judged in proportion to that of a similar area, which is white or highly transmitting. *Saturation* is another correlate of colourfulness, judged in proportion to the area's brightness.

2.6.2.2 Colorimetry

As shown in table 2.1, various approaches may be used in the objective quantification of colour. Spectral colour definition is a purely physical approach, for example quantifying the relative amounts of wavelengths contained within a stimulus (the *spectral power distribution*), the spectral absorption and reflection characteristics of surfaces or the spectral absorption properties of a photoreceptor. Colorimetry (the 'measurement of colour'), by its strictest definition, is a means of predicting a colour match between two light sources of different spectral power distributions under specified conditions for a standard 'average' observer.

The Commission Internationale D'Eclairage (CIE), in 1931, recommended a colour measurement system for absolute specification of such colour matches, which formed the basis of modern colorimetry. By specifying colour matching functions for a standard observer (for a 2°, and later a 10° visual field) and providing SPDs for a variety of standard illuminants, CIE colorimetry allows absolute specification of colours that is independent of device or system [19].

CIE colorimetry is a trichromatic system, and is based on the assumption that colour vision at *photopic* levels operates in the first instance by trichromatic matching of colours by the addition of the responses of the three types of cone receptors in the retina. The cone responses span three broad bands of wavelengths, with peaks at 580nm, 540nm and 440nm and are often denoted L, M and S (long, medium and short wavelengths, respectively), or ρ,γ and β. There is significant overlap between the responses of the different types of cones in certain parts of the spectrum.

CIE colorimetry aims to specify and describe all colours in terms of three variables, termed *tristimulus values*. These values are amounts of three additive primary stimuli that combine to match the colour being specified [52]. According to Grassman's laws of additive colour mixture, a colour match can be specified by [50]:

$$C \equiv \mathrm{R}(\boldsymbol{R}) + \mathrm{G}(\boldsymbol{G}) + \mathrm{B}(\boldsymbol{B}) \qquad (2.5)$$

Where R(***R)*** is R units of the ***R*** primary, etc. The CIE RGB colour matching functions, *r(λ), g(λ)* and *b(λ),* shown in figure 2.5, were obtained experimentally for the standard (2°) observer matching three primary monochromatic stimuli: R(700nm), G(546.1nm) and B(435.8nm). The curves in the figure indicate the relative amounts of the three illuminants required at any wavelength, to invoke a visual sensation equal to unit amounts of power a monochromatic source of the same wavelength.

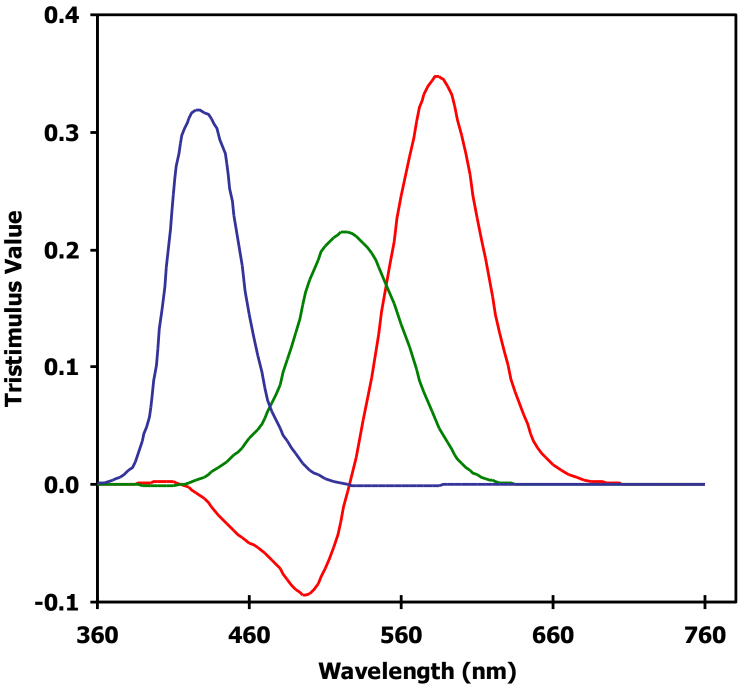


Figure 2.5: CIE 1931 RGB colour matching functions, from Triantaphilldou [38]

To calculate the tristimulus values for any stimulus with a spectral power distribution ϕ(λ), the generalised equations are [50]:

$$R = \int_{\lambda} \phi\,(\lambda)\bar{r}(\lambda)d\lambda \qquad (2.6a)$$

$$G = \int_{\lambda} \phi\,(\lambda)\bar{g}(\lambda)d\lambda \qquad (2.6b)$$

$$B = \int_{\lambda} \phi\,(\lambda)\bar{b}(\lambda)d\lambda \qquad (2.6c)$$

As described by Fairchild [50], the, $\phi(\lambda)$ term is dependent on the stimulus being measured. For a self-luminous source, it is usually the relative spectral power distribution, whereas for a reflective material, it is the product of the spectra of the source and the object.

As figure 2.5 illustrates, some areas of the colour matching functions are negative. As Hunt [52] explains, this is because there is some overlap in the response of the three types of cone receptors, which means that it is never possible to stimulate the γ cones alone. To eliminate the negative values in the colour matching functions the CIE converted the ***RGB*** primaries using linear transformations to a further set of imaginary primaries, ***XYZ***, which could match all stimuli without negative values. These primaries were selected so that one of them, the ***Y*** primary, would match *V(λ),* the CIE photopic luminous efficiency function, by choosing the other two primaries to produce no luminance response [50]. The ***XYZ*** primaries have colour matching functions *x(λ), y(λ* and *z(λ)* respectively and general equations for calculating tristimulus values are:

$$X = k\int_{\lambda} \phi(\lambda)\bar{x}\,(\lambda)d\lambda \qquad (2.7a)$$

$$Y = k\int_{\lambda} \phi(\lambda)\bar{y}\,(\lambda)d\lambda \qquad (2.7b)$$

$$Z = k\int_{\lambda} \phi(\lambda)\bar{z}\,(\lambda)d\lambda \qquad (2.7c)$$

Where *k* is a normalising constant. *k* is defined as 683 lumens W^{-1} for *absolute colorimetry,* which is used mainly for self-luminous stimuli. For *relative colorimetry*:

$$k = \frac{100}{\int_{\lambda} S_{\lambda}\bar{y}(\lambda)d\lambda} \tag{2.8}$$

Where *S(λ)* is the relative spectral power distribution of the illuminant (a relative SPD is one that has been normalised). This normalisation constant yields tristimulus values in the range 0-100 for most materials; if relative colorimetry is used for a light source, the Y tristimulus value will always be 100.

The colour matching functions described above were determined from experiments using a 2° visual field. In 1964 the CIE published data for the *supplementary standard colorimetric observer* with a 10° visual field.

2.6.2.3 CIE Uniform Colour spaces

Tristimulus values provide a model of colour measurement directly associated with the trichromatic model of human visual processing. However, when attempting to visualise the colour of a stimulus, they are not particularly intuitive descriptors, as they do not establish the relationship with perceptual attributes relating to colour. Tristimulus values are commonly represented graphically using a two-dimensional *chromaticity diagram*. The conversion from XYZ tristimulus values to *x,y*, or *u'v'* chromaticity co-ordinates is achieved by normalisation of the values to remove luminance information. They cannot therefore represent many aspects of the colour appearance of stimuli [50].

The *CIELAB* and *CIELUV* colour spaces, recommended by the CIE in 1976, were developed with the aim of providing visually uniform spaces for the measurement of colour differences, as this cannot be achieved using tristimulus values or chromaticity co-ordinates for the reasons stated above. The CIE colour spaces are three-dimensional, expressed in dimensions which approximately correlate with the perceived, hue, chroma and lightness of a stimulus[37,38]. Correlation and visual uniformity are achieved by incorporating features that account for non-linear visual responses and chromatic adaptation. Both spaces have a uniform lightness scale *L**, and two chromaticity co-ordinates: *a*, b** in the CIELAB space approximately represent redness-greenness and yellowness-blueness; *u*, v** are the equivalent co-ordinates in the CIELUV space.

The transformations from XYZ tristimulus values to the co-ordinates of the two colour spaces are not dissimilar. CIELAB values are obtained as follows [51]:

$$L^* = 116f\left(\frac{Y}{Y_N}\right) - 16 \tag{2.9a}$$

$$a^* = 500\left[f\left(\frac{X}{X_N}\right) - f\left(\frac{Y}{Y_N}\right)\right] \tag{2.9b}$$

$$b^* = 200\left[f\left(\frac{Y}{Y_N}\right) - f\left(\frac{Z}{Z_N}\right)\right] \tag{2.9c}$$

Where X_n, Y_n and Z_n are tristimulus values of the reference white, and *f(x)* is defined differently depending on the normalised values as follows:

For normalised values (i.e. x= $\frac{X}{X_N}$, or $\frac{Y}{Y_N}$ or $\frac{Z}{Z_N}$) >0.008856:

$$f(x) = (x)^{\frac{1}{3}} \tag{2.10}$$

For normalised values ≤ 0.008856:

$$f(x) = 7.7871(x) + \frac{16}{116} \tag{2.11}$$

The CIELAB space is illustrated in figure 2.6. Perceived chroma, (C^*_{ab})and hue (h_{ab}) are cylindrical co-ordinates in the same three-dimensional space and are also depicted. They may be derived from a* and b*:

$$C^*{}_{ab} = (a^2 + b^2)^{1/2} \tag{2.12}$$

$$h_{ab} = tan^{-1}\left(\frac{b}{a}\right) \qquad (2.13)$$

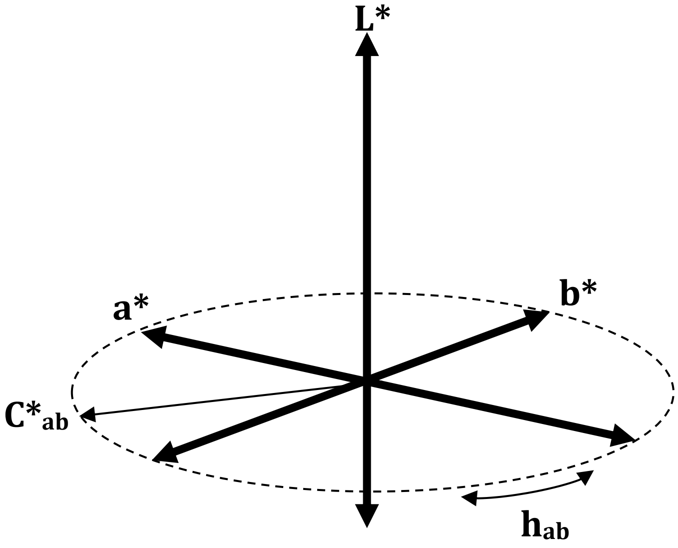


Figure 2.6: Three-dimensional representation of the CIELAB uniform colour space. Cylindrical coordinates, C^*_{ab} and h_{ab} are also illustrated. From Triantaphillidou [38]

2.6.2.4 Colour Difference Formulae

Colour-difference values are important in objective image quality assessment, performing various functions depending upon the approach and methods being implemented. As a measure of differences between images, they can be used for the quantification of errors in a full reference distortion metric. Evaluating pixel-by-pixel colour differences produces a colour difference map, which is useful in highlighting image content that is particularly susceptible to distortion. Colour differences may also be pooled to produce a single value metric [50]. Furthermore, perceptibility and acceptability thresholds may be investigated in terms of colour differences. Because of the perceptual uniformity of the CIE colour spaces, colour differences may be evaluated using the Euclidean distance between the co-ordinates of two stimuli. In the CIELAB space, this is expressed as [50]:

$$f(x) = 7.7871(x) + \frac{16}{116} \qquad (2.14)$$

$$\Delta E^*{}_{ab} = [(\Delta L^*)^2 + (\Delta a^*)^2 + (\Delta b^*)^2]^{\frac{1}{2}} \qquad (2.15)$$

Colour differences may also be expressed in terms of chroma and hue:

$$\Delta E^*{}_{ab} = [(\Delta L^*)^2 + (\Delta C^*{}_{ab})^2 + (\Delta H^*{}_{ab})^2]^{\frac{1}{2}} \qquad (2.16)$$

$$\Delta H^*{}_{ab} = [(\Delta E^*{}_{ab})^2 - (\Delta L^*)^2 - (\Delta C^*{}_{ab})^2]^{\frac{1}{2}} \qquad (2.17)$$

As noted by Fairchild [50], CIELAB colour differences are not perceptually uniform throughout the colour space. Because of this, a number of modifications have been made, to improve colour-difference measurement uniformity. These modifications have resulted in CIE recommendations for two newer models for colour-difference measurement, CIE94 and CIE-DE2000.

The CIE94 model, published in 1995, is an adaptation of the equations used to calculate CIELAB colour differences and is based on a set of reference conditions defining illumination, observer, background and sample characteristics [50]. Parametric factors are used to weight hue, chroma and lightness components according to the position of the stimulus in the CIELAB space, to adapt for conditions different to the reference. $\Delta C^*{}_{ab}$ and $\Delta H^*{}_{ab}$ decrease with increasing $C^*{}_{ab}$ CITATION Hun98 \l 1033 [52].

The CIE-DE2000, which was published in 2001 and has been proposed for adoption by the CIE, incorporates some aspects of CIE94 and a previous rather complex model known as the CMC module (developed in the mid-1980's, CMC also weighted the different components according to the position of the stimulus). The CIE-DE2000 colour difference, denoted ΔE_{00}, incorporates a weighting function which is dependent on hue [49] and a term that is dependent on the hue and chroma difference product. It also rescales the a* axis prior to computation of hue and chroma. Again, a rather complex formula, Sharma [49] notes that there are some concerns about CIE-DE2000. Although well-behaved in most regions of CIELAB, some distortions are evident particularly in the blue regions of the colour space. There are additional concerns about the validity of the scaling functions under some conditions, as they are based upon different (colour difference) datasets. A more general concern about all colour difference formulae is that the visual datasets that they are based on involve comparisons of stimuli on a fixed and uniform background [49,53] [54]. This raises questions over their applicability in imaging applications where colours are surrounded by other, often-similar colours. Further, in image quality studies, it is desirable to be able to evaluate perceived differences between colours in complex scenes.

2.6.2.5 Colour appearance modelling

Colour appearance modelling aims to specify, according to Fairchild [50]: ‘the colour appearance of stimuli under a wide variety of viewing conditions’. As well as the influence of viewing conditions, a number of visual phenomena and environmental conditions influence colour appearance: chromatic adaptation, light adaptation, luminance level, background colour and surround colour. Of these, chromatic adaptation is the most important.

Chromatic adaptation is a mechanism of *colour constancy* (one of a number of visual constancies that improve the cognitive processing of visual information and the recognition and interpretation of scenes). It enables the HVS to maintain the perceived colour of objects under illuminants with different characteristics and white points [51]. It is analogous with white balance in digital cameras. Chromatic adaptation is achieved via transforms (CATs) which compute the corresponding colours under a reference illuminant for a stimulus defined under a test illuminant.

Non-Uniform(NU)/ Uniform(U) Colorimetric Measures	**Input Data CIECAM02**	**Output Data CIECAM02 (Perceptual Correlates)**
NU: Luminance factor *L/Ln* U: CIE 1976 lightness L*	*X,Y,Z*: Relative tristimulus values of color stimulus in the *source* conditions	*J*: Lightness
NU: Luminance *L* U: None	L_A: Luminance of the adapting field (cd/m^2)	*Q*: Brightness
NU : None U: CIE 1976 C^*_{uv} or C^*_{ab}	X_w,Y_w,Z_w: Relative tristimulus values of white	*C*: Chroma
NU: Excitation Purity p_e U: CIE 1976 Saturation s_{uv}	$Y_{b:}$ Relative luminance of the background	*s*: Saturation
NU: None U: None	*c*: Impact of surround	*M*: Colourfulness
NU: Dominant wavelength λ_d U:CIE 1976 hue-angle h_{uv} or h_{ab}	N_c: Chromatic induction factor	*h*: hue angle
	F_{LL}: Lightness contrast factor	*H*: hue composition
	F : Degree of adaptation factor	a_M,b_M: Cartesian colour coordinates derived from colourfulness and hue
		a_C,b_C: Cartesian colour coordinates derived from chroma and hue
		a_s,b_s*:* Cartesian colour coordinates derived from saturation and hue

Table 2.3: Relationship between colorimetric measures and the input and output data from CIECAM02 (based on Hunt [52], and Triantaphillidou [51])

There are a number of different Colour Appearance Models (CAMs), which share some general concepts, summarised by Fairchild [50]: Stimulus and viewing conditions are specified in terms of CIE XYZ tristimulus values, which are converted to cone responses via (usually) linear transformation, to allow more accurate modelling of the physiological aspects of the HVS.

Characteristics of the viewing environment are also considered; including tristimulus values of the adapting illuminant and in some cases various other data such as absolute luminance level and details of the background and surround. A chromatic adaptation transform is then applied, additionally incorporating data about the adapting stimulus and viewing conditions. After this stage, results are combined into higher-level signals modelling the opponent-colours theory of the HVS as well as various non-linearities and threshold effects. The final signals are combined in various ways to produce predictors of perceptual attributes.

Early colour appearance models, developed by Hunt (1982,1985) and Nayatani (1986), lead to the first recommendation from the CIE, CIECAM97s. More recently, a simpler model has been recommended, CIECAM02. A summary of colorimetric measures, including those in CIECAM02 and their perceptual correlates is given in table 2.3. The table illustrates how the physical attributes (measured by the input data in the second column) affect the perceptual correlates. Basic colorimetric measures alone are not able to predict how stimuli will be perceived in complex scenes and images, particularly in terms of the interactions between attributes – as indicated by the lack of measures in the last few rows of the table. However, they can be adapted and combined with other variables to produce an overall model of colour appearance.

2.6.3 Resolution

Resolution is concerned with the reproduction of detail within an image. Like sharpness, it describes micro-image spatial properties, but where sharpness is concerned with edges alone, which are sudden transitions in intensities, resolution describes the finest level of detail that may be captured or represented within the image [55]. Rather confusingly, the term is used to describe many different aspects of imaging systems, particularly digital imaging systems.

In capture devices, resolution often describes the number of (imaging) pixels on a sensor, with the aim of its use as a comparative figure of merit for consumers. This is more correctly termed *digital* resolution [47] or *pixel* resolution. Such a definition is relatively meaningless without reference to the sensor size; together the two measures determine the *addressable resolution* which is the number of individual image elements per unit distance for a device or image, often described by dots per inch (dpi) or pixels per inch (ppi). The addressable resolution is also a descriptor for output devices, although the final subjective impression of resolved detail (and sharpness) will also depend upon viewing distance and viewing conditions.

The 'true' resolution of an image or imaging system, the *spatial* resolution is fundamentally dependent upon the size of the basic imaging element; *the point spread function* (PSF) the shape of the image of a point of light, which is influenced by all stages of the imaging process [40]. These include the imaging optics; anti-aliasing filters; optical aberrations; sensor characteristics; optical spreading within a photographic emulsion, termed *turbidity*; optical spreading from micro-lenses in a digital sensor; interpolation processes, as averaging will always blur images and reduce resolution; and environmental conditions at image capture. The PSF may be viewed as a 'blurring function' imposed by the imaging system. Its size and shape determines how closely spaced two distinguishable image points can be. As noted by Burns [56] two factors determine the resolution of a digital system: the PSF and the sampling frequency (in samples per unit distance, analogous to the addressable resolution described above), which defines the upper resolution limit that may be achieved.

A traditional measure of resolution in photographic systems is *resolving power*, where an image of a test target containing blocks of equally spaced bars of increasing frequency (Figure 2.7) is visually inspected to identify the highest number of lines per unit distance that can be clearly distinguished by an observer. This is a form of assessment of *limiting resolution,* determined as the point at which finely spaced features are no longer detectable.

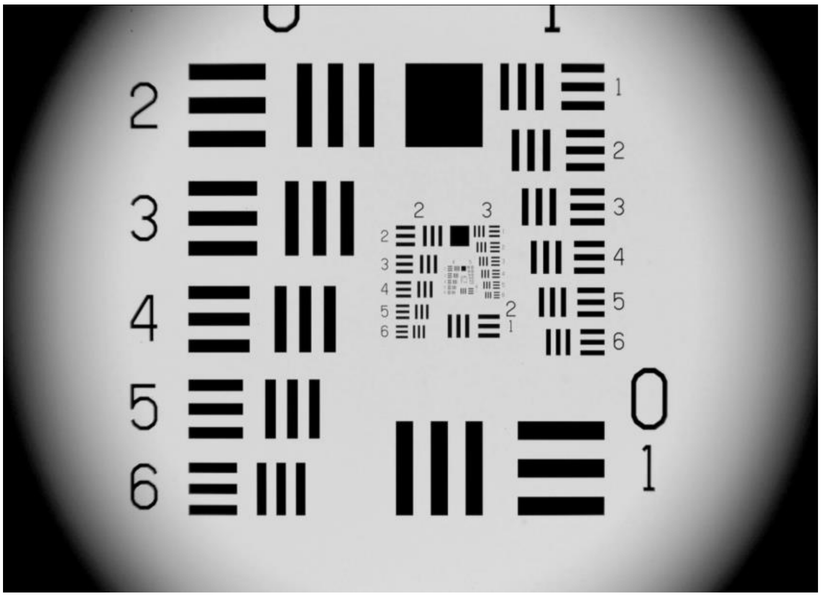


Figure 2.7: The USAF 1951 Lens Test Target

While simple to implement and understand, resolving power has a number of disadvantages, in that it ultimately relies on subjective evaluation, the results are dependent on the contrast of the target, and it only identifies a threshold criterion [56]. While useful and widely used therefore as an imaging performance measure, resolving power is less suitable as an indicator of image quality. Because resolution and sharpness are contrast dependent, a more complete measure of both can be obtained by investigating contrast or *modulation* as a function of spatial frequency, in the *modulation transfer function* (MTF), or the *spatial frequency response* (SFR), both of which are described in section 2.6.4.

In considering images compressed using lossy algorithms such as the JPEG baseline algorithm, or the JPEG 2000 algorithm, the issue of resolution is complicated by the fact that colour images are often pre-processed to convert from an RGB colour encoding to a luminance-chrominance colour space, at which point the chrominance channels are sub-sampled [33]. As Ford points out [19], both colour space and sub-sampling method may be unknown. When the image is decompressed, the sub-sampled channels must be interpolated to provide missing values. Although the final number of samples will be unaffected, artefacts may be introduced by this process and spatial detail lost, which then leads to questions about the accuracy of resolution measures, in this context, and particularly their correlation with or influence upon image quality. Furthermore, lossy algorithms such as JPEG are notoriously non-linear, and as system linearity is one of the fundamental assumptions in the evaluation of MTF, this complicates the measurement and interpretation of MTF curves in assessment of such images [55].

The subjective impression of resolution within an image is highly influenced by other image quality attributes, most notably image sharpness and contrast. Although resolution and sharpness correlate physically [19] (i.e. edge sharpness is highly dependent on how well the edge is resolved) it has been found that an increase in contrast can increase the perceived sharpness in a low-resolution image, relative to a lower contrast higher resolution image [19].

2.6.4 Sharpness

Image sharpness is concerned with the micro-image reproduction of edges. Image edges may be defined in terms of two variables, the edge gradient, which defines the spatial extent of the edge and the magnitude, which defines the contrast of the edge. A traditional photographic measure of sharpness is *acutance,* which is evaluated from the mean squared gradient of an edge. However it can be subject to errors and only partially correlates with visual sharpness [43]. Because edge reproduction is a function of both the PSF and micro-image contrast, the MTF, which evaluates the contrast with respect to frequency content of the edge, is a more complete measure.

When considering a sinusoidal intensity pattern of fixed spatial frequency, the modulation *M*, which may be thought of as a measure of contrast, is defined as:

$$M = \frac{I_{max} - I_{min}}{I_{max} + I_{min}} \tag{2.18}$$

When a sinusoidal pattern is input into a linear spatially invariant imaging system, the result will be an output sinusoidal pattern of the same frequency, but with a reduction in modulation (equivalent to a change in contrast, see Figure 2.8); the modulation reduction increases with increasing spatial frequency. The change in modulation for a given angular frequency ω is known as the *modulation transfer factor*:

$$M(\omega) = \frac{M_{out}(\omega)}{M_{in}(\omega)} \tag{2.19}$$

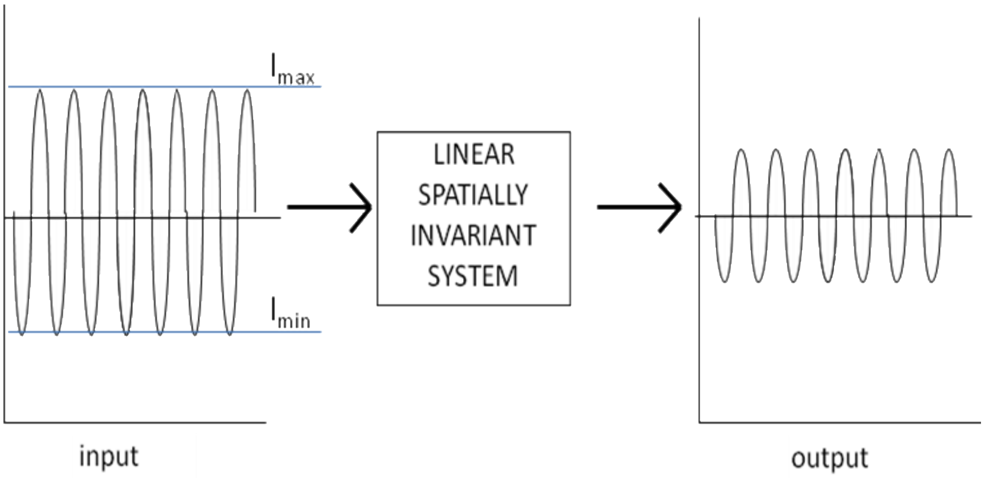
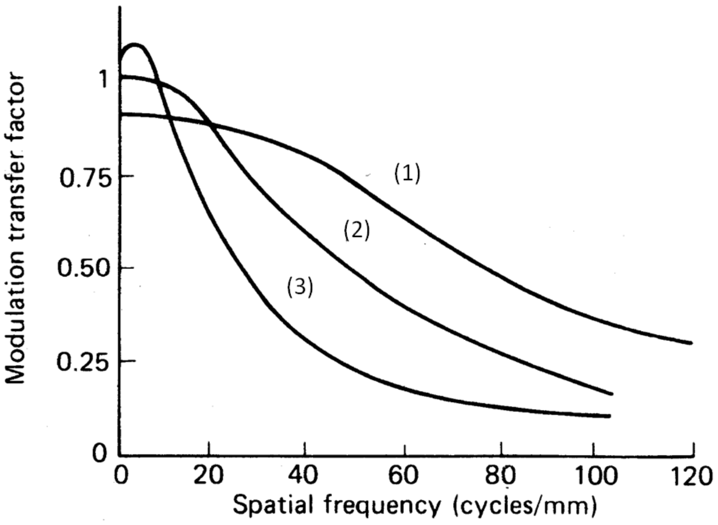


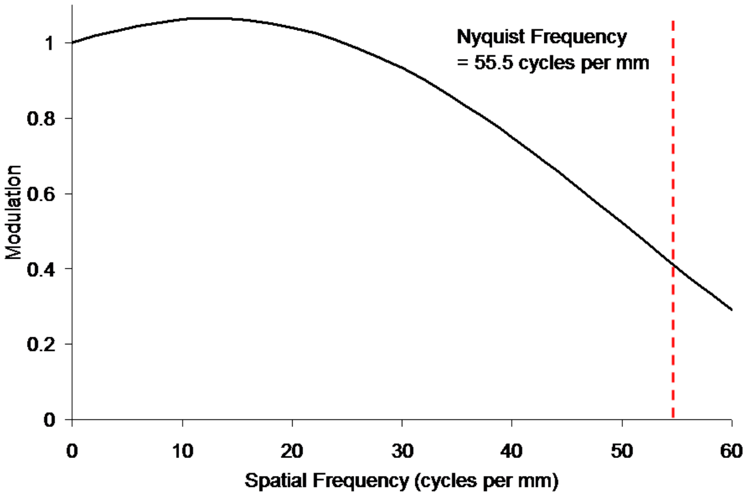
Figure 2.8: A sinusoidal input when passing through a linear spatially invariant system, results in another sinusoid of equal frequency, but with reduced modulation (and possibly a change in phase)

The modulation transfer function is a plot of the modulation transfer factor versus spatial frequency and can be evaluated by imaging a range of sinusoidal input targets and calculating the modulation transfer factor for each (the *sine wave recording method*). As described earlier, the MTF is based on the assumption that the system is linear and spatially invariant. In reality most imaging systems are non-linear, and digital systems are also spatially variant [57](as mentioned earlier, these are particular problems with lossy compressed images). In photographic systems the shape of the characteristic curve has a significant linear portion and as long as the test images are exposed on this area, they may be treated as quasi-linear. In digital systems the non-linearities are typically accounted for using the Opto-Electronic Conversion Function (OECF).

Some examples of MTFs are illustrated in figure 2.9. All show the characteristic shape of an MTF curve, illustrating an increasing loss of modulation as the spatial frequency increases. The higher that the level of the graph is, the less there is a loss of contrast. A theoretical 'ideal' imaging system would produce an MTF of unity across the full range of frequencies illustrating no change in modulation. The film curves illustrate the relationship between grain size and MTF; the fine grain film has the highest MTF across the full range of frequencies compared to the other two. This can be explained in terms of both resolution and sharpness. A film with fine grain will have a narrower PSF than one with larger grain structure. Hence its ability to resolve high spatial frequencies is better, and its resolution of edges and maintenance of edge contrast is improved. A further point to note is that the curves for both the fast panchromatic film (a(1)) and the CCD/CMOS imager in (b) increase above unity. This represents an increase in contrast from input to output and equates to a sharpening process. In the film, it is the result of *adjacency effects,* which are a chemically induced non-linearity of densities on either side of an edge. In the digital MTF, the rise is typical of a sharpening process, which may well be as a result of a sharpening operation (typically a *laplacian* filter) applied to compensate for the blurring introduced by the various interpolation processes and anti-aliasing filters typically implemented with image sensors [47].



(a)



(b)

Figure 2.9: (a) Typical MTF curves for (1) a fine grain film (2) medium speed film (3) fast panchromatic film (b) for a typical digital CCD or CMOS camera with a 9um pixel (from Jenkin [57])

For image sensors, evaluation of the MTF from sinusoidal inputs is problematic, due to imperfect alignment in-phase of the sensor elements with the sinusoidal pattern on higher frequency targets, a situation made worse by noise [58]. An alternative method is the use of an edge input, because a 'perfect' edge theoretically contains an infinite number of frequencies and edges are easy to find in images. The *edge spread function* (ESF) is obtained by scanning or sampling the edge. As illustrated in figure 2.10 the *line spread function* (LSF) may then be obtained by differentiating the ESF. The MTF is then calculated by taking the modulus of the Fourier Transform (FT) of the LSF:

$$M(\omega) = \left|\int_{-\infty}^{+\infty} l(x)e^{-2\pi i\omega x}dx\right| \tag{2.20}$$

Where *l(x)* is the line spread function.

Note that the LSF is also related to the PSF, and can be obtained by integrating the PSF in one direction.

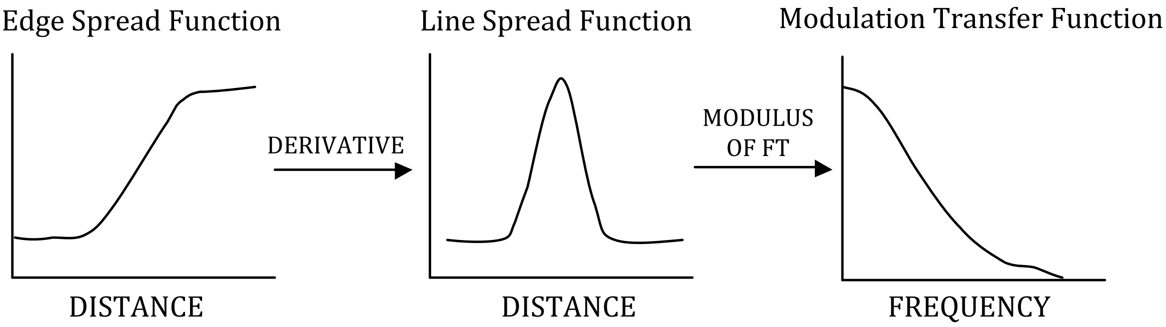


Figure 2.10 Relationships between the edge spread function, the line spread function and the modulation transfer function

In practice in digital systems an adaptation of the edge technique may be used, *the slanted edge method* defined by ISO 12233 [48] [49]. This is a method designed for use with sampled images. A region of interest is selected from an area across a slanted edge of specified contrast, usually obtained by imaging a specially designed test target and the image transformed to compensate for a non-linear transfer function (typical of digital devices) using its OECF, before channels are weighted to produce a luminance record. The edge samples are combined to allow creation of a 1D edge profile, which is effectively *super-sampled*, reducing aliasing and allowing accurate evaluation of the MTF beyond the sensor's Nyquist frequency. As for the edge method described above, the edge profile is then differentiated, followed by Fourier transformation of the resulting LSF and the modulus taken. The result is termed the *spatial frequency response (SFR)* and is distinct from MTF (although equivalent) because it takes no account of the frequency content of the target.

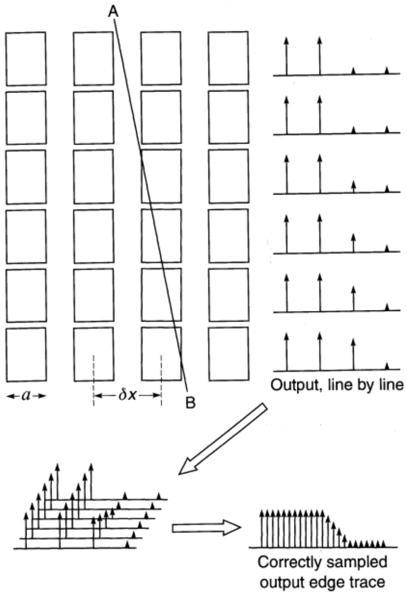


Figure 2.11: Supersampling an edge in the slanted edge method for SFR measurement © Jenkin [58]

The point at which the MTF/SFR drops to 10% is defined as the resolution limit or *effective resolution* [56] [58]. Each component within a system (including image processing as well as optical components) will have an associated MTF; the system MTF is a combination of all the individual MTFs. This *cascading property* means that the MTFs of system components are multiplied together. The effects of an individual component may hence be removed by dividing the system MTF by the individual MTF, which can be very useful in system design and evaluation.

MTF measurements are incorporated into many image quality and fidelity metrics and measures. They are also useful in providing models of the HVS, as for example in modelling the effect of the optics of the eye as a contributing component of the *contrast sensitivity function (CSF),* which describes the overall spatial frequency response of the eye. Note however that the CSF is sometimes described as the MTF of the eye, which is an incorrect definition. See section2.7.1 for more information on the CSF.

2.6.5 Noise

Noise has numerous causes and is an inherent feature of all imaging systems. It is defined as unwanted fluctuations in intensity over the image area and may be introduced by many aspects of the imaging system, processes, or as part of the signal itself. The main sources in photographic images are Poisson exposure noise, which is present in the signal (due to the random distribution of photons in a nominally uniform exposure) and therefore affects both photographic and digital images, and random partitioning of the exposing light due to the photographic grain structure. In digital imaging systems, sources additional to Poisson exposure noise include fluctuations as a result of photoelectric conversion, electronic and thermally generated noise, and errors introduced by the process of quantization in analogue to digital conversion. Although more commonly considered to be a random pattern which is superimposed on top of the image signal, in digital systems various sources of fixed pattern noise also exist. The artefacts introduced by processes such as lossy image compression are often considered as a form of noise and there is a close relationship between the simple distortion measures mentioned earlier and some of the noise measures described below.

Because noise is generally a random pattern, it may be treated as a random variable, which can be described by a probability density function (pdf). Simple measures of noise are therefore concerned with first order statistics, such as the mean of the associated pdf, which helps to describe the area of the tonal range most affected, and variance based measures which can quantify the amount of noise present. An example is a traditional measure of noise in photographic systems, *granularity,* which is obtained by taking microdensitometry traces across uniform areas of the image and assessing fluctuations in density. Assuming a uniform clean sample, Selwyn granularity, ($G = \sigma \sqrt{(2A)}$, where σ is the standard deviation of density fluctuations and A is the sampling aperture area of the microdensitometer) is found to correlate well with the corresponding subjective attribute, graininess [58]. However, as a first order statistical measure, which produces a single number quantification of the amount of noise present, granularity is not informative about either the structure of the noise pattern or its visibility. Hence, like other statistical measures of distortion, it has poor correlation with image quality. Other measures, such as the *autocorrelation function* provide a more complete description of the spatial structure of noise. The autocorrelation function is defined mathematically as:

$$C(\tau)) = \lim_{x \to \infty} \frac{1}{2x} \int_{-x}^{x} \Delta D(x) \Delta D(x+\tau) dx \qquad (2.21)$$

This describes a process of correlation of a 1-dimensional density trace *D(x)* with itself displaced by a distance, τ. At each value of τ, the product of the density deviations are calculated. Note that the term density is traditionally used, as the function is obtained from density traces, but this may be generalised to $\Delta I(x)$, to represent intensity fluctuations, when dealing with traces from digital images. The result is a function that decreases with increasing displacement. The shape of the function will be dependent upon the amount of noise present and its spatial structure. A noise trace containing many high frequencies will tend to drop to zero more quickly than one with a more low frequency structure such as that contained in a large grained emulsion.

Another important noise measure is the *noise power spectrum (NPS)*, which describes the noise characteristics in frequency space and can be obtained from the autocorrelation function via its Fourier transform [58]. The measured NPS, $\mathrm{N}_{\Delta \mathrm{I}}(\omega)$, can be directly calculated from the intensity fluctuations, $\Delta I(x)$ using:

$$N_{\Delta I}(\omega) = \lim_{X\to\infty}\left(\frac{L}{X}\left|\int_{\frac{-X}{2}}^{\frac{x}{2}} \Delta I(x)e^{-2\pi i\omega x}dx\right|^{2}\right) \qquad (2.22)$$

Where L is the length of the measuring slit, $\Delta I(x)$ the measured intensity fluctuations at displacement x, and < > represents an average of the ensemble of values across the range.

As mentioned above, the autocorrelation function and the noise power spectrum are Fourier transform pairs. The former is useful for identifying the structure and causes of noise, while the latter helps in evaluating the effects; by combining the image power spectrum (i.e. a frequency representation of the signal) with the noise power spectrum, it is possible to gain information about the signal to noise ratio in terms of spatial frequency. This is used in image processing and image restoration and is the basis of the *Wiener filter* [4]. Note also that the variance of the measured intensity fluctuations is equal to the value of the NPS at zero spatial frequency.

2.7 Modelling visual perception

Early approaches to image quality assessment focused upon the understanding and quantification of the signal, or of measurable physical characteristics of the imaging system. More recent research has attempted to better quantify the visual effects of changes to image attributes, by using knowledge of the properties of the human visual system and underlying neural processes. Significant advancements, particularly in the last decade, in our understanding of the human visual system have allowed the development of better models for perceptual criteria used in image quality evaluation [59]. Much of this research has stemmed from fields of visual neuroscience and visual psychophysics.

In practical applications, image quality can be modelled by the perceptual weighting of all relevant visual attributes, the relative significance of each determined by the specific imaging context and purpose [60]. Study of human vision is a vast and complex subject, but some of the psychophysical properties of the human visual system of particular relevance in modelling image quality are detailed below.

As described by Pappas et al, [61], human visual sensitivity to variations in luminance depends on a number of factors, including light level, spatial frequency and signal content. The perceptual models most used within image quality relate to lower order aspects of vision. These are based upon an understanding of:

- The optics of the eye, the physical characteristics and modelling of the visual pathway until visible radiation reaches the retina;
- Photoreceptor responses (trichromatic, and opponent), scotopic and photopic vision, light and dark adaptation and initial mechanisms of colour vision;
- Ganglion cell receptive fields, which exhibit *centre-surround antagonism* [62] and can facilitate various signal processes, such as dynamic range enhancement, edge detection and enhancement and compression of redundant visual information;
- The lateral geniculate nucleus, a part of the thalamus, which has a fundamental role in temporal and spatial correlations of signals received on the visual pathway from different areas of the visual field, from both eyes, to produce a three-dimensional representation of space. LGN processing is complex and not yet well understood, but LGN cells and groups of cells also have receptive fields and are believed to play a role in the coordination of visual attention [63];
- The striate cortex, where more complex encoding of visual information is performed. Area V1 for example, contains cells (or groups of cells) that are tuned to respond to specific stimuli important to higher order visual understanding, such as specific bands of spatial frequencies, or specifically oriented edges; the cells also interact and combine responses to build up more complex perceptions important for visual phenomena such as size and shape constancy, motion or depth perception [62].

Modelling of the functions described above allows better predictions of their effects on visual sensitivity, which can be incorporated into perceptual image quality metrics. Some of the models more commonly used in image quality are summarised below; Chandler in [59] provides a more in-depth discussion.

2.7.1 Contrast Sensitivity Function

The variation in response of the human visual system to contrast as a function of spatial frequency is known as the Contrast Sensitivity Function (CSF). Contrast sensitivity is the inverse of the contrast threshold at a particular spatial frequency; that is, as defined by Johnson and Fairchild [64] as 'The level of contrast necessary to elicit a perceived response by the human visual system'. As mentioned earlier, the CSF is sometimes described incorrectly as the MTF of the HVS. The primary difference however is that MTF is based upon linear systems theory, whereas the CSF is a measure of the contrast response of the entire visual system, which is neither linear, nor spatially invariant.

The CSF is different for achromatic and chromatic channels. There are a number of different models for the achromatic CSF, which vary in complexity, and there have been various experiments to measure the CSF, which have provided good agreement with the models. One of the most well-known of the more complex models is that of Barten, described in detail in his book 'Contrast Sensitivity of the Human Eye and its Effect on Image Quality' [65]. His model considers contrast sensitivity as a function of the internal noise (photon and neural noise) of the eye, lateral inhibition processes, also influenced by external noise, image size, pupil diameter, colour temperature of illuminant and luminance level. Daly's CSF [66] is also complex, and includes parameters for orientation, luminance, radial spatial frequency, image size and viewing distance.

A somewhat simpler model is that of Movshon and Kiorpes [67], which is approximated by a three-parameter exponential function and is viewing-distance dependent. It is a version of this that has been adapted for the Modular Image Difference Metric by Johnson and Fairchild [68] described in more detail in (7.2), and implemented in Chapter 7 of this thesis.

Common to all of these functions is a shape that is bandpass in nature, peaking at somewhere between 4 and 8 cycles per degree (cpd). The effect of this, as shown in figure is to enhance frequencies around the 4cpd level while attenuating the very low frequencies and the much higher ones. Models such as that of Movshon are isotropic, making them easy to implement. Daly's is anisotropic, providing a larger response to horizontal and vertical frequencies compared to diagonal frequencies [64].

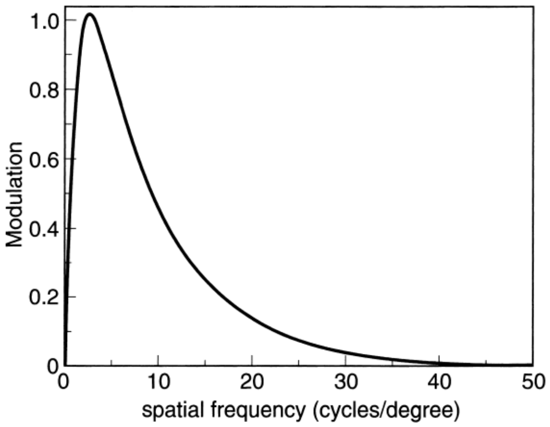


Figure 2.12 Bandpass shape of achromatic contrast sensitivity function ©R Jenkin, from [63]

There are fewer models available for the chromatic channels. One that is described in the modular image difference model [68] [69] takes a similar form to the Movshon-Kiorpes model, but has the form of a low pass rather than band pass function.

CSFs are often included at an early stage within an image quality metric as a form of spatial frequency filter prior to error calculations. The functions may be approximated through convolution filtering in the spatial domain, or more accurately modelled by multiplication with the modulus of the Fourier transform of the image in the frequency domain. Their effect is to modulate frequencies beyond the limits of the HVS at a particular viewing distance, hence reducing errors within these frequencies that would have no significant visual effect.

It should be noted however that measured and modelled CSFs of this form are derived from the detection of simple targets (often consisting of one or only a few different spatial frequencies) against a plain background. More recent research in image quality metrics, detailed later, is exploring different approaches to the modelling of constrast sensitivity.

2.7.2 Visual Masking

Masking describes the effect that the presence of a signal can have in suppressing the visibility of another signal, and hence the ability of an observer to detect that signal. This is a cause of *scene susceptibility,* a form of scene dependency (see 3.6.1). Luminance masking may be considered to be a function of *amplitude non-linearity* [61]as a result of *Weber's law*, which describes the increase in a threshold of detection or just noticeable difference as a result of the increase in background luminance:

$$\frac{\Delta L}{L} = k \tag{2.23}$$

Where L is luminance and ΔL describes the change in luminance required to produce one just noticeable difference in luminance (i.e. the threshold) and k is a constant.

Pattern masking is another form of masking where the presence of a pattern or a texture masks the visibility of distortions, particularly if they are similar or close in frequency. Noise masking can occur as a result of additive noise in the image or the process, which disrupts the detection of, for example, edges, or boundaries of objects. Distortions caused by digital processes can also be considered to be noise.

Contrast masking [59] is also described as spatial frequency adaptation [70]. Spatial frequency adaptation occurs as a process of desensitisation to particular frequencies and those in the adjacent octaves as a result of the presence of the frequency in the visual field. The effect is described as a 'Dipper effect' by Chandler [59] because of the impact that it has on the shape of the contrast sensitivity function.

2.7.3 Multichannel model of the HVS

Experimental work by Campbell and Robson [71] found that the contrast detection thresholds of complex gratings (in this case square waves) were lower than that for a sine wave of the fundamental frequency. This led them to the conclusion that the HVS decomposes a signal into multiple separate spatial frequency channels. This *multi-channel model* may be used in image quality metrics to isolate and modulate specific bands of frequencies. The wavelet transform that is the basis of JPEG 2000 may be viewed as a form of multi-channel decomposition.

2.8 Types of Metrics

As mentioned in section 2.3, classical signal processing approaches to image quality view the imaging process from the bottom-up. Here the problem is seen as one of modelling the physical parameters that affect image quality and using them to process the image signal. The range of metrics that have been developed is enormous and new metrics are continually being developed in the search for a speedy and robust approach to accurately predict perceived image quality as an alternative to the need for time consuming psychophysical experiments. Here some of the broad approaches to metrics are summarised. Those relevant to this work are described in more detail in a later section.

Jacobson describes *Image Quality Metrics* (IQMs) as [72]: 'single numbers (figures of merit) derived from physical measurements of the system, which relate to perceptions of image quality'. These methods combine measures of system attributes only or their related perceptual attributes, without accounting for the visual system. Examples include *Minkowski metrics,* used to quantify errors between images [73], or in a multivariate formalism to combine different perceptual attributes, such as the combination of sharpness and graininess by Bartleson [74]. Wang et al [20] highlight some of the limitations of Minkowski metrics: they are based upon the implicit assumption that signal samples are independent of one another, which in natural images is not the case. Hence they are incapable of predicting the *visual* errors across the image, which are influenced by the structure of the image as a result of dependencies between samples (for example errors may be masked in areas of containing fine detail). Other types of IQMs include colour difference formulae as described in section 2.6.4. Although some versions include viewing conditions, they do not incorporate visual models and therefore are incapable of modelling colour appearance.

For more detail about current and recent research in image quality metrics, Chandler [59] provides a useful and detailed overview of recent developments in image quality and metrics, classified into full-reference (FR) and reduced reference/no-reference (RR/NR) algorithms. He further divides the full-reference methods (which are of relevance to this work) into the following categories [59]:

2.8.1 Methods based on HVS models

Visual Image Quality Metrics extend classical metrics to incorporate models of the human visual system. They combine physical measures of image attributes (usually sharpness and noise as a minimum), which are weighted for the human visual system using some form of contrast sensitivity function and possibly models of other visual phenomena.

These metrics commonly include masking, and spatial frequency decompositions, to produce some form of error measure between distorted and undistorted images. These models tend to work best at threshold levels of distortion, because the knowledge and models for the CSF and other perceptual phenomena is better defined, than that for suprathreshold evaluations. Visual image quality metrics and modular image difference models are examples of these types of metrics. Examples include the Square-Root Integral with noise ($SQRI_N$). Pictorial Information Capacity (PIC) and Effective Pictorial Information Capacity (EPIC). However these have certain limitations, in that they are linear metrics, and digital imaging systems exhibit many non-linearities, limiting the validity of results and requiring caution and adaptation in their application. More complex perceptual metrics [75] use better models of visual processing, including frequency analysis to decompose the image into sub-bands of different frequencies and orientations and incorporate non-linear mechanisms for masking effects, followed by error pooling. They are commonly used for evaluating thresholds or distortions. Examples include Daly's *Visible Differences Predictor* [76].

Colour science has been another source of VIQMs, with the spatial extension by Zhang and Wandell [77]] of the standard CIELAB ΔE equation, known as the S-CIELAB metric. This model spatially filters the image in an opponent colour space to approximate the CSF, prior to calculating colour differences. Johnson and Fairchild more recently introduced the Modular Image Difference Model [68], which extends S-CIELAB with several pre-processing steps, to model various appearance phenomena such as adaptation and local and global contrast detection.

Some VIQMs have also been developed to incorporate some form of saliency modelling to determine visual importance of specific areas within the image, to allow weighting of the models.

2.8.2 Methods based upon image structure

Recently image quality assessment has been approached from a quite different (information processing) perspective. Structural approaches to image quality are based upon a number of fundamental assumptions as described by Sheikh and Bovik [78]: (i) That natural image signals contain significant amounts of structural information and correlation (ii) the human visual system is highly adapted to extract useful structural information from natural scenes. Therefore a measure of structural similarity between a processed and reference image should be a good indicator of perceptual distortion. The Structural Similarity Index (SSIM) [78] [79] achieves this by calculating relatively simple *luminance comparison*, *contrast comparison* and *structure comparison* functions from first-order statistical information. The three comparisons are weighted and combined to produce the SSIM index value. Another similar measure is the Visual Information Fidelity Measure (VIF) [73] [80], which aims to quantify the 'mutual information' shared between the test and reference image and to relate this to the visual quality of the test image.

Alternative structural methods have been developed and use other structures within the image, for example gradient-based methods or methods which include some form of wavelet, DCT or Fourier transform, which may predict, for example, the presence or degradation of textures within an image.

2.8.3 Other Issues

2.8.3.1 Universal applicability of metrics

There is a wide range of alternative methods, using image statistics or machine learning, to identify structures or distortions and predict their effects within images. The success of the methods depends of course upon the nature of any distortions within the image, and how they interact with the specific processes within the image quality algorithm. This has meant that some methods will work well on particular types of distortions but not on others, therefore they are not universally useful. The same can be said for the sample images tested, if they are all of one type or sharing common characteristics. But perhaps this is a reflection on the complexity of image quality evaluation: the pace of development in imaging modalities and technologies (for example High Dynamic Range Imaging) means that often metrics are either developed to address a particular need as a new technology is released, or are playing 'catch-up' in testing and comparing new imaging technologies against existing ones after they are released. Giocca et al [60] provide a useful review of currently available metrics, including Full-Reference (FR), Reduced-Reference (RR) and No-Reference (NR), which includes descriptions of their operation, approach and implementation, and performance. The No-Reference methods tend for the most part to be developed for the evaluation of specific algorithms or artifacts.

2.9 Subjective Evaluation: Psychophysics

As introduced in section 2.3, subjective evaluation of images uses psychophysics, which is described by Sharma [49], as the 'study of the relationships between physically measured stimuli and the sensations and perceptions of those stimuli'.

Engeldrum [24] highlights the importance of distinguishing between observer responses that are judgements, from those that are preferences*.* The former are often considered to be to some extent more 'objective' than the latter, with less expected variation in observer response. Preference may be regarded as more 'subjective', being much more closely linked to an individual's quality criteria; therefore response is more variable across a group of individuals. To obtain meaningful results, care must be taken during experimental set-up, with consideration to observer demographic and their experience in image judgement, careful scene selection and attention in the wording of instructions given to the observers to ensure that the requirements of the task are clear.

Psychophysical experiments are generally of two types: scaling experiments, which investigate relationships between stimuli and perceptual magnitudes; and threshold experiments, which involve detection and discrimination of small visual changes in stimuli, or alternatively that identify visual matches.

Scaling experiments encompass suprathreshold magnitudes (judgement and preference) and may evaluate individual attributes, combinations or overall image quality. Threshold experiments are more commonly concerned with image fidelity, (which may be correlated with image *naturalness*) or acceptability (which may be correlated with image *usefulness,* although this is not always the case). Keelan [16] suggests a framework using Just Noticeable Differences (JNDs) derived from threshold experiments, as natural units for image quality evaluation, for example in calibrating numerical scales of quality, allowing the comparison of different image quality scales or scales of different attributes. He also suggests that [16]: 'The term just noticeable difference might be better generalised to just significant difference or something similar' when considering preference. Such a term is particularly applicable when considering artefactual attributes, such as errors introduced by compression. Sharp [81] notes:

'It is relatively easy to recognise a difference between the appearance of two images…however, appearance is not necessarily synonymous with image quality'

Sharp comments on this difference in the context of medical imaging, where image usefulness in the context of object detection and recognition may require that image processing be applied to enhance certain image characteristics, while sacrificing overall image quality. However his point may equally be applied to 'beauty contest' image quality judgements, where detection of a distortion is not necessarily to the detriment of image quality.

2.9.1 Types of Scales

There are four different types of scales generated in psychophysical scaling experiments, originally defined by Stevens [82] in 1946. The scale types and characteristics are summarised in table 2.4.

Scale	**Basic Empirical Operations** **Stevens' definitions** [82]	**Scale Characteristics**
Nominal	Determination of equality	No magnitude or direction. Samples are placed in categories only, which may be identified by numerals or descriptors, but the numbers do not relate to quantities. A form of classification, rather than quantification.
Ordinal	Determination of greater or less	Direction without magnitude. Samples are rank ordered in terms of attribute(s) being scaled, but without information about distances along the scale.
Interval	Determination of equality of intervals or differences	Quantitative scales. Samples are placed numerically along the scale, to allow determination of distances (differences) between them. Interval scales have no absolute zero point and are floating scales, meaning that numerical values are not fixed and have meaning only relative to one another.
Ratio	Determination of equality of ratios	Quantitative scales with a fixed zero point. More difficult to generate experimentally; and it is conceptually difficult to define a zero point for many attributes and for overall image quality.

Table 2.4: Description of the four classic types of psychometric scale, with reference to Stevens [82] definitions

The scales increase in the level of complexity from top to bottom in the table, with each one encompassing the qualities of the ones above. Experimentally, the scaling process becomes more difficult for observers, and data analysis also becomes more complex, but much more information may be gained, and the scale comparison becomes more achievable. Interval scales are probably the most widely used in the study of image quality, representing a good balance between complexity and usefulness. The scales are illustrated in figure 2.13.

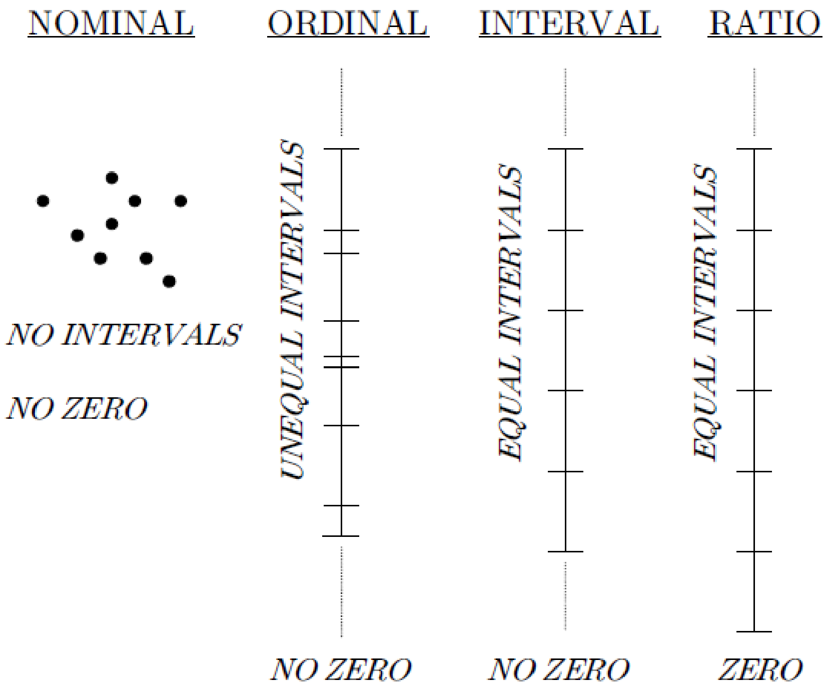


Figure 2.13: Illustration of Stevens' psychometric scales, from Boynton [83]

2.9.2 Scaling Methods

There are a number of approaches to scaling, which differ in terms of complexity to set up and the type of scale derived [40] [84]. All have limitations, which has led to the recent development of the quality ruler method, described in ISO 20462 [84] [85] and later in this chapter.

2.9.2.1 Rank Ordering and Paired Comparison

Both of these approaches require comparison between image samples, the main difference being in the number of samples considered simultaneously. In rank ordering an entire sample set will be considered and ranked, which requires a large working space for hard copy and is usually impractical for soft copy samples. Paired comparisons consider the sample set in a series of pairs and in each pair they identify which is greater or lesser in terms of quality or the attribute or 'ness' under consideration. Paired comparison takes longer, as samples must be presented multiple times, but results are generally more robust than rank ordering. This, and the limitations in the application of rank ordering mean that paired comparison is one of the more widely used approaches. Data from both may be converted to interval or ratio scales based upon Thurstone's Law of Comparative judgement, assuming that perception of two samples is modelled by a pair of Gaussian distributions and using z-deviates to determine differences between them [16] [18] [40] [84]. However, if the difference exceeds 1.5 JNDs, the magnitude of the difference cannot be reliably estimated [16] [84]. Paired comparisons are also widely used in the evaluation of thresholds and JNDs as described in section 2.7.3. ISO 20462 additionally specifies a triplet comparison method, which 'combines elements of paired comparison, rank ordering, and categorical sort methods'. [86].

2.9.2.2 Categorical Sort

Here the stimulus is classified into a category identifying either different levels of quality or attributes. The categories have descriptors to identify the levels. This method is relatively simple to set up and as an observer task it is easy to understand, meaning that assessments can be quite rapid. However it has a number of limitations [16]. Observers do not tend to use the end categories much, and the results are often dependent on the number of categories, rather than image properties. The descriptors used can influence the results, and the number of categories is relatively few. While simple to produce an ordinal scale, the conversion to an interval scale is complex, and the lack of a standardised set of adjectives as category descriptors makes comparison between experiments difficult.

2.9.2.3 Magnitude Estimation

Observers are required to assign a numerical value to a sample stimulus, usually by comparison with a reference sample used to anchor the range of values. This is a method often used to generate a ratio scale, which, as defined in ISO-20462 part 1 [84],

'…Is a scale in which a constant percentage change in value corresponds with one JND. In practice, modest deviations from this behaviour occur, complicating the transformation of the rating scale into units of JNDs without inclusion of unidentified reference stimuli (having known quality)'

2.9.3 Thresholds and Just Noticeable Differences

Engeldrum [18] describes the difference in the understanding of thresholds and JNDs in classical and modern psychophysics. Classically, thresholds are defined as the amount of a physical image parameter required to be perceptible or to evoke a just noticeable difference. Modern psychophysics broadens the approach, so that thresholds and JNDs may be evaluated for more complex 'nesses' such as image quality and are not tied to a single physical image parameter. However, on a practical level this requires that mechanisms must be devised to link the 'ness' scale of JNDs to physical image parameters, using visual algorithms.

Threshold experiments are useful when evaluating processes that introduce distortion, such as compression. They require binary yes' or 'no' responses from the observer, who may be presented with a single image and asked whether they can detect a 'ness' or attribute; or they may asked to compare the image with a standard or reference and asked whether they can either see the 'ness', or whether the test image is acceptable in comparison with the reference. The first two examples identify a threshold of detection, which is a process of determining the sensitivity of the perceptual system to a stimulus, its *perceptibility*. By contrast, the third example identifies a threshold of *acceptability,* which is a process of determining the point at which a change in the stimulus becomes bothersome, and so is about the impact of the stimulus on quality.

It should be noted that the meaning of acceptability depends upon the context; it can be correlated with *usefulness* for example, if acceptability refers to whether artefactual attributes are detrimental to detection and recognition (for example in diagnostic imaging); whereas in a more general imaging context (the ‘beauty contest’ as defined by Engeldrum [24]) acceptability may correlate more with a combination of *naturalness* and *fidelity.*

Although the observer task is different in perceptibility and acceptability threshold evaluation, the process of data analysis is similar. The experiment described in chapter 5 explores the relationship between perceptibility and acceptability in compressed images.

Gescheider [87] highlights the complexity of determining thresholds of perceptual attributes:

‘Biological systems are not fixed… but rather are variable in their reaction and therefore when an observer is presented on several occasions with the same stimulus, he is likely to respond “yes” on some trials and “no” on others. Thus the threshold cannot be defined as the stimulus below which detection never occurs and above which detection always occurs’

His point is that observations vary not only across groups of observers because of the variability in their perceptual systems, but also in individual observers when presented with the same stimuli at different times. Therefore the decision making process is not deterministic but probabilistic; the threshold value is a random variable. Various steps may be taken in experimental design, such as careful selection of observers, stimuli and presentation of stimuli multiple times across multiple observers to mitigate this variability and reduce noise in the results, but observer responses in perceptual experiments will nevertheless be distributed upon a perceptual continuum. A probability density function is used to model the variability of observer responses and it is found experimentally [87] that the variation tends to be normally distributed.

2.9.3.1 The Psychometric Function

The psychometric function is a graph constructed from the proportion of ‘yes’ responses, *p,* plotted against *X,* the stimulus intensity or another parameter, such as compression ratio. A curve is fitted to the data points, and often, if produced from enough responses, the curve will tend towards an *ogive* function, which is a particular type of s-shaped curve [87]. The curve is a cumulative distribution function of the probability density function describing user responses [18]. If the model fits the data well enough, parameters describing it may be defined and these can be used to extract various useful information. If the proportions of 'yes' responses follow a Gaussian (normal) distribution, then the proportions may also be expressed as z-scores, and the relationship between X and z is linear. The relationship between the normal distribution and the psychometric function is illustrated in figure 2.13.

The *absolute threshold* is defined as the smallest amount of the stimulus that is perceptible and is commonly taken as the point at which observers respond 'yes' in 50% of trials [18]. This is also known as the *point of subjective equality* (PSE), referring to the point in a JND study at which two images will be seen as equal. Effectively, it is the point at which the stimulus is perceptible to the more sensitive of observers, but without certainty, and takes into account the probability of some positive responses as a result of observer guessing. In figure 2.13, it is the amount of X that produces a proportion p of 0.5, and is also the area under the normal distribution curve to the left of the z=0 point (which is 50% of the total curve).

The *difference threshold,* or *just noticeable difference* is usually taken to be the 0.75 proportion point on the psychometric curve. ISO 20462 defines a JND in relation to the responses in a forced choice paired comparison test, as the point at which there is a 75%:25% proportion of 'yes' responses.

Keelan [16] describes the *JND increment* as 'the number of units of an objective metric or rating scale required to produce a sample difference of one JND. In other words, it is one JND from the point of subjective equality and means that by this point there is enough certainty in its perceptibility for the majority of observers to perceive it. It is the 0.75 line in Figure 2.14, which corresponds to 75% of the area under the normal distribution (to the left of the second dashed line at z=0.67).

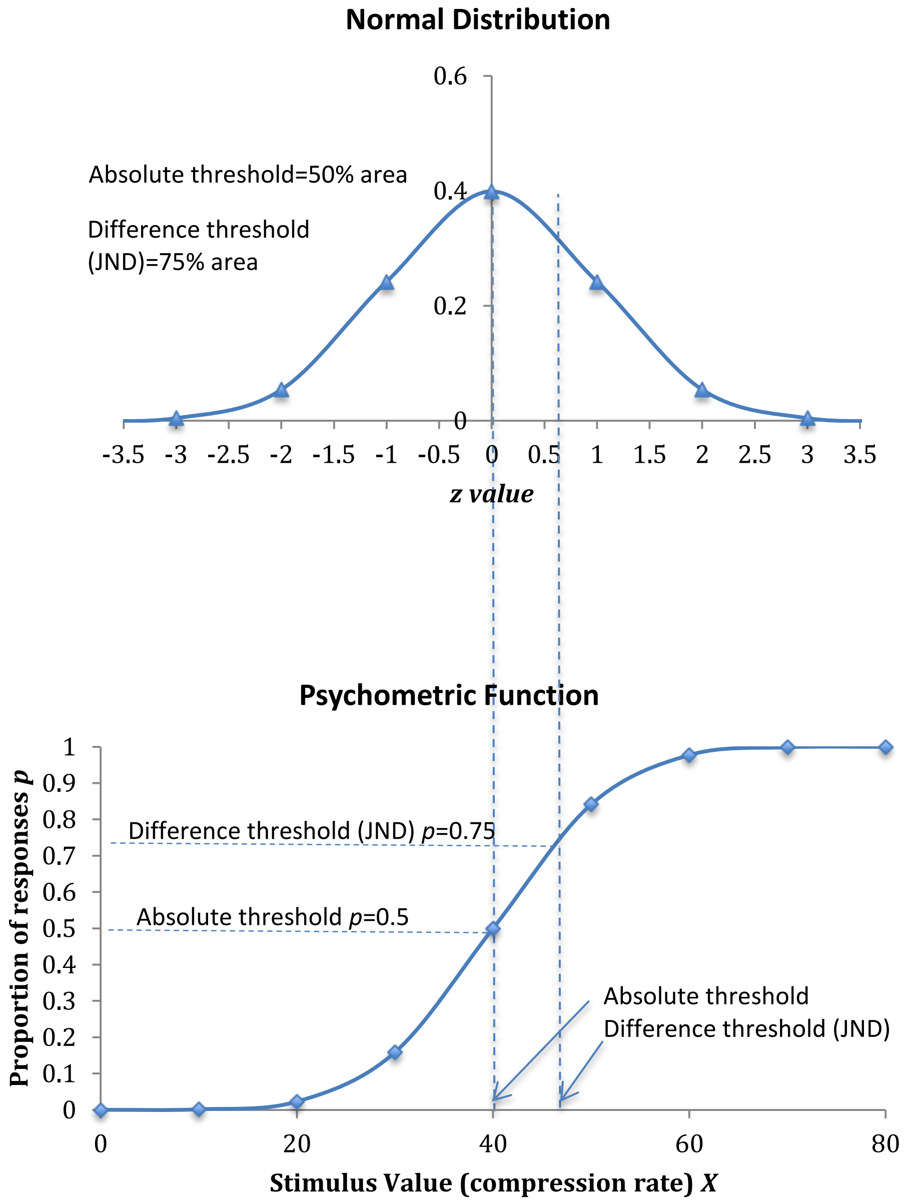


Figure 2.14: The relationship between the psychometric function and the normal distribution. When proportions on the psychometric curve are transformed to z-values, thresholds on the psychometric curve correspond to areas under the normal distribution to the left of the z-values. Adapted from Gescheider [87]

ISO 20462 differentiates the two different types of JND:

- An *attribute JND* requires that only one attribute in the image is varying, and is a measure of the detectability of that change.
- A *quality JND* is a measure of the effect of changes of combinations of image attributes upon image quality.

Engeldrum [18] distinguishes between the absolute threshold and the difference threshold in terms of the question being asked of the observer. In the first case they are being asked to identify the first point at which the ‘ness’ or attribute is detectable, whereas in the second they are identifying at what level of the ‘ness’ or attribute the image is seen as different from a reference or standard. Determination of the absolute threshold can be particularly useful when scaling some ‘nesses’: conceptually the idea of a zero point is difficult to imagine for complex ‘nesses’ which do not have a clear physical correlate (for example, naturalness, or image quality), so they tend to produce interval scales. The absolute threshold may be used instead as a theoretical zero point with JNDs as unit increments, allowing better interpretation of scale values.

2.9.3.2 Approaches Used In Threshold Experiments

There are three main approaches used in classical psychophysics for the evaluation of thresholds, which determine how sample stimuli are prepared and presented to observers [18] [87].

In the *method of limits* observers are presented with a series of sample stimuli, which begin with a sample that is either well above the detection threshold (all observers should be able to detect the distortion or ‘ness’) or well below the threshold (no observers can detect it). The range of sample stimuli should be determined beforehand using a pilot study. The stimuli are manipulated by the observer, in a descending or ascending series towards the threshold, until they reach the transition point between detection and non-detection, before the process is repeated from the opposite end of the series, so that observers are detecting the point at which the distortion or ‘ness’ becomes perceptible in one direction and the point it which it becomes imperceptible in the other. The transition point will be midway between the two.

The *method of adjustment* is similar to the method of limits, but the observer is not restricted in terms of the way that stimuli are presented and controls the stimuli themselves from a random starting point in the series. Engeldrum [18] points out that the ‘active involvement of the observer raises interest, reduces boredom and tediousness, and generally improves the quality of the data’.

The *method of constant stimuli* is so called because it uses a set of fixed stimuli from within a range of stimulus variation, that is presented to the observer multiple times in a random order. The range is determined by a pilot study to ensure that at one end the 'ness' or distortion is always detected and at the other that it is never detected. The method of constant stimuli is one of the most widely used approaches, as it is simpler in many cases to use fixed samples but also because it allows the formation of a psychometric function, described below, from the collected data. The method of constant stimuli may be applied as a *no-reference* method, in which single samples are presented without a standard or reference, and an absolute threshold is determined; or as a *full-reference* method, using a reference or standard in a comparison with the sample in a paired comparison experiment.

The various common approaches for both scaling and threshold experiments are detailed in table 2.5, which illustrates the information that may be derived from them.

Method	Threshold	JND	Ordinal Scale	Interval Scale	Ratio Scale
Method of Limits	X	X			
Method of Adjustments	X	X			
Method of Constant Stimuli	X	X	X	X	
Rank order			X		
Categorical scaling			X	X	
Rating scaling				X	X
Magnitude estimation				X	X

Table 2.5: Common psychometric methods for threshold estimation and scaling, from Triantaphillidou [40]

2.9.4 The Quality Ruler

The quality ruler method is defined in ISO 20462 part 3 [85] as a 'psychophysical method that involves quality or attribute assessment of a test stimulus against a series of ordered, univariate reference stimuli that differ by known numbers of JNDs'.

The images in any one ruler are of the same scene, and are presented to the observer so that they may compare a ruler image side-by-side with a test image (usually, but not necessarily, of the same scene), to identify the ruler image closest in quality to the test image. Because the ruler images are calibrated to differ by known amounts of quality JNDs, the ruler method allows the immediate association of a numerical value with the test stimulus when the observer identifies the ruler image which best matches the test image in terms of image quality [88]. The reference samples are closely spaced in terms of subjective quality and span a large range overall, to reduce the inaccuracy introduced by other psychophysical methods as a result of either extrapolation (when the range is small) or interpolation (when the range of samples is large and spaced out). It also avoids the problem of saturation effects, which may occur in threshold experiments with sample comparisons spaced by more than 1.5 JNDs, resulting in the proportions appearing in the tails of the assumed underlying Gaussian distribution (which tend towards positive and negative infinity). Therefore the quality ruler is suitable for both threshold and suprathreshold quality evaluations.

The ruler images vary in a single attribute, which prevents the ambiguity in determination of image quality [16] that can be introduced as a result of interactions between multiple attributes; this also simplifies the process of ruler generation. There are several requirements of such an attribute: Its variation needs to have a known impact on image quality; Varying the attribute must produce results that are relatively robust in terms of observer sensitivity and variation across scenes. Keelan [16] also suggests that it should be an attribute that varies widely in practical imaging systems (so fulfilling naturalness criteria). Finally, the attribute needs to be easily simulated and characterised.

An attribute fulfilling these criteria particularly well is sharpness, and therefore sharpness is used in the model within the standard. Sharpness is easy to manipulate through digital image processing and its correlation with the MTF means that the user can generate a set of ruler images for a particular test scene varying in sharpness, by manipulation of the system MTF.

A set of reference prints varying in sharpness (the *standard reference set,* (SRS)) has been made available through the I3A, and these are calibrated to the *standard quality scale* (SQS) [85]. This is:

'…a fixed numerical scale of quality having the following properties:

a) the numerical scale is anchored against physical standards;
b) a one unit increase in scale value corresponds to an improvement of one JND of quality; and
c) 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'

The reference images can be used to calibrate user-generated rulers to the SQS values (note that hard copy rulers calibrated to the SRS are quantified in primary SQS_1 values, whereas soft copy quality rulers, calibrated to the DRS are termed secondary SQS_2 values). If not calibrated against the SRS or DRS, *scene dependent* ruler calibration [85] should be applied, to obtain results that are not biased as a result of scene content.

Alternatively, attributes other than sharpness may be varied in rulers, but these must be calibrated using a quality ruler varying in sharpness. The attributes should be artefactual attributes (as defined earlier, those that when visible, generally have a negative impact upon image quality) rather than preferential attributes, such as relative colourfulness, because these are much more subject to observer variation.

The standard describes both hardcopy and softcopy implementations of quality rulers. The softcopy method is of interest in the context of image compression and is the method employed in chapter 6.

2.9.4.1 Generation of Ruler Images

It is most useful to be able to generate ruler images for different scenes. Although images can be evaluated by comparison with different but similar scenes, the task for observers is conceptually simpler if they are comparing the same image. Ruler images of the scene are therefore generated by manipulation of the system MTF.

The system MTF of the complete imaging system generating the ruler images is first characterised for horizontal and vertical orientations, and this is checked for conformance with the shape of the monochromatic MTF of a diffraction-limited lens defined by:

$$m(v) = \frac{2}{\pi}\left(cos^{-1}(kv) - kv\sqrt{1-(kv)^2}\right) \qquad kv \leq 1$$

$$m(v) = 0 \qquad kv > 1 \qquad (2.24)\ [85]$$

Where *m* is the modulation of the imaging system, *v* is the spatial frequency in cycles per degree at the eye of the observer and *k* is a constant. The *aim MTF* is identified as the one modelled from equation 2.23 where the area under the curve best matches that of the measured system MTF over the range of 0 to 30 cycles per degree. The value of *k* for this aim MTF may be regarded as reciprocally related to the system bandwidth.

By varying the value of the *k* constant, a series of MTF curves may be generated from equation (2.24) [85], each of which corresponds to a differing amount of blurring or sharpening of the system MTF. Values of *k* may be selected to provide required JND increments using [15]:

$$JNDs = \frac{17{,}249 + 203,792k - 114{,}950k^2 - 3{,}571{,}075k^3}{578 - 1{,}304k + 357{,}372k^2} \qquad (1 \leq 100k \leq 26)$$

(2.25)

Some examples of MTF curves generated from equation (2.24) [85] and differing by one quality JND [85] are illustrated in Figure 2.15 A series of MTFs generated from equation(2.24) [85], spaced by increments of 1 JND.

The JNDs values calculated from are relative JND values; the difference between scale values calculated for two reference stimuli are equivalent to differences in quality JNDs, but the scale values are not absolute. However, they are also expressed as Standard Quality Scale (SQS_2) values in the later version of ISO 20462-3 [85]. SQS_2 values are ‘obtained through assessments traceable to the Digital Reference Stimuli (DRS) or the average scene relationship’ [85].

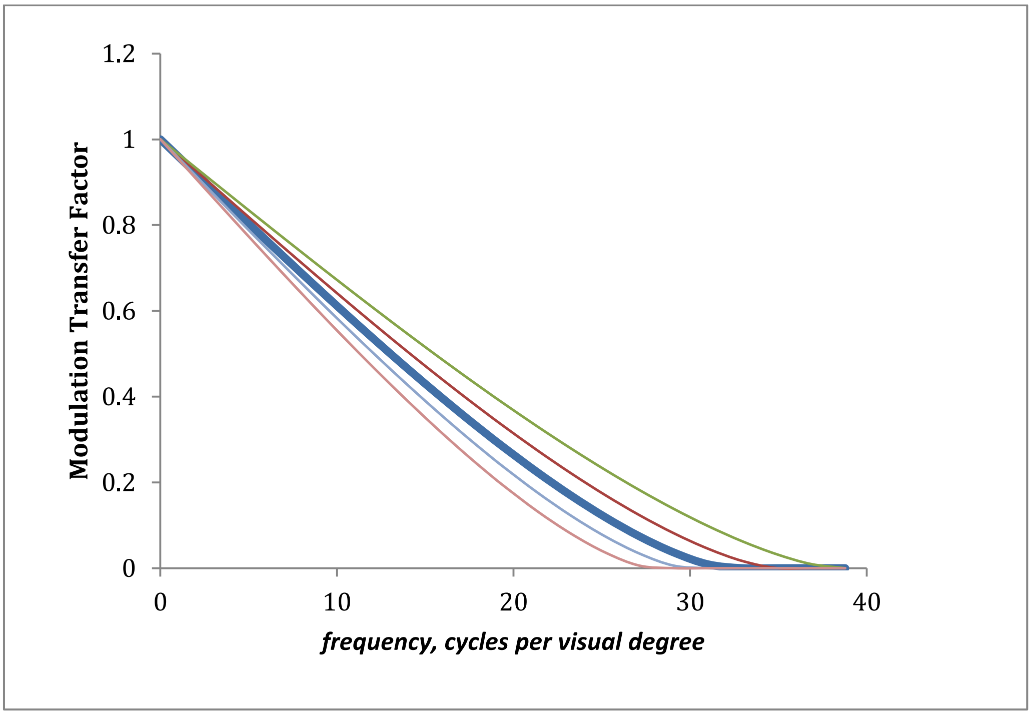


Figure 2.15 A series of MTFs generated from equation(2.24) [85], spaced by increments of 1 JND.

The system MTF of the original reference image is modified to approximate the aim MTFs for the series of JND increments required. This modification is achieved using linear spatial filtering by Jin, Keelan et al [38], and by filtering in the frequency domain by Young-Park [89]. An alternative implementation of a frequency domain approach is presented in chapter 6

3 Image Compression and Image Quality

3.1 Redundancy in Images

The process of encoding an image involves the assigning of a *code word*, which is a set of binary digits, to a *source symbol*, which may be a single pixel value or a group of pixel values, or alternatively, to transformed information such as frequency coefficients [5].

Image compression is possible because of the existence of redundancies within natural images. Gonzales and Woods [4] broadly classify these redundancies into three types:

Interpixel redundancy, also called *spatial redundancy*, exists due to correlation within the structure of the image itself. Note that the term 'spatial' does not specifically refer to the spatial image plane, but to the relationships between a value and the values around it (which could be in the frequency domain). For example, identical or similar consecutive values within a frame, or between values in the same location in consecutive frames in a moving image (interframe redundancy), or between colour channels (spectral redundancy). Interpixel redundancies are common in areas of low frequency, where tones and colours are unchanging or are changing smoothly, in a predictable manner. These correlations mean that each value need not be explicitly encoded. Values can be encoded in groups, as differences, or using models to predict their values [7]. The result may be lossless or lossy, depending upon the accuracy of the prediction.

Coding redundancy, or statistical redundancy, is a data redundancy as a result of the unequal frequency distribution of digital levels in natural images (as illustrated by peaks within image histograms). Using the same fixed length of code for every value is inefficient, and therefore variable length coding methods [7] [5] are used, to assign shorter codes to the most commonly occurring values.

For a single channel, the average length of code per pixel, L_{ave}, may be calculated from:

$$L_{ave} = \sum_{i=0}^{L-1} P_i \, l(i) \qquad (3.1)\ [5]$$

Where P_i is the probability of a pixel taking value *i, L* is the number of possible levels that a pixel may take and $l(i)$ is the code length assigned to level i. Coding redundancy is exploited to minimise L_{ave}.

Psychovisual redundancy is redundancy in image information due to the limits of the human visual system. This may be regarded as perceptually irrelevant information [6] [7]. These methods often include a frequency transform, to allow attenuation of frequencies that are beyond the limits of the Contrast Sensitivity Function (CSF), or to reduce the magnitude of frequencies that are less visually important within a scene. Other areas of redundancy occur as a result of the reduced colour discrimination compared to tonal discrimination of the HVS and the non-linear nature of many aspects of human vision (see section 2.7). As noted by Sayood [7]:

"…the mind does not perceive everything the eye sees. We can use this knowledge to design compression systems such that the distortion introduced by our lossy compression scheme is not noticeable."

Lossy compression schemes therefore include one or more quantization stages in which information is removed. The quantization is designed to exploit psychovisual redundancy, with the degree of loss linked to a user -set quality level, so that the level of distortion introduced is controlled.

3.2 Image structure and information content

Entropy is a measure of the information content of an image. The first order entropy, *H,* is the *average self-information* per pixel of the image where [5]:

$$H = \sum_{i=0}^{L-1} P_i \, log_2 \, \frac{1}{P_i} \; bits \qquad (3.2)$$

Or alternatively:

$$H = -\sum_{i=0}^{L-1} P_i \, log_2 \, P_i \; bits \qquad (3.3)$$

Entropy may also be calculated for blocks of symbols or pixels; for a block of length b, the entropy equals b times the entropy of a single symbol. P_i is approximated by the normalised image histogram.

Information capacity is a measure of the maximum amount of information that may be transmitted from an information source per symbol or group of symbols [58]. The information capacity, C, of an image containing m pixels per unit area is:

$$C = m \, log_2(L) \; bits \; per \; unit \; area \qquad (3.4) \; [58]$$

The maximum amount of information will be conveyed if all outcomes from the information source are equally likely. In the case of an image containing L possible pixel values, if all pixel values have equal probabilities, then all P_i= 1/L, which sum to 1, and equation (3.3) becomes:

$$H = \; \log_2(L) \; bits \; per \; pixel \qquad (3.5)$$

In this case, entropy is at a maximum and is equal to the information capacity. It is assumed in this expression that all image values are independent of one another. When the values are unequally distributed (as can be seen by peaks in the image histogram), for example when only a few values have a very high probability, the entropy decreases (an image containing pixels with all the same value has an entropy of 0).

Shannon [90], in his *noiseless coding theorem*, defines the theoretical limit for compression, using variable length coding alone. Based upon the assumption of the image as a zero-memory information source, Shannon defines the lower bound for L_{ave} as equal to the first order image entropy *H.* Variable rate coding methods are also termed *entropy coding*, and are used as a final stage in almost all compression, lossless or lossy. The difference between the information content and the information capacity defines the amount of redundancy within the image and hence the possible compression that may be achieved from coding redundancy alone.

Tavokoli in [91] describes two different measurements of the information content of a source; in terms of either its statistical properties or its predictive properties. The statistical properties refer to the independent probabilities of source values; first order entropy is a statistical property measure but cannot give a measure of the predictability attached to the values.

As described above, there are many types of correlation existent in natural images. Correlations indicate predictability; some values are predictable based upon knowledge of preceding values, therefore they carry less information. Higher order entropies (which consider the probabilities of groups of values occurring sequentially, rather than independent values) are predictive measures [91]. Second order entropy, for example, is calculated as follows:

$$H_2 = H(A,B) - \sum_{i=0}^{L-1}\sum_{j=0}^{L-1} P(A_i,B_j)\, log_2\, P(A_iB_j)\ bits \qquad (3.6)\ [91]$$

Where (*A,B*) are two-value sequences of random variables $A_i$and B_j, and $P(A_iB_j)$ is their joint probability density function. Second, and higher order entropy values are found to be consistently lower than first order entropy in natural images [91], because there is more redundancy in an image if values can be predicted from their neighbours.

Entropy, and particularly higher order entropy, can provide measures of the structure within an image. A high value of first order entropy indicates many pixel values evenly distributed. A low value indicates that some pixel values are more probable than others and therefore more compression is possible. Higher order entropy can be a measure of the predictability of image structure. Pixel values are least correlated and predictable in an image of random noise. High frequency areas in images, and fine random texture are less correlated than uniform areas, and low frequency tonal gradations, such as those found in areas of sky in landscapes, or in skin tones in portraits. Texture can also be predictable, if it follows a regular pattern. As well as being used extensively in the design of compression algorithms, entropy and associated measures can be used in the analysis of images, to predict both the degree of potential compression, and image areas most susceptible to information loss.

3.3 Transform based Lossy Compression

A well-developed class of lossy compression algorithms, these methods rely upon the transforming of image data from the spatial to the frequency domain. This allows the data to be grouped according to frequency and orientation. Colour images may also be transformed at a pre-processing stage into a luminance-chrominance colour space to better exploit the redundancy in the chrominance channels (taking advantage of the fact that the human visual system is less sensitive to colour discrepancies than to changes in tone), and to provide better, more channel-specific models matched to human contrast sensitivity [14].

The frequency transform stage itself is usually lossless with the loss incurred when the frequency information is quantized. The final stage involves (usually) lossless encoding, which employs both interpixel and coding redundancy to achieve further data compression.

A generalised model of the compression-decompression process is shown in Figure 3.1

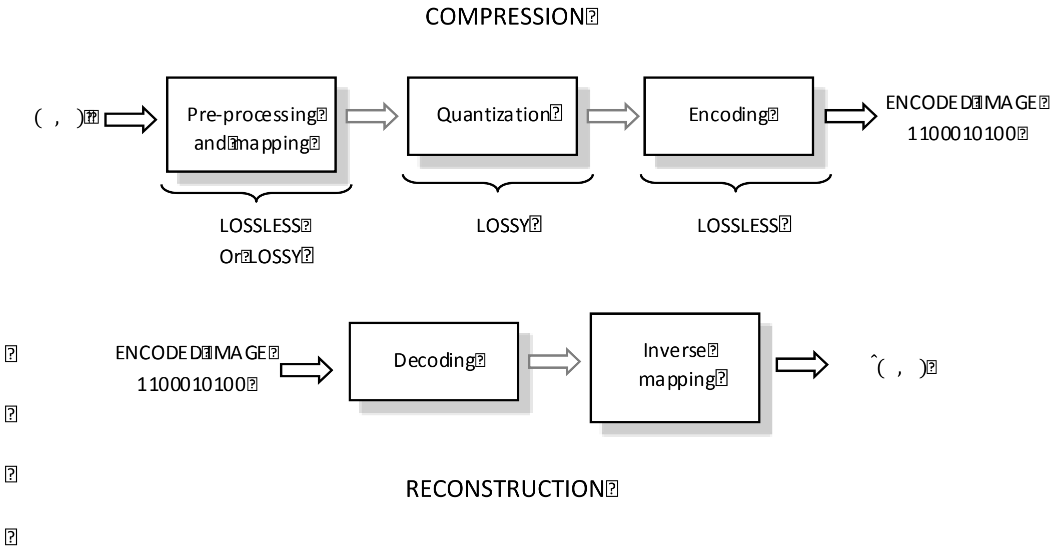


Figure 3.1 Generalised model of image compression and reconstruction, adapted from [4]

The JPEG and JPEG 2000 compression schemes are well known examples of transform-based compression. They are described briefly here to offer some insight into the nature and characteristics of the distortions that they introduce. More in-depth treatment can be found in [9] [92] [93] and in numerous documents available upon the JPEG committee webpage at [8].

3.3.1 JPEG Compression

The JPEG standard was the first international digital image compression standard for continuous tone still images [9]. It was developed from 1986 [8] by the Joint Photographic Experts Group committee, which is a collaboration between the International Organisation for Standardisation (ISO), and the International Telephone and Telegraph consultative committee (CCITT), (which is now the International Telecommunications Union's Telecommunication Standardization Sector (ITU-T)). The standard was developed in response to advances in digital technology and the need for a reduction in image file sizes with minimum loss of visual quality.

The latest version was standardised in 1994 [8] and the specification has five parts, and a number of different modes of operation, including a lossless mode based upon predictive coding and a lossy progressive encoding mode. However it is the lossy baseline sequential mode that has proved to be the most widely used. Baseline JPEG applies a forward DCT to 8 x 8 pixel blocks or sub-images, followed by quantization and encoding, as illustrated in figure 3.2.

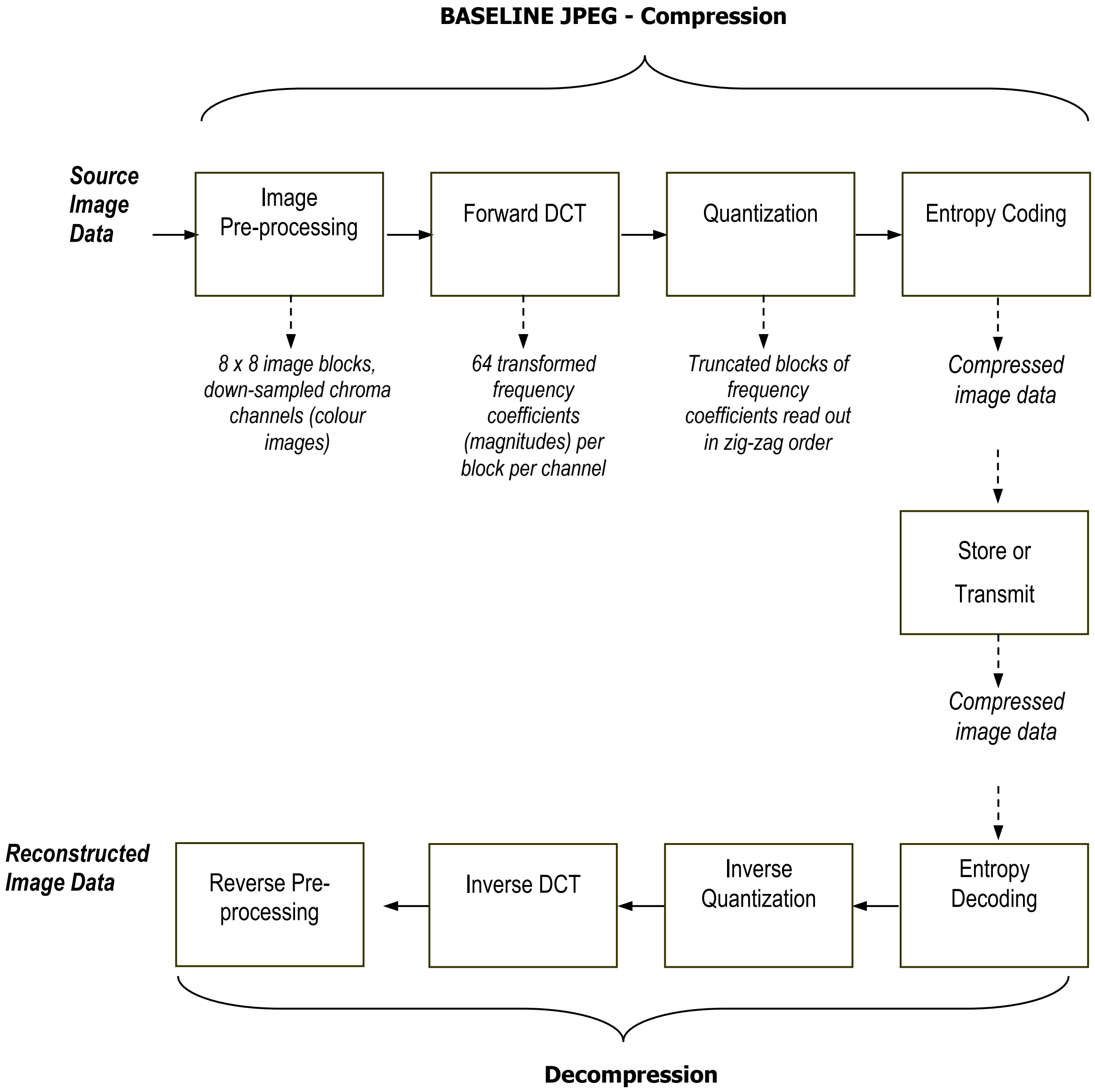


Figure 3.2: Stages in JPEG compression algorithm. Items in italics indicate output from each stage of compression process. Adapted from a diagram by Wallace [9]

Baseline JPEG allows lossy compression rates of up to 100:1 (although achievable compression rate is very scene dependent). This incurs some information loss, however the perceptibity of such loss is minimised by the architecture of the standard. In colour images, image data is converted to a luminance-chrominance colour space, YCbCr, and chrominance samples are down-sampled.

Following colour transformation, the Forward Discrete Cosine Transform (DCT) is applied to image blocks of 8 x 8 pixels. This produces blocks of 64 coefficients, representing the magnitudes of cosine basis functions of frequencies within that image block.

As continuous tone images tend to contain large areas of slowly changing tone, most of the information will be concentrated in the lower frequencies, as illustrated in Figure 3.3 and many of the high frequency coefficients are likely to be zero or close to zero.

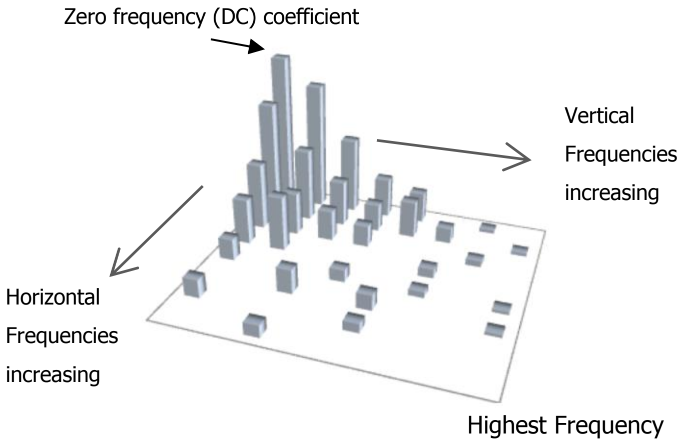


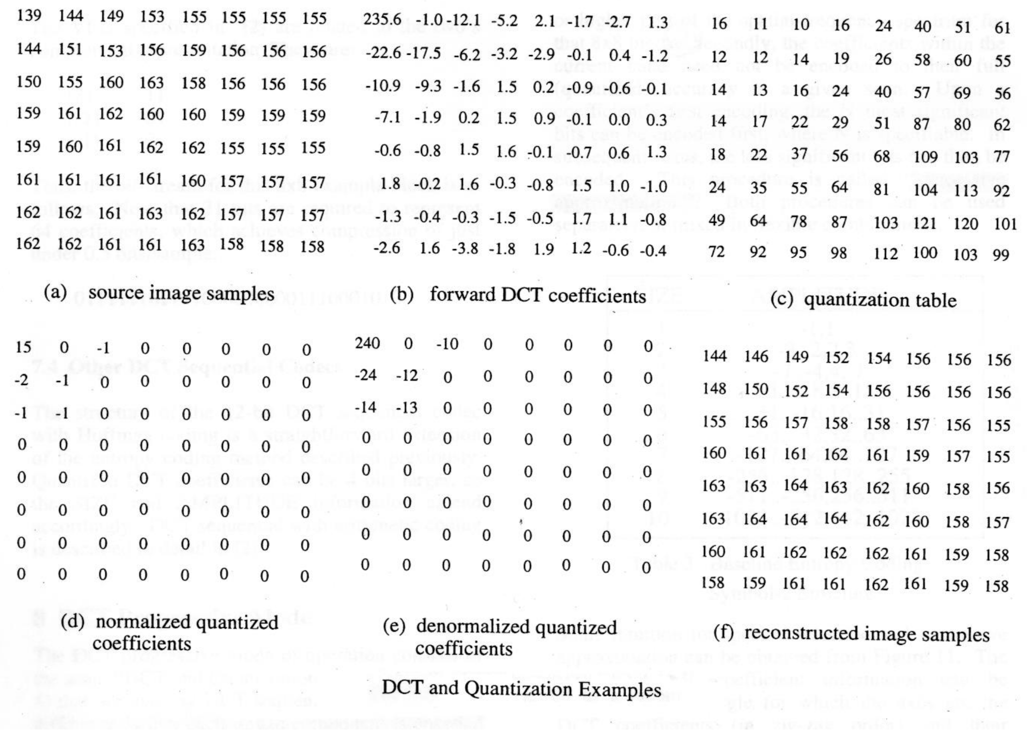
Figure 3.3: 3-dimensional representation of typical layout of coefficient magnitudes from a Discrete Cosine Transform after reordering. The zero frequency component at the top left, has the highest magnitude and is surrounded by low frequency coefficients. The frequencies increase in a diagonal zig-zag, to the highest frequency component in the bottom right. The highest frequency components are of very small or zero magnitudes. [5]

These coefficients are quantized using a 64-element quantization table, the values in the table being defined within the application according to the required quality level set by the user.

The quantization process is applied using:

$$F^Q = integer\ round\ \left(\frac{F(u,v)}{Q(u,v)}\right) \qquad (3.7)\ [94]$$

Where *F(u,v)* is a value from the block of frequency coefficients output from the DCT stage, *Q(u,v)* is the value from the quantization table in the same position and F^Q is the quantized output value. An example quantization table is illustrated in figure 3.4 (c). Because the highest values in the quantization table are in the position of the highest frequency coefficients, the division followed by rounding sets these coefficients to zero.



139	144	149	153	155	155	155	155
144	151	153	156	159	156	156	156
150	155	160	163	158	156	156	156
159	161	162	160	160	159	159	159
159	160	161	162	162	155	155	155
161	161	161	161	160	157	157	157
162	162	161	163	162	157	157	157
162	162	161	161	163	158	158	158

(a) source image samples

235.6	-1.0	-12.1	-5.2	2.1	-1.7	-2.7	1.3
-22.6	-17.5	-6.2	-3.2	-2.9	-0.1	0.4	-1.2
-10.9	-9.3	-1.6	1.5	0.2	-0.9	-0.6	-0.1
-7.1	-1.9	0.2	1.5	0.9	-0.1	0.0	0.3
-0.6	-0.8	1.5	1.6	-0.1	-0.7	0.6	1.3
1.8	-0.2	1.6	-0.3	-0.8	1.5	1.0	-1.0
-1.3	-0.4	-0.3	-1.5	-0.5	1.7	1.1	-0.8
-2.6	1.6	-3.8	-1.8	1.9	1.2	-0.6	-0.4

(b) forward DCT coefficients

16	11	10	16	24	40	51	61
12	12	14	19	26	58	60	55
14	13	16	24	40	57	69	56
14	17	22	29	51	87	80	62
18	22	37	56	68	109	103	77
24	35	55	64	81	104	113	92
49	64	78	87	103	121	120	101
72	92	95	98	112	100	103	99

(c) quantization table

15	0	-1	0	0	0	0	0
-2	-1	0	0	0	0	0	0
-1	-1	0	0	0	0	0	0
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0

(d) normalized quantized coefficients

240	0	-10	0	0	0	0	0
-24	-12	0	0	0	0	0	0
-14	-13	0	0	0	0	0	0
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0

(e) denormalized quantized coefficients

144	146	149	152	154	156	156	156
148	150	152	154	156	156	156	156
155	156	157	158	158	157	156	155
160	161	161	162	161	159	157	155
163	163	164	163	162	160	158	156
163	164	164	164	162	160	158	157
160	161	162	162	162	161	159	158
158	159	161	161	162	161	159	158

(f) reconstructed image samples

DCT and Quantization Examples

Figure 3.4: An example of a 64 pixel block going through the different stages of the baseline JPEG algorithm, from [9]

The higher the quality setting, the more coefficients will be retained. The quantization tables are designed perceptually, based upon the spatial frequency response of the Human Visual System, the Contrast Sensitivity Function (CSF), giving more weight to the frequency components that are more visually important.

Quantization is followed by (or combined with) a reordering of the coefficients into a zigzag sequence, with the zero frequency component (DC coefficient) in the top left hand corner and the highest frequency component in the bottom right (figure 3.5). The DC coefficients are losslessly encoded separately from the other coefficients using differences between the coefficients of consecutive blocks (DPCM), to ensure that the mean intensity value of each block, which is proportional to the DC coefficient, is maintained. The remaining AC coefficients are entropy coding using either Huffman coding or arithmetic coding. The resulting bit stream may be stored or transmitted as required.

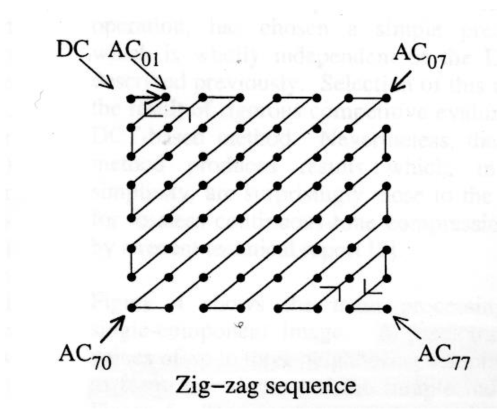


Figure 3.5: Zigzag rearrangement of DCT coefficients from a single 64 x 64 pixel block in baseline JPEG. The DC value corresponds to the zero frequency, and the AC coefficients are numbered from lowest to highest frequencies horizontally and vertically. From Wallace [9]

The decompression stages of the algorithm produce a reconstructed and viewable image. Each stage of the compression is reversed, as shown in the bottom row of figure 3.2. True inverse quantization is not possible as many of the coefficients will have been truncated to 0 and therefore the resulting frequency coefficients will be an approximation of the original. Additionally, if the chroma channels were down-sampled at the pre-processing stage, then upsampling them at the final stage may cause colour distortion due to interpolation and rounding errors. The artefacts produced by JPEG are introduced later in this chapter and in more depth in chapter 4.

3.3.2 JPEG 2000

The growing requirements of technologies and applications producing and using digital imagery, in particular the expansion of the Internet and multimedia applications, prompted a call for contributions to a new image compression standard in 1997. Areas in which JPEG and other image standards had failed to deliver were to be addressed with certain requirements of the new standard. JPEG2000 Part 1 was standardised in 2000 [95] with the following features [93]:

- *Superior rate distortion and subjective image quality performance at low bit rates, to that of existing standards.* This is a key requirement of network image transmission and remote sensing applications.
- *The ability to encode different types of images.* The ability to compress bi-level, grey scale, colour and multi-component images is a feature of JPEG2000. This allows the compression of documents containing both images and text. JPEG suffers from artefacts when compressing binary images and is not successful at compression of text in particular. This expands the range of possible applications for JPEG2000.
- *The ability to encode images with different characteristics.* For example, natural images, images from scientific and medical applications, images containing text or computer graphics. These applications have very different requirements from a compression algorithm; therefore flexibility of operation is one of the key features of JPEG2000.
- *Lossless and lossy encoding.* This allows the use of JPEG2000 by applications such as medical imaging where lossless reconstruction is required. Progressive lossy to lossless decompression means that an image may be compressed losslessly, but then decompressed to required lossy compression levels or quality levels. Effectively, a single compressed version of the image may be used in multiple contexts.
- *Progressive transmission by pixel accuracy or spatial resolution.* This is particularly important for image archives and web-browsing applications.
- *Robustness to bit errors.* This is important for transmission over wireless communication channels.
- *Special features to improve flexibility*, such as region-of-interest coding and protective image security.

JPEG2000 images do not suffer from blocking artefacts characteristic of JPEG unless the image has been tiled. The decomposition of an image into frequency *sub-bands* using a wavelet transform is quite different to the structure of block based DCT transformation used in the JPEG architecture and is illustrated in figures 3.6, 3.7 and 3.8. The wavelet transform decomposes the image by dividing a set of image samples into different sub-bands of down-sampled high-pass and low-pass samples. This is achieved using 1D filters, termed *wavelet* filters (giving the high-pass output) and *scaling* filters (giving the low-pass output) in horizontal and vertical directions. The result is shown in figure 3.6. Each image in the top left quadrant of a sub-band is one quarter of the size of the original (having been down-sampled in both the horizontal and vertical directions) and contains the coefficients representing the low-pass output. On the top right is the high frequency information in the horizontal direction, bottom left are high vertical frequency and bottom right relates to high diagonal frequencies. These four images make up one 'scale' of the original image.

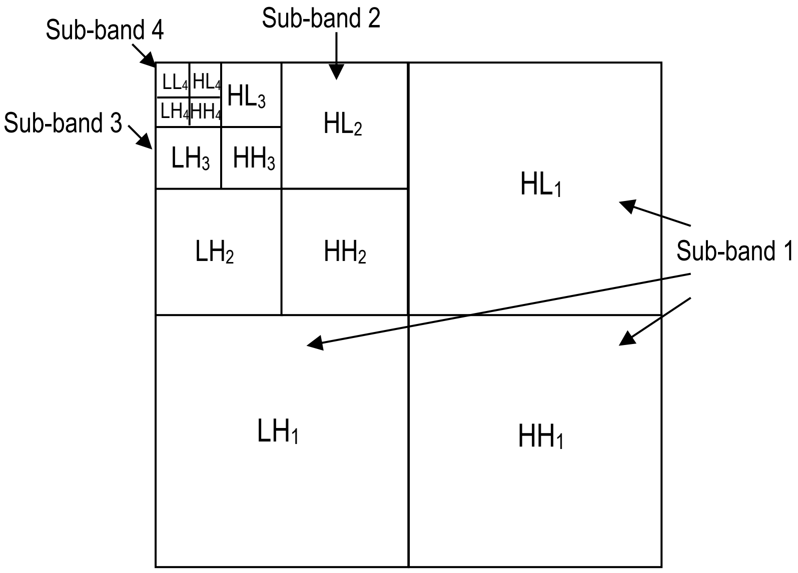


Figure 3.6 A four scale sub-band wavelet decomposition, based on a diagram from [96]

At each level, the process is repeated iteratively on the low-pass samples. This recursion continues to further levels, the final output consisting of a small block of low-pass samples at the lowest resolution in the top left, with the remainder of the image made up of high frequency detail coefficients at different resolutions. An example of the effect on an image is shown in figure 3.7. The low-pass output may be seen to be a version of the original image at a lower spatial resolution.

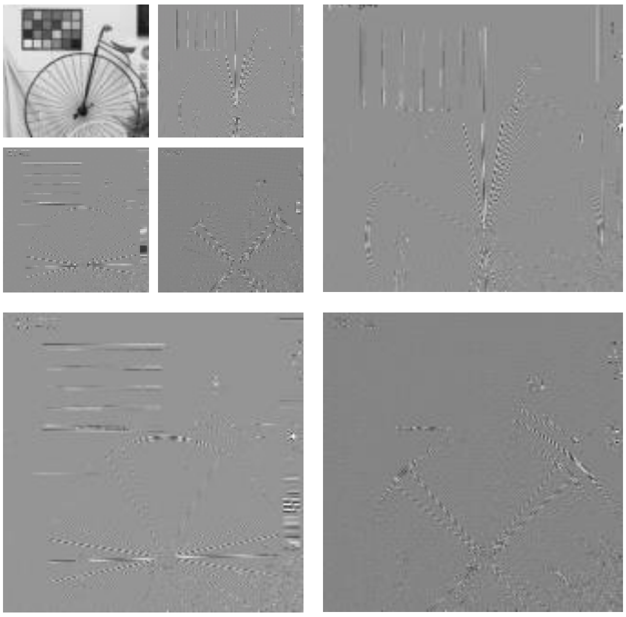


Figure 3.7 A three scale wavelet decomposition upon an image © David Taubman UNSW, from [92] The low pass version of the image is top left. All other squares contain high frequency components at different scales.

Following decomposition, the sub-bands are quantized separately. The image is then further sub-divided before entropy coding into precincts and code blocks. The inputs into the entropy coder are the code blocks by bit plane, from a precinct scanned in raster order [93] [95]. A precinct, identified in figure 3.8, consists of three blocks of code blocks, one from each sub-band at that resolution level. The diagram above illustrates the difference in the structure of the two algorithms. Where the values in a reconstructed block within a JPEG compressed image will be dependent upon the frequencies present at that spatial location within the original image and the quantization table selected, the reconstruction from the JPEG2000 image is made up of code blocks from precincts in the same spatial position relative to the edge of a particular sub-band, from all the different sub-bands. Each sub-band may be viewed as a version of the image at a different scale or resolution. Because the quantization step-size is different in different sub-bands, errors build up in a very different way, being much less uniform over a spatial location by comparison with JPEG blocks.

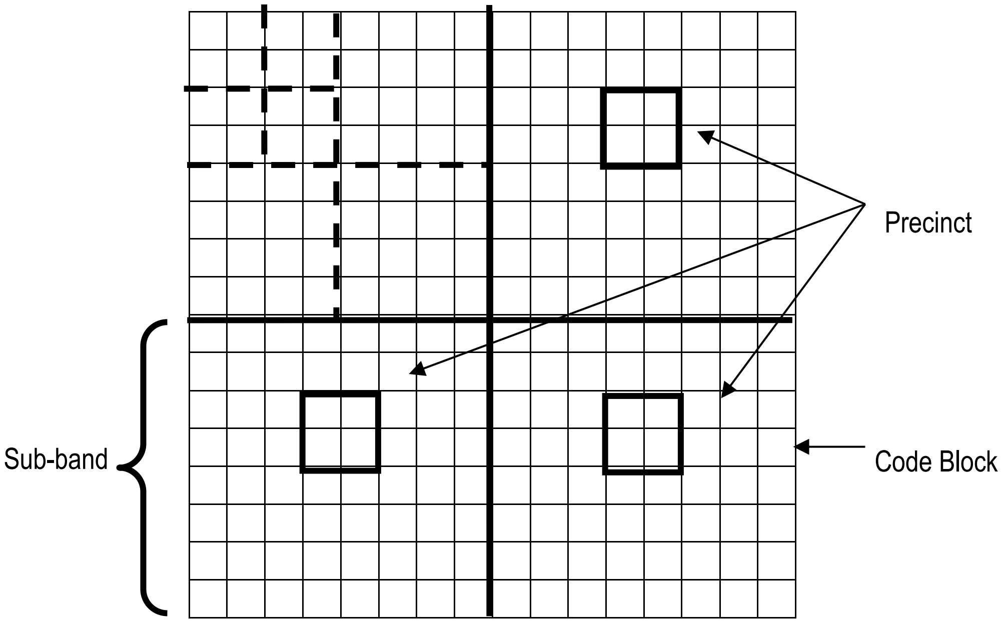


Figure 3.8 Diagram illustrating the partition of a tile or image component into code blocks and precincts (based on a diagram by Skodras et al. [93]

3.4 Quantifying Distortion

Measuring the performance of a lossy compression scheme requires some quantification of the degree of loss, in determining the usefulness of the system. Error measures can be used to understand the balance between information loss and compression rate.

Rate-distortion theory, derived from Shannon's work on information theory [90], defines a relationship between compression rate and some measure of distortion. Rate distortion theory can be used in the development of lossy compression systems as a basis for optimizing their performance. It establishes theoretical limits for an achievable compression rate based upon minimal entropy, without exceeding a given maximum level of distortion. The relationship can be expressed in a *rate-distortion function*. The distortion measures below are some of those most commonly employed [5].

Mean Absolute Error (MAE) is a measure of the average amount of error per pixel across the image:

$$MAE = \frac{1}{N}\sum_{n=1}^{N}|x_n - \hat{x}_n| \qquad (3.8)\ [7]$$

Where N is the number of pixels in the image, x_n is the value of the pixel at position n in the original image, and $\hat{x}_n$ is the value of the pixel at position n in the compressed image. The other distortion measures below are adaptations of the MAE.

Mean Squared Error (MSE):

$$MSE = \frac{1}{N}\sum_{n=1}^{N}(x_n - \hat{x}_n)^2 \qquad (3.9)$$

Signal to Noise Ratio (SNR) compares the power of the original signal, *P,* with the MSE:

$$P = \frac{1}{N}\sum_{n=1}^{N}(x_n)^2 \qquad (3.10)$$

$$SNR = \frac{P}{MSE} \qquad (3.11)$$

SNR is often expressed in decibels (dB) as:

$$SNR(dB) = 10log_{10}(SNR) \qquad (3.12)$$

Peak Signal to Noise Ratio, (PSNR), defines the ratio between the theoretical maximum power (defined by the bit depth, $2^{\text{no. of bits}}$-1) of the signal and the noise in terms of MSE. It is expressed in decibels as follows:

$$PSNR(dB) = 20\ log_{10}\ \left(\frac{max}{\sqrt{MSE}}\right) \qquad (3.13)$$

These measures are simple to compute, requiring only the compressed image and the original, but are not good at predicting perceived image quality performance. Their limitation is that they generalise error; all distortions are treated equally. They provide a measure of the magnitude of the difference between two images without considering the type of error or its context.

For example [97], it is possible to obtain the same value for MSE from calculations between an original image and a JPEG compressed version of it; or a version that has been blurred; as the MSE value calculated from a version that has been rotated; or where the contrast has been increased. The first two processes usually result in quality loss; rotation is unlikely to affect quality unless aliasing is introduced; contrast is a preferential attribute and might actually result in an increase in perceived quality if the original image was too low in contrast to start with. Furthermore, the viewing conditions, observer demographic and quality criteria, the purpose of the image, and the nature and characteristics of the original do not affect the MSE value, yet all are important aspects of the 'image quality context'.

Distortions are not equal in their perceptual effect for various reasons, as discussed later in relation to scene dependency; some distortions are more bothersome than others; most distortions vary in terms of their visibility depending upon the spatial configuration of the scene, and some can mimic preferential attributes at certain levels on specific types of image content.

Wang and Bovik [97] explain this by highlighting a number of assumptions made about perceived image quality that are implied by the use of distortion metrics as predictive measures:

1) That perceptual image quality is unaffected by the spatial relationships between image elements; in effect pixels are spatially independent from one another. There is no consideration of image structure and content.
2) The relationship between original signal and error signal is unimportant; it does not account for the effect of the error on the image, which depends upon the structure of the original.

3) Distortion is measured as a magnitude, without direction. Perceived quality will be affected by the way in which a signal changes, as well as by how much it changes.
4) Image samples are equally weighted in terms of visual importance.

These assumptions break down as soon as the characteristics of real scenes and the response of the human visual system to them are considered.

3.5 Scene Dependency and Image Quality

Scene dependency may be defined as a variation in subjective image quality evaluation, which is directly related to the scene content. This means that two images of different scenes, which have the same or similar values for physical characteristics but different scene content, may be scaled differently in subjective image quality, or in terms of particular perceptual attributes. Scene dependency represents a challenge in image quality quantification.

Engeldrum describes the dependence of judgements of images upon the *spatial configuration* of the images [18] and defines spatial configuration and object content as important ‘context factors’. This description of scene dependency is usually defined as independent from the aesthetic aspects of the images, and is specifically in relation to preferential and artefactual attributes. Of course, it is very difficult to eliminate observer bias based upon scene content. Engeldrum [18] describes the ‘emotional involvement, or potential emotional involvement’ of observers with scene content as another context factor, but suggests that a potential strategy is to include images from various different sub-classes (for example, portraits, landscapes, abstract) to reduce the effect of observer bias.

3.5.1 Sources of Scene Dependency

Triantaphillidou [34] identifies three different sources of scene dependency affecting subjective image quality in image compression:

- *Scene Dependency as a result of observers’ quality preference* An example is the difference in sharpness preference between architectural and portrait scenes [ref], whereby portraits may be considered acceptable if slightly blurred, whereas architectural scenes are generally preferred slightly sharper overall.

- *Scene dependency due to variable visibility of an artefact in some image areas compared to others.* This may be the result of *masking effects,* whereby the image content masks the particular artefact (for example, areas of fine detail can mask noise, and noise can mask contouring artefacts [98] [40]). Alternatively it may be as a result of an artefact being more or less visible as a result of visual adaptation.
- *Scene dependency of digital processes or image processing algorithms.* Sometimes termed *scene susceptibility.* Many algorithms are designed to selectively enhance or suppress specific image characteristics. Image compression is a particular example. As described in the previous sections, lossy transform based algorithms tend to remove or reduce both colour resolution and high frequency magnitudes. Therefore scenes with more of this type of content will compress less well and will contain more artefacts when compressed to the same level as a scene with fewer high frequencies, for example.

3.5.2 Scene Dependency and Compression

As described in 3.2 the information content of an image is fundamental to compression performance. The structure of the image determines the predictability of image content, which in turn defines the amount of redundancy (spatial, coding or psychovisual) within the image. Because compression is based on the elimination of redundancy, it follows that it is inherently scene dependent. This means that when using a particular lossless compression method, images of the same original file size with different scene content, will compress by different amounts [5]. Conversely, in lossy compression, the two images when compressed to the same file size will differ in image quality.

Because lossy compression removes information as well as reducing data, the distortion introduced will be dependent upon the scene content; specifically the interaction between the scene content and the particular compression mechanism. Different lossy compression schemes introduce their own distortions; so the scene dependency varies depending upon the algorithm. The artefacts introduced by complex algorithms tend to be very specific and recognisable, and this can make them less acceptable than artefacts that are more general and from a range of different sources. Certain types of image content can also mask or enhance the artefacts.

Finally, and perhaps the most important point, distortions may affect observers' quality criteria. In particular, observers may become habituated to certain types of artefacts over time, if they are exhibited by widely adopted algorithms and processes, and this may have the potential to increase perceived image quality, when compared to an artefact that is newer, less familiar and perhaps therefore more apparent. It does not seem unreasonable to suggest that such habituation may have happened, for example, in response to the blocking artefact characteristic of JPEG, particularly because it is so prevalent in compressed moving images from the MPEG algorithm, but also because JPEG has been in use since 1992 [94].

Scene dependency is a particular area of interest in compression research, not least because it can help to explain why algorithms do not perform as might be predicted. Characterisation of images in terms of spatial configuration and an understanding of scene dependent interactions can help to determine the suitability of an algorithm for a particular class of image if it has a limited range of specific characteristics (for example in images of fingerprints, which have a lot of high frequency content).

Keelan [99] suggests that in the case of artefactual rather than preferential attributes, variability in results across observers occurs due to their sensitivity to the artefact. Determination of this sensitivity is more straightforward than it is for a preferential attribute, because preference is likely to be much more variable across observers. By definition, an artefact is a defect; therefore as all observers will find the artefact bothersome in the majority of images, optimum quality becomes the threshold point at which the artefact is not visible. In this case, variability across scenes is entirely a result of scene susceptibility to the digital process. Keelan proposes the quantification of quality loss for subsets of observers and scenes to obtain a better fit with psychometric data [100], grouping according to observer sensitivity and/or scene susceptibility (see section 6.6.3).

3.5.3 Scene selection

Various approaches may be used to reduce the bias introduced as a result of the susceptibility of certain scenes to either distortion or the visibility of the distortion, by averaging results over a range of different types of scenes, or by limiting the characteristics of the scenes used in the evaluation. These strategies are entirely appropriate to ensure that the results reflect the performance of systems under average conditions for most scenes.

Engeldrum [18] notes that the 'selection of image samples is governed by the objective of the scaling study'. As pointed out by Keelan, [99] [101]this is particularly true if the study is evaluating an artefactual attribute; where scenes have different levels of susceptibility they should be selected to ensure the inclusion of scenes spanning the full range of artefact visibility, from those where the artefact is well masked to those where it is emphasised by scene content. The distortions in lossy compression are a case in point. Keelan also suggests that it does not matter if selection of scenes to emphasise or suppress certain characteristics means that the sample set is not representative of the overall population of 'customer images', if the aim is to 'improve the signal in a psychometric experiment', because this will mean that the results will be more able to test the efficacy and accuracy of an objective metric, assuming that the metric is developed for such a restricted sample set.

Bartleson in [102] proposed five categories for sample image selection, as summarised in table 3.1. The categories represent a narrower and more specific approach to image selection as they move down the table. All of the categories are used in product development, dependent upon product and context. The ISO has developed reference sets of images in ISO 20462 part 3 [85], the Standard Reference Set (SRS) of hard copy images and the Digital

Reference Set (DRS) of digital images for use with the quality ruler, which are examples of incident samples (to be made available through www.imaging.org [38]). In exploring scene dependency, it useful to select a purposeful set of samples, by evaluation of various attributes deemed to influence image quality preference. Approaches to characterising images for this type of sample selection, are discussed in section3.7.

Sample Category	Properties and Characteristics
Random and Independent	Images have equal chance of selection; Selection of each image independent of selection of others.
Stratified	Population classified in terms of distinction of interest, and numbers in each class reflect population statistics. Common in psychometric experiments.
Contrast	Stratified, with extra images in classes of interest, common in product development testing. Ignores irrelevant classes.
Purposeful	Represents a specific population OR varies independently in some attribute. Also common in product development.
Incidental	Random sample representing sub-groups of particular interest OR special (unique) existing collection (e.g. ISO standard images) Reference sets; 'sacred samples'.

Table 3.1 Sample Categories by Bartleson [102], adapted from Engeldrum [18]

3.6 Digital Imaging Artefacts

The objective characterisation of digital imaging systems and processes is more complex than that of analogue, because, in addition to the physical attributes described in chapter 2, a number of digital artefacts (detailed in table 3.1) are introduced, as a result of the various processes involved in digitisation and throughout the imaging chain.

As described by Keelan [103], these artefactual attributes are not always evident in an image (unlike a preferential attribute such as colourfulness), but when present, they are nearly always detrimental to image quality. However, in previous work by the author [33], detailed in chapter 4, it has been found that the presence of small amounts of particular artefacts sometimes appears to enhance subjective image quality. Such cases, although isolated, are interesting examples of scene dependency, resulting from the interactions of the algorithm with particular scene characteristics. It should be noted that the visual effect of the artefact (*ringing* as a result of JPEG compression) mimics that of a preferential artefact (sharpening).

The common digital artefacts are detailed in table 3.1, along with scene areas which are either more susceptible or in which they are more visible. The final column details some of the interactions with other image attributes, which may serve to emphasise or mask them, and some potential methods to reduce or correct them.

3.6.1 Variation of artefacts across an image

Artefactual and preferential attributes may affect an image globally or locally [101]. Global artefacts are equal in magnitude and character across the image plane, both objectively and perceptually, (colour misregistration in large amounts is an example, see table 3.1). But many artefacts (and preferential attributes) are variable within an image, meaning that they affect the image plane unequally. This variability, as noted in section 3.3.1, is a significant source of scene dependency. Keelan describes this variation as dependent on one or more of three factors:

- Signal level (many attributes vary as a function of image intensity);
- Location (for example radial position)
- Orientation or direction

Artefacts can be localised perceptually, whilst affecting the image globally, meaning that they are linear in an objective metric space, but non-linear in a perceptual space. Therefore the visual system has a variable response to them across the image plane, which is often dependent upon signal level. Keelan [101] describes the Detail Visibility Function (DVF) which characterises the visibility of certain artefacts (for example noise, streaking and banding) weighting them in relation to the signal level. Thus, although they are distributed equally across the tonal range of the image, their visibility is affected by the visual density response to the artefact.

Artefact	**Causes**	**Scene Susceptibilities**	**Masking/Interactions/ Reduction**
Contouring	Poor quantisation; multiple colour conversions; chroma sub-sampling	Large areas of slowly varying tone or colour (low frequencies)	Strong interaction with noise
Jaggedness /Pixelisation	Insufficient spatial resolution; aliasing from down-sampling	Diagonal edges and lines	Can be masked by blurring or noise; dithering of edges may reduce visibility
Aliasing	Sampling	Areas with periodic high frequency information (high frequency lines)	May be masked by random noise. Pre-filter prior to down-sampling to limit bandwidth
Chromatic aliasing	Differential sampling of colour channels, resulting in differences in Nyquist frequencies of RGB channels	Areas with periodic high frequency information (high frequency lines), colour moiré more apparent in neutral areas	May be masked in highly chromatic areas
Blocking	DCT compression	Areas with high frequency information; More visible in uniform and slowly varying areas	May be masked by fine detail
Smudging	DWT compression	Edges and lines; areas of texture and high frequencies.	
Colour bleeding	DWT compression	Adjacent colour areas; More visible in relatively uniform low chromatic areas	
Ringing	Abrupt truncation of frequencies. Examples include ideal filtering in the frequency domain or DCT compression	Appears as an echo or a ripple around edges and lines	Use of more gradual frequency attenuation filters; windowing
Halo artefact	Exaggerated 'overshoot' and 'undershoot' at edges caused by digital sharpening in the spatial domain	Appears as a halo around edges and lines. Can appear similar to the ringing artefact	Affected by (but not the same as) sharpening[3]
Patterning	Dithering	Slowly varying areas except for pure blacks and pure whites	May be masked by noise
Streaking	Pixel-to-pixel non-uniformity in linear arrays (mostly in digital printing devices)	Uniform areas, slowly varying areas	Weak interaction with noise
Banding	Periodic variations in digital printing devices	Uniform areas, slowly varying areas	
Colour Misregistration	Spatial shift between colour records of an image. A global artefact.	Small amounts: edges, lines may appear unsharp; large amounts, ghosting in all areas	

Table 3.2 Digital imaging artefacts, scene susceptibility and masking (adapted from Triantaphillidou [34] [40])

[3] Oversharpening and sharpness are distinct from each other. The sharpness of an edge is determined by the steepness of the slope of the edge (its spatial extent orthogonal to the edge direction); oversharpening is an artefact, the effect of which is determined by the magnitudes of the overshoot and undershoots. However once oversharpening is present, it will be affected by any processes that affect image sharpness

While individual artefacts may be detrimental to image quality, their interactions with other attributes may mask their presence, and some processes within the imaging chain are introduced for precisely this reason. Finally, some artefacts are highly localised in objective space and therefore unpredictable in their effects upon quality and their visibility. The smudging artefact in JPEG 2000 is an example.

For attributes that are variable as a result of position, some form of location (field position) weighting may be applied within metrics. Keelan and Jin in [104] found that subject matter that particularly affected sharpness was often positioned closer to the centre of the image, whilst judgements of noisiness were more affected by scene content in the periphery of the image. If the positional variability is well defined and consistent in an objective space, it can be relatively straightforward to divide the field position into various regions and weight them according to their areas. A similar approach can be used for orientation dependent attributes or measures. An example is the modulation transfer function, which for lenses is positionally dependent (optical performance is best at the centre of the lens), and for sensors is often orientation dependent, due to different horizontal and vertical spacing and dimensions of sensor elements.

3.6.2 Visual saliency

While the discussion above is in relation to measures or attributes changing in objective space as a result of spatial position, field position can be linked to the idea of *saliency,* which is more of a reflection upon the visual response to particular image features and is dependent upon their position within the image. Saliency is a term used in computer vision to identify features or locations within an image that are more distinguishable from their surroundings, or are more visually important [105]. It is not difficult to imagine that if a distortion, particularly a compression artefact is both bothersome and very visible, in a particular area within an image, that area will become a salient feature.

Kadir and Brady [105] describe visual saliency as referring to the idea that ‘certain parts of a scene are pre-attentively distinctive’. This is based upon work from Neisser [106] from 1964, which divides the early stages of human vision into pre-attentive and attentive stages. Pre-attentive vision identifies key features (so-called *salient features*) within a scene; so any features that are distinctive will be picked out. Attentive vision follows and takes in the rest of the scene, finding relationships between the features and grouping objects and attributes for interpretation. This is analogous to the ordering of the two stages of vision proposed in Hochstein and Ahissar’s Reverse Hierarchy Theory [30], described in section 2.2.

Saliency is linked to the concept of *foveated vision* [107], which describes the variable resolution of the retinal image, produced as a result of the unequal distribution of photoreceptors and ganglion cells on the retina. The fovea is the point on the retina coincident with the optical axis and contains the highest density of cone receptors, which are responsible for visual acuity. The retinal image is sharply focused in this area, subtending 2 degrees of visual angle [108] and image sharpness gradually decreases away from the optical axis. To create the final perceived fully sharp image, the eye undergoes small involuntary jerking movements known as *saccades* to direct the fovea to different areas of the scene, lead by objects of interest, or salient points.

Saliency feature extraction attempts to identify and extract salient areas within an image, which may then be used for weighting attributes in metrics. Saliency metrics can be tested by identifying the spatial distribution of important subject matter in an image, for example from experiments with observers using eye tracker technology to track visual attention.

3.7 Scene Classification

Scene classification has long been a subject area of interest to researchers from various disciplines; imaging science can use an understanding of the characteristics of images and imaging systems combined with the human visual response to guide this exploration.

Scene classification has significance in a number of different contexts in imaging science; its relevance to the results of image quality studies [22] [34] [33], its usefulness in classifying images on the internet [109], and in image taxonomy for image content based retrieval systems [110] are just a few. The approaches used vary widely in their complexity depending upon discipline and purpose. For image quality studies it is useful to explore methods that are relatively simple to implement and analyse, but that correlate well with human perceptions of quality (or other attributes). Many approaches use simple first order or second order statistical measures applied to luminance or chrominance data, or simple feature extraction approaches to edges or texture.

Experimental work by Triantaphillidou in [34], investigates scene classification for image quality, exploring the issues of scene susceptibility and scene dependency in relation to subjective image quality judgements. A number of simple scene analysis measures are used to group and classify test stimuli in relation to some of their inherent scene properties. The measures used aim to quantify global image content in relation to tone, colour and spatial characteristics. A range of measures are employed, including first order statistical measures applied mainly to the CIELAB L* channel; chroma variance derived from C^*_{ab}; a multi-stage image segmentation metric to evaluate image *busyness* (as a ratio of busy, or detailed areas to non-busy areas within the image) and two methods to determine the number of lines within the image.

The results are compared against interval scales from a paired comparison experiment, to explore correlations between low-level visual features and human image quality criteria. The paired comparison is described in chapter 4, where the relationship between the visual descriptions of image quality attributes and the scene susceptibilities of digital artefacts are explored in relation to JPEG and JPEG 2000.

Mancusi adapts this approach in [111] to develop a multi-dimensional image selection and classification system, expanding upon the range of measures used in [34] to classify images in terms of: global lightness and lightness contrast; colourfulness, colour contrast and dominant hue; and scene busyness.

Hoon [112] applies scene descriptors both globally and to local image regions of interest based upon Kadir and Brade's saliency model [105] to explore scene susceptibility with respect to sharpness and noisiness. First order statistical measures are employed to quantify tone, contrast, colour and information content. Second order statistics extract textural information. Edge detection filters are used to extract gradient images and quantify the number and strength of edges in the image, which might be susceptible to sharpness degradation. In this study, measures applied globally appear to correlate better to subjective quality evaluation than local measures, however, this may be affected by positional variations in the effects of the attributes, as described in [104].

Various scene descriptors have been used to classify colour attributes of images. Hasler and Susstrunk [113] use statistical measures of the chrominance channels of the CIELAB space to develop a number of metrics, which correlate well with colourfulness in natural images. Scene descriptors relating to colour have also been used to explore the relationship between gamut variation and JPEG and JPEG 2000 compression [114]. Scenes are classified according to the size of the image gamut, colourfulness, spatial frequency response of the chroma channels, and the number of unique colours in RGB colour space.

4 Comparison of JPEG and JPEG 2000

4.1 Background to the investigation

Since its adoption as a standard in 1992, [9] the JPEG compression scheme has become the most widely used method for the lossy compression of digital images [115]. As discussed in the previous chapter there are a number of artefacts that are characteristic of JPEG at high compression rates, most notably *blocking* and *ringing*. Nevertheless as the predominant format in, for example, camera phone applications, which probably generate the largest volume of consumer digital images globally, the JPEG algorithm is so widely used, that it might be argued that its artefacts have, to some extent, become an accepted characteristic of contemporary images.

The development of JPEG 2000 was not, therefore, primarily about failings in terms of image quality from JPEG. Both algorithms are 'perceptually lossless' at low compression rates, making them satisfactory for many applications. Key motivations prompting ongoing research to improve upon JPEG, described in the call for proposals for the new standard were to [11] 'provide an open system approach to image compression', to 'provide capabilities to markets that currently do not use compression' and to 'address areas where current standards failed to produce the best quality or performance'. This last refers in particular to the levels of distortion and perceived image quality at low bit rates, but fundamentally JPEG 2000 was designed to address the requirements of modern digital imaging and its expansion into new territories. For some of the disciplines in question, such as medical and forensic imaging, image fidelity and acceptable levels of distortion are critical.

Traditionally, evaluation of lossy compression has involved simple distortion metrics [7,14]. JPEG 2000 distortion has been assessed objectively in numerous studies, such as [96,116,117,11,93,118]. Those investigations incorporating comparisons of the performance of JPEG 2000, JPEG and other compression standards have also focused on distortion metrics [118].

However, as discussed in section 3.4 [97], because they are based upon a number of possibly questionable assumptions about image quality evaluation, distortion metrics do not always correlate well with perceived quality.

Objective measures that incorporate some model of the human visual system have been found to correlate better with subjective image quality scaling studies [22,119]. However the number of psychophysical investigations of the quality of JPEG 2000 against which to evaluate them is far more limited than those for JPEG.

A few subjective comparison studies have been implemented. Steingrimsson and Simon [120] implemented a three-image discrimination test to determine the perceptibility of distortions, with three perceptibility thresholds (66%, 75% and 90%) identified and compared for the two algorithms. In the same study, suprathreshold quality was investigated using paired comparisons of JPEG and JPEG 2000 images. Their results for both studies were found to be scene dependent.

This chapter describes experimental work evaluating and comparing JPEG and JPEG 2000. The study involved a paired comparison of subjective image quality between JPEG and JPEG 2000 to establish whether JPEG 2000 demonstrated significant improvements in visual quality [33]. The derived quality interval scales were obtained using paired comparison of images displayed under calibrated viewing conditions. A particular focus of this work was the inherent scene dependency of the two algorithms and their influence on subjective image quality results. Further work on the characterization of scene content was carried out in a connected study [34].

4.2 JPEG and JPEG 2000

It is useful to consider the aspects of the architecture of the two algorithms that lead to their specific artefacts. JPEG 2000 workflow is similar to that of JPEG (which is shown in figure 3.2). Key differences in JPEG 2000 are in the use of the Discrete Wavelet Transform (DWT) instead of the Discrete Cosine

Transform (DCT), and in the separate quantization and encoding of sub-bands rather than blocks.

The differing architectures lead to characteristic errors. Both algorithms are scene dependent, resulting in better performance on certain types of scenes than on others, producing higher compression ratios with less visible loss. In addition, the types of artefacts are more visible in some image areas than in others [33]. This work aims to compare the two compression schemes across a set of images varying in scene content and to explore scene susceptibility and scene dependency in relation to image quality.

The wavelet function has been extensively researched as an alternative to other frequency transforms for signal decomposition [121,122,123,124,125]. Wavelets have a number of distinctive properties, in that they are localised, and have translational and scaling properties, meaning that they can provide a multi-resolution representation of an image through a set of wavelets at different scales. An advantage of the DWT is that it can be applied using a filter bank, which simplifies the transform process [96] [121]. The use of a pyramidal filter bank provides the ability to encode the image at different scales. This is extremely useful in image compression, as it allows a single image to be compressed and then decompressed at different pixel resolutions as required. Additionally, the use of wavelets in compression minimises the blocking artefacts inherent in schemes such as JPEG, but comes with certain other associated distortions. At low bit rates, the nature and visibility of their artefacts and their scene dependencies result in differences in preference between the two algorithms.

In both, the frequency transformation stages decorrelate the data prior to quantization and encoding. Baseline JPEG allows lossy compression rates of up to 100:1 (although achievable compression rate is very scene dependent). The perceptibility of this information loss is minimised by the use of visually optimized quantization tables (see figure 3.4).

A detailed description of the architecture of both is provided in chapter 3; the key differences in the operation of the two algorithms at each stage are summarised in Table 4.1. The provision of the required features and functionalities expressed in the call for proposals has resulted in an architecture that is more complex in JPEG 2000 than that of baseline JPEG.

Compression Stage	**Baseline JPEG**	**JPEG 2000**
Pre-processing	Conversion of RGB image to YCbCr and down-sampling of chroma channels Division of image into 8 x 8 pixel sub-images.	Image 'tiling' (OPTIONAL) Division of image into non-overlapping image tiles. (varying sizes). Reversible or irreversible colour transformation.
Frequency Transformation	Discrete Cosine Transform, resulting in 64 coefficients representing magnitudes of different frequencies for each sub-block.	Reversible or irreversible Discrete Wavelet Transform (for lossless or lossy compression respectively). The image or image tile is decomposed into a number of 'sub-bands'. Each sub-band consists of coefficients describing horizontal and vertical frequency components at a particular resolution.
Quantization	Coefficients are re-ordered using a zig-zag sequence though *each block*. Frequency coefficients from each block are quantized using visually weighted quantization tables, resulting in the highest frequency components and lowest magnitude components being removed. This results in a string of zero magnitudes for the highest frequency coefficients at the end of each block.	*Sub-bands* of coefficients are quantized separately using a uniform scalar quantizer with the option of different quantizer step sizes for different sub-bands, based upon the dynamic range of the sub-band. Quantization step size will be 1 if lossless compression is required.
Entropy Coding	Differential Pulse Code Modulation (DPCM) of DC coefficients of *all blocks*. Huffman or Arithmetic coding of *each block* of AC coefficients (those left after truncation). Run-length coding of remaining string of zero magnitudes.	Sub-bands are divided into *precincts* and *code blocks* Each code block is input independently in raster order into the entropy coder. Code blocks are coded by individual bit plane, using three passes of an arithmetic coder.

Table 4.1 Comparison of compression stages in Baseline JPEG and JPEG 2000

The frequency transformation stage, because of the resulting configuration of frequencies, may be considered to be the most important factor in the differences differences between the distortions introduced. Although both decompose the image into image into frequency coefficients, the arrangement of the coefficients is different. It is different. It is commonly assumed that the improvements demonstrated in quality quality comparison trials of JPEG 2000 against JPEG are due to the use of the wavelet wavelet transform instead of the DCT. However Steingrímsson [120] suggests that it is not that it is not the choice of the transform but the differences in the way in which the image the image is sub-divided in the stages before entropy coding that might be the key to key to quality improvements. The output from the DCT stage in JPEG is an array consisting array consisting of blocks of 64 coefficients arranged so that they relate to the magnitudes magnitudes of frequencies in the same spatial region in the original image. The blocking blocking artefacts produced at higher compression ratios, as illustrated in

Figure 4.1, which arise as a result of coarse quantization in individual blocks of pixels and may be seen as one of the main causes of data loss and unrecoverable distortions in JPEG [126].

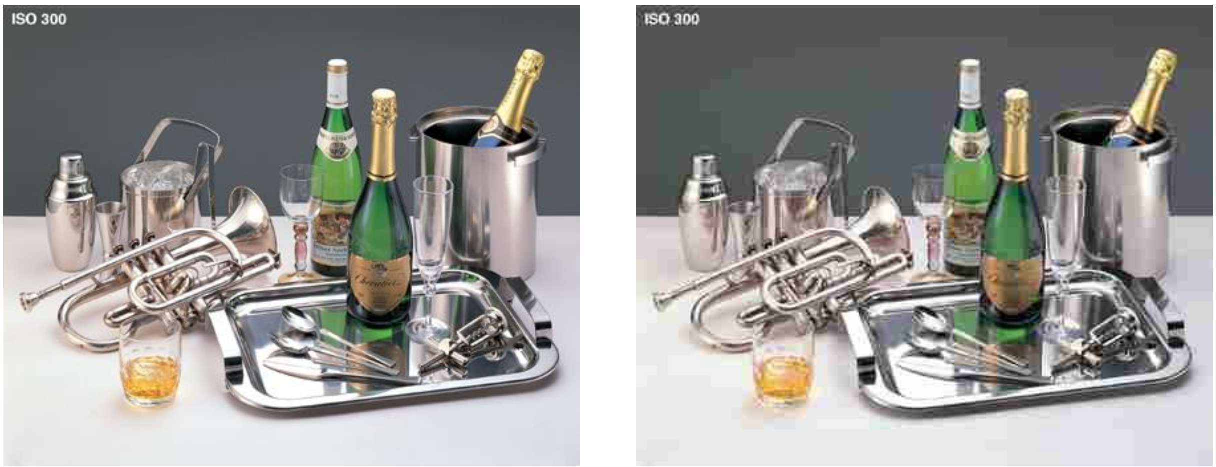


Figure 4.1 Left: Uncompressed reference image Right: Test image *'ISO Table'* displays blocking artefacts in the background at a JPEG compression rate of 50:1.

JPEG 2000 images do not suffer from blocking artefacts unless the image has been tiled. The decomposition of an image into sub-bands using a wavelet transform results in a lower-resolution version of the original and high-frequency information in horizontal, vertical and diagonal directions [96] (see figures 3.6 and 3.7). The sub-bands are quantized separately and further sub-divided into *code blocks* before entropy coding. A code block is a rectangular section of a sub-band and *a precinct* (see figure 3.8), consists of three groups of four code blocks, each group from the same position in each high-frequency sub-band at that decomposition level [127]. The inputs into the entropy coder are the code blocks by bit plane, from a precinct, scanned in raster order.

The values in a reconstructed block within a JPEG compressed image are dependent upon dependent upon the quantization table, (which is determined by the implementation and implementation and the quality setting) and the frequencies present at that spatial spatial location. By contrast, the reconstruction from the JPEG 2000 image will be made up of code blocks from precincts in the same spatial position relative to the edge of a particular sub-band, from all of the different sub-bands. Because the quantization step-size is different in different sub-bands, the errors build up variably, affecting an image area at different scales and being much less uniform over a spatial location by comparison with JPEG blocks. 'Smoothing' or 'smudging' artefacts appear at higher levels of compression. These appear as a blurring of small regions within the image as shown in

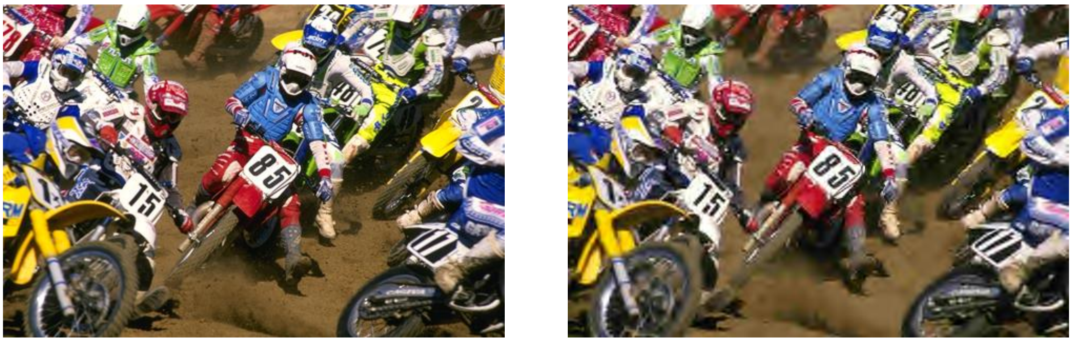
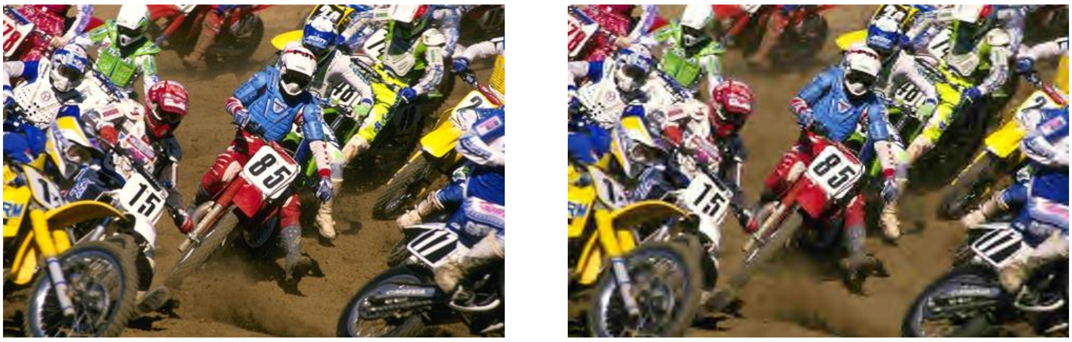


Figure 4.2.



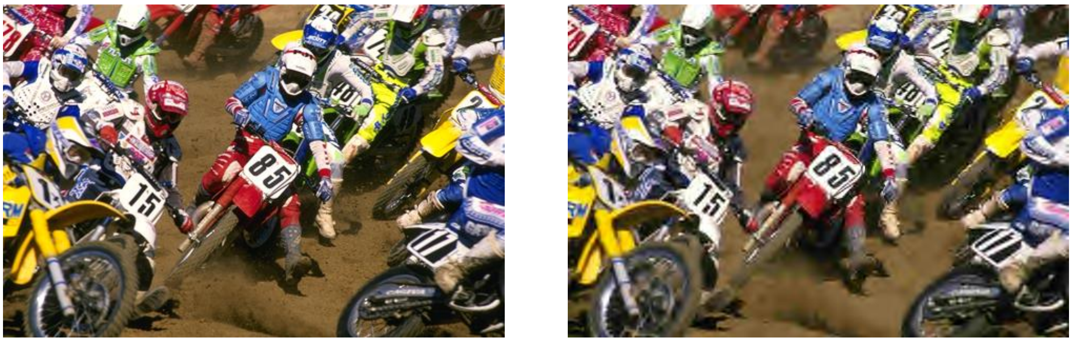


Figure 4.2 Left: Original image. Right: Test image *'Motorace'* at compression rate of 80:1 displays severe smoothing artefacts from the JPEG 2000 algorithm. Note however that numerical information is well preserved.

The lossy versions of both compression schemes suffer from ringing artefacts. These artefacts are a result of abrupt truncation of high-frequency coefficients during quantization. The effect of this may be modelled in the frequency domain as a frequency transformed image (which may be a block, or a sub-band) being filtered (multiplied, as a result of the convolution theorem) by a 2-dimensional version of the *rect* function, a *top-hat function* [128], which is the equivalent to an *ideal filter* (see figure 6.2(a)). During inverse transformation back to the spatial domain, a *rect* function becomes a *sinc* function (the *sinc* function and the *rect* function are a Fourier transform pair [129]). This is equivalent to convolving the image with a *sinc* function in the spatial domain; it affects the appearance of edges in particular and is evident as oscillations or 'ripples' around high-contrast edges [130] as shown in Figure 4.3.

The visual effects of ringing in JPEG 2000 may be reduced (although they are still evident) because of the arrangement of sub-bands, meaning that they are less localised and the errors are distributed across the image. They tend therefore to be less noticeable than the smoothing artefact illustrated in Figure 4.2. Smoothing artefacts may also mask ringing to some extent.

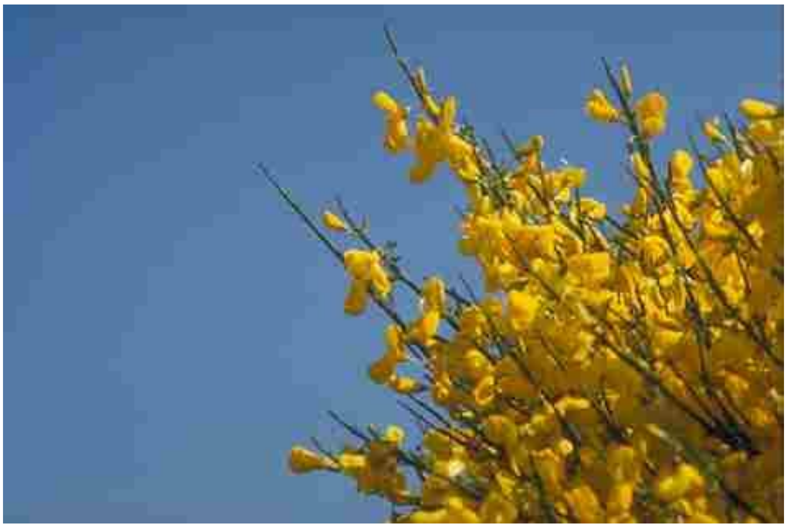


Figure 4.3 Ringing artefacts around edges in test image 'Yellow Flowers' compressed at 60:1 (JPEG compression). Severe blocking and colour distortions are also evident.

In JPEG images, because the errors from a block in a specific spatial area will affect the same area in the reconstructed image, ringing is much more visible. This can be identified as one of the key reasons that JPEG compresses text poorly (Figure 4.4).

At low compression rates, ringing can give a similar visual impression to sharpening. The density oscillations around an edge can lead to the slight overshoot or undershoot of density on either side of an edge in the same way as a result of the use of a sharpening filter (the *halo* artefact).

Both algorithms suffer from colour artefacts. These are caused by various factors, including the sub-sampling of chroma channels in the JPEG algorithm and the irreversible colour transformation in JPEG 2000, as well as reconstruction errors from quantization. The visual effect of these errors is a 'colour bleeding', which affects smoothly graduating areas and neutral tones, as seen in figure 4.5, and in the background in figure 4.1.

Figure 4.4 The ringing and blocking artefacts in this JPEG image compressed at a ratio of 40:1 make the text unreadable.

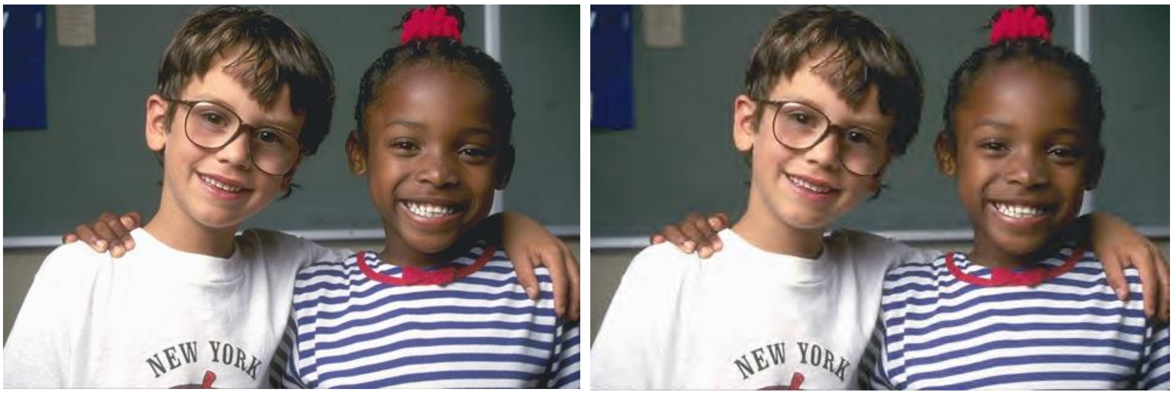


Figure 4.5 Colour distortions are evident in the less chromatic areas of the background of this JPEG 2000 compressed image.

4.3 Artefacts and Scene Dependency

As discussed in chapter 3, lossy compression of stimuli with different scene characteristics produces output images with different levels of image quality at the same compression rate. Similarly, if two stimuli of the same size but with different scene properties are compressed to the same quality level, they will usually have different file sizes.

The degree of error in the output image is primarily dependent upon the amount of compression; in JPEG the user set quality level defines the file size and the quantization table used; therefore the level of reconstruction errors. JPEG 2000 allows the user to compress to a desired file size or quality level, although it should be noted that the quality settings in JPEG and JPEG 2000 are specific to each algorithm and are unrelated to each other’s scales. Because the quantization stage is performed in the frequency domain, it is also dependent upon the frequency characteristics of the image contained within each image area or image block. The other main area of compression in both algorithms, the reduction in colour resolution, also affects the amount of distortion introduced.

The error difference as a result of scene property variations is a result of the scene dependency of the algorithm, which can be quantified to some extent by distortion metrics. But they cannot predict the visibility of the errors, which will be determined by scene characteristics and any masking effects.

The susceptibility of a scene to a particular artefact influences the results of an image quality study, in both objective and subjective assessments. Scenes with large areas of smoothly graduating tone may show more obvious blocking artefacts at high levels of compression, therefore such scenes might be expected to produce poorer results for JPEG. Meanwhile scenes containing many edges and straight lines, for example those typical in architectural images might suffer more visual degradation from the smoothing artefacts of JPEG 2000. These images may also suffer ringing artefacts.

4.4 Subjective Image Quality Assessment

This investigation aims to provide a comparison of the subjective image quality of JPEG versus JPEG 2000 in relation to scene content.

Thurstone's Law of Comparative judgement [37] assumes that the discriminal process (the process by which observers make judgements of samples) is a random variable with a probability density function following a Gaussian or normal distribution on the perceptual attribute scale (or 'ness'); in this case the image quality scale. Thurstone postulated that the proportion of times that a stimulus is judged greater than another stimulus might be viewed as an indirect measure of the distance between the two stimuli on the 'ness' scale being evaluated. Normalizing the difference between two mean scale values, by dividing by the standard deviation of the probability density function describing the values, produces results in terms of z-values. The z-scores may then be used to generate an interval scale of image quality [131].

Interval scales provide numerical values for a perceptual attribute, or image quality, against the physical properties of the image, in this case, compression rate. Distances between values on the scale are proportional to distances in perceived image quality, allowing predictions of differences between samples [131]. They may be compared in terms of relative magnitudes of differences.

In this investigation, a paired comparison experiment was performed in which observers were asked to select an image from a pair displayed on screen, based upon their preferred image quality. Ten observers, six male and four female, with some experience of visual assessment of images carried out the tests. Observers had normal or corrected vision, and normal colour vision. Each uncompressed image was compressed to the same range of four compression rates using both algorithms. All of the compressed images for a particular scene were then compared with all of the others from the same dataset. The dataset also included the original uncompressed TIFF version of the scene. The total number of unique pairs was 36 per scene.

4.4.1 Test Images

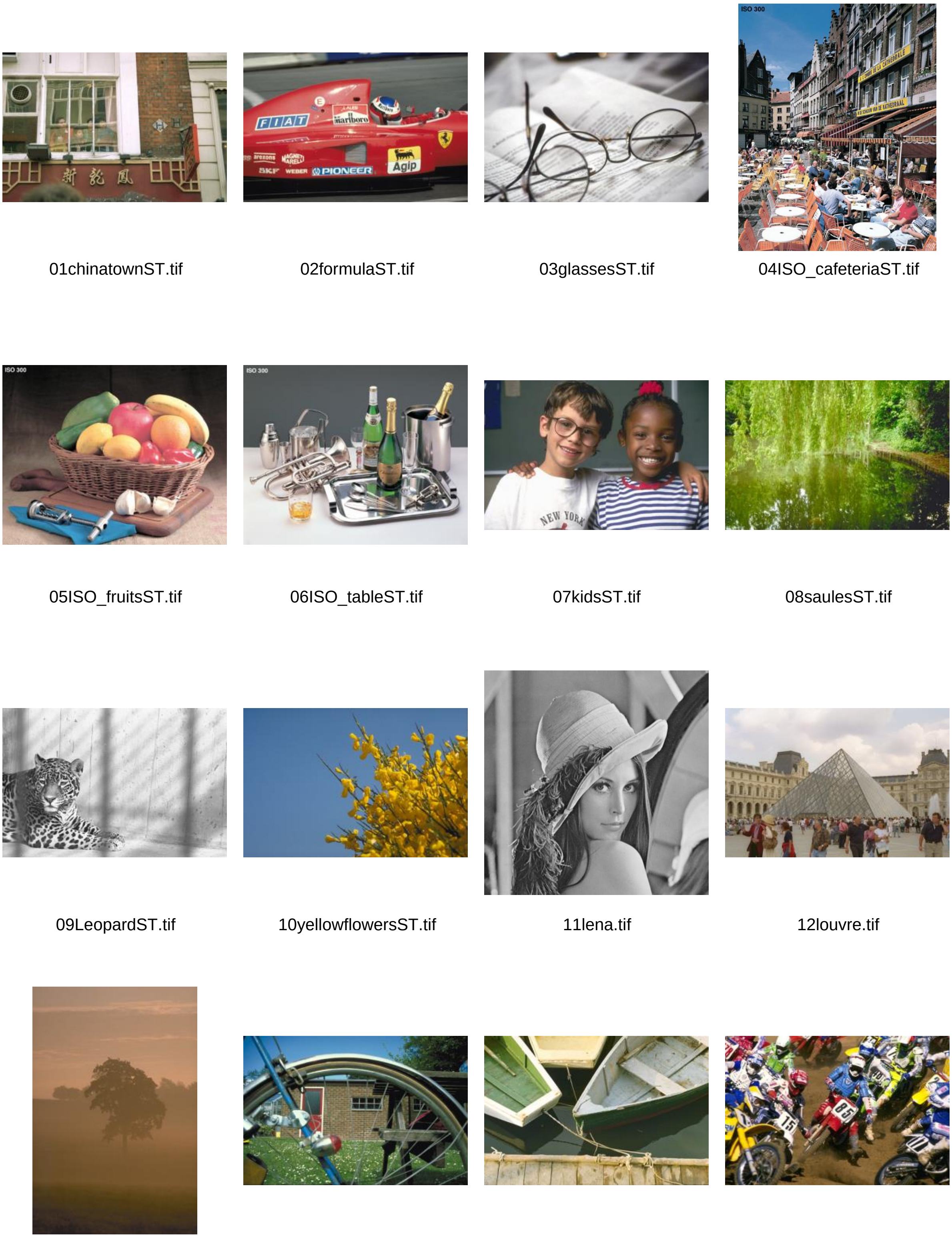
Sixteen original images were used in the investigation. Twelve were selected from a Kodak Photo-CD collection, three from the ISO 12640:1997 standard image set and the final one was *'Lena',* an image commonly used in compression quality investigations. The images were selected to cover a range of image content and characteristics.

The data set included:

- A range of different scenes, such as portraits, natural scenes, architectural.
- Scenes containing smoothly graduating tones, in which blocking distortions might be highly visible at higher compression rates.
- Scenes containing text, which might be susceptible to the ringing distortions inherent in JPEG.
- Highly chromatic scenes and some with a very low chromatic content.

- Scenes containing a large amount of fine detail, which might be particularly susceptible to JPEG 2000 smoothing artefacts.

The majority of images were colour, although two grayscale images were also used, including *'Lena'*. The Kodak Photo-CD images were opened at a resolution of 512 by 786 at 72dpi in CIELAB 16 bit per channel colour space. They were converted to sRGB colour space, down-sampled to 317 by 476 pixels and finally saved as TIFFs. The original images were all the same size, approximately 445Kb, to be displayed at 100% resolution. The selected size allowed two images to be displayed side-by-side on screen without any effects from further interpolation by the graphics card for the display. The images are shown in Figure 4.6.



01chinatownST.tif

02formulaST.tif

03glassesST.tif

04ISO_cafeteriaST.tif

05ISO_fruitsST.tif

06ISO_tableST.tif

07kidsST.tif

08saulesST.tif

09LeopardST.tif

10yellowflowersST.tif

11lena.tif

12louvre.tif

13africantree.tif

14bike.tif

15boats.tif

16motorace.tif

Figure 4.6: The 16 images used in the psychophysical comparison of JPEG and JPEG 2000

4.4.2 Image Compression

It was necessary to set a maximum compression rate based upon JPEG rather than JPEG 2000 because of the more limited compression capabilities of JPEG. After a pilot test, the images were compressed at intervals of 80:1, 60:1, 40:1, 20:1. This set of compression rates was selected to cover a range that might conceivably be used in everyday imaging across a range of applications, particularly consumer imaging applications and the web.

The JPEG compressed images were processed using Advanced JPEG Compressor v4.1, a stand-alone software application by Winsoftmagic Development [132]. The software compressed the images using baseline JPEG standard compression, while allowing specification of output file size, quality setting or compression rate. The JPEG 2000 images were compressed using Lurawave Smart Compress 3.0, developed by Algo Vision Luratech GmbH [133]. Default settings were used in both methods of image compression. Image optimisation was not used.

4.4.3 Display Characterisation

The images were displayed on a 15" NEC Multisync M500 CRT monitor, with a Matrox Graphics MGA Millenium graphics card adapter at screen resolution of 1024 x 768 pixels. To ensure correct colour rendition, the monitor was characterised [14], before being calibrated to the sRGB standard [134,135,136].

The viewing conditions were first adjusted to the sRGB recommended viewing conditions: the reference ambient illuminance was set to an average of 64 lux at the faceplate of the monitor and reference average chromaticities of the surround were equivalent to D50 (x=0.3457 and y=0.3585) [137].

The display allowed independent adjustment of the R, G and B channel signal levels. After a warm-up time of 30 minutes, the uncalibrated white point of the display was measured using a Minolta Chromameter CL200 with a CL-A11 head, taking measurements from a generated white patch covering 50% of the display area. Measurements were taken with a hood around the display under the reference viewing conditions. The white point was adjusted through the RGB channel signals to match the CIE x,y, chromaticity co-ordinates of the CIE D65 standard illuminant [137]: x=0.3127, y=0.3290.

The tone reproduction characteristics of the three channels were measured from a series of patches generated [138] on screen covering a range of pixel values from 0-255 for each channel, in increments of 5. For the individual channel measurements, the patches covered 50% of the displayed area and the other 50% was filled with the complementary colour (i.e. if a patch was displayed with RGB values of [128, 0, 0], the surround was filled with a uniform pixel value of [0,128,128]). Each measurement was taken three times from different areas and the results were averaged. Figure 4.7 illustrates the results. Power functions fitted to the individual plotted datasets are on the graphs with their R^2 values, indicating that the derived functions fitted the data to a satisfactory level, and that the channels combined to produce a gamma value of close to 2.2, as specified for the sRGB colour space.

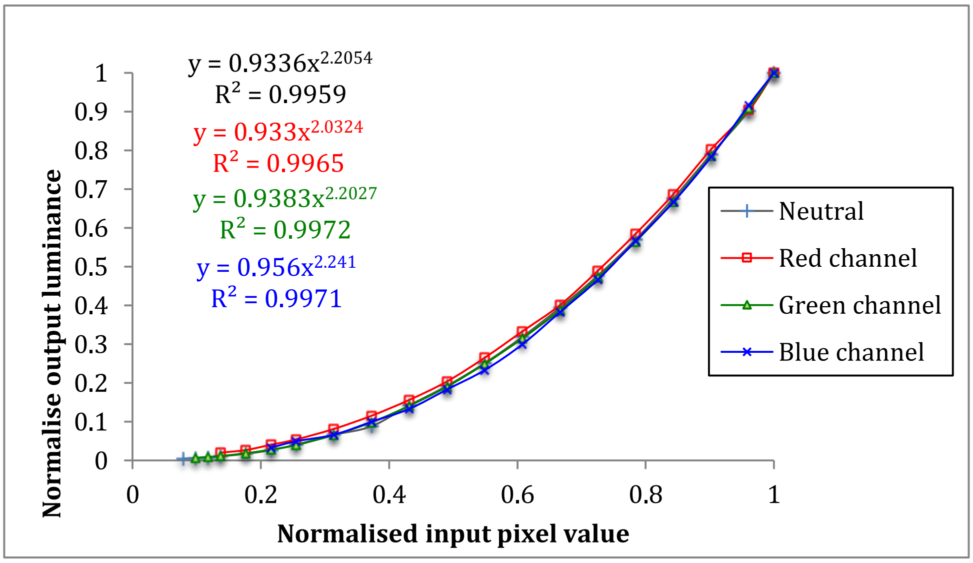


Figure 4.7 Tone characteristics of the NEC Multisync M500 CRT monitor.

4.4.4 Psychophysical Display

The paired comparison software was written in Visual Basic™ 6 [139]and run on an IBM compatible HP Vectra VA platform. Observers sat approximately 60cm from the display, although the distance was not controlled by a headrest for this experiment. Images were displayed side-by-side on the screen in a random sequence, and randomised in position from left to right. For each scene there were four compressed images for each algorithm and one uncompressed image. These gave a total of 36 unique pairs per scene and 576 comparisons in total.

Observers were provided with written instructions prior to the test and were asked to take a break from the test if they felt fatigued. They were asked to select the image in the pair that in their opinion had the highest quality. They were not time-restricted in making each selection and all observers completed the test in less than an hour. Results from the observations were recorded into a text file, which was saved directly on the computer hard drive.

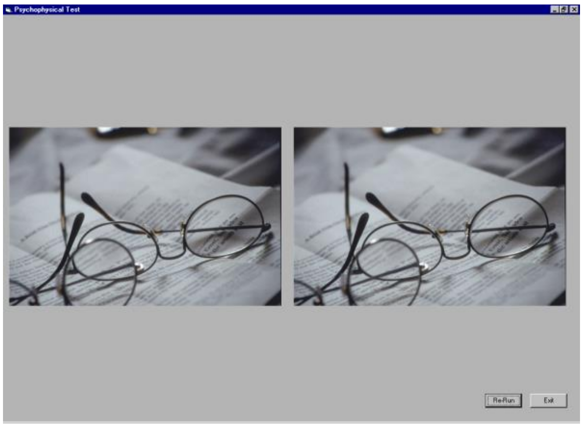


Figure 4.8 Psychophysical display.

4.4.5 Image Distortion

Additional to the subjective investigation, values for peak-signal-to-noise ratios (PSNR) were calculated between each original image and all compressed versions. As described in (3.4), PSNR is defined as:

$$PSNR(dB) = 20\ log_{\ 10}\ \left(\frac{max}{\sqrt{MSE}}\right) \tag{4.1}$$

4.5 Scene Characterisation

The scenes used in this experimental work were characterised by Triantaphillidou in a linked experiment [34]. A summary of the methodology and results are provided here in relation to the psychophysical experiement. More details of some of the scene measures used are included in chapter 5.

In the study, simple *scene metrics* (summarised in Table 4.2) were used to quantify global image content in terms of tone, colour and information content. The measures were applied to the uncompressed images with the aim of quantifying the scene properties and identifying any correlations between the results and the performance of the compression algorithms. The measures were applied in the CIELAB colour space.

Type of measure	**Measure**	**Predicted scene characteristics**
First order statistical measures *applied to the CIE L* channel*	Mean value *m* Median value *md* Skewness *s* Variance *v* First order entropy *e*	Global average intensity Global average intensity Imbalance of the probability density function Global scene contrast Information content, random changes in the scene
applied to image C_{ab}^* calculated from the CIE a b* channels*	Chroma variance $V\ C_{ab}^*$	Colourfulness
Image Segmentation	Busyness *b*	Ratio of busy areas to uniform areas within the scene
Line Detection: Edge detection followed by radon transformation	Log$_{10}$ of the number of lines $log_{10}(f)$	Amount of lines and busyness of the scene

Table 4.2 Scene measures used to quantify scene properties [34]

The measures were applied in the MATLAB environment [140]. The L* channel of the images was converted to an 8-bit greyscale image. The *busyness* metric involved edge detection using the Sobel operator to form a gradient image, followed by thresholding and the application of morphological operations to separate homogenous and inhomogenous regions of the image (Figure 4.9).

The scenes were ranked according to their values for the individual measures. The images were grouped according to their interval scale results and correlations were sought with the results from the scene measures.

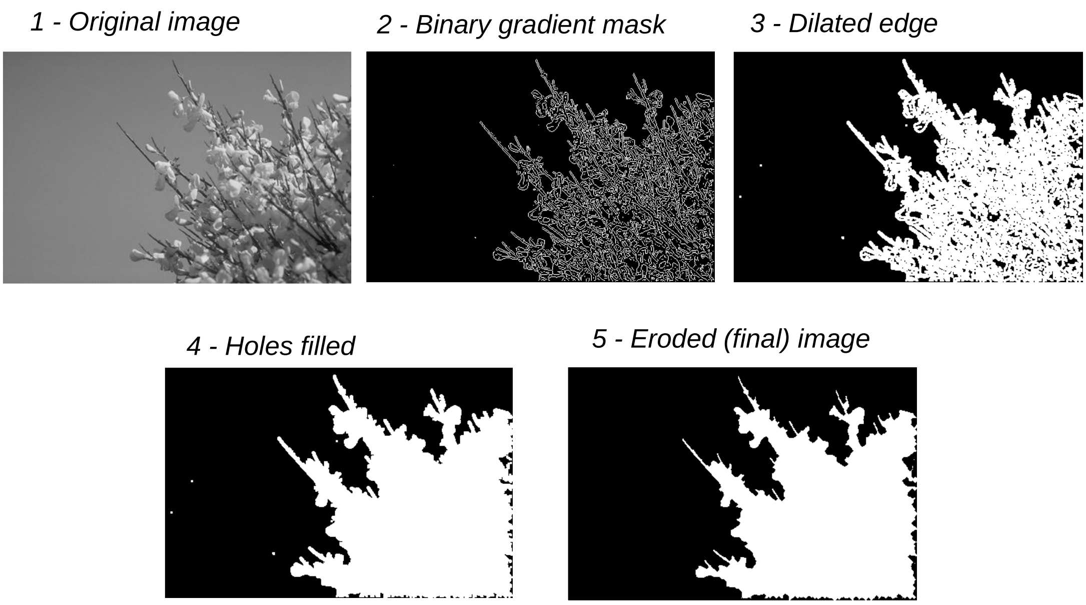


Figure 4.9 The stages in the busyness metric, by Triantaphillidou, reproduced from [34]

4.6 Interval Scale Generation

Engeldrum describes the practical derivation of interval scales from the proportion of observer responses in detail in [131]; a summary is provided here for completeness. According to Thurstone [37], the relationship between the z-deviates and scale values for samples A and B is defined by [131]:

$$S_A - S_B = Z_{A-B}\sqrt{\sigma^2{}_a + \sigma^2{}_b - 2\rho\sigma_a\sigma_b} \qquad (4.2)$$

Where $\mathbf{S_A} - \mathbf{S_B}$ is the difference between scale values, $\mathbf{z_{A-B}}$ is the z-value produced, $\sigma_\mathbf{A}$ and $\sigma_\mathbf{B}$ are the standard deviations of the observers' responses for the two samples and ρ is the correlation between the two samples.

In the Case V solution to this expression, it is assumed that the variances are equal and that there is zero correlation between samples, which simplifies the expression to:

$$S_A - S_B = Z_{A-B}\sigma\sqrt{2} \qquad (4.3)$$

The measured proportions from observer responses are used as an estimate of the probability of sample *A* being preferred over sample *B*. Therefore [131]:

$$P(A > B) = H(S_A - S_B) \qquad (4.4)$$

Where H is an underlying cumulative density function (probability distribution function). The inverse of H will give the scale difference values [131]:

$$S_A - S_B = H^{-1}[P(A > B)] \qquad (4.5)$$

If $\sigma\sqrt{2}$ is set equal to 1, then $H^{-1}[P(A > B)]$ is equal to Z_{A-B}.

In the case where there is unanimous agreement among observers about a particular pair of images giving p=1, there will always also be a value of p=0. These proportions indicate that the scale values for the two images are so far apart that there is no confusion between the two. A proportion of zero or one results in a z-value of $\pm\infty$, due to the extended tail of the underlying cumulative distribution function [141]; this means that scale estimates are not robust, particularly if the number of observers are small. One option is to increase the number of observers significantly to improve accuracy, but this is an impractical solution. Engeldrum [142] suggests adopting a strategy proposed by Noether [143], in which a proportion of p=1 is substituted by 1-1/(2n) and p=0 by 1/(2n), where n is the number of observers.

An example of the scale value calculation matrix for image '*Lena'* is shown in Table 4.3. The values in the data matrix are the relative scale differences between two samples, calculated from Z_{A-B}. To reduce the number of comparisons, images were not compared with themselves. Instead the assumption was made that because observers would not be able to distinguish between the images, in a forced choice experiment they would guess, resulting in a proportion of 0.5 and a z-deviate of 0, as shown on the diagonal of the matrix. The final scale values are a sum of the relative scale differences for each image treatment.

	TIFF	JPEG20	JPEG40	JPEG60	JPEG80	JP2K20	JP2K40	JP2K60	JP2K80
TIFF	0.00	0.00	-1.64	-1.64	-1.64	0.52	-0.52	-1.64	-1.64
JPEG 20	0.00	0.00	-1.28	-1.64	-1.64	1.28	-0.25	-1.28	-1.64
JPEG40	1.64	1.28	0.00	-1.64	-1.64	1.64	1.64	0.00	-1.28
JPEG60	1.64	1.64	1.64	0.00	-1.64	1.64	1.64	1.64	0.52
JPEG80	1.64	1.64	1.64	1.64	0.00	1.64	1.64	1.64	1.64
JP2K20	-0.52	-1.28	-1.64	-1.64	-1.64	0.00	-0.25	-1.64	-1.64
JP2K40	0.52	0.25	-1.64	-1.64	-1.64	0.25	0.00	-1.64	-1.64
JP2K60	1.64	1.28	0.00	-1.64	-1.64	1.64	1.64	0.00	-1.28
JP2K80	1.64	1.64	1.28	-0.52	-1.64	1.64	1.64	1.28	0.00
Standard deviation	0.91	1.03	1.44	1.16	0.55	0.69	1.01	1.44	1.21
Mean Scale Difference	0.91	0.72	-0.18	-0.97	-1.46	1.14	0.80	-0.18	-0.77
Scale Value	**8.22**	**6.47**	**-1.64**	**-8.75**	**-13.16**	**10.28**	**7.19**	**-1.64**	**-6.97**

Table 4.3 Scale value differences for image *'Lena'*

4.7 Results and Observations

The results are presented as plots of interval scale values against compression ratio. Scenes producing similar trends in their interval scales have been grouped for clarity.

The first two groups indicated preference for JPEG 2000 over JPEG across most of the range, with significantly better performance at higher compression ratios. Group 1 '*Lena*', *'Glasses'* and *'Leopard'* produced the highest interval scale ratings for JPEG 2000 compared to JPEG across the majority of the compression range. They demonstrated little perceptible quality loss until a compression ratio of 40:1 for JPEG 2000 compression, with only a gradual loss in quality (Figure 4.10) from 40:1. The results for the JPEG compression of these images produced steeper curves, showing more overall quality loss, particularly at high compression levels. There was more perceived quality loss overall; the average loss on the interval scale for the JPEG was 21.16 scale points, compared to only 14.14 for JPEG 2000. The difference between the results from the two algorithms was more marked at the bottom of the range, confirming better performance of JPEG 2000 at lower bit rates. Image 'Leopard' indicated only a slight preference at 40:1 but with a relatively large standard deviation this could be an anomaly; the remainder of the range again indicated a clear preference for JPEG 2000.

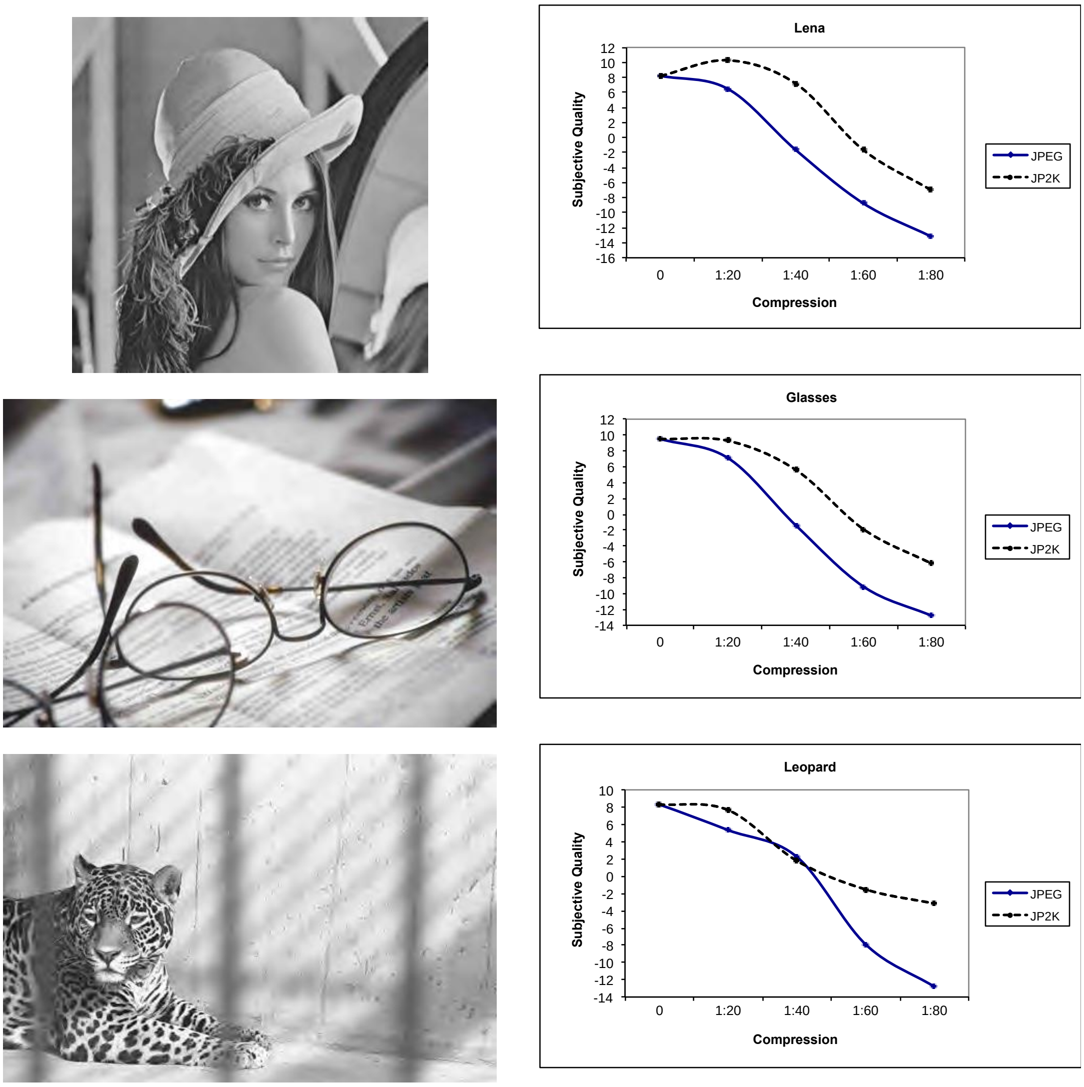


Figure 4.10 Group 1: images compressed at 60:1 by JPEG 2000 and interval scales. Images that were the least susceptible to JPEG 2000 artefacts

The scenes in the groups shared certain obvious characteristics and the similarities were confirmed by the common results from the scene metrics:

Average global intensity *m, md*	**Skewness *s***	**Colourfulness VC_{ab}***	**Busyness *b***	**Information content *e***
High	Negative	Low	Average	Average

Table 4.4 Group 1: Common scene characteristics

Only one of the images (*glasses*) was a colour image. The high global intensity values and negative value for skewness indicated that their histograms were dominated by light tones. The lack of chromatic information in the images meant that they did not suffer from colour artefacts, although this was a characteristic of both algorithms and therefore could not explain the difference in results. Overall, the quality of the scenes was more robust under JPEG 2000. An example is illustrated in Figure 4.11

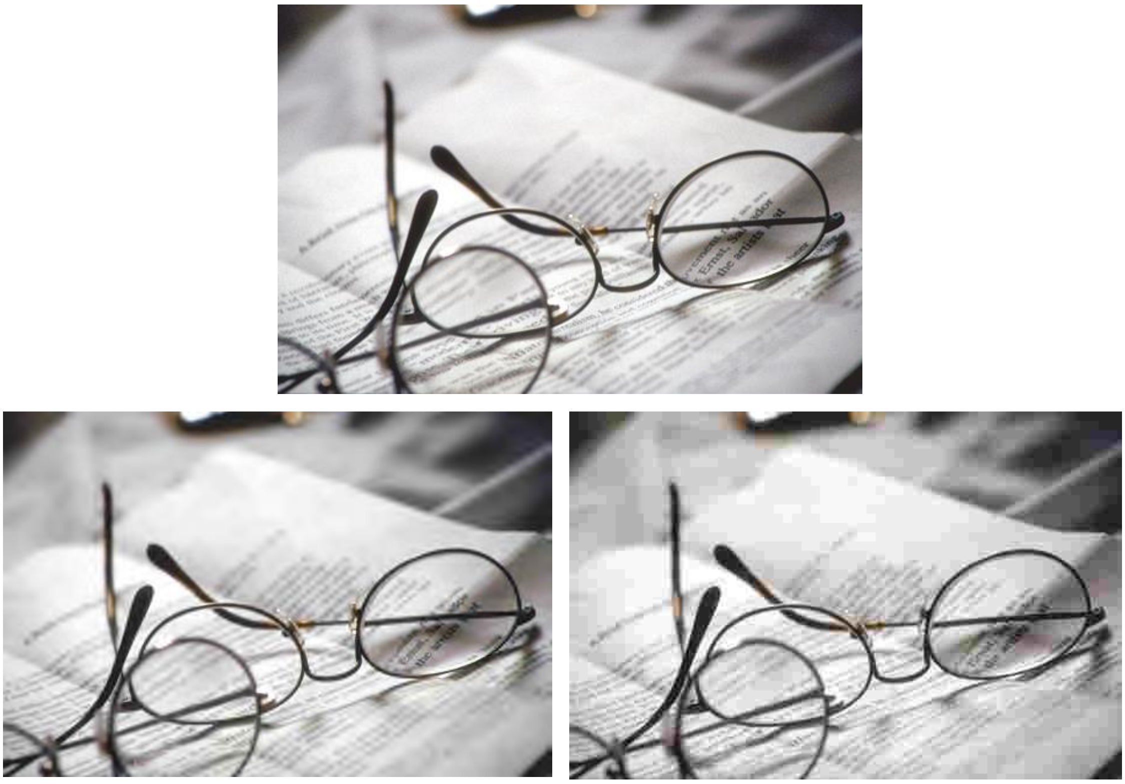


Figure 4.11 Compression of the *glasses* image at 80:1 compression ratio. Top: Original (uncompressed) Bottom Left: JPEG 2000, Bottom Right: JPEG

The most obvious explanation is the prevalence of blocking, which is very noticeable in the smooth areas of graduating tone. The scene measures do not reflect this well, but some form of combination of global intensity measures and quantification of regions of low frequencies might be able to identify these types of images.

Of interest within this group is the *'Lena'* image. This image was preferred to the uncompressed image at a JPEG 2000 compression of 20:1. This is not surprising in the context of previous scene dependency studies, confirming the observations from the work of Biederman [144], that portraits are often preferred in terms of quality when slightly blurred.

The results for the images in group 2 (Figure 4.12), *'Kids'*, *'Formula'*, *'Motorace'* and *'ISO Cafeteria'* indicate better perceived quality for JPEG 2000 compared to JPEG across the entire compression range, although with a less marked improvement than group 1; quality decreased at a relatively constant rate for both algorithms.

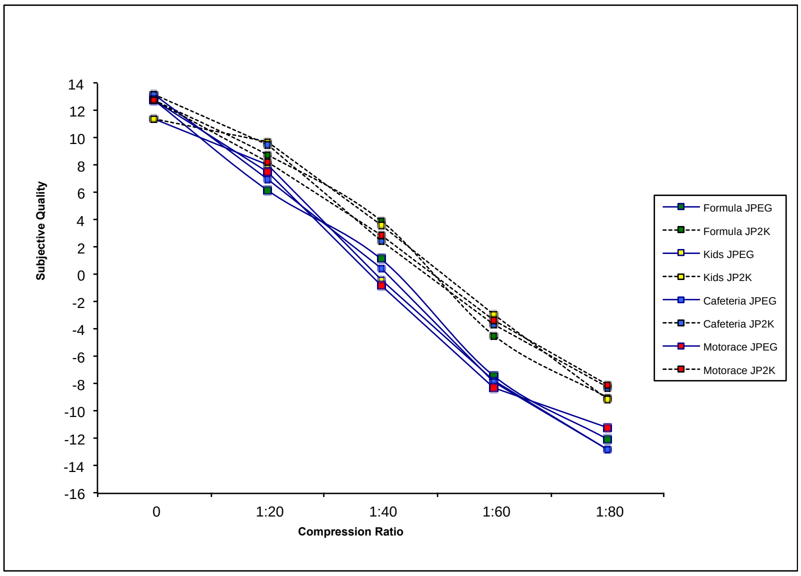


Figure 4.12 Group 2 : Preference for JPEG 2000 across the entire range, significant difference in quality loss for the two algorithms

The overall loss was large for both algorithms, with an average loss of 24.14 relative scale points for JPEG and 21.13 for JPEG 2000. This indicated a greater loss in image quality at high compression ratios than observed in the first group. These images all had a high chromatic range and contained key areas of fine detail, where loss of high frequencies was more noticeable. All four images also contained text or numerical data.

Average global intensity *m, md*	**Colourfulness *VC_{ab}***and/or Variance V**	**Busyness *b*, Information content *e***	**Amount of lines *$Log_{10}(f)$***	**Common scene content**
Low to average	Very high	High to very high	Very high	Text

Table 4.5 Group 2 scene characteristics

The high frequency content of the images, indicated by high values for *b, e* and *log$_{10}$(f)* might account for the large loss in quality. This is an example of algorithm scene dependency, as a result of both algorithms assigning less perceptual importance to high frequencies. The high errors in these areas resulted in very apparent blocking, smoothing artefacts and ringing. These scenes could therefore be identified as being highly susceptible to transform based compression and the scene measures were successful in identifying this.

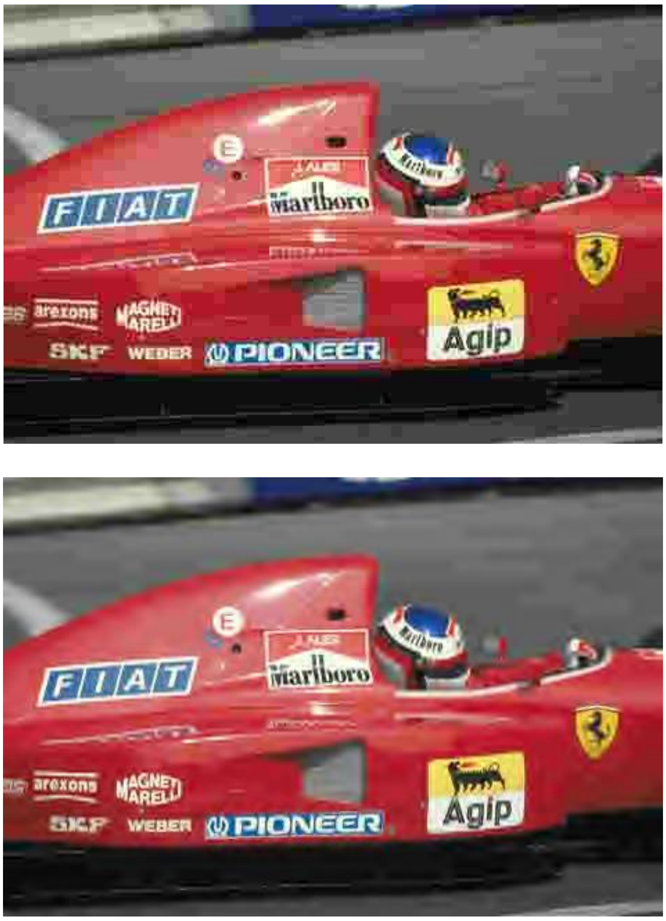


Figure 4.13 Comparison of artefacts at the highest compression rates on the Formula image: Top: JPEG 80:1, Bottom: JPEG 2000 80:1

The difference between the results from JPEG and JPEG 2000 indicated that blocking artefacts were more bothersome than smoothing artefacts in these scenes. This is illustrated in

, which shows the results for the *'Formula'* image at a compression ratio of 80:1. It is clear from this image that the ringing artefact is more evident in the

JPEG version. Text and numerical data would also be fixation points, therefore, artefacts in these areas would be expected to be more noticeable.

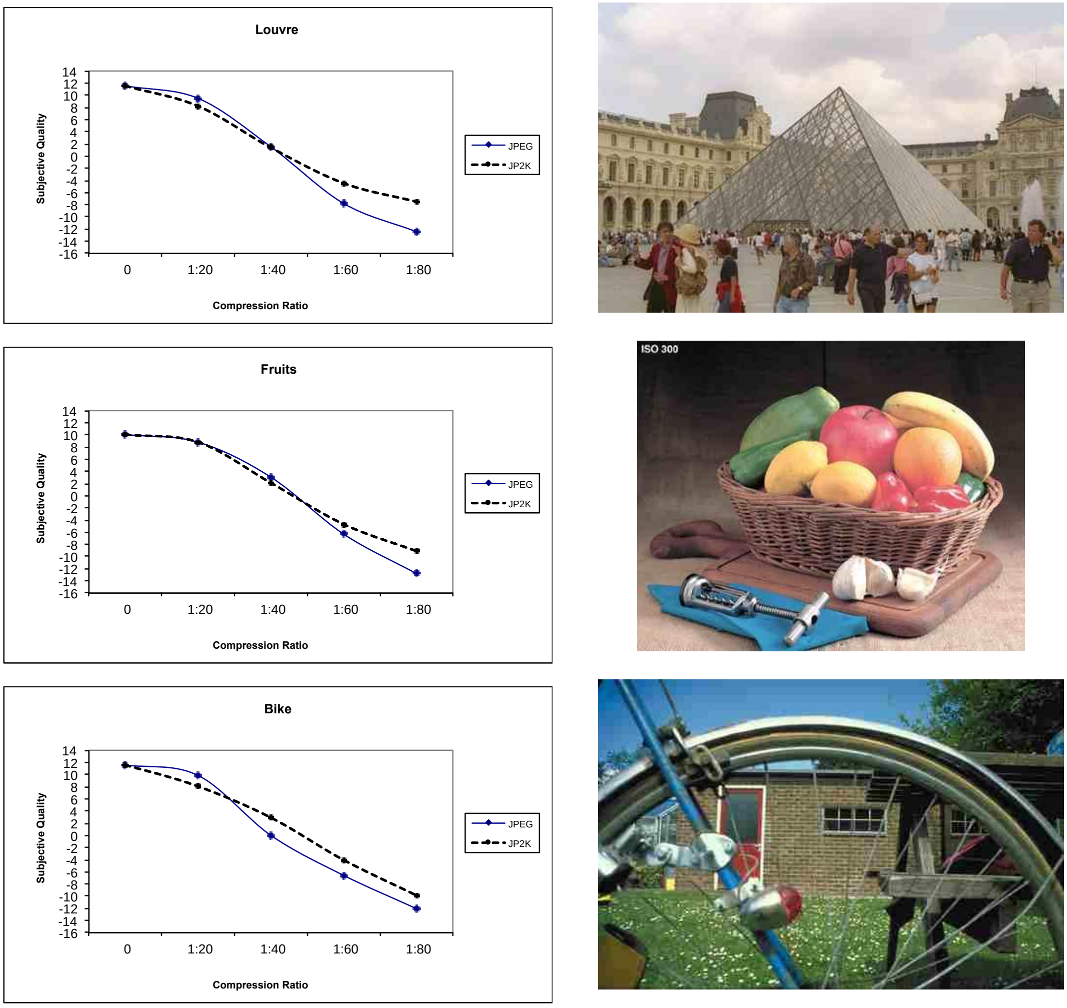


Figure 4.14 Group 3: Preference for JPEG at low CR, preference for JPEG 2000 at high CR

The next few groups of scenes produced more ambiguous results, shown in Figure 4.14 and Figure 4.15. The scene measures for groups 3-5 show no distinctive correlations in scene properties, which were in most cases average in their rankings [34].

Figure 4.14 indicates that perceived subjective quality is similar for both algorithms at low compression rates for this image group, with a slight preference for JPEG over JPEG 2000, but the reverse is true at higher compression rates. Scenes *'ISO Fruit'* and *'Louvre'* had a slightly smaller range of quality loss range for JPEG 2000 compression than JPEG than the scene *'Bike'*, meaning that there is less quality loss at higher JPEG 2000 compression rates. The results are very close for both algorithms, but the slight preference for JPEG at low levels of compression suggests that the sharpening effect from JPEG ringing artefacts might be preferable to the blurring caused by JPEG 2000. As compression is increased the perceived quality decreases and JPEG 2000 is preferred. At high compression rates blocking becomes more visible and ringing more severe. These scenes contain both flat areas and high frequencies; therefore the effects of either might be less preferable or more noticeable than the smoothing of JPEG 2000.

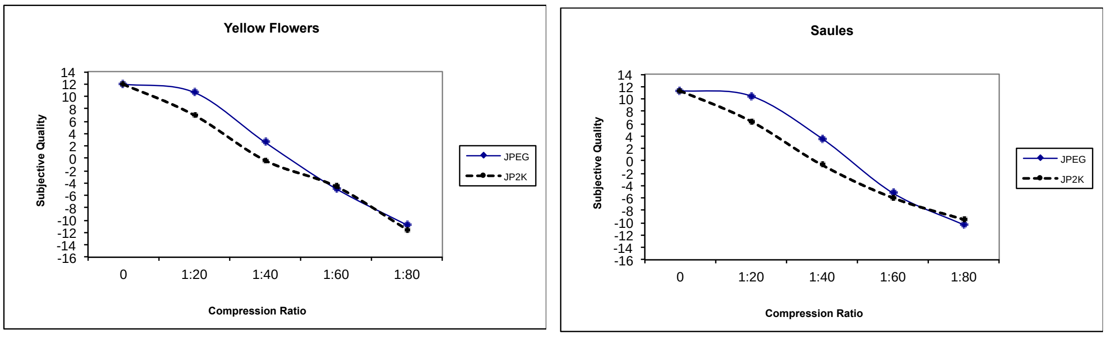


Figure 4.15 Group 4: Preference for JPEG at lower compression levels

The results for group 4 are unexpected. In both of these scenes, JPEG produces much better subjective quality than JPEG 2000 across most or all of the compression range. The curves (Figure 4.15) are extremely similar. The quality range for both compression algorithms is almost identical; however at compression ratios from 1:20 to 1:40, JPEG demonstrates improved quality over JPEG 2000. Both of these images contain large areas of fine and random detail. This detailed information is of one predominant colour in both scenes. Blocking, ringing and smoothing artefacts are present in the images produced by both algorithms; however the level of ringing is similar and may be discounted. Because there is so much fine detail within the scenes, the smoothing artefact is highly visible, however the blocking artefact is somewhat masked (Figure 4.16).

.

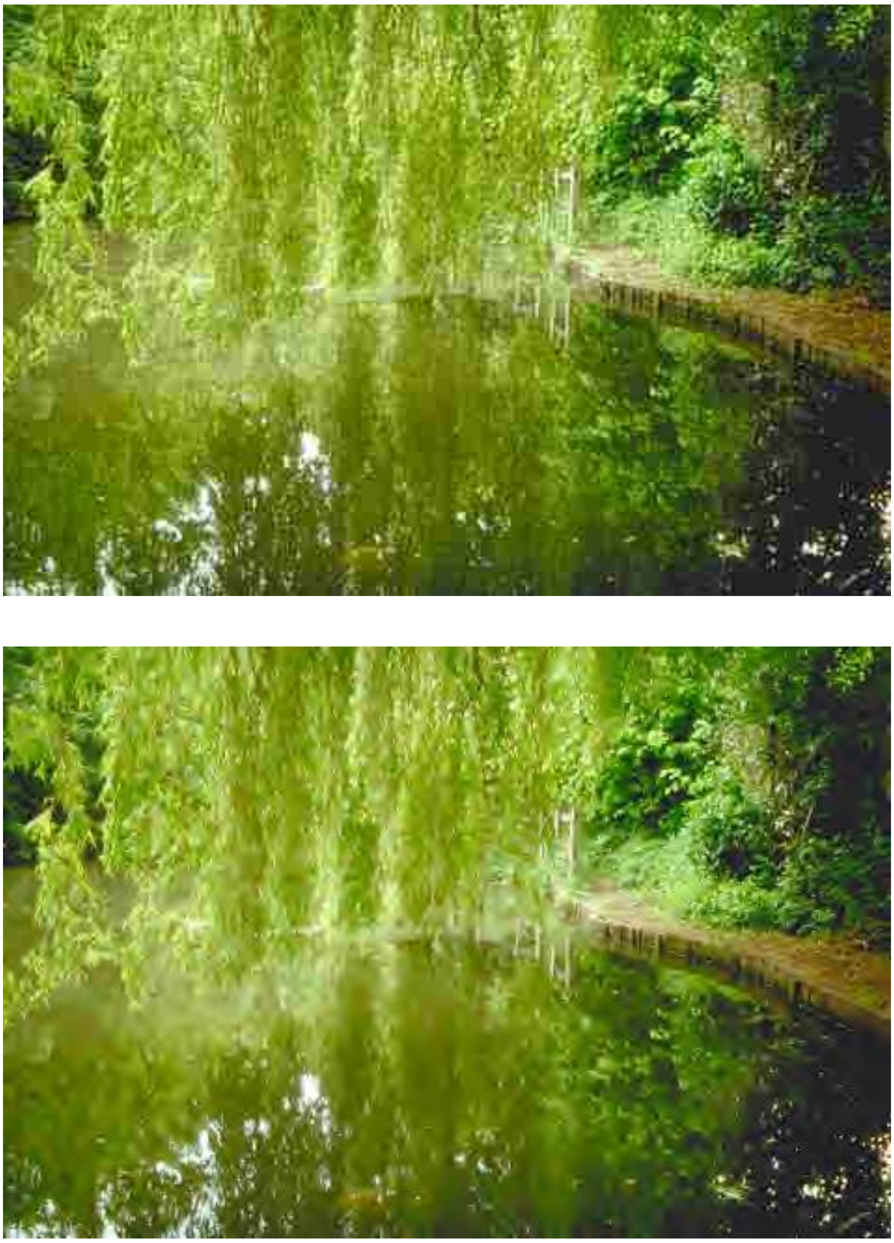


Figure 4.16 Compression ratio 40:1.Smoothing is clearly evident in the JPEG 2000 image (top), however the fine detail within the image appears to mask blocking artefacts produced by JPEG (bottom).

The images in group 5 (Figure 4.17) produced extremely similar curves for both algorithms, although in *'Chinatown'* there was a slight preference indicated for JPEG 2000, whereas JPEG seemed to be preferred for the *'Boats'* image. The *'ISO Table'* scene produced results that vary in preference for one algorithm or another across the scene.

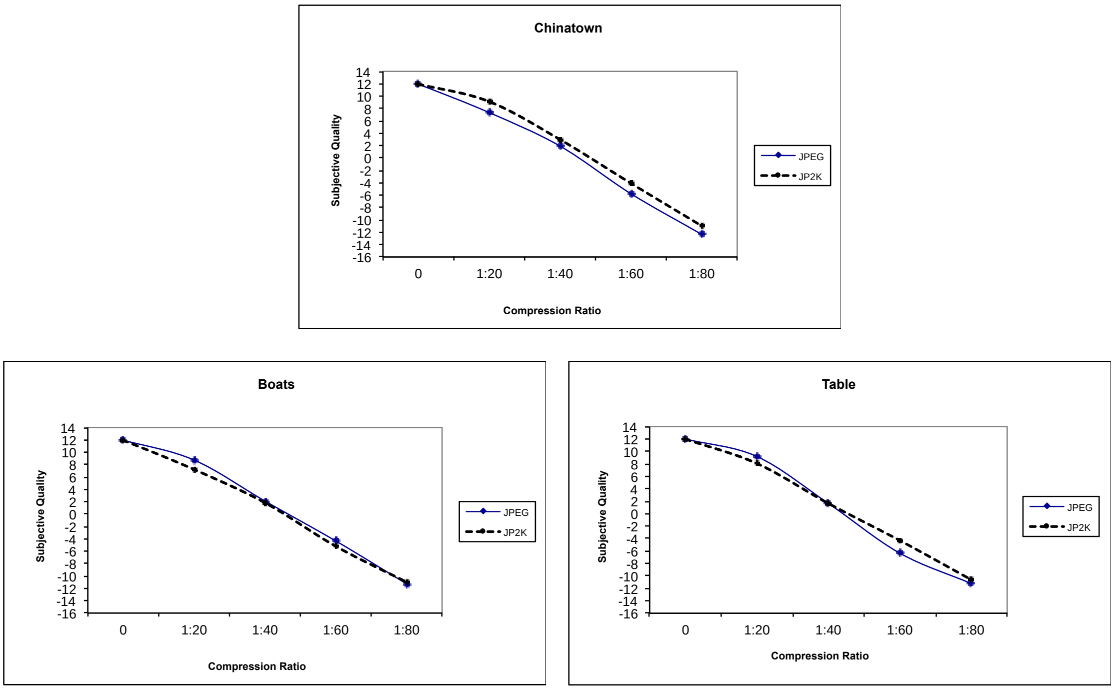


Figure 4.17 Group 5 results: Quality loss results similar for both algorithms

The images in group 5 might be considered to be 'average' scenes, average in all scene characteristics and producing very similar quality loss across the range, with no clear susceptibility to the artefacts of one compression algorithm or the other. These are the types of scenes that are often used in image quality studies while excluding the more susceptible scenes.

The most unusual results are produced from the *'African Tree'* image (Figure 4.18).

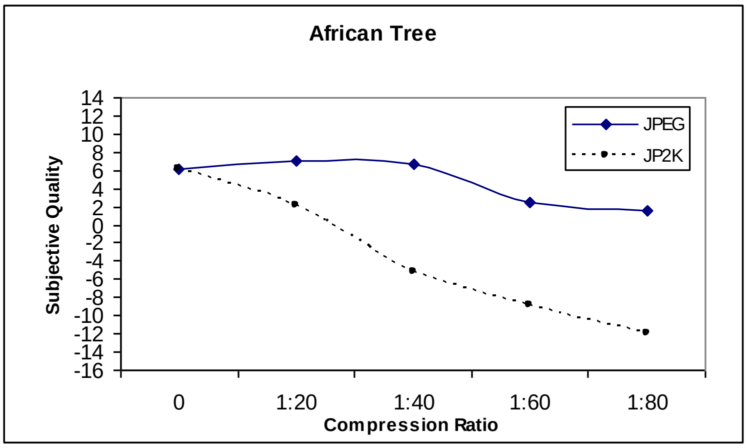


Figure 4.18 The *'African Tree'* scene produces the most anomalous results, with JPEG being preferred to both the original and JPEG 2000 at low compression rates.

In this scene, the JPEG images at compression ratios 1:20 and 1:40 have higher quality than both the uncompressed original and JPEG 2000 versions. For the rest of the range, JPEG is preferred to JPEG 2000 and there is little quality loss from JPEG.

This is the only image that has a positive quality scale value at JPEG compression of 80:1. Examining the scene characteristics, it is clear that this image is quite different to the other scenes, having low chroma, low contrast and virtually no fine detail. Significantly, the scene is an image of a tree in mist, and therefore contains soft edges. The blurring artefacts produced by JPEG 2000 therefore represent a loss in image quality, whereas the slight sharpening produced by JPEG might be viewed as an improvement. This result is similar to the results from one image in Steingrimmson's study, where JPEG was also preferred across the range [120].

Figure 4.19 shows the average interval scale across most scenes. The values for *'African tree'* have not been included, as they are so unusual compared to the rest of the images and cause a large increase in the standard deviation of the distribution. From these curves it is quite clear that JPEG 2000 outperforms JPEG across most of the range, with much more significant differences at high compression ratios. At lower compression ratios, the large standard deviations indicate a large spread of results and there seems much less of a performance advantage using JPEG 2000. JPEG was originally developed to be visually lossless at low compression rates and this perhaps indicates that both algorithms perform well at these lower levels of compression.

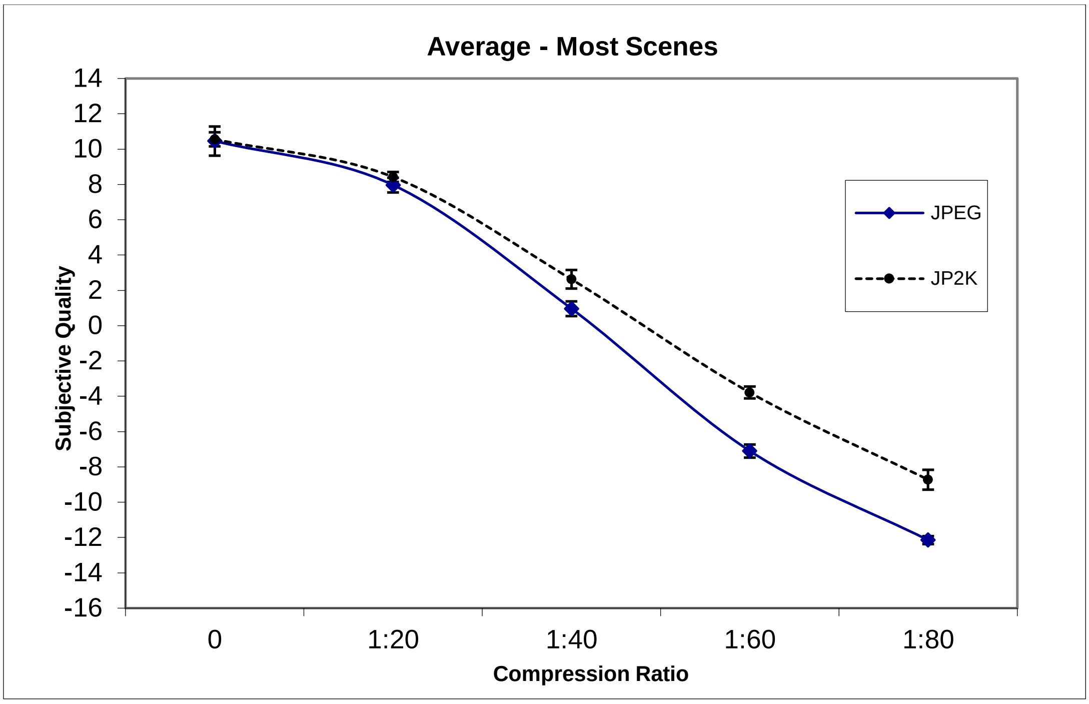


Figure 4.19 Average interval scales for all scenes, excluding *African Tree*

PSNR provides a measure of the error within an image compared to the original. Because the images are all of a standard size, the values between different scenes are comparable. Interval scales provide a measure of image quality loss, across a range of compressed images compared to the original, but do not provide information about the relative perceived quality of different scenes, as their zero point is not fixed and absolute. For this reason, error measures can be a useful method for quantifying the effects of an algorithm across different scenes and may predict the types of scenes that will produce fewer artefacts when compressed.

The results for the two algorithms, shown in Table 4.6, indicate higher PSNR, for JPEG 2000 compared to baseline JPEG across all scenes at all compression rates, apart from the two highest compression rates for the '*African Tree*' image. This confirms the results from previous similar investigations [96,121,117,11,93,118] indicating that JPEG 2000 has better error resilience than JPEG. Figure 4.20 shows the average PSNR results across all scenes except '*African Tree'*, which was again removed due to results anomalous with the remaining images. The average results for both subjective and objective evaluations confirm that JPEG 2000 outperforms JPEG; however, PSNR does not predict the scene dependency influencing the perceptual results. This confirms the assertion that PSNR and associated error measures are limited in their value as a tool in image quality studies.

PSNR (db)					
Compression Ratio		**20:1**	**40:1**	**60:1**	**80:1**
AFRICAN TREE	JPEG	44.0	42.3	41.4	40.6
	JPEG 2000	45.5	42.2	41	40.2
BIKE	JPEG	28.1	25.4	23.9	22.8
	JPEG 2000	30.1	26.5	24.6	23.6
BOATS	JPEG	31.8	28.4	26.9	25.7
	JPEG 2000	34.2	30	28	26.6
ISO CAFETERIA	JPEG	23.8	21.2	19.8	18.8
	JPEG 2000	25.4	21.9	20.4	19.5
CHINATOWN	JPEG	32.5	29.1	27.3	26
	JPEG 2000	35	30.5	28.2	26.6
FORMULA	JPEG	32.6	28.5	26.8	25.6
	JPEG 2000	36.7	31.1	28.3	26.6
ISO FRUITS	JPEG	32	29.2	27.7	26.5
	JPEG 2000	34.8	30.8	28.8	27.3
GLASSES	JPEG	36.5	32.9	30.6	29.1
	JPEG 2000	38.7	35.1	32.8	31.3
KIDS	JPEG	34.9	30.9	28.8	27.1
	JPEG 2000	38.3	33.3	30.8	28.8
LENA	JPEG	38.7	34.3	32.1	30.2
	JPEG 2000	41.7	37	34.4	32.3
LOUVRE	JPEG	32.2	29.4	27.7	26.8
	JPEG 2000	34.7	30.6	28.4	27.4
MOTORACE	JPEG	25.3	22.5	21	19.8
	JPEG 2000	27.5	23.6	21.8	20.6
SAULES	JPEG	25.4	23.4	22.4	21.9
	JPEG 2000	26.8	24	22.8	22.1
ISO TABLE	JPEG	32.3	28.3	26.2	24.9
	JPEG 2000	35.5	30.2	27.5	25.6
LEOPARD	JPEG	32	28.8	27.3	26.4
	JPEG 2000	34.8	30.8	28.9	27.8
YELLOW FLOWERS	JPEG	30.6	27.6	26.3	25.2
	JPEG 2000	34.4	30.1	28.1	26.5

Table 4.6 PSNR ratio results

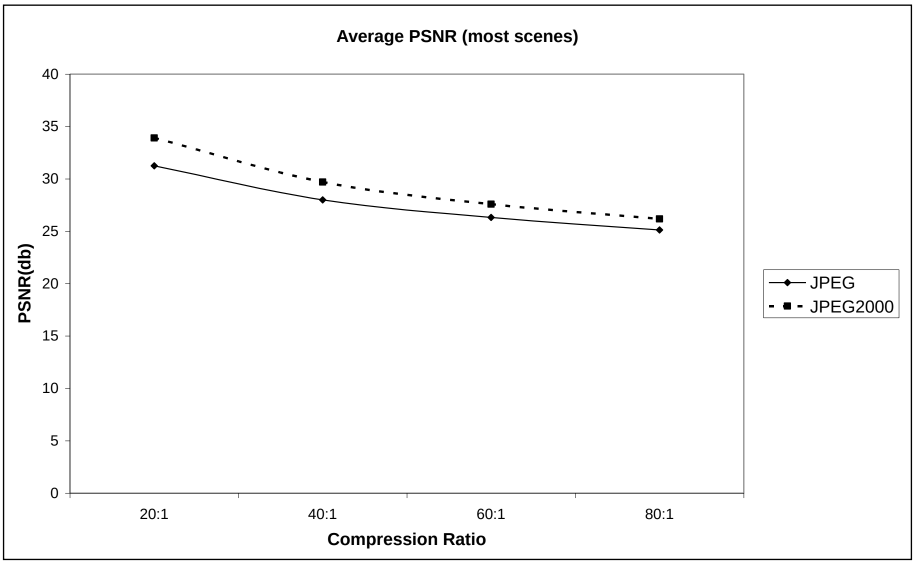


Figure 4.20 Average PSNR results for most scenes

Finally, PSNR results do not correlate with the conclusion from the subjective investigation that in some scenes, JPEG results are preferred to either JPEG 2000 or to the original. Error measures may be considered to in some way quantify the scene dependency of the algorithm, but as there are a number of other influencing factors, they cannot predict the perceived image quality results.

4.8 Summary

This chapter describes a psychophysical experiment to evaluate the image quality of the baseline JPEG algorithm against lossy compressed JPEG 2000 and original undistorted TIFF images, producing interval scales of the results. The results showed a small preference for JPEG 2000 over JPEG for most scenes. Examination of individual scene results demonstrated clear scene susceptibilities for each algorithm. An associated piece of work by Triantaphillidou et al [34] explored scene characteristics and some initial groupings were made based upon common behaviour of the two algorithms with particular images (for example where one algorithm proved to be significantly better than the other across the entire compression range). Correlations were identified between particular types of scene content and quality loss as a result of compression, particularly in relation to scene busyness. A further investigation explored perceptibility thresholds in a pilot study using JPEG 2000 compression only.

The results from the scene metrics were useful in that they demonstrated the potential to identify and predict scene susceptibilities in groups of images in relation to a particular distortion or algorithm with a relatively simple combination of scene descriptors. The results prompted the work in chapter 5, with a larger scale experiment, using a set of high quality and high-resolution test images. The work from this chapter has been presented at conferences and has also been published in a two-part paper (see chapter 10).

5 JPEG 2000 Thresholds

5.1 Perceptibility and Acceptability Thresholds

Chapter 4 explored the influence of scene content upon image quality and its importance in understanding the comparative performance of compression schemes. The next two chapters examine the image quality of JPEG 2000 in more detail in two psychophysical experiments; in this chapter an investigation into thresholds of perceptibility of distortion (image fidelity) and acceptability of the distortion (which is a suprathreshold judgement). The next chapter describes the implementation of ISO 20462 part 3: Quality Ruler Method [85] using the same set of sample stimuli.

Image fidelity studies are an important aspect of image quality evaluation for lossy processes, allowing the determination of the perceptibility of artifacts introduced into the image [145,41]. The *absolute threshold* in a psychophysical study of perceptibility is sometimes termed the point of subjective equality (PSE) [146]. This is the amount of a physical stimulus or image parameter that produces a response in 50% of observers when asked whether they can detect a difference between two images. The *visual difference threshold* is the point that is one JND from the absolute threshold, and is normally taken to be corresponding to a response from 75% of observers [84].

As explored in chapter 3, image compression processes are scene dependent in various ways. The artefacts from JPEG 2000 are very specific, and are localised, affecting, and being visible in, some parts of the image more than others. Chapter 4 illustrated that there is no clear quality preference between baseline JPEG and JPEG 2000 across all images, which may be partly as a result of the differences in the distortions introduced by the two algorithms. There will always be some scenes that will be more robust and suffer less from distortion when compressed with a particular algorithm, and neglecting this in image quality evaluation may give an incomplete picture of predicted performance.

Psychophysical studies of thresholds and JNDs of distortion, contribute to the development of guidelines around the use of image processes for imaging applications where fidelity may be critical (such as forensic and medical imaging [147] [148]), as well as providing reference data against which image quality metrics may be benchmarked. For less specialised imaging applications, image fidelity is not always a requirement and the *acceptability* of image degradations in a given context is also useful. Image fidelity studies involve observer judgements about perceptibility thresholds and just noticeable differences (JNDs) beyond the threshold. Judgements of acceptability are exclusively suprathreshold and are concerned with image quality; in this case distortions are visible, but may or may not be bothersome to the observer.

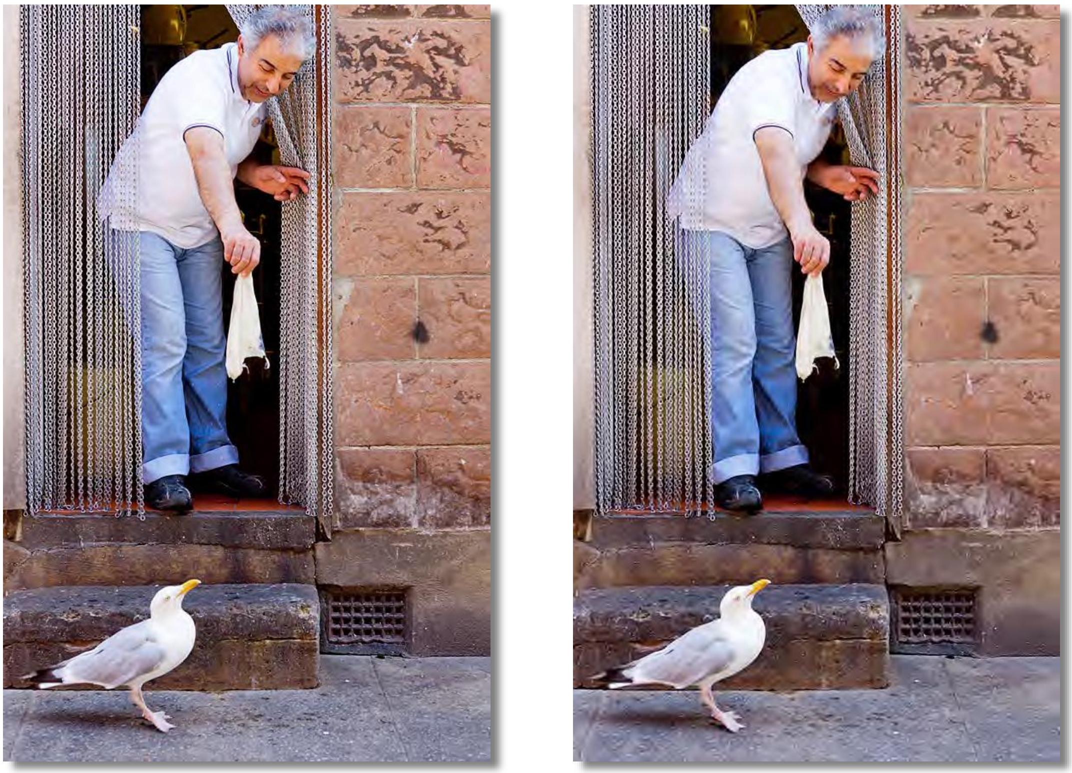


Figure 5.1 An example of an image at the perceptibility threshold (left) and the acceptability threshold (right) from [149]

Figure 5.1 shows an example of an image at its perceptibility and acceptability thresholds for JPEG 2000. The left hand image is the one in which most observers can detect a difference from the original (errors can be seen under close examination in the textured background and ringing around the edge of the man's arm). For this scene, the perceptibility threshold is low, at a CR of approximately 14:1. The image on the right is compressed to a level at which 75% of observers deem it to be unacceptable (for this image at around a CR of 22:1). In this image the distortions are more obvious and affect a larger area of the background, also ringing artifacts are beginning to affect the area behind the seagull, which is a focal point in the image, so they become more bothersome. From these results it can be assumed that for this scene, for compression ratios of between 14:1 and 22:1, distortions will be evident, but the image quality will be acceptable for most observers.

The problems caused by scene dependency affecting image quality evaluations raise the following questions pertinent to this investigation:

- What aspects of scene content affect the perceptibility and acceptability of JPEG 2000 distortions?
- Do scene characteristics affect fidelity and acceptability in the same manner?
- How does scene content affect the relationship between the perceptibility and acceptability thresholds?

It would seem that the susceptibility of scenes to particular distortions would be key to understanding the effects of scene dependency on thresholds. The interactions of the algorithms with scene properties and the *visibility characteristics* of the distortions (the way in which they are masked or emphasised by scene content) will determine the impact of the distortions.

In the case of acceptability, the *observers' preference criteria* must play a fundamental role in the decision as to what point a distortion becomes unacceptable. This of course relates to the imaging context and the relative specialism of the observer group. For example, the preference criteria of forensic specialists looking at fingerprint images will focus on the specific scene characteristics important to the extraction of key fingerprint features. In this case, variation in imaging conditions, scene content and resulting image characteristics will be restricted. In more general applications of imaging, preference criteria will be affected by variation in characteristics that make a scene more or less ‘pleasing’.

The images used in this investigation were captured purposefully to represent a range of scene content to allow further investigation of the scene dependency of the algorithm. This chapter includes a description of image acquisition and the image-processing pipeline used to prepare the images for the psychophysical investigation. Characterisation of the devices and workflow relevant to the experimental work is also detailed, and the process of scene classification using simple *scene metrics* [34] introduced in chapter 4, has been developed and employed in the selection and classification of the final set of test images.

Images were compressed using JPEG 2000 to a range of compression ratios, progressively introducing distortion to levels beyond the threshold of detection. Twelve observers took part in a paired comparison experiment to evaluate the perceptibility threshold compression ratio. A further psychophysical experiment was conducted using the same scenes, compressed to higher compression ratios, to identify the level of compression at which the images became visually unacceptable. Images were ranked for the two thresholds and were further grouped, based upon the relationships between perceptibility and acceptability. Scene content and the results from the scene descriptors were examined within the groups to determine the influence of specific common scene characteristics upon both thresholds.

5.2 Image Acquisition and Processing

The images used in chapter 4 were from a number of different sources, including some ISO images, and a number of images from a Kodak photo CD™. The capture systems were therefore unavailable and not characterised. The images were also of rather low digital resolution.

It was decided to investigate the performance of JPEG 2000 on images captured at higher digital resolution, using a contemporary professional level digital single-lens reflex camera, for display on a high-resolution monitor. It was also considered essential to be able to investigate the performance of the image capture system, and subsequent image processes prior to compression. Therefore the images used for the remainder of the experimental work in this research have all been captured on the same system by the author, with workflow carefully controlled and with due consideration to scene content to allow exploration of the scene dependency of the JPEG 2000 algorithm.

The focus during the preparation of the sample image set was to obtain images of optimal quality prior to compression. The selection of a RAW workflow, good exposure, use of a high quality lens, and maintaining native resolution and high bit-depth until late in the image processing pipeline were decisions made with image quality in mind. The aim was to minimise unwanted distortions introduced by other image processes where possible. The images were selected to encompass a range of scene content, allowing investigation into the scene dependency of the algorithm. Different scene types were included in the test set, providing good variation in scene characteristics, and captured under a range of different lighting conditions typically encountered in consumer photography.

5.2.1 Image Acquisition

An original sample set of 44 images was captured in raw file format (.cr2) using a Canon EOS 5D mkII D-SLR camera, which has a full frame sensor with resolution of 21Mp (5616 x 3744 pixels), and a Canon EF 24-70mm L II USM lens. Use of the raw file format enabled careful control of the image-processing pipeline prior to image compression. The images were captured using a range of focal lengths and apertures to provide variation in scene content and types of images. The image capture settings are summarised in Table 5.1.

Camera	Canon EOS 5D MkII
Lens	Canon EF 24-70mm L II USM
Pixel resolution	5616 x 3744
ISO	200-3200
Focal length	24-70mm
Aperture	f2.8-f22
File format	Canon Camera Raw (CR2)
Processed colour space and bit depth	sRGB, 16 bits per channel

Table 5.1 Initial image capture of test images. Images were captured using a range of focal lengths and apertures, although not all were used in the final selected test set.

5.2.2 Image Processing

The images were processed in an sRGB viewing environment, on the same calibrated display that was used later in the psychophysical investigation. The steps in the processing pipeline were based on a typical camera processing pipeline, but using linear rather than adaptive processing methods.

Using an external raw processing pipeline (Adobe CS 5.1 Camera Raw), the images were optimized scene-by-scene to correct exposure and white balance. A standard medium contrast tone curve was applied. Colour noise reduction was applied using the filter in the raw processor, with sRGB as output color space and the images were down-sampled to the minimum size possible in the raw processor (from 5616 x 3744 pixels to 1536 x 1024 pixels for uncropped images).

The raw workflow is shown in the top row of Figure 5.2. Initial demosaicing was carried out using the colorimetric interpretation from the camera profile, which produced an image in a linear camera RGB space [150]. The image was previewed on screen in Camera Raw. The preview process is similar to that used in soft-proofing: the camera profile is the source profile and the destination profile a user selected output profile, selected from four Adobe workspace options: Adobe RGB 1998, sRGB, ProPhoto RGB, Colormatch RGB, which are standard (calibrated) RGB colour spaces [151]; used with the display profile to correctly display the image. White balancing was achieved by adjusting two sliders. In Adobe Camera Raw, these 'blend' two colorimetric interpretations [150] for the camera from two profiles, each for a different white point (D65 and standard illuminant A). The resulting image was now in a large linear gamut colour space for processing, which has the same primaries and white point as Pro-Photo RGB (also known as Reference Output Medium Metric RGB (ROMM RGB)).

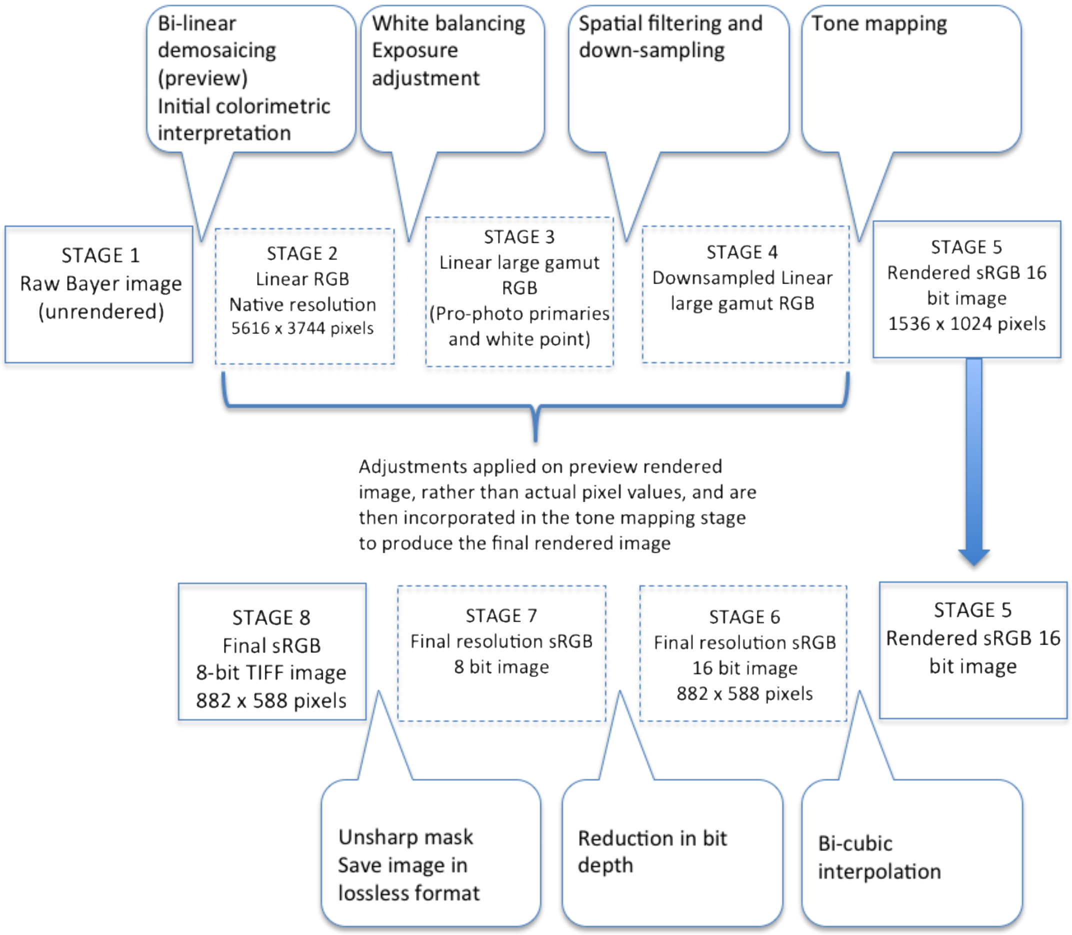

Bi-linear demosaicing (preview) Initial colorimetric interpretation
White balancing Exposure adjustment
Spatial filtering and down-sampling
Tone mapping
STAGE 1 Raw Bayer image (unrendered)
STAGE 2 Linear RGB Native resolution 5616 x 3744 pixels
STAGE 3 Linear large gamut RGB (Pro-photo primaries and white point)
STAGE 4 Downsampled Linear large gamut RGB
STAGE 5 Rendered sRGB 16 bit image 1536 x 1024 pixels
Adjustments applied on preview rendered image, rather than actual pixel values, and are then incorporated in the tone mapping stage to produce the final rendered image
STAGE 8 Final sRGB 8-bit TIFF image 882 x 588 pixels
STAGE 7 Final resolution sRGB 8 bit image
STAGE 6 Final resolution sRGB 16 bit image 882 x 588 pixels
STAGE 5 Rendered sRGB 16 bit image
Unsharp mask Save image in lossless format
Reduction in bit depth
Bi-cubic interpolation

Figure 5.2: Image processing pipeline for raw images captured using Canon EOS 5d MkII and processed using Adobe Camera RAW and Adobe Photoshop. Adapted from [88]

After raw processing, the rendered images were opened in Adobe Photoshop CS 5 version 12.1 X64. Further downsampling was applied (using bi-cubic interpolation) to optimize for the psychophysical display, with final image sizes of approximately 588 x 882 (some images were cropped to slightly different dimensions). The bit depth was reduced from 16- to 8-bits and a final sharpening stage was applied using an unsharp mask. The unsharp mask was applied to the lightness channel only in CIELAB space, before converting back to RGB; the images were then saved as uncompressed TIFF files. Figure 5.3 shows one of the final images after initial raw rendering and in its final processed state, which is much closer to the JPEG rendered image captured by the camera (all images were captured as RAW and JPEG).

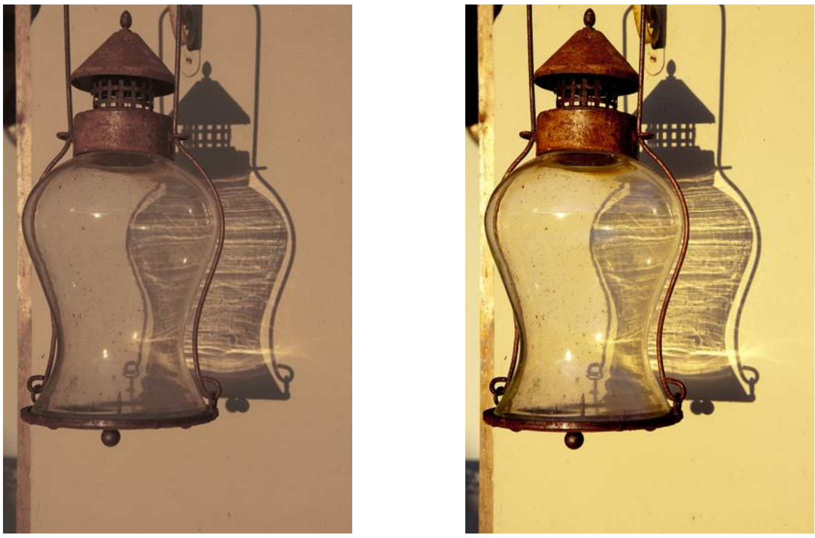


Figure 5.3: 'Lamp' image. The image on the left is as previewed with default rendering in the RAW processor prior to optimisation. With a linear tone curve and no exposure correction, the image is dark and desaturated. The image on the right is processed using the image-processing pipeline shown in Figure 5.2.

5.3 Characterisation of Devices and Workflow

5.3.1 Camera-Lens System Tone Reproduction

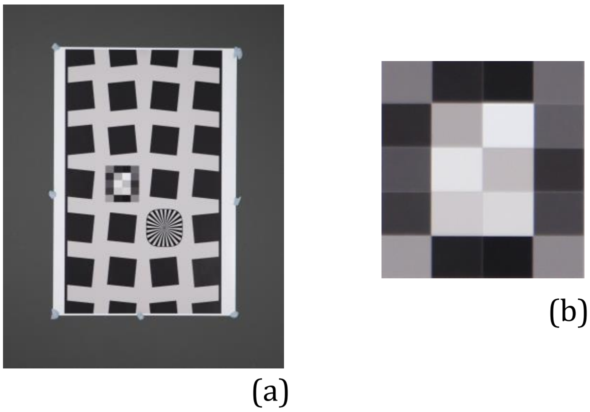
The opto-electronic conversion function (OECF) for the camera and lens system was measured according to ISO 14524 (2009) [152]. The measurements were taken prior to compression, because the image processing prior to compression was most likely to affect tone reproduction; and because an accurate estimation of the camera OECF was required for linearisation of images prior to calculating the system MTF, a necessary stage of the Soft-Copy Quality Ruler methodology described in chapter 6. The camera OECF is defined in ISO 14254 [152] as the: 'relationship between the input scene log luminances and the digital output levels for an opto-electronic digital image capture system'

The standard defines target characteristics, illumination, and a number of different methods for determining the OECF of a camera system. These include two methods (one with and one without the lens) for determining the focal plane OECF in terms of focal plane log exposure; and a further method, for determining full camera system OECF, in which a test target is captured under controlled conditions, and the independent input variable is scene log luminance. Because the OECF was to be used to linearise the images to determine the system SFR after image processing, the second method was chosen, producing the full camera OECF. The test chart used was the SFR plus test target, illustrated in Figure 5.4, which has a central area containing 20 neutral patches which are differentiated in approximate 0.1 increments.

The chart reflection densities were measured using a Macbeth TR924 densitometer, with three readings taken from the centre of each patch and averaged, resulting the densities illustrated in Figure 5.4(c). The chart densities were converted into log luminances using:

$$\mathrm{L_i} = \frac{10^{-D_i}E}{\pi} \qquad (5.1)$$

Where D_i is the grey scale patch density, and E is the illuminance incident on the chart in lux.



0.81	1.56	1.73	0.91
1.45	0.53	0.07	1
1.1	0.12	0.42	1.38
1.29	0.33	0.22	1.2
0.62	1.65	1.83	0.72

(c)

Figure 5.4: (a) Test target used for camera system OECF determination (cropped image) (b), area used for density measurements (c) density measurements corresponding to patch arrangement.

The chart was displayed vertically with its surface normal to the optical axis of the camera, and was illuminated using two tungsten lights positioned at an angle of approximately 45° to the target to provide even illumination. The illuminance incident upon the chart was measured using a Minolta Chromameter CL200 at the chart surface, at the position of the grey scale, and at 8 locations around it, on the slanted squares and the sinusoidal star directly adjacent. The measurements are shown in Table 5.2 and indicate that both illuminance and colour temperature were within 2% of the mean, with mean values of 866.3 lux and 2678.7 Kelvin.

Position	Illuminance (lux)	CT (Kelvin)
1	875	2680
2	871	2681
3	869	2678
4	864	2674
5	871	2681
6	871	2681
7	858	2679
8	861	2678
9	857	2676
Mean	**866.3**	**2678.7**

Table 5.2 Measured illuminance and colour temperature across the surface of the chart

The mean illuminance was used in equation (5.1) to generate input log luminance values from the measured chart densities.

The test target was photographed at a target-to-camera distance of 203cm. Nine images were captured at a speed of 160 ISO, focal length 70mm, f2.8 and 1/40 second shutter speed. The images were processed in Adobe Lightroom 3. All tone adjustment settings were set to zero, and the tone curve was set to linear. Noise removal and sharpening were turned off, and a lens correction was applied. The white balance was set manually to 2679 kelvin. The images were saved as 16 bit TIFF files with an sRGB colour profile.

A 51 x 51 pixel area was selected from each patch on the greyscale and using the region-of-interest manager in Image J, the mean pixel values recorded for each patch in all nine images, and the results averaged. The results are shown in Figure 5.5 and Figure 5.6.

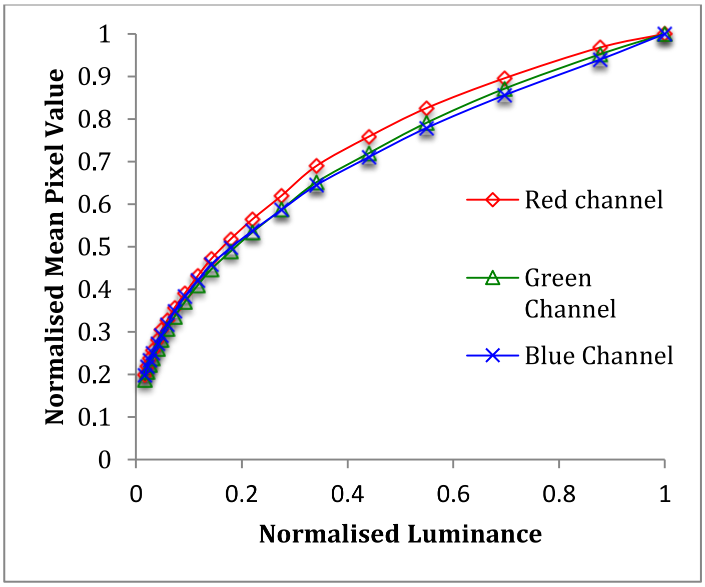


Figure 5.5: Opto-electronic Transfer Function for Canon EOS 5d MkII and Canon EF 24-70mm L II USM plotted in linear units, for minimally processed image (linear tone curve, 16-bit, sRGB rendering, TIFF from raw file)

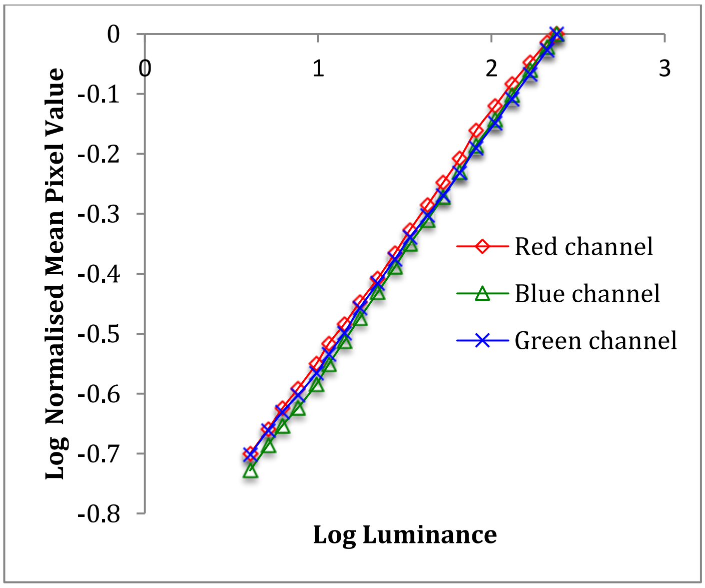


Figure 5.6: Opto-electronic Transfer Function for Canon EOS 5d MkII and Canon EF 24-70mm L II USM plotted in $\log_{10}$ units, for minimally processed image (linear tone curve, 16-bit, sRGB,rendering, TIFF from raw file)

5.3.2 Camera-System-Image Processing Pipeline: Tone Reproduction

The tone reproduction functions in the previous section characterise the OECF of the camera-lens systems with linear output and minimal processing of the RAW files. The image-processing pipeline described earlier, aimed to optimise the images perceptually, as well as down-sampling the images in preparation to be displayed on screen in the psychophysical tests. As illustrated in Figure 5.3, if the test images had been presented to observers with linear tonal adjustment from the raw files as described they would have appeared rather flat in terms of contrast, and too dark. The images also needed to be down sized to fit on the screen, which involved several stages of interpolation, and they required some colour noise reduction and sharpening to counteract the many sources of blurring (as a result of multiple interpolations, and the anti-aliasing filter over the sensor).

It should be noted that the image processing was kept to a minimum and the processes applied are similar to those that would be applied automatically in a rendered camera workflow if the images were being output in a fully rendered format (for example as JPEG files) [116]. The impact of these processes affected the image attributes in various ways, in particular introducing non-linearities into the tone characteristics. Therefore the same test images captured and detailed in section 5.3.1, were further processed using the same image-processing pipeline as the sample set of images to measure the effect of the processing on the OECF. The only significant difference was in the amount of optimisation required in terms of exposure, because the exposure of the test targets had been carefully controlled in the laboratory. The same procedure was used for generating the OECF. Because the images were significantly smaller than the full resolution version, and the grey scale test area was a small part of the image, the selected areas from each patch were only 6x6 pixels, to ensure that patch edges were not included in the calculation.

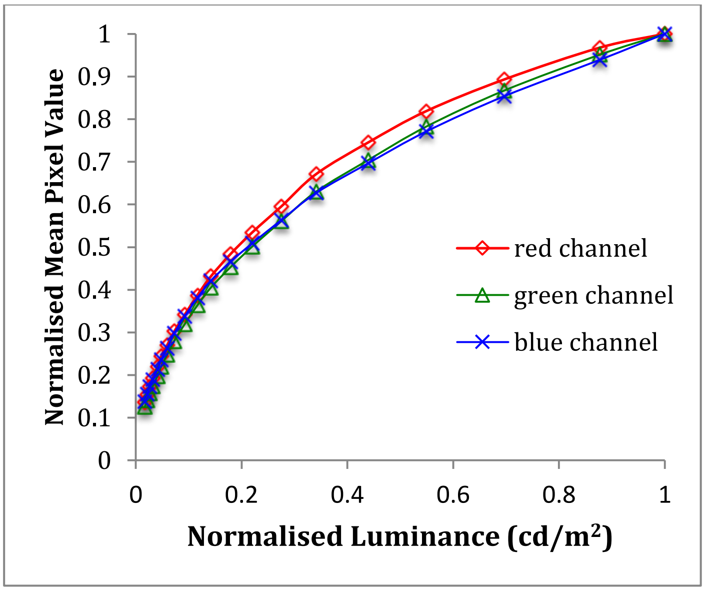


Figure 5.7: Opto-electronic Transfer Function for Canon EOS 5d MkII and Canon EF 24-70mm L II USM plotted in linear units, using the image-processing pipeline that was used to render final psychophysical test images (medium contrast tone curve, 8-bit, sRGB rendering, down-sampled, noise reduction, sharpened, TIFF image)

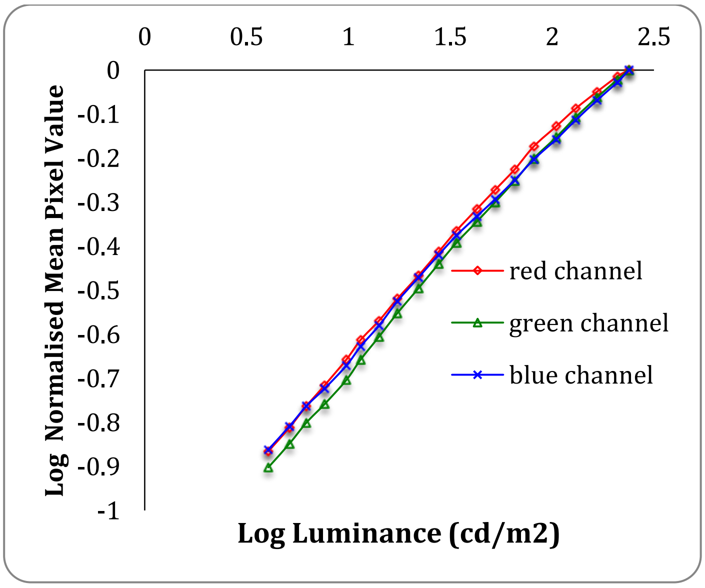


Figure 5.8: Opto-electronic Transfer Function for Canon EOS 5d MkII and Canon EF 24-70mm L II USM plotted in $\log_{10}$ units, using the image-processing pipeline that was used to render final psychophysical test images (medium contrast tone curve, 8-bit, sRGB rendering, down-sampled, noise reduction, sharpened, TIFF image)

The gamma values for the curves were obtained as before and are shown in Table 5.3.

5.3.3 Comparison of camera OECF before and after processing

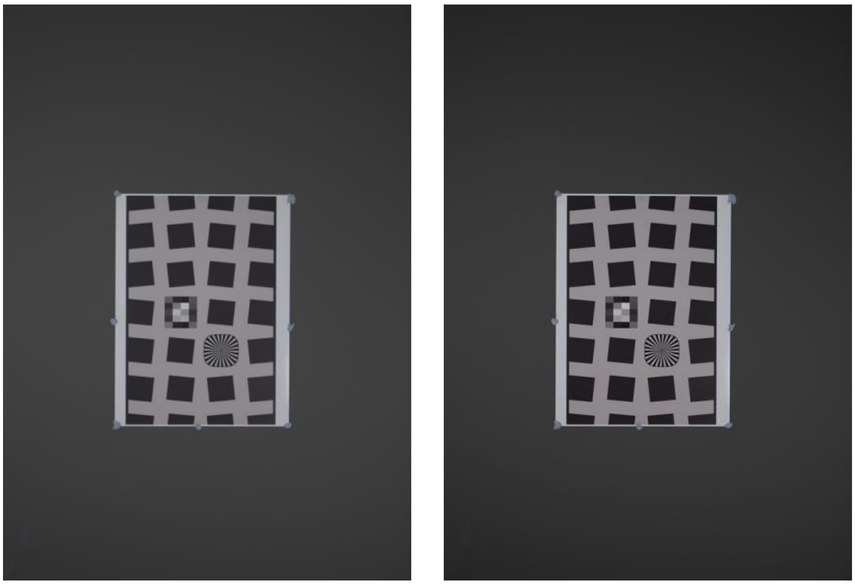


Figure 5.9 Visual comparison of one of the images used in the OECF measurements processed linearly before applying the image-processing pipeline (left hand side) and non-linearly after the image-processing pipeline (left hand side). Note that the images have been compressed for this document.

Table 5.3 shows the effective gamma values for the red, green and blue channels and the R^2 coefficients of the functions fitted to the data, for the test images before and after the image-processing pipeline.

	Full resolution, linear tone reproduction		**Fully processed, down-sampled image**	
Colour Channel	**Gamma (γ)**	**R^2**	**Gamma (γ)**	**R^2**
Red	0.4049	0.9993	0.5004	0.99627
Green	0.4171	0.9997	0.5192	0.99777
Blue	0.3974	0.9996	0.4892	0.99787

Table 5.3: Derived gamma values for the RGB channel responses for target image before and after processingFigure 5.9 illustrates the difference in appearance of the test images before and after the image-processing pipeline was applied.

It can be seen from the graphs of the functions that the three colour channels are fairly consistent in their tone reproduction, but that the normalised responses do not perfectly match across the three channels, with a relative increase in the response in the red channel at higher luminances, which is most pronounced above a normalised luminance of 0.5, and a lower relative response in the green channel at lower values, which is a reflection upon the increased gamma value in the green channel, and therefore slightly higher dynamic range. This deviation is more evident after processing than before and can be seen particularly in Figure 5.8. The R^2 values are lower for the data produced after processing, indicating that the functions used to calculate gamma values did not fit the data as well as they did on the linearly processed images.

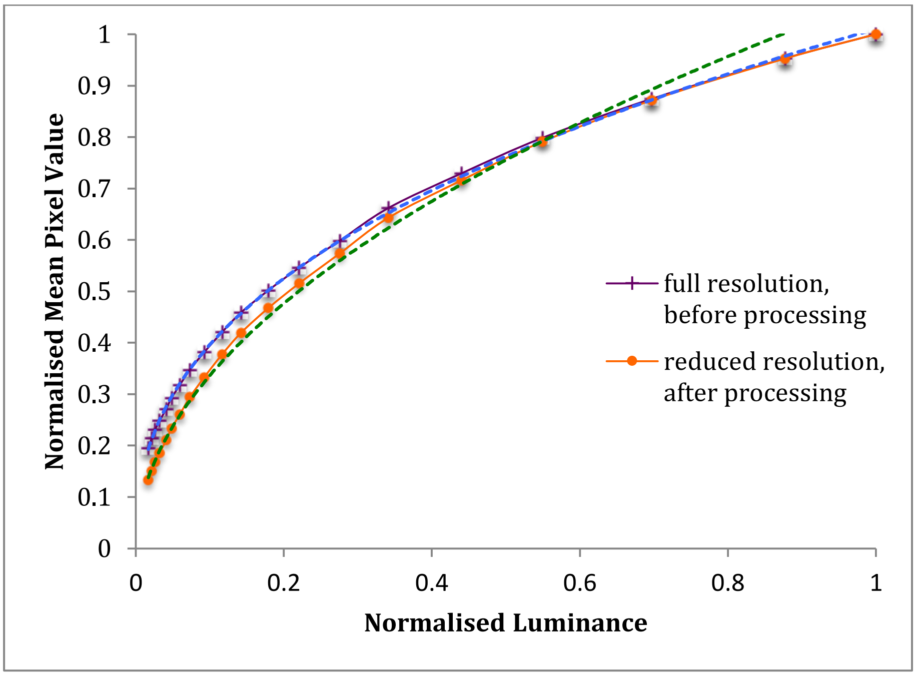


Figure 5.10 Camera OECF for combined RGB channels before and after processing expressed in linear units. Dashed lines indicate power function trend lines fitted to the data

The combined transfer function for the three colour channel responses before and after the image-processing pipeline was taken by averaging (equally weighting) the RGB values for each patch. The results are plotted in linear units in Figure 5.10, with dashed lines (blue for full resolution before processing, green for after processing) indicating the trend lines used to fit each data set. The gamma value for the combined channels prior to processing was 0.4064, with an R^2 value of 0.99927; after processing the gamma value was 0.5027 with an R^2 value of 0.99747. The difference in R^2 values is seemingly small (0.00228), however the trend lines in Figure 5.10 are interesting. The trend line for the data before image processing (blue dashed line) is consistent with the plotted data across the whole dataset, however the trend line for the data after image processing (green dashed line) appears to fit the data reasonably well at low luminance values, but deviates quite significantly from the data from 0.2. This is not unexpected, as some non-linearities have been introduced by the processing. The medium contrast tone curve applied to the images in the raw processor for example, was s-shaped (see Figure 5.11), with some tonal compression at very low luminances and very high luminances and tonal expansion in the centre of the range. The shape of the OECF of the processed images corresponds to this, with an increased dynamic range between 0.1 and 0.5 normalised luminance, compared to that of the linear processed response. Because there are a greater number of measurement points in the lower part of the response (in linear units, because the target contained patches spaced in approximately equal logarithmic units), the trend-line fits well to this data, and less well to the sparsely populated higher values.

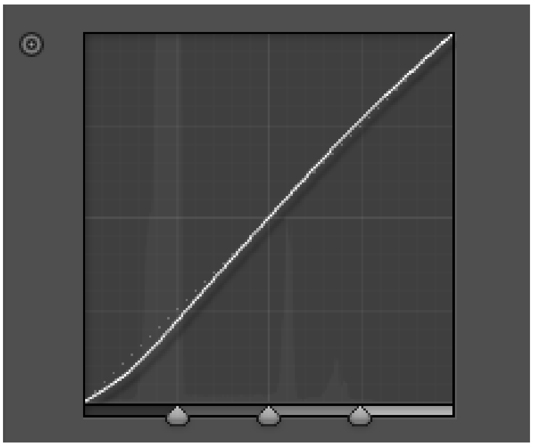


Figure 5.11: Medium contrast tone curve applied to images in image processing pipeline. This was the only tone correction applied to images of grey scale patches for calculation of OECF

The overall difference before and after processing is better illustrated in the log-log graph in Figure 5.12. The function produced from the image after processing is much steeper, indicating higher contrast, as expected from the higher gamma value, and it can be seen that a single linear function does not fit the data well (again indicated by blue and green dashed lines).

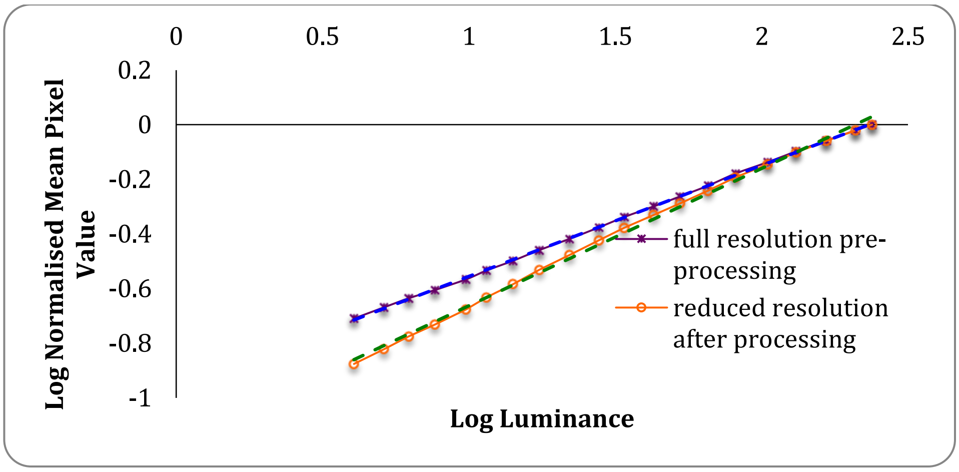


Figure 5.12: Camera OECF for combined RGB channels before and after processing expressed in $\log_{10}$ units

5.3.4 Display Characterisation

Another PhD student within the department characterised the display in accordance with *BS EN 61966 part 4: Equipment Using Liquid Crystal Display Panels* [153] [89]. The author assisted with measurements for the OECF, which are detailed in the next section, and the results from the remaining measurements in relation to the standard are included in Appendix B. The measurement of the SFR of the display is detailed in chapter 6.

The measurement and calibration of the display device was carried out under controlled conditions in accordance with BS EN 61966:4-2000 [153], as shown in Table 5.4.

Devices	
Display:	EIZO CG245W 21.4”
PC:	Dell Optiplex 760 with an ATI Radeon HD 3450 graphics card
Calibrator:	GretagMacbeth i1Pro display calibrator & Built-in calibration sensor
Colorimeter:	Konica-Minolta CS-200 tele-chromameter
Environmental Conditions	
Temperature:	20 degrees celcius
Relative Humidity	N/A
Illumination:	Total darkness
Warm up time:	1 hour
Object distance:	150 cm (Effective screen height: 32.4 cm, width: 51.84cm)

Table 5.4: Devices and environmental conditions used in display characterisation

A series of uniform patches of pixel values, with values as defined in [153] were displayed in the centre of the screen, and measurements taken using the Kodak-Minolta CS-200, placed parallel to the screen, at a distance of 150 cm, set with a 0.2° field of view.

Measurements were taken from four sets of 32 generated uniform patches, one set for each colour channel (in which only that channel was on, and the other two channels were set to zero), and one set of neutral patches (equal values in all three channels). The patches spanned a range of values from 8-255 in increments of 8. Each patch was 240 x 240 pixels.

Three measurements of luminance and tristimulus values for each patch were taken, and the results were averaged for each. The normalised luminance values are plotted against normalised input pixel values for the red, green and blue colour channels in Figure 5.13.

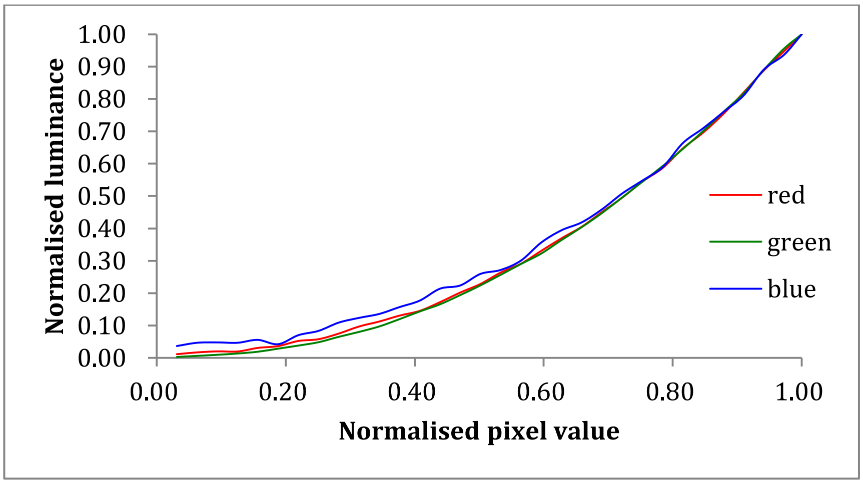


Figure 5.13 Tone reproduction characteristics of EIZO CG245W, plotted in linear units

The measurements taken from the neutral patches are plotted in Figure 5.14. The display gamma value ($\gamma_{display}$= 2.1849) was obtained from this curve, with a power function fitted to pixel values of 48 and above (normalised to 0.188).

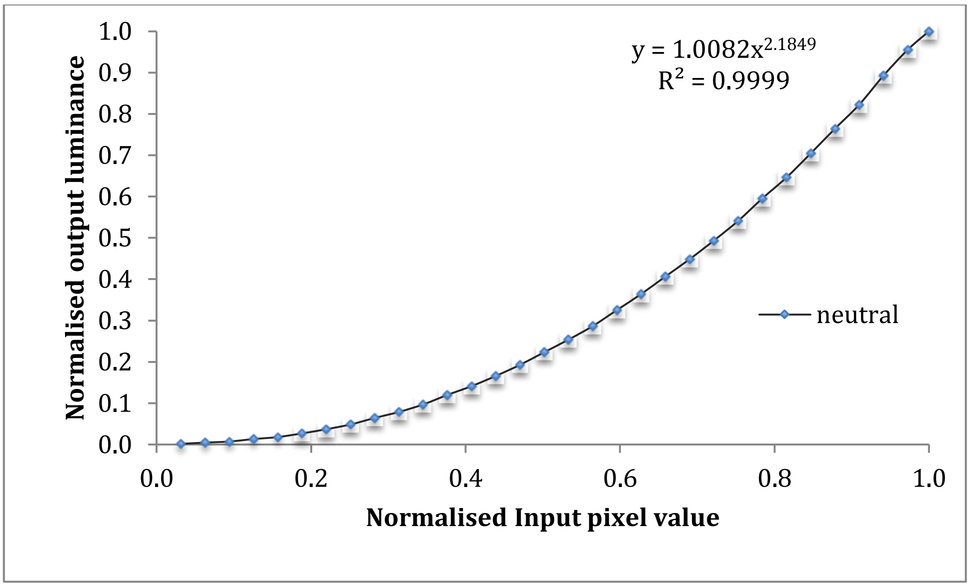


Figure 5.14 Tone reproduction characteristics of EIZO CG245W, measured from neutral displayed patches, plotted in linear units

5.4 Psychophysical Experiment

5.4.1 Quantification of Scene Characteristics

An aim of this experimental work was the exploration of the relationship between scene characteristics and thresholds. Based upon previous work by Triantaphillidou et al [34] and Hoon et al [35], a series of scene metrics were selected to classify the scenes and seek correlations with the results from the psychophysical experiment.

The optimized images were converted to CIELAB for scene analysis. A range of simple image analysis tools were used [7], to evaluate and rank selected scene characteristics in the test images and provide relevant visual scene descriptors (i.e. metrics used to quantify each scene characteristic). These included:

First-order statistical measures (median *md*, variance *V*, and skewness *s*); derived from the probability density function (PDF) of the L* channel [154]:

$$Variance = V = \sum_{a=1}^{L-1} P(a)(a - \bar{a})^2 \qquad (5.2)\ [154]$$

Where a is a particular pixel value, $P(a)$ is the probability of a pixel taking value a (estimated from the normalised histogram), L is the number of levels in the channel, and $\overline{a}$ is the mean pixel value.

$$skewness = s = \frac{1}{{\sigma_a}^3}\sum_{a=1}^{L-1} P(a)(a-\bar{a})^3 \qquad (5.3)\ [154]$$

A busyness metric developed by Triantaphillidou [34], involving the following steps, reproduced here for completeness:

(i) The L* channel of the image was filtered with horizontal and vertical Sobel filter masks and a threshold of 0.04.

(ii) The resulting gradient image was thresholded and dilated with a line shaped structuring element of length three pixels in horizontal and vertical directions. This filled gaps in the detected edges and amplified the 'busy' areas (detected edges).

(iii) A flood filling operation was applied to fill in 'holes' in the identified busy regions.

(iv) The result was eroded using a diamond shaped structuring element to eliminate noisy pixels.

The stages of the metric were implemented in MATLAB [155] and are illustrated in Figure 4.9. The output of the metric was a ratio of detailed areas to overall image area.

Chroma variance *VC*_{ab},* the variance of CIELAB C*_{ab} ; derived from the a* and b* channels.

These measures allowed the images to be broadly classified according to their overall lightness (*md, s*), global contrast (*V*), spatial content / amount of detail (*busyness metric, b*) and colour contrast (*VC*_{ab},*). Histogram skewness was included as a measure of global scene lightness, because as well as correlating with median values, it indicates a predominance of light or dark tones within the image (i.e. the 'key' of the scene).

5.4.2 Scene Selection

The original set of 44 images was ranked according to each of the scene descriptors and the final set of images selected as follows:

(i) The mean value for the scene descriptor was determined.

(ii) Images were classified according to whether they fell into the *average* category for the selected scene characteristic, the *greater than* or *less than average* categories (based upon distance from mean)

(iii) Scenes for were selected to ensure that all from all five scene descriptors were represented.

The final set consisted of 25 images, which can be seen in Appendix A. Figure 5.15 illustrates the categorisation of images for *md* and *s* descriptors, in this case with additional categories for very low and very high descriptor values.

Median (md) **Skewness (s)**					
Category	Extremely high	Higher than average	Average	Lower than average	Extremely Low
Range of scene descriptor	$x>\mu + 1\sigma$	$x<\mu+1\sigma$ $x>\mu + 0.5\sigma$	$x<\mu + 0.5\sigma$ $x>\mu -0.5\sigma$	$x<\mu -0.5\sigma$ $x>\mu -1\sigma$	$x<\mu -1\sigma$

Figure 5.15 Classification of scenes into five categories for *md* and *s* and category bounds for the scene descriptors

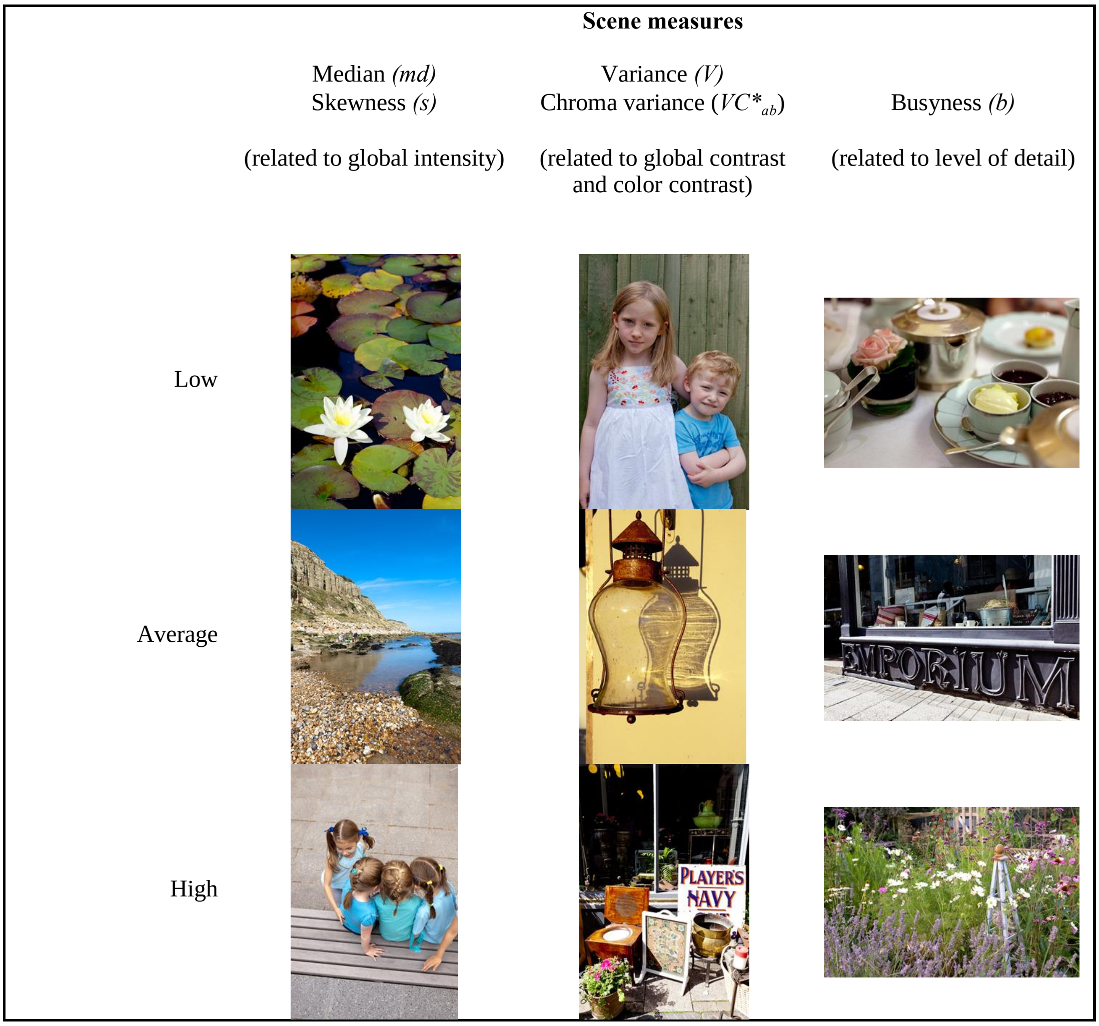


Figure 5.16 Examples of scenes from the three classes (less than average, average, greater than average) for the five scene characteristics. Images are (left to right, top to bottom), *Lilies, Kids, Afternoon Tea, Cliffs, Lamp, Emporium, Huddle, Players Navy, Flower Garden.*

5.4.3 Image Compression

The processed sRGB TIFF files were compressed as JPEG 2000 files in the MATLAB environment. Default settings were used for the compression parameters (i.e. lossy compression, single quality layer, tile size equal to image size) and the images were compressed to a set of defined compression ratios (CR).

A study by the author in 2004 evaluated thresholds of perceptibility for the image set used in the investigation in the previous chapter [156]. Compression ratios for all but one of the images were found to fall in the range from 10:1 to 35:1. Therefore the selected compression ratios for this perceptibility test were 5, 10, 15, 20, 25, 30 and 40. An additional pilot test using two observers was carried out to ensure that the range was suitable under the current experimental conditions. Results indicated that three images (‘Afternoon Tea’, ‘Fred’ and ‘Bride’) had potentially high perceptibility thresholds, and therefore their range was extended up to a CR of 60:1. The sample set consisted of 190 images in total.

The acceptability test, being a suprathreshold evaluation, required a larger range of compression. The images were inspected at a range of higher CRs and 70:1 was established as a rate at which most images became unacceptable. The images were therefore further compressed to the following compression ratios: 45, 50, 55, 60, 65, and 70, resulting in a total of 150 images. The acceptability test for an individual observer consisted of all images in which the observer had noted a difference during the perceptibility test, in addition to the 150 images at higher compression ratios.

5.4.4 Psychophysical Display and Viewing Conditions

The display used for the investigation of thresholds (this chapter) and the Soft Copy Quality Ruler experiment (chapter 6) was calibrated every day during the period of the test to the sRGB standard, *BS EN 61966 part 2-1: Colour Management – Default RGB Colour Space - sRGB* [137].

The viewing environment was calibrated to closely match the specification in BS ISO 3644:2009, [157] section 4.5, with neutral surround, elimination of strong environmental colours and veiling glare, an ambient colour temperature of 5000K and an ambient illuminance of 64 lux, measured on a Minolta Chromameter CL200.

5.4.5 Paired Comparison Test

Perceptibility and acceptability thresholds were evaluated through a two-part paired comparison test. The test interface was developed using MATLAB. The test images were displayed side-by-side, one compressed and the other an uncompressed original. The images were presented in a random sequence and the original and compressed versions were randomized in their presentation on the left or right of the screen. The effective screen size was 518.4 mm wide by 324.0 mm; each image took up approximately 45% of the half-screen area on a mid-grey background. Image size was selected to ensure that there would be no interpolation when they were displayed. The viewing distance was fixed at 60cm, giving an angle of subtense of 22.450 degrees of arc (0.392 radians). The time to view the images was unrestricted, the observer controlling when they would move on to the next image using a push button.

The perceptibility test was performed first and based upon the threshold identified by the observer, the set of images to be used for the acceptability test were identified. Any images in the set that were above the perceptibility threshold were then presented to the observer in the acceptability experiment. This meant that observers had different sets of compressed images for the acceptability experiment, tailored to their sensitivity level.

Twelve experienced observers carried out the perceptibility test, and eleven of them completed the acceptability test. All had normal, or corrected vision.

In the first section of the test, observers were asked to provide a 'yes' or 'no' answer, to the question of whether they could perceive a difference between the two displayed images of the same scene. Observers were given the opportunity to take a break between the perceptibility and acceptability tests and were asked to stop if they felt tired. In the second half of the test, observers were asked whether they found the compressed image acceptable when compared to the original.

The PSE is defined as the statistical point at which observers perceive two images to be equal. Here the PSE corresponded to the compression rate at which 50% of the observers responded 'yes' to perceiving a difference between the two images. Corresponding to this, the absolute acceptability threshold is the point at which 50% of the observers find the images unacceptable (i.e. respond 'no' to the question 'Do you find the image acceptable.').

5.5 Determination of Thresholds

5.5.1 Functional Form of the Psychometric Curve

When dealing with a small number of observers, as is often the case in psychophysical tests, it can be difficult to generate a smooth psychometric function from the data. Therefore a functional form for the psychometric curve is hypothesized and this is fitted to the data. A detailed treatment can be found in [158], but key steps are summarised here.

The probability of a sample being judged to have more of a ness than a reference is assumed to be a linear function of the form [158]:

$$P_{js} = F(\alpha_s + \beta_s x_{js}) \qquad (5.4)\ [154]$$

Where j is the sample, s is the standard P_{js} is the probability of j being preferred over s, and $x_{js}$is the stimulus level (in this case compression rate) corresponding to P_{js} on the psychometric function. F is the function used to model the psychometric curve. Various functions can be used for F; the Gaussian model and the logistic model (applied here) are two that are widely used. The parameters of the function α_s and β_s can be determined and then used to estimate thresholds.

The logistic function is defined as:

$$P_{js} = \frac{1}{1 + e^{-(\alpha_s + \beta_s x_{js})}} \qquad (5.5)\ [154]$$

And:

$$\ln\left(\frac{P_{js}}{1 - P_{js}}\right) = \alpha_s + \beta_s x_{js} \qquad (5.6)\ [154]$$

The proportions of responses were used to generate estimated psychometric curves using the Palamedes MATLAB® Toolbox [159] using the logistic function.

The psychometric curves directly related the proportion of observers' responses to the original compression ratio and were used to evaluate various points of interest: the point of subjective equality (PSE), or absolute threshold, defined as the 0.5 proportion point; the just-noticeable-difference (JND) (defined by convention as the stimulus increment between the PSE and the 0.75 proportion); and the detection threshold, defined here as the 0.75 proportion point.

The PSE is defined as the statistical point at which observers perceive two images to be equal [146,160]. Here the PSE corresponded to the compression rate at which 50% of the observers responded 'yes' to perceiving a difference between the two images. Corresponding to this, the absolute acceptability threshold was the point at which 50% of the observers find the images unacceptable (i.e. responded 'no' to the question 'Do you find the image acceptable.').

The compression ratio identified at the 0.75 proportion was the point at which 75% of observers could either perceive a difference between the two images, or found the differences unacceptable; in this study it is this value that is referred to as the threshold of perceptibility/acceptability.

5.5.2 Error estimation and goodness of fit of the psychometric curve

The curve fitting procedure produced a maximum likelihood estimate of the parameters (α, corresponding to threshold at a probability (P) of 0.5 and β, corresponding to the slope) of the psychometric curve for each image, based upon the observer responses across the compression range [160]. A goodness-of-fit test was performed when the estimated curve was generated, based on 1000 simulations of the data [160] and resulting in a ρ-value, the probability that the observed data could be part of the population generated from the estimated model. ρ-values of below 0.05 were deemed to be an unacceptably poor fit. The results of the goodness-of-fit test were used as a means of determining the approach to be used in the error estimation.

To estimate standard error, for a ρ-value of greater than or equal to 0.05, a bootstrapping procedure was again used to generate 400 hypothetical sets of data, based on the parametric description of the observed experimental data (i.e. the estimated curve). A logistic function was fitted to each set of simulated data to derive the new α and β parameters. Finally the sample standard deviation (i.e. the standard error, equal to population standard deviation divided by number of samples) for each parameter was calculated from their distributions in the simulated sets of data. This *parametric bootstrap* approach used the estimated curve as the starting point and its parameters were the mean values from which the standard deviation was calculated.

In the case of images where the ρ-value <0.05, a *non-parametric bootstrap* was used to evaluate the errors. The experimental data was used instead of the parameters of the curve in the simulations, obtaining hypothetical data based on actual data rather than an 'average' estimated function [160]. The standard deviation for the α and β parameters was again determined from the sampling error across the simulated datasets.

The error estimation procedures did not generate an accurately fitted function ('failed fits') for a few of the images, Figure 5.17. In these cases, the failed datasets were excluded in the calculations of standard error, as their estimated parameters were deemed an inaccurate fit to the data and would have biased the results.

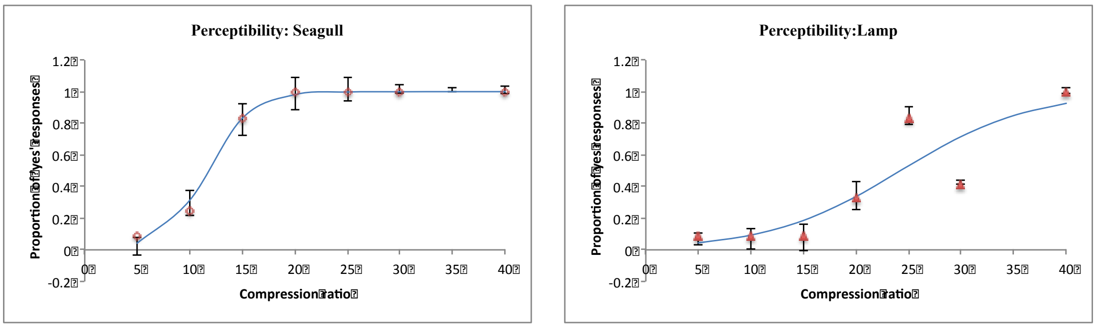


Figure 5.17 Examples of psychometric curves: Seagull (top), ρ-value 0.559 (good fit) error bars fitted to curve, Lamp (bottom) ρ-value 0.026 (poor fit) error bars fitted to data

5.6 Results

5.6.1 Scene ranking from objective measures

The images were ranked according to each scene descriptor. The correlations between the individual scene descriptors were evaluated to determine whether there was a predictable relationship between them for the sample image set. They were calculated from the original scenes using Spearman's correlation coefficient [161]$r = \frac{6\sum d^2}{n(n^2-1)}$, where d=the difference in rank between the two descriptors for each image, and n is the total number of scenes. For 24 degrees of freedom (from n-1), the coefficient has a greater than 95% chance of being significant at a value > 0.406 [161].

The results are shown in Table 5.5. The coefficient of the correlation between median and skewness (in bold) is the only coefficient that indicates a significant correlation (for this number of images). This is an expected result as a histogram skewed in one direction or another will have a median value in the same direction. The results for all images and scene characteristic ranks are shown in Table 5.6.

Scene Measure		Median	Skewness	Variance	Busyness	Chroma Variance
		m	*s*	*V*	*b*	***VC*_{ab}***
Median	*m*	1.00				
Skewness	*s*	**0.89**	1.00			
Lightness Variance	*V*	-0.18	-0.32	1.00		
Busyness	*b*	-0.05	-0.24	0.33	1.00	
Chroma Variance	*VC*ab*	-0.29	-0.33	0.03	0.30	1.00

Table 5.5 Correlations between scene descriptors

Image		**Rank**		**Rank**	%	**Rank**		**Rank**		**Rank**
	md	***md***	*s*	***s***	*b*	***b***	*V*	***V***	VC^*_{ab}	VC^*_{ab}
accordion	120	6	0.05	9	49.67	11	5760.00	19	144.64	12
afternoon tea	154	18	-0.80	22	2.59	1	3712.70	7	123.24	10
beach goods	128	11	0.17	4	62.80	20	5573.10	17	405.31	24
bride	138	14	-0.40	17	18.23	5	4745.80	14	252.32	20
cliffs	140	16	-0.25	14	61.95	19	2899.60	5	338.50	23
crockery	125	9	-0.10	11	74.24	23	6625.40	21	175.04	15
crown antiques	172	22	-0.48	18	65.23	21	8644.40	24	226.21	17
emporium	60	1	0.54	1	55.49	14	7884.70	23	34.66	1
flags	154	19	-0.73	21	9.60	4	1750.60	1	133.30	11
flower garden	127	10	0.07	7	93.38	25	2863.10	4	270.20	21
formal	129	12	0.13	5	59.11	17	5585.10	18	246.59	18
fred	205	24	-1.42	24	21.61	6	6043.70	20	65.57	2
hive beach	109	3	0.45	3	29.86	8	4743.50	13	118.20	9
huddle	186	23	-1.60	25	23.26	7	1843.60	2	100.49	6
kids	134	13	-0.06	10	6.43	3	2491.90	3	70.98	3
lamp	214	25	-1.36	23	43.50	10	4274.80	10	152.37	13
lilies	110	4	0.11	6	35.22	9	4425.80	12	303.90	22
marle sculpture	158	20	-0.32	15	50.43	13	7565.20	22	79.43	4
pink flowers	124	7	-0.24	13	49.80	12	2977.50	6	250.70	19
players navy	60	2	0.46	2	61.60	18	9484.10	25	443.74	25
pool	153	17	-0.63	19	59.07	16	4984.30	15	208.66	16
seagull	139	15	-0.33	16	72.02	22	4110.10	8	104.93	7
serpent	124	8	-0.19	12	56.68	15	4157.30	9	115.77	8
stones ii	167	21	-0.64	20	84.05	24	5002.30	16	98.19	5
summer	119	5	0.05	8	3.04	2	4300.60	11	157.97	14
average	13.2		-0.32		47.74		4954.05		185.96	
SD	7.2		0.56		24.05		2094.79		107.23	

Table 5.6 Images ranked according to objective measures: Median (*md*), Skewness (*s*), % Busyness (*b*), Variance (*V*), Chroma Variance (VC^*_{ab})

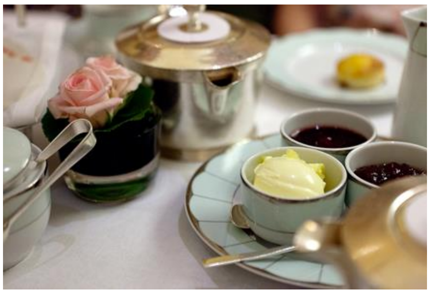
5.6.2 Perceptibility and acceptability thresholds

The results for the thresholds for perceptibility and acceptability for all images are shown in Table 5.7. The images are presented as ranked according to their perceptibility thresholds. The ranks indicate a not unexpected positive correlation between the perceptibility and acceptability thresholds, which can also be seen in Figure 5.19. Of interest is the relationship between perceptibility and acceptability thresholds, particularly in images where there are significant differences in their rankings for perceptibility and acceptability.

	Perceptibility Thresholds				**Acceptability Thresholds**			
	PSE	**Threshold**	**JND**	**Rank**	**PSE**	**Threshold**	**JND**	**Rank**
	P(0.5)	P(0.75)	P(0.75)-P(0.5)		P(0.5)	P(0.75)	P(0.75)-P(0.5)	
Afternoon Tea	44.8	54.6	9.8	**25**	NA	NA	NA	**24***
Fred	34.0	47.9	13.9	**24**	NA	NA	NA	**25***
Summer	24.9	32.4	7.5	**23**	36.5	48.2	11.8	**22**
Lamp**	24.2	31.1	6.9	**22**	44.0	55.9	11.9	**23**
Lilies	23.1	29.7	6.6	**21**	38.0	48.1	10.1	**21**
Huddle	19.3	25.5	6.2	**20**	30.1	36.2	6.1	**16**
Bride	16.6	24.4	7.7	**19**	34.5	45.7	11.2	**20**
Flags	19.2	24.0	4.8	**18**	28.0	31.5	3.5	**11**
Emporium	17.9	22.3	4.3	**17**	34.6	44.6	10.0	**19**
Pink Flowers	17.5	21.7	4.2	**16**	25.8	30.7	4.9	**10**
Kids	14.9	21.6	6.7	**15**	32.5	40.4	7.9	**17**
Serpent	15.5	21.6	6.1	**14**	26.2	35.1	8.9	**14**
Crockery	17.0	21.6	4.6	**13**	26.3	33.2	6.9	**12**
Accordion	14.7	20.4	5.7	**12**	31.2	41.5	10.3	**18**
Marle Sculpture	16.2	20.2	4.0	**11**	26.2	33.7	7.5	**13**
Pool	14.5	18.2	3.6	**10**	24.1	36.2	12.0	**15**
Flower Garden	12.5	17.4	4.8	**9**	22.1	25.4	3.3	**5**
Hive Beach	13.7	16.1	2.4	**8**	20.8	25.6	4.8	**6**
Formal	12.9	16.0	3.1	**7**	20.7	24.2	3.5	**4**
Beach Goods**	12.4	15.7	3.3	**6**	21.4	27.6	6.2	**8**
Crown Antiques	11.3	15.6	4.2	**5**	20.9	29.8	8.9	**9**
Players Navy	13.8	15.4	1.6	**4**	22.2	27.6	5.4	**7**
Seagull	11.6	13.9	2.3	**3**	18.9	21.7	2.8	**2**
Stones II**	11.2	13.3	2.0	**2**	17.5	21.9	4.5	**3**
Cliffs**	8.0	9.3	1.3	**1**	12.8	16.7	3.9	**1**

Table 5.7 Perceptibility and acceptability thresholds for all images. Ranks are based on the (0.75) threshold.

Two scenes with the highest rank in terms of perceptibility threshold (*Afternoon Tea*, and *Fred*, marked with an * in table 3) do not have corresponding values for acceptability thresholds. These images proved to be extremely robust under JPEG 2000 compression. Neither reached the 0.75 proportion point for acceptability at the maximum compression ratio evaluated, meaning that psychometric curves and derived thresholds could not be estimated. The order of their ranking in terms of acceptability is assumed, based upon the proportions of observers who gave 'no' responses at the maximum compression ratio tested (0.4 for *Fred* and 0.7 for *Afternoon Tea*).



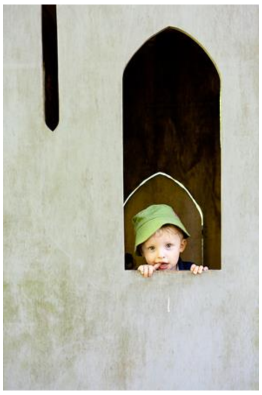


Figure 5.18 *Afternoon Tea* (left) and *Fred* (right). These were images which had high perceptibility thresholds and did not reach the threshold of acceptability for the given compression range. These images might be classified as the least susceptible scenes.

Four images (marked with **) are highlighted in Table 3 because they had ρ-values below the 0.05 threshold point in the perceptibility test, indicating that the estimated curve was of an unacceptably poor fit.

5.6.3 Grouping Scenes

Because significant correlations had not been found between the rankings of the scene descriptors for this particular image set, the thresholds were instead used to find groupings.

The perceptibility and acceptability JND thresholds (p=0.75) were evaluated using k-means clustering. Various numbers of cluster groups and distance measures were evaluated, and the final clustering used 7 groups and a squared Euclidean distance measure. The scenes in each group were then compared in terms of their scene measures to search for correlations and common characteristics.

Figure 5.19 illustrates the relationship between perceptibility and acceptability thresholds and shows some of the identified groups.

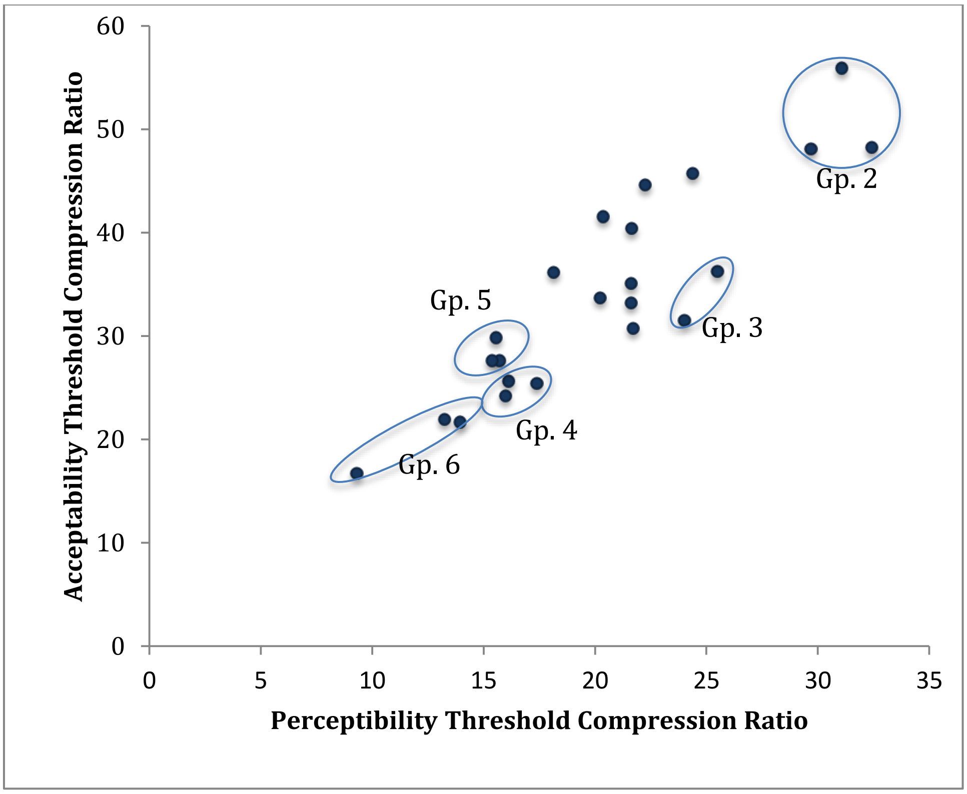


Figure 5.19 Relationship between perceptibility and acceptability threshold compression ratios.

5.6.4 Correlations between thresholds and scene metrics

Spearman's rank correlation coefficients [161] were calculated to examine the relationships between the perceptibility and acceptability thresholds with each scene descriptor. The results are presented in Table 5.8.

	Perceptibility	**Acceptability**
md	0.09	0.09
V	-0.25	0.03
s	0.32	0.27
b	**-0.80**	**-0.73**
Vc	-0.28	-0.30

Table 5.8 Spearman's correlation coefficients calculated between scene metrics and subjective thresholds.

The high negative coefficients corresponding to the busyness metric with both subjective measures are significant. The negative sign of the coefficient indicates that as scene busyness increases, the thresholds of perceptibility and acceptability decrease. This implies scene dependency of the JPEG 2000 algorithm, meaning that it performs less well in images containing lots of detail. Figure 5.20 illustrates the effects of the algorithm on the most and least busy of the images. The majority of scene measures suggest weak correlation with the subjective thresholds. This implies that scene global lightness, contrast, and colour contrast do not play a significant role in JPEG 2000 compression when considered across the entire sample set of images, although correlations may exist within the image groups indicated on Figure 5.19, as shown in Table 5.9.

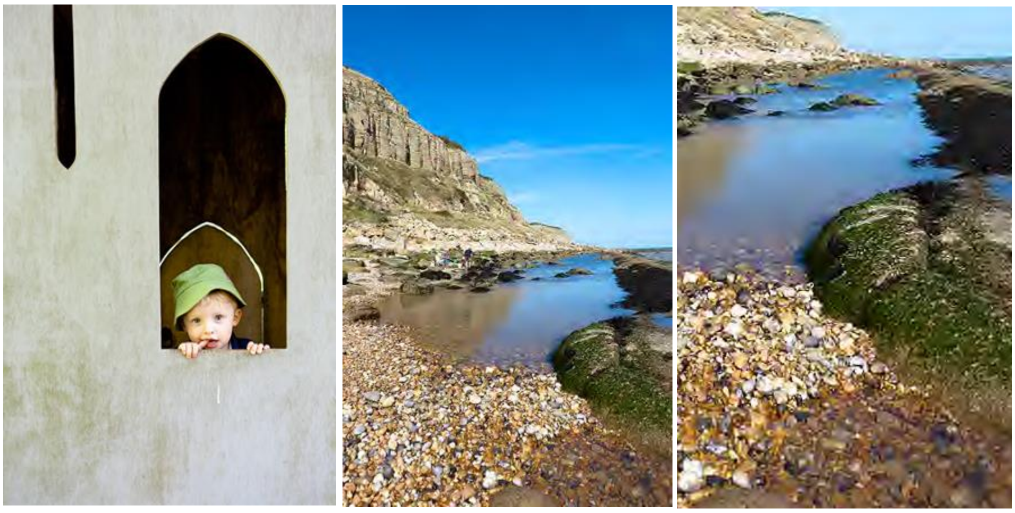


Figure 5.20 High and low ranking images in terms of the busyness metric. From left to right: *Fred*, compression ratio 70, perceptibility 54.5, acceptability threshold not reached; *Cliffs*, compression ratio 70, perceptibility 9.3, acceptability threshold 16.7; Close up of *Cliffs* illustrating significant distortion.

5.6.5 Scene grouping and correlations

Images were grouped in terms of their threshold levels, and the relationship between their perceptibility and acceptability thresholds. Correlations between scene descriptors or scene content within groups were identified. The thresholds and results from the visual scene descriptors for the six groups are shown in Table 5.9. The remaining images were not found to correlate strongly with one another.

	Perceptibility CR	Acceptability CR	Median *m*	Skewness *s*	Busyness *b*	Variance *V*	Chroma Variance VC^*_{ab}
Group I							
Afternoon Tea	54.56 (25)	BR	154 (18)	-0.80 (22)	2.59 (1)	3712.70 (7)	123.24 (10)
Fred	47.87 (24)	BR	205 (24)	-1.42 (24)	21.61 (6)	6043.70 (20)	65.57(2)
Group II							
Lamp	31.07 (22)	55.92 (23)	214 (25)	-1.36 (23)	43.50 (10)	4274.80 (10)	152.37(13)
Summer	32.43 (23)	48.22 (22)	119 (5)	0.05 (8)	3.04 (2)	4300.60 (11)	157.97 (14)
Lilies	29.70 (21)	48.11 (21)	110 (4)	0.11 (6)	35.22 (9)	4425.80(12)	303.90 (22)
Group III							
Huddle	25.50 (20)	36.24 (16)	186 (23)	-1.60 (25)	23.30 (7)	1843.60 (2)	100.49 (6)
Flags	24.02 (18)	31.53 (11)	154 (19)	-0.73 (21)	9.60 (4)	1750.60 (1)	133.30 (11)
Group IV							
Flower Garden	17.39 (9)	25.41 (5)	127 (10)	0.07 (7)	93.38 (25)	2863.10 (4)	270.20 (21)
Formal	15.99 (7)	24.21 (4)	129 (12)	0.13 (5)	59.11 (17)	5585.10 (18)	246.59 (18)
Hive Beach	16.1(8)	25.6(6)	109(3)	0.45 (3)	29.86(8)	4743.50(13)	118.20(9)
Group V							
Beach Goods	15.72 (6)	27.63 (8)	128 (11)	0.17 (4)	62.80 (20)	5573.10 (17)	405.31 (24)
Players Navy	15.39 (4)	27.60 (7)	60 (2)	0.46 (2)	61.60 (18)	9484.10 (25)	443.74 (25)
Crown Antiques	15.56 (5)	29.83 (9)	172 (22)	-0.48 (18)	65.23 (21)	8644.40 (24)	226.21 (17)
Group VI							
Seagull	13.94 (3)	21.67 (2)	139 (15)	-0.33 (16)	72.02 (22)	4110.10 (8)	104.93 (7)
Stones II	13.25 (2)	21.94 (3)	167 (21)	-0.64 (20)	84.05 (24)	5002.30 (16)	98.19 (5)
Cliffs	9.31 (1)	16.71 (1)	140 (16)	-0.25 (14)	61.95 (19)	2899.60 (5)	338.50 (23)
Group VII							
Serpent	21.6 (14)	35.08 (14)	124(8)	-0.19 (12)	56.68 (15)	4157.3(9)	115.77(8)
Marle Sculpture	20.2(11)	33.7 (13)	158(20)	-0.32(15)	50.43 (12)	7565.20 (22)	79.43 (4)
Pool	18.2(10)	36.2 (15)	153 (17)	-0.63 (19)	59.07 (16)	4984.30 (15)	208.66 (16)

Table 5.9 Perceptibility and acceptability thresholds and objective scene descriptors and ranks for images in groups. Numbers in parentheses indicate rank. BR=Beyond Range

Some adjustment was made to groups based upon their content. The images in groups did not all exhibit similar values or rankings when considering scene descriptors. However it was clear that their susceptibilities to compression within the groups linked them. This is more apparent in Figure 5.21, which shows the grouped images and the similarity in their thresholds and the relationships between the thresholds.

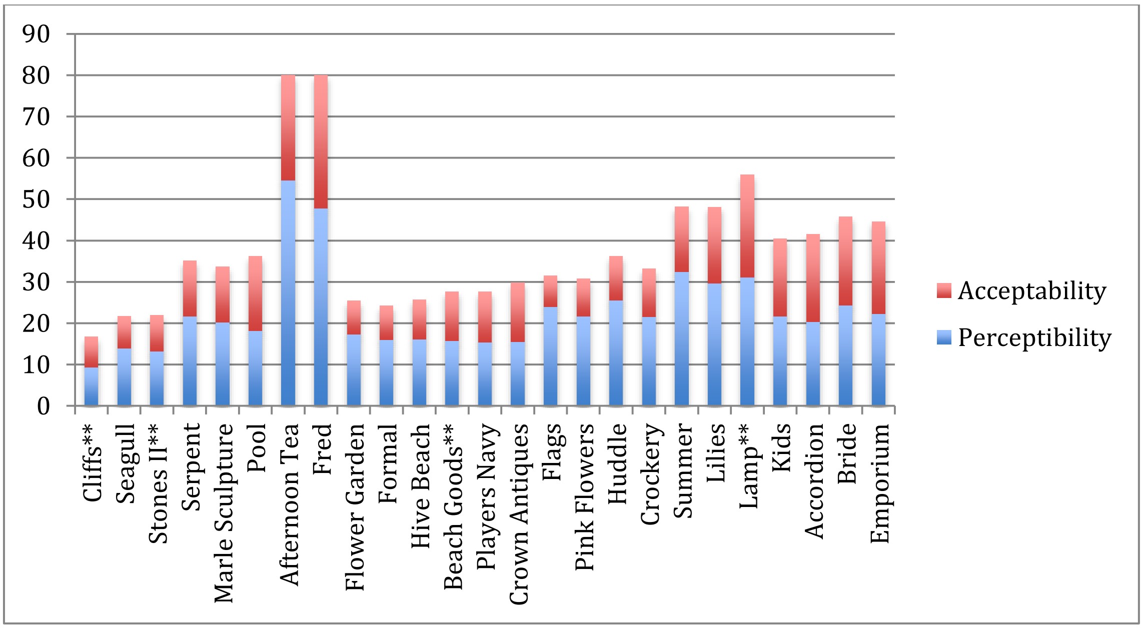


Figure 5.21 Graph showing relationships between perceptibility and acceptability for all images. Images are arranged according to kmeans cluster groupings

The cases where there was not a clear similarity in scene measures therefore warrant further consideration, as discussed later.

5.7 Discussion

5.7.1 Group 1

Images with very high thresholds for perceptibility and acceptability

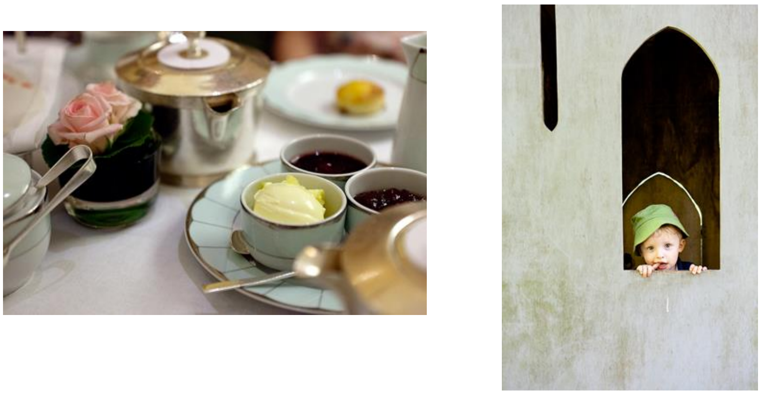


Figure 5.22 Group I images *Lamp (left), Afternoon Tea (center), Fred (right)*

Two images belong to this category: *Fred,* and *Afternoon Tea*. Both have very high acceptability thresholds (neither reached the acceptability threshold within the experimental CR range). The perceptibility thresholds for *Fred,* and *Afternoon Tea* are higher than the acceptability thresholds for 20 out of the other 23 images. *Afternoon Tea* has the highest perceptibility threshold of all the images.

The scene descriptors Table 5.9 shows similarity in the ranks for median, skewness and busyness, with high ranks for median and skewness, illustrating the predominance of light tones in the images and low ranks for busyness because of the lack of high frequency detail in each of them. The images were all average or below average for chroma variance.

The psychometric curves for the images indicated a degree of noise in the observers' responses, also confirmed by discussion with the observers after the test. Distortions were difficult to detect and did not tend to affect any of the salient features within the images. The lightness and lack of contrast in significant areas of the image also meant that there was less contrast in the distortions and the blurring distortions were less visible, which may partly account for the very high thresholds of these images.

5.7.2 Group II

Images with high thresholds for perceptibility and acceptability

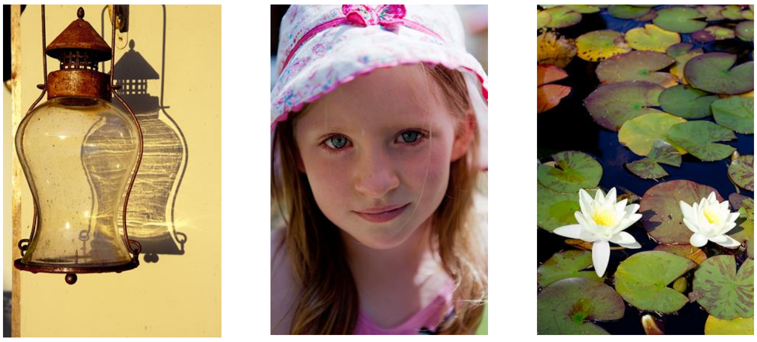


Figure 5.23 Group II images, with high thresholds for both perceptibility and acceptability *Lamp* (left) *Summer* (middle), *Lilies* (right)

The three images in this category, Figure 5.23, had high values for both perceptibility and acceptability. The scene descriptors for two of the group II images indicate similarity in terms of median and skewness (lower than average), busyness (lower than average), and variance (average). The other image, *lamp* does not correlate in this way, other than in variance, which is very similar. However, the significant features in the images are not busy, and therefore less susceptible to the localized blurring and ringing artifacts introduced by JPEG 2000. Like the images in group I, these images contain relatively large areas of higher than average or lower than average lightness. The effect on distortions on these areas is similar to that in the light areas in the images from group I; the distortion contrast is reduced and so they are potentially less visible.

5.7.3 Group III:

Images with high perceptibility thresholds but lower in acceptability threshold rank

These two images (Figure 8) show similarity across all of the scene descriptors. Both images have high median and skewness rankings, low rankings for busyness and variance, and average to low rankings for chroma variance. As for group 1, the dominant light areas of similar color and tone covering much of the image area do not appear to be very susceptible to visible distortion. At CRs beyond the perceptibility threshold for these images, the important features (the flags in the first image, the blue clothing in the second) are affected by very visible ringing as well as blurring; this may account for the reduction in acceptability once the distortion becomes visible.

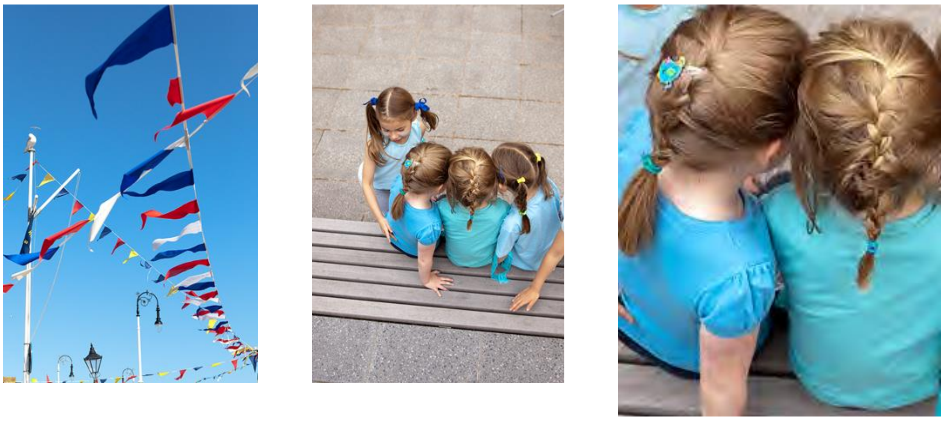


Figure 5.24 Group III images: Flags (left) Huddle (center) Huddle showing artifacts at CR 70 (right)

5.7.4 Group IV

Images with low perceptibility and acceptability thresholds and lower acceptability rank

The landscapes of group IV are dominated by natural textures. Two of the images, *flower garden* and *formal* correlated in all scene descriptors apart from variance, with high values for busyness and chroma variance and relatively low values for median and skewness. While *hive beach* did not correlate, it is clear that the scene content is equally susceptible to the artefacts. The blurring artefact is very noticeable in these scenes, particularly in the foreground areas where the textures are degraded significantly (Figures 9 and 10). The low thresholds for these images are unsurprising; the reduction in acceptability threshold rank compared to perceptibility is an indicator that the distortion is bothersome once perceived. Better scene descriptors identifying the natural textures in these images might provide more correlation.

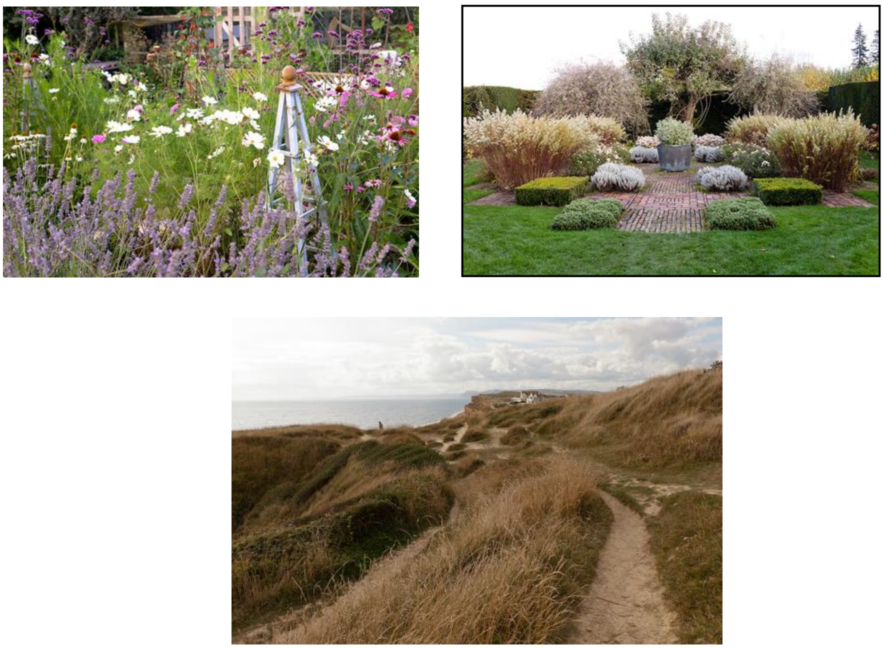


Figure 5.25 Group IV images: *Flower Garden* (top left) *Formal* (top right) *Hive Beach* (below)

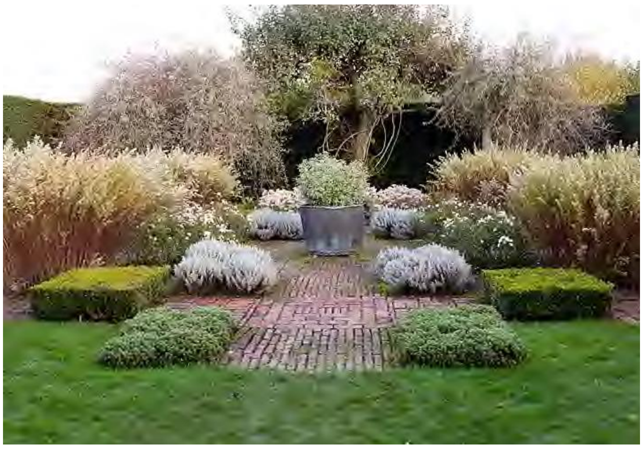


Figure 5.26 Group IV images: *Formal* at compression ratio 70, showing severe and highly visible distortion.

5.7.5 Group V

Images with low perceptibility thresholds, but an increase in acceptability rank.

These images are very low in terms of perceptibility threshold, but exhibit improved acceptability rankings. The images have high busyness, lightness variance and chroma variance rankings. As well as containing visually important textural features, which proved susceptible to distortion, all three images contain text. The areas of texture are proportionally less than those in Group VI, which may account for their comparatively improved acceptability rankings.

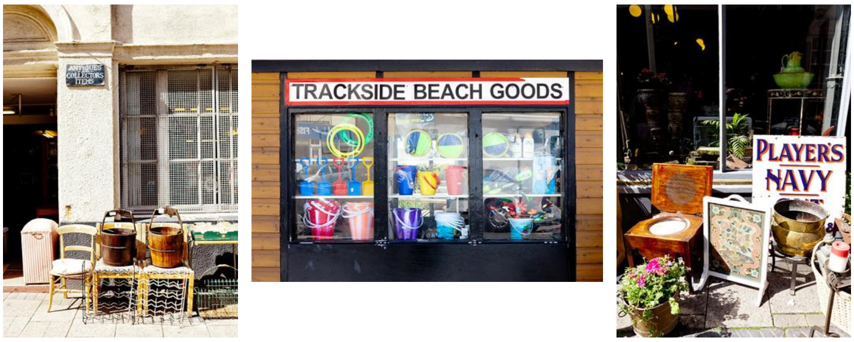


Figure 5.27 Group V images: Crown Antiques (left) Beach Goods (center) Players Navy (right)

5.7.6 Group VI

Images with very low perceptibility and acceptability thresholds

With the lowest thresholds of all the images, these three images are found to be very busy, with significant proportions of the image areas dominated by texture. The texture was affected by blurring artifacts at very low compression ratios. Because of its visual importance within the images, this must be seen as a significant factor influencing the results.

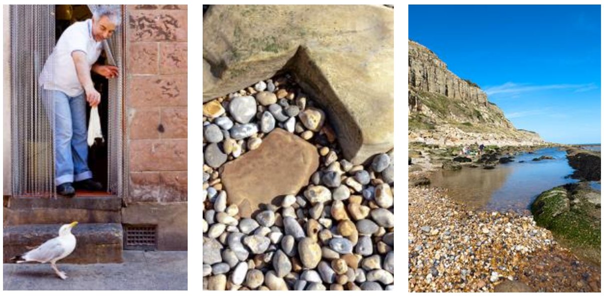


Figure 5.28 Group VI images: Seagull (left), Stones II (center), Cliff (right)

It is of interest to note that in cases where scene descriptors were not well correlated, groupings of scenes still appeared robust because of the scene content. For example, although the images in Group IV did not match on all descriptors, all had similar areas of natural texture. Equally the scenes in group II had large areas of high or low intensity low frequencies.

5.8 Summary

The experimental work in this chapter was carried out in several stages. The first stage involved the acquisition of a set of Raw images, captured upon a professional level digital SLR and processed using a carefully tested workflow to minimise the introduction of distortion from sources other than JPEG 2000 compression where possible. The tone reproduction of the imaging chain from capture to display was evaluated in stages. The images were characterised using a selection of the scene descriptors that had been used in the work in Chapter 4. The busyness metric was adapted based upon empirical observations, to the larger, higher resolution images and the change in viewing conditions in this experimental work compared to those used in the previous experiment. The images were classified based upon their position from each descriptor mean value and divided into categories according to whether they had a high, low or average measure of the descriptor. A subset of the original set of images was selected to ensure that there was a range of different types of scene content, and that all the different categories of each scene descriptor were represented. A psychophysical paired comparison experiment was carried out to evaluate thresholds of perceptibility and acceptability for the images. The results were evaluated using k-means clustering between the perceptibility and acceptability thresholds to identify groups of images that exhibited the same behaviour in terms of their thresholds (for example, with high perceptibility but relatively low thresholds and other variants). Finally the cluster groups were evaluated in terms of their scene characteristics to explore and identify further scene dependencies and their possible causes. This lead to the work in chapter 6, which used the same image set to investigate whether the soft-copy quality ruler produced similar results in terms of scene groupings.

This work further demonstrates the usefulness of some form of scene characterisation prior to image quality studies. It also explores the relationship between perceptibility and acceptability thresholds across a range of scene content. This approach might have potential as a mechanism for exploring distortion visibility and masking within particular image types. Making the images available with scene characteristic data would be useful for further research.

6 Soft Copy Quality Ruler

6.1 The Soft Copy Quality Ruler

As described in chapter 2, the soft-copy quality ruler method is a recently standardised approach to subjective image quality evaluation, specified in ISO 20462-3 [85], which allows assessment of images against a reference set, producing numerical results that are directly equivalent to interval scale relative differences in terms of JNDs.

The soft-copy ruler is a series of images of the same scene, varying in a single attribute (in this case sharpness), and spaced in known JNDs of quality [85]. The images are used as a reference (the 'ruler') against which the visual effects of imaging systems or processes may be compared and matched.

The ruler images are created using a *shaping function*, which is a filter that shapes the system MTF to a set of required aim MTFs, each corresponding to a different level of sharpness. Observers are presented with a test image and a ruler image (the start level is randomised), and using a slider, match the quality loss in the test image introduced by the process or system under investigation, with the quality loss as a result of the change in sharpness in the ruler images. In this manner, a relative JND can be defined in real time for each image, without the significant data analysis required from other methods such as paired comparison experiments.

The images used in this investigation were selected from the set used in the thresholds experiment in chapter 5, to allow comparison between the results from the two investigations, and in particular to explore the results in terms of the groupings identified in the threshold experiment. The processing pipeline prior to compression was the same, and so this needed to be accommodated in the measurements of the system MTF.

Although the task of scaling the images against the quality ruler was predicted to be faster for observers than the paired comparison experiments described in chapters 4 and 5, the experimental set-up and preparation of ruler images was more complex.

The steps involved in generating the quality rulers (one per scene) were as follows:

- Acquisition of test images
- Characterisation of imaging chain OECF and SFR
- Formation of system MTFs for required focal lengths and aperture stops
- Identification of aim MTF closest to system MTF for the particular aperture / focal length combination
- Shaping system MTF to aim MTF (if non-conforming)
- Development of filters from shaping function to modify scene sharpness by known JND increments
- Filtering of the original scenes to create a set of ruler images

The original image set (consisting of 45 images) had been captured using a range of ISO speeds, focal lengths and apertures, which meant that the system MTF was not the same for all images. Therefore a restricted range of images was selected, covering a limited range of focal lengths and apertures, whilst ensuring that images from all groups identified from the experimental work in chapter 5 were represented.

6.2 MTF Modification

The image of a scene may be modelled in the frequency domain as a scene frequency spectrum in which the magnitudes of the frequencies have been variably attenuated by the components of the imaging system. The system MTF can therefore be thought of as a form of frequency domain filter. This is relatively straightforward to understand when considering its spatial domain counterpart; the point spread function (PSF). The system PSF can be determined by the convolution of a single point of light, an impulse function (represented mathematically by the dirac delta function) with the system. Any point in the image is produced by the convolution of the system PSF with the spatial configuration of scene luminances, defined by the *imaging equation* [57]*:*

$$Q'(x_p, y_p) = \iint_{-\infty}^{\infty} Q(x,y)P(x_p - x, y_p - y)dxdy \qquad (6.1)$$

Where $Q'(x_p, y_p)$ is the output image, $Q(x_p, y_p)$is the input image and *P(x,y)* is the system PSF.

The process is shown in the top row of Figure 6.1.

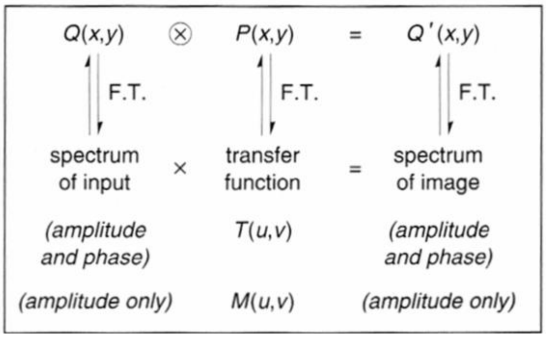


Figure 6.1 The imaging equation (convolution) and the spatial frequency equivalent, from Jenkin [57] Q(x,y) is the input scene, P(x,y) is the system PSF and Q'(x,y) is the output image. Note that T(u,v) is correctly termed the *optical transfer function* and M(u,v), the modulus of the optical transfer function is the MTF.

The PSF of the system is the result of the convolution of a number of different PSFs (from lens, sensor, processing and so on). Assuming a linear system, the convolution theorem states that 'the Fourier transform of a convolution of two functions is the product of the Fourier transforms of the same two functions' [57]. The relationships between spatial and frequency domain are illustrated in Figure 6.1. Therefore the system MTF defined by the process of cascading MTFs (which is the product of the constituent MTFs) is the equivalent in the frequency domain, of the convolution of the scene luminances with the individual component PSFs in the spatial domain. Modification of the MTF may be achieved in the spatial domain by convolution with a spatial mask, or in the frequency domain by multiplying with a filter *transfer function*. Both approaches are known as linear filtering methods, and both modify the frequency content of the image.

Convolution filtering is relatively straightforward to implement on digital images, as the convolution integral defined for continuous functions becomes the summation of products for discrete functions [162] [128]. The effects of convolution filters are determined by the filter coefficients and the shape and extent of the convolution mask. However, it is difficult to predict the exact effect upon frequencies from the spatial domain mask, and the number of arithmetic operations can become extremely large depending upon the size of the image and the extent of the filter. The alternative implementation in the frequency domain involves far fewer calculations [128]; the filter is the same size in the frequency domain as the image spectrum, and the filtered image is obtained by a point-by-point multiplication of filter transfer function and the Fourier transform of the image. Most usefully, it is possible to manipulate image frequencies very precisely. Frequencies in the image are attenuated by values between 0 and 1 in the filter and boosted by filter values of greater than 1. Low pass and high pass filters 'pass' low and high frequencies respectively, attenuating other frequencies. Band pass and band stop filters work on specific ranges of frequencies; notch filters have a very narrow stop band so can remove very specific ranges of frequencies. The shape and extent of the filter functions in the frequency domain determines the effect in the spatial domain. Some examples of frequency domain filters are shown in Figure 6.2.

Figure 6.2(a) and (b) are of interest, because although able to precisely remove some remove some frequencies whilst allowing others to pass unchanged, the shape of the filter of the filter function causes ringing artefacts in the spatial domain image. This can be can be explained because the shape of the filter is a *rect* function, and the Fourier Fourier transform of a *rect* function is a *sinc* function. Ringing appears as a rippling rippling artefact, which is particularly noticeable around high contrast edges. The abrupt The abrupt truncation of frequencies in JPEG during quantization is equivalent to the to the application of an ideal low pass filter. Many frequency domain filters suffer from suffer from ringing to some extent. Careful shaping of the filter function by selection of selection of suitable functions, or by *windowing* the function, to provide a more gradual

gradual transition to zero, can help to alleviate the problem. A highboost filter such as the such as the one shown in

Figure 6.2 (c) maintains low frequency information whilst boosting high frequencies. This is a form of sharpening filter, similar to the unsharp mask (which is usually applied through convolution in the spatial domain).

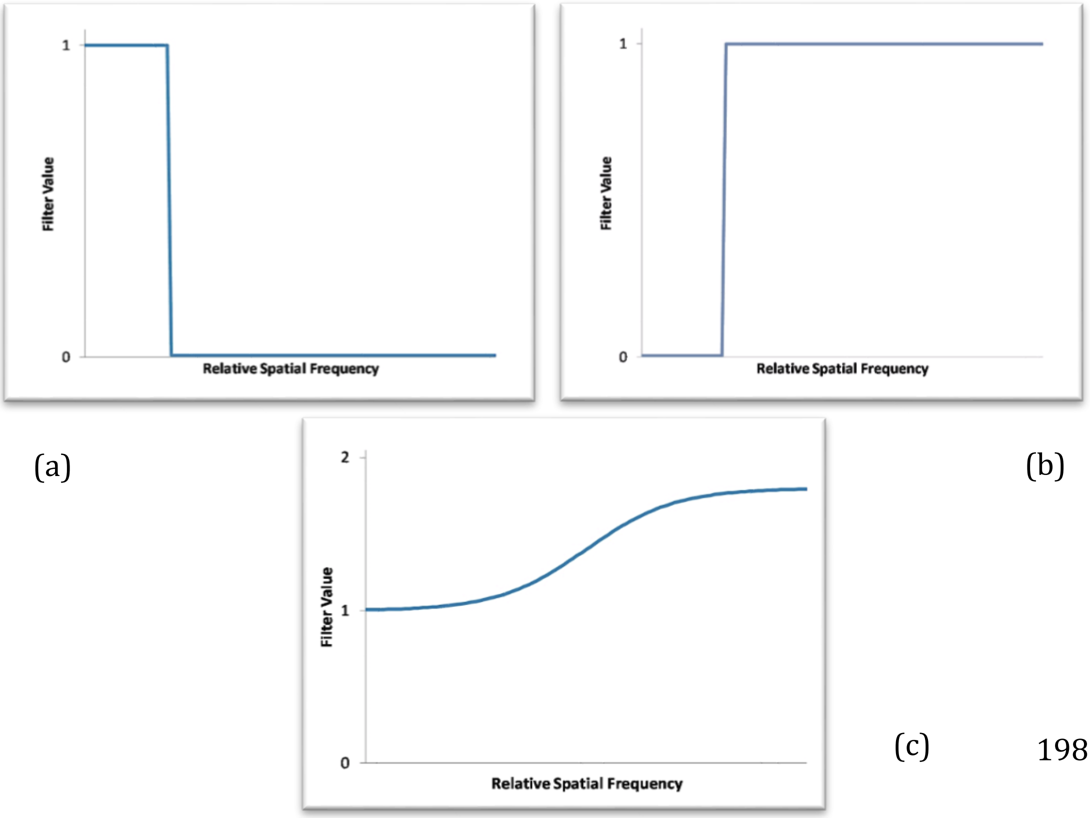


Figure 6.2 Examples of frequency domain filter transfer functions (from Jenkin [128])
(a) Ideal low-pass (b) Ideal high pass (c) High boost

6.3 Measurement of the system MTF

6.3.1 Cascading the system MTF

Each component or process within an imaging chain can be described by its own MTF, and the cascading property of MTF means that the component MTFs multiply together to produce the system MTF, assuming linearity throughout [58]. This also means that the effect of the MTF of any individual component from the imaging chain can be removed from the system by dividing the system MTF by the component MTF. This allows individual component MTFs to be modelled even if they haven't been measured in isolation, as long as enough of the other components are known.

The cascading of the imaging chain used in this experimental work is described by:

$$M(\omega)_{sys} = M(\omega)_{IS} \times M(\omega)_{L} \times M(\omega)_{P} \times M(\omega)_{DS} \times M(\omega)_{D} \qquad (6.2)$$

Where $M(\omega)_{sys}$ is the MTF of the system, IS is image sensor, L is lens, P is processing DS is downsampling, and D is display. Equation 6.1 defines the system MTF for the final processed and downsized images as a combination of the MTFs of the components, each of which might be varied. It was necessary to obtain a series of separate system MTFs for each focal length and aperture combination used.

The simplest approach to measuring the system MTF is to capture an image of a test chart using the camera and lens, process the image as the other test images were, display it on screen and photograph the image on the screen. However, this would mean that the camera-lens MTFs would be included in the system MTF twice, and would need to be extracted to correctly calculate the system MTF. It also required that the process be repeated for each aperture and focal length combination.

Jin et al [88] implemented the soft-copy quality ruler method in 2009 to evaluate images from two digital camera systems. In their experiment, they measured the display MTF from a captured point source image using a calibrated monochrome camera.

Without access to specialist equipment, the display MTF had to be measured using the camera within the imaging chain. Obtaining a good MTF from an image captured from the display proved to be a complex process, requiring careful alignment of the camera and numerous tests, to minimise the interactions between the arrangement of display pixels and the Bayer array on the image sensor, which caused spurious chromatic aliasing. Increasing the potential sources of aliasing by using a downsampled image would be likely to compound these issues. It was decided that a simpler approach would be to generate the display target, and capture an image of the displayed image, to reduce the length of the imaging chain and the effect of other component MTFs.

An alternative method was thus implemented to obtain MTFs for separate parts of the imaging chain, which were cascaded together to obtain the system MTFs, as follows:

(1) *Camera-lens MTF:* This was evaluated by measurement from an image of a printed *SFR plus* test target. The image was captured as a raw file and processed with a linear transfer function. This was repeated for all focal length/aperture combinations required.

(2) *Display MTF:* An *SFR plus* test target was generated in Imatest™ and displayed on screen before being photographed. The Camera-lens MTF was extracted from the result to obtain the display MTF.

(3) *Processed Image MTF:* The images captured in (1) were processed using the image-processing pipeline illustrated in figure 5.1. The processed images were used to obtain cascaded camera-lens-processing MTF.

(4) *System MTF:* The system MTF for use in generating the ruler images was obtained by cascading the relevant processed image MTF (3), with the display MTF (2).

Care had to be taken throughout to ensure correct translation of frequencies, with the final system MTF being expressed in cycles per visual degree.

6.3.2 SFR software and Test Target

As described in section 2.6.4, measurement of MTF for digital systems is generally achieved using a modified method of the edge input method, known as the *slanted edge method*, [58] to obtain the spatial frequency response (SFR), due to the difficulties in evaluating the MTF of sampled systems by traditional sine wave recording or edge methods. The slanted edge method has been adopted as an ISO standard; the latest version is ISO 12233: 2014 [163].

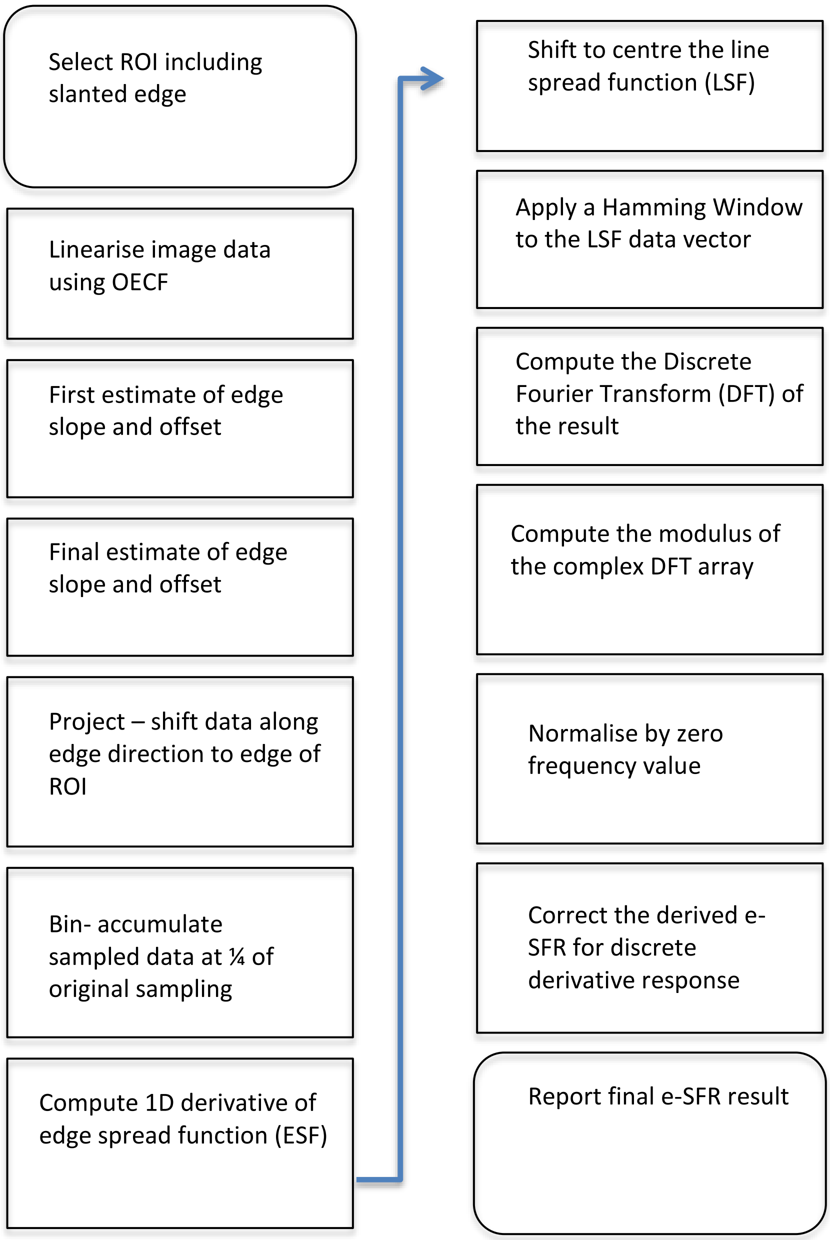

Select ROI including slanted edge
Linearise image data using OECF
First estimate of edge slope and offset
Final estimate of edge slope and offset
Project – shift data along edge direction to edge of ROI
Bin- accumulate sampled data at ¼ of original sampling
Compute 1D derivative of edge spread function (ESF)
Shift to centre the line spread function (LSF)
Apply a Hamming Window to the LSF data vector
Compute the Discrete Fourier Transform (DFT) of the result
Compute the modulus of the complex DFT array
Normalise by zero frequency value
Correct the derived e-SFR for discrete derivative response
Report final e-SFR result

Figure 6.3 Flow diagram for edge based SFR algorithm from [163]

The method uses an edge projected at a slight angle to the vertical or horizontal onto a sensor. A single row of array elements from an image sensor will undersample the edge, but when rows in an area are combined, because the slanted edge is slightly offset in each row, if the pixel values are interlaced they form a single *super-sampled* edge trace (see figure 2.11), which avoids the problems of aliasing. The super-sampled edge is differentiated and the Fourier Transform applied (see Figure 2.11) to obtain a measure of the spatial frequency response (SFR) for that orientation to frequencies beyond the

Nyquist frequency. The SFR method is now widely used and is implemented in a number of standalone software applications. The flow chart for the implementation of SFR in the standard is shown in

Figure 6.3.

The SFR software used in this experimental work was Imatest™ Version 4.0-beta Master, available from [164], which tests a range of image quality factors, including SFR, tone, sharpness and noise, using a number of different charts and implementing ISO standard methods.

A scalable vector graphics (SVG) test target, with a contrast of 20:1 and gamma 2.2 was downloaded from the Imatest® website and printed on an Epson Stylus Pro 3880 inkjet printer. The chart contained a focus star, a grey scale step chart and a number of squares slanted slightly to provide horizontal and vertical slanted edges. The chart was printed to give a target contrast ratio of 20:1 across the slanted edges.

6.3.3 Image Acquisition

The test target was displayed and photographed as described in section 5.3.1 as the same images were used for OECF and SFR measurements. Correct target to camera distance was calculated to ensure that the MTF from the inkjet printer did not affect the overall MTF measurement. The distance was calculated according to guidelines given on the Imatest website for measurement of high quality inkjet prints [165], so that that the captured image should have no more than 140 sensor pixels per inch of target. The distances were evaluated by defining an image size in the captured image to conform to this requirement, and this meant that the camera had to be moved with each change in focal length, to maintain approximately the same size of projected image on the sensor. Focal lengths and target to camera distances are shown in table 6.1.

Focal length (mm)	Target to camera distance (mm)
24	1000
34	1200
50	1670
70	2170

Table 6.1: Target to camera distance for focal lengths used

The images were captured at an ISO speed of 160, and apertures of f/2.8, f8.0 and f22.0 for all four focal lengths. Exposures were bracketed by +/- one stop for each aperture/focal length combination. The raw images were opened in Adobe Lightroom and the best exposures selected from the histograms to ensure a good spread of values without clipping.

6.3.4 Camera-Lens MTF

As for the OECF calculations, two sets of the test images were processed, one with linear processing at maximum resolution and, and the other set with the final processing from the image processing pipeline. For measurement of the SFR of the camera-lens system, the images were selected, opened in the raw processor and processed minimally as shown in table 6.2:

Settings	Value
Resolution	3744 x 5616 (sensor)
Bit depth	16
Brightness & Contrast	0
White balance	Custom 2679 K
Tone Correction	Linear
Lens correction	Canon EF 24-70mm f/2.8 L USM
Sharpening & Noise	off
Output colour profile	sRGB
Output format	TIFF

Table 6.2 Settings for linear processed full resolution output images

A MATLAB routine was written to construct a lookup table to linearise the RGB channels in the output images [166]. Gamma correction was achieved using the reciprocal of the individual channel γ values calculated from the measured OECFs (Figure 5.5 and 5.6). The pixel values from the greyscale step chart values from the linearised image were plotted against the input normalised luminance values to check for linearity.

A further MATLAB routine combined the linearised RGB channels in the images into a single luminance image. For RGB colour spaces, such as sRGB, [137] which use the ITU-R BT.709-3 reference primaries, relative luminance can be calculated from linear RGB values using a weighted average of the three channels [167]:

$$\mathrm{Y} = 0.2126R + 0.7152G + 0.0722B \qquad (6.3)$$

The linearised luminance images were processed in Imatest, using a selection window of 120 x 200 pixels. As specified in ISO 20462 [85] the SFR was measured from on-axis and off-axis positions for both horizontal and vertical orientations.

For each focal length/aperture combination, a minimum of two slanted edge measurements were taken from within 6% of the centre of the image, and a further two to four measurements at the edge of the target, which was between 20% and 40% from the centre of the image (Figure 6.4). The on-axis and off-axis measurements were averaged separately and the overall MTF for each orientation was calculated by giving the on-axis average a weight of 3/7 and the off-axis average a weight of 4/7.

The horizontal and vertical orientation SFRs were combined to give an overall MTF, by weighting them so that the orientation with the lower MTF (calculated as the area under the function between frequencies of 0 and 0.7 cycles per pixel, equivalent to 0-30 cycles per visual degree in the final system MTF) was given a weighting of 2/3 and the orientation with the higher MTF a weighting of 1/3.

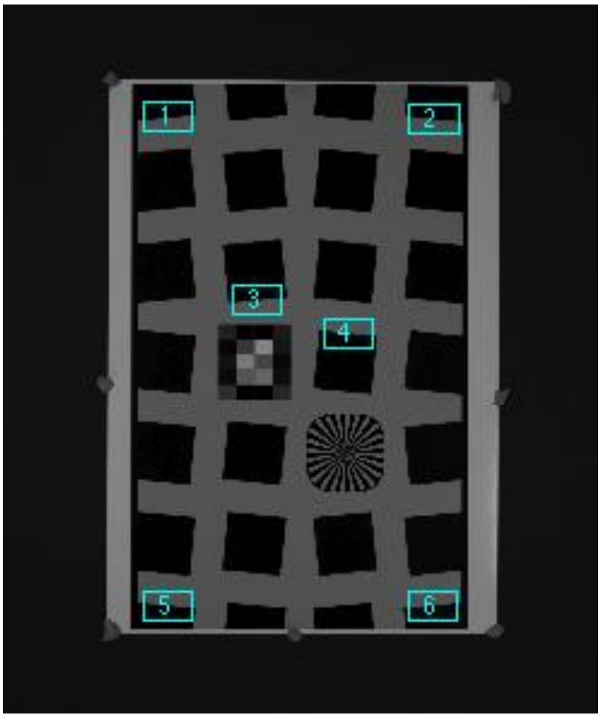


Figure 6.4 Example of areas selected on and off axis for vertical SFR evaluation (image has been cropped)

Figure 6.5 shows the measured SFRs for the different focal lengths at each of the major aperture stops. Because the SFRs were calculated from a weighted average of a number of different edge samples, the error was calculated as a weighted standard deviation and error bars show +/- 1 standard deviation at each point.

The weighted mean may be defined as follows:

$$\bar{x}_w = \frac{\sum_{i=1}^{N} w_i x_i}{\sum_{i=1}^{N} w_i} \tag{6.4}$$

Where $\bar{x}_w$is the weighted mean from all the measurements, w_i is the weight for the *ith* value, x_i and N is the number of samples.

The weighted standard deviation is then calculated from [168]:

$$SD_w = \sqrt{\frac{\sum_{i=1}^{N} w_i (x_i - \bar{x}_w)^2}{\frac{(N' - 1)\sum_{i=1}^{N} w_i}{N'}}} \tag{6.5}$$

Where *N'* is the number of non-zero weights (in this case the same as *N*).

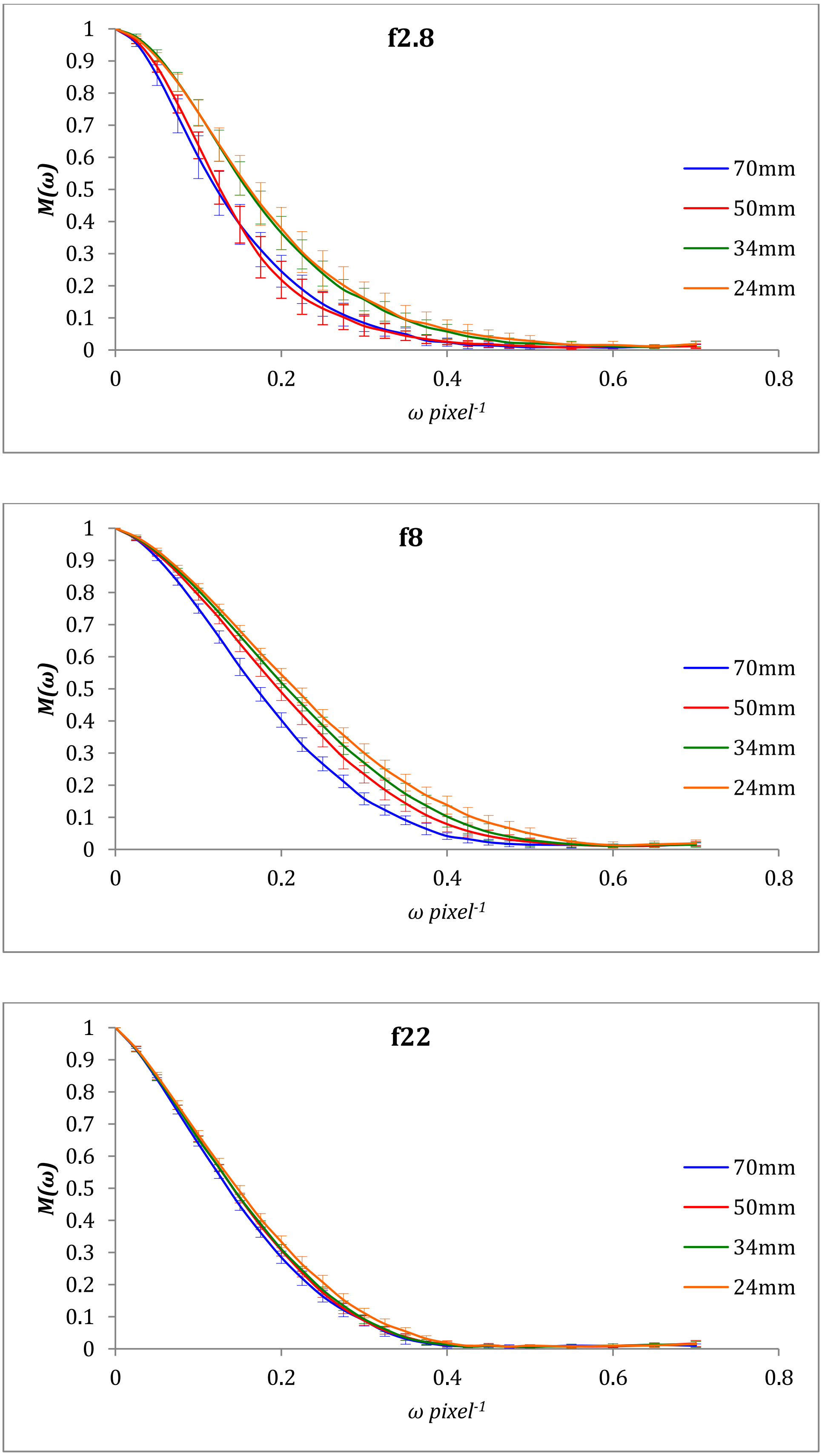


Figure 6.5 : SFR for camera-lens system (minimal processing), showing differences between performance at different focal lengths for the three measured apertures

The standard deviation was largest at f2.8 for all focal lengths; and the difference in mean SFR between focal lengths was larger at f2.8, possibly because at shorter focal lengths the lens has more off-axis correction. Differences in mean SFR, and standard deviation were comparably smaller for all focal lengths at f22.

The variation of the SFR with aperture stop for a single focal length is shown in Figure 6.6. As would be expected, optimum performance for all focal lengths was at an aperture f8.0. The variation in SFR between apertures at different focal lengths was significant, meaning that it was necessary to calculate individual SFRs in defining the filters to create ruler images.

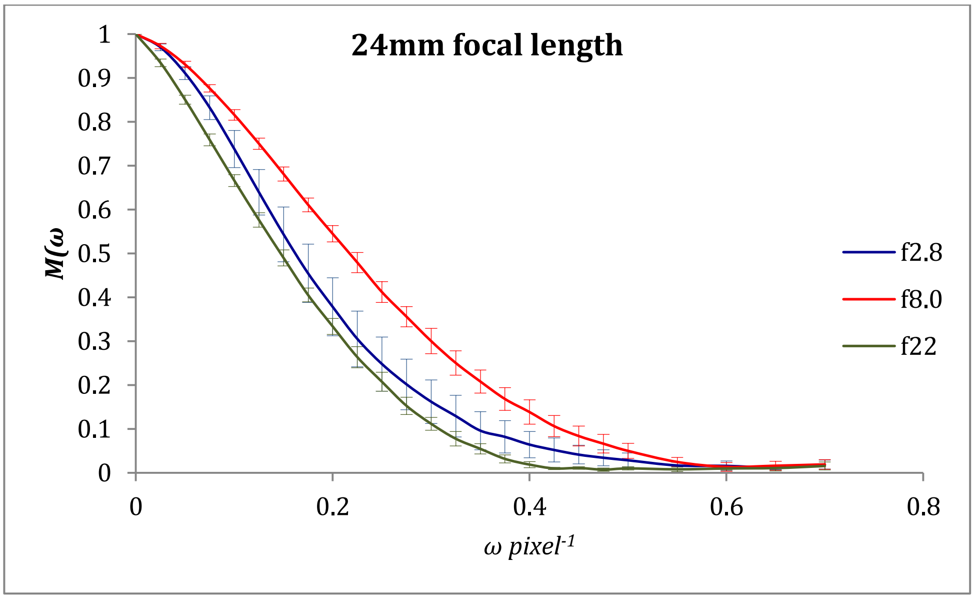


Figure 6.6: Variation of SFR with aperture at a focal length of 24mm

6.3.5 Derivation of Display MTF

For the measurement of the display MTF, an *SFR plus* test target containing edges with contrast ratios of both 10:1 and 20:1 [163], was generated in Imatest® and displayed on the calibrated screen at full size, providing a 1:1 pixel correspondence, to ensure that no interpolation was introduced by the display.

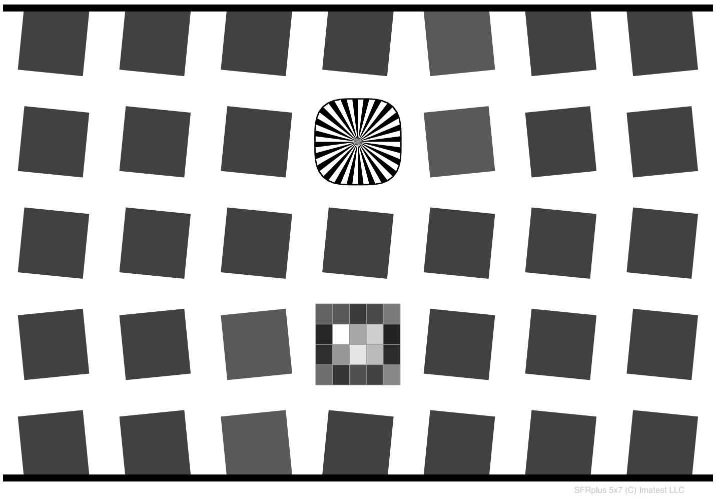


Figure 6.7 *SFR plus* target displayed on screen for measurement of display MTF

The display was photographed in darkness, with the camera placed 1 metre away from the surface of the screen on a tripod. The display target was 1548 x 1042 pixels. With the lens set at a focal length of 50mm, this distance produced an image of the target with dimensions of 3360 x 2256 camera pixels, meaning that each display pixel was sampled by 2.17 camera pixels, enough to ensure adequate sampling, while blending the pattern of the display pixel arrangement [47].

The image was captured at an ISO speed of 160, aperture f/8 and shutter speed of 1/4 s, with the reflex mirror locked up to prevent camera shake. Careful alignment was required to ensure that the camera was parallel to the screen surface, to minimise aliasing as a result of the interactions between the display pixel pattern and the Bayer array of the camera.

The captured image was processed using the same settings as used for the camera processing described in Table 6.2, the only differences being that the image was converted to greyscale during raw conversion, to minimise chromatic aliasing, and that the white balance was left ‘as shot’ (4750K). The output image was a 16bit RGB TIFF file, with equal RGB channels.

As the SFR for the display was to be derived from a combined camera plus display SFR, the combined OECF (display and camera) was calculated from the image of the 20-step greyscale in the test target displayed on screen. For each patch six 38 x 38 pixel regions were selected and the mean RGB channel values recorded for each. The results for each patch were averaged. The normalised mean output pixel values were plotted against normalised input pixel values to obtain the combined gamma value of 1.0251, as shown in Figure 6.8. As would be expected the gamma value was very close to unity, as the camera and display correct for each other. The graph includes standard error of the mean (SEM) values.

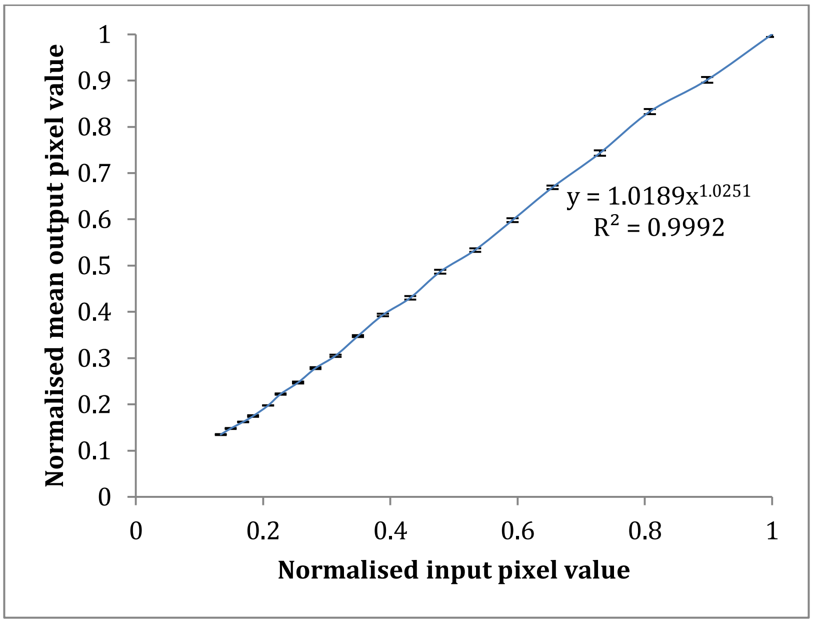


Figure 6.8 Combined tone reproduction characteristics of camera and display, plotted in linear units.

The image of the target was linearised by gamma correcting for this value ($\frac{1}{\gamma} = 0.9755$). The SFR was calculated using a combination of on-axis and off axis selection areas for both vertical and horizontal directions.

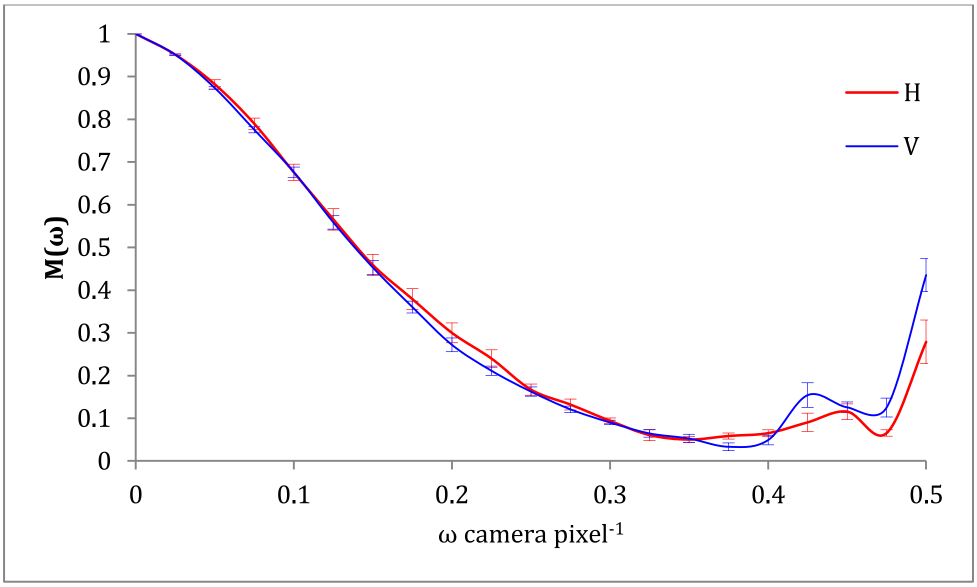


Figure 6.9 Horizontal and vertical SFRs for the camera x display system, expressed in cycles per camera pixel

The mean horizontal and vertical camera-display SFRs (with SEM estimates) are shown in Figure 6.9. Expressed in cycles per camera pixel, these plots sample far beyond the Nyquist frequency of display, hence the appearance of the curve beyond 0.35 cycles per pixel, which indicates aliasing. The conversion of frequencies to cycles per display pixel, by multiplying the frequency by the number of camera pixels sampling each display pixel (≈2.17) gave the display Nyquist frequency at 0.238 cycles per camera pixel.

The display MTF was extracted using the cascading property of the MTF:

$$M(\omega)_{\text{display}} = \frac{M(\omega)_{\,system}}{M(\omega)_{camera}} \qquad (6.6)$$

Where $M(\omega)_{\,system}$ refers to the SFR measured from the camera-lens x display system, and $M(\omega)_{camera}$ is the SFR measured from the camera-lens system (as detailed in 6.3.4). The $M(\omega)_{camera}$ used was the one measured for the same focal length and aperture as the image was captured with (50mm, f8.0). The SFRs for system and camera in cycles per camera pixel are illustrated in Figure 6.10. The frequencies were converted to cycles per display pixel as described above. Figure 6.11 shows the resulting display MTF.

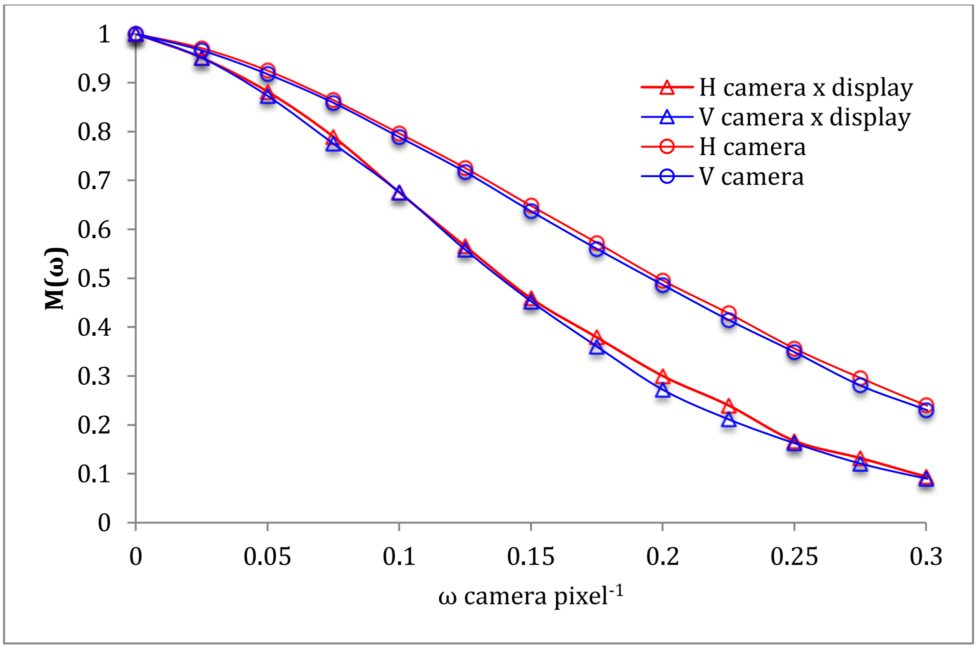


Figure 6.10 SFRs for camera x display and camera-lens system only at a focal length of 50mm and f8.0, for horizontal and vertical orientations, used in deriving the display MTF.

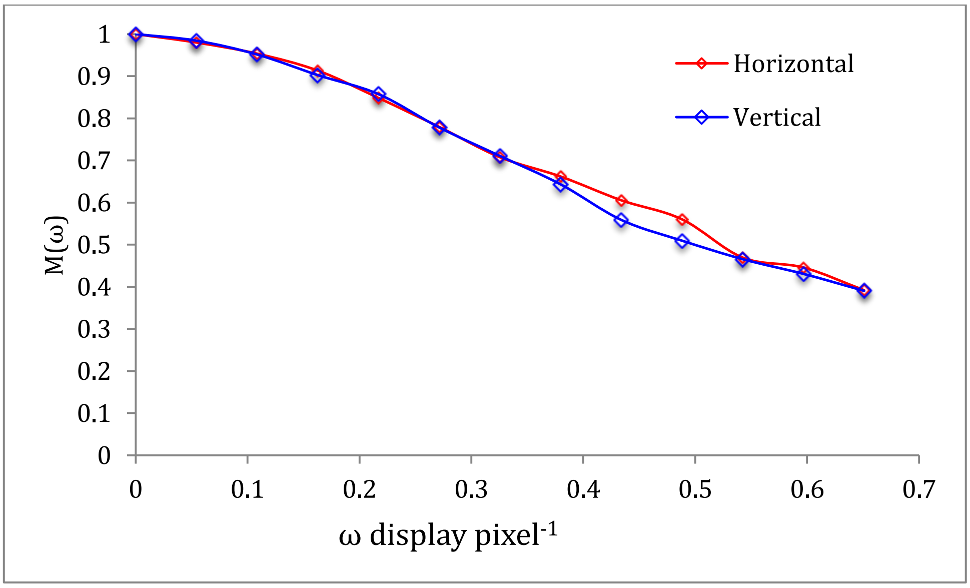


Figure 6.11 Derived SFR for the display for horizontal and vertical orientations

The process of defining the system MTF required the display SFR to be cascaded with the processed image SFR. The processed images had been downsized so that they could be displayed on screen without interpolation introduced by the graphics card. Therefore one pixel in the processed image would map to one pixel in the display. For the purposes of cascading the measured SFRs it was useful to express them in the same frequency increments, therefore a third degree polynomial (shown in Table 6.3) was fitted to each curve.

$M(\omega)_{\text{display}}(\text{H})$	$2.2139x^3 - 2.3999x^2 - 0.3087x + 1.0050$	$R^2 = 0.99728$
$M(\omega)_{\text{display}}(\text{V})$	$3.6256x^3 - 3.6530x^2 - 0.0822x + 0.9999$	$R^2 = 0.99909$

Table 6.3 Polynomials used to fit display SFRs

6.3.6 Camera-Lens-Processing MTF

The images used in the psychophysical experiments had been processed using the pipeline defined in figure 5.1. The processing is summarised in Table 6.4.

Settings	**Value**
Scene Specific Adjustments	
White balance	As shot, or custom to correct
Exposure correction	Minimal, if necessary
Levels correction	As necessary
Applied to all images	
Initial pixel resolution (uncropped, native)	*3744 x 5616*
Down sampling (raw)	1024 x 1536
Down sampling (bicubic interpolation)	588 x 882
Bit depth	8
Tone Correction	Medium contrast curve
Lens correction	Canon EF 24-70mm f/2.8 L USM
Noise	25% colour noise reduction
Sharpening	Unsharp mask applied to 75% opacity L*channel, radius and threshold dependent upon ISO
Output colour profile	sRGB
Output format	TIFF

Table 6.4: Processing applied to finalised images

The processing is separated in the table into processes that were applied scene by scene, to optimise tone and colour, and those that were applied to all scenes. It was not possible to account for scene-by-scene optimisation in the SFR measurements; in capturing the test target images, illumination and exposure was carefully controlled. Therefore only the processes applied to all images were applied to the test target images.

Applying tone correction, noise reduction and sharpening were the processes deemed most likely to affect the OECF and the SFR of the system. As described in section 5.3.2, the OECFs of the processed images were measured from the greyscale step charts in the processed images. The RGB channel γ values were used to linearise the images, and the linearised RGB channels combined to form single channel luminance images from which the SFRs were measured. The SFRs were measured as before, but using a smaller sampling aperture of 32 x 22 pixels (limited by the smaller size of the slanted squares, to ensure that the edge fully covered the selection region). Slanted edges were measured both on and off axis and weighted as previously. The horizontal and vertical SFRs were not combined at this stage, because they were to be cascaded with the display SFR, which was slightly different in the two orientations.

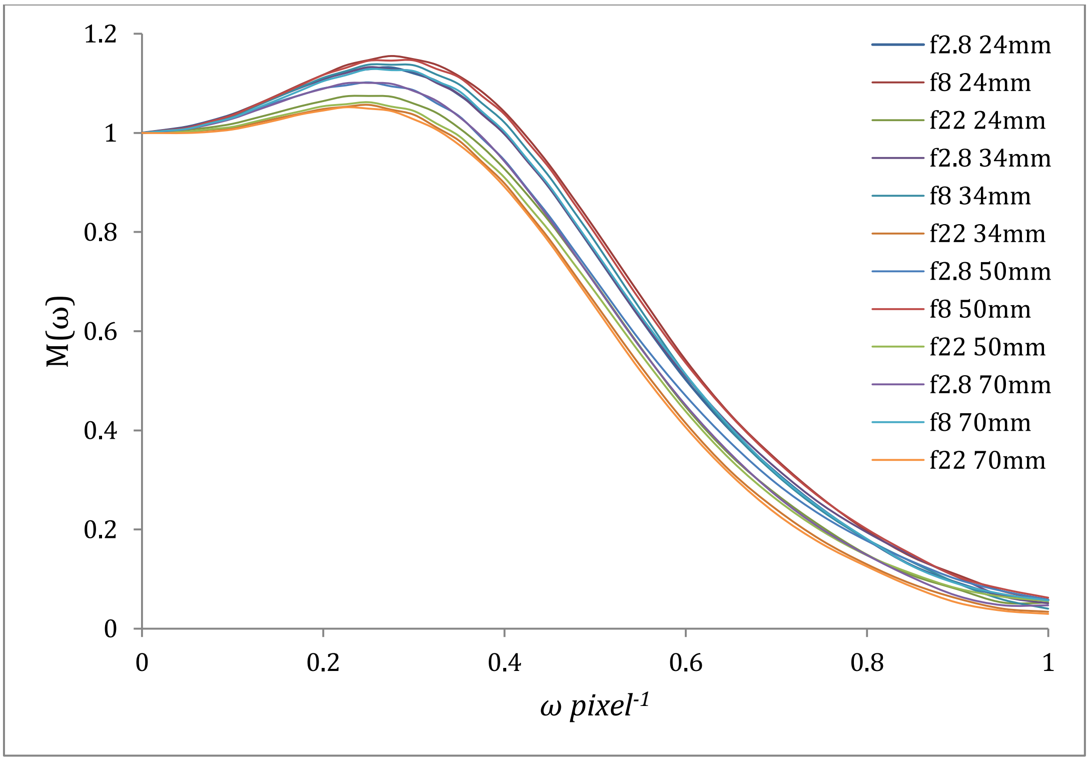


Figure 6.12 Horizontal SFR for camera-lens-processing (SEM omitted for clarity)

The SFRs for all aperture-focal length combinations (horizontal orientation shown in Figure 6.12) followed the same pattern as the camera lens MTFs in terms of focal length and aperture, with the best performance for all focal lengths at f8. The frequencies were expressed in cycles per pixel. In this case one 'pixel' was a pixel in the down-sampled image, which was interpolated from 6.4 pixels in a single orientation in the original full resolution image. These frequencies corresponded to cycles per display pixel.

The SFRs all show an increase above unity between 0.1 and 0.4 cycles per pixel, peaking at 0.3 cycles per pixel, which is typical of an MTF produced as a result of the unsharp mask being applied to the image [169], illustrating an artificial boosting of these frequencies.

6.3.7 System MTF

A subset of 16 of the 25 psychophysical test scenes used in the threshold experiment was selected. It was deemed necessary to reduce the number of scenes to ensure that the quality ruler experiment was not too long for observers. This limitation meant that the system MTF was required for a small number of focal length-aperture combinations: 24mm, 50mm and 70mm focal lengths all at f/8, and 70mm at f2.8. Where images did not exactly match a focal length-aperture combination, the nearest was used.

The measured SFR for the processed image was multiplied point-by-point with the display SFR for horizontal and vertical orientations. It was necessary to express the frequencies in cycles per visual degree (CPD) from the position of the observer [85].

The frequencies were converted from cycles per pixel (CPP) on the display to CPD using:

$$\text{cycles per degree} = \text{cycles per pixel} \times \frac{\pi}{180} \times \frac{d}{p} \qquad (6.7)\ [170]$$

Where *d* is the distance from observer to screen in some unit measure, and *p* is the centre-to-centre pixel pitch in the same units. In this experiment, *d*=600mm and *p*=0.27mm.

The horizontal and vertical orientation SFRs were combined and weighted as before, by 1/3 and 2/3, with the higher weighting given to the orientation with the lower modulation between 0 and 30 CPD [85]. The system MTFs for the four focal length aperture combinations are shown in Figure 6.13.

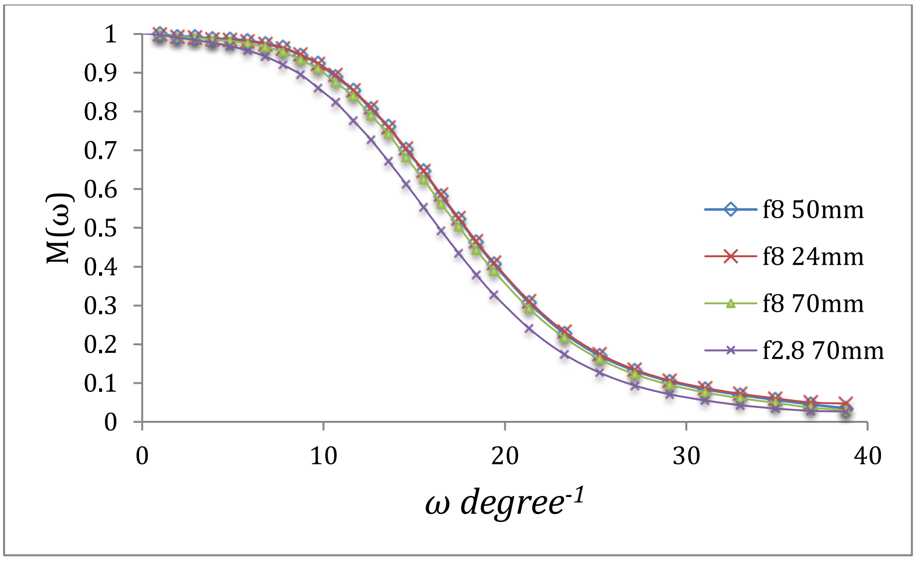


Figure 6.13 System MTFs for the focal length-aperture combinations used in the soft copy quality ruler

6.4 Creation of Ruler Images

6.4.1 Determination of aim MTF and shaping function

The aim MTF was defined using equation 2.24, reproduced here using the nomenclature used in this chapter:

$$M(\omega) = \frac{2}{\pi}\left(cos^{-1}(k\omega) - k\omega\sqrt{1-(k\omega)^2}\right) \qquad for\ k\omega \leq 1 \qquad (6.7)\ [85]$$

$$M(\omega) = 0 \qquad for\ k\omega \leq 1$$

The equation describes the MTF of a diffraction-limited lens. A series of curves were modelled with different values of *k*. They were compared to the measured system MTF to check whether the system conformed adequately to the aim MTF and was suitable for use. ISO 20462-3 describes conformance between the functions if [85]:

‘The mean fractional modulation transfer of the system and aim MTFs over each of the frequency bands 0 to5, 5 to 10…, and 25 to 30 CPD agree to within 0.05’

The increase in $M(\omega)$ of the system MTF between 5 and 10 CPD as a result of the unsharp mask, affected the entire MTF and meant that it was not possible for the system MTF to conform to the aim MTF as required. Therefore an approach used by Jin [88] and Young-Park [89] was adapted, defining a filter to modify the system MTF so that it conformed more closely to the aim MTF. Jin defined a spatial filter to achieve this, while Young-Park modified the MTF in the frequency domain.

First, an aim MTF was selected, for the system MTF to be modified to fit. MTF modification was described by:

$$\mathrm{M}(\omega)_{aim} = \mathrm{M}(\omega)_{system} \times \mathrm{M}(\omega)_{filter} \tag{6.8}$$

And therefore:

$$\mathrm{M}(\omega)_{filter} = \frac{\mathrm{M}(\omega)_{aim}}{\mathrm{M}(\omega)_{system}} \tag{6.9}$$

The filter function was found by dividing the $\mathrm{M}(\omega)_{aim}$ by the measured $\mathrm{M}(\omega)_{system}$ at all of the measured frequencies. The resulting data were plotted and a sixth degree polynomial function was fitted to the graph. This polynomial was used to precisely generate a frequency domain filter transfer function.

6.4.1.1 Shaping the system MTF to the aim MTF

A program was written in MATLAB to apply the filter to the image of the test target (the same image that had been used to derive the system MTF). The process consisted of the following steps:

1) The image was padded with zeros so that its dimensions were to a power of 2. Padding was necessary to avoid wraparound error, because of the assumed periodicity of the image when using the discrete Fourier transform (DFT) [171]), and the image dimensions were converted to powers of 2 to facilitate implementation using the Fast Fourier transform (FFT) algorithm. Code from Gonzales and Woods [172]was used for the purpose.

2) An array of ones was created to the same dimensions as the padded image.
3) The polynomial function was applied radially to the array from (2), so that the values corresponding to the zero frequencies were at the centre of the image and the function was radially symmetrical.
4) The image FFT was computed and centred. The result was separated into magnitude and phase components.
5) The 2D magnitude image was multiplied with the filter array.
6) The result was recombined with the phase and the inverse FFT applied (followed by centring and removal of the padded values) to obtain the filtered image.

The aim MTF, defined for the 70mm f2.8 image is shown, with the system MTF, in Figure 6.14. Note the deviation between the two functions, most pronounced between 0.2 and 0.3 CPP. The functions are expressed in cycles per display pixel, as the derived filters were expressed in pixels.

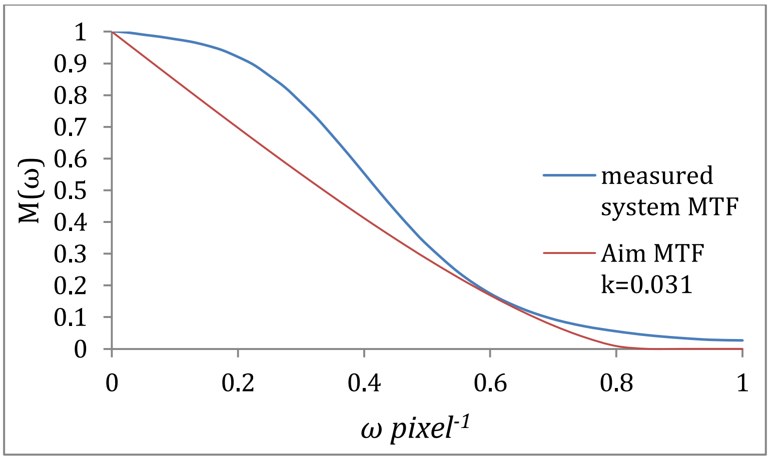


Figure 6.14 System MTF and selected aim MTF (determined from equation 6.7 using a value of k=0.031) for a focal length of 70mm and aperture f2.8

The selection of the aim MTF for a particular camera setting was based upon trial and error as follows: Different values of *k* were used to determine various possible aim MTFs, and their corresponding filter functions were created and used to filter the test image. The conformance of the resulting MTF with the aim MTF after it had been ‘shaped’ was checked for the frequency ranges defined in ISO 20462. It was found that the best result was obtained if the aim

MTF matched the system MTF at near the Nyquist frequency. The resulting shaping function, determined from this combination of aim MTF and system MTF is shown in Figure 6.15. The polynomial was fitted to data up to 0.75 CPP.

Figure 6.15 shows a one-dimensional representation of the filter transfer function, effectively a cross-section of the radius of the filter. Figure 6.16 shows the appearance of the filter as a two-dimensional image. The tones of the image correspond to the filter values, between 0 (black) and 1(white).

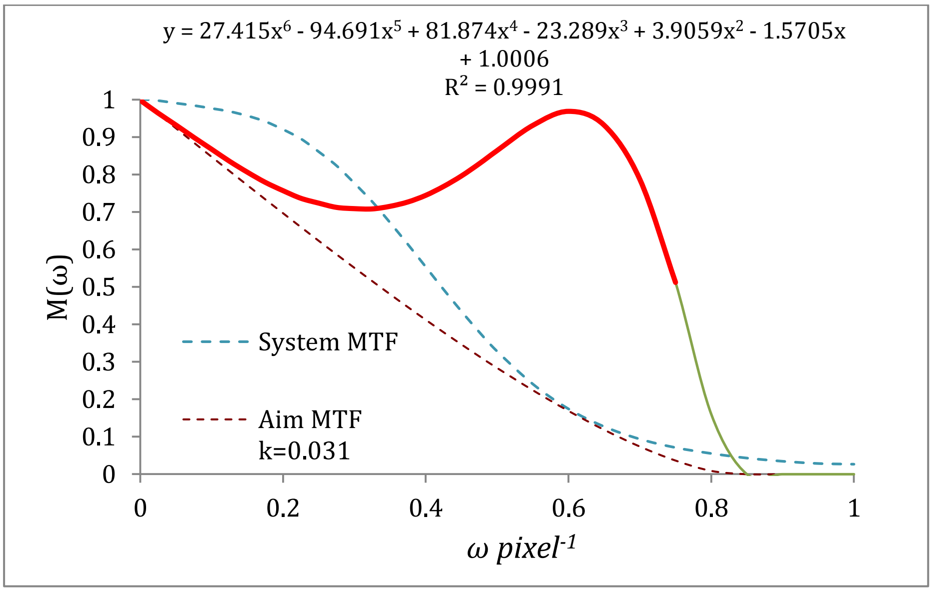


Figure 6.15 Shaping filter and derived polynomial function (filter shown by red solid line)

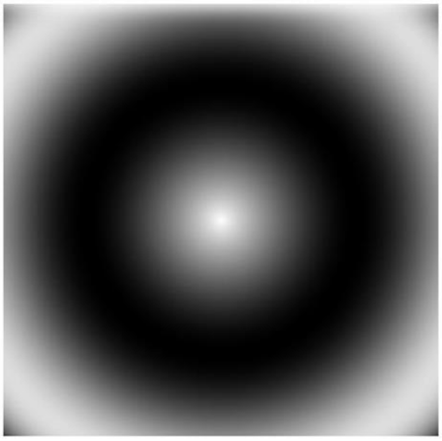


Figure 6.16 A two-dimensional representation of the shaping filter used for 70mm f2.8. The centre of the image corresponds to zero frequency, with highest frequencies at the edges.

The MTF of the filtered image was quantified and is shown in Figure 6.17, indicating satisfactory conformance to the aim MTF, determined from a k value of 0.031.

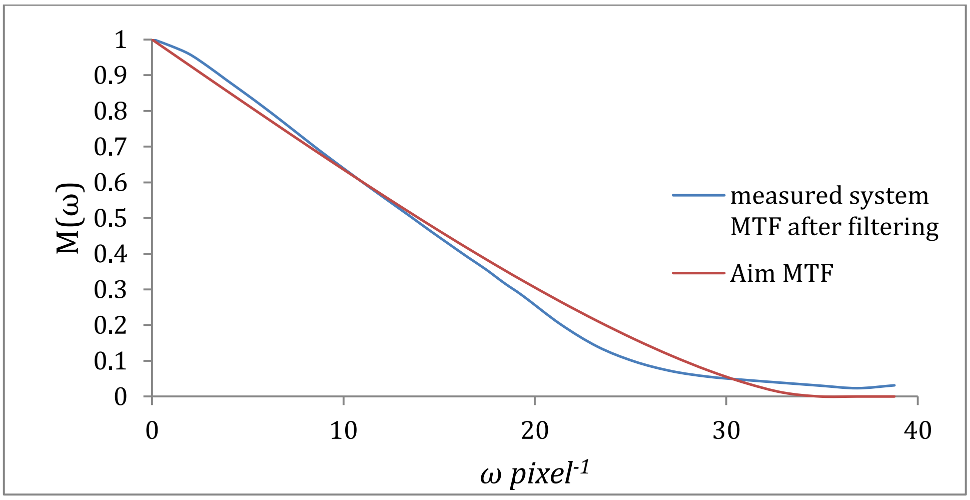


Figure 6.17 System MTF for 70mm f2.8 after filtering with the shaping function, compared with the aim MTF

The process was repeated for all four focal length/aperture combinations. The k values for the best fitting aim MTFs are shown in Table 6.5. The SQS2 value for each was calculated from [85]:

$$SQS_2 = \frac{17{,}249 + 203{,}792k - 114{,}950k^2 - 3{,}571{,}075k^3}{578 - 1{,}304k + 357{,}372k^2} \qquad (1 \leq 100k \leq 26)$$

(6.10)

Focal length /Aperture	***k***	$\boldsymbol{SQS_2}$
24mm f8.0	0.029	27.3276
50mm f8.0	0.029	27.3276
70mm f8.0	0.030	26.9175
70mm f2.8	0.031	26.5033

Table 6.5 *k* values defining aim MTFs for different focal length aperture combinations.

6.4.2 Development of the JND filters

From the determined aim MTF k value for each focal length and aperture, a set of functions could be defined. Using equation (6.10), the k values were found which produced a range of SQS2 values in increments of one JND (for almost all, see below) from the SQS2 value of the aim MTF. These represented a set of aim MTFs (Figure 6.18) of varying sharpness. The lower k values correspond to higher MTFs. The bold line shows the 'original aim MTF' for a k value of 0.031.

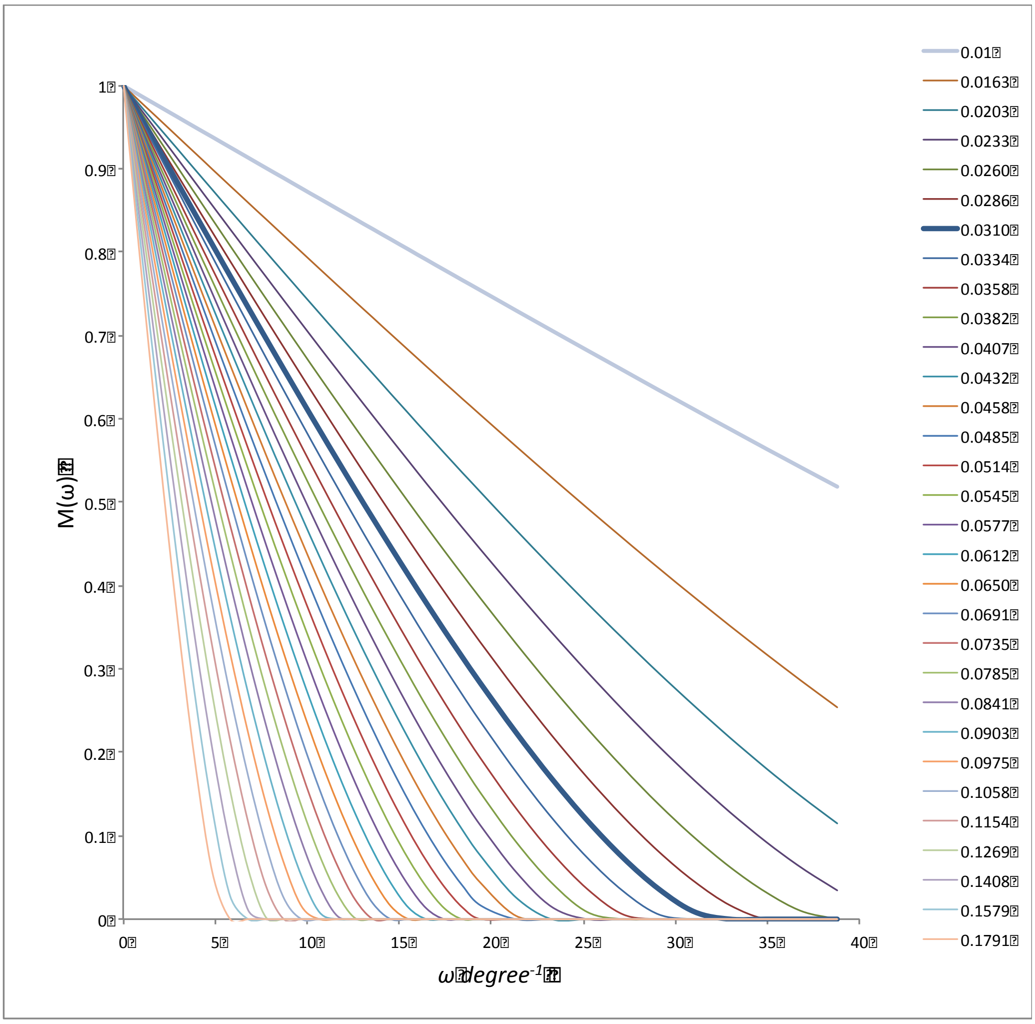


Figure 6.18 Aim MTFs spaced in 1JND increments for a series of filters for the 70mm f2.8 system MTF. The legend shows the k value used to generate each curve.

It is important to note that the highest MTF, for a k value of 0.0100, was the limit according to equation (6.10), which is only defined for $(1 \leq 100k \leq 26)$. This one curve was less than one JND increment from the curve below (0.58

JND difference for the 70mm f2.8 images; the difference varied depending upon the starting point for *k*), but it was important to include it, to allow the largest range of sharpness to be represented. As the process of modelling to the original aim MTF necessitated a slight reduction in sharpness from the system MTF (by removal of the increase in the magnitude of frequencies between 0.1 and 0.4 cycles per degree), it was possible that some of the less compressed images might be matched to a ruler image with a higher MTF than the original aim MTF.

The 'JND filters' were generated in the same manner as the one generated to shape the system MTF to the aim MTF, described in section 6.4.16.4.1, by dividing the system MTF by the required filtered image MTF, and fitting a polynomial, which could be applied as a filter to the image FFT. A total of 31 filters were created for the 70mm f2.8, settings ranging from ≈+6JND to -24 JNDs (relative to the original aim MTF). These covered a range of *k* values from 0.01 to 0.1791 for the filters. For the other camera settings, the number of filters was one less (+5JNDs to -24 JNDs) because the initial *k* values were higher, covering a *k* range from 0.01 to 0.1613 for 50mm and 24mm at f8.0 and a *k* range of 0.0136 to 0.1697 for 70mm at f8.0. Some examples of the JND filters are shown in Figure 6.19.

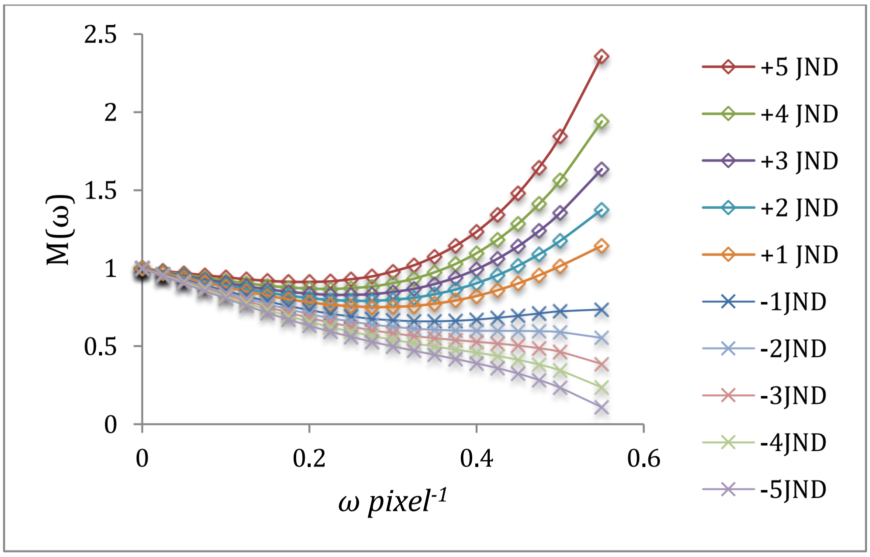


Figure 6.19 A sub-set of the JND filters developed for 70mm f2.8 system. Positive JND filters sharpened the image relative to the original aim MTF, negative JNDs blurred the image.

6.4.3 Application of the JND filters to the test images

The JND filters were generated for each focal length-aperture combination using the appropriate set of k values and the measured system MTF. The MATLAB routine developed in 6.4.1 was adapted to apply the filters to the final set of processed images to be used in the experiment. These were the reference images, so had been processed using the image processing pipeline as described in Table 6.4, but had not been compressed. The program separated the RGB colour channels, normalised them, and converted them to linear sRGB values as follows:

$$\mathrm{C}_{sRGB} = \left[\frac{(\mathrm{C'}_{sRGB}+0.055)}{1.055}\right]^{2.4} \qquad \text{if } \mathrm{C'}_{sRGB} > 0.04045 \qquad (6.11)\ [137]$$

$$\text{and } \mathrm{C}_{sRGB} = \frac{\mathrm{C'}_{sRGB}}{12.92} \qquad \text{if } \mathrm{C'}_{sRGB} \leq 0.04045$$

Where C is the colour channel $\mathrm{C'}_{sRGB}$ is the non-linear sRGB value for that colour channel and $\mathrm{C}_{sRGB}$is the equivalent linear sRGB value. Steps 1-6 from section 6.4.1 were applied separately to the linear RGB colour channels. The filter was applied using the appropriate fifth or sixth degree polynomial function (fifth degree polynomials were found to fit adequately for many of the functions and sixth degree polynomials were only used where necessary for accuracy). The polynomial functions including their cut-off frequencies are tabulated in Appendix C. The resulting images were mapped back to the spatial domain, the sRGB transfer functions were applied to produce non-linear sRGB values, and the images were saved as 8 bit TIFF files.

6.5 Psychophysical Investigation

6.5.1 Psychophysical Display and Viewing Conditions

The display was an EIZO CG245W 24.1" LCD, driven by a Dell Optiplex® and was calibrated daily during the period of the test to the sRGB specification [137]. The viewing environment was also calibrated to closely match the sRGB specification, with ambient color temperature of 5000K and an ambient illuminance of 64 lux.

6.5.2 Interface Design

The quality ruler interface was designed in MATLAB 7.12 using GUIDE. The images were presented side-by-side on screen with the ruler image on the left, and the test (compressed) image on the right. The effective screen size was 518.4mm wide by 324.0 mm; the images took up approximately 45% of the half-screen area on a mid-grey background. The slider was below the images. The test images were presented in a random order, being randomised each time the test was run. The slider contained no numerical information, and was marked at either end by ‘beyond high range’ (beyond the sharpest images) and ‘beyond low range’ (beyond the most blurred image). The increments on the slider adapted in size depending upon the total number of images in each set. The interface is shown in Figure 6.20.

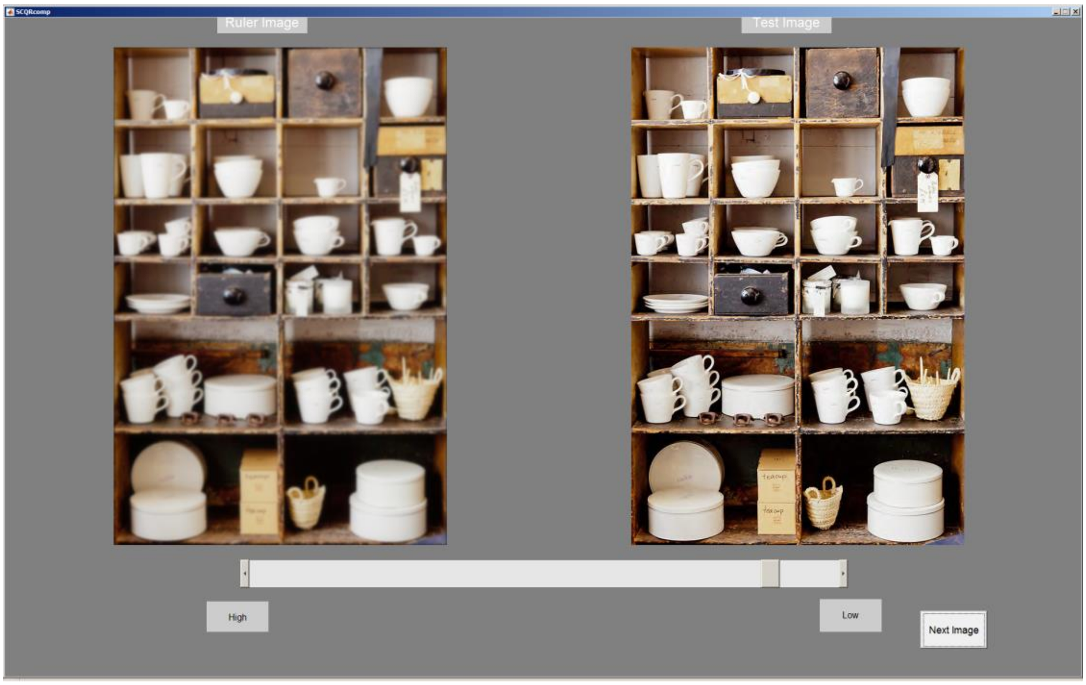


Figure 6.20 Soft-Copy Quality Ruler Interface

When the ‘next’ button was pressed, a new pair of images was presented to the observer. The initial level of sharpness of the ruler image and its associated slider position were randomised, so the ruler image might be very close or quite far from the test image in quality. The observer was then able to move the slider left for a sharper ruler image and right for a less sharp image, until they felt that the images matched in quality level.

The application stored the results for each observer in a text file, which was named by the observer in a dialogue box at the beginning of the test.

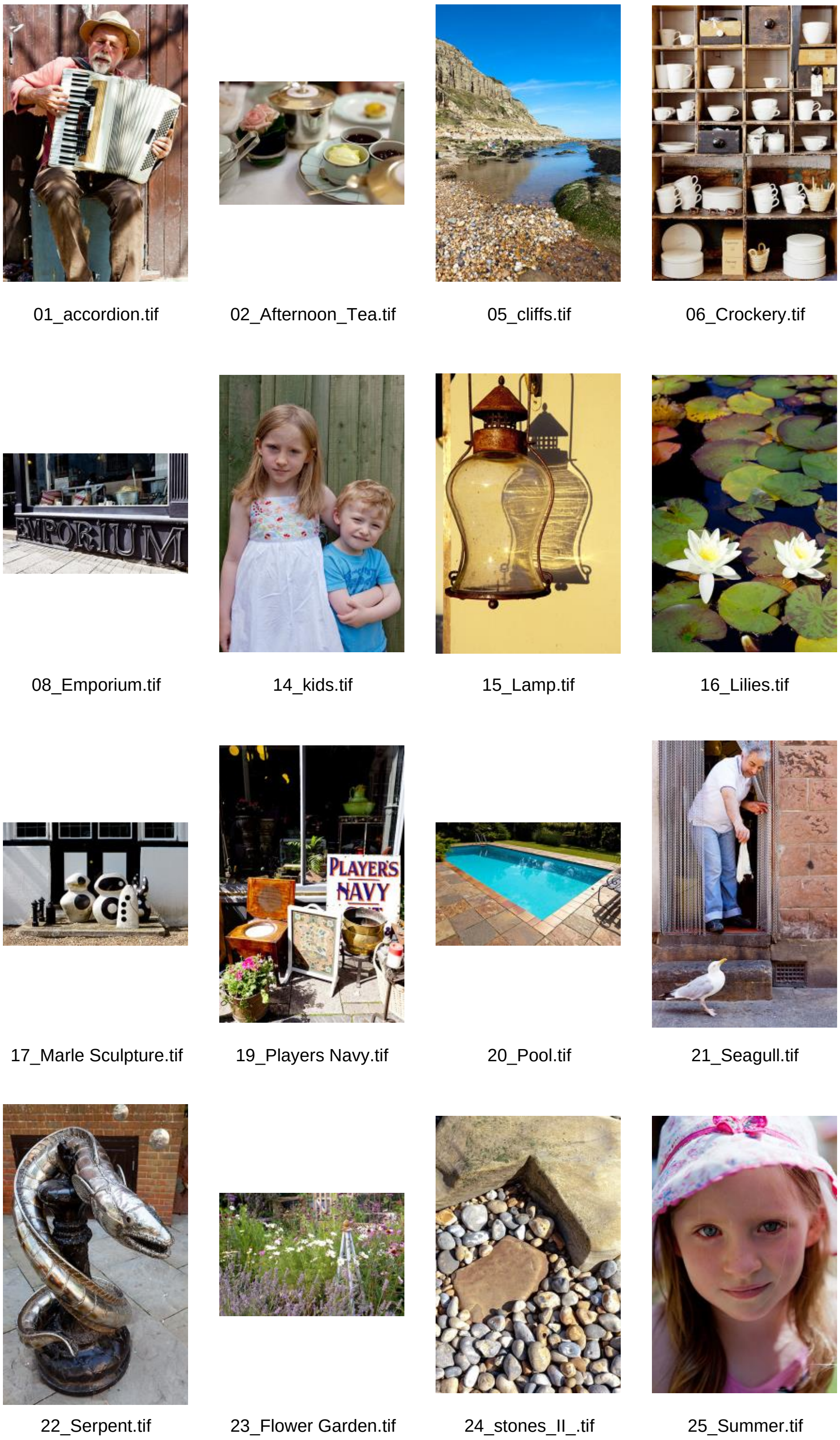
6.5.3 Observers and Test Images

14 observers with experience in image evaluation, from a variety of (image related) backgrounds, completed the test. The observers ranged in age from 19 to 50 years old. All had normal or corrected vision.

The test consisted of 16 different scenes, (Figure 6.21). For most scenes compression ratios of 10:1 to 60:1 were used. This range was selected because it had been found that the perceptibility and acceptability thresholds were contained in this range for all scenes. For two of the scenes that had been found to have low perceptibility thresholds in the previous experiment, additional compression ratios of 5:1 and 15:1 were used. This meant that there were a total of 101 test images.

Observers were given detailed verbal instructions about how to conduct the experiment and were shown examples of a couple of test images and the process of scaling them. The types of artefacts characteristic of JPEG 2000, and the scene areas that were most susceptible were highlighted. They were also given written instructions, adapted from [85]. These can be found in Appendix D.

The time taken by observers to complete the test varied, but all were completed within one hour.



01_accordion.tif

02_Afternoon_Tea.tif

05_cliffs.tif

06_Crockery.tif

08_Emporium.tif

14_kids.tif

15_Lamp.tif

16_Lilies.tif

17_Marle Sculpture.tif

19_Players Navy.tif

20_Pool.tif

21_Seagull.tif

22_Serpent.tif

23_Flower Garden.tif

24_stones_II_.tif

25_Summer.tif

Figure 6.21Images evaluated by the Soft Copy Quality Ruler

6.6 Results and Discussion

Each observer defined a ruler value for each test image, by selecting the point on the ruler at which they judged the ruler image to be identical in terms of image quality to the test image. The results were output as a text file containing the matched ruler and test image file names. From these results, SQS_2 values were identified for the selected ruler image. The results from all observers were averaged to give a final SQS_2 value for each image. Initially, these results were used to investigate scene dependency, but it is important to note that the rulers at this point had not been calibrated (see section 6.6.4). The calibration process as detailed in section 7.2 of [85] provides a method of reducing scene dependent effects and is therefore necessary when the results are to be interpreted in JNDs. However the uncalibrated results were deemed more useful for investigating scene dependency (as to some extent the scene dependency was more exaggerated prior to calibration) and were also therefore of interest.

6.6.1 Overall average, all scenes

The results for all images and all observers are shown in Figure 6.22, which is effectively a quality loss function in SQS_2 values against compression ratio.

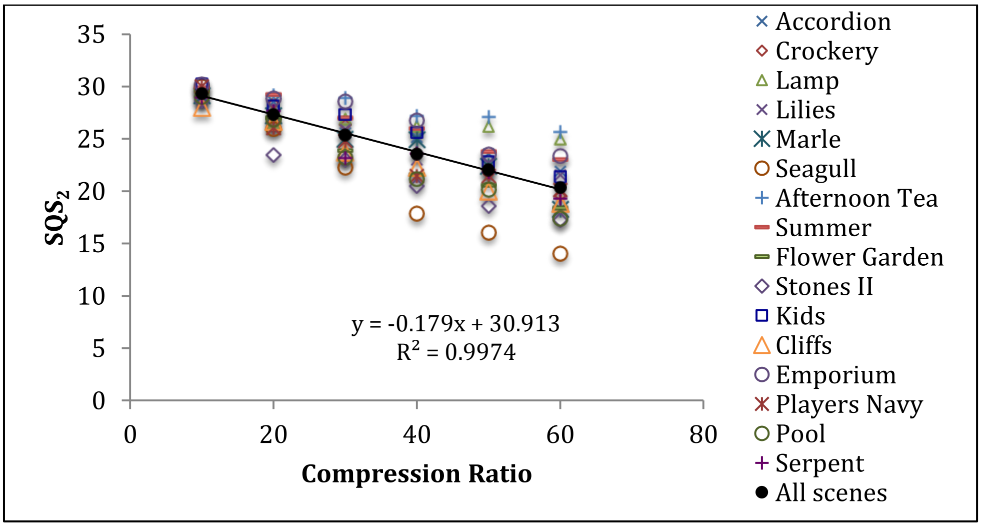


Figure 6.22 SCQR results for all scenes and all observers.

The average quality loss, determined from compression ratios of 10: 1 to 60:1 was equivalent to 9 SQS_2 values. This was calculated from the average scale values of all scenes. However it is clear from Figure 6.22 that there was significant variation across scenes. The slope of an individual set of scene values will determine the amount of quality loss for that scene. Significant deviations of the slope can identify scenes that are more or less susceptible to the distortions introduced by the process in question.

As would be expected the results are very close for all scenes at low compression rates, because the distortions are not highly visible, with much more variation at higher compression rates, indicating the difference in the impact of distortion on different types of images.

The ruler images included an 'original' image, which was filtered to the original aim MTF. It might be assumed that this would be regarded as the highest quality image as it is not affected by the compression. Therefore the compressed images would be expected to be lower than this original. The original image SQS_2 values (shown in Table 6.5) were between 26.5 and 27.5.

Because of the low *k* values of the aim MTF, the original image was not at the centre of the JND range of the ruler images (see Figure 6.18), but was near the top, meaning that there were only five or six ruler images that were sharper than the original and many more that were blurred. It can be seen from Figure 6.22 that the majority of images at the lowest compression ratio have an SQS_2 value which is higher than the original SQS_2 value. This is somewhat misleading as it implies that the quality of the images at 10:1 compression ratio was universally regarded as better than the original.

This result can be explained by the fact that the shaping function for the original image removed the increase in the system MTF between 0.1 and 0.4 cycles per pixel. This is clearly illustrated in Figure 6.14 and Figure 6.17, which show the system MTF before and after filtering. Effectively the image filtered to the aim MTF was slightly less sharpened (or oversharpened) than the original. However, the test images incorporated the sharpening and therefore were perceived at a higher SQS_2 value than the 'original' image. The relative quality of the compressed but sharpened images had effectively been artificially increased by the sharpening process.

Unfortunately the unfiltered and uncompressed original image (the real original, which had not been filtered with the shaping function) was not included in the set of test images in anticipation of this effect. A possible strategy for dealing with this problem would be to include this image and in some way use the scaling value from it to rescale all the other values down (i.e. to identify the difference between the known SQS_2 value of the original image which has been shaped to the aim MTF, and the observer scaled SQS_2 of the 'real' original image. However it is not clear how much the sharpening would have artificially boosted the scale values across the range. This is an area that warrants further investigation.

6.6.2 Individual Scene Results Prior to Ruler Calibration

The results below show the results for individual scenes, broadly grouped by comparison with the average function. Trend lines were calculated using linear regression, which was found to be a good fit to the data in the majority of cases. Error bars indicate standard error of +/- 1.

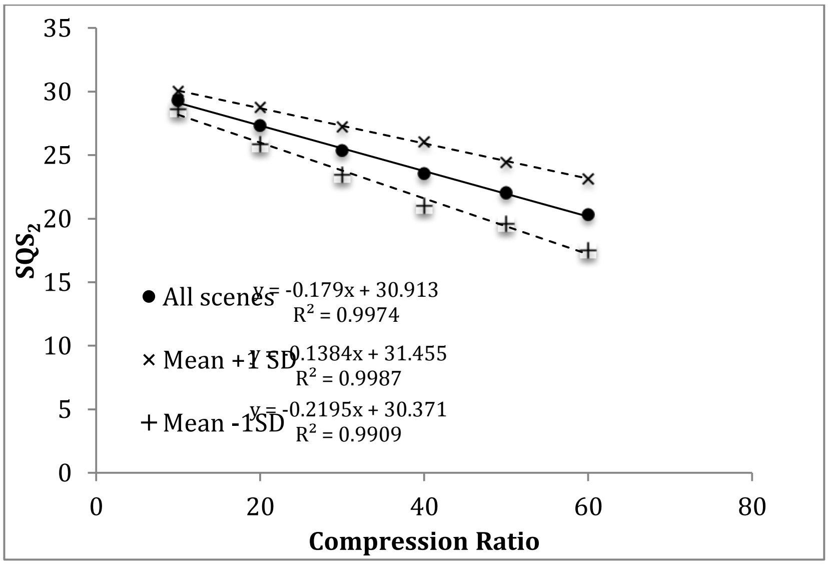


Figure 6.23 Quality loss functions for all scenes including the scene average and the functions + and – 1 standard deviation from the mean

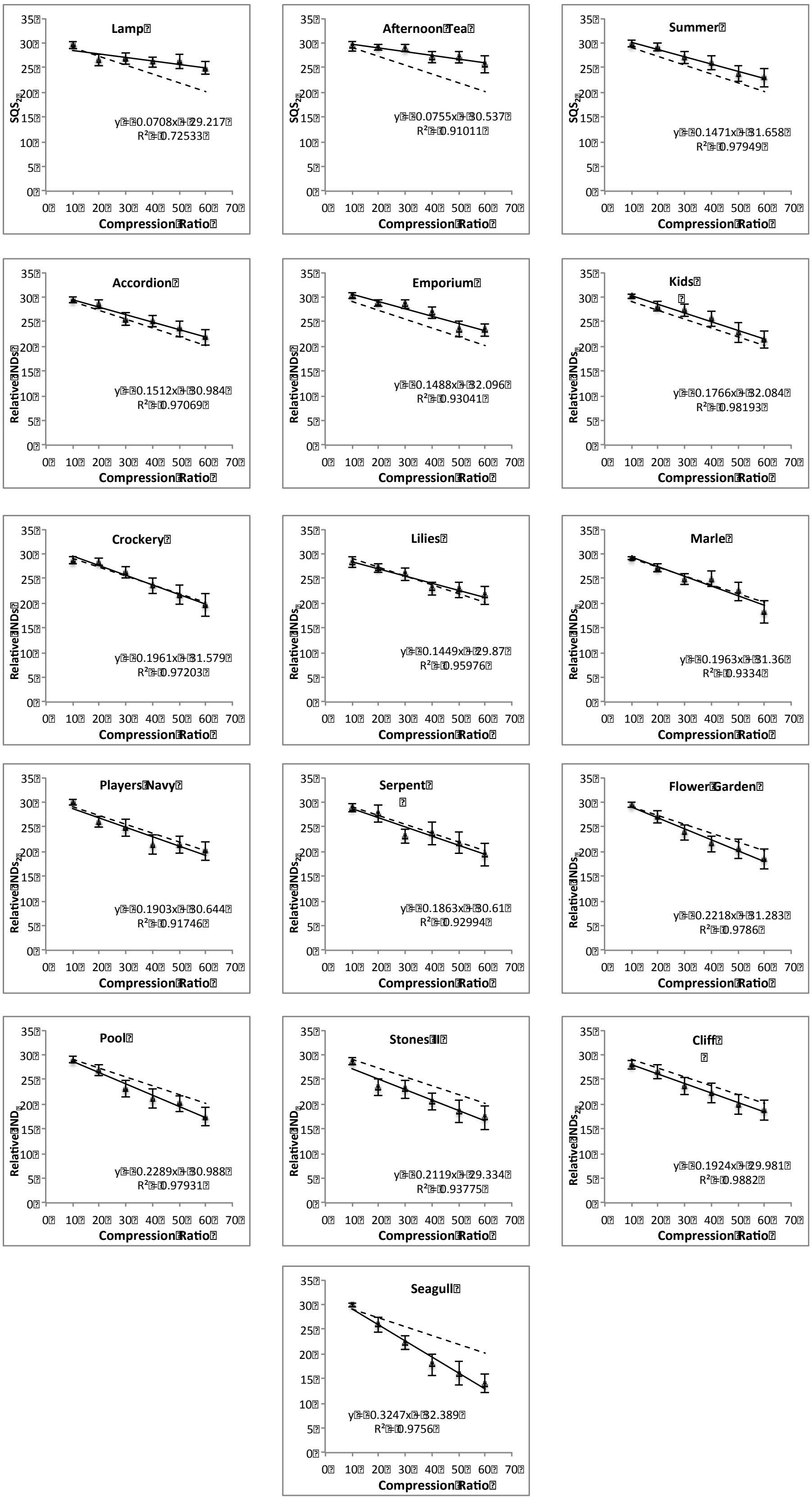


Figure 6.24 All Scene Results SCQR prior to calibration

6.6.2.1 Scene Susceptibility

As discussed by Keelan [99], psychophysical tests are inevitably affected by observer sensitivity and scene susceptibility. It can be useful to quantify these effects by evaluating individual scenes or observer results against the average for all. Figure 6.23 shows the average quality loss function for all scenes and all observers. The standard deviation is calculated from the variation in scale values across all scenes at a particular compression level. The dotted lines show + and -1 standard deviation. The larger standard deviations at the lower quality end of the scale illustrate the scene dependent effects of the distortion as a result of its increased perceptibility. The slope of the lines indicates more or less quality loss. It seems that these could be useful in exploring scene susceptibility. Figure 6.24 illustrates the results for all of the individual scenes.

6.6.2.2 Scene results close to the all scene average

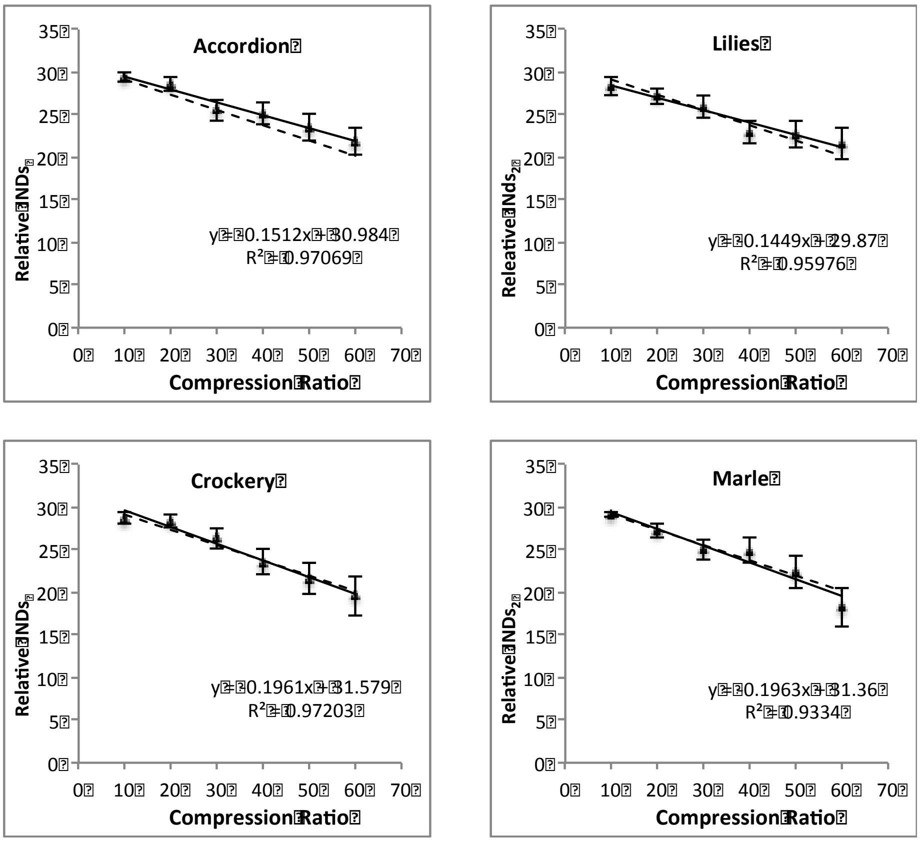


Figure 6.25 Scenes that are close to the all scene average quality loss in the SCQR

The scenes in Figure 6.25 produced results that were close to the initially calculated all scene average. The error bars increase from lowest compression to highest, and this is a consistent result across all the scenes, indicating more variation in observer response as the distortions became more visible, whereas there was remarkable agreement between observers when compression was low.

6.6.2.3 Scenes more or less susceptible to image distortions

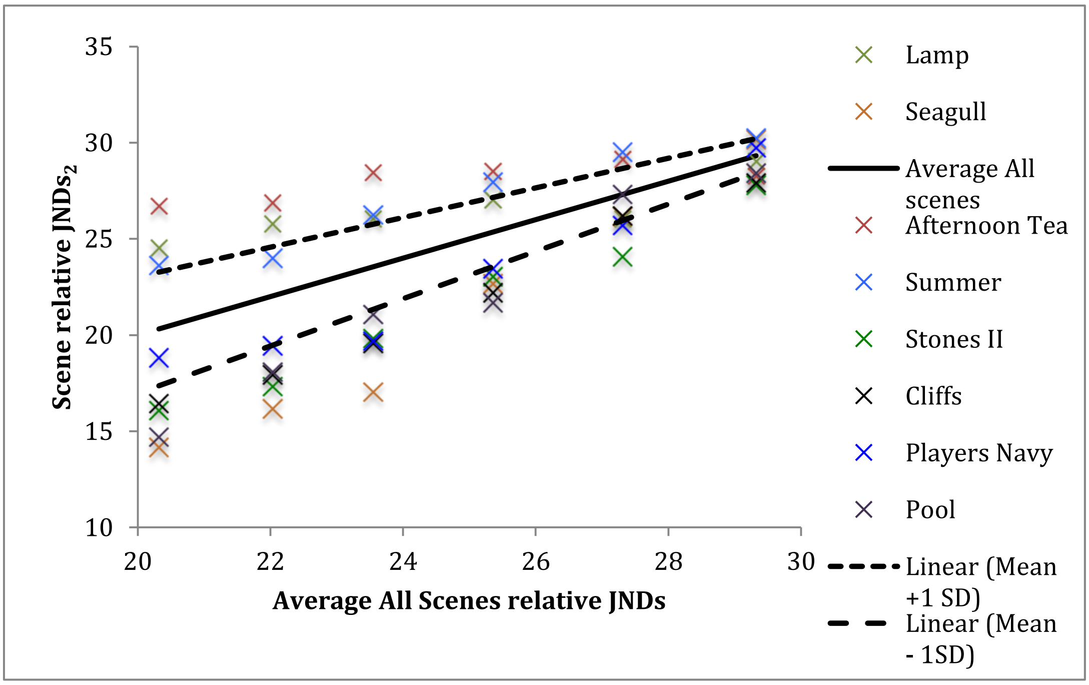


Figure 6.26 Individual scene SQS$_2$ values for a selected set of scenes plotted against average SQS$_2$ values for all scenes

A potential approach to exploring scene susceptibility is illustrated in Figure 6.26. In this graph, relative JND values for individual scenes are plotted against the scene average relative JND values. The values are reversed from the previous figures in that they start from lowest quality on the left hand side. The line for average all scenes is a plot of the average against itself, so the values on both axes are the same; if on equal scale axes this line would be at 45°. The (mean + σ) and (mean − σ) lines are also plotted. These would seem to be useful limits with which to identify unusual scenes.

The slope of the functions can be used (see regression lines and equations on Figure 6.23) Figure 6.23) to distinguish between more or less susceptible scenes. A steeper gradient gradient indicates a greater quality loss by the individual scene compared to the original, the original, while a more robust scene would have less overall quality loss and a less a less steep function gradient than the average. There are some exceptions however, for however, for example the *summer* image plotted in light blue, has a similar gradient to the gradient to the average but is consistently higher, indicating that it is judged as better better than average quality throughout the compression range. The individual scale values scale values have been included in the figure for the scenes where at least three of the three of the values fall outside the ± 1σ lines. The images are shown in

Figure 6.28 and

Figure 6.29.

The images in Figure 6.27 (lamp, summer, afternoon tea) are the scenes that were least susceptible to quality loss. Their individual quality loss functions show that they are relatively flat across the whole compression range compared to the average (shown by the dashed line). These are most of the images from Groups I and II, identified from the results of the threshold experiment in chapter 5, confirming that both methods can accurately identify the least susceptible scenes.

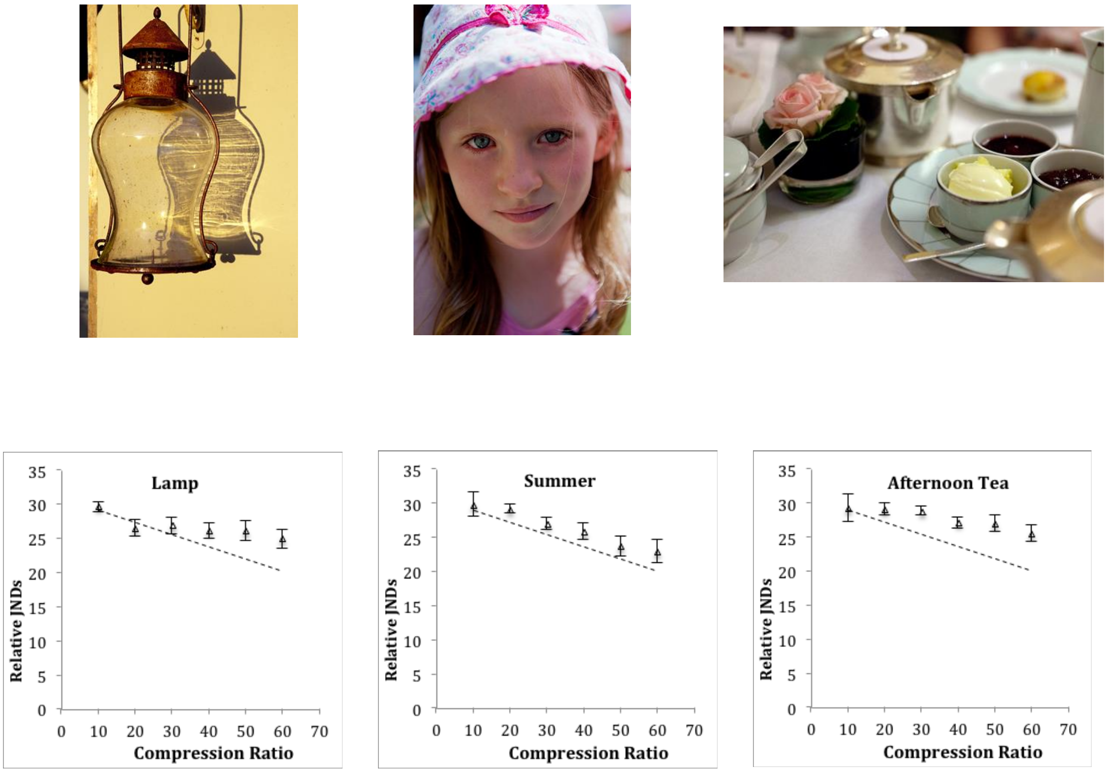


Figure 6.27 Scenes that are less susceptible to quality loss than average in the SCQR

Figure 6.28 (*seagull, stones II, cliffs, players navy*) are those below the (mean – σ) line in Figure 6.24 and therefore can be considered to be the most susceptible scenes. Again, the results correlate with those from chapter 5; the first three images are those from the group with the lowest thresholds.

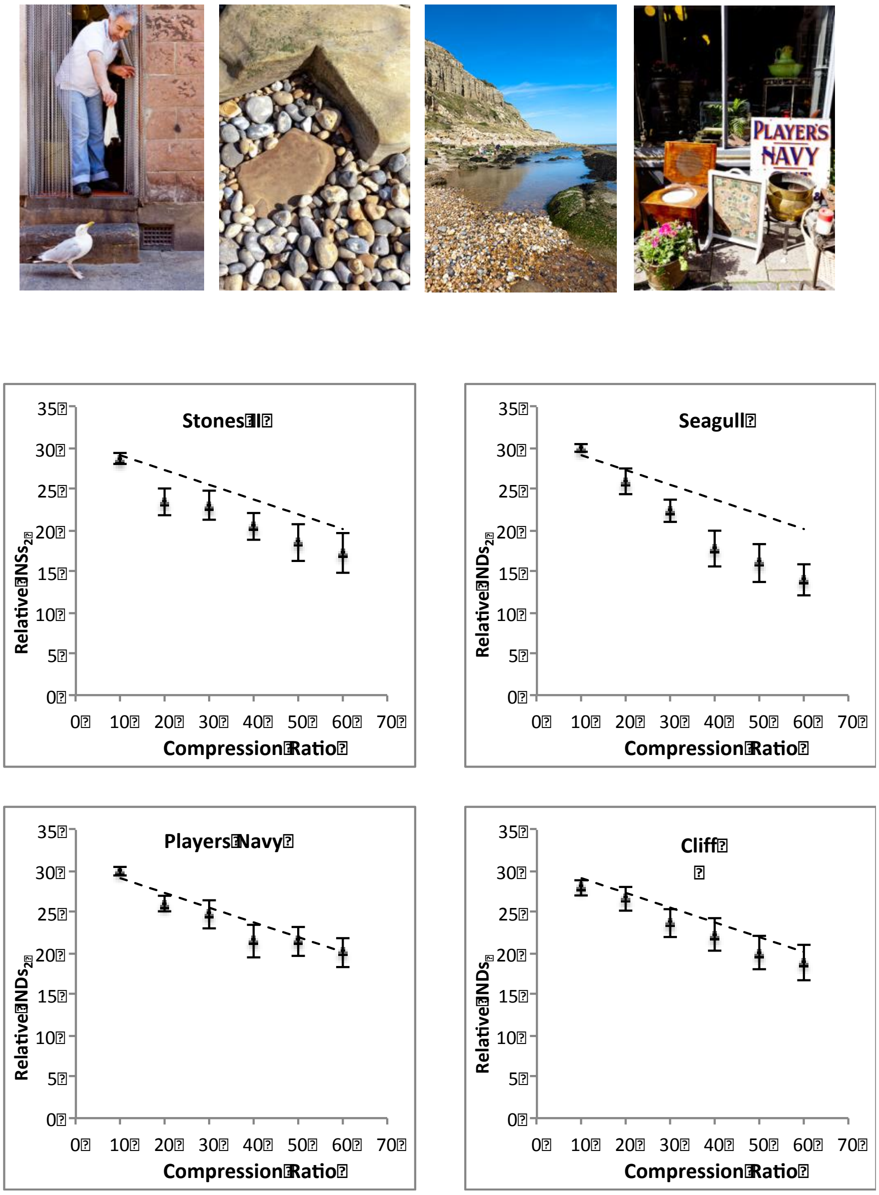


Figure 6.28 Scenes found to be more susceptible to quality loss than average in the SCQR experiment

The *pool* image is also within this group, shown in Figure 6.29 The pool image compressed to a compression ratio of 60:1, which observers noted as particularly difficult to scale on using the quality ruler. The reason given was that the image has fine random detail in the foreground, while being slightly blurred in the background, and observers found it difficult to match the quality of the ruler image to both parts of the test image.

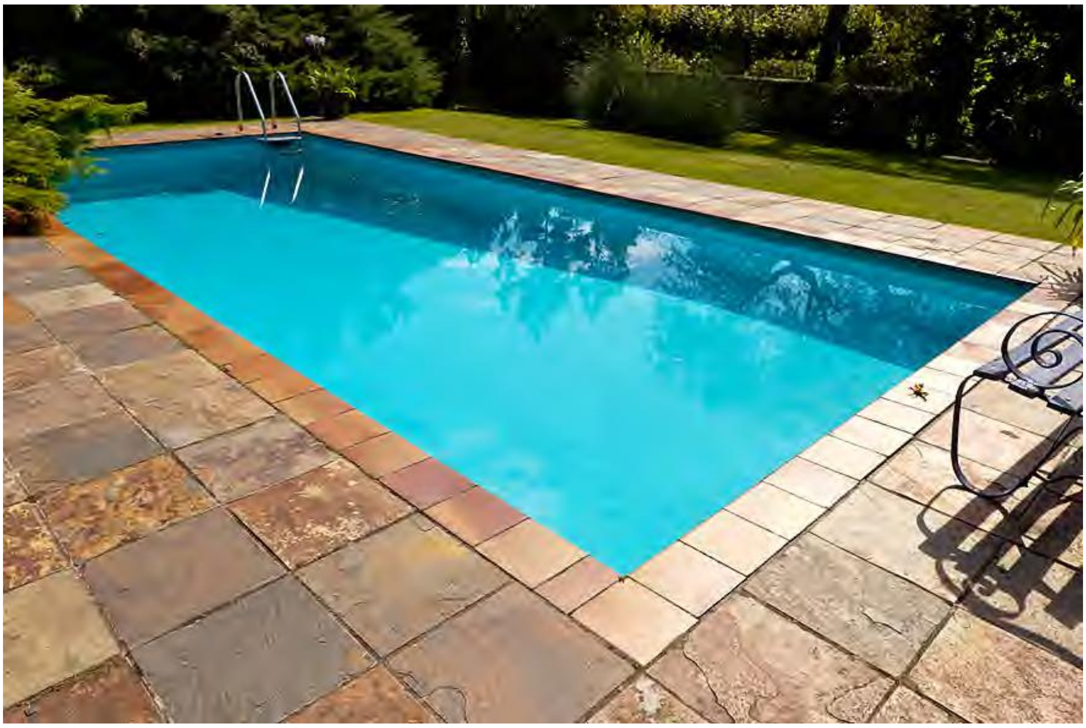


Figure 6.29 The pool image compressed to a compression ratio of 60:1, which observers noted as particularly difficult to scale on using the quality ruler

6.6.3 Observer Sensitivity

The results from individual observers were also evaluated (Figure 6.30). The total and average scale values over all scenes were calculated for each observer. Using the same approach as used to identify non-average scenes, the mean total for all scenes was calculated and the standard deviations from the observer totals. This gave limits for observer response.

It was found that six observers fell outside the (μ ± 1σ) range, three at either end. The least sensitive observers consistently ranked the test images at a higher level than the average and their responses were flat across the range. The most sensitive observers ranked the images with the greatest quality loss across the range, producing average quality loss functions, which were much steeper than the average. It is interesting to note that these three observers were researchers with a particular interest in image quality, therefore might be expected to be more critical in their evaluations.

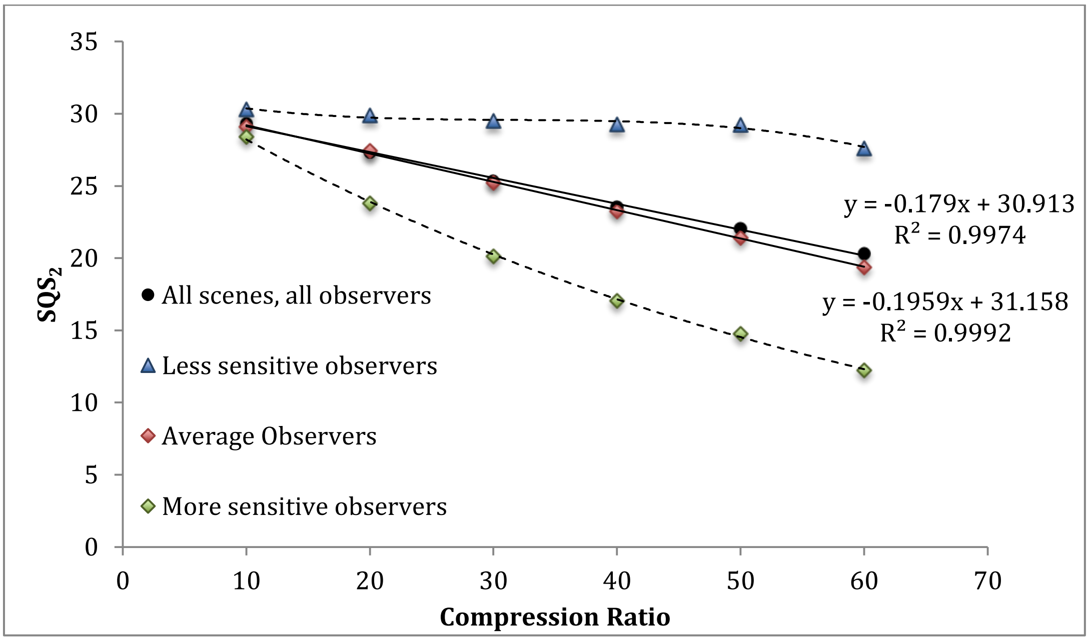


Figure 6.30 Observer sensitivity in the SCQR results

The two solid lines plotted in Figure 6.30 are the average quality loss function for all observers, and the function for average observers when the most sensitive and least sensitive observers were excluded. This indicates that observer sensitivity has not biased the results in this investigation, as expected, because there are an equal number of more and less sensitive observers.

6.6.4 Calibration of the Rulers

The ISO standard 20462-part 3 [85] defines the methodology for validating and calibrating quality rulers. For soft copy quality rulers, user generated scene-dependent rulers created using equation (6.10), may be calibrated by direct comparison of the ruler images with the Digital Reference Stimuli (DRS), or using the *average scene relationship*. In this case, ruler results shall be averaged against at least two other quality rulers from different scenes, to reduce bias introduced as a result of a scene having particular dependency upon the attribute being varied. The quality rulers used for the averaging in this study should themselves not exhibit strong sharpness related scene dependency.

For the purposes of ruler calibration a number of images were selected which had quality loss functions close to the overall average (Table 6.6)

	Trend line equation	**R^2 value**	**Used to calibrate**
Average of all scenes	y = -0.179x + 30.913	$R^2 = 0.99744$	Calibration image 1 & Calibration image 2
Calibration Image 1 (Marle)	y = -0.1963x + 31.36	$R^2 = 0.9334$	All other images except calibration image 2
Calibration Image 2 Crockery	y = -0.1961x + 31.579	$R^2 = 0.97203$	All other images except calibration image 1

Table 6.6 Ruler calibration

The calibration was carried out as follows:

(i) For all but the two calibration images: The results for each compressed image were averaged with those of the same level of compression in the two calibration images.

(ii) For the two calibration images: The results for each compressed image were averaged against the all scene average.

(iii) The calibrated results for each image were re-plotted and a linear trendline was fitted to the data. In this case a common starting image was assumed by extrapolating the data from all scenes (i.e. the y-intercept of the regression of all scenes). These functions were derived to allow modelling of the results from chapter 5 as SQS_2 values.

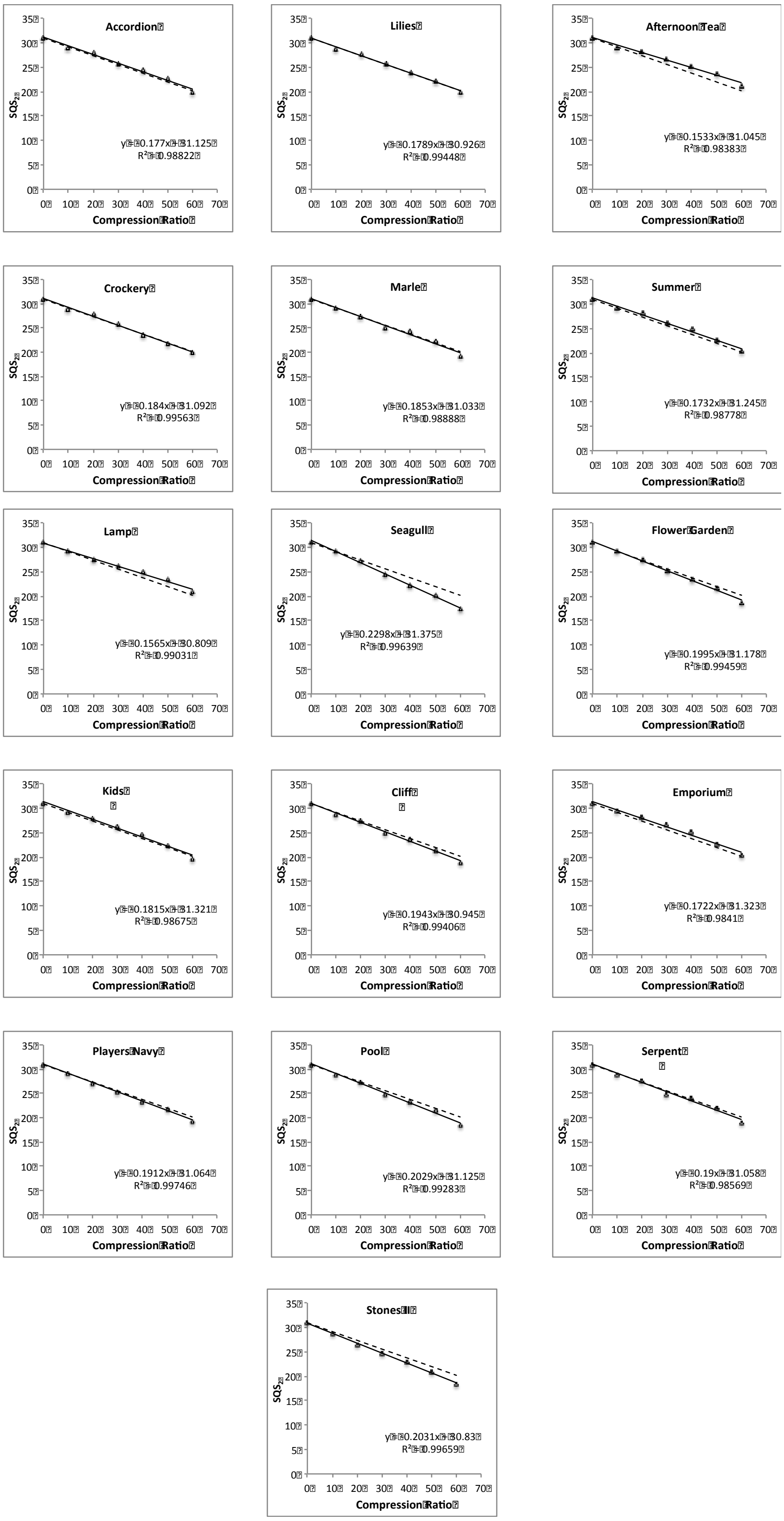


Figure 6.31 Quality Rulers after calibration

6.7 Summary

This chapter summarises an experiment to implement the soft copy quality ruler according to ISO 20462 part 3 [85] [85]for JPEG 2000 compressed images. The modelling of the system MTF was evaluated, and an aim MTF was defined. The approach used to adapt the system MTF to the aim MTF involved precise modelling in the frequency domain followed by the creation of a filter using a polynomial function. The ruler images were then created using the equation defined in the ISO standard to create a set of ruler images for each scene, varying in terms of sharpness. The quality rulers were used to implement a psychophysical experiment to determine interval scales of quality for JPEG compression using a sub-set of images from the set used for the investigation in chapter 5. The data clearly highlighted scene susceptibilities. The rulers were also useful in determining observer sensitivity. The ruler results were also calibrated in accordance with ISO 20462, which all but eliminated the scene dependencies in the results.

The SCQR ruler approach has clear application in future image quality studies. Because the image set has been characterised by scene and thresholds have also been separately evaluated, the subjective quality of the images is well defined, and the relationship with scene characteristics should make them useful as a test set of images in developing and adapting metrics. This is the subject of the experimental work in the next chapter.

7 Image Quality Metrics

The experiments in chapters 4, 5 and 6 have investigated a number of different psychophysical methods including the new soft copy quality ruler ISO standard for evaluating image quality. The results have provided an overview of the quality performance of JPEG 2000 and have also illustrated the complexity of implementing psychophysical studies and the analysis and interpretation of the results.

In many practical imaging applications there is not the time to implement a psychophysical study on a large enough scale to produce results quickly enough to keep apace of developing technologies. Therefore, as described in chapter 2, research into robust, predictive and easily applied image quality metrics remains an important topic, which has developed a great deal in the last decade. Section 2.8 gave an overview of just a few categories of the vast range of image quality metrics now available.

7.1 Selecting the metrics for use in this work

In selecting metrics for the final part of this experimental work, literature research was undertaken with a view to identifying and testing metrics that:

- Illustrate a variety of different approaches to metric design.
- Have been developed for general image quality applications, rather than for specific types of algorithms or artifacts.
- Are relatively straightforward to implement.
- Are flexible enough to allow some adaption.
- Have been tested and shown to be effective.

Four metrics were finally selected. These were:

1) The Modular Image Difference Metric (MIDM).
   This metric, developed first by Fairchild and Johnson [70] [68] was selected because of its modular structure, meaning that individual modules could be left out, added or tuned to explore the use of the

scene metrics used earlier in the research (chapters 4 and 5). Furthermore, some research, supervised by the author, had already been carried out using the MIDM on the images and the interval scaling results from in chapter 4 [173], which indicated that it might be effective as a model, particularly in predicting scene dependency in the psychophysical results.

2) The Structural Similarity Index Model (SSIM) [174]

   The SSIM has found widespread use since its introduction by Wang et al in 2004 [174]. It has been found to be as effective in predicting subjective quality as much more complex VIQMs. Many variants of SSIM have been developed, and there are readily available software implementations in Java and MATLAB, making it easy to implement. As it focuses on relatively simple scene descriptors to evaluate 'structure' within an image, it would also seem to be suitable for a scene dependency investigation.

3) The Multi-scale Structural Similarity Model (MSSIM)

   The MSSIM [175] is a variant of structural similarity, which includes a multichannel decomposition step prior to application of the metric. This is achieved through a down-sampling and low pass filtering process, not dissimilar to filter banks used in wavelet processing, therefore it might be considered highly applicable to JPEG 2000.

4) The Three-component Weighted Structural Similarity Model (WSSIM)

   The WSSIM [176] is an adaptation of either the SSIM or the MSSIM, in which the image is segmented into three broad categories: texture, edges and uniform areas. The SSIM or MSSIM map is calculated in the normal manner, and the segmentation is then used to weight the SSIM map components, according to their assumed differing perceptual importance.

7.2 The Modular Image Difference Model

The modular image difference model is a framework for a colour image difference metric [69], which has been incorporated into the ICAM appearance model [177]. The original framework was based upon the S-CIELAB [178] spatial extension to the CIELAB colour space. S-CIELAB was initially developed to adapt the traditional colour difference equations with a pre-processing step to simulate the human CSF. In the application in [178] the CSF was approximated using convolution filtering, with the aim of reducing the high frequency colour patterns (present for example in half tone printing) that are beyond the limits of the CSF to provide a better model of colour appearance and colour differences.

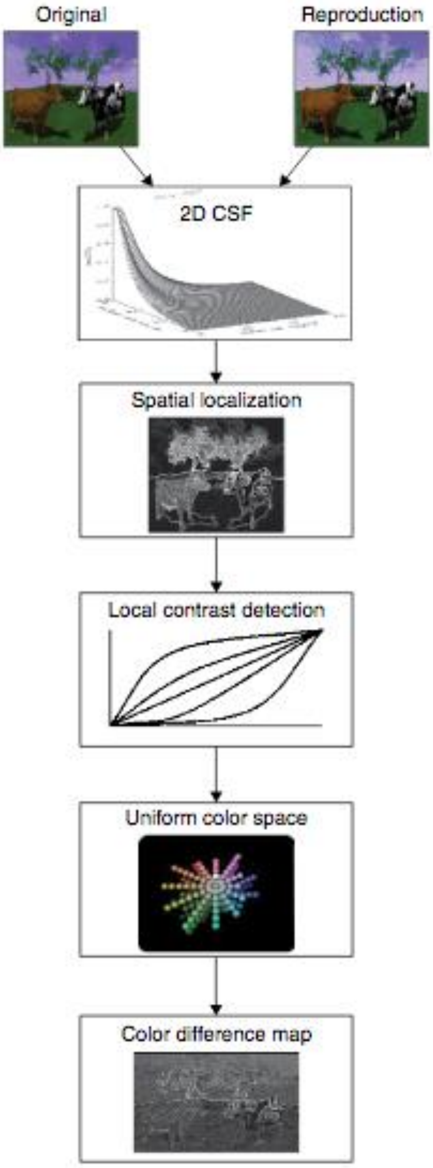


Figure 7.1 Flowchart of a modular image difference metric (image © M.D. Fairchild, from [70])

The modular image difference model has extended S-CIELAB, adding modules to take into account other aspects of visual processing. In MIDM [68] [70] both chromatic and achromatic CSFs are applied in the frequency domain, and there are further options for spatial frequency adaptation, spatial localisation of edges, and local contrast detection. The same set of modules is applied to both original and distorted images, and an error map is then calculated between the resulting images. The error map can then be used to generate a metric. A number of statistics may be used including the mean, median, maximum, which may describe different aspects of the data [70]. The steps in the modular image difference model are illustrated in Figure 7.1.

7.3 Implementation of MDIM in this work

7.3.1 Pre-processing: Colour space conversion

The model was implemented using the IPT colour space, an opponent colour space (I is the luminance channel and P and T are chroma channels), first proposed by Fairchild and Ebner [179]because of its uniformity of hue and suggested for use in the iCAM model.

The processing steps from sRGB values to non-linear IPT values were as follows:

1) Image normalised by dividing by maximum value
2) RGB channels linearised by application of the inverse of the XYZ to sRGB transfer functions
3) Linear sRGB values ➔XYZ (D65 adapting white point)
4) XYZ➔LMS (another cone response space)
5) Application of non-linear transfer curves LMS➔L'M'S'
6) L'M'S'➔IPT

All transfer functions and transform matrices are detailed in Appendix E. Step 5 can be optionally omitted, and in this case it was, to ensure that linear values were used for frequency space processing.

7.3.2 Application of CSFs and Spatial Frequency Adaptation

A variation of the Movshon-Kiorpes CSF was used for filtering the achromatic L channel defined by:

$$\mathrm{csf}_{lum}(\mathrm{f}) = \mathrm{a}.f^{c}.e^{-b.f} \qquad \text{where a=75, b=0.2 and c=0.8} \qquad (7.1)\ [70]$$

Where f is spatial frequency in cycles per degree of visual angle. The chromatic CSFs were defined by

$$\mathrm{csf}_{chrom}(\mathrm{f}) = \mathrm{a}_1.e^{-b_1.f^{c_1}} + \mathrm{a}_2.e^{-b_2.f^{c_2}} \qquad (7.2)\ [70]$$

Where (a_1, b_1, c_1, a_2, b_2, c_2) are (109.14, 0.00038, 3.424, 93.60, 0.000367, 2.1680) respectively.

The spatial frequency adaptation model used was the model proposed in [64] based upon the Natural Scene Assumption, also known as the 1/f approximation. This assumes that the probability of occurrence of any given frequency within a natural scene is inversely proportional to the frequency. This may be defined by:

$$\text{frequency of occurence(f)} = \frac{1}{f} \qquad (7.3)\ [70]$$

Therefore to model the effects of spatial frequency adaptation, it is assumed that each frequency can be divided by its frequency of occurrence, meaning that those frequencies most commonly occurring will be the most attenuated. This is achieved by manipulating the luminance CSF as follows:

$$\mathrm{csf}_{adapt}(\mathrm{f}) = \frac{\mathrm{csf}_{lum}(\mathrm{f})}{\left({}^{1}/_{f}\right)^{1/3}} = f^{1/3}.\mathrm{csf}_{lum}(\mathrm{f}) \qquad (7.4)\ [70]$$

The adaptation exponent of 1/3 is included to prevent too much attenuation of low frequencies and emphasis of high frequencies.

The adaptation was applied to the modelled CSF prior to it being applied to the image in the frequency domain. The shape of this function after normalisation is shown in Figure 7.2. Note that the DC value (at zero cpd) is maintained at a value of 1 in both cases. The original function when calculated from (7.1) [70] consisted of values ranging from 0 to approximately 120 for the peak value. The f term multiplier in the equations means that the DC value is going to be zero after application of the CSF filter. The DC value defines the average luminance value of the image, therefore it should remain unchanged during frequency space processing or the overall brightness of the image will change. After consultation by email with the authors of [64] the shape of the functions below was achieved by shifting the original luminance CSF towards the x-axis and by 1 cpd (also recommended in [64]). This meant that the zero cpd value was now what had been the 1cpd value (a value of approximately 60). This new value was used to normalise the entire curve. Although it meant that the peak of the CSF was shifted by 1cpd it was still within the range of 4-8cpd. Once normalised the curve maintained its band pass shape even after adaptation.

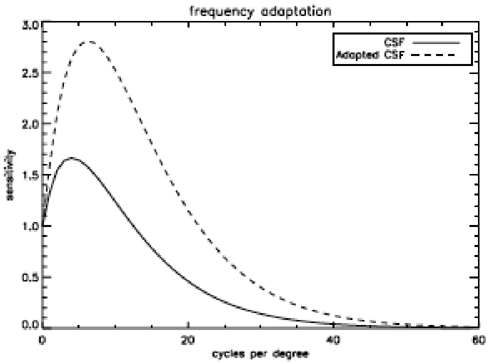


Figure 7.2 The Movshon-Kiorpes achromatic CSF before and after adaptation from [64].

Application of the CSFs was implemented in MATLAB as follows:

1) The CSFs and adapted CSF functions were defined as two-dimensional, rotationally symmetrical functions of the same dimensions as the image.

2) The image channels were separated and transformed to frequency space using the fast Fourier Transform function.
3) The magnitude and the phase of the frequency spectrum were separated. The CSFs were applied to the magnitude component only.
4) The inverse fft2 of the result was taken to produce a filtered image.

7.3.3 Conversion to Non-Linear IPT space

The frequency space processing had been applied to a linear image. The remainder of the processing was to be applied to images with perceptual non-linear transfer curves applied. The images were therefore transformed from IPT space to LMS values using the inverse of the forward LMS to IPT transform matrix. The LMS non-linear transfer functions were applied before the images were again transformed into non-linear IPT values.

7.3.4 Edge enhancement

The busyness metric derived by Triantaphillidou in [34] had been adapted for the work in chapters 5 and 6 to the different image size and viewing conditions and had been found to correlate with the subjective results of the psychophysical experiments. The metric consists of five stages:

1) The image is thresholded using Otsu's method to define a global contrast threshold. The method minimises intra-class variance between black and white pixels in the result. The position of the threshold is scene dependent.
2) An additional threshold is set, which is used as a multiplier for the threshold defined in (1). This multiplier defines which parts of the image will be included as part of the 'busy' segment. It is a function of image size and viewing distance and was determined empirically both in the work in chapter 4 [34] and in chapter 5, by applying the metric and evaluating the result. A threshold of 0.04 was used in the images in chapter 4, and a value of 0.13 was found to work best for the images used in chapters 5, 6 and 7.
3) Sobel filters are applied in the horizontal and vertical directions.

4) For the busyness metric, the resulting binary image was dilated, and holes filled to produce the segmented image.

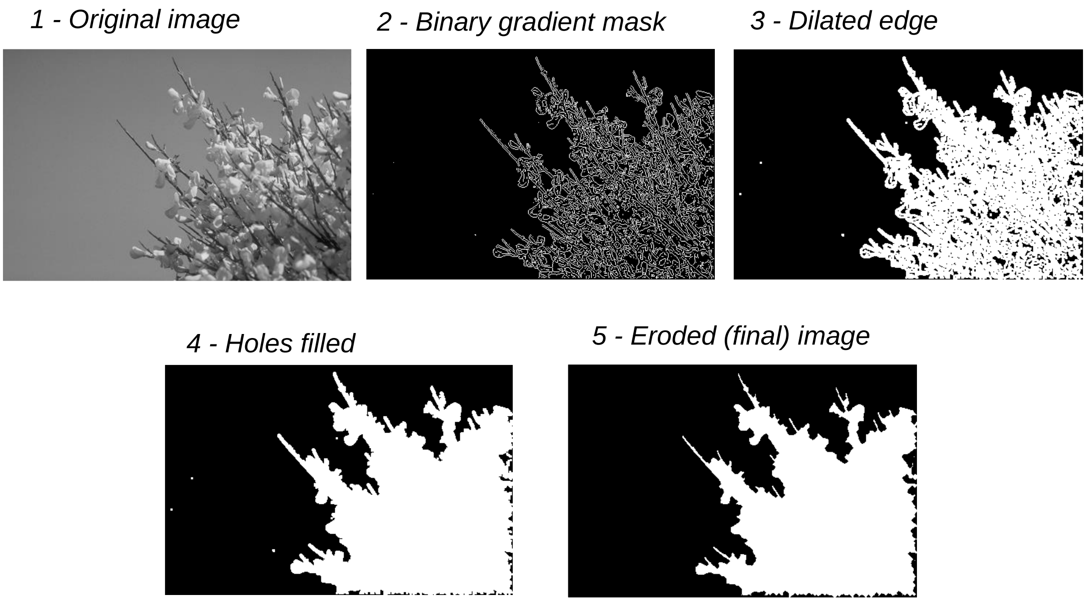


Figure 7.3 Stages of busyness metric (reproduced from chapter 4 and [34] for reference)

A version of this was adapted to provide edge localisation to the luminance image channel.

The adapted version was as follows:

1) The Otsu algorithm was applied to the original image to obtain the image specific threshold. The original threshold of 0.13 was maintained. This produced a thresholded mask image.
2) Horizontal and vertical 3 x 3 sobel filters were applied by convolution to the processed image to produce two gradient images.
3) The magnitude of the two edge images was taken, using: $gr\ ad=(gh^2+gv^2)^{0.5}$, where gh and gv were the horizontal and vertical gradient images.
4) The outputs from (1) and (3) were multiplied together to produce an edge mask. This was then added to the image to sharpen it.
The results from 1), 3) and 4) can be seen in Figure 7.4.

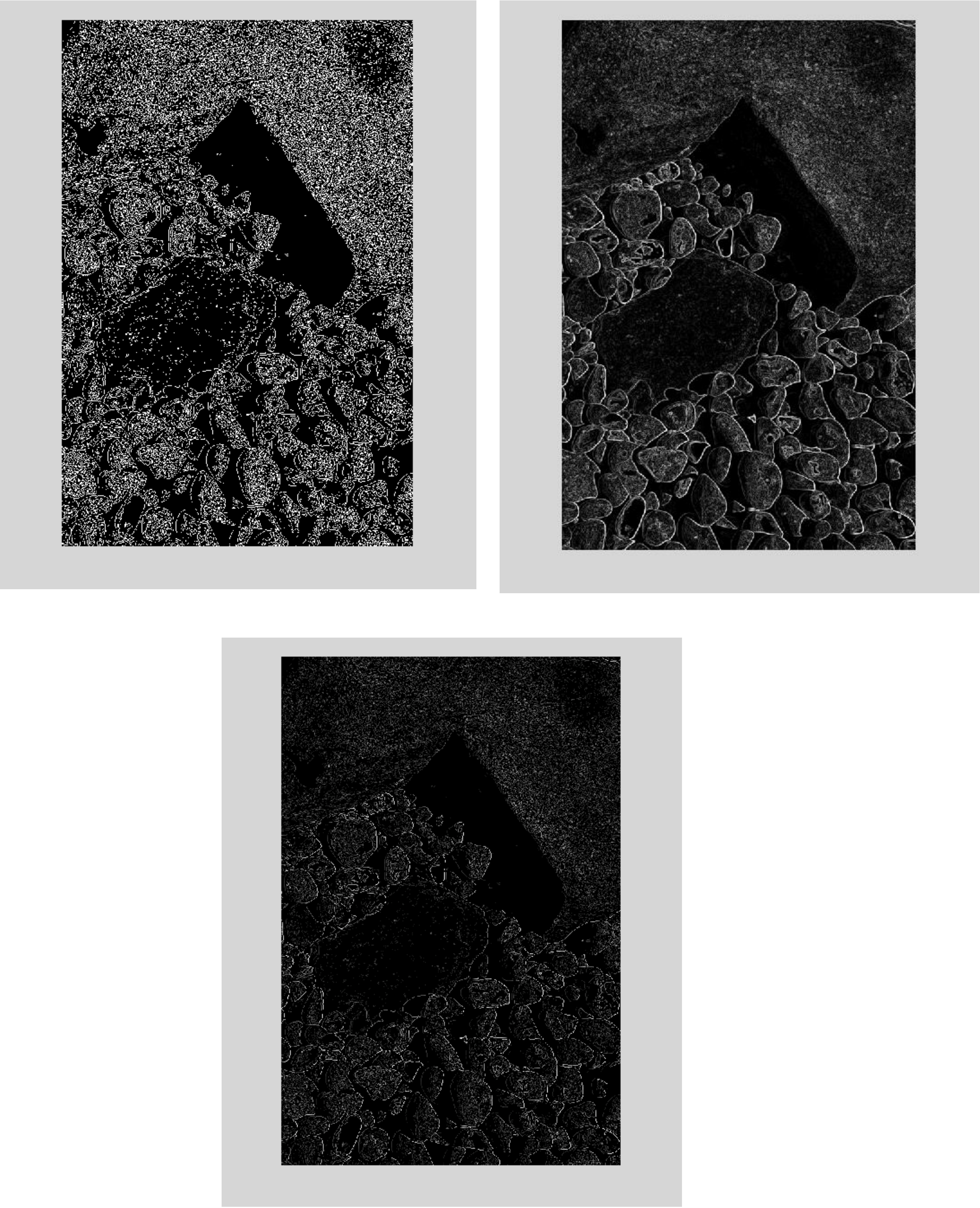


Figure 7.4 Edge masks developed from the busyness metric and used for spatial localisation module in the MDIM. Top left: thresholded output after applying Otsu's method to gain an image specific threshold. Top right: edge magnitude image. Bottom: resultant edge mask image, obtained by multiplying the other two images together.

7.3.5 Local contrast detection

A final module was used to implement local contrast detection based upon a method by Moroney [180]. A low pass filtered version of the image (obtained by applying a 10 x 10 averaging convolution filter) was used as a mask and the following equation was applied to provide local contrast correction:

$$output = max\left(\left(\frac{input}{max}\right)^{2^{\left(\frac{median-mask}{mask}\right)}}\right) \qquad (7.5)\ [180]$$

This function results in individual tone reproduction curves being generated per pixel – mask values greater than the median value result in an exponent of less than 1 and vice versa. Mask values equal to the median will give an exponent of 1 and the input value will be unchanged.

7.3.6 Calculation of error metric

The entire cascade of modules was applied to both reference and distorted images. From these I, P and T channels of the output of distorted image were subtracted from the output of the original to produce an error map. At each pixel position in the image, the following image difference measure was calculated, to give a single difference image. From this a mean value was calculated to give the final metric value.

$$\Delta Im = \sqrt{\Delta I^2 + \Delta P^2 + \Delta T^2} \qquad (7.6)\ [70]$$

7.4 Structural Similarity Approaches to Image Quality

According to Wang et al in [181]: ' The most fundamental principle underlying structural approaches to image quality assessment is that the HVS is highly adapted to extract structural information from the visual scene, and therefore a measurement of structural similarity (or distortion) should provide a good approximation to perceptual image quality'.

Structural similarity methods are top-down full-reference metrics, which aim to quantify distortions without the necessity to know anything of the specifics of the imaging system (other than the HVS).

As described in chapter 3, traditional distortion measures such as MSE do not correlate well with perceived distortion because they provide a simple difference measure, evaluating the magnitude of errors without reference to their direction, their visual impact, or the impact of masking processes; and based upon the assumption that individual image values are statistically uncorrelated and independent of each other. The problem with this approach is that very different distortions can produce the same level of errors, for example, if an image is slightly rotated with respect to the original, it will produce a very high MSE value, but have little impact upon perceived image quality.

The sophistication of the HVS is illustrated by various psychological phenomena known as *visual constancies.* These are learned principles of perception, which allow us to perceive objects in the world as the same regardless of the different images that they may project onto the retina. Examples include shape and size constancy, and colour and brightness constancy. The local effects of change to luminance or contrast, for example when an object is partially in shadow, do not prevent the HVS from understanding an image, as if the HVS is subtracting their effects during object recognition.

The structural similarity approach aims to separate out image quality attributes and in particular the influence of illumination, which may have impact upon local variations in luminance and contrast, from structural distortions [181], which are likely to have more of an impact on quality. A summary diagram of this approach is shown in Figure 7.5.

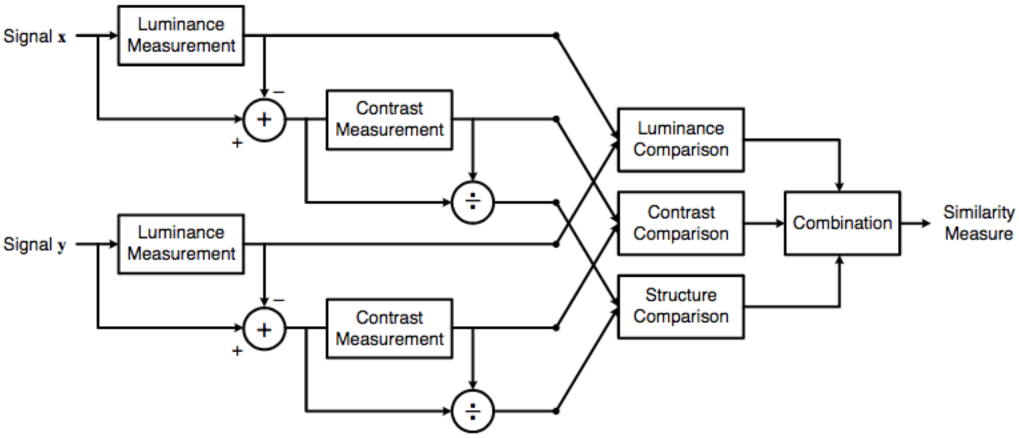


Figure 7.5 ‘Details of the Structural Similarity Measurement System’ from [174]

7.4.1 Structural Similarity Index Metric (SSIM)

The SSIM algorithm is usually applied only to a luminance channel, although this is not a requirement. Therefore for the purposes of this experiment, the image was converted from sRGB to the non-linear version of IPT (as described earlier and in appendix F). The metric was applied to the I channel only.

The SSIM metric was implemented in MATLAB using the SSIM.m code provided by Wang et al and available to download from [182]. The code includes a down sampling stage, in which the image is first low pass filtered and then down sampled by a factor determined by the image size. This is based upon advice given in the readme file on suggested usage. In this implementation, images were down-sampled by a factor of 2.

Let x and y denote the original and the distorted images respectively. The algorithm is applied to the image using a series of windows, the default size of which is 11 x 11 pixels, and the distortion map is formed from the combination of the outputs from all the windows. Using smaller windows enables the effects of local luminance and contrast differences to be evaluated using local statistics.

In each window, the following comparison measures are applied [181]:

1) The luminance of the signal is estimated by the mean intensity:

$$\mu_x = \frac{1}{N}\sum_{i=1}^{N} x_i \qquad (7.7)\ [181]$$

Where N is the total number of pixels and x_i is the intensity of the pixel at position i. The luminance comparison is a function of the mean intensities from the original and the distorted image (denoted by x and y subscripts) in the same window:

$$l(x, y) = l(\mu_x, \mu_y) \qquad (7.8)\ [181]$$

2) The standard deviation within a local window is used as an estimate of image contrast:

$$\sigma_x = \left(\frac{1}{N-1}\sum_{i=1}^{N} (x_i - \mu_x)^2\right)^{0.5} \qquad (7.9)\ [181]$$

And the contrast comparison is:

$$c(x, y) = c(\sigma_x, \sigma_y) \qquad (7.10)\ [181]$$

3) The signal is now normalised by dividing by its own standard deviation and the structure comparison is as follows:

$$s(x, y) = \frac{2\sigma_{xy} + C_3}{\sigma_x \sigma_y + C_3} \qquad (7.11)\ [181]$$

The constant C_3 is introduced top and bottom to prevent numerical overflow caused by a zero denominator.

The SSIM Indices are computed by a combination of the three measures with weighting parameters $\alpha, \beta\ and\ \gamma$:

$$SSIM(x,y) = \left[l(x,y)^{\alpha}.c(x,y)^{\beta}.s(x,y)^{\gamma}\right] \qquad (7.12)\ [181]$$

If α, β *and* γ are set to 1 and C_3= $C_2/2$, then a specific SSIM index is produced:

$$SSIM(x,y) = \frac{(2\mu_x, \mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu^2{}_x + \mu^2{}_y + C_1)(\sigma^2{}_x + \sigma^2{}_y + C_2} \qquad (7.13)\ [181]$$

This final index was applied to each window, to produce an SSIM map the same size as the down-sampled image. As for the modular image difference model the statistics from this map can be used to provide single valued measures of quality. In this experiment the mean of the SSIM index map was taken.

7.4.2 Multi-scale Structural Similarity Index Metric (MSSIM)

The MSSIM [175]uses a lifting structure to compute the SSIM at different scales. This could be interpreted as incorporating an approach similar to the multichannel model of the HVS into the metric. As JPEG 2000 is based upon a multi-resolution transform, this might prove to be a particularly appropriate metric.

The metric uses multiple stages of low-pass filtering followed by down-sampling as shown in

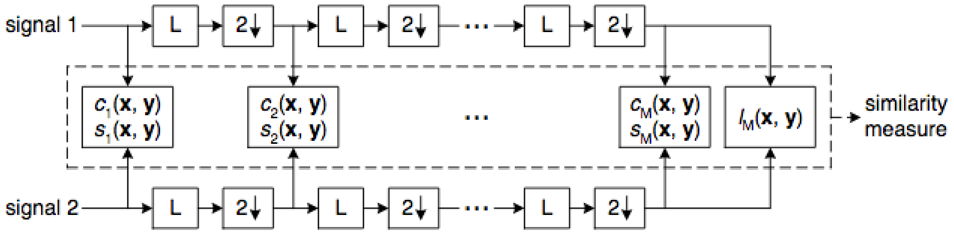


Figure 7.6 'Multi-scale structural similarity measurement system' reproduced from [175]

The original and distorted images are each passed through a low pass filter, and then down-sampled by 2. At each scale the individual comparison measures are computed (although not all of them are computed at every scale).

If the original scale is scale 1 (i.e. before any down-sampling has taken place) and the highest scale is scale M, the MSSIM is expressed as follows:

$$MSSIM(x,y)=\left[l_m(x,y)\right]^{aM}\overset{M}{\underset{j=1}{\tilde{O}}}[c_j(x,y)]^{bj}[s_j(x,y)]^{gj} \quad (7.14)\ [175]$$

The luminance comparison is only computed at scale M, while the contrast and structure comparisons are computed at all scales. The weighting factors are separate for each scale. As described in [175] the weightings could be linked to the CSF, but it is pointed out that the CSF functions are calculated at visibility thresholds whereas the images being evaluated are complex and the distortions are supra-threshold.

In this work, the MSSIM was applied again to the image in a non-linear IPT colour space on the*I* channel only. The algorithm was applied in MATLAB, adapted from code written by the authors of [175] and downloaded from [182].

7.4.3 The Three-Component Weighted SSIM (WSSIM)

The three-component weighted SSIM was developed by Li and Bovik and is described in [176]. While the SSIM separates out and weights components of the scene, this adaptation considers the scene in terms of three types of 'features': edges, textures and uniform areas. The three areas can be unequally weighted depending upon the distortion being tested. The reasoning behind this separation is based upon the relative importance of edges for object recognition, and the high sensitivity of the HVS to edge distortions, particularly in the luminance channel. If an image has a large area of texture however, this may mask any distortions. Although distortions will show up in uniform areas of an image, these can be the least susceptible areas to errors and therefore they can be considered to have less of an impact on image quality.

In this research the busyness metric was used to segment the image in terms of its texture. As in 7.3.4, horizontal and vertical Sobel filters were used to identify edges, and were combined to produce a magnitude image. From testing it became clear that a 3 x 3 Sobel was not adequate to identify significant edges in all images. The gradient magnitude image was therefore processed with a dilation filter above a threshold, which was determined from the edge image again using the MATLAB graythresh function, which identifies a scene specific threshold using Otsu's method. This had the effect of identifying and enhancing the strongest edges within the image (assumed to be the most visually important). The output of this operation was a binary edge image mask.

The busyness metric as described in 5.4.1 was again used with a threshold set at 0.13, which had been determined visually through inspection of the segmentation of a number of images. The output at this stage was a binary mask, which segmented the image into busy and non-busy areas. The edge mask image was subtracted from this mask, leaving a mask, which identified busy areas, but did not include their boundaries, which were part of the edge mask.

The final stage was to compute the uniform areas binary mask. This was computed by subtracting the other two masks from a mask image, and so included whatever was left. The output masks for the three image areas are shown in Figure 7.7.

The SSIM metric was applied to the entire image to create an SSIM index map. The masks were then each multiplied by the map to get SSIM maps for each feature. The average value for each SSIM map was then computed using only the non-zero elements (i.e. those included in the mask). Finally the mean values of the masks were weighted and combined. In the original implementation of this method the weights proposed in [176] were: 0.25 for textured and uniform areas, and 0.5 for edges. However, it was found through testing that this approach gave too much weight to uniform areas. Alternative weights were therefore tested and the eventual weightings were: Edges=0.6; textures=0.3 and uniform areas=0.1.

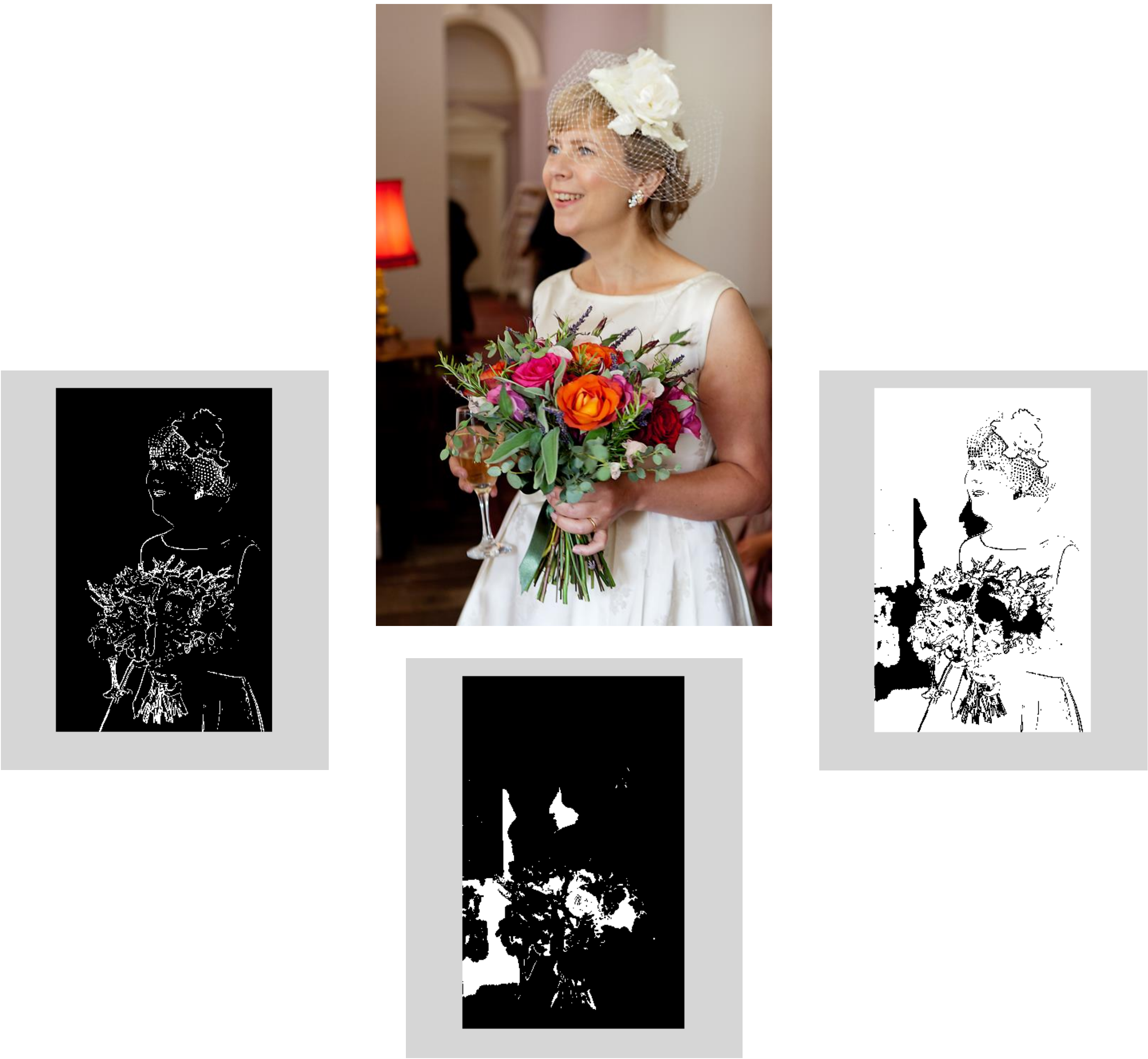


Figure 7.7 Example of image segmentation in the Three-Component Weighted Structural Similarity model. Anti-clockwise, from the left are the masks for edges, texture areas and uniform areas. White pixels indicate that pixels are classified in that image feature.

7.5 Results

The error metrics were calculated for all of the images and the results plotted against the quality ruler results from chapter 6. Because the MDIM is a measure of errors, or differences from the (group average) 'original' in terms of JNDs, the quality ruler values were calculated in terms of quality loss from the original (by subtracting the original average value from each) and these were used in the plot instead of the raw data. For the three structural similarity metrics, where the values reflect a measure of similarity rather than difference, the values were plotted directly against the JNDs.

The JNDs values used were from the raw un-calibrated data. This was because one of the aims of the work was to explore scene susceptibilities and the calibration process removed these to a large extent.

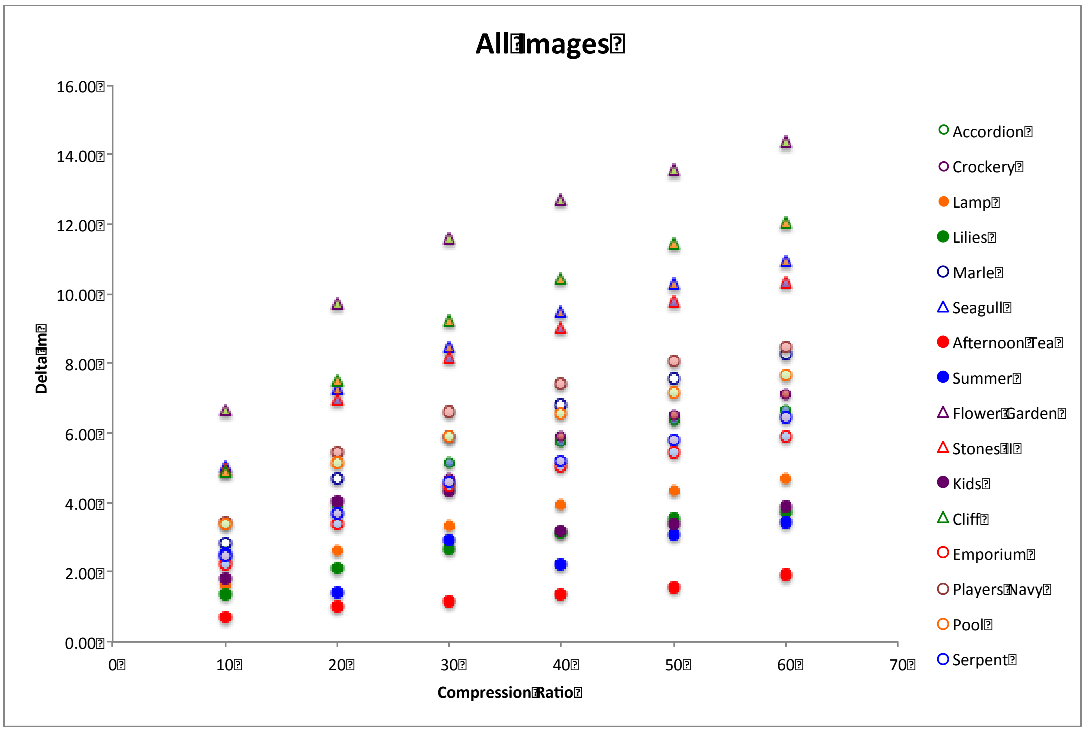


Figure 7.8 Results for the modular image difference metric. High values indicate larger visual differences

The results from the modular image difference model were plotted against compression ratio. Figure 7.8 illustrates a clear separation in the response to by the metric to different image types. In particular the scenes at the bottom of the plot are those that show the least difference from the original. The three lowest images were 'Lamp', 'Summer' and 'Kids'. The images at the top of the plot, indicating the largest differences were images with lots of texture: 'Flower Garden', 'Seagull', 'Cliff' and 'Stones II'.

The results for all four metrics are shown as scatter plots, with objective metric plotted against subjective ratings, or difference ratings, in figures

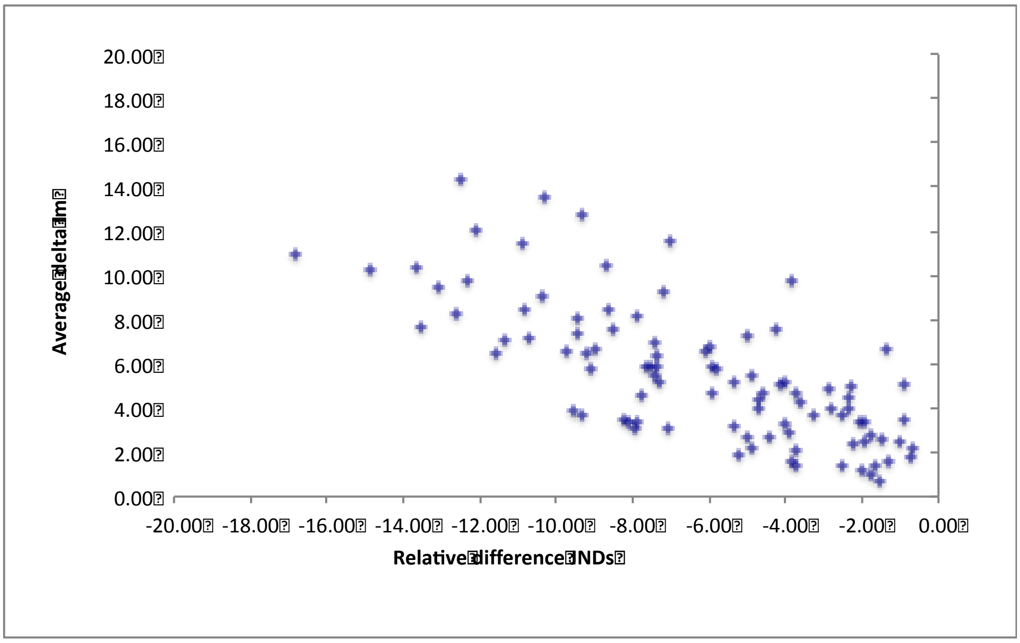


Figure 7.9 Average ΔIm from the MIDM values for all images plotted against relative difference JNDs from the data from the soft copy quality ruler in chapter 6.

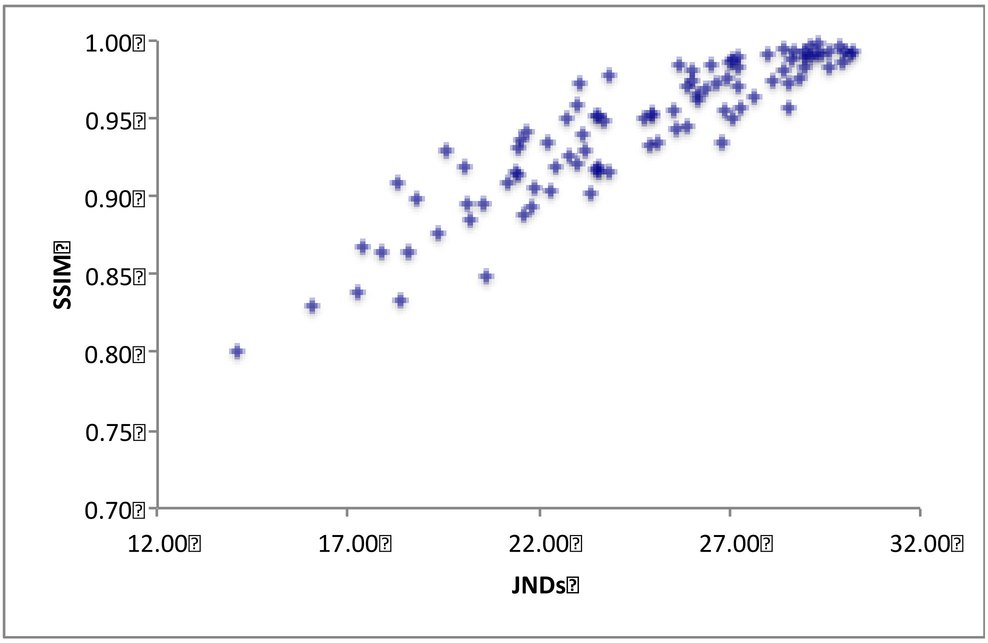


Figure 7.10 Mean SSIM values for all images plotted against JNDs obtained from the quality ruler experiment in chapter 6.

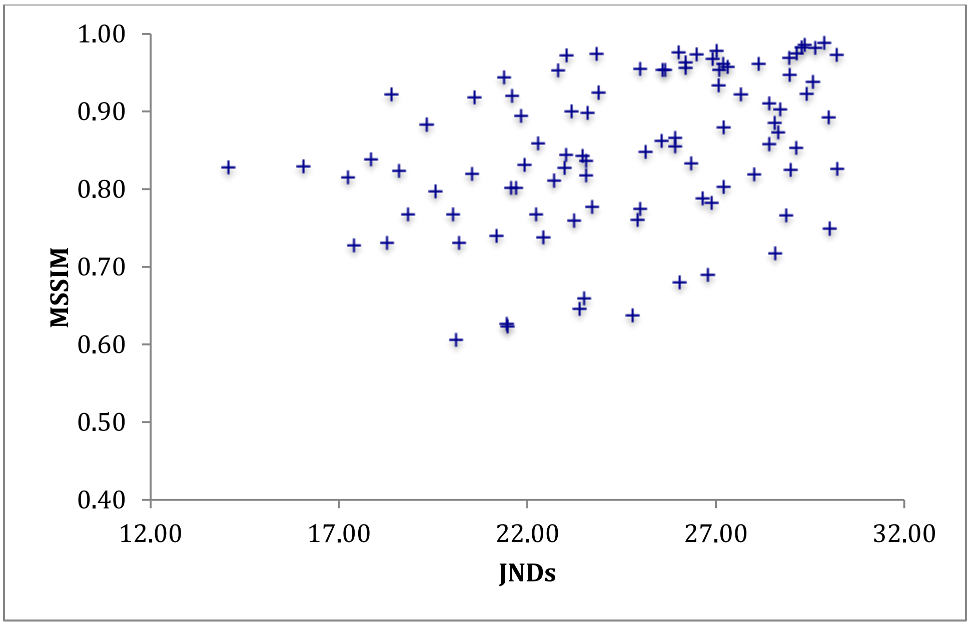


Figure 7.11 Mean MSSIM values for all images plotted against JNDs obtained from the quality ruler experiment in chapter 6.

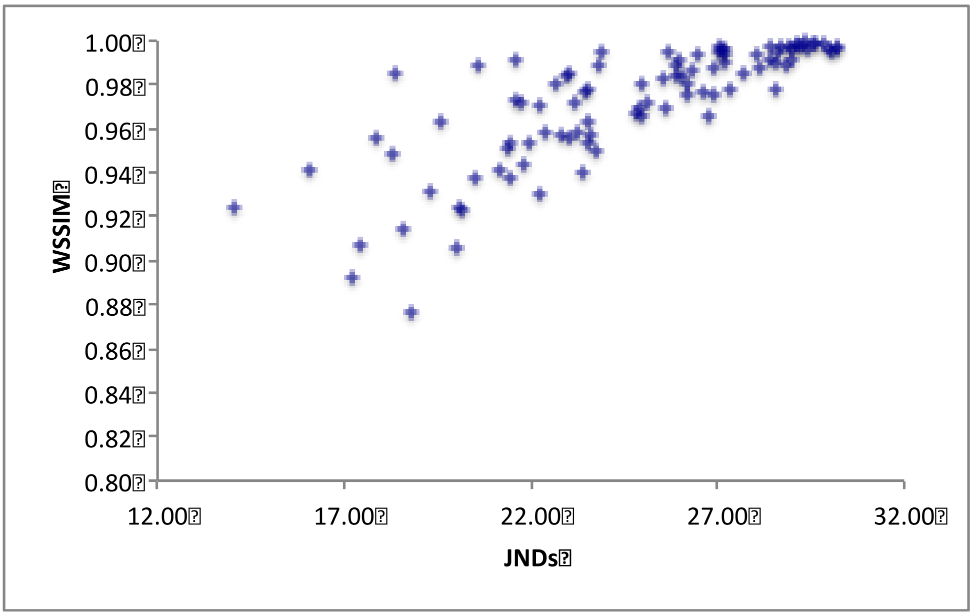


Figure 7.12 Mean WSSIM values for all images plotted against JNDs obtained from the quality ruler experiment in chapter 6.

The data in all cases appears to exhibit a correlation between the metric scores and the subjective JNDs. Therefore correlations between the data were calculated using Pearson's correlation coefficient for linear correlation and Spearman's Rank correlation coefficient. In all cases p-values were calculated for the null hypothesis (i.e. a measure of how likely the results would be if the null hypothesis was true). The correlation coefficients and p-values are shown in Table 7.1, and indicate strong correlations between the SSIM, MIDM and WSSIM and the subjective data and moderate correlation between the MSSIM and subjective scores. P-values against the null hypothesis indicate that all the results are significant.

	Modular Image Difference Metric (MIDM)	Structural Similarity Index (SSIM)	Multi-scale Structural Similarity Index (MSSIM)	Weighted Structural Similarity Index (WSSIM)
Pearson's Linear Correlation Coefficient	-0.7333	0.9123	0.3855	0.7978
ρ value	2.067 x 10^{-17}	3.716 x10^{-38}	1.0500 x 10^{-4}	2.2618 x 10^{-22}
Spearman's Rank Correlation	-0.7296	0.909900	0.438300	0.841700
ρ value	3.4039 x 10^{-17}	1.0831 x 10^{-37}	7.9384 x 10^{-6}	6.7691 x 10^{-27}

Table 7.1 Correlation coefficients and p values for the four metrics against subjective JNDs of quality from the experiment in chapter 6

7.6 Summary

Four objective metrics were tested with a subset of the image set generated for the experimental work in chapter 5. Two types of metrics were chosen: the Modular Image Difference Model (MIDM), which may be regarded as a bottom-up appearance modelling metric, which uses modules to model the effects of the HVS on data and then evaluates the difference between the original and distorted image; and three types of Structural Similarity Index Metrics, which evaluate similarity between original and distorted image by using separate terms for luminance, contrast and image structure.

The MIDM was the most complex to implement but allowed a great deal of flexibility in terms of allowing individual modules to be included, omitted or 'tuned' for a particular image type or context.

8 Discussions

The aim of this research has been to explore image quality evaluation in relation to lossy compression as an example of one of the many sources of digital artefacts that may be introduced during the imaging chain. There are many different approaches to image quality evaluation depending upon disciplinary perspectives and imaging context. When usefulness or fidelity are not key requirements of an imaging process, acceptable image quality is more difficult to define. This research has focused particularly upon scene dependency inherent in image quality evaluation. Although JPEG 2000 is no longer a new compression method, and numerous studies have been done to evaluate its performance since its introduction, the algorithm has not been as widely adopted as baseline JPEG, other than in specialist imaging applications such as medical or forensic imaging and it is in these contexts that research has tended to focus. There is relatively little research that has explored subjective image quality of JPEG 2000 in a general purpose imaging application. This may be because of the increased use of raw workflows in professional imaging, partly facilitated by improvements in processing and storage. It may also be because the JPEG algorithm is good enough when applied to high resolution and high quality images, to allow it to remain the de facto standard image format for multiple imaging applications. Nevertheless, as the focus of this research has included the impact of scene dependency on image quality studies, JPEG 2000 has proved to be a useful case study.

8.1 Quality Comparison of JPEG and JPEG 2000

Although some comparative studies of JPEG and JPEG 2000 have been implemented, they have tended to focus upon distortion metrics, and so there are not many results from subjective studies available.

The performance of JPEG and JPEG 2000 was compared in this experimental work using paired comparison tests and a series of images of varying scene content. A particular focus of the work was the scene dependency of the two algorithms. Scene dependency is well known to affect image quality evaluation.

Jacobson and Triantaphillidou in previous work [22] proposed that scene classification using simple metrics might be a useful approach to evaluating scene dependent effects in image quality. Steingrimmson et al [120] suggested that some less predictable results in a JPEG and JPEG 2000 comparison study might be linked to the scene susceptibility of the algorithm.

The architecture of the two compression schemes causes certain specific and different artefacts. In JPEG the blocking artefact is the most obvious, caused by inaccuracies in reconstruction of image blocks as a result of quantization resulting in discontinuities at block edges. Like contouring (and sometimes with a similar appearance), blocking was highly visible in uniform or low frequency areas. JPEG 2000 does not suffer from blocking unless the image is tiled. The separate encoding of different scales of the image as a result of the multi-resolution nature of the DWT and the variable size of the quantizers used in the different sub-bands mean that the most obvious JPEG 2000 artefact is a localised blurring artefact (see Figure 4.2). Both algorithms also suffer from ringing artefacts and colour distortions. JPEG is more susceptible to ringing; it is the result of the truncation of high frequencies, which in JPEG are localised by image block. The effects of ringing in JPEG 2000 are at different scales, and this may be the reason that JPEG 2000 is better at compressing text.

A set of images was selected for the experiment with varying scene content and characteristics, to enable the exploration of the effects of scene dependency of the algorithms. The scenes were classified in a separate study [34]using a range of different scene metrics to quantify attributes of tone, colour and scene busyness and information content. The original images were obtained from various sources, including some of the ISO standard test images and some images from a Kodak Photo CD. They were obtained in their original form at a low enough resolution so that they might be displayed at full size on screen without requiring resizing and incurring interpolation artefacts. The low resolution of the images meant that they were highly susceptible to compression artefacts from both algorithms.

The different appearance of the artefacts meant that they were more or less obvious in different types of scene content. Blocking tended to be more visible in uniform areas. Blurring artefacts became very evident around edges where they appear smudged, and in areas of random texture. Ringing was problematic near high contrast edges.

The results of the paired comparison test indicated that on average JPEG 2000 images were preferred over JPEG for most scenes across the entire compression range. The results for individual scenes however indicated clear scene dependency of the two algorithms and as observers assessed all images suggested that this was less to do with observer quality criteria but predominantly as a result of the scene dependency of the algorithms and masking effects. The results from this and the associated study indicated that scenes could be grouped according to their relative susceptibility to the artefacts of one or the other algorithm. The scene measures used supported these results. The images used in this study were very low resolution and therefore it could be suggested that they were more susceptible to errors than the file sizes more typical from contemporary digital cameras.

JPEG 2000 was found to clearly outperform JPEG in scenes that were predominantly light in tone, achromatic, and possessing large uniform or low frequency areas. These scenes were successfully grouped from their results from scene metrics quantifying scene global intensity and busyness.

The largest quality loss from lowest to highest compression ratio was found in scenes that shared a range of characteristics including significant amounts of fine detail in areas which might be considered to be focal points. These scenes were correlated in their results from the scene measures, with high values and rankings for busyness, colourfulness, and number of lines, and average to low rankings for global intensity. The subjective ratings for both algorithms were very low at high levels of compression. JPEG 2000 performed better than JPEG possibly because all had text or periodic patterns within them, which were very susceptible to ringing.

There were some scenes in which JPEG was preferred to JPEG 2000 either at very high compression ratios or across the entire range. These images were average for many of the scene measures but contained some texture or multi-coloured detail, which was blurred by JPEG 2000, while blocking artefacts were masked.

The results indicate that the nature and visibility of artefacts are of primary importance in image quality evaluation of compression. Their interaction with scene content produces results that are difficult to predict using simple distortion metrics. Scene metrics proved useful in grouping images according to their scene content and susceptibility to the processes.

Key findings from the experimental work in chapter 4 are as follows:

JPEG 2000 is capable of achieving much higher compression ratios than JPEG across most images. The results from both JPEG and JPEG 2000 are highly scene dependent, due to the nature of their characteristic artefacts. Such scene dependencies are mainly due to the architecture of the algorithms and their operation on specific scene content, as well as the visibility of the artefacts in particular scenes.

For most scenes, there are small gains in quality for JPEG 2000 compared to baseline JPEG across the entire compression range (up to 80:1). JPEG 2000 outperforms JPEG in terms of subjective quality for the majority of images at high compression rates (>60:1). This is likely to be due to the localisation of errors within JPEG, and the visibility of the blocking artefacts produced by JPEG compared to the smoothing produced by JPEG 2000. The differences in performance between the two algorithms is much less noticeable at lower compression ratios, (<40:1) and indeed, the slight sharpening effect of the increased ringing artefacts in JPEG is judged as a quality improvement in some images.

At high compression ratios blocking artefacts are generally more bothersome than smoothing artefacts in images containing large areas of flat tone or low frequencies. However the opposite is the case in areas of texture, which do not mask the blurring artefacts so well and where fine detail can start to break down very visibly.

JPEG 2000 produces less distortion of text and numerical data than JPEG. Large areas of fine detail within images may mask blocking artefacts, and in such images the smoothing artefacts produced by JPEG 2000 may reduce perceived image quality.

JPEG 2000 outperforms JPEG in terms of error resilience across most images and most of this compression range. PSNR is an inadequate predictor of subjective image quality and in particular the scene dependency affecting image quality studies.

8.1.1 Additional experimental work based upon the first image set

Following the work detailed in this chapter, a number of additional studies were undertaken using the same set of images.

The first was an investigation into perceptibility thresholds for JPEG 2000, the results of which were presented at the Royal Photographic Society Digital Futures conference (see Related Work, 10.2). Using a paired comparison test to evaluate thresholds of perceptibility of distortions, this work laid the foundations for the work undertaken in chapter 5. Key findings from this work were that the perceptibility thresholds were highly scene dependent, that the average 50% JND threshold (i.e. where 75% of observers noticed a difference) for this image set was close to 20:1 and that, as had been demonstrated in the experiment from chapter 4, the thresholds were correlated with image busyness.

Additional work was carried out by Orfanidou et al [183] used the same set of images and the interval scaling results to investigate the use of the Modular Image Difference Model for image quality evaluation. The implementation was somewhat different to the version described in chapter 7 of this research. Orfanidou found that the metric was able to usefully predict the scene dependencies in the subjective results.

The images used in chapter 4 were of rather low resolution. Increases in potential bandwidth and improved sensor technologies even in mobile phone cameras mean that typical image resolutions tend to be higher than the ones investigated here.

8.1.2 Recent Research into JPEG, JPEG 2000 and other standards

Research has been ongoing into JPEG 2000, with particular reference to its applicability and performance in specialist imaging applications. As a compression format it has been adopted, for example, in various applications in medical imaging. It is the preferred format [184] for the lossless encoding of Digital Mammogram images, providing many features that are useful in reliable image transmission and fast image database access.

Recent research [185] has explored the use of JPEG 2000 as a visually lossless format for Remote Image Browsing. The format lends itself particularly to this application, as it is now often a requirement to allow panning and zooming when viewing images remotely. The resolution scalability of JPEG 2000 helps to facilitate this process. The research proposed embedding scalable quantization step sizes into the JPEG compressed bit stream, to allow the image to easily be zoomed at multiple resolutions. The researchers found that the use of the JPEG 2000 resulted in significant reductions in required bandwidth compared to other formats.

A recently published study [186]compared the performance of JPEG 2000 and JPEG with a newly developed adaptation of baseline JPEG 1992 algorithm, the CSI-JPEG, which uses cubic spline interpolation. The performance of the three algorithms were evaluated in terms of compression rate, colour accuracy and visual quality and found that JPEG 2000 and CSI-JPEG outperformed baseline JPEG 1992 for small colour differences. These results were correlated with visual data.

The Joint Photographic Experts Group recently brought out a new standard, JPEG XT (for JPEG Extension) [187], which has a significant advantage in that it is backwards compatible with JPEG (whereas JPEG 2000 is a completely different algorithm). This new format makes use of the additional functionality in the original JPEG specification that has not found previous widespread use. For example, there is a high precision 12-bit depth mode in the original specification that was not included in the JPEG File Interchange Format (JFIF) (one of the most widely adopted implementations of baseline JPEG). JPEG XT offers a number of useful features, such as this higher bit-depth encoding and, enhancements to allow high dynamic range encoding which will allow the original 8-bit version of an HDR image to be stored in the baseline format with an extension layer for the High Dynamic Range version of the image. JPEG XT also supports lossless encoding. Although there is a lossless mode in the original JPEG standard it has not been widely adopted and therefore is not included in many implementations of JPEG. Further features include the coding of opacity information and a privacy and security standardisation initiative to enable control of image distribution across networks. Potential future extensions include the possibility of encoding 360-panoramic images and animated JPEGS.

8.2 Perceptibility and Acceptability of JPEG 2000

This study explored the relationship between perceptibility and acceptability thresholds of compression across a range of different scenes. The results indicated a significant correlation, for most images, between perceptibility and acceptability thresholds for JPEG 2000.

Although the acceptability context was not clearly defined to observers prior to the experiment, the results for acceptability thresholds across the observers were relatively consistent and the derived psychometric curves fit the observed acceptability data reasonably well, for the majority of images.

Scene characteristics of the test images were evaluated using simple scene descriptors (median, variance, histogram skewness, busyness, chroma variance) [35,34] A strong statistical correlation was found between the busyness descriptor and both perceptibility and acceptability thresholds across all the scenes, demonstrating the susceptibility of highly textured scenes to JPEG 2000 distortions as well as the scene dependency of the algorithm, due to the localized nature of the blurring distortions [33] [34]

The images were grouped according to the level of their perceptibility and acceptability thresholds, and the values from the scene metrics. Scene characteristics other than busyness did not correlate with the thresholds consistently across all of the images, but there was good correlation within the image groups, particularly with the descriptors for scene lightness (median and skewness).

Images with high thresholds were found to have low busyness and either higher than average, or lower than average lightness. In these cases, the contrast of the blurring distortions affecting light or dark areas within the images was low and therefore less visible. The reduced contrast also meant that the ringing artifact was not very visible.

Images within groups with low thresholds were also found to correlate across the scene descriptors. Busyness was the biggest influencing factor, but its effect on the thresholds depended upon the visual importance of the busy areas within the image. If the image area contained a large proportion of busy areas, or if important features were very detailed, distortions (particularly the blurring artifact) became both visible and bothersome. The majority of images with low thresholds were low to average in terms of lightness.

The other scene descriptors (variance and chroma variance) were also reasonably well correlated within the groups. Chroma variance in particular seemed to be higher in (most of) the images with low thresholds. This might be because either the blurring artifact is reducing contrast, or that the blurring and ringing artifacts are obvious in these images, which contain lots of high contrast and highly chromatic areas.

Research by Alers et al [188] has shown that image regions are unequally weighted in terms of visual significance by observers in image quality studies. The scene dependency of JPEG 2000 and the localization of its distortions mean that it affects some image areas more than others. The distribution of salient features [189](i.e. significant focal points in the image), their area in relation to the overall image area, and their susceptibility to distortion as an influence upon image quality warrants further investigation in this context.

The results from scene descriptors indicate that in some cases they are useful for predicting performance. But the choice of metrics will define this, and other metrics might better identify natural texture, for example. The metrics have proven useful in grouping scenes, and particularly in identifying scenes that are more or less susceptible to JPEG 2000 distortion artefacts.

8.3 Soft Copy Quality Ruler

This investigation implemented the soft copy quality ruler according to ISO 20462 part 3 for a subset of the scenes compressed with JPEG 2000 for the investigation in chapter 5. The results indicate that the quality ruler can usefully be used to determine quality loss as an alternative to a paired comparison test. The SCQR was complex to prepare but once the JND filters were defined it was found easy to create quality rulers from a range of different scenes.

Certain aspects of the imaging workflow meant that the final measured system MTF did not conform well to the aim MTF. In particular, the processes of sharpening and downsizing for display meant that the system MTF deviated significantly from the aim MTF particularly at lower frequencies between 0 and 10 cycles per degree. Further investigation into the effects of non-linear processing on ruler results is warranted. The effects of the processing upon the MTF were not anticipated at image capture, because the image capture and processing was carried out before the decision to use the SCQR was made and the non-linearities in the system MTF therefore had to be accommodated in modelling of the aim MTF, which added an extra layer of complexity to the work.

The approach used to both shape the system MTF to the aim MTF and to produce the JND filters is novel. The method was highly successful in modelling the aim MTF. However, the best-fit polynomial function had to be determined by a process of trial and error. One of the issues was the need to ensure that there was no sharp drop off of values in the MTF, which would have introduced ringing in the image. Further work could be done to investigate other methods for producing a smooth function in the frequency domain, for example the use of a windowing function.

These complications could be avoided if the implementation of the SCQR was known beforehand, by further limiting the processes affecting the system MTF. This would be perfectly possible if the images were going to be developed as a reference set. But it is more difficult to apply such restrictions if quality rulers are to be generated for many typical image processing workflows.

Observers found the ruler relatively intuitive to use once they fully understood the task. Their comments indicated that some images were more difficult to scale than others, particularly those with differential sharpness in the scene, which might be a limitation in its application. The test was quite time consuming for some observers, perhaps because it took a while for them to understand the task. It would be useful to develop a test set of images, which could be used as a training set. A number of images could be randomly repeated in some of the sets of ruler images to investigate the consistency of observer results.

+/-1 standard errors were used as error bars and these were plotted on all of the initial results. They indicated close agreement in the results at the top end of the quality ruler closer to the original, whereas the errors were much larger for all of the images at lower quality levels. This indicates that there is significant variability in tolerance of visible distortion across groups of observers.

The quality ruler was successful in identifying most and least susceptible scenes, with results that correlated well with those from the paired comparison thresholds experiment in chapter 5. The ruler was also useful in identifying the most and least sensitive observers. The results indicated however that in this experimental work these observers did not bias the overall results.

The majority of ruler results appeared to fit a linear function. This was a somewhat surprising result, indicating that the perceived quality loss changes approximately linearly with compression ratio. However, the compressed images spanned a relatively wide range, and with only six images compressed images per scene it is possible that different results would have been achieved with a more finely sampled range of compression rates.

The lack of inclusion of an original uncompressed image within the group of rulers is acknowledged as a weakness in these results; however, the scene average linear regression was extrapolated to model an average original scene. This was felt to be a pragmatic approach, particularly as there was good agreement in the scaling of the images at the top end of the scale (below the perceptibility distortion threshold) leading to low errors across the entire set of images. The use of the additional (hypothetical) uncompressed image was included and the trend lines were recalculated. It was found that there was very little difference in the gradient or position of the trend line in nearly all cases, but that the correlation of the data to the regression line was improved Figure 8.1.

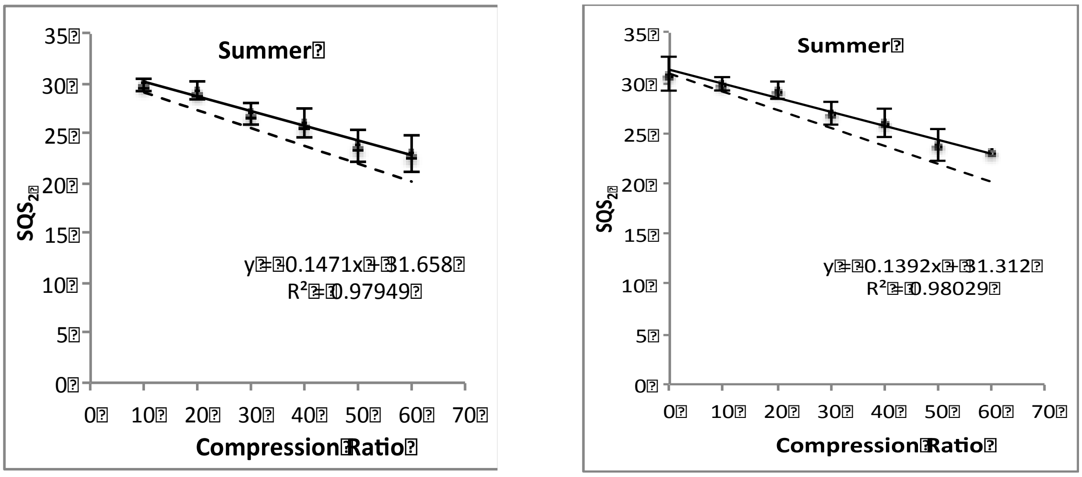


Figure 8.1 Quality ruler scales with and without a hypothetical average starting point.

The quality rulers were calibrated according to the *average scene relationship* detailed in ISO 60462 [85] and the results can be seen in Figure 6.31. What is most interesting to note is that the scene susceptibility has almost entirely disappeared as a result of the averaging process. While it is useful to have calibrated rulers for future work, it is less informative in this work when trying to explore scene susceptibilities. It is suggested that the rulers should additionally be calibrated against the DRS, to check whether this calibration method is accurate.

On a conceptual level the quality ruler is an interesting approach. There are questions however about using a ruler based upon sharpness when one of the distortions that is being explored has a significant effect on sharpness.

8.4 Image Quality Metrics

The final experimental work in this research explored a number of objective metrics. Two types of metrics were investigated, one which attempted to model some of the perceptual effects of the HVS from the bottom up, and three variants of another type that simply looked at the structure of the image itself.

The Modular Image Difference Metric was somewhat complicated to implement, but once the basic framework was created, its modular nature (which was reflected in the way that it was programmed) meant that it was quite easily adaptable, and there is potential for it to be further developed. The individual modules were not tested beyond the implementation due to time constraints. Aspects of the busyness metric were included in the framework for spatial localisation. The plotted results indicate that the modular image difference metric has clearly identified the scene dependencies within this image data set. Scenes that suffered a large amount of quality loss are further spread along the x-axis and reach higher in terms of the image metric. In particular the use of the adapted busyness metric at the spatial localisation stage of the metric appears can be seen to be reflected in the results, with the busiest images, (Cliffs, flower Garden, and Stones) forming a group highest on the plot, indicating that they scored more highly in the image difference metric. Images of less susceptible scenes are clustered closer to the centre, indicating less loss in both objective and subjective ratings.

The results for individual images when plotted against compression ratio Figure 7.8 show a large vertical separation, which can be directly related to the busyness metric and possibly indicate that its effects need to be somehow attenuated. Of note is the fact that the image identified by the MIDM (see Figure 8.2) as having the highest value (i.e. the lowest in quality) was not one of the ones with very lowest subjective ratings, although it was one that was lower than average. This is likely to be due to the texture throughout the image having a masking effect for observers, which the MIDM did not model effectively. The texture dominates the entire image and does not dominate as a feature of interest. The objective metric however highlights it and this is likely to be because of its high level of detail meaning that it has a high busyness rating. In this case, busyness does not directly correlate with image quality. In its current form, the MDIM is too sensitive to busyness. The results also indicate that using the current structure it is not able to predict masking.

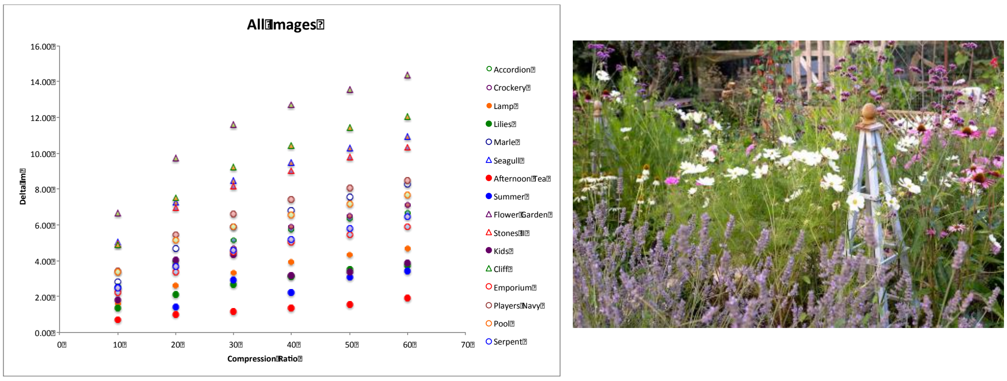


Figure 8.2 *Flower Garden* The modular image difference metric gives the highest value for this image, which does not correlate with the subjective results.

The results from the modular image difference metric indicate that it is monotonic with respect to subjective quality loss and is also able to distinguish between more or less susceptible scenes. The results illustrate the usefulness in tuning the metric with the addition of modules (or adaptation of existing modules) to highlight particular image attributes. In this case the spatial localisation module has been adapted using the busyness metric, which was found to provide good correlation with subjective results in previous chapters, to facilitate better prediction of quality results based on the busyness of images.

The filtering module used a relatively simple model for the achromatic CSF and was quite straightforward to implement. However, it is acknowledged that the CSF used, which was band pass in nature, was based upon CSF models for test targets, whereas the distortion metric is attempting to quantify suprathreshold detection in a complex image. Haun and Peli [190]suggest that the best CSF to use in image quality investigations might in fact be low pass rather than band pass (note that the chromatic CSFs used in chapter 7 were low pass). They suggest that when image quality is defined in terms of the errors between two images, the higher order statistics of images, which define the structures in the image (and relate to higher level cortical processing) 'swamp the contributions of lower-order statistics'. Recently, Triantaphillidou et al [191] described the measurement and modelling of more complex visual functions including the *Contextual Contrast Sensitivity Function* (cCSF); that is the contrast detection thresholds when measured within an image. Their results also suggest that the cCSF is lower than the *Isolated Contrast Sensitivity Function*, which is modelled from detection of single frequency targets against a plain background. Therefore the models used within this model may not be as suitable and it is an area that warrants further investigation.

The strength of the MIDM in terms of its adaptability is also one of its weaknesses. It is easy to adapt and expand and therefore can quickly become complex by the addition of extra modules. As it becomes more complex however, the interactions between different modules become harder to predict. Each stage requires adequate testing both individually and in context. Therefore as a metric, it has great potential, but it is not straightforward to implement and requires careful testing to define the optimum modules and parameters for a given image quality context. The model is presented as a general framework; it would be useful to test some different structures to find the optimum workflow/modular structure for different types of image quality applications (for example, to tune it for images with specific characteristics).

SSIM metrics have found widespread use over the last several years [176] [78] in part because they are so successful at predicting image quality. The results for all of the metrics for all of the images showed that they had a correlation with the subjective results, both in terms of prediction and ranking.

It is surprising however that the implementation of the MSSIM produced the least correlation and this warrants further investigation. The model was implemented using default settings and it might be that using a different scale might change the results. The scatter plot in Figure 7.11 indicates a wide spread of values. Further analysis is required to determine whether this is a result of scene dependency and if in fact the separation is a result of clustering of images according to their characteristics.

The Weighted Structural Similarity Index is simple to implement, but in the current implementation gives lower correlation than the un-weighted version. Elements of the busyness metric were included in both the MIDM and the Weighted Three Component SSIM. The SSIM metrics were all relatively straightforward to implement but more difficult to adapt.

It should be noted that these results were performed on the initial scene data before ruler calibration. Future work will include analysis of the objective results in terms of the calibrated rulers

The Sobel filter proved useful for a number of the segmentation tasks.

$$g_H = \begin{bmatrix} 1 & 2 & 1 \\ 0 & 0 & 0 \\ -1 & -2 & -1 \end{bmatrix} \quad g_V = \begin{bmatrix} 1 & 0 & -1 \\ 2 & 0 & -2 \\ 1 & 0 & -1 \end{bmatrix}$$

Figure 8.3 Filter kernels for the Sobel edge detection filter.

The filter is particular useful for edge detection because of its low susceptibility to noise. This is generally attributed to the fact that it is directional and while one direction is emphasising high frequencies by the application of positive and negative coefficients, there is an averaging process going on in the other direction in the same filter. The filters were useful when applied for edge enhancement in the spatial localisation module of the Modular Image Difference Model, but the response was not strong enough when used for feature extraction and therefore the results had to be dilated.

8.5 General Comments

Chandler et al [192] [59] describe 'Seven Challenges for Image Quality Research', in which they comment on the general shift in the focus of image quality research away from the objective of gaining better knowledge and more accurate models of human visual processing and more towards metrics that fit the ground truth data currently available. Some of the key issues they discuss include how to improve masking models for use on natural images; how to deal with suprathreshold distortion; how to model the effects of distortions on an image's appearance; and how to deal with multiple simultaneous distortions.

The metrics tested in the final part of this work indicate different approaches to the image quality problem and highlight some of these challenges. As an example, the busyness metric has proved very successful at identifying images with more or less busyness. Therefore in the first few stages of this research it was useful in predicting image quality, because it was found that busy areas in the images were particularly susceptible to the blurring errors from wavelets. This can be very problematic for image quality if textured areas are salient and the distortions are affecting a salient part of the image (for examples, see the *cliffs* image). But conversely, if the entire image is busy, then the details can act as a contrast mask, as was seen in the *Flower Garden* image. Fundamentally, the effects of texture within an image depend upon its context within the structure of the image, for example whether it is in salient area, whether it is an area of texture isolated from the background, or whether the texture is in fact the background. Therefore an approach whereby visual masking is considered only in terms of its effects upon the CSF does not give a complete picture of image quality. Chandler suggests that masking models require more research into 'the interplay between recognition and masking' [192], commenting that 'recognisability has been argued to influence detection thresholds', raising the threshold if the observer is unfamiliar with the content, i.e. it is not easily recognisable.

The results in subjective tests must also be considered in terms of these issues of masking. Feedback from the observers who took part in the investigations in chapters 5 and 6 indicated that once they became familiar with the parts of an image most affected by a particular distortion they found the image quality judgement easier and faster. The distortions became more recognisable and therefore their detection threshold went down.

Chandler's other questions are all highly relevant to this research. The results from the perceptibility and acceptability thresholds experiment indicate the effects of scene dependency upon thresholds.

Structural Similarity models do not consider the imaging chain or visual processing and are far more concerned with the structure within the image itself. They have proved very successful in predicting image quality. One of the problems however with a top down approach is that although it can be very successful in predicting image quality effects, the lack of system modelling makes it more difficult to identify the causes of those effects, and to adapt systems or system components as a result of them.

It would seem that a useful image quality approach would be to include aspects of both top down structural approach and the bottom up system modelling of more traditional signal processing approaches. Perhaps the modular image difference model or something similar can provide a framework for this.

9 Conclusions and Further Work

Image quality knowledge and understanding has developed a great deal since this research began, particularly in terms of our understanding of visual perception, but also in terms of the approaches found to be most successful for image quality metrics and for subjective scaling. Metrics capable of accurately predicting image quality continue to be the 'holy grail' of imaging research, as a result of which many hundreds of metrics and their variants have been developed and tested over the last decade. Subjective scaling methods have also developed, but the time required for psychophysical experiments cannot keep up with the speed of technological development. The development of databases of images with known subjective quality ratings have also been useful for testing image quality metrics.

9.1 Conclusions

This research has resulted in the following conclusions:

- The experiment described in chapter 4 investigated image quality of JPEG compared to JPEG 2000 and found that for most images JPEG 2000 was slightly preferred to JPEG although the difference in performance was rather small until compression ratios of >80:1, at which point JPEG 2000 was most commonly preferred. JPEG 2000 was found to be better in particular at compressing images containing text or periodic patterns than JPEG.
- Both algorithms were found to be scene dependent and certain types of scenes were more susceptible than others. Image quality was generally more robust across the compression range for images with low busyness and those that were predominantly lighter or darker in tone. In all three of the subjective studies the images that performed least well in terms of image quality were those containing areas of texture in or near a focal point of the image.
- Scene metrics are useful for grouping images and can be used to identify the characteristics of scenes that are more or less susceptible to perceived quality loss, either as a result of the interaction of the distortion with scene content, or due to masking effects. Scene groupings are algorithm specific and can be seen as a result of either scene susceptibility to distortions, or the characteristics of the distortions and

their visibility. The most useful scene metrics are likely to vary however depending upon the type of distortion(s).

- The types of distortions affecting image compression, particularly JPEG 2000 compressed images, are somewhat difficult to model as they are often combinations of more than one distortion (e.g. ringing,, blurring and colour distortions). However the use of scene metrics to evaluate the scenes that perform poorly in terms of image quality for a particular algorithm can be useful to identify susceptible scene characteristics and therefore some form of modelling of the *visibility characteristics* of the distortion.
- Perceptibility thresholds provide information about the visibility of distortions, therefore scenes with very low or high perceptibility thresholds can help to inform us about visual masking and interactions either visually, or structurally between an algorithm and specific types of scene content. This can be informative when designing metrics, allowing the prediction of the effects of an algorithm on a particular scene type and incorporating elements (such as the busyness metric) into the metric to correctly weight those aspects of image content.
- Acceptability thresholds are much more variable than perceptibility thresholds. Where images have a small difference between perceptibility and acceptability thresholds this indicates that the scene content results in distortions being highly visible. A larger difference between the two thresholds might indicate that although the distortion is visible it is not affecting visually important image areas. The variability may also be as a result of observer preference and quality criteria within the given imaging context.
- The Soft Copy quality ruler is a relatively new approach to subjective quality scaling, which once set up, provides a method to extract quality JNDs directly from the results of a psychophysical experiment without lengthy analysis. The SCQR provides results that are consistent with those from the thresholds experiment in terms of image groupings.
- The SCQR is usefully able to identify both scene susceptibility and observer sensitivity in an image quality study.
- The approach used in the SCQR to model the system MTF to the aim MTF and subsequently to develop the JND filters was successful in producing the ruler images. Care must be taken however at image capture and during processing to try to

minimise non-linearities where possible because of the limitations imposed by the shape and range of the function used to model the JND filters.

- The SCQR can be calibrated using the average scene, but this reduces scene variability and therefore may not be useful to accurately predict quality with respect to scene dependence.
- The metrics used in chapter 7 all had moderate to very good correlation with the subjective data from the Soft Copy Quality Ruler in terms of prediction and ranking of image quality. The MSSIM performed the worst of the four and the SSIM was the best. However the analysis was carried out over all scenes and it would be useful to evaluate the results in terms of scene clusters derived from earlier psychophysical tests.
- The modular image difference framework is adaptable and allows the addition and testing of further modules. In this implementation, aspects of the busyness metric were used to filter the images. However the results indicated that the metric was oversensitive to busyness and did not model masking effects.
- The WSSIM used the busyness metric to segment the image in to textured areas versus edges and uniform areas. Values for the weightings were determined by a process of trial and error. These weightings can be tuned to the particular image content.

9.2 Recommendations for Future Work

- Create further sets of Soft Copy Quality Rulers from the remaining images in the set used in chapter 5. Test the existing SCQR images against the Digital Reference Set to check the efficacy of the calibration thus performed. Once calibrated the rulers can be made public as a resource for other researchers.
- Investigate improvements to the implementation of the SCQR when creating the JND filters, for example using a windowing function.
- Make the sample set of reference images available as a database, together with information about perceptibility and acceptability thresholds and scaled to the SCQR.
- Develop some practical guidelines to support others in producing their own SCQR images using the method developed in chapter 5.
- Further analyse the results from the metrics by scene to investigate which of the metrics are best at highlighting scene dependencies.

- Explore alternative scene metrics or combinations of metrics for different scene types. Provide scene metric values with the images developed for the SCQR.
- ‘Tune’ the different modules used in the MIDM to provide less emphasis on busyness.

10 Related Work

10.1 List of publications (Primary Author)

Allen, Elizabeth, Triantaphillidou, Sophie and Jacobson, Ralph E. (2007) *Image quality comparison between JPEG and JPEG2000. I. Psychophysical investigation.* Journal of Imaging Science and Technology, 51 (3). pp. 248-258. ISSN 1062-3701

Allen, E., Triantaphillidou, S. and Jacobson, R. (2014) *Perceptibility and acceptability of JPEG 2000 compressed images of various scene types*. In: Proceedings of SPIE Electronic Imaging: Image Quality and System Performance XI, Jan 2014, San Francisco, USA.

10.2 Presentations at Conferences and Symposia

Allen, E. (presenter), Triantaphillidou, S., Jacobson, R., Attridge, G.G. *Image quality comparison between JPEG and JPEG2000. Part I. Subjective Evaluation*, at: Digital Futures Conference 2003, Royal Photographic Society, 14 October 2003, National Physical Laboratory, Teddington, Middlesex, UK

Allen, E. *An investigation into perceptual thresholds of distortion detectability for JPEG 2000 encoded images,* at: Digital Futures Conference 2004, Royal Photographic Society, 26 October 2004, National Physical Laboratory, Teddington, Middlesex, UK

Allen, E. *JPEG and JPEG 2000 compression* at: Good Picture 2004 - Digital Demystified, Royal Photographic Society Symposium, 15 December 2004, University of Westminster, Regents Street Campus, London, UK

Allen, E. *Image Processing from Capture to Output* at: Good Picture 2006- Management and Manipulation, Royal Photographic Society Symposium, December 2006, University of Westminster, Regents Street Campus, London, UK

Allen, E. (author), Triantaphilidou, S. (presenter), Jacobson, R., *Perceptibility and acceptability of JPEG 2000 compressed images.* At: SPIE Electronic Imaging: Image Quality and System Performance XI, Jan 2014, San Francisco, USA

10.3 Awards

Selwyn Award 2005, Royal Photographic Society Imaging Science Group. Presented at the Royal Society, London, UK

10.4Related Publications (not primary author)

Triantaphillidou, Sophie, Allen, Elizabeth and Jacobson, Ralph E. (2007) *Image quality comparison between JPEG and JPEG2000. II. Scene dependency, scene analysis, and classification*. Journal of Imaging Science and Technology, 51 (3). pp. 259-270. ISSN 1062-3701

Mancusi, Francesco, Triantaphillidou, Sophie and Allen, Elizabeth (2010) *Multidimensional image selection and classification system based on visual feature extraction and scaling.* In: Image quality and system performance VII : 18-19 January 2010, San Jose, California, United States. Proceedings of SPIE (7529). SPIE, Bellingham, Wash., A1-A11. ISBN 9780819479228

Orfanidou, Maria, Triantaphillidou, Sophie and Allen, Elizabeth (2008) *Predicting image quality using a modular image difference model.* In: Image quality and system performance V : 28-30 January 2008, San Jose, California, USA. Proceedings of SPIE (6808). IS&T - The Society for Imaging and Science and Technology and SPIE, F1-F12. ISBN 9780819469809

References

[1] Junichi Nakamura, *Image Sensors and Signal Processing for Digital Still Cameras*, Junichi Nakamura, Ed. Boca Raton, Florida: CRC Press, Taylor and Francis Group, 2006.

[2] Canon UK. http://www.canon.co.uk/for_home/product_finder/cameras/digital_slr/eos_5d_mark_ii/#p-specification21. [Online]. http://www.canon.co.uk/for_home/product_finder/cameras/digital_slr/eos_5d_mark_ii/#p-specification21

[3] S Triantaphillidou, "Digital Colour Reproduction," in *The Manual of Photography*, 10th ed., E J Allen and S Triantaphillidou, Eds. Oxford, UK: Focal Press, an Imprint of Elsevier Ltd, 2011, p. 411.

[4] R C Gonzales and R E Woods, *Digital Image Processing*, 2nd ed., R C Gonzales and R E Woods, Eds. Boston, USA: Addison-Wesley Longman Publishing Co, 2001.

[5] E J Allen, "Chapter 29: Image Compression," in *The Manual of Photography*, E J Allen and S Triantaphillidou, Eds. Oxford, UK: Focal Press, 2011, pp. 536-537.

[6] A C Bovik, *Handbook of Image and Video Processing*, 2nd ed. Burlington, USA: Elsevier Academic Press, 2005.

[7] K Sayood, *Introduction to Data Compression*, K Sayood, Ed. San Francisco, CA/USA: Morgan Kauffman, 2006.

[8] Joint Photographic Experts Group. http://www.jpeg.org/about.html. [Online]. http://www.jpeg.org/about.html

[9] G K Wallace, "The JPEG Still Picture Compression Standard," *IEEE Transactions on Consumer Electronics*, vol. 38, no. 1, pp. xviii - xxxiv, February 1992.

[10] E J Allen, "Chapter 17: Digital Image File Formats," in *The Manual of Photography*, E J Allen and S Triantaphillidou, Eds. Oxford, UK: Focal Press, 2011, p. 319.

[11] M Rabbani and R Joshi, "An Overview of the JPEG 2000 Still Image Compression Standard," *Signal Processing: Image Communication* , vol. 17, pp. 3-48, 2002.

[12] P Burns, S Houchin, K Parulski, and M Rabbani, "Using JPEG 2000 in Future Digital Cameras: Advantages and Challenges," in *Proceedings of International Congress of Imaging Science*, Tokyo, 2002, pp. 371-373.

[13] Q Ghulam, X Zhao, and A T Ho, "Estimating JPEG2000 Compression for Image Forensics Using the Benford's Law," in *Proc. SPIE 7723, Optics, Photonics, and Digital Technologies for Multimedia Applications*, vol. 7723, 2010, pp. 77230J-1-77230J-10.

[14] A Ford, PhD Thesis: Relationships between Image Quality and Still Image Compression, 1997.

[15] International Standards Organisation, Photography - Psychophysical experimental methods for estimating image quality - Part 3, Quality Ruler Method, 2005.

[16] B Keelan, *Handbook of Image Quality*, B Keelan, Ed. New York, USA: Marcel Dekker Inc, 2002.

[17] R Jacobson, "An Evaluation of Image Quality Metrics," *Journal of Photographic Science*, vol. 43, 1995.

[18] P G Engeldrum, *Psychometric Scaling: A toolkit for Imaging Systems Development*. Winchester: Imcotek Press, 2000.

[19] A Ford, "Determination of Compressed Image Quality," in *Colour Imaging: Vision and Technology*, L W MacDonald and M R Luo, Eds.: John Wiley and Sons Ltd, 1999, pp. 315-337.

[20] Z Wang and A Bovic, "Why is Image Quality Assessment So Difficult?," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2002.

[21] T Leisti, J Radun, T Virtanen, R Halonen, and J Nyman, "Subjective Experience of Image Quality: Attributes, Definitions and Decision Making of Subjective Image Quality," in *Image Quality and System Performance IV Proceedings of SPIE-IS & T Electronic Imaging*, 2009.

[22] R E Jacobson and S Triantaphillidou, "Metric Approaches to Image Quality," in *Colour Image Science*, L W MacDonald and M R Luo, Eds.: John Wiley and Sons Ltd, 2002, pp. 371-391.

[23] A J Ahumada and C H Null, "Image Quality: A Multidimensional Problem," in *Digital Images and Human Vision*, A B Watson, Ed. Cambridge, Massachusetts, USA: MIT Press, 1993, pp. 140-148.

[24] P G Engeldrum, "A Short Image Quality Model Taxonomy," *Journal of Imaging Science and Technology*, vol. 48, no. 2, pp. 160-164, March/April 2004.

[25] T Janssen and F Blommaert, "The Semantics of Image Quality," in *The Fifth Color Imaging Conference: Color Science, Systems and Applications*, Arizona, 1997, pp. 126-130.

[26] S Yendrikhovski, "Image Quality and Color Categorisation," in *Color Image Science*, L W MacDonald and M R Luo, Eds.: John Wiley and Sons Ltd, 2002, pp. 393-419.

[27] C J Bartleson, "Memory Colors of Familiar Objects," *Journal of the Optical Society of America*, vol. 50, pp. 73-77, 1960.

[28] R W G Hunt, I T Pitt, and L M Winter, "The Preferred Reproduction of Blue Sky, Green Grass and Caucasian Skin in Color Photography," *Journal of Photographic Science*, vol. 50, pp. 144-150,

1974.

[29] S M Newhall, R W Burnham, and J R Clark, "Comparison of Successive with Simultaneous Colour Matching," *Journal of the Optical Society of America*, vol. 47, pp. 43-56, 1957.

[30] S Hochstein and M Ahissar, "View from the Top: Hierarchies and Reverse Hierarchies in the Visual System," *Neuron*, 2002.

[31] S Yendrikhovskij, F Blommaert, and H de Ridder, "Towards Perceptually Optimal Colour Reproduction of Natural Scenes," in *Colour Imaging: Vision and Technology*, L W MacDonald and M R Luo, Eds.: John Wiley and Sons Ltd, 1999, pp. 363-382.

[32] P G Engeldrum, "Image Quality Modelling: Where Are We?," in *PICS*, pp. 251-255.

[33] E J Allen, S Triantaphillidou, and R E Jacobson, "Subjective Quality Comparison Between JPEG and JPEG 2000: Psychophysical Investigation," *Journal of Imaging Science and Technology*, vol. 51, no. 3, pp. 248-258, May-June 2007.

[34] S Triantaphilldou, E J Allen, and R E Jacobson, "Subjective Quality Comparison Between JPEG and JPEG 2000: Scene Classification," *Journal of Imaging Science and Technology*, vol. 51, no. 3, pp. 259-271, May-June 2007.

[35] O Hoon, S Trinataphillidou, and R E Jacobson, "Scene Classification with Respect to Image Quality Measurements," in *Proceedings SPIE 7529 752908*, 2010.

[36] T N Pappas, R J Safranek, and J Chen, "Perceptual Criteria for Image Quality Evaluation," in *Handbook of Image and Video Processing*, A Bovic, Ed.: Elsevier Academic Press, pp. 939-957.

[37] L L Thurstone, "A Law of Comparative Judgement ," *Psychological Review (reprinted from 1927)*, vol. 101, no. 2, pp. 266-270, 1994, This is a reprint of the original publication in Psychological Review in 1927.

[38] E W Jin, B W Keelan, J Chen, J B Phillips, and Y Chen, "Softcopy Quality ruler Method: Implementation and Validation," in *Image Quality and System Performance VI, Proceedings SPIE-IS&T Electronic Imaging SPIE* , vol. 7242.

[39] Technical Committee 42, International Organisation for Standardisation, Photography - Psychophysical Experimental Methods for Estimating Image Quality - Part 3: Quality Ruler Method, 2005.

[40] S Triantaphillidou, "Chapter 19: Image Quality," in *Manual of Photography 10th Edition*, E J Allen and S Triantaphillidou, Eds. Oxford, UK: Focal Press, 2010.

[41] D A Silverstein and J E Farrell, "The Relationship Between Image Fidelity and Image Quality," in *International Conference on Image Processing*, Lausanne, 1996, pp. 881-884.

[42] M P Eckhert and A P Bradley, "Perceptual Quality Metrics Applied to Still Image Compression,"

Signal Processing: Special Issue on Image and Video Quality Metrics, vol. 70, no. 3, pp. 177-200, 1998.

[43] J Egglestone, *Sensitometry for Photographers*. London, UK: Focal Press, 1984.

[44] G G Attridge , *The Manual of Photography, Ninth Edition*. London, UK: Focal Press, 2000, pp. 222-228.

[45] G G Attridge, "Chapter 8: Sensitometry," in *The Manual of Photography, tenth edition*, E J Allen and S Triantaphillidou, Eds.: Focal Press, 2010.

[46] S Triantaphillidou, "Chapter 21: Tone Reproduction," in *The Manual of Photography, tenth edition*, E J Allen and S Triantaphillidou, Eds. UK: Focal Press, 2010.

[47] S Triantaphillidou, Aspects of Image Quality in the Digitisation of Photographic Collections, PhD Thesis, 1999.

[48] C Poynton and TO BE COMPLETED,.

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

[50] M D Fairchild, *Color Appearance Models*. Chichester, England: John Wiley and Sons Ltd, 2004.

[51] S Triantaphillidou, "Chapter 5: Colour Science," in *The Manual of Photography, tenth edition*, E J Allen and S Triantaphillidou, Eds. Oxford, UK: Focal Press, 2010.

[52] R W G Hunt, *Measuring Colour*, 3rd ed. UK: Fountain Press Ltd, 1998.

[53] Touradj Ebrahimi, P Schelkens, A Skodras, and T Ebrahimi. (2009) The JPEG 2000 suite.

[54] G Sharma, W Wu, and E N Dalal, "The CIEDE2000 Color-Difference Formula: Implementation Notes, Supplementary test Data, and Mathematical Observations," *Color Research and Application*, vol. 30, no. 1, February 2005.

[55] R Jenkin, R E Jacobson, and K Maclennan-Brown, "Determination of the MTF of JPEG Compressed Images Using the ISO 12233 Spatial Frequency Response Plugin," in *Imaging Science and Technology PICS conference*, 2000.

[56] P Burns, Evaluating Digital Scanner and Camera Imaging Performance for Digital Collections, , 2006.

[57] R Jenkin, "Chapter 7: Images and Image Formation," in *Manual of Photography*, 10th ed., E J Allen and S Triantaphillidou, Eds. UK: Focal Press, 2010.

[58] R Jenkin, "Chapter 24: Resolution, Sharpness and Noise," in *The Manual of Photography*, 10th ed., E J Allen and S Triantaphillidou, Eds. UK: Focal Press, 2010.

[59] D M Chandler. (2013) Review Article: Seven Challenges in Image Quality Assessment: Past, Present and Future Research.

[60] G Ciocca, S Corchs, F Gasparini, and R Schettini. (2014) How to assess image quality within a workflow chain: an overview.

[61] T N Pappas, R J Safranek, and J Chen, "Perceptual Criteria for Image Quality Evaluation," in *Hanbood of Image and Video Processing*, A Bovik, Ed.: Elsevier Academic Press, 2005, pp. 939-959.

[62] M D Fairchild, "1: Human Color Vision," in *Color Appearance Models*, 3rd ed.: John Wiley and Sons Ltd, 2013.

[63] R J Jenkin, "Chapter 4: The Human Visual System," in *The Manual of Photography*, 10th ed., E J Allen and S Triantaphillidou, Eds. Oxford: Focal Press, an imprint of Elsevier Ltd, 2011, pp. 59-76.

[64] G M Johnson and M D Fairchild, "On Contrast Sensitivity in an Image Difference Model," in *Proceedings of the I S & T PICS Conference*, 2001, pp. 18-28.

[65] P G J Barten, *Contrast Sensitivity of the Human Eye and Its Effect on Image Quality*. Bellingham: SPIE, 1999.

[66] S J Daly, "Visible differences predictor: an algorithm for the assessment of image fidelity.," in *Proc. SPIE 1666, Human Vision, Visual Processing, and Digital Display III,* 1992, Available from the SPIE digital library.

[67] T Movshon and L Kiorpes, "Analysis of the development of spatial sensitivity in monket and human infants," *JOSA*, vol. 5, 1998.

[68] G M Johnson and M D Fairchild, "Measuring Images: Differences, Quality and Appearance," in *SPIE/IS&T Electronic Imaging Conference*, 2003.

[69] G M Johnson and M D Fairchild, "Darwinism of Color Image Difference Models," in *Proc. of the IS&T/SID 9th Color Imaging Conference*, 2001.

[70] M D Fairchild, "Chapter 20: Image Appearance Modelling and the Future," in *Colour Appearance Models*, 3rd ed.: John Wiley and Sons Ltd, 2013.

[71] F W Campbell and J G Robson, "Application of Fourier Analysis to the Visibilty of Gratings," *Journal of Phsyiology*, vol. 197, pp. 551-566, 1968.

[72] R E Jacobson, Image Quality Measurements, Necessity, Numbers and 'Nesses', November 3, 2009.

[73] H R Sheik and A Bovik, "Chapter 8.4: Information Theoretic Approaches to Image Quality Assessment," in *Handbook of Image and Video Processing*, A Bovik, Ed. Orlando, Florida, USA: Elsvier Academic Press Inc, 2005.

[74] J Bartleson, "The Combined Influence of Sharpness and Graininess on the Quality of Colour Prints," *Journal of Photographic Science*, vol. 30, pp. 33-38, 1982.

[75] H Pappas, R J Saffranek, and J Chen, "Chapter 8.2: Perceptual Criteria for Image Quality Evaluation," in *Handbook of Image and Video Processing*, A Bovik, Ed. Orlando, Florida, USA: Elsevier Academic Press Inc, 2005.

[76] S Daly, "The Visible Differences Predictor: An Algorithm for the Assessment of Image Fidelity," in *Digital Images and Human Vision*, A B Watson, Ed. Cambridge, Massachussets, USA: MIT Press, pp. 179-206.

[77] X Zhang, D Silverstein, J E Farrell, and B A Wandell, "Color Image Quality Metric S-CIELAB and its Application on Halftone Texture Visibility," *COMPCON97 Digest of Papers*, pp. 44-48, 1997.

[78] H R Sheikh and A Bovik, "Chapter 8.3: Structural Approaches to Image Quality Assessment," in *Handbook of Image and Video Processing*, Bovik A, Ed. Orlando, Florida, USA: Elsevier Academic Press Inc, 2005.

[79] C Li and A Bovik, "Three-Component Weighted Structural Similarity Index," in *SPIE/IS&T Electronic Imaging*, vol. 7242, San Jose, USA, 2009, pp. 72420Q1-Q9.

[80] H R Sheikh and A Bovik, "Image Information and Visual Quality," *IEEE Transactions on Image Processing*, vol. 15, no. 2, pp. 430-444, 2006.

[81] P F Sharp, "Quantifying Image Quality," *Clinical Physics and Physiological Measurement*, vol. 11, no. Supplment A, pp. 21-26, 1990.

[82] S S Stevens, "On the Theory of Scales of Measurement," *Science*, vol. 103, no. 2684, June 1946.

[83] R M Boynton, "Chapter 6," in *Optical Radiation Measurements, Volume 5, Visual Measurements, Part II*. New York, USA: Academic Press, 1984, vol. 5, pp. 335-366.

[84] International Standards Organisation, Photography - Psychophysical experimental methods for estimating image quality- Part 1: Overview of psychophysical methods, 2005.

[85] International Standards Organisation, Photography - Psychophysical experimental methods for estimating image quality - Part 3: Quality ruler method, 2012.

[86] International Standards Organisation, Photography - Psychophysical experimental methods for estimating image quality- Part 2: Triplet Comparison Method.

[87] G A Gescheider, "Chapter 2: The Classical Psychophysical Methods," in *Psychophysics Method and Theory*. Hillsdale, New Jersey, USA: Lawrence Erlbaum Associates, John Wiley and Sons, 1976, pp. 20-38.

[88] E W Jin, B W Keelan, J Chen, and J B Phillips, "Softcopy quality ruler method: implementation and validation," in *Proceedings SPIE 7242 Image Quality and System Performance IV*, 2009.

[89] J Young-Park, Evaluation of changes in image appearance with changes in displayed image size; PhD Thesis, 2013.

[90] C E Shannon, "A Mathematical Theory of Communication," *The Bell System Technical Journal*, vol. 27, October 1948.

[91] N Tavakoli, "Short communication: Entropy and Image Compression," *Journal of Visual Communication and Image Representation*, vol. 4, no. 3, pp. 271-278, September 1993.

[92] D Taubman, JPEG 2000 Compression Standard for Interactive Imaging, March 27, 2003, accessed from http://maestro.ee.unsw.edu.au/~taubman/seminars_files/IEEE_IEA_J2K.pdf on 17/08/2015 10.32am.

[93] A Skodras, C Christopoulos , and T Ebrahimi, "The JPEG 2000 Still Image Compression Standard," *IEEE Signal Processing Magazine*, pp. 37-58, September 2001.

[94] G K Wallace, The JPEG still picture compression standard, 1991.

[95] International Standards Organisation, ISO/IEC 15444-1, December 15, 2000.

[96] S Lawson and J Zhu, "Image compression using wavelets and JPEG 2000: a tutorial," *Electronics and Communication Engineering Journal*, pp. 112-121, June 2002.

[97] Z Wang and A Bovik, "1.2 What's wrong with MSE?," in *Modern Image Quality Assessment*.: Morgan and Claypool, 2006.

[98] B Keelan, "Chapter 17: Attribute Interaction Terms in Objective Metrics," in *Handbook of Image Quality: Characterisation and Prediction*. New York, USA: Marcel-Dekker Inc, 2002.

[99] B Keelan, "Chapter 10, Scene and Observer Variability," in *Handbook of Image Quality*. New York, USA: Marcel-Dekker Inc., 2002.

[100] B Keelan, "Chapter 20: Preference in Colour and Tone Reproduction," in *Handbook of Image Quality: Characterisation and Prediction*, Inc Marcel-Dekker, Ed. New York, 2002.

[101] B Keelan, "Chapter 15: Weighting Attributes that Vary Across an Image," in *Handbook of Image Quality: Characterisation and Prediction*. New York, USA: Marcel-Dekker Inc, 2002.

[102] C J Bartleson, *Optical Radiation Measurements Volume 5*.: Academic Press, 1984, vol. 5.

[103] B Keelan, "Chapter 1: Can Image Quality be Usefully Quantified?," in *Handbook of Image Quality: Characterisation and Prediction*. New York: Marcel-Dekker Inc, 2002.

[104] B Keelan and E W Jin, "Weighting of Field Heights for Sharpness and Noisiness," in *Proceeedings SPIE Image Quality and System Performance VI 2009*, vol. 7242, 2009.

[105] T Kadir and M Brady, "Saliency, Scale and Visual Description," *International Journal of Computer*

Vision, vol. 45, no. 2, pp. 83-105.

[106] U Neisser, "Visual Search," *Scientific American*, vol. 210, no. 6, pp. 94-102, 1964.

[107] Z Wang and A Bovik, "2.2.2.4 Foveated Vision," in *Modern Image Quality Assessment: Synthesis Lectures on Image, Video and Multimedia Processing*.: Morgan & Claypool Publishers, 2006, pp. 25-27.

[108] M D Fairchild, "Human Color Vision," in *Color Appearance Models*.: Addison-Wesley, 1998, p. 7.

[109] L Ciocca, C Cusano, R Schettini, and C Brambilla, "Semantic labelling of digital photos by classification," in *Proceedings of SPIE - the International Society for Optical Engineering*, 2003.

[110] G Ciocca, C Cusano, and R Schettini, "Semantic Classification, Low Level Features, and Relevance Feedback for Content-based Image Retrieval Systems," in *Proceedings SPIE 7255: Multimedia Content Access: Algorithms and Systems III*, San Jose, 2009.

[111] F Mancusi, S Triantaphillidou, and E Allen, "Multidimensional image selection and classification system based on visual feature extraction and scaling," in *Proc. SPIE 7529, Image Quality and System Performance VII, 75290A* , vol. 7529, 201.

[112] O Hoon, S Triantaphillidou, and R E Jacobson, "Scene Classification with Respect to Image Quality Measurements," in *Proceedings SPIE 7529 752908*, 2010.

[113] D Hasler and S Susstrunk, "Measuring Colourfulness in Natural Images," in *IS&T/SPIE Electronic Imaging*, 2003, pp. 87-95.

[114] Kyung-Woo Ko, Tae-Yong Park, and Yeong-Ho Ha, "Analysis of Relationship between Image Compression and Gamut Variation: JPEG and JPEG2000," *Journal of Imaging Science and Technology*, vol. 53, no. 6, pp. 060402(5) -060402(12), 2009.

[115] J L Mitchell and W B Pennebaker, "Evolving JPEG Color Data Compression Standards," *Standards for Electronic Imaging Systems, Critical Reviews*, vol. CR37, pp. 68-95.

[116] J E Adams, Jr. and J F Hamilton, Jr., "Chapter 3: Digital Camera Image Processing Chain Design," in *Single Sensor Imaging: Methods and Applications for Digital Cameras*. Boca Raton, Florida, USA: CRC Press, Taylor and Francis Gp, 2009, p. 69.

[117] D Santa-Cruz, R Grosbois, and T Ebrahimi, "JPEG 2000 Performance evaluation and assessment," *Signal Processing: Image Communication*, vol. 17, pp. 113-130, 2002.

[118] D Santa-Cruz and C Ebrahimi, "A study of JPEG 2000 still image coding versus other standards," in *Proceedings of the Tenth European Signal Processing Conference*, vol. 2, 2001, pp. 673-676.

[119] R Jacobson, "An Evaluation of Image Quality Metrics," *Journal of Photographic Science*, vol. 43, no. 1, pp. 7-16, 1995.

[120] U Steingrimsson and K Simon, "Perceptive Quality Estimations: JPEG 2000 vs JPEG," *Journal of Imaging Science and Technology*, vol. 47, no. 6, pp. 572-585, 2003.

[121] J C Russ, *The Image Processing Handbook*, 2nd ed.: CRC Press Inc., 1995, pp. 130-132.

[122] S G Mallat, "A Theory for Multiresolution Signal Decomposition: The Wavelet Representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence* , vol. 11, no. 7, 1989.

[123] C Edwards, "Wavelet Analysis Transforms Data Processing," *Scientific Computing World*, June 1996.

[124] B A Cipra, "Wavelet applications come to the fore," *SIAM News*, November 1993.

[125] G Strang, "Wavelet transforms versus Fourier transforms," *Bulletin of the American Mathematical Society*, vol. 28, no. 2, pp. 288-305, April 1993.

[126] K Sayood, "13.6 Application to Image Compression - JPEG," in *Introduction to Data Compression*. San Francisco, USA: Morgan-Kauffman Publishers Inc, 1996, p. 416.

[127] D Buckley, "Color Imaging with JPEG 2000," in *IS&T/SID Proceedings of the Ninth Color Imaging Conference*, 2001, pp. 113-119.

[128] R Jenkin, "Chapter 28: Digital image processing in the frequency domain," in *The Manual of Photography*, 10th ed., E Allen and S Triantaphillidou, Eds. Oxford, UK: Focal Press, an imprint of Elsevier, 2011, p. 524.

[129] R C Gonzales and R E Woods, "4.3 Smoothing Frequency Domain Filters," in *Digital Image Processing*. Boston, USA: Addison-Wesley Longman Publishing Co, 2001, p. 171.

[130] B W Keelan, "18.3 Reconstruction Artifacts," in *Handbook of Image Quality: Characterisation and Prediction*. New York, USA: Marcel-Dekker, Inc, 2002, pp. 258-259.

[131] P G Engeldrum, "Chapter 8: Indirect Interval Scaling - Case V and Paired Comparison," in *Psychometric Scaling: A Toolkit for Imaging Systems Development*. Winchester, MA, USA: Imcotek Press, 2000, pp. 93-108.

[132] WinSoftMagic Development. Software: Advanced JPEG Compressor™.

[133] Algo Vision Luratech GmbH. Lurawave Smart Compress 3.0.

[134] M Anderson, R Motta, S Chandrasekar, and M Stokes, "Proposal for a Standard Default Color Space for the Internet - sRGB," in *IS&T/SID Proceedings, Fourth Color Imaging Conference*, 1996, pp. 127-134.

[135] S Susstrunk, R Buckley, and S Swen, "Standard RGB Color Spaces," in *IS&T/SID Proceedings Seventh Color Imaging Conference*, 1999, pp. 127-134.

[136] E Bilissi, R E Jacobson, and G G Attridge, "Perceptibility and Acceptability of Gamma Differences of Displayed sRGB images," in *PICS 2003*, Rochester, NY, 2003, pp. 120-125.

[137] BSi British Standards, Multimedia Systems and Equipment - Colour Measurement and Management - Part 2-1: Colour Management-Default RGB Colour Space-sRGB, accessed from www. bsol-bsigroup-com.

[138] E Bilissi, Interface for generating colour patches for display characterisation, RGB. exe.

[139] E Bilissi, Paired Comparison Software written in Visual Basic 6.

[140] Mathworks, MATLAB - The Language of Technical Computing, Version 6.1.

[141] B Keelan, "Chapter 2: The Probabilistic Nature of Perception," in *The Handbook of Image Quality: Characterization and Prediction*. New York: Marcel-Dekker Inc., 2002, p. 26.

[142] P Engeldrum, "Chapter 9: Indirect Interval Scaling - Generalisations of Thurstone's Case V," in *Psychometric Scaling: A Toolkit for Imaging Systems Development*. Winchester, MA, USA: Imcotek Press, 2000, p. 115.

[143] G E Noether, "Remarks about a Paired Comparison Model," *Psychometrika*, vol. 25, p. 357, 1960.

[144] E M Biederman, *Photographic Korrespondent*, vol. 25 and 41, p. 103, 1967.

[145] S A Klein, "Image quality and image compression, a psychophysicist's viewpoint," in *Digital Images and Human Vision*, A Watson, Ed., 1997.

[146] P Engeldrum, "Chapter 5: Thresholds and Just Noticeable Differences," in *Psychometric Scaling: A Toolkit for Imaging Systems Development*. Winchester, MA, USA: Imcotek Press, 2000, p. 56.

[147] K Kim Joong et al., "Objective index of image fidelity for JPEG2000 compressed body CT images," *Medical Physics*, vol. 36, no. 3218, 2009.

[148] F Guarneri, M Vaccaro, C Guarneri, and T Cannavo, "JPEG vs JPEG 2000: benchmarking with dermatological images," *Skin Research and Technology, early view, online version*, 2013.

[149] E J Allen, S Triantaphillidou, and R E Jacobson, "Perceptibility and Acceptability of JPEG 2000 compressed images of various scene types," in *Proceedings of the SPIE/IS&T Electronic Imaging: Image Quality and System Performance XI*, vol. 90160W, 2014.

[150] J Schewe and B Fraser, "From RAW to Colour; Chapter 2: How Camera RAW works," in *Real World Camera RAW using Adobe Photoshop CS5*. Berkeley, California, USA: Peachpit Press, Pearson Education, in association with Adobe Press, 2011.

[151] S Susstrunk, "Standard RGB Color Spaces," in *Seventh Color Imaging Conference: Color Science, Systems and Applications*, vol. 7, Arizona, 1999, pp. 127-134.

[152] International Standards Organisation, Photography — Electronic still-picture cameras — Methods for measuring optoelectronic conversion functions (OECFs), 2009.

[153] BSi British Standards, Multimedia Systems and Equipment - Colour Measurement and Management - Part 4: Equipment Using Liquid Crystal Display Panels.

[154] W K Pratt, "Part 5: Image Analysis: Histogram Amplitude Features," in *Digital Image Processing: PIKS Scientific Inside*, 4th ed.: Interscience, 2007, p. 539.

[155] Mathworks, MATLAB - The Language of Technical Computing, Version 2011a.

[156] E J Allen, Thresholds of Perceptibility in JPEG 2000, 2004, Presentation available from the author allene@wmin.ac.uk.

[157] International Standards Organisation, Graphic Technology and Photography - Viewing Conditions.

[158] P G Engeldrum, "5.6.3 Psychometric Models: Basic Data Analysis," in *Psychometric Scaling: a Toolkit for Imaging Systems Development*. winchester, MA, USA: Imcotek Press , 2000, pp. 64-68.

[159] N Prins and F Kingdom, Palamedes Toolbox version 1.8.0 for MATLAB, downloaded from http://www.palamedestoolbox.org/index.html.

[160] F Kingdom and N Prins, *Psychophysics: A Practical Introduction*.: Elsevier Ltd, 2010.

[161] Royal Geographic Society. www.rgs.org. [Online]. http://www.rgs.org/NR/rdonlyres/4844E3AB-B36D-4B14-8A20-3A3C28FAC087/0/OASpearmansRankExcelGuidePDF.pdf

[162] E Allen, "Chapter 27: Spatial image processing," in *The Manual of Photography*, 10th ed., E Allen and S Triantaphillidou, Eds. Oxford, UK: Focal Press, an imprint of Elsevier, 2011, p. 513.

[163] International Standards Organisation, Photography - Electronic Still Picture Imaging - Resolution and Spatial Frequency Responses.

[164] Imatest website. [Online]. http://www.imatest.com

[165] Imatest. Imatest.com. [Online]. http://www.imatest.com/docs/sfrplus_instructions/#distance

[166] C Poynton, ""Gamma" and its Disguises: The Nonlinear Mappings of Intensity in Perception, CRTs, Film, and Video," *SMTPE Journal*, December 1993.

[167] International Telecommunication Union, Recommendation ITU-R BT.709-6 , June 2015, downloaded from https://www.itu.int/dms_pubrec/itu-r/rec/bt/R-REC-BT.709-6-201506-I!PDF-E.pdf on 17/08/2015.

[168] National Institute of Standards and Technology, US Department of Commerce. (2015, Aug.) http://www.itl.nist.gov/div898/software/dataplot/refman2/ch2/weightsd.pdf. [Online].

http://www.itl.nist.gov/div898/software/dataplot/refman2/ch2/weightsd.pdf

[169] Imatest. Imatest.com. [Online]. http://www.imatest.com/docs/mtf_appearance/

[170] A Floren and A C Bovik, "Chapter 14: Foveated Image and Video Processing and Search," in *Academic Press Library in Signal Processing*, R Chellappa and S Theodoridis, Eds.: Elsevier Ltd, 2014, vol. 4, p. section 4.14.3.7.

[171] R C Gonzales and R E Woods, "Chapter 4: Image Enhancement in the Frequency Domain," in *Digital Image Processing*, 2nd ed. New Jersey, USA: Prentice-Hall, 2002, pp. 199-205.

[172] R C Gonzales, R E Woods, and S L Eddins, "4.3 Filtering in the Frequency Domain," in *Digital Image Processing using MATLAB*, 1st ed. New Jersey, USA: Pearson Prentice Hall, 2004, p. 117, function paddedsize.m.

[173] M Orfanidou, S Triantaphillidou, and E Allen, "Predicting Image Quality Using a Modular Image Difference Model," in *Image Quality and System Performance V. SPIE Proceedings*, vol. 6808, San Jose, USA, 2008, pp. 68080F-68080-12.

[174] Z Wang, A Bovic, H R Sheikh, and E P Simoncelli, "Image Quality Assessment: From error Visibility to Structural Similarity," *IEEE TRANSACTIONS ON IMAGE PROCESSING*, vol. 13, no. 4, April 2004.

[175] Z Wang, E P Simoncelli, and A Bovic, "MULTI-SCALE STRUCTURAL SIMILARITY FOR IMAGE QUALITY ASSESSMENT," in *Proceedings of the 37th IEEE Asilomar Conference on Signals, Systems and Computers*, 2003.

[176] C Li and A Bovic, "Three-component weighted structural similarity index," in *Proc. SPIE 7242, Image Quality and System Performance VI,* vol. 7242, 2009.

[177] M D Fairchild and G M Johnson, "iCAM framework for image appearance, differences, and quality," *Journal of Electronic Imaging*, vol. 13, no. 1, pp. 126-138, January 2004.

[178] X Zhang, D Silverstein, J Farrell, and B Wandell, "Color Image Quality Metric S-CIELAB and Its Application on Halftone T exture Visibilit y".

[179] F Ebner and M D Fairchild, "Development and Testing of a Color Space with Improved Hue Uniformity," in *Conference: Proceedings of the 6th Color Imaging Conference*, 1998.

[180] N Moroney, "Local Colour Correction Using Non-Linear Masking," in *IS&T/SID Eigth color imaging conference*, 2000.

[181] Z Wang, A C Bovik, and E P Simoncelli, "Chapter 8.3: Structural Approaches to Image Quality Assessment," in *Handbook of Video and Image Processing*, 2nd ed., A Bovik, Ed. Burlington, USA: Elsevier Academic Press, 2005, pp. 961-974.

[182] Z Wang, A C Bovik, H R Sheikh, and E P Simoncelli. The SSIM Index for Image Quality Assessment. [Online]. https://ece.uwaterloo.ca/~z70wang/research/ssim/

[183] M Orfanidou, S Triantaphillidou, and E Allen, "Predicting Image Quality using a Modular Image Difference Model," in *Image Quality and System Performance V*, vol. 6808, 2008.

[184] Krishnan S. Khademi A1, "Comparison of JPEG 2000 and Other Lossless Compression Schemes for Digital Mammograms.," in *Conf Proc IEEE Eng Med Biol Soc.* , 2005, pp. 3771-4.

[185] H Oh, A Bilgin, and M Marcellin, "Visually Lossless JPEG 2000 for Remote Image Browsing," *Information*, vol. 7, no. 3, p. 45, July 2016.

[186] Muhammad Safdar, Ming Ronnier Luo, and Xiaoyu Liu, "Performance Comparison of JPEG, JPEG 2000 and newly adopted CSI-JPEG by adopting different color models," *Color Research and Application*, November 2016.

[187] T Richter, A Artusi, and T Ebrahimi, "JPEG XT: A New Family of JPEG Backward Compatible Standards," *IEEE Multimedia*, vol. 23, no. 3, July 2016.

[188] H Alers, J Redi, H Liu, and I Heynderick, "Studying the effect of optimizing image quality in salient regions at the expense of background content," *Journal of Electronic Imaging*, vol. 22, no. 4, 2013.

[189] R Achanta, S Hemami, F Estrada, and S Susstrunk, "Frequency-tuned Salient Region Detection," in *IEEE International Conference on Computer Vision and Pattern Recognition*, pp. 1597-1604.

[190] A Haun and Eli Peli, "Is Image Quality a function of contrast perception?," in *SPIE Conference 8651 Human Vision and Electronic Imaging XVIII*, vol. 8651.

[191] S Triantaphillidou, J Jarvis, and G Gupta, "Spatial Contrast Sensitivity and Discrimination in Pictorial Images," in *Proc. SPIE 9016, Image Quality and System Performance XI,* 2014.

[192] D Chandler, M M Alam, and T D Phan, "Seven Challenges for Image Quality Research (Keynote paper)," in *Human Vision and Electronic Imaging XIX*, vol. 9014, 2014.

[193] International Electrotechnical Committee. (1998) Part 2-1: Default RGB Colour Space - sRGB.

[194] Jenkin R, "Chapter 7: Images and Image Formation," in *The Manual of Photography*, 10th ed., E J Allen and S Triantaphillidou, Eds. UK: Focal Press, 2010.

[195] A C Bovik, *Handbook of Image and Video Processing*, 2nd ed. Orlando, Florida, USA: Elsevier Academic Press, Inc, 2005.

Appendices

Appendix A Test Images Chapter 5

Final set of 25 images, used in the thresholds experiment (captured by the author)

01_accordion.tif

02_Afternoon_Tea.tif

03_beach goods.tif

04_bride.tif

05_cliffs.tif

06_Crockery.tif

07_Crown_Antiques.tif

08_Emporium.tif

09_Flags.tif

10_formal.tif

11_Fred.tif

12_Hive Beach.tif

13_huddle.tif

14_kids.tif

15_Lamp.tif

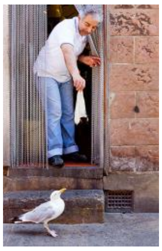
16_Lilies.tif

17_Marle Sculpture.tif

18_pink flowers.tif

19_Players Navy.tif

20_Pool.tif


21_Seagull.tif

22_Serpent.tif

23_Flower Garden.tif

24_stones_II_.tif

25_Summer.tif

Appendix B Display Characterisation

Display characterisation of the CG245W was performed by Jae Young-Park. Figures, tables and captions are from [89]

10.4.1.1 Devices

Display: EIZO CG245W
PC: DELL Optiplex 760 with an ATI Radeon HD 3450 graphics card
Calibrator: GretagMacbeth i1Pro display calibrator & Built-in calibration sensor
Colorimeter: Konica-Minolta CS-200 (Field of view was set to 0.2 degree)

10.4.1.2Environmental conditions

Temperature: 20 degrees Celsius
Relative humidity: N/A
Illumination condition: Total darkness
Warm up time: 1 hour (calibrated before the measurement of each characteristic)
Object distance: 150 cm (Effective screen height: 32.4 cm, width: 51.84cm)

Display calibration
D_{65}, gamma of 2.2, 120 cd/m²

	EIZO CG245W
Displayable area (cm)	51.8 (H) x 32.4 (V)
Native pixel resolution	1920(H) x 1200(V)
Display colour	24bits (DVI) / 30bits (DP) from a palette of 48bits
Viewing angle (°)	178 (H), 178 (V)
Pixel pitch	0.27mm (H), 0.27mm (V)
Maximum brightness	270cd/m²
Maximum brightness for calibration and experiments	120cd/m²
Colour representation	sRGB

Table B1.Technical specifications of display device and the settings used during calibration and experiments.

	CG245W				
	X'	Y'	Z'	u'	v'
Peak red	41.45	21.39	2.44	0.448	0.521
Peak green	37.43	73.18	11.98	0.128	0.562
Peak blue	19.29	7.61	99.56	0.179	0.158
Peak white	95.59	100.00	113.49	0.197	0.465

Table B2 CIE 1931 tristimulus values and CIE 1976 chromaticity coordinates for the peak colours and the peak white from the display device.

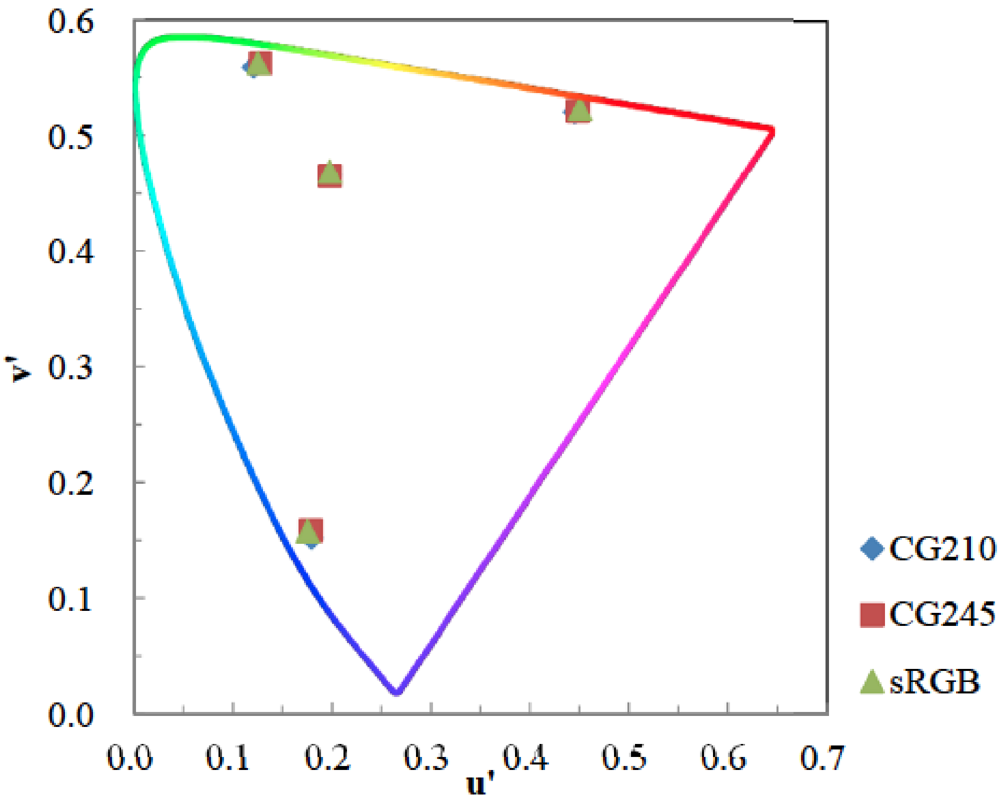


Figure B1: Reproduction of the peak patches on display devices and their corresponding values in sRGB colour space, plotted on u’, v’ diagram.

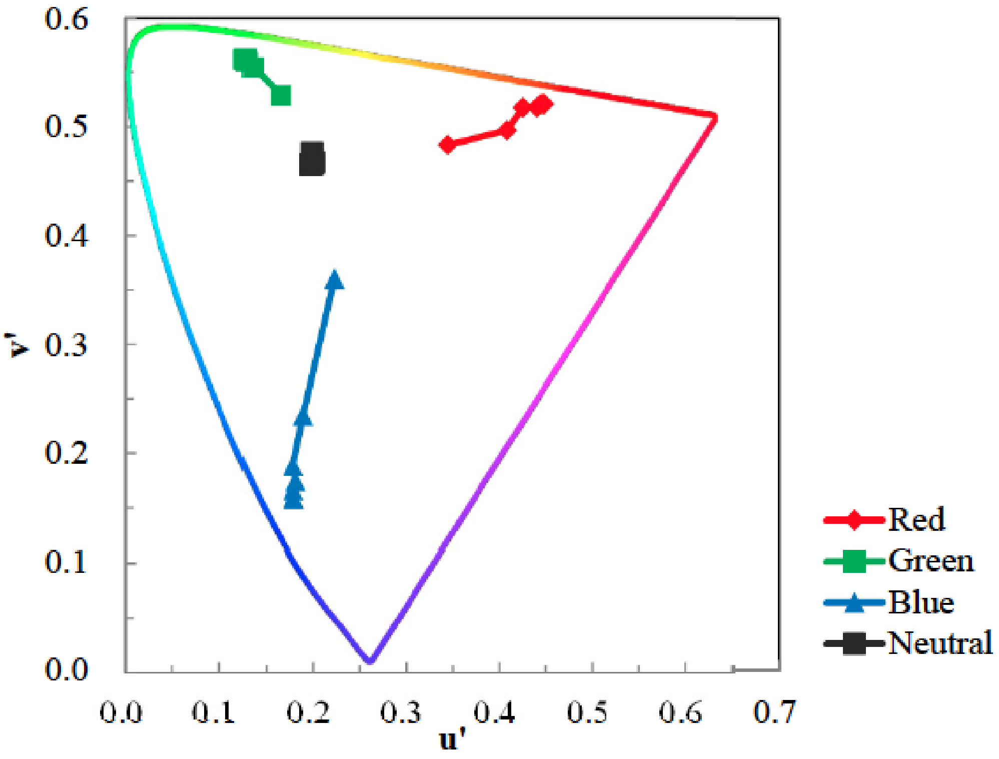


Figure B2: Colour tracking characteristics of the EIZO CG245W. Reproduced primary colours and neutral patches were plotted on u’, v’ diagram.

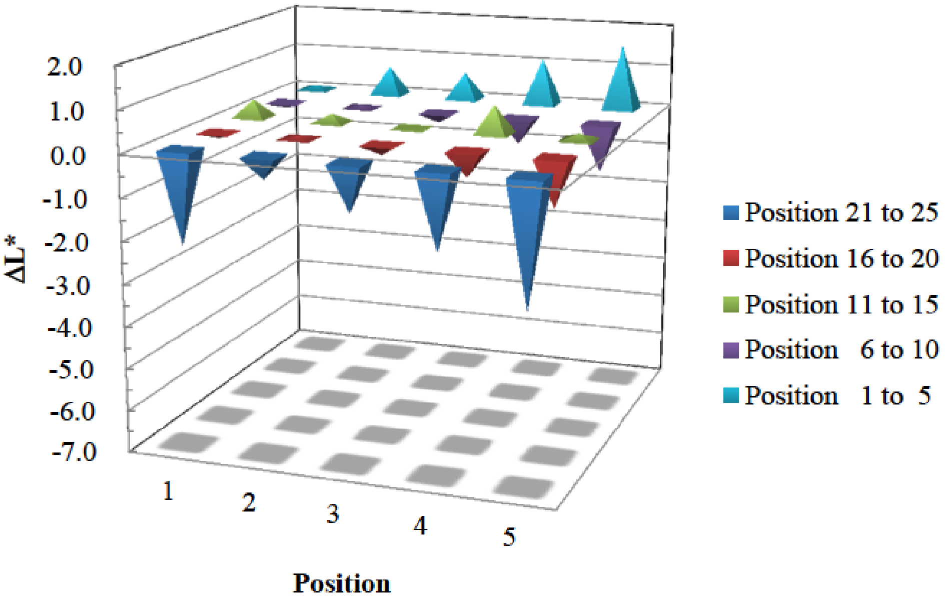


Figure B3: Positional non-uniformity. Lightness L^*_{ab} differences, from the reference point to the measured points across the screen for the CG245W.

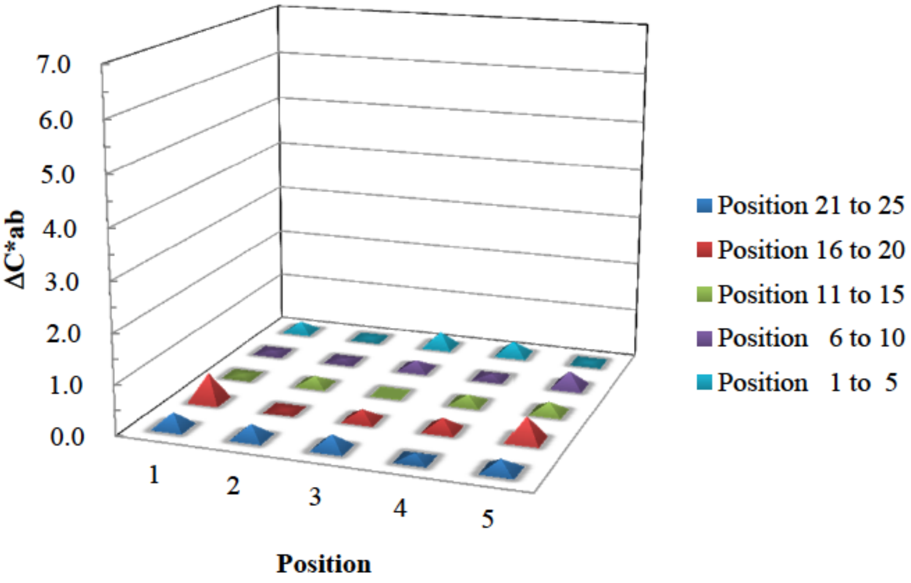


Figure B4: Chromatic differences,$\Delta C^*{}_{ab}$ from the reference point to the measured points across the screen.

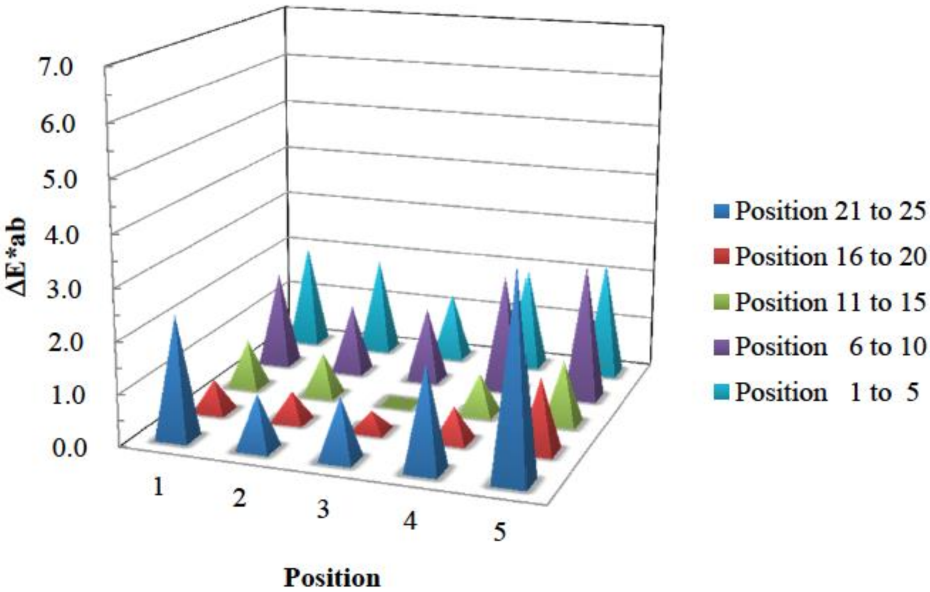


Figure B5: Figure 3-14. Colour differences, $\Delta E^*{}_{ab}$ from the reference point to the measured points across the screen. The CG210 (top) and the CG245W (bottom).

Display Model	Black Background			White Background			ΔE^{*}_{ab}	ΔE^{*}_{00}
	L*	a*	b*	L*	a*	b*		
CG245W	106.79	-8.81	-8.12	106.73	-8.85	-8.26	0.16	0.18

Table B3: Dependency on background Measured CIELAB values and evaluated colour differences

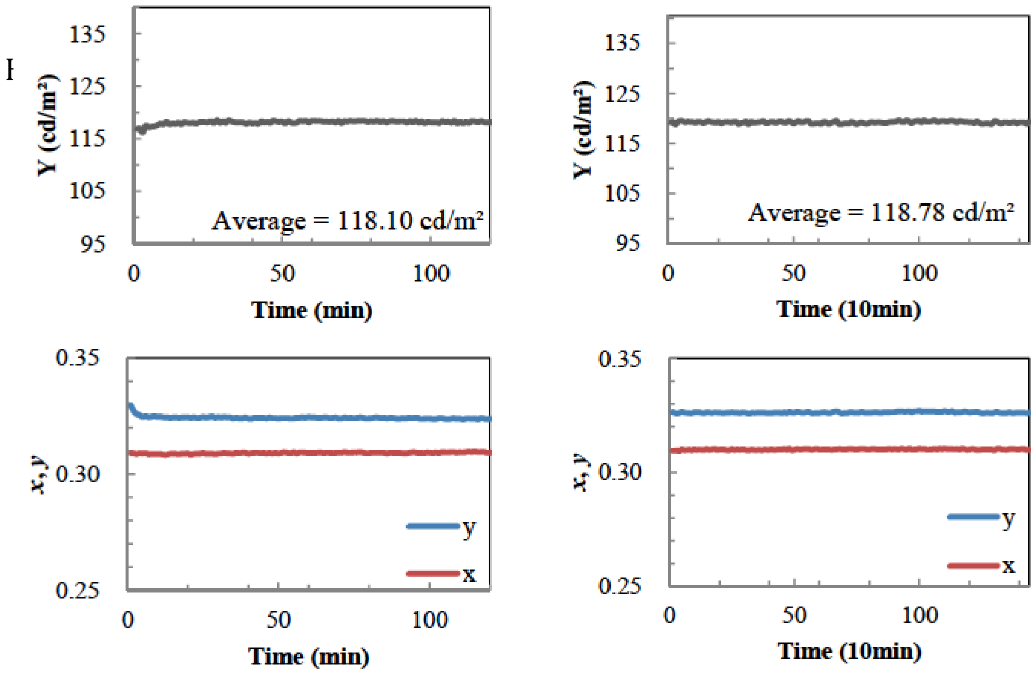


Figure B6 Temporal Stability: Short-term stability (left) and mid-term stability (right) in luminance (top) and x, y chromaticity coordinates (bottom)

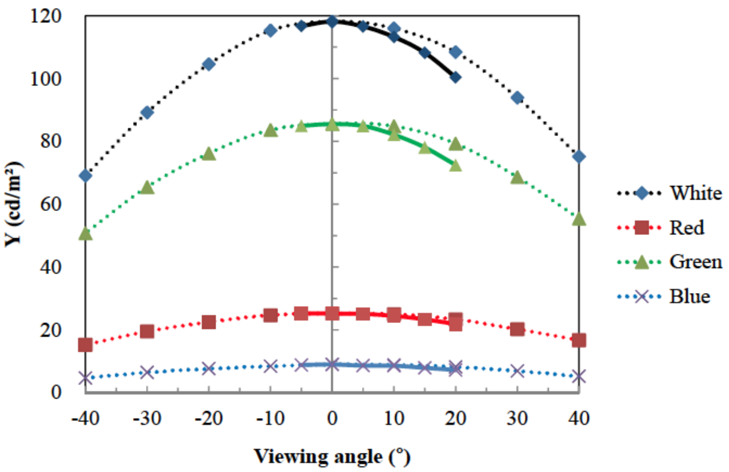


Figure B7: Viewing Angle Dependency: Luminance output of the peak colours and the peak white at the various horizontal and vertical viewing angles. Solid lines represent vertical luminance and broken lines represent horizontal luminance.

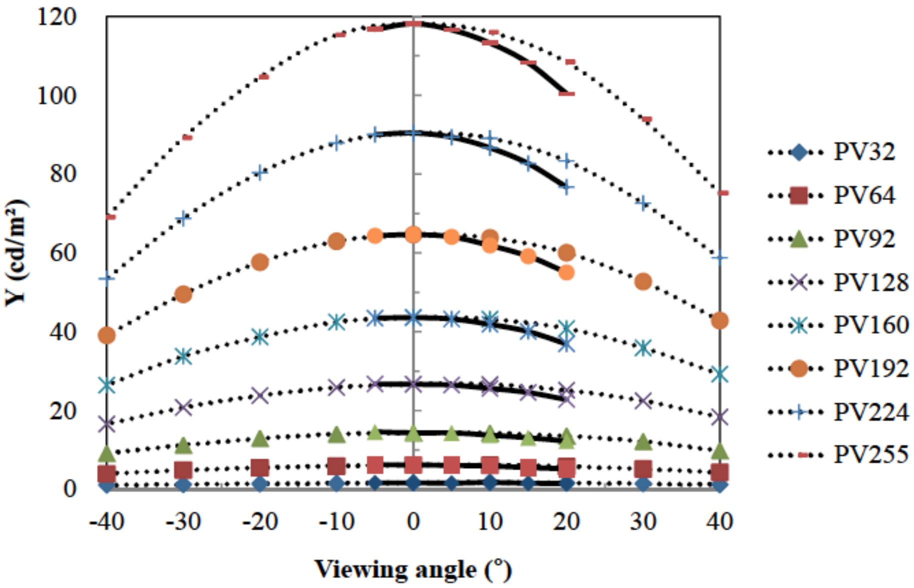


Figure B8: Changes in luminance output of neutral patches at various horizontal and vertical viewing angles. Solid lines represent vertical luminance and broken lines represent horizontal luminance.

Appendix C Polynomial Functions for JND filters

filter	**K**	**Polynomial**	**Fit**	**Cut off**
1	0.0136	$y = -63.02367191x^5 + 113.97895044x^4 - 52.00773711x^3 + 11.07355158x^2 - 1.40184154x + 1.01064112$	$R^2 = 0.99993$	0.750
2	0.0188	$y = -58.73796286x^5 + 93.65961602x^4 - 39.39053803x^3 + 8.08615063x^2 - 1.41755594x + 1.00718922$	$R^2 = 0.99991$	0.750
3	0.0221	$y = -53.36260637x^5 + 77.60844792x^4 - 29.97326990x^3 + 5.93005133x^2 - 1.41102187x + 1.00478963$	$R^2 = 0.99989$	0.750
4	0.0249	y = -46.38712024x5 + 61.07951578x4 - 20.65973874x3 + 3.85101399x2 - 1.38793134x + 1.00254428	$R^2 = 0.99987$	0.750
5	0.0275	$y = -15.87321704x^6 - 1.81634528x^5 + 12.81190066x^4 + 0.99507901x^3 - 0.41046558x^2 - 1.20759534x + 0.99859113$	$R^2 = 0.99982$	0.750
6	0.0300	$y = -24.46936163x^5 + 19.36492875x^4 + 1.79736436x^3 - 1.01817742x^2 - 1.26310970x + 0.99746839$	$R^2 = 0.99936$	0.750
7	0.0324	$y = -5.54745780x^5 - 12.22736986x^4 + 18.38422331x^3 - 4.56494292x^2 - 1.11089337x + 0.99381815$	$R^2 = 0.99909$	0.750
8	0.0348	$y = 171.59217189x^6 - 348.50653465x^5 + 244.99265044x^4 - 72.27251211x^3 + 10.18214873x^2 - 2.13592599x + 1.00344549$	$R^2 = 0.99925$	0.750
9	0.0372	$y = 170.37909143x^6 - 320.00123012x^5 + 205.87423181x^4 - 54.28397739x^3 + 6.82938134x^2 - 2.03492146x + 1.00109897$	$R^2 = 0.99972$	0.700
10	0.0396	$y = 150.14490436x^6 - 261.29275147x^5 + 151.67613740x^4 - 33.86913584x^3 + 3.54347126x^2 - 1.96584090x + 0.99936427$	$R^2 = 0.99993$	0.650
11	0.0421	y =293.09375575x6 - 487.20690727x5 + 284.63700094x4 - 70.37121929x3 + 8.12727343x2 - 2.30094886x + 1.00079378	$R^2 = 0.99985$	0.650
12	0.0447	$y = 312.13692010x^6 - 482.57568309x^5 + 261.06957064x^4 - 58.70041652x^3 + 6.08424398x^2 - 2.30824397x + 0.99971378$	$R^2 = 0.99993$	0.600
13	0.0474	$y = 271.12623183x^6 - 386.11268679x^5 + 187.99409745x^4 - 35.49621618x^3 + 2.94492612x^2 - 2.28999946x + 0.99865599$	$R^2 = 0.99999$	0.550
14	0.0502	$y = 413.61667631x^6 - 569.88146769x^5 + 276.55807074x^4 - 55.36316921x^3 + 5.01059963x^2 - 2.50758296x + 0.99903828$	$R^2 = 0.99998$	0.550
15	0.0532	$y = 525.96307272x^6 - 683.02096555x^5 + 315.03241258x^4 - 60.25709956x^3 + 5.16600934x^2 - 2.64675302x + 0.99888438$	$R^2 = 0.99998$	0.500
16	0.0563	$y = 640.76948777x^6 - 786.90238084x^5 + 345.52419406x^4 - 63.00578501x^3 + 5.12505871x^2 - 2.79038092x + 0.99874452$	$R^2 = 0.99998$	0.475
17	0.0597	$y = 782.84345556x^6 - 908.72633777x^5 + 379.73577360x^4 - 65.94453657x^3 + 5.09554349x^2 - 2.94624244x + 0.99862845$	$R^2 = 0.99999$	0.450
18	0.0634	$y = 972.55765608x^6 - 1,066.81318725x^5 + 424.45010739x^4 - 70.30541384x^3 + 5.18628784x^2 - 3.11974053x + 0.99854559$	$R^2 = 0.99999$	0.425
19	0.0673	$y = 1,244.66990568x^6 - 1,290.08726435x^5 + 489.12534847x^4 - 77.53331643x^3 + 5.50068036x^2 - 3.31573077x + 0.99849722$	$R^2 = 0.99999$	0.400
20	0.0716	$y = 1,630.55986946x^6 - 1,594.10481012x^5 + 575.06007293x^4 - 87.17135452x^3 + 5.96076636x^2 - 3.53371561x + 0.99846501$	$R^2 = 1.00000$	0.375
21	0.0764	$y = 2,190.96306548x^6 - 2,015.03849800x^5 + 689.79860578x^4 - 99.77887920x^3 + 6.57709873x^2 - 3.77689682x + 0.99844368$	$R^2 = 1.00000$	0.350
22	0.0817	$y = 2,951.70920543x^6 - 2,541.85617308x^5 + 821.85995324x^4 - 112.70265793x^3 + 7.13422100x^2 - 4.04407872x + 0.99841980$	$R^2 = 1.00000$	0.325
23	0.0876	$y = 3,934.89195086x^6 - 3,156.10487854x^5 + 959.44669135x^4 - 123.82055112x^3 + 7.50652718x^2 - 4.33886522x + 0.99839293$	$R^2 = 1.00000$	0.300
24	0.0944	$y = 5,026.04401925x^6 - 3,745.39193664x^5 + 1,070.71538290x^4 - 129.48954970x^3 + 7.55646348x^2 - 4.66789975x + 0.99836925$	$R^2 = 1.00000$	0.275
25	0.1022	$y = 5,074.91891260x^6 - 3,486.23657818x^5 + 943.66211985x^4 - 104.97355783x^3 + 6.03273045x^2 - 5.01914562x + 0.99833226$	$R^2 = 1.00000$	0.275
26	0.1112	$y = 10,602.04308614x^6 - 6,820.58270544x^5 + 1,717.86413763x^4 - 184.10790689x^3 + 9.97630288x^2 - 5.53319137x + 0.99837552$	$R^2 = 1.00000$	0.250
27	0.1219	$y = 20,488.25552529x^6 - 12,043.17609474x^5 + 2,770.41229761x^4 - 275.48433999x^3 + 13.86394645x^2 - 6.11630522x + 0.99838761$	$R^2 = 1.00000$	0.225
28	0.1347	$y = 38,313.19862711x^6 - 20,282.61222064x^5 + 4,218.79235047x^4 - 383.00404814x^3 + 17.83682373x^2 - 6.79752130x + 0.99838740$	$R^2 = 1.00000$	0.200
29	0.1504	$y = 75,063.37910175x^6 - 35,875.83039594x^5 + 6,768.47474611x^4 - 563.42843673x^3 + 24.28069130x^2 - 7.64657265x + 0.99839239$	$R^2 = 1.00000$	0.175
30	0.1697	$y = 7,405.54025388x^6 - 2,221.48715901x^5 + 797.49897709x^4 - 61.92697426x^3 + 6.66109728x^2 - 8.38018151x + 0.99838364$	$R^2 = 1.00000$	0.150

Appendix D Observer Instructions for the SCQR

Instructions for softcopy ruler experiment

Thank you for participating in today's evaluation.

In this experiment, you will be assessing the overall quality of a series of images using a psychophysical technique called the softcopy quality ruler. Please remember there are no right or wrong answers; image quality is defined by observer perception, which varies among individuals. We are interested in your personal impression.

When you first enter the opening screen, please click the start button and enter your name in the dialogue box that opens. This will save your results in a text file.

Here is how we are asking you to evaluate the test images:

a) A pair of images will be presented on the monitor in front of you. The image on the left is labelled 'Ruler Image' and the image on the right is labelled 'Test Image'. For each test image on the right, we ask you to adjust the ruler image on the left so that the quality of the two is matched.

b) The test images shown on the right represent different amounts of compression artifacts. The distortion is in the form of localized blurring artifacts, ringing (which appears as halos) and other areas of texture distortion. During this session you will be evaluating between 6 and 8 different levels of compression in each of sixteen different scenes.

c) You will be comparing each test image on the right with a series of ruler images on the left, which can be varied by moving the slider bar. These ruler images differ only in sharpness. You will be balancing the quality loss due to unsharpness in the ruler images to the quality loss due to compression in the test images. When you are comparing test and ruler images, ask yourself which image you would keep if this were a treasured image and you were allowed only one copy. If you prefer the test image, then you should move the slider bar to the left for a sharper ruler image. If, instead, you prefer the ruler image, then you should move the slider bar to the right for a more blurred ruler image. When you have finished adjusting the ruler, the two images should be equal in your preference. Your response will be recorded when you press the 'Next' button.

Appendix E Colour Space Transformations

Conversion from sRGB to 1931 CIE XYZ values [193]**:**

sRGB values are first normalised by dividing all values by the maximum level.

They are then transformed to linear sRGB values using the following transfer functions:

(0.1

Let C denote R, G or B in the following:

$$If\ C_{srgb} \leq 0.4045$$

$$C_{linear\ sRGB} = \frac{C_{srgb}}{12.92}$$

else:

$$C_{linear\ sRGB} = \left(\left(C_{srgb} + 0.055\right)/1.055\right)^{2.4}$$

Linear sRGB values are then transformed to 1931 CIE XYZ values as follows:

$$\begin{bmatrix} X \\ Y \\ Z \end{bmatrix} = \begin{bmatrix} 0.4124 & 0.3576 & 0.1805 \\ 0.2126 & 0.7152 & 0.0722 \\ 0.0193 & 0.1192 & 0.9505 \end{bmatrix} \begin{bmatrix} R_{srgb} \\ G_{srgb} \\ B_{srgb} \end{bmatrix}$$

Conversion from 1931 CIE XYZ values (adapted to a D65 white point) to IPT opponent colour space [70]**:**

Step 1: convert XYZ to LMS:

$$\begin{bmatrix} L \\ M \\ S \end{bmatrix} = \begin{bmatrix} 0.4002 & 0.7075 & -0.0807 \\ -0.2280 & 1.1500 & 0.0612 \\ 0.0 & 0.0 & 0.9184 \end{bmatrix} \begin{bmatrix} X_{D65} \\ Y_{D65} \\ Z_{D65} \end{bmatrix}$$

Step 2: Apply transfer curves:

Let C denote L, M, or S in the following:

$$C' = {C_{linear}}^{0.43};\ If\ C_{linear} \geq 0$$

$$else\ C' = -|C_{linear}|^{0.43}$$

Step 3: conversion from LMS to IPT colour space (LMS values may be linear or non-linear L',M',S')

$$\begin{bmatrix} I \\ P \\ T \end{bmatrix} = \begin{bmatrix} 0.4000 & 0.4000 & 0.2000 \\ 4.4550 & -4.8510 & 0.3960 \\ 0.8056 & 0.3572 & -1.1628 \end{bmatrix} \begin{bmatrix} L \\ M \\ S \end{bmatrix}$$